Instituto Superior
de Engenharia
Politécnico de Coimbra

DEPARTMENT OF SYSTEMS AND COMPUTER ENGINEERING

Face Morphing Detection in Identification Documents

Project Report to fulfill the Master's degree in Informatics Engineering

Specialization in Software Engineering

Author

Pedro Jorge Fernandes Morais

Supervisor

Inês Campos Monteiro Sabino Domingues

Co-Supervisor

Teresa Raquel Corga Teixeira da Rocha

INSTITUTO POLITÉCNICO
DE COIMBRA

INSTITUTO SUPERIOR
DE ENGENHARIA
DE COIMBRA

Coimbra, December 2025

Resumo

Os ataques de *morphing* facial constituem uma ameaça crítica aos sistemas de autenticação biométrica, permitindo a criação de imagens compostas que correspondem a múltiplas identidades. Este trabalho investiga a deteção de ataques de *morphing* (MAD) baseada em *deep learning*, com foco nas arquiteturas *Vision Transformer* (ViT) e na interpretabilidade através de *heatmaps* baseados em oclusões.

Uma revisão sistemática de 23 estudos evidenciou limitações recorrentes: reduzida disponibilidade de código, análise de interpretabilidade insuficiente e dificuldades na generalização entre diferentes *datasets*. Estes resultados motivaram cinco experiências que avaliaram estratégias de pré-processamento, composição de dados, equilíbrio de classes e robustez entre *datasets*, incluindo uma experiência com um extrator ViT seguido de um classificador *Support Vector Machine* (SVM).

Os resultados demonstram que o isolamento da face melhora consistentemente o desempenho, e que a diversidade dos dados aumenta a generalização. Embora o modelo ViT tenha alcançado precisão quase perfeita em avaliações *intra-dataset*, o desempenho degradou-se substancialmente em conjuntos de dados não incluídos no treino, evidenciando o impacto da mudança de domínio. A inclusão de técnicas avançadas de *morphing* durante o treino melhorou a deteção de ataques mais simples, enquanto que a sua não inclusão proporcionou uma generalização limitada.

Para responder aos requisitos de interpretabilidade em contextos críticos, este trabalho propõe uma plataforma de análise padronizada baseada em oclusões, combinando estratégias semânticas e espaciais. Aplicados ao modelo ViT-SVM, os *heatmaps* indicam que as regiões do nariz e da boca são especialmente relevantes, com o modelo a focar-se em áreas de transição de profundidade. A plataforma é independente do modelo e reproduzível, oferecendo uma abordagem sistemática para identificar regiões faciais que impactam o desempenho do classificador.

Em suma, este trabalho valida arquiteturas ViT para MAD, analisa os fatores chave que influenciam a robustez dos modelos e desenvolve ferramentas de interpretabilidade para implementar sistemas de segurança biométrica fiáveis.

Palavras-chave: Deteção de Ataques de Morphing, Morphing Facial, Deep Learning, Interpretabilidade, Heatmaps.

Abstract

Face morphing attacks pose a critical security threat to biometric authentication systems by enabling the creation of composite facial images that can be matched to multiple identities. This work investigates deep learning-based Morphing Attack Detection (MAD), with a focus on Vision Transformer (ViT) architectures and on improving model interpretability through occlusion-based heatmap analysis.

A systematic review of 23 recent studies highlighted several persistent research limitations: scarce code availability hindering reproducibility, insufficient interpretability analysis, and ongoing challenges in cross-dataset generalization. These findings informed five experiments evaluating preprocessing strategies, dataset composition, class balancing, and cross-dataset robustness, with one experiment specifically employing a ViT feature extractor coupled with a Support Vector Machine (SVM) classifier.

The results show that face cropping consistently improves detection performance, and that dataset diversity enhances generalization capability. While the ViT-based model achieved near-perfect accuracy in intra-dataset evaluations, performance degraded substantially on unseen datasets, underscoring the impact of domain shift. Additionally, training on advanced morphing techniques improved detection of simpler ones, whereas training on more simple techniques offered limited transferability.

To address interpretability requirements in security-critical contexts, this work proposes a standardized occlusion-based analysis framework that combines landmark-based (semantic) and grid-based (spatial) strategies. When applied to the ViT-SVM model, the resulting heatmaps indicate that the nose and mouth regions are especially influential for detection decisions, with the model focusing on depth-transition areas. The framework is model-agnostic and reproducible, offering a clear and systematic approach to identifying facial regions that contribute to MAD classifications.

Overall, this work advances the field by validating ViT architectures for MAD, analyzing key factors affecting detection robustness, and developing interpretability tools essential for the deployment of trustworthy deep learning-based biometric security systems.

Keywords: Morphing Attack Detection, Face Morphing, Deep Learning, Interpretability, Heatmaps.

Agradecimentos

Gostaria de expressar a minha profunda gratidão a todas as pessoas que, de diferentes formas, contribuíram para a realização deste projeto e para o meu percurso académico.

Em primeiro lugar, agradeço à minha orientadora, Professora Doutora Inês Campos Monteiro Sabino Domingues, pelo acompanhamento constante, pela orientação rigorosa e pelo apoio contínuo ao longo de todo este trabalho, cuja disponibilidade e visão foram fundamentais para a concretização deste projeto. Manifesto igualmente a minha gratidão à minha coorientadora, Professora Doutora Teresa Raquel Corga Teixeira da Rocha, pelo apoio e acompanhamento disponibilizados ao longo do percurso, e agradeço também ao Professor Doutor Simão Pedro Mendes Cruz Reis Paredes pela sua disponibilidade e atenção.

Gostaria também de expressar o meu agradecimento ao IEEE Access pelo apoio financeiro concedido para a publicação do artigo científico na revista. Agradeço igualmente ao comité do ISPA 2025 – 14th International Symposium on Image and Signal Processing and Analysis pelo apoio financeiro na submissão do artigo, assim como ao Instituto Superior de Engenharia de Coimbra (ISEC) pelo financiamento da minha participação na 31st Portuguese Conference on Pattern Recognition (RecPad 2025), permitindo a divulgação e discussão dos resultados deste projeto.

Não poderia deixar de agradecer à minha namorada, Beatriz, pelo apoio incondicional, pela compreensão nos momentos mais exigentes e por toda a ajuda prestada ao longo deste percurso; a tua presença tornou este caminho muito mais leve e motivador.

Aos meus colegas e amigos do curso, especialmente à Ágata e ao Rodrigo, agradeço o companheirismo, o apoio mútuo nas aulas e nos trabalhos de grupo, bem como os momentos de partilha que enriqueceram estes anos de formação. Estendo este agradecimento a todos os outros colegas que, de alguma forma, contribuíram para este percurso.

Por fim, um agradecimento muito especial à minha família e aos meus amigos pelo apoio constante, pela motivação contínua e por acreditarem sempre em mim. Sem vocês, este percurso não teria sido possível.

A todos, o meu sincero obrigado.

Contents

Resumo i

Abstract ii

Agradecimentos iii

Contents iv

List of Figures vii

List of Tables ix

List of Equations x

List of Abbreviations xi

1 Introduction 1
- 1.1 Morphing Attacks in the Context of Facial Manipulation 2
- 1.2 Research Objectives and Contributions 3
- 1.3 Work Schedule 4
- 1.4 Document Structure 5

2 State of the Art 7
- 2.1 Background Concepts 8
  - 2.1.1 D-MAD and S-MAD 8
  - 2.1.2 Bona Fide and Attack Images 9
  - 2.1.3 Evaluation Metrics 9
- 2.2 Literature Review Methodology 11
- 2.3 Literature Review Analysis 14
  - 2.3.1 Datasets Used 15
  - 2.3.2 Pre-Trained Models 16
  - 2.3.3 Architecture 17
  - 2.3.4 Experimental Setup 34
  - 2.3.5 Results 36
- 2.4 Findings and Trends 46

2.5 Challenges and Gaps . . . 48
2.6 Future Research Directions . . . 49
2.7 Conclusions . . . 51

3 Methodology . . . 53
3.1 Use of Publicly Available Code . . . 53
3.2 Vision Transformer Exploration . . . 54
3.3 Dataset Selection . . . 54
3.4 Dataset Preprocessing . . . 55
3.5 Model Explainability and Interpretability . . . 56
3.5.1 Occlusion Mapping Methodology . . . 58
3.5.2 Heatmap Visualization . . . 59
3.6 Evaluation Metrics . . . 60
3.6.1 Error Rates . . . 60
3.6.2 Explainability Tools . . . 60
3.7 Conclusions . . . 61

4 Experiments . . . 63
4.1 Experiment 1 — Fang et al. (2022) . . . 64
4.1.1 Datasets . . . 64
4.1.2 Class Balancing . . . 65
4.1.3 Evaluation Metrics . . . 65
4.1.4 Experimental Setup . . . 65
4.1.5 Experimental Results . . . 65
4.1.6 Conclusions . . . 68
4.2 Experiment 5 — Zhang et al. (2024) . . . 69
4.2.1 Datasets . . . 70
4.2.2 Class Balancing . . . 70
4.2.3 Occlusion Mapping . . . 70
4.2.4 Evaluation Metrics . . . 71
4.2.5 Experimental Setup . . . 72
4.2.6 Experimental Results . . . 72
4.2.7 Conclusions . . . 78

5 Occlusion-based Interpretability . . . 79
5.1 Method Overview . . . 79
5.2 Technical Implementation and Features . . . 80
5.2.1 Configurable Landmark Precision . . . 80
5.2.2 Configurable Grid Resolution . . . 81
5.2.3 Flexible Heatmap Modes . . . 81
5.3 Application to MAD Model Interpretability . . . 82

5.3.1 Heatmap Examples . . . 82
5.3.2 Heatmap Improvements in the Standardized Framework . . . 83
5.4 Discussion . . . 84
5.4.1 Impact of Preprocessing and Class Balancing . . . 84
5.4.2 Insights from Selected Landmarks . . . 85
5.4.3 Methodological Considerations . . . 87
5.5 Conclusions . . . 90

6 Conclusions . . . 91
6.1 Summary of Contributions . . . 91
6.1.1 Literature Review and Research Gap Identification . . . 91
6.1.2 Experimental Findings . . . 92
6.1.3 Standardized Interpretability Framework . . . 94
6.2 Future Work . . . 95
6.2.1 Extended Landmark Analysis . . . 95
6.2.2 Comparative Interpretability Studies Across Architectures . . . 95
6.2.3 Geometry-Aware Detection Methods . . . 95
6.2.4 Domain Adaptation and Generalization . . . 96
6.2.5 Advanced Morphing Techniques and Adversarial Robustness . . . 96
6.2.6 Cross-Domain Detection: Morphing Attacks and Deepfakes . . . 96
6.2.7 Operational Deployment Considerations . . . 97
6.2.8 Standardization and Benchmarking . . . 97
6.3 Final Remarks . . . 98

Bibliography . . . 99

Annexes . . . 112
Annex A - IEEE Access Journal Paper . . . 113
Annex B - ISPA International Conference Paper . . . 144
Annex C - RecPad Paper . . . 151
Annex D - State of the Art: Tables . . . 154
Annex E - Selected Datasets Sample Images . . . 158
Annex F - Unsuccessful Experiments . . . 161
Annex G - Experiment 1: Experimental Code . . . 164
Annex H - Experiment 5: Experimental Code . . . 167
Annex I - Experiment 5: Results . . . 171
Annex J - Occlusion-based Interpretability Method: Source Code . . . 172

List of Figures

1.1 Example of a morphed image generated using FaceShape face morph tool. 1
1.2 Gantt diagram illustrating the temporal organization of tasks and deliverables. 5

2.1 Proportion of papers employing S-MAD and D-MAD techniques. 9
2.2 Methodology applied in the literature review. 12
2.3 Distribution of papers by year of publication. 12
2.4 Distribution of selected papers by average citations per year. 13
2.5 Citation count per year, ordered by total number of citations. 13
2.6 Number of papers using each dataset type and their overlap. 15
2.7 Distribution of pre-trained models used in the reviewed literature. . . . 16
2.8 Distribution of papers in the reviewed literature by architecture type. . . 17
2.9 CNN architecture. 18
2.10 Autoencoder architecture. 22
2.11 GAN architecture. 23
2.12 Siamese network architecture. 25
2.13 ViT architecture. 28

3.1 Methodology workflow: overview. 53
3.2 Methodology workflow: literature review flowchart. 54
3.3 Methodology workflow: experiments flowchart. 55
3.4 Example of dataset preprocessing applied to a sample image. 57
3.5 Methodology workflow: interpretability flowchart. 57
3.6 Overview of the occlusion mapping methodology. 58
3.7 Three heatmap visualizations from occlusion mapping. 61

4.1 Experiment 1: methodology overview. 64
4.2 Experiment 1: datasets. 65
4.3 Experiment 1: train flowchart. 66
4.4 Experiment 1: evaluation flowchart. 66
4.5 Evolution of D-EER across training epochs of the model trained and evaluated on the SMDD dataset. 66
4.6 Evolution of D-EER across training epochs of the model trained on the SMDD dataset and evaluated on the FRLL-Morphs dataset. 67

4.7 Sample images and D-EER from SMDD dataset. 68
4.8 Sample images and D-EER for each morphing technique from FRLL-Morphs dataset. 69
4.9 Experiment 5: methodology overview. 69
4.10 Experiment 5: datasets. 70
4.11 Experiment 5: train and evaluation flowchart. 71
4.12 Experiment 5: interpretability flowchart. 75
4.13 Visualization of occlusion heatmaps. 76
4.14 Visualization of outlier in occlusion heatmaps. 77

5.1 Landmark precision modes example. 80
5.2 Grid resolutions example. 81
5.3 Occlusion heatmaps generated by the standardized framework. 83

E.1 Sample images from the SMDD dataset. 158
E.2 Sample images from the FRLL dataset. 159
E.3 Sample images from the FRLL-Morphs dataset. 159
E.4 Sample images from FFHQ dataset. 160
E.5 Sample images from the FFHQ-Morphs dataset. 160

List of Tables

2.1 Training setups and strategies. 35
2.2 Best reported results on standardized MAD benchmarks. 44

3.1 Summary of datasets used. 56

4.1 Summary of experiments performed. 64
4.2 Performance, presented in terms of D-EER (%). 67
4.3 Intra-dataset testing results using Zhang et al. (2024) code. 73
4.4 Cross-dataset testing results using Zhang et al. (2024) code. 73

D.1 Major datasets referenced in the literature reviewed. 155
D.2 Major pre-trained models referenced in the literature reviewed. 156
D.3 Best results reported in each study included in the literature review. . . . 157
I.1 Results from experiments using Zhang et al. (2024) code. 171

List of Equations

2.1 APCER formula. 10
2.2 BPCER formula. 10
2.3 D-EER formula. 10
2.4 ACER formula. 11
2.5 Accuracy formula. 11

List of Abbreviations

3D-MAD	3D Morphing Attack Detection
ABC	Automated Border Control
ACC	Accuracy
ACER	Average Classification Error Rate
AMONOT	Augmented Morphing Network Optimization Tool
APCER	Attack Presentation Classification Error Rate
AUC	Area Under the Curve
bbox	Bounding Box
BM3D	Block Matching and 3D filtering
BOEP	Bologna Online Evaluation Platform
BPCER	*Bona Fide* Presentation Classification Error Rate
BSIF	Binarized Statistical Image Features
BW	Black and White
CAE	Convolutional Autoencoder
CCA	Canonical Correlation Analysis
cGAN	Conditional Generative Adversarial Network
CNN	Convolutional Neural Network
CPU	Central Processing Unit
D-EER	Detection Equal Error Rate
D-MAD	Differential Morphing Attack Detection
DET	Detection Error Tradeoff
DFT	Discrete Fourier Transform
DnCNN	Denoising Convolutional Neural Networks
DNN	Deep Neural Network
DWT	Discrete Wavelet Transform
EER	Equal Error Rate
FAR	False Acceptance Rate
FATE	Face Analysis Technology Evaluation
FFHQ	Flickr-Faces-HQ
FL	Federated Learning
FRGC	Face Recognition Grand Challenge
FRLL	Face Research London Lab
FRR	False Rejection Rate
FRVT	Face Recognition Vendor Test

GAN	Generative Adversarial Network
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
HOG	Histogram of Oriented Gradients
IEC	International Electrotechnical Commission
IEM	Interactive-Enhancement Module
ISO	International Organization for Standardization
ISPA	International Symposium on Image and Signal Processing and Analysis
KLD	Kullback-Liebler Divergence
KNN	K-Nearest Neighbors
LBP	Local Binary Patterns
MACER	Morphing Attack Classification Error Rate
MAD	Morphing Attack Detection
MBF	Multiresolution Bilateral Filtering
MLP	Multi-Layer Perceptron
MONOT	Morphing Network Optimization Tool
MS-CAN	Multi-scale Context Aggregation Network
MTCNN	Multi-task Cascaded Convolutional Networks
NIST	National Institute of Standards and Technology
OTF	On-The-Fly
P-CRC	Probabilistic Collaborative Representation Classifier
P&S	Print-Scanned
PMDB	Progressive Morphing Database
PW-MAD	Pixel-Wise Morphing Attack Detection
RecPad	Portuguese Conference on Pattern Recognition
RGB	Red Green Blue
ROC	Receiver Operating Characteristic
S-MAD	Single-image Morphing Attack Detection
SE	Squeeze-and-Excitation
SEM	Self-Enhancement Module
SENet	Squeeze-and-Excitation Network
SHAP	SHapley Additive exPlanations
SLERP	Spherical Linear Interpolation
SMDD	Synthetic Morph Detection Dataset
SOTA	State-of-the-Art
SOTAMD	State-of-the-Art Morphing Detection
SPL	Self-Paced Learning
SPL-MAD	Self-Paced Anomaly Detection for Morphing Attack Detection
SVM	Support Vector Machine
V-MAD	Video-based Morphing Attack Detection

ViT	Vision Transformer
WD	Wavelet Denoising

1 Introduction

The falsification of identification documents presents a threat to public security, personal data protection, and fraud prevention. A growing concern within this domain is the use of morphed facial images, a deceptive technique that has become more prevalent with the rapid advancement of artificial intelligence and image processing technologies [1, 2].

Morphed images are created by blending the facial features of two or more individuals to create a composite image that resembles all contributors, as illustrated in Figure 1.1. This manipulation enables multiple individuals to use a single identification document, for example, to gain unauthorized access across international borders [3, 4].

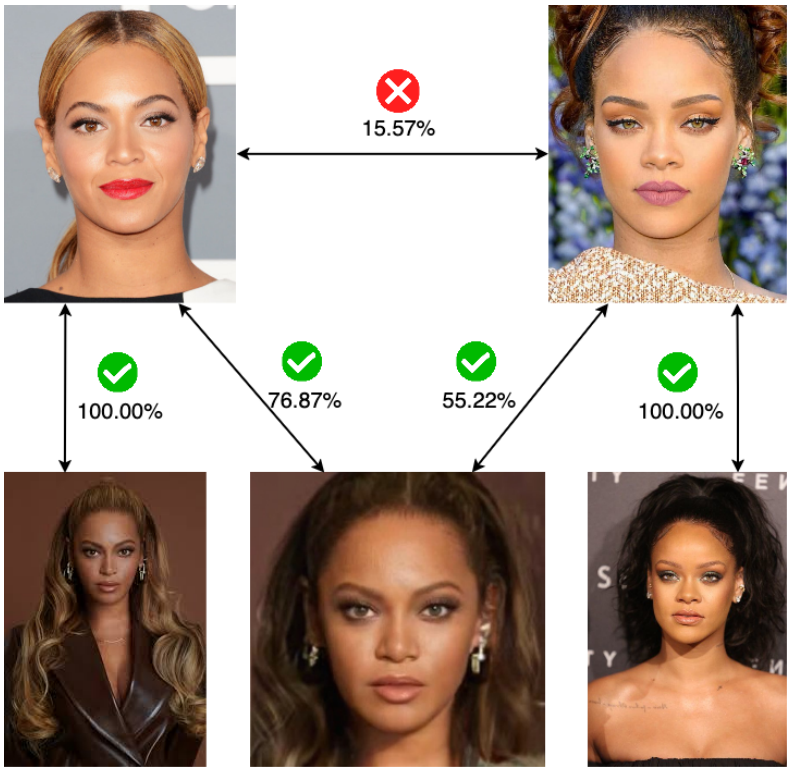


Figure 1.1: Example of a morphed image generated using the online FaceShape face morph tool available at [5] with source images [6, 7]. The resulting morph was compared to reference images [8, 9] using the FaceShape compare faces tool available at [10]. Additional icons used in the diagram are credited to [11, 12]. The methodology was inspired by the survey presented by Scherhag et al. (2019) [4].

Detecting these morphed images poses a considerable challenge. Humans are often unable to detect the manipulation, and even advanced facial recognition systems strug-

gle [13]. As the sophistication of morphing techniques continues to increase, attackers are able to produce more convincing morphs, making detection increasingly difficult. This requires ongoing research and innovation to develop more effective detection methods and keep up with the rapidly advancing techniques used in morphing attacks.

The practical implications of Morphing Attack Detection (MAD) techniques are particularly critical in the context of biometric security systems used at border checkpoints, automated identity verification in e-government services, and forensic analysis for criminal investigations. As morphed images undermine the reliability of facial recognition systems, robust MAD solutions are essential for preventing identity fraud, enforcing immigration policies, and supporting legal proceedings that rely on the integrity of biometric evidence. In border control scenarios, for example, the inability to detect a morphing attack can lead to unauthorized individuals crossing borders undetected, posing a risk to national and international security. In forensic analysis, undetected morphs may compromise the attribution of digital identities in criminal cases, which could affect both investigations and judicial outcomes. Therefore, the development and deployment of accurate and generalizable MAD techniques have practical consequences for public safety, legal accountability, and the trustworthiness of modern identity systems [14, 15].

1.1 Morphing Attacks in the Context of Facial Manipulation

The rapid advancement of artificial intelligence and generative models has enabled increasingly sophisticated forms of facial manipulation and synthetic media generation. While techniques such as deepfakes and face swapping have gained attention for creating synthetic media [16, 17], morphing attacks present a distinct challenge with different objectives and detection requirements.

Unlike deepfakes, which aim to impersonate specific individuals or create synthetic personas, morphing attacks create composite images that blend facial features from multiple individuals to produce a single image that can be verified as matching multiple different people by face recognition systems [18]. This fundamental difference in attack objective, targeting biometric authentication systems in high-security applications such as passport control, leads to distinct detection challenges. While deepfakes often exhibit temporal inconsistencies and boundary artifacts, morphing attacks produce subtle blending characteristics in single images [18, 19].

Despite these differences, both fields share common technical foundations in deep learning-based image forensics and artifact detection [16, 18]. This work focuses specifically on MAD, addressing the critical security challenges posed by morphing attacks

in biometric authentication contexts.

1.2 Research Objectives and Contributions

Given the security threats posed by morphing attacks and the challenges in detecting them, the present work investigates deep learning-based approaches for MAD, with a particular focus on understanding model behavior, improving detection performance, and developing interpretable systems suitable for deployment in security-critical applications. The research addresses three interconnected objectives:

1. **Systematic literature analysis**: Conduct a review of recent advances in deep learning-based MAD to identify current approaches, evaluate their effectiveness, and determine research gaps that require further investigation.
2. **Experimental validation and performance analysis**: Replicate and validate existing MAD approaches to evaluate the impact of preprocessing strategies, dataset composition, class balancing, and architectural choices on detection performance and generalization capability.
3. **Interpretability framework development**: Design and implement a standardized methodology for analyzing which facial regions contribute most to MAD model decisions, enabling transparent and trustworthy systems suitable for high-security applications.

To address these objectives, this work makes several interconnected contributions that advance the State-of-the-Art (SOTA) in MAD.

The first contribution is a comprehensive literature review that analyzed 23 recent studies on deep learning-based MAD, examining diverse approaches including Convolutional Neural Networks (CNNs), Autoenconders, Generative Adversarial Networks (GANs), Siamese Networks, Transformer-based models, and hybrid architectures. This systematic analysis revealed several findings that directly informed the experimental design and methodological choices presented in subsequent chapters. The literature review was published as a journal article in IEEE Access [20] (see Annex A for the full publication text).

Building on the insights from the literature review, the second contribution comprises a series of five experiments that systematically investigate factors affecting MAD system performance. These experiments examine the impact of preprocessing strategies (particularly face cropping), dataset composition and diversity, cross-dataset generalization capabilities, and class balancing techniques. The fifth experiment specifically applies interpretability analysis to understand which facial regions contribute most to model decisions. The experimental work provides empirical evidence on which design choices most influence detection accuracy and model robustness across different

morphing techniques and data distributions. The interpretability findings from Experiment 5 were accepted and presented at two conferences: the International Symposium on Image and Signal Processing and Analysis (ISPA) 2025 (see Annex B), and the Portuguese Conference on Pattern Recognition (RecPad) 2025 (see Annex C); at the date of submission of this document, these findings have been published in the ISPA 2025 proceedings[1], with the RecPad proceedings still pending publication.

To enable broader application of the interpretability techniques developed in Experiment 5, the third contribution is a generalized, reusable occlusion-based interpretability framework. This standardized methodology can be applied to any MAD model to systematically analyze which facial regions contribute most to detection decisions. The framework combines landmark-based (semantic) and grid-based (spatial) occlusion strategies to provide insights into model behavior, enabling researchers to develop transparent and trustworthy MAD systems suitable for security critical applications. The framework is fully documented with demonstration notebooks (see Annex J).

1.3 Work Schedule

To address the research objectives outlined in the previous section, the work was organized into five interconnected tasks, each producing specific deliverables that contributed to the overall research goals. Figure 1.2 presents the temporal organization of these tasks and their associated deliverables throughout the research period.

The research workflow comprised the following major tasks:

- **T1 - Literature Analysis**: Comprehensive review of recent literature on deep learning-based MAD techniques, examining methodologies, datasets, evaluation protocols, and identifying research gaps.
- **T2 - Dataset Analysis and Selection**: Systematic evaluation and selection of publicly available datasets suitable for training and evaluating MAD models, considering factors such as morphing techniques, image quality, and dataset diversity.
- **T3 - Experimental Validation**: Replication and extension of existing MAD approaches using publicly available implementations, investigating the impact of preprocessing strategies, dataset composition, and cross-dataset generalization.
- **T4 - Interpretability Framework Development**: Design and implementation of a standardized occlusion-based interpretability framework for analyzing which facial regions contribute most to MAD model decisions.
- **T5 - Thesis Writing**: Integration of all research findings, experimental results, and methodological contributions into the final thesis document.

[1] https://ieeexplore.ieee.org/document/11259485

These tasks produced six key deliverables that document the research contributions:

- **D1 - Journal Article (IEEE Access)**: Submission of the literature review manuscript to IEEE Access, including both the initial submitted version and the final revised version, analyzing 23 recent studies on deep learning-based MAD, published in IEEE Access [20] (Task T1).
- **D2 - Mid-Term Workshop Report**: Submission of the Mid-Term Workshop Report, including a summary of the literature review, an analysis of the SOTA, and the preliminary findings from the dataset exploration and initial experimental results (Tasks T1, T2, and T3).
- **D3 - ISPA 2025 Conference Paper**: Submission of the manuscript to ISPA 2025, including the initial draft and the final Camera-Ready version, presenting the interpretability analysis findings [21] (Task T3).
- **D4 - RecPad 2025 Conference Paper**: Submission of the manuscript to RecPad 2025, including both the initial draft and the final Camera-Ready version, presenting the additional interpretability results (see Annex C) (Task T3).
- **D5 - Interpretability Framework**: Creation of a public repository[2] for the formalized occlusion-based interpretability framework, including fully documented implementation and demonstration notebooks(Task T4).
- **D6 - Thesis Document**: Submission of the Final Report consolidating all research phases, experimental results, and main contributions (Task T5).

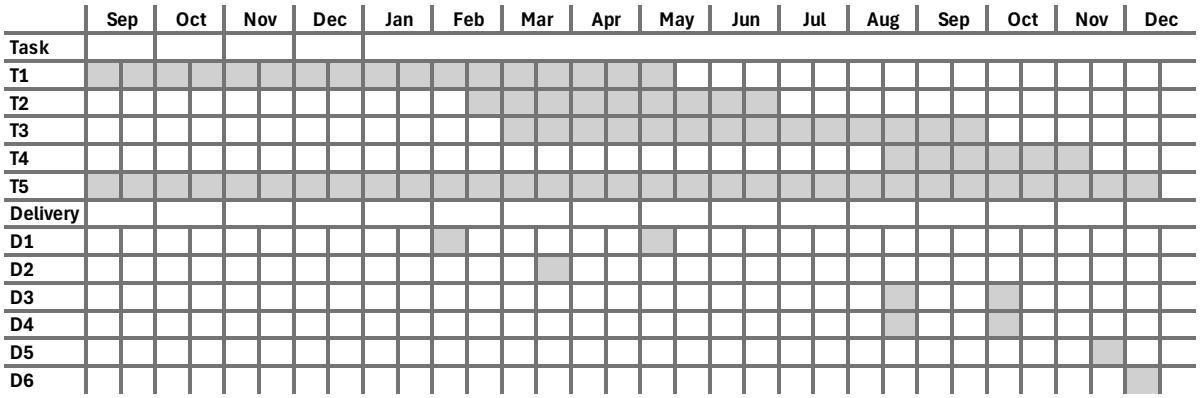


Figure 1.2: Gantt diagram illustrating the temporal organization of tasks (T1–T5) and deliverables (D1–D6) throughout the project timeline.

1.4 Document Structure

The remainder of this document is organized as follows:

[2]https://github.com/pedrojfmorais/MEI-Thesis

- **Chapter 2 - State of the Art**: Presents a literature review of deep learning-based MAD techniques, analyzing 23 recent studies to identify current approaches, evaluate their effectiveness, and determine research gaps. This chapter served as the basis for a journal article published in IEEE Access.
- **Chapter 3 - Methodology**: Outlines the experimental methodologies adopted in this research, including dataset selection, preprocessing strategies, model architectures, evaluation metrics, and the occlusion-based interpretability approach.
- **Chapter 4 - Experiments**: Details five experiments investigating different aspects of MAD system behavior, including replication studies, preprocessing impact analysis, cross-dataset generalization evaluation, and interpretability analysis. The interpretability findings from Experiment 5 were published at ISPA 2025 and RecPad 2025.
- **Chapter 5 - Occlusion-based Interpretability**: Formalizes the occlusion-based interpretability framework as a reusable, standardized methodology, discussing its implementation, capabilities, limitations, and potential applications for analyzing MAD models.
- **Chapter 6 - Conclusions**: Summarizes the main contributions, discusses the implications of the findings, and identifies promising directions for future research in MAD.

This document also includes several annexes containing supplementary materials: the full text of the journal article published in IEEE Access (Annex A); conference papers presented at ISPA 2025 and RecPad 2025, both accepted and presented, with the ISPA 2025 paper published at the date of submission of this document (Annex B, Annex C); tables of datasets, pre-trained models, and experimental results from the literature review (Annex D); sample images from the datasets selected in the methodology (Annex E); unsuccessful experiments (Annex F); experimental results and code (Annex G, Annex H, Annex I); and the source code for the occlusion-based interpretability framework (Annex J).

2 State of the Art

As established in the previous chapter, morphing attacks pose a threat to biometric authentication systems, particularly in high-security contexts such as border control and identity verification. To address this challenge, numerous studies have explored the field of MAD [22, 23, 24, 25, 26, 27, 28], with varying degrees of success in detecting morphed images.

The objective of this chapter[1] is to provide a comprehensive review of recent advances in MAD, with a particular focus on deep learning techniques. By critically analyzing current approaches and results, this literature review aims to identify the existing research gaps and highlight areas where further investigation is needed.

While this review focuses on deep learning-based MAD, it is important to acknowledge the evolution of detection methods. Before the adoption of deep learning models, researchers explored forensic approaches such as symmetry analysis and compression artifacts to detect morphs, though these proved limited in practical scenarios [29]. Handcrafted features like Local Binary Patterns (LBP) and Binarized Statistical Image Features (BSIF) were also used, but their performance degraded under compression and cross-domain variation [30, 31]. Related methods have also addressed broader biometric security tasks such as spoofing detection and template protection using deep learning techniques [32].

This chapter seeks to address the following research question:

RQ: How can deep learning techniques be used to improve MAD?

To answer this question, a literature review was conducted, evaluating the effectiveness of deep learning-based approaches in MAD. This review assesses their strengths and limitations, highlighting areas where further research is needed to advance the field.

The key contributions of this chapter are as follows:

- Understanding the role of deep learning techniques in MAD.
- Identifying key areas for future research to improve detection capabilities.

The structure of the chapter is as follows: Section 2.1 explains the background concepts. Section 2.2 outlines the methodologies used for the literature selection. Section 2.3 reviews recent studies on deep learning techniques for MAD. Section 2.4 summarises

[1]The work presented in this chapter served as the basis for the publication: P. Morais, I. Domingues, and J. Bernardino, "Deep learning techniques for detecting morphed face images: A literature review", IEEE Access, vol. 13, pp. 105 952–105 981, 2025. [20] (see Annex A for the full publication text).

the key findings from the literature. Section 2.5 identifies the major research gaps in the field. Section 2.6 proposes future research directions. Finally, Section 2.7 presents the main conclusions of the chapter.

2.1 Background Concepts

In this section, we introduce the key concepts necessary for understanding MAD, particularly the approaches and metrics discussed in the literature review.

2.1.1 D-MAD and S-MAD

MAD can be broadly categorized into two types: Differential Morphing Attack Detection (D-MAD) and Single-image Morphing Attack Detection (S-MAD).

D-MAD involves comparing a suspected morphed image with a trusted reference image, typically obtained from a live capture. The goal of D-MAD is to identify discrepancies between the two images, which could indicate that the suspected image has been morphed. This approach is widely explored in studies, such as those by Borghi et al. (2021) [33] and Singh and Ramachandra (2022) [34].

S-MAD, on the other hand, focuses on detecting whether a given single facial image has been morphed without any reference image available for comparison. It is a more challenging task due to the lack of a ground-truth reference, relying solely on inherent features of the image itself. Methods like the one proposed by Ferrara et al. (2021) [35] and Medvedev et al. (2023) [22] exemplify S-MAD approaches.

While both D-MAD and S-MAD aim to detect morphing attacks, their effectiveness and applicability differ depending on the operational context. D-MAD generally achieves higher detection Accuracy (ACC) by leveraging a trusted reference image to spot subtle inconsistencies. However, it is only feasible in verification scenarios, such as border control, where a live capture or previously enrolled image is available. In contrast, S-MAD is crucial in standalone settings, like passport issuance, where only the submitted image can be analyzed. Despite being more widely applicable, S-MAD methods face greater challenges, especially when dealing with high-quality morphs or images affected by compression and resolution loss [36, 18]. Thus, the choice between D-MAD and S-MAD should be guided by the availability of reference data and the specific requirements of the biometric application.

Figure 2.1 illustrates the adoption of S-MAD and D-MAD techniques by the reviewed literature, with some papers exploring both approaches.

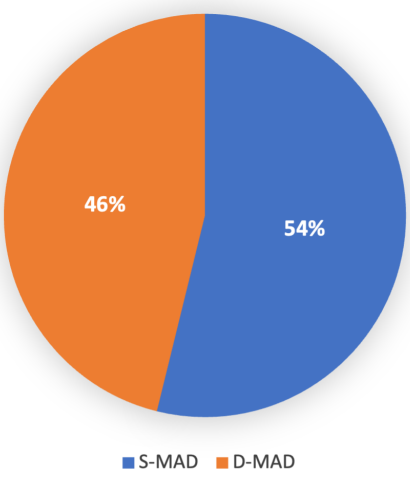


Figure 2.1: Proportion of papers employing S-MAD and D-MAD techniques.

2.1.2 Bona Fide and Attack Images

In the context of MAD, two types of images are commonly referenced:

- ***Bona Fide* Image**: A *bona fide* image is a genuine, unaltered facial photograph used in identification or verification systems. It represents the actual appearance of an individual without any manipulation or morphing. The goal of the detection system is to correctly identify these images as authentic [37].
- **Attack Image**: An attack image refers to a manipulated or morphed facial image used with the intent to deceive identification systems. In the context of MAD, these are the images created by blending features from multiple individuals to create a composite that resembles each contributor. Detecting these morphs is critical to prevent fraudulent access or identity misuse. Studies such as those by Makrushin et al. (2017) [38] and Sarkar et al. (2022) [39] investigate the characteristics and challenges of detecting these attack images.

2.1.3 Evaluation Metrics

Various metrics are used to evaluate the performance of MAD systems, including those specified by ISO/IEC 30107-3, which provides guidelines for the testing and performance evaluation of biometric systems in detecting morphing attacks [40]. Key metrics commonly referenced include:

- **Attack Presentation Classification Error Rate (APCER)**: represents the proportion of attack samples (i.e., morphed images) that are incorrectly classified as *bona fide* by the MAD system. It is conceptually similar to the False Acceptance Rate (FAR), as both metrics indicate the system's tendency to erroneously accept attack samples [40, 41, 42]. The formula for APCER is:

$$\text{APCER} = \frac{\text{Number of false acceptances}}{\text{Total number of attack presentations}} \tag{2.1}$$

In the context of morphed images, APCER and Morphing Attack Classification Error Rate (MACER) are often treated as equivalent, as both measure the FAR for morphed faces. In this way, note that the name of the metric has been updated to MACER in some contexts, in accordance with a new ISO standard currently under development.

- ***Bona Fide* Presentation Classification Error Rate (BPCER)**: measures the proportion of *bona fide* samples that are incorrectly classified as attacks by the system. It is analogous to the False Rejection Rate (FRR), indicating the system's propensity to falsely reject genuine samples [40, 41, 42]. The formula for BPCER is:

$$\text{BPCER} = \frac{\text{Number of false rejections}}{\text{Total number of } \textit{bona fide} \text{ presentations}} \tag{2.2}$$

- **BPCER @ APCER**: in MAD tasks, BPCER is typically assessed at specific APCER values that reflect common operating points in real-world applications. These thresholds are chosen to balance the trade-off between false acceptance and false rejection. This approach emphasizes security, as it prioritizes minimizing the acceptance of attacks by fixing APCER and evaluating the impact on BPCER. The relationship between APCER and the corresponding BPCER is often illustrated using a Detection Error Tradeoff (DET) curve, which also facilitates the estimation of the Equal Error Rate (EER) [43].
- **APCER @ BPCER**: although less commonly used than BPCER @ APCER, this metric evaluates APCER while maintaining a fixed BPCER, focusing on usability by minimizing the impact on *bona fide* users. It measures how effectively the system detects attacks while ensuring an acceptable rate of false rejections [44].
- **EER / Detection Equal Error Rate (D-EER)**: refer to the operational point at which APCER equals BPCER. While EER is commonly used across biometric evaluation tasks, D-EER is specifically applicable to MAD, representing the point where the attack and *bona fide* errors are balanced. A lower D-EER signifies better overall system performance, as it shows a reduced error rate across both attack and *bona fide* classifications [41, 30]. The D-EER value is determined as the point where:

$$\text{APCER} = \text{BPCER} \tag{2.3}$$

- **Average Classification Error Rate (ACER)**: average of APCER and BPCER, providing a balanced measure of the system's performance. Although ACER provides a single performance metric for MAD systems, it assumes that both types of errors are equally important. However, in practical applications, APCER and BPCER may have different priorities, and ACER may not fully capture the trade-offs between these error rates [42]. The formula for ACER is:

$$\text{ACER} = \frac{\text{APCER} + \text{BPCER}}{2} \tag{2.4}$$

- **ACC**: measures the overall performance by calculating the ratio of correctly classified samples to the total number of samples [27]. While ACC provides an overview of performance, it is not commonly used in MAD evaluations. In imbalanced datasets, ACC can be misleading, as a high ACC may occur by simply favoring the majority class (*bona fide* samples), neglecting the importance of detecting attacks [45]. The formula for ACC is:

$$\text{ACC} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}} \tag{2.5}$$

2.2 Literature Review Methodology

A systematic literature review was conducted to identify relevant research on face MAD. The methodology followed a structured approach involving data source selection, search query formulation, and the application of inclusion and exclusion criteria to filter relevant studies, as illustrated in Figure 2.2. For detailed information about the data sources, search query, and inclusion and exclusion criteria, refer to Annex A.

As shown in Figure 2.2, a total of 265 papers were initially identified from three data sources: Scopus (146), IEEE Xplore (96), and Google Scholar (46). These papers were published between 2019 and 2024, a timeframe chosen to ensure relevance to the current landscape of the field.

After removing duplicates, the total number of papers was reduced to 163. As shown in Figure 2.3, face MAD research has experienced steady growth, with a moderate number of publications in 2019 and 2020, and an increase starting in 2022. This trend highlights the recognition of face morphing as a serious security threat, and the continued relevance of MAD techniques.

The review process narrowed down the selection to include only those papers that were most relevant to the study, focusing on their contribution to the detection of face

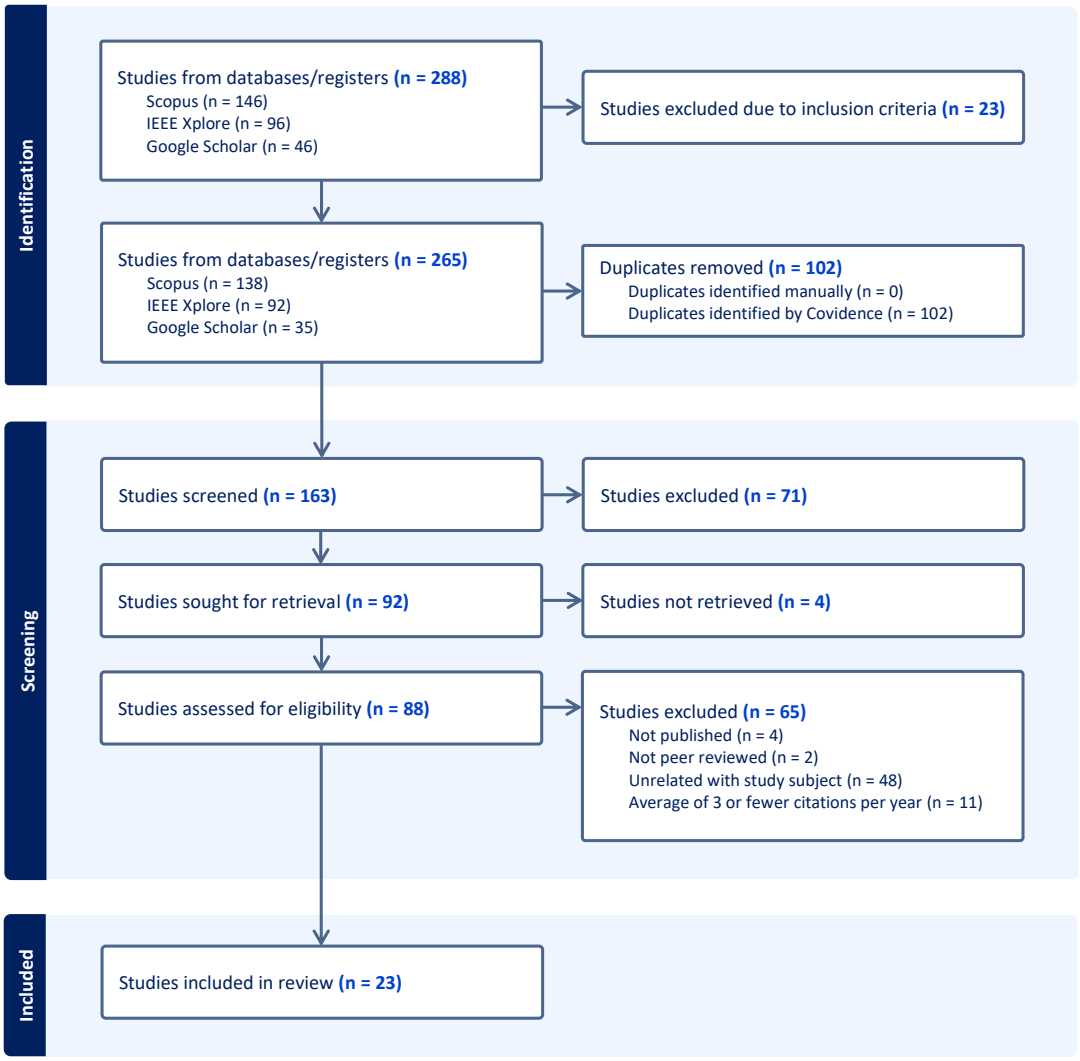


Figure 2.2: Methodology applied in the literature review.

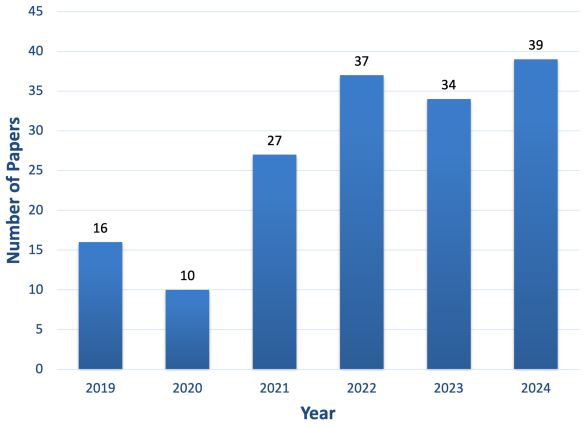


Figure 2.3: Distribution of papers by year of publication.

morphing attacks using deep learning techniques.

Using the exclusion criteria, the number of papers was further reduced to 23. One of the exclusion criteria was to remove papers with an average of three or fewer citations

per year. The distribution of papers by average number of citations per year is shown in Figure 2.4, where six papers with [0-1] citations per year correspond to those published in 2024.

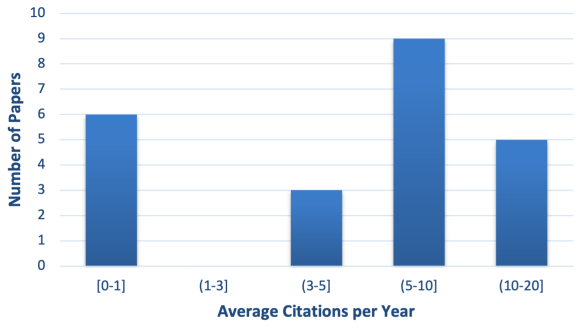


Figure 2.4: Distribution of selected papers by average citations per year.

The distribution of citations per year for each of the final 23 papers selected is illustrated in Figure 2.5. In this graph, the papers are ordered by their total citation count, from highest to lowest, to highlight the relative impact of each work over time. The citation counts were collected from Google Scholar as of 27 April 2025. This visualization provides an overview of how the relevance and influence of each paper has progressed, reflecting both the maturity of the research and the growing interest in face morphing detection techniques using deep learning.

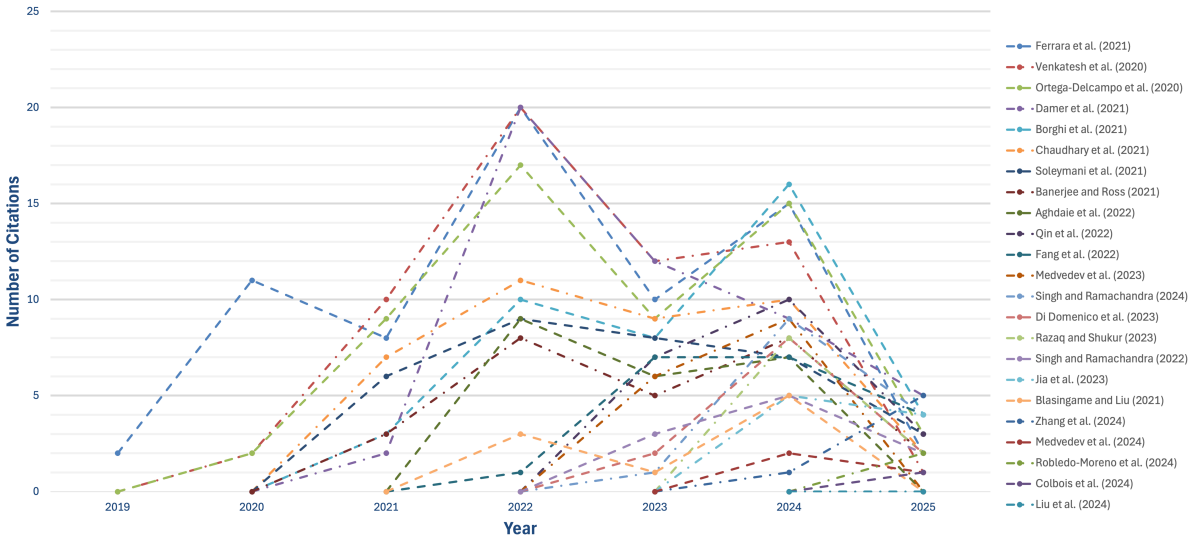


Figure 2.5: Selected papers citation count per year, ordered by total number of citations.

The citation analysis presented in Figure 2.5 reveals that Ferrara et al. (2021) [35] stands out as the most cited work across the reviewed period (2019–2025), consistently accumulating a high number of citations and peaking at around 20 citations in 2022, followed by sustained influence through 2024.

Other highly cited papers include Venkatesh et al. (2020) [46], which exhibited a sharp

rise and a strong peak in 2022, as well as Ortega-Delcampo et al. (2020) [47] and Damer et al. (2021) [48], both of which showed citation growth between 2021 and 2023. Similarly, Borghi et al. (2021) [33] and Chaudhary et al. (2021) [49] demonstrated steady citation increases over the years, indicating ongoing relevance within the community.

These highly cited papers dominate the field, suggesting that they have had a foundational impact on advancing research in MAD using deep learning approaches.

Looking at the overall citation evolution, some general trends can be observed:

- There is a clear peak around 2022, where the majority of the papers reached their highest yearly citation counts.
- After 2022, a gradual decline in yearly citations becomes apparent, although several papers continued to attract moderate attention through 2024 and into 2025.
- Recent papers published in 2023 and 2024 (e.g., Medvedev et al. (2023) [22], Singh and Ramachandra (2024) [28]) are starting to accumulate citations but have not yet reached the impact levels of earlier publications, which is expected given their limited exposure time.
- Papers from 2024, including Liu et al. (2024) [50] and Colbois et al. (2024) [51], currently show low citation counts, consistent with their recent release.

The overall trends suggests that the research area experienced a surge in activity and visibility around 2022, likely reflecting an intensified research interest during that period, followed by a phase of stabilization. This pattern is typical of maturing fields where initial breakthroughs drive a surge in publications and citations, subsequently leading to a more steady pace of contributions.

2.3 Literature Review Analysis

This section presents an analysis of the major deep learning approaches to face MAD as identified in the reviewed literature. These approaches include CNNs, Siamese Networks, Transformer-Based Models, and Hybrid Architectures. Each method offers unique capabilities and addresses different aspects of MAD, such as identity verification, artifact detection, and feature representation. The review focuses on evaluating their strengths, limitations, and contributions to the advancement of MAD techniques in various experimental settings. Through this analysis, this section aims to highlight the diversity of strategies and the potential for combining approaches to enhance robustness and generalization in addressing morphing attacks.

2.3.1 Datasets Used

The papers reviewed in the following subsections refer to a variety of datasets, many of which are critical to experiments or serve as the basis for creating new datasets. Table D.1 in Annex D summarizes the major datasets mentioned in the literature, highlighting those used for experimental evaluation or dataset creation. Note that not all datasets are publicly available; in some cases, it may be necessary to contact the authors of the dataset to gain access. Figure 2.6 presents a Venn diagram illustrating the distribution of datasets by their typologies (Public, Restricted, and Created) across the reviewed literature. It highlights the number of papers using each dataset type and visualizes the overlap between them.

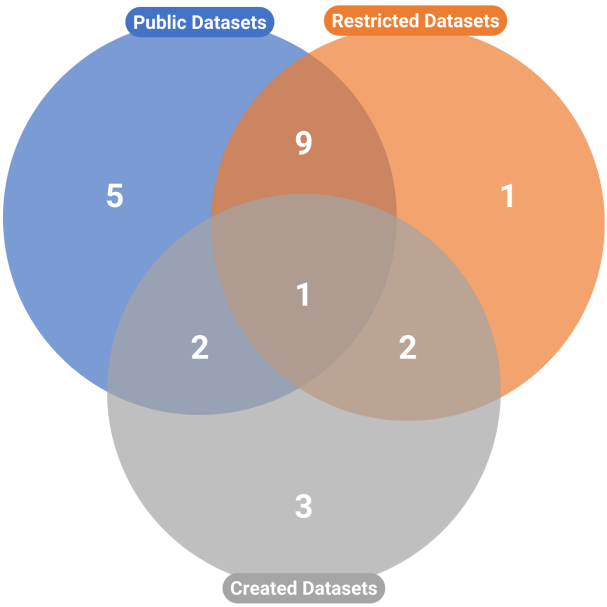


Figure 2.6: Number of papers in the literature review using each dataset type (Public, Restricted, and Created) and their overlap.

Regarding dataset availability, the datasets can be categorized into several types:

- **Publicly Available**: Datasets such as the AMSL Face Morph [52, 53, 54], VGGFace2 [55], and FRLL-Morphs [52, 56, 39] are made accessible to the public, often under licenses like Creative Commons. These datasets are generally free to use, although they may have conditions such as requiring citation or restricting commercial use.
- **Requires Authorization**: Some datasets (e.g. MorphDB [57], MorGAN [58], and the LMA Dataset [58]) are only accessible after receiving explicit authorization from the dataset creators or associated organizations. Access to these datasets typically involves submitting a formal request and providing a justification for their use, along with possible non-disclosure agreements.
- **Sequestered Evaluation Platform**: Datasets such as the State-of-the-Art Morph-

ing Detection (SOTAMD) Dataset [41] are accessible exclusively through a controlled evaluation platform, where researchers submit their models for assessment, ensuring the data remains secure and is used within specified parameters.

- **Suspended**: Certain datasets, such as the FRGC-Morphs [56], have been withdrawn from distribution due to concerns regarding their use or changes in distribution policies.

In response to these access constraints and growing privacy concerns, recent research has increasingly relied on synthetic datasets. These datasets are generated to resemble real-world biometric data while preserving privacy, and have become an essential resource for both the creation and evaluation of MAD systems. Notable examples include the Synthetic Morph Detection Dataset (SMDD) [59], Morphing Network Optimization Tool (MONOT) [60], and Augmented Morphing Network Optimization Tool (AMONOT) [61] datasets. These synthetic datasets offer high-quality, privacy-compliant biometric images and have been instrumental in advancing MAD techniques. By enabling experimentation without compromising real individuals' privacy, they present a valuable alternative in contexts where real data is restricted or unavailable [59, 60, 61].

2.3.2 Pre-Trained Models

The pre-trained models discussed in the reviewed literature are commonly used for feature extraction, transfer learning, and fine-tuning in MAD tasks. Table D.2 in Annex D provides an overview of the major pre-trained models referenced, summarizing their characteristics and applications. These models include widely used architectures such as AlexNet [62] and ResNet [63], as well as more specialized frameworks such as ArcFace [64] and PointNet [65]. Figure 2.7 provides a visualization of the pre-trained models used in the literature, illustrating the distribution of papers using each pre-trained model.

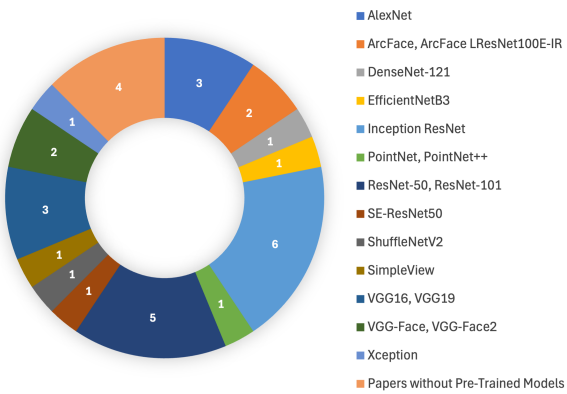


Figure 2.7: Distribution of pre-trained models used in the reviewed literature, highlighting the number of papers using each model.

2.3.3 Architecture

The architecture of MAD models is critical to their effectiveness and adaptability. This subsection examines key approaches, including CNNs, Autoencoders, GANs, Siamese Networks, Transformer-Based Models, and Hybrid Architectures. Each architecture offers unique strategies for dealing with morphing attacks, from pairwise comparisons to the use of different features and representations. Figure 2.8 illustrates the distribution of papers in the literature review by architecture type, showing the prevalence of CNN-based models, followed by hybrid models, siamese networks, autoencoders, GAN-based models, and transformer-based models. The following subsections explore these innovations and their contributions to the advancement of MAD.

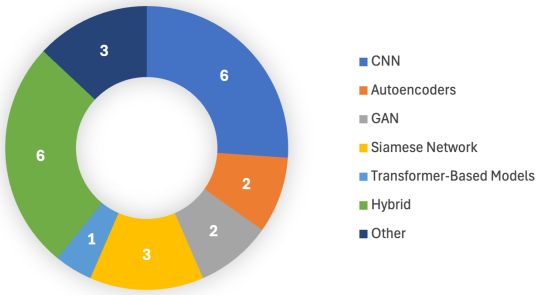


Figure 2.8: Distribution of papers in the reviewed literature by architecture type.

Convolutional Neural Networks

CNNs are a type of deep learning model that is widely used in computer vision for its ability to learn features from images. A CNN typically consists of convolutional layers that apply filters to input images to detect features such as edges, textures, and patterns. These features are then passed through pooling layers, which reduce spatial dimensions while preserving essential information, and fully connected layers for classification or regression tasks [66, 67]. The structure and flow of a basic CNN are illustrated in Figure 2.9, which show how convolutional, pooling, and fully connected layers work together to process and classify input data. This ability to automatically extract features from raw image data makes CNNs effective for MAD, where detection of subtle variations is essential. Several studies have adapted CNN architectures to improve their robustness and generalization in MAD.

Venkatesh et al. (2020) [46] propose a method to detect morphed face attacks using residual noise extracted with a deep Multi-scale Context Aggregation Network (MS-CAN). The approach involves the aggregation of multiple denoising techniques — Wavelet Denoising (WD), Block Matching and 3D filtering (BM3D), Multiresolution Bilateral Filtering (MBF), and Denoising Convolutional Neural Networks (DnCNN) — to create a fused denoised image. The residual noise from the aggregated denois-

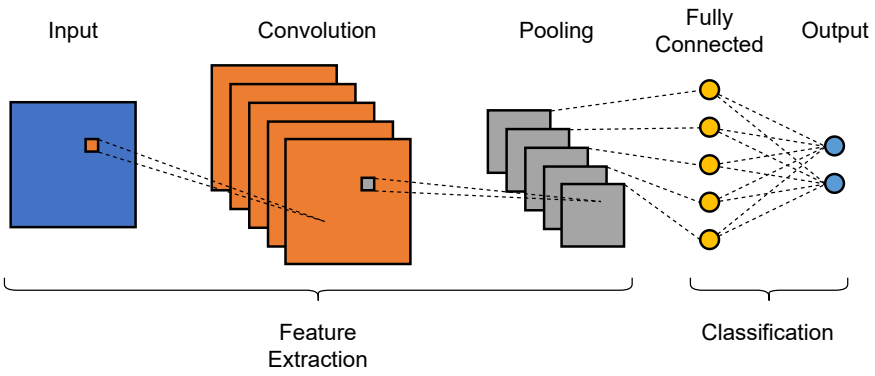


Figure 2.9: CNN architecture, inspired by [67].

ing process is then analyzed using a pre-trained AlexNet [62] to extract discriminative texture features, which are classified using a Probabilistic Collaborative Representation Classifier (P-CRC). In addition to using CNN-based feature extraction, the authors also incorporate texture analysis by employing advanced descriptors such as BSIF, Histogram of Oriented Gradients (HOG), and LBP. These texture descriptors capture discriminative local features, enhancing the robustness of MAD when combined with residual noise features. The evaluation of the method was conducted on three distinct morphed face datasets, each constructed using unique morphing techniques and compositions. The datasets included both *bona fide* and morphed images and were divided into training and testing partitions to ensure robust validation. The evaluation followed ISO/IEC 30107-3 standards, using metrics such as D-EER, BPCER, and APCER. Performance was evaluated in three experimental setups: a) training and testing on the same dataset; b) merging all datasets to evaluate scalability; c) cross-dataset evaluation to test generalizability. This evaluation strategy assessed the adaptability of the proposed MS-CAN framework across different morphing scenarios and varying image quality conditions.

Damer et al. (2021) [48] present a pixel-wise supervision approach for S-MAD, called Pixel-Wise Morphing Attack Detection (PW-MAD). Unlike traditional methods that provide a single binary classification label for an entire image, PW-MAD employs pixel-wise labels to guide the network during training, thereby focusing on more subtle morphing artifacts that may be distributed throughout the image. The proposed method uses a modified DenseNet-121 [68] architecture, adapted to include two dense blocks, two transition blocks, and a convolutional layer that outputs a feature map for pixel-wise supervision. The evaluation of the PW-MAD framework was performed using the LMA-DRD [48] dataset. Experiments were performed under different training and testing configurations, including: a) intra-dataset settings (training and testing on the same data type); b) cross-dataset settings (training on digital images and testing on re-digitized images, and vice versa). Metrics such as the APCER and the BPCER, defined by ISO/IEC 30107-3 standards, were used to evaluate performance. Receiver

Operating Characteristic (ROC) curves were also plotted to visualize the trade-off between APCER and BPCER at different operating points. This evaluation strategy aimed to measure the ability of the PW-MAD method to generalize across different types of morphing attacks and image acquisition scenarios.

Ferrara et al. (2021) [35] present an approach for face MAD that focuses on the challenges posed by Print-Scanned (P&S) images and heterogeneous image sources. Their method uses CNNs, specifically pre-trained models such as AlexNet [62], VGG19 [69], VGG-Face [70], and VGG-Face2 [55], which are fine-tuned to adapt to the morphing detection task. To address the difficulties of training on real P&S images, they simulate the P&S process during data augmentation, which involves modeling the image degradation typically caused by the printing and scanning stages. The evaluation of the proposed method was conducted using multiple datasets, including MorphDBD [57], MorphDBP&S [57], and the SOTAMD [41] benchmark. These datasets represent different morphing scenarios, from digitally morphed images to those processed through printing and scanning pipelines. The evaluation metrics included ACC, EER, and error trade-offs between APCER and BPCER, according to ISO/IEC 30107-3 standards. The method was tested under both intra-dataset and cross-dataset conditions, emphasizing generalizability and robustness. In addition, simulations of the P&S process were validated by comparing real and synthetic P&S images to ensure the reliability of the augmentation strategy. These extensive experiments provided insight into the performance of the method under different acquisition and morphing scenarios.

Qin et al. (2022) [71] introduce a face MAD and localization approach based on feature-wise supervision, designed for both S-MAD and D-MAD. Unlike traditional score-level losses, their method employs a CNN with feature-wise loss to better capture the morphing patterns present in different face morphing attacks. For S-MAD, the approach uses a fine-grained classification loss that allows both the detection of morphing attacks and the precise localization of morphed areas within an image. For D-MAD, the method constructs differential loss functions based on similarity and distance, which improves generalization to unseen morphing attacks and increases robustness to low-resolution and non-frontal probe images. The evaluation of this approach was conducted using a combination of self-collected and publicly available datasets, with images processed to simulate morphing techniques such as complete, splicing, retouching, and combined morphing. Several protocols were designed to test different aspects of the system, including: a) overall detection performance to intra-technique (same morphing technique) morphing types; b) generalization to unseen morphing types; c) cross-dataset performance; d) robustness to varying probe image qualities; e) robustness to non-frontal probe face images. Metrics such as EER, BPCER, and APCER were used following ISO/IEC 30107-3 standards. Furthermore, visualization techniques were used to validate the localization of morphed areas, providing inter-

pretable results for potential human verification.

Medvedev et al. (2023) [22] present MorDeephy, a MAD method that employs a fused classification approach using two CNNs trained in parallel. The proposed method is designed for S-MAD and aims to discriminate between *bona fide* and morphed images by learning deep facial features that encode authenticity information. The architecture uses two ResNet-50 [63] networks, each of which is trained with different identity labels for the morphed images. During training, both networks learn high-level identity-specific features, and their outputs are combined through a dot product, followed by a sigmoid activation for morph detection. The authors also employ a unique dataset filtering strategy using VGGFace2 [55], generating selfmorphs to reduce the bias towards detecting superficial morph artifacts. The evaluation of MorDeephy was conducted using multiple publicly available benchmark protocols, such as FRGC-Morphs [56] and AMSL [52, 53, 54]. These benchmarks incorporate different morphing techniques, including landmark-based and GAN-generated morphs. Metrics such as APCER and BPCER were used to evaluate performance at different operational thresholds. DET curves were plotted to analyze the trade-offs between detection ACC and error rates across different datasets. The evaluation covered intra-dataset and cross-dataset scenarios, testing MorDeephy's robustness to: a) unseen morphing attacks; b) varying image acquisition conditions, such as P&S images. In addition, the experiments included comparisons with SOTA methods to assess generalizability and scalability.

Medvedev et al. (2024) [23] propose an extension of the MorDeephy approach [22] to D-MAD using a fused classification strategy. The method employs a dual CNN-based setup, similar to MorDeephy, where two distinct networks process input image pairs and perform identity classification independently. This extension adapts the previous no-reference setup to handle D-MAD scenarios. To improve detection performance, the authors introduce a data mining strategy leveraging GAN-based and landmark-based morphs and employ a class allocation method for morphed samples to improve feature discriminability. The evaluation was conducted on multiple benchmark datasets, including custom protocols designed for D-MAD scenarios. These protocols used a mix of *bona fide* and morphed image pairs, generated from datasets such as Face Research London Lab (FRLL) [72] and VGGFace2 [55], and included various morphing techniques, including StyleGAN and landmark-based approaches. Metrics such as APCER and BPCER, in accordance with ISO/IEC 30107-3 standards, were used for evaluation. DET curves were plotted to visualize performance trade-offs across different thresholds and morphing scenarios. Additionally, the study included comparative experiments with baseline binary classification methods and alternative fused strategies to assess generalization capabilities and performance across unseen datasets. This evaluation was designed to examine the adaptability of the method to different morphing attack techniques and operating conditions.

The CNN-based approaches reviewed here illustrate different ways in which convolutional architectures can be tailored for MAD. Venkatesh et al. (2020)[46] use a MS-CAN and residual noise analysis to improve robustness by aggregating denoising techniques, while Damer et al. (2021)[48] focus on pixel-wise supervision to detect subtle artifacts, using a modified DenseNet-121 architecture for PW-MAD. Ferrara et al. (2021)[35] adapt pre-trained CNNs to detect morphing in P&S images, addressing the unique challenges of image degradation in real-world conditions. Qin et al. (2022)[71] propose a feature-wise supervision approach for both S-MAD and D-MAD, emphasizing fine-grained classification and morphing localization. Medvedev et al. (2023)[22] and their extension Medvedev et al. (2024)[23] use parallel CNNs for both S-MAD and D-MAD, with an emphasis on cross-protocol generalizability by training on identity-specific features.

Together, these CNN-based models showcase a wide range of techniques, from adversarial learning to pixel-wise supervision and feature-wise loss, each offering a distinct approach to handling different aspects of MAD. These methods reflect the versatility and adaptability of CNN architectures in addressing different challenges in morphing attack scenarios.

Autoencoders

Autoencoders, introduced by Hinton and Salakhutdinov (2006) [73], are a type of neural network designed for unsupervised learning, with a primary focus on dimensionality reduction. The network is composed of two main components: an encoder and a decoder. The encoder compresses the input data into a lower-dimensional latent representation, while the decoder reconstructs the original input from this compressed form. The primary objective during training is to minimize the reconstruction error, enabling the model to learn an efficient and compact representation that retains the most important features of the data [73]. As illustrated in Figure 2.10, the autoencoder architecture facilitates this process by using the encoder to map the input into a latent space and the decoder to reconstruct the input, with both parts working together to reduce the reconstruction error.

Ortega-Delcampo et al. (2020) [47] propose a de-morphing approach for MAD at Automated Border Control (ABC) systems, using CNNs. The method is based on the comparison of two images: the potentially morphed chip image from an electronic passport and an *in vivo* image captured at the ABC system. The de-morphing process, facilitated by an autoencoder-based CNN, attempts to unravel the morphing and isolate the original identity by reconstructing an image that highlights any morphing artifacts. The evaluation of the proposed system has been performed using the FRAV-ABC dataset, which includes images captured under real ABC conditions. The dataset consists of chip images, *in vivo* images, and P&S versions of the morphed images. The

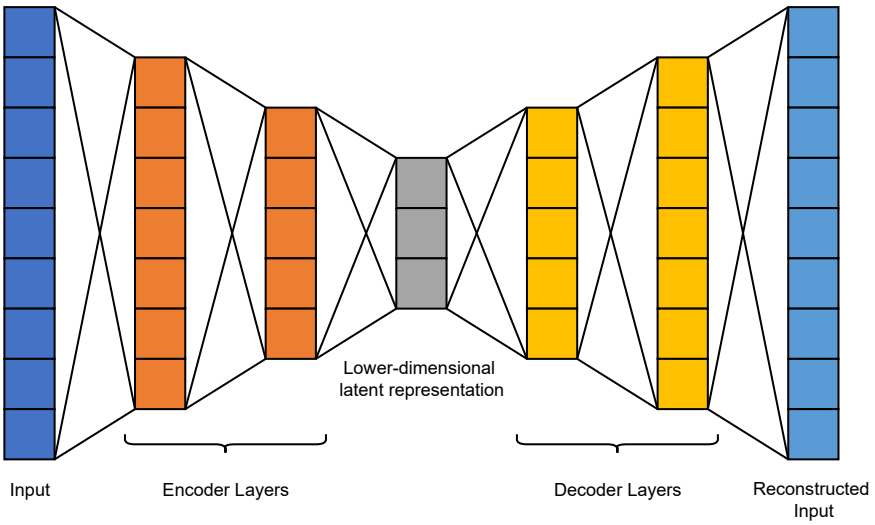


Figure 2.10: Autoencoder architecture, inspired by [74].

authors used various evaluation metrics, including the EER and DET curves, to assess the effectiveness of their approach. The evaluation was performed on both digital and P&S images to assess the system's ability to operate under different image quality conditions and real-world scenarios.

Fang et al. (2022) [75] propose an unsupervised MAD method called Self-Paced Anomaly Detection for Morphing Attack Detection (SPL-MAD). This method uses a Convolutional Autoencoder (CAE) to analyze reconstruction errors as an indicator of morphing artifacts. The CAE is trained on large-scale, general face recognition datasets that may contain unlabeled and unintentionally manipulated samples, referred to as "contaminated data". These contaminated samples, which are not intentionally labeled as morphed, may contain changes or manipulations that are not immediately apparent. The model learns to distinguish between *bona fide* and morphed images by adapting to these subtle differences in reconstruction. The Self-Paced Learning (SPL) paradigm is incorporated to iteratively filter out suspicious samples during training, thereby widening the reconstruction gap between *bona fide* and attack images. The evaluation of SPL-MAD was conducted on eight public MAD datasets, including both landmark- and GAN-based morphing techniques, to test its generalizability. Metrics such as APCER and BPCER were used, along with EER for overall performance comparison. The model's reconstruction error was analyzed using ROC curves to visualize performance across various operating points. Cross-dataset evaluations sought to assess the robustness of SPL-MAD against unknown attacks, while ablation studies aimed to demonstrate the contribution of the SPL paradigm in improving detection ACC by increasing the reconstruction error gap between *bona fide* and attack samples.

The autoencoder-based approaches reviewed here illustrate two strategies to address the challenges of MAD. Ortega-Delcampo et al. (2020)[47] present a de-morphing

approach that uses an autoencoder-based CNN to isolate and reveal hidden identities within morphed images. Fang et al. (2022) [75] present an unsupervised anomaly detection approach using a CAE and SPL, emphasizing the importance of adapting to unknown morphing attacks without the need for extensive labeled data.

Together, these autoencoder-based methods emphasize the effectiveness of this architecture in unsupervised learning for MAD.

Generative Adversarial Networks

GANs are a type of deep learning model introduced by Goodfellow et al. (2014)[76]. GANs consist of two neural networks: a generator and a discriminator, which are trained together in an adversarial manner. The generator learns to create synthetic data, such as images, by trying to mimic the distribution of real data. The discriminator's role is to distinguish between real and fake data generated by the model. The generator and discriminator are in a continuous feedback loop, with the generator trying to fool the discriminator into classifying fake data as real, and the discriminator improving its ability to correctly identify real and generated data. This adversarial training process results in the generator producing increasingly realistic data over time. GANs have been widely used in image generation and enhancement tasks, demonstrating their ability to produce high-quality, realistic results [76]. The architecture of GANs, including the interactions between the generator and the discriminator, is shown in Figure 2.11.

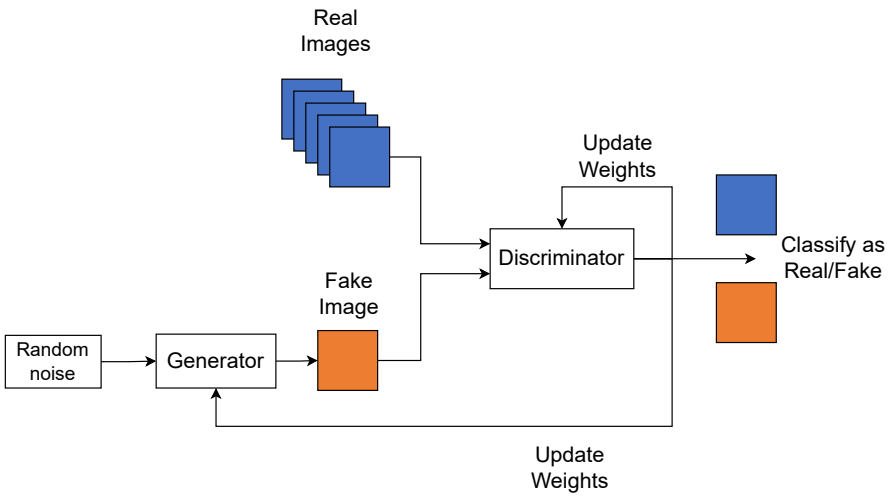


Figure 2.11: GAN architecture, inspired by [77].

Banerjee and Ross (2021) [78] propose a D-MAD method that uses a Conditional Generative Adversarial Network (cGAN) for identity disentanglement. The approach aims to determine whether a document image is morphed by learning to disentangle identities from the morphed image, conditioned on a trusted reference image. Using a cGAN, the model is trained to reconstruct the reference image from the mor-

phed document image, removing the contribution of the second identity present in the morph. The disentangled output is then compared with the reference using a biometric comparator to generate a similarity score, which is used for morph detection. Evaluation was performed on the AMSL [52, 53, 54], MorGAN [58], and EMorGAN [79] datasets, which include both landmark-based and GAN-based morphing techniques. Intra-dataset evaluations measured the effectiveness of the method under controlled conditions, while cross-dataset tests assessed its robustness and generalizability to unseen morphing scenarios, including high-quality morphs, cross-domain variations, and hybrid morphing techniques. Metrics such as APCER and BPCER were used, consistent with ISO/IEC 30107-3 standards. Additionally, the system's ability to handle identity disentanglement in scenarios where document and reference image pairs are highly correlated was analyzed. These experiments evaluate the framework's strength in detecting morphing attacks without requiring explicit labels during training.

Blasingame and Liu (2021) [80] propose a S-MAD method using an adversarial learning approach. The architecture leverages a GAN to enhance the training process of the detector, which is designed to be more robust to various morphing attacks. The GAN architecture includes an encoder, a generator, and a critic, with an additional component called the "combinator" that combines latent codes to generate morphs. Latent codes are vector representations in a compressed latent space that encapsulate essential features of the original images, facilitating morph generation by mixing these features in the latent space. The critic, trained using adversarial techniques, is then repurposed as the morphing attack detector, with transfer learning used to fine-tune it for the S-MAD task. The evaluation was performed on two datasets containing both landmark-based and deep 5learning-based morphs. Several experiments were conducted to assess the robustness and generalization of the proposed method. These experiments included: a) training and testing on the same dataset; b) evaluating the system across different morphing techniques and datasets. To ensure reliable evaluation metrics such as D-EER and DET curves were used as recommended by ISO/IEC 30107-3 standards.

The methods discussed here highlight the flexibility and creativity of GANs for MAD, with each approach incorporating different strategies and improvements. Banerjee and Ross (2021) [78] use a cGAN for identity disentanglement, allowing the system to reconstruct and compare identities to detect morphing, while Blasingame and Liu (2021)[80] apply adversarial learning with GANs to improve the model's resistance to various morphing attacks.

Together, these GAN-based models demonstrate how adversarial learning and identity disentanglement can be used to improve MAD performance.

Siamese Networks

Siamese Networks are a popular architecture for D-MAD because they are well-suited for comparing pairs of images to find similarities or discrepancies. A Siamese Network consists of two identical neural network branches that process each input image separately but share the same weights and parameters. This design allows the network to learn comparable feature representations for each image [81]. By evaluating the distance or similarity between these features, the network can determine whether the images are genuine or manipulated. The architecture of a Siamese Network, as shown in Figure 2.12, highlights the dual-branch structure and the feature comparison process.

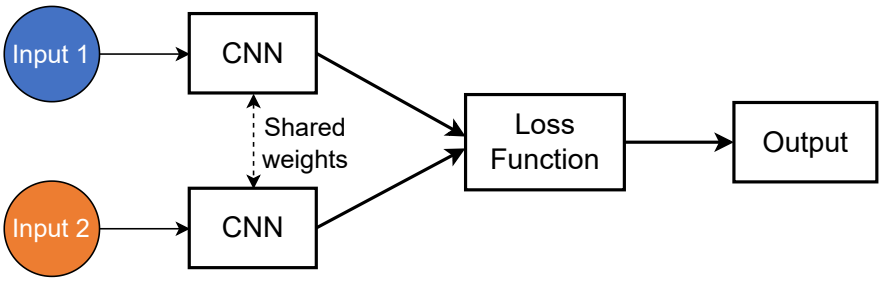


Figure 2.12: Siamese network architecture, inspired by [81].

Borghi et al. (2021) [33] present a framework for D-MAD using a double Siamese architecture. The proposed method consists of two parallel deep learning-based modules: the Identity block, which uses a ResNet50 [63] model pre-trained on the MS-Celeb-1M and VGGFace2 [55] datasets to extract identity-specific information from input face images, and the Artifact block, which employs a Squeeze-and-Excitation (SE)-ResNet50 [82] model to identify artifacts caused by morphing. The output features of both blocks are merged using a final fully connected layer, resulting in a prediction score that indicates whether the probe image is *bona fide* or morphed. The evaluation was conducted on three datasets—Progressive Morphing Database (PMDB) [57], MorphDB [57], and AMSL [52, 53, 54]—using a cross-dataset validation strategy. The PMDB [57] dataset was used for training, while MorphDB [57] and AMSL [52, 53, 54] were employed for testing to assess the framework's generalization capability. Performance was measured using metrics such as EER, BPCER, APCER, with DET curves plotted for performance analysis. The evaluation included different test scenarios involving variations in morphing factors and types of probe images (criminal or accomplice). These scenarios provided insight into the robustness of the framework against different morphing attack conditions and its ability to handle artifacts and identity inconsistencies.

Chaudhary et al. (2021) [49] present a D-MAD method using an undecimated 2D Discrete Wavelet Transform (DWT) and a deep Siamese network. The approach focuses

on identifying morphing artifacts in the wavelet domain, where they are easier to detect than in the spatial image domain. The method involves decomposing images into 48 mid- and high-frequency wavelet sub-bands and applying Kullback-Liebler Divergence (KLD) to select the 22 most discriminative sub-bands for training. A Siamese network, using a pre-trained Inception ResNet v1 architecture [83] is then trained on these selected sub-bands, with contrastive loss applied to maximize the dissimilarity between morph and genuine images. In addition to using Siamese Networks, the authors used advanced texture descriptors, such as BSIF and LBP, to extract highly discriminative features for MAD, which were then combined with Support Vector Machines (SVMs) for classification. These texture descriptors capture subtle variations in local textures, which are indicative of morphing artifacts, and complement the deep learning-based approach to improve robustness. By combining these texture-based features with wavelet domain analysis, the method ensures the enhanced detection of localized and frequency-based morphing inconsistencies. The evaluation of the proposed framework includes two primary approaches: BW-22 and RGB-66. BW-22 works on grayscale images, where the image is converted to black-and-white, and wavelet decomposition is applied to the grayscale channels. A total of 22 highly discriminative sub-bands are selected from the mid- and high-frequency components of the wavelet decomposition. This approach emphasizes the texture and structural features that emerge in the wavelet domain, making it suitable for scenarios where color information is less critical. RGB-66, on the other hand, decomposes each of the three color channels (Red Green Blue (RGB)) of the image separately, resulting in 144 wavelet sub-bands. After selecting the 22 most discriminative sub-bands from each channel, a total of 66 sub-bands are used for training. This method preserves the color information and provides a basis for evaluating whether color features contribute to morph detection. The evaluation of both methods was performed on three datasets—VISAPP [38], MorGAN [58], and LMA [58]—each containing different morphing techniques, such as splicing and GAN-based generation. The model was trained on a universal dataset, which combined the training sets from all three datasets, and then tested individually on each dataset as well as on the universal dataset. In addition, the method was tested on AMSL [52, 53, 54] to validate its ability to generalize to unseen morphing scenarios. Metrics such as APCER and BPCER were used, following ISO/IEC 30107-3 standards, with DET curves plotted to illustrate performance over different thresholds.

Soleymani et al. (2021) [84] propose a D-MAD approach using a Siamese network architecture with a pre-trained Inception ResNet v1 [83] as the base model. The model is fine-tuned to compare image pairs by contrastive loss, which emphasizes the similarity or dissimilarity between images in a latent embedding space. The evaluation was performed on two morphing datasets, VISAPP17 [38] and MorGAN [58], each containing images generated using different morphing techniques. The evaluation included three decision-making strategies: a) Euclidean distance between the embeddings; b) feature

difference combined with a SVM; c) feature concatenation with an SVM. Metrics such as D-EER, APCER, and BPCER were used to evaluate the performance of the framework according to ISO/IEC 30107-3 standards. Additionally, cross-dataset validation was performed to evaluate the robustness and generalizability of the approach across different morphing techniques and datasets. The experiments also compared the proposed method against baseline techniques, including traditional handcrafted features such as BSIF and LBP, as well as other deep learning-based approaches.

The approaches described in [33],[49], and[84] demonstrate the versatility and innovation of Siamese networks for D-MAD, each employing different strategies and extensions. These different strategies highlight distinct contributions to improving MAD, with each approach excelling in different aspects. Borghi et al. (2021) [33] combine identity verification and artifact detection through a double Siamese framework with specialized modules, an Identity block and an Artifact block. This combination enables the model to simultaneously leverage the strengths of identity matching and artifact analysis, making it particularly robust in cases where high-quality morphed images contain minimal visible artifacts. On the other hand, Chaudhary et al. (2021) [49] adopt a unique preprocessing strategy by analyzing morphing artifacts in the wavelet domain, using 2D DWT to focus on specific frequency-based characteristics. This approach provides an advantage in identifying subtle morphing artifacts that may be less distinguishable in the spatial domain, thus increasing the robustness of the model to different morphing techniques. Finally, Soleymani et al. (2021) [84] take a different path by focusing on embedding analysis, using a simple Siamese architecture with a pre-trained Inception ResNet v1 [83] to produce high-quality latent representations that enable effective classification using multiple decision strategies.

Together, these approaches provide complementary perspectives on the challenges of MAD, emphasizing embedding quality, artifact-specific feature extraction, and multifaceted analysis to advance the field.

Transformer-Based Models

Transformer-based models for image processing, such as Vision Transformers (ViTs), are a type of deep learning architecture introduced by Dosovitskiy et al. (2020)[85]. These models are based on the Transformer architecture originally developed for natural language processing, but adapted for image classification and other vision tasks. Instead of using convolutional layers, ViTs divide an image into fixed-size patches, which are then linearly embedded into a sequence of tokens. These tokens are processed by the Transformer's self-attention mechanism, which allows the model to capture global relationships between image patches. Like traditional Transformers, ViTs consist of an encoder-decoder structure, where the encoder processes the sequence of image tokens, and the decoder generates the output, typically the class label for image clas-

sification tasks. The self-attention mechanism in the encoder helps the model focus on relevant parts of the image, while the decoder performs the final classification. ViTs have shown success in image processing tasks, outperforming traditional CNNs in various benchmarks, particularly when trained on large datasets[85]. The architecture of ViTs, including the partitioning of images into patches and the self-attention layers, is illustrated in Figure 2.13.

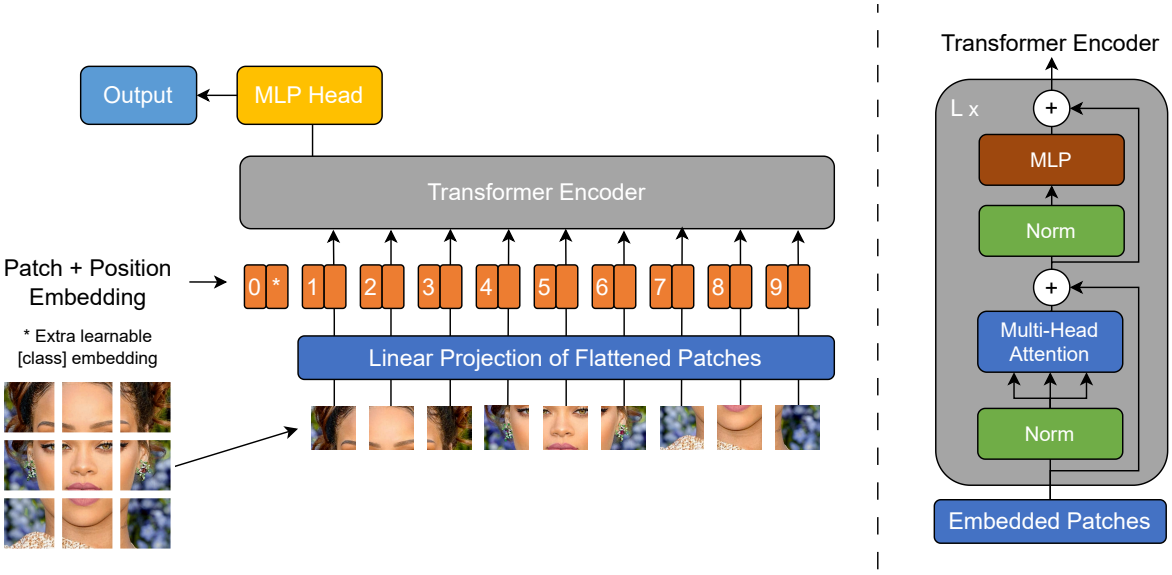


Figure 2.13: ViT architecture, inspired by [85].

Zhang et al. (2024) [24] propose a generalized S-MAD algorithm leveraging deep representations from a ViT. Unlike traditional CNN-based methods, the ViT-based approach integrates local and global information using self-attention, which is well-suited for detecting subtle and distributed morphing artifacts. The ViT-L model, pre-trained on ImageNet [86], is used to extract features from face images, which are classified using a linear SVM for morph detection. The evaluation was conducted on datasets containing different morphing algorithms, including GAN-based and landmark-based methods, with different image types: digital, P&S, and print-scanned-compressed. The study used cross-dataset testing to evaluate the generalizability of the method to unseen scenarios, and intra-dataset testing to evaluate performance in known contexts. Metrics such as D-EER and MACER, equivalent to the ISO/IEC-defined APCER, at fixed BPCER were used. The method was benchmarked against SOTA algorithms, including both feature-based and deep-learning-based approaches. Experiments were conducted to evaluate the system's ability to generalize to different morphing techniques and operational scenarios, especially in digital contexts. Additionally, statistical analysis, including mean and standard deviation of D-EER values, were performed to quantify variability and robustness across datasets and morphing techniques.

The model proposed by Zhang et al. (2024) [24] shows the potential of transformer-based architectures for MAD. Using a ViT, the method integrates both local and global

information through self-attention, which is well-suited for detecting subtle and distributed morphing artifacts. This approach explores how transformer models can improve the robustness and adaptability of MAD systems, especially in the context of digital images.

Hybrid Architectures

Hybrid architectures for MAD combine multiple complementary methods to leverage the strengths of each, thereby increasing the overall robustness and effectiveness of detection models.

Aghdaie et al. (2022) [87] present a face MAD framework that aims to improve detection by leveraging structured group sparsity in conjunction with wavelet domain analysis. This method decomposes face images into 48 spatial-frequency wavelet sub-bands and isolates morphing artifacts in the high-frequency components. A Deep Neural Network (DNN) equipped with structured group sparsity selects the most discriminative wavelet sub-bands during training. By imposing a group sparsity constraint on the weights of the first convolutional layer, irrelevant sub-bands are discarded, leaving only the most informative features for morph detection. The remaining sub-bands are retrained to fine-tune the network, optimizing its ability to distinguish *bona fide* from morphed images. These textural features, combined with the DNN's high-level spatial feature extraction, aim to provide a representation of morphing inconsistencies. The evaluation of this framework was performed using three datasets: VISAPP17 [38], LMA [58], and MorGAN [58]. Two experimental setups were used: a) using individual datasets for training and testing; b) using a universal dataset that combines samples from all three datasets for training, and each of the aforementioned datasets for testing. Metrics such as D-EER, APCER, and BPCER were used in accordance with ISO/IEC 30107-3 standards. Additionally, DET curves were plotted to illustrate the model's performance at various operational thresholds. These evaluations were designed to assess the framework's ability to generalize across datasets and adapt to different morphing techniques and image types.

Di Domenico et al. (2023) [25] propose a combined approach for D-MAD that fuses identity features and artifact analysis, using both D-MAD and S-MAD features. The proposed system uses an Inception ResNet v1 [83] as the S-MAD module, fine-tuned on various morphing datasets, and a ResNet-50 [63] network as the D-MAD module using ArcFace loss for identity verification. The features extracted by both modules are fused, and a Multi-Layer Perceptron (MLP) classifier is used to produce a final morphing detection score. This work builds on the concept of combining identity and artifact features for morphing detection, a direction also explored by Borghi et al. (2021) [33], but extends it by explicitly fusing S-MAD and D-MAD modules to address the limitations of standalone approaches. The evaluation was conducted using several

datasets, including PMDB [57], MorphDB [57], Idiap Morph [39, 56], and FEI [88], which cover a variety of morphing techniques and quality levels. They also developed the FEI-Morphs dataset [89], which is based on the FEI dataset [88]. The dataset contains morphed images generated using three morphing algorithms (FaceFusion, UTW, NTNU) and two morphing factors (0.3 and 0.5). Three experimental scenarios were used: a) assessing performance with the criminal subject; b) with the accomplice subject; c) across both. Metrics such as EER, BPCER, and APCER were used in accordance with ISO/IEC 30107-3 standards. Feature fusion strategies, including Min-Max normalization, were explored to align the contributions of the S-MAD and D-MAD features. Additionally, the cosine similarity of embeddings from the D-MAD module was incorporated to further refine the classification. The purpose of these evaluations was to assess the framework's ability to handle challenging morphing scenarios, especially when the accomplice's image was involved in the comparison.

Jia et al. (2023) [26] propose a MAD method based on a hybrid architecture that combines high-frequency feature extraction with a progressive enhancement learning framework. This hybrid approach merges texture-based and deep learning methodologies to address the challenges of detecting subtle morphing artifacts. High-frequency features are first extracted from the three color channels (RGB) of an image using a Discrete Fourier Transform (DFT) and high-pass filtering. These features, which capture fine-grained texture variations indicative of morphing artifacts, are subsequently fused with standard RGB image data through a two-stream neural network architecture. The progressive enhancement learning framework consists of two modules: a Self-Enhancement Module (SEM) and an Interactive-Enhancement Module (IEM). The SEM enhances the individual features of each stream, while the IEM facilitates cross-stream interaction between RGB and high-frequency data, progressively refining feature representations. This combination allows the model to capture both global and local morphing traces. The evaluation of the proposed method was conducted using the FEI [88] and HNU [90] datasets, which contain different morphing scenarios with varying pixel fusion and position fusion factors. The experiments included a single dataset and cross-dataset tests to assess the generalization ability of the model. Standard metrics such as EER, APCER, BPCER, and ACER were used in accordance with ISO/IEC 30107-3 standards. Ablation studies were also performed to analyze the contributions of the SEM and IEM, in order to investigate their role in improving feature representation and interaction. The evaluations aim to determine the robustness of the framework in capturing subtle morphing traces and its adaptability to different dataset conditions.

Razaq and Shukur (2023) [27] propose a face MAD framework that combines wavelet scatter transform with CNN features. This dual feature extraction approach provides an analysis of morphing artifacts by capturing both subtle textural variations and com-

plex spatial patterns. In the feature extraction phase, the wavelet scatter transform is used to capture frequency-domain features, while a CNN extracts high-level spatial features. These features are fused using Canonical Correlation Analysis (CCA) to provide an enhanced feature set, which is then classified using multiple classifiers, including a DNN and SVM. The evaluation was conducted on a custom dataset created using a combination of publicly available and self-generated morph images, as well as the AMSL dataset [52, 53, 54]. These datasets included different morphing techniques, such as automatic and manual landmark selection, and StyleGAN-based morphing. The model was evaluated using metrics such as ACC, FAR, and FRR, with a focus on minimizing errors in both *bona fide* and morphing classifications. These evaluations aimed to measure the robustness of the combined wavelet-CNN approach, particularly in detecting morphing artifacts in high-resolution images and scenarios with minimal visible artifacts, and to evaluate its suitability for border control applications.

Colbois et al. (2024) [51] propose a hybrid approach for MAD that combines attack-agnostic feature extraction with supervised and one-class classification. The attack-agnostic feature extraction uses pre-trained vision models such as CLIP, DINOv2, and AIM to extract features from *bona fide* images. These attack-agnostic features, which are not specific to any particular attack, are used to detect morphing attacks across different techniques, including landmark-based, GAN-based, and diffusion-based morphs. By using pre-trained models trained exclusively on *bona fide* data, the system can generalize across a variety of morphing methods. For classification, the authors employ two methods: a supervised binary SVM classifier trained on the extracted features and a one-class Gaussian Mixture Model (GMM) trained only on *bona fide* data to detect out-of-distribution morphs. The approach was evaluated on several types of morphing attacks and datasets, including Face Recognition Grand Challenge (FRGC) [91], Flickr-Faces-HQ (FFHQ) [92], and FRLL [72], with a focus on generalization to unseen attacks and datasets, as well as performance on P&S data. The evaluation used D-EER to assess the system's ability to distinguish between real and morphed images in various scenarios.

Liu et al. (2024) [50] propose a D-MAD approach that combines triplet-based metric learning with artifact analysis. This hybrid approach integrates two complementary techniques: identity-based feature extraction and artifact detection. The identity-based feature extraction uses metric learning to separate morphed images from their contributing subjects in the feature space. This is achieved through a triplet loss function, which increases the distance between the morphed image and its source subjects, making it easier to distinguish between genuine and morphed images. In parallel, the artifact detection module focuses on identifying the latent artifacts introduced during the morphing process. These artifacts, such as inconsistencies in lighting, texture, or facial features, are captured and analyzed to further validate whether an image has been

morphed. The outputs of the two modules are processed independently before being combined at the score level to make the final decision. The evaluation of this approach involved training the model on datasets such as PMDB [57], Idiap Morph [39, 56], and MorphDB [57], while FEI Morph [89, 25] was used for testing. The performance of the system was tested in three different experimental scenarios: a) using a trusted live image of the accomplice subject, b) using an image of a criminal subject, and c) a combined ("Both") scenario involving both accomplice and criminal subjects. The evaluation used standard metrics such as BPCER and EER, in accordance with ISO/IEC 30107-3 standards.

The hybrid architectures discussed in [87, 25, 26, 27, 51, 50] are distinguished by their unique approaches to combining different techniques for MAD. Aghdaie et al. (2022) [87] integrate structured group sparsity with wavelet domain analysis, which allows the method to focus on high-frequency components in the wavelet domain, enabling a more targeted detection of morphing artifacts. Di Domenico et al. (2023)[25] adopt a fusion strategy combining S-MAD and D-MAD features, using different neural networks (Inception ResNet v1 [83] for S-MAD and ResNet-50 [63] for D-MAD) and incorporating ArcFace loss for identity verification, creating a more refined approach for morphing detection in difficult scenarios. Jia et al. (2023)[26] propose a two-stream architecture that combines high-frequency feature extraction with RGB image data through a progressive enhancement learning framework, allowing for both individual feature enhancement and cross-stream interaction. Razaq and Shukur (2023)[27] use a dual feature extraction strategy by combining wavelet scattering transform and CNNs, which allows them to capture both frequency-domain and spatial features, with the features fused using CCA for better representation. Colbois et al. (2024)[51] use attack-agnostic features extracted using pre-trained models such as CLIP, DINOv2, and AIM, which are then used for supervised and one-class detection to identify morphing attacks without being biased by specific attack types. Liu et al. (2024)[50] combine triplet-based metric learning with artifact analysis. They integrate identity-based feature extraction, which uses triplet loss to separate morphed images from their contributing subjects, with artifact detection to identify inconsistencies introduced by morphing.

Taken together, these hybrid models highlight the potential of combining different techniques and capabilities to advance MAD. Each study presents a distinct approach to integrating identity and artifact information, underscoring the value of hybrid strategies in addressing the complexities of MAD across diverse datasets and challenging scenarios.

Other Approaches

Other approaches include noise analysis, deep learning-based artifact analysis, shape-based detection methods, and additional deep learning strategies that do not fit neatly into the previously discussed categories.

Singh and Ramachandra (2022) [34] present a D-MAD method tailored for On-The-Fly (OTF) border control scenarios with varying image resolutions and capture distances. The proposed algorithm integrates Spherical Linear Interpolation (SLERP) with hierarchical fusion techniques to enhance the detection of morphing attacks. Features are extracted from six pre-trained CNN models, including AlexNet [62], VGG16 [69], VGG19 [69], Xception [93], ResNet50 [63], and ResNet101 [63]. These features undergo hierarchical fusion at both score and feature levels to combine complementary information. The evaluation was performed using the SCFace-Morph dataset, derived from the SCFace [94] dataset, which simulates real-world border control conditions with variations in camera resolutions and capture distances. Separate evaluation protocols were developed for intra- and cross-medium scenarios, comparing digital and P&S images. Metrics such as EER, APCER, and BPCER were used, following the ISO/IEC 30107-3 standards.

Robledo-Moreno et al. (2024) [95] propose a MAD approach based on Federated Learning (FL), a technique that enables the training of models across distributed locations without centralizing sensitive biometric data, in order to preserve privacy. This approach addresses privacy concerns associated with biometric data sharing, a key challenge in the field. The authors evaluate FL for both S-MAD and D-MAD tasks and compare its performance to traditional centralized training methods. For the S-MAD task, an Inception-ResNet [96, 83] model is trained on images generated by several morphing algorithms. In the D-MAD task, the authors use ArcFace [64] embeddings to compare trusted and suspect images for detecting morphing attacks. The evaluation process involves training the models across multiple locations in a federated environment and evaluating their ability to detect morphing attacks. The evaluation uses standard metrics for MAD, such as BPCER and EER, to measure the performance of the FL-based models in comparison to centralized models.

Singh and Ramachandra (2024) [28] propose three different 3D Morphing Attack Detection (3D-MAD) algorithms to detect 3D face morphing attacks. These algorithms are based on deep features extracted from pre-trained 3D point-based CNN models: PointNet [65], PointNet++ [97], and SimpleView [98]. PointNet [65] works directly on 3D point clouds, extracting features from the classification layer. PointNet++ [97] builds upon this approach by using hierarchical networks to capture more detailed geometric structures, while SimpleView [98] uses multi-view projections of point clouds, extracting a 40-dimensional feature vector from each depth map for detection analysis. These detection methods take advantage of depth information to better distinguish

between *bona fide* and morphed face models. To evaluate the proposed 3D-MAD algorithms, the authors developed a dataset (3DMD) containing both *bona fide* and morphed 3D face models. This dataset includes 3D point clouds of 41 individuals, which were processed using various 3D morphing techniques to create morphed models. Additionally, the publicly available Facescape [99, 100, 101] dataset was used to extend the evaluation with a broader set of subjects and 3D face models. The evaluation used APCER, BPCER, and D-EER as performance metrics, following the ISO/IEC 30107-3 standards. The features extracted by PointNet [65], PointNet++ [97], and SimpleView [98] were classified using a linear SVM. DET curves were also plotted to analyze the trade-off between APCER and BPCER at different thresholds.

The approaches reviewed here highlight a variety of strategies for MAD, each offering a unique perspective. Singh and Ramachandra (2022) [34] integrate deep features from multiple CNNs using hierarchical fusion, which enables effective detection across different image resolutions and challenging OTF border control settings. Robledo-Moreno et al. (2024) [95] introduce a FL-based approach for MAD that addresses privacy concerns by enabling model training across distributed locations without centralizing sensitive biometric data. Singh and Ramachandra (2024) [28] focus on 3D morph detection, providing insights into the use of depth information to improve detection robustness in both 2D and 3D scenarios.

Taken together, these studies illustrate the diversity of approaches that can be employed to address MAD, ranging from GAN-based identity reconstruction to unsupervised learning techniques, demonstrating the potential for innovative solutions beyond traditional supervised learning paradigms.

2.3.4 Experimental Setup

This section presents the hardware configurations and training procedures adopted in different studies. Table 2.1 summarizes the key specifications, including GPU models, batch sizes, learning rates, training strategies, number of epochs, and training times, for the papers that reported them.

The training setups reported across the reviewed studies exhibit notable diversity in architecture types, hardware configurations, hyperparameters, and methodological strategies. The key findings are summarized below:

- **Hardware configurations** range from high-end setups to more modest consumer-grade GPUs:
  - High-performance GPUs were used by Blasingame and Liu (2021) [80] (4x NVIDIA V100), Aghdaie et al. (2022) [87] (2x TITAN X (Pascal)), and Chaudhary et al. (2021) [49] (3x TITAN X (Pascal)).
  - In contrast, Jia et al. (2023) [26] and Robledo-Moreno et al. (2024) [95] used

Table 2.1: Training setups and strategies.

Author	Architecture	Hardware	Batch Size	Learning Rate	Training Strategy	Epochs	Training Time
Aghdaie et al. (2022) [87]	Hybrid	2x 12 GB TITAN X (Pascal) GPUs	-	0.001[1]	Learning rate scheduling, no batch size or training time details	150	-
Blasingame and Liu (2021) [80]	GAN	4x NVIDIA V100 GPUs	256	0.00001	Full training with 200 epochs	200	17 hours
	Combinator			0.001	Network trained with 150 epochs	150	6 hours
Chaudhary et al. (2021) [49]	Siamese Network	3x 12 GB TITAN X (Pascal) GPUs	64[2]	0.0001	Learning rate reduction on validation loss plateau; early stopping after 35 epochs without improvement	150	-
			32[3]				
Jia et al. (2023) [26]	Hybrid	GeForce GTX 1060Ti	16	0.0001	Standard training with small batch size	20	-
Robledo-Moreno et al. (2024) [95]	Other Approaches	Nvidia GeForce GTX 1070 6GB (per site)	-	0.001	FL using FLARE; model updates and aggregation each epoch	-	-

[1] Learning rate divided by 10 every 20 epochs.
[2] For BW-22.
[3] For RGB-66.

consumer-grade GPUs (GTX 1060Ti and GTX 1070, respectively).

- **Architectures** varied across studies:
  - Blasingame and Liu (2021) [80] evaluated two models: a GAN, trained for 200 epochs, and a combinator, trained for 150 epochs.
  - Aghdaie et al. (2022) [87] and Jia et al. (2023) [26] used hybrid architectures.
  - Chaudhary et al. (2021) [49] implemented a siamese network.
  - Robledo-Moreno et al. (2024) [95] explored other approaches under a FL framework.
- **Learning rates** and their adaptation strategies were also varied:
  - Blasingame and Liu (2021) [80] applied 0.00001 for the GAN and 0.001 for the Combinator model.
  - Aghdaie et al. (2022) [87] began with 0.001, reducing it by a factor of 10 every 20 epochs.
  - Chaudhary et al. (2021) [49] and Jia et al. (2023) [26] both used a fixed learning rate of 0.0001.
  - Robledo-Moreno et al. (2024) [95] also used 0.001, with no further strategy specified.
- **Batch sizes** were reported as follows:
  - Blasingame and Liu (2021) [80] used a batch size of 256 across both models.
  - Chaudhary et al. (2021) [49] reported batch sizes of 64 for BW-22 and 32 for RGB-66.
  - Jia et al. (2023) [26] used a small batch size of 16.
  - Aghdaie et al. (2022) [87] and Robledo-Moreno et al. (2024) [95] did not report batch size information.

- **Training durations and epoch counts** varied:
  - Blasingame and Liu (2021) [80] trained their GAN for 200 epochs (17 hours) and their Combinator model for 150 epochs (6 hours).
  - Aghdaie et al. (2022) [87] and Chaudhary et al. (2021) [49] both trained for 150 epochs, though only the latter included early stopping after 35 epochs without improvement.
  - Jia et al. (2023) [26] trained for only 20 epochs, likely constrained by hardware.
  - Robledo-Moreno et al. (2024) [95] did not specify training duration or number of epochs.
- **Training strategies** reflected the experimental goals and limitations of each study:
  - Aghdaie et al. (2022) [87] and Chaudhary et al. (2021) [49] incorporated learning rate scheduling based on validation performance.
  - Blasingame and Liu (2021) [80] performed full supervised training for both architectures.
  - Jia et al. (2023) [26] used standard training with a small batch size.
  - Robledo-Moreno et al. (2024) [95] applied a FL framework (FLARE), aggregating model updates after each epoch across decentralized sites.

In summary, the reviewed studies showcase diverse training setups, varying in hardware, architecture, learning rates, and strategies. These differences highlight the flexibility of approaches used to meet specific research goals and constraints, offering insights into optimizing future training configurations.

2.3.5 Results

This section outlines the key findings reported in the reviewed literature on MAD. By focusing on the results, it provides a concise summary of the performance metrics and conclusions drawn from various studies. These findings will serve as a foundation for the discussion that follows, where trends, strengths, and limitations of MAD techniques are analyzed in more detail.

Reported Results from Reviewed Literature

Table D.3 in Annex D presents the best performing results obtained in each reviewed paper for easier comparison. Note that results are not directly comparable, as they are obtained on different test sets. For results from standard benchmarks, we refer to Table 2.2.

The study by Aghdaie et al. (2022) [87] highlights the effectiveness of structured

group sparsity in selecting discriminative features within the wavelet domain, facilitating robust morph detection in diverse datasets. Texture analysis plays a pivotal role in this approach, as the decomposition into wavelet sub-bands captures subtle, localized texture variations that are characteristic of morphing artifacts. While the individual dataset training setup excels at detecting intra-dataset morphs, especially in VISAPP [38] and MorGAN [58], the universal dataset training also shows strong generalization to unseen morphing techniques in the aforementioned datasets, with slightly less satisfactory results on LMA [58].

The method proposed by Banerjee and Ross (2021) [78] achieves comparable results to SOTA approaches despite not requiring explicit supervision or labeled training data to discriminate between morphed and non-morphed images. In particular, it outperforms existing methods in cross-dataset evaluations when trained on the AMSL [52, 53, 54] dataset and tested on the MorGAN [58] dataset, surpassing previous benchmarks. However, the method exhibits reduced performance when trained on the MorGAN [58] dataset and tested on AMSL [52, 53, 54] dataset, which may be due to the low correlation and overall biometric utility of the images in the MorGAN [58] dataset. The method also shows strong potential for decoding the second identity in morphed images.

The approach proposed by Blasingame and Liu (2021) [80] showed strong performance across different experimental setups, excelling in 35 out of 36 experiments. It showed remarkable resilience to changes in datasets and morphing algorithms, although performance decreased when both variables were changed simultaneously. The incorporation of the combinator into the training process improved the effectiveness of the method over alternatives without it. However, the contribution of the combinator was ultimately overshadowed by the advantages provided by the GAN framework and triplet loss, which played a more central role in the success of the method. This suggests that the performance improvements are less about the combinator itself and more about the unique training approach used. While the method showed reduced generalization to high-quality StyleGAN2 [102] morphs, this was not attributed to a failure of the model but rather to the inherent strength of these morphs, which limits detectable artifacts. Interestingly, when trained on StyleGAN2 [102] morphs, the method performed exceptionally well against weaker morphs such as FaceMorpher [103] and OpenCV [104], supporting the hypothesis that models trained on more sophisticated morphs can generalize to weaker ones.

The framework proposed by Borghi et al. (2021) [33] achieved strong performance in D-MAD, particularly in the "criminal" scenario, where identity features were more discriminative. However, its effectiveness varied between datasets due to differences in the composition of morphs. In the MorphDB [57] dataset, morphed images closely resembled the accomplice, whereas in the AMSL [52, 53, 54] dataset, morphs exhibited

a balanced mix of criminal and accomplice features. This highlights the adaptability of the framework to different morphing scenarios. Furthermore, the results emphasize that the performance of the framework is closely related to the quality of the training dataset. When artifacts are less prominent, such as in JPEG2000-compressed images, or when the framework encounters unseen image generation techniques, its performance declines, thus highlighting the need for diverse datasets that include different morphing tools to improve generalization. The study demonstrated the importance of combining identity- and artifact-related features for robust detection. While the individual modules of the framework, focusing on identity and artifacts respectively, performed worse in isolation, the integration through the final fully connected layer was critical. This layer balanced the contributions of the modules, allowing the framework to achieve high ACC across datasets.

Chaudhary et al. (2021) [49] obtained a particularly strong performance of the RGB-66 on the MorGAN [58] dataset, surpassing that of BW-22. This highlights the role of color information in detecting GAN-based morphs, which inherently contain more color details than landmark-based morphs such as LMA [58] and VISAPP [38]. However, the wavelet analysis reveals that color information is less critical in this domain, as evidenced by the strong performance of BW-22, especially on challenging morphs from the LMA dataset [58]. The VISAPP dataset [38] showed consistent performance across all frameworks, likely due to its smaller size and uniform morphing technique, making it easier for models to learn. In contrast, LMA [58] morphs were harder to detect, consistent with previous findings that landmark manipulation morphs pose greater detection challenges than GAN-generated morphs. The framework also performed well on an unseen dataset, AMSL [52, 53, 54], demonstrating the robustness of wavelet decomposition with selective sub-bands in detecting morphs across different techniques.

In the study by Colbois et al. (2024) [51], the baseline experiment showed that all methods achieved near-perfect results on the constrained FRGC [91] dataset, although their effectiveness declined when applied to the more diverse FFHQ [92] dataset. Generalization to unseen attacks confirmed the adaptability of the model, while detectors trained on the FFHQ [92] dataset generally outperformed those trained on FRGC [91] when tested on different source datasets. However, the P&S scenario proved challenging, as artifacts were degraded during the printing and scanning process, which affected the detection performance. One-class detection was highly effective on the FRGC [91] dataset but underperformed on FFHQ [92] data. Among the features tested, DNADet excelled at one-class digital modeling, but struggled with P&S generalization due to its pretraining. AIM features performed well for unseen attacks but were inconsistent in other scenarios, limiting their broader applicability. DINOv2 excelled at P&S generalization, making it suitable for scenarios where training on P&S data is im-

practical. CLIP features, while rarely the best, performed consistently well across all scenarios, making them very versatile for complex, multi-challenge environments.

Damer et al. (2021) [48] demonstrate that the proposed PW-MAD framework outperforms baseline methods in both intra-dataset and cross-dataset evaluations, particularly in scenarios involving re-digitized morphing attacks. The pixel-wise supervision approach allowed the framework to focus on localized manipulation effects, improving generalizability to unseen variations, including re-digitized morphs. In the most challenging settings, such as training on digital morphs and testing on re-digitized ones, PW-MAD achieved lower error rates compared to existing methods, demonstrating its robustness.

The results obtained by Di Domenico et al. (2023) [25] show that while D-MAD generally outperforms S-MAD when the live image contains the criminal, its performance deteriorates when the accomplice is present due to greater identity similarity. The proposed fusion method addresses this limitation by integrating S-MAD's artifact-based features and D-MAD's identity-based features. However, combining these features proved to be non-trivial. The study showed that features from S-MAD and D-MAD occupy different regions in the feature space, and normalization using the Min-Max strategy improved performance. Incorporating cosine similarity as an additional feature further improved detection, provided it was kept in its original domain. Experimental results indicate that the proposed method outperformed competitors in scenarios involving the accomplice, highlighting the value of S-MAD features in improving D-MAD performance. Conversely, performance with the criminal remains an area for improvement, suggesting the need for tailored fusion methods and separate handling of criminal and accomplice scenarios.

Fang et al. (2022) [75] conclude that the inclusion of the SPL paradigm improves the performance of the unsupervised SPL-MAD approach. Models trained with SPL consistently outperformed those trained without SPL in various scenarios, including cases with uncontaminated and contaminated training data. While the benefit of SPL was less evident for certain types of attacks, such as FRLL-based [52, 56, 39] attacks with contaminated training data, the overall result underscores its important role in improving MAD performance. The SPL-MAD approach demonstrated comparable results to supervised methods while maintaining greater consistency and avoiding extreme variations in performance. Additionally, its unsupervised nature allowed for effective treatment of unknown morphing attacks, demonstrating higher generalizability compared to supervised methods.

Ferrara et al. (2021) [35] show that the effectiveness of MAD varies depending on the type of images and pre-training of the networks. For digital morphs, networks pre-trained on natural images, such as AlexNet [62] and VGG-19 [69], perform well, using general filters that detect digital artifacts and manipulation traces. However, these net-

works struggle with P&S images because the P&S process removes most digital artifacts, rendering *bona fide* and morphed images indistinguishable. In contrast, networks pre-trained on large face datasets, such as the VGG-Face [70] models, perform better on P&S images due to their ability to detect fine, face-specific features that survive the printing and scanning process. Incorporating simulated P&S images into training further improves performance, allowing networks like VGG-Face16 [70] to adapt to the specific challenges of P&S morph detection, although AlexNet [62] does not benefit from this additional training. Finally, while the proposed approach has shown superiority over existing methods in specific benchmarks such as National Institute of Standards and Technology (NIST) Face Recognition Vendor Test (FRVT) MORPH [105] and SOTAMD [41], overall performance remains modest.

Jia et al. (2023) [26] conclude that the proposed approach effectively enhances MAD by leveraging high-frequency features and a progressive enhancement learning framework. High-frequency features proved superior to RGB-based features in distinguishing between *bona fide* and morphed images, capturing subtle morphing traces that RGB color space has difficulty identifying. The fusion of high-frequency and RGB features in a two-stream network outperformed single-branch networks, highlighting the benefits of combining different feature types for comprehensive representation. The inclusion of self-enhancement and interactive-enhancement modules further improved feature interaction and representation, enabling the detection of subtle morphing artifacts. Experiments demonstrated robust performance across datasets with varying pixel and position fusion factors, highlighting the adaptability of the method. Cross-dataset evaluations confirmed the strong generalizability of the approach, which maintained high effectiveness even when tested on unseen data.

The model proposed by Liu et al. (2024) [50] shows superior performance in the identification of morphed images, particularly in scenarios involving accomplices, while maintaining a slightly lower but competitive effectiveness in criminal cases. It excels in the "Both" scenario and achieves superior results that confirm its overall reliability and applicability in practical contexts. The authors attribute its success to the innovative use of metric learning to disrupt identity pathways, and the integration of artifact features, which increases the stability of the model and ensures consistent performance across different evaluation scenarios.

Medvedev et al. (2023) [22] conclude that the proposed MorDeephy framework demonstrates robust performance and generalizability in detecting face morphing attacks. It effectively generalizes to unseen scenarios, including GAN-generated morphs and P&S morphs, highlighting its adaptability. The use of selfmorphs during training reduces the reliance on detecting blending artifacts and shifts the focus to deep feature behavior, which enhances its ability to handle diverse and previously unseen attack types. The method outperforms other approaches in detecting landmark-based and manually

generated morphs while maintaining stable performance across protocols, without bias toward specific generative morphing techniques. Importantly, the method achieves these results without extensive modifications for realistic scenarios, highlighting the effectiveness of focusing on the authenticity of deep facial features for morph detection.

Medvedev et al. (2024) [23] build on the previous study. The approach outperforms baseline methods and demonstrates strong generalization capabilities, achieving robust performance across diverse datasets and unseen scenarios without bias towards specific morphing generative strategies. The V2 strategy, which separates morphs from their original classes, further improves performance compared to the V1 strategy, where morphs are linked to their original classes. The integration of face recognition signals with the fused classification detector improves detection, achieving superior results in all benchmarks tested.

The results reported by Ortega-Delcampo et al. (2020) [47] indicate that superior MAD results are consistently obtained with digital images compared to P&S images, which perform similarly regardless of resolution. Furthermore, the results surpass those reported in other studies in the literature for both digital and P&S images. The research also confirms the suitability of the CNN-based approach for real-world scenarios, providing robust and accurate detection capabilities. Finally, the study highlights the adaptability of the method to ABC system requirements and its potential for broader applications, such as revealing hidden imposter identities in morphing attacks.

In the experiments conducted by Qin et al. (2022) [71], the proposed method achieves reliable detection for complete, spliced and combined morphs when training and testing on the same morphing technique, with slightly higher error rates observed for retouched morphs. The method demonstrates effective localization of morphed regions, although it struggles with complete morphs due to their fully manipulated nature. D-MAD consistently outperforms S-MAD by using probe image information to enhance detection features. The method generalizes well to unseen morphing attacks, particularly in D-MAD, while S-MAD encounters higher error rates for complete morphs, again due to their uniquely tampered nature. It also exhibits stable performance in cross-dataset scenarios, with little degradation caused by variations in datasets, imaging conditions, and morphing parameters. Additionally, the model shows resilience to low-resolution probe images but experiences reduced performance with non-frontal images due to reduced biometric information. Ablation studies demonstrate the importance of components such as feature-wise differential loss that help improve detection ACC and enhance generalization to unseen morphing attacks and varied datasets.

Razaq and Shukur (2023) [27] demonstrate the effectiveness of combining wavelet transform features and high-level features from a CNN for face MAD. DNN outperformed all other classifiers, closely followed by SVMs. Other classifiers, such as K-

Nearest Neighbors (KNN), Decision Trees, and Naïve Bayes, showed lower performance. The proposed method also performed effectively on the AMSL [52, 53, 54] dataset, identifying morphs with high ACC, even when the changes targeted specific facial landmarks such as eyes, nose, or mouth.

Robledo-Moreno et al. (2024) [95] showed that FL performs comparably to centralized approaches, with better performance when the same morphing technique is used for both training and testing. In S-MAD, FL effectively handles unbalanced datasets and benefits from the inclusion of low-quality morphs, but struggles with high-quality morphs. In D-MAD, FL often outperforms centralized methods, particularly when artifacts in the training data degrade the performance of centralized models. Nevertheless, centralized methods have lower maximum error rates and benefit from augmentation techniques not explored in this study.

The approach proposed by Singh and Ramachandra (2022) [34] achieves optimal results in intra-evaluation protocols where the same medium (digital or P&S) is used for training and testing, demonstrating robustness across different cameras and capture distances, illustrating its robustness in real-world scenarios. In cross-evaluation protocols, the method performs better when trained on digital data and tested on P&S images but struggles when the training and testing mediums are reversed, likely due to differences in image quality between digital and P&S data.

Singh and Ramachandra (2024) [28] demonstrate the feasibility of MAD in 3D images, with the SimpleView [98] network achieving the best performance, highlighting its strength in using depth-based features to detect 3D morphing attacks.

Soleymani et al. (2021) [84] demonstrated that fine-tuning their proposed Siamese network on the training portions of the VISAPP17 [38] and MorGAN [58] datasets improved its detection capabilities, reducing error rates. However, error rates remained higher for the MorGAN [58] dataset, reflecting the inherent difficulty of the dataset due to its lower resolution and generative properties. Texture descriptors outperformed key-point-based approaches in capturing morphing artifacts. The Siamese network's ability to detect subtle differences between images was further strengthened through pre-training on the WVU Twins Day dataset[2] [107, 108, 109], resulting in improved detection ACC. Additionally, the integration of SVM-based decision-making frameworks, using feature concatenation and difference further enhanced the model's performance.

The model proposed by Venkatesh et al. (2020) [46] achieves the best detection performance across three datasets in the intra-dataset experiment. This superior performance is attributed to its innovative aggregated denoising and fusion scheme, which uses sub-band selection for optimal results, and the robustness of MS-CAN in extracting noise features from natural images. The research highlights the advantages of deep fea-

[2]The WVU Twins Day Dataset 2010–2015 is only available to US nationals and researchers based in the USA [106].

tures over non-deep feature-based techniques, with AlexNet [62] and ResNet101 [63] achieving the best results among the deep learning models. Moreover, the method shows robustness to increasing sample sizes and varying image characteristics, confirming its adaptability to diverse datasets. It also improves computational efficiency by consolidating multiple denoising operations into a single coherent framework.

The method proposed by Zhang et al. (2024) [24] achieves superior performance for digital images with the lowest mean and standard deviation of D-EER among the tested methods, showing an improvement in robustness. However, performance degrades for P&S and print-scan-compression scenarios due to the pre-training of the model on digital images and the lower resolution of P&S data. The approach demonstrates strong generalization capabilities in cross-dataset testing, particularly for GAN-based morphs, where features are well-separated from *bona fide* samples. In contrast, landmark-based morphs are more difficult to discriminate due to their less separable features. They also conclude that post-processing of morphs increases the strength of attacks, making them difficult to detect.

Benchmark Evaluations from the Reviewed Literature

This subsection presents the performance of selected approaches when evaluated on common benchmarks, following consistent evaluation protocols. These benchmark results offer direct comparability between methods under controlled conditions and complement the reported results extracted from the literature.

There are two main platforms: Bologna Online Evaluation Platform (BOEP) [110], which is part of the FVC-OnGoing framework [111, 112], and NIST Face Analysis Technology Evaluation (FATE) MORPH [113]. Each platform offers multiple benchmarks (e.g., SOTAMD [114] or iMARS for BOEP, and different evaluation tiers for NIST).

FVC-onGoing [111, 112] is a long-standing benchmarking platform developed by the Biometric System Laboratory at the University of Bologna. Originally focused on fingerprint recognition, it has expanded to encompass broader biometric challenges, including face morphing detection [115].

Building upon this infrastructure, the BOEP [110] was developed to offer a standardized evaluation environment specifically tailored for MAD algorithms. BOEP [110] facilitates the submission and automated evaluation of algorithms, ensuring consistent and reproducible assessments across different MAD approaches [110].

Within this ecosystem, the SOTAMD benchmark [114] was introduced to evaluate MAD systems under realistic and challenging conditions, such as varying image qualities and morphing techniques. SOTAMD leverages the BOEP framework to provide a testing environment that reflects operational scenarios encountered in border control and identity verification processes [110].

In parallel, NIST FATE Morph [113] is an independent benchmark developed by the United States NIST as part of the FATE program, formerly part of the FRVT framework. This benchmark focuses on evaluating face morphing detection algorithms in high-security identity verification contexts, such as border control and credential issuance. It employs proprietary datasets and rigorously defined evaluation protocols, emphasizing operational effectiveness, robustness to image quality variations, and resistance to diverse morphing techniques. Although developed independently from European frameworks like BOEP and SOTAMD, NIST FATE Morph serves as a complementary international benchmark, enabling broader cross-national comparisons and fostering standardization across institutions engaged in MAD research [113].

Table 2.2 presents the best performing results from the reviewed papers that report performance on standardized MAD benchmarks.

Table 2.2: Summary of the best results reported from each study included in the literature review on standardized MAD benchmarks. For comparison purposes, only the best results per evaluation system were selected.

Evaluation System	MAD Type	Benchmark	Reviewed Literature	BPCER (%) @ APCER 1%	BPCER (%) @ APCER 5%	BPCER (%) @ APCER 10%	APCER (%) @ BPCER 1%	D-EER (%)
NIST	S-MAD	UNIBO Automatic Morphed Face Generation Tool v1.0	Ferrara et al. (2021) [35]	-	-	-	5	-
FVC	S-MAD	SMAD-SOTAMD_-P&S-1.0		-	100.00	100.00	-	37.10
		SMAD-BIOLAB-1.0	Zhang et al. (2024) [24]	-	-	78.41	-	40.45
	D-MAD	DMAD-SOTAMD_-D-1.0	Di Domenico et al. (2023) [25]	-	-	-	-	∼10
			Liu et al. (2024) [50]	34.1	16.8	8.0	-	9.2
		DMAD-SOTAMD[1]	Robledo-Moreno et al. (2024) [95]	-	-	28.8	-	18.2

[1] Not further specified.

Ferrara et al. (2021) [35] show strong performance with VGG-Face16 fine-tuned on digital images from MorphDB, achieving high ACC (97.4%) and low error rates (0.0% BPCER). However, performance drops on P&S images, with BPCER rising to 43.8% at 1% APCER, and ACC falling to 93.5%. While the model performs well on the NIST S-MAD benchmark using the UNIBO Automatic Morphed Face Generation Tool v1.0, achieving an APCER of 5% at a BPCER of 1%, this result may be misleading. On four other benchmarks within the same NIST S-MAD evaluation, the model records an APCER as high as 91% at the same BPCER threshold. Furthermore, in the FVC S-MAD SMAD-SOTAMD_P&S-1.0 test, BPCER hits 100% at 5% and 10% APCER, reflecting a complete failure in attack detection.

Zhang et al. (2024) [24] present moderate performance in controlled scenarios but

struggle under standardized evaluation. In a cross-technique evaluation on digital morphs from the FRGC dataset, their method achieves a D-ERR of 13.63%, indicating a fair ability to detect morphs generated through varying techniques. However, in the FVC S-MAD SMAD-BIOLAB-1.0 benchmark, which simulates more realistic and diverse conditions, performance degrades substantially, with a BPCER of 78.41% at 10% APCER and a D-ERR of 40.45%. These results highlight a limited generalization capability when moving from controlled datasets to real-world applications.

Di Domenico et al. (2023) [25] report consistent but moderate performance across both controlled and standardized settings. When trained on a universal dataset and tested on FEI accomplice images using Min-Max feature fusion and cosine similarity, their method achieves a D-ERR of 12.5% and a BPCER of 46.8% at 1% APCER, improving to 23.7% at 5% APCER, indicating challenges in correctly classifying *bona fide* images under stricter thresholds. In the standardized FVC D-MAD DMAD-SOTAMD D-1.0 benchmark, the model performs slightly better with a D-ERR of approximately 10%, suggesting reasonable generalization and reliability in more diverse and realistic conditions.

Liu et al. (2024) [50] achieve strong results in a controlled criminal scenario, with a D-ERR of 3.7% and a low BPCER of 11.5% at 1% APCER, improving to just 2% at 5% APCER, demonstrating effective morph detection and minimal false rejections. However, in the standardized FVC D-MAD DMAD-SOTAMD D-1.0 benchmark, performance drops, with BPCER increasing to 34.1% at 1% APCER and a D-ERR of 9.2%, indicating reduced robustness and difficulty in maintaining ACC under more realistic and varied conditions.

Robledo-Moreno et al. (2024) [95] demonstrate outstanding performance in a cross-technique evaluation for S-MAD, achieving 0.00% BPCER and D-ERR at both 1% and 5% APCER, suggesting highly effective morph detection under controlled technique variation. However, such perfect detection may also indicate possible overfitting, especially given the contrast with other evaluations. In a cross-dataset evaluation for D-MAD, performance decreases notably, with BPCER rising to 34.3% at 1% APCER and a D-ERR of 7.1%, pointing to limited generalization across different data sources. In the standardized FVC D-MAD DMAD-SOTAMD benchmark—where the type of images used (digital, P&S, or others) is not specified—performance drops further, with BPCER reaching 28.8% at 10% APCER and D-ERR increasing to 18.2%, suggesting reduced robustness under more realistic and varied conditions.

Moving on to direct comparisons between studies using the same benchmark, the FVC D-MAD DMAD-SOTAMD D-1.0 benchmark, conducted on high-resolution digital images, provides an interesting contrast. Di Domenico et al. (2023) [25] achieves a D-ERR of approximately 10%, while Liu et al. (2024) [50] slightly outperforms with a D-ERR of 9.2%. Liu's results also provide more information, with BPCER values of 34.1% at 1%

APCER, 16.8% at 5%, and 8% at 10%, showing a notable improvement in performance as more attack errors are tolerated. Di Domenico et al. (2023) [25], while similar in overall detection performance, does not report BPCER values, limiting the ability to assess its handling of *bona fide* samples. Both studies exhibit moderate robustness on high-resolution digital images, with Liu et al. (2024) [50] offering a slightly better balance between attack detection and user convenience.

2.4 Findings and Trends

The literature review uncovers several findings and emerging trends in the application of deep learning techniques to MAD, providing valuable insights into how these techniques can improve MAD performance. Building on these findings, this section directly addresses the Research Question: "How can deep learning techniques be used to improve MAD?" The following key points summarize the ways in which deep learning techniques can improve MAD:

- **Feature Extraction and Hybrid Approaches**: Feature extraction is a critical step in improving MAD. Deep learning models, particularly CNNs, automate this process, allowing the identification of subtle traces of manipulation that traditional methods may miss, thus improving the detection ACC for various morphing techniques. According to Venkatesh et al. (2020) [46], deep learning techniques provide features that yield better performance, allowing models to better distinguish between morphed and genuine images. This highlights the importance of using advanced methods that capture subtle patterns or inconsistencies in morphed images. Several studies — Aghdaie et al. (2022) [87], Chaudary et al. (2021) [49], and Razaq and Shukur (2023) [27] — have shown how models use wavelet domain analysis to identify discriminative features. In addition, the combination of identity and artifact-based features, as demonstrated by Borghi et al. (2021) [33], proves effective for robust detection. The trend of incorporating multiple feature types, such as high-frequency and RGB image data, as seen in Jia et al. (2023) [26], is gaining traction for its ability to improve detection ACC across different morphing techniques.
- **Cross-Dataset Generalization**: A key trend is the ability of deep learning models to generalize across different morphing techniques and datasets. Damer et al. (2021) [48] highlight the importance of generalization for real-world applications, where unseen morphing techniques frequently occur. However, Banerjee and Ross (2021) [78] and Blasingame and Liu (2021) [80] showed that many models trained on one dataset often struggle when tested on another, especially when different morphing techniques are used. In addition, variations in environmental factors such as lighting, angle, and distance across datasets can further

hinder the generalizability capability of these models, as they are often trained on controlled datasets that do not adequately capture such real-world conditions.

- **Unsupervised and Semi-Supervised Learning**: Unsupervised learning methods have emerged as a critical trend in MAD, enabling the detection of unknown morphing techniques without requiring large labeled datasets. Studies such as those by Fang et al. (2022) [75] and Banerjee and Ross (2021) [78] highlight the benefits of unsupervised approaches that improve flexibility and detection of new morphing attacks.

- **GAN-Based and Landmark-Based Morphs**: A notable trend is the challenge of detecting high-quality landmark-generated morphs compared to GAN-based morphs. Chaudhary et al. (2021) [49] and Zhang et al. (2024) [24] found that landmark-based morphs are harder to detect due to more subtle artifacts created by manipulating facial features. On the other hand, GAN-based morphs produce more pronounced features that are easier to detect. This points to the need for deep learning models that are capable of detecting both strong and weak morphs across different types of attacks.

- **Performance Variations with Image Types**: The performance of MAD techniques varies considerably depending on the image type. The studies by Ferrara et al. (2021) [35] and Ortega-Delcampo et al. (2020) [47] demonstrate that digital images typically yield better results than P&S images, which are often used in real-world scenarios such as passport verification. The challenge posed by P&S morphs, which may lack detectable artifacts, highlights the importance of training models on different image types to ensure robustness and improve detection in different scenarios.

- **Advanced Morphing Techniques**: With the increasing sophistication of morphing techniques, the development of MAD models that can keep up with these advances has become a major concern. Enhanced post-processing and similar techniques aim to enhance the quality of morphed images, making them more difficult to detect. Zhang et al. (2024) [24] discuss how these attacks complicate the detection process, requiring more advanced methods capable of detecting subtle traces of morphing.

In summary, deep learning techniques have played a critical role in advancing MAD by automating feature extraction, improving generalization across different datasets, and enabling the identification of images created by advanced morphing methods. Approaches that incorporate hybrid models, unsupervised learning, and multi-modal features continue to evolve and provide solutions to the growing complexity of morphing attacks.

2.5 Challenges and Gaps

This section discusses the challenges and gaps identified in the reviewed studies. Despite advances in MAD, several limitations remain in the proposed methods. The key challenges that hinder the effectiveness and applicability of current MAD techniques are:

- **Generalization to Unseen Morphing Techniques**: Many existing methods perform well on specific datasets or types of attacks but struggle when faced with novel morphing techniques, particularly those that involve subtle or complex image manipulation. Furthermore, the ability of these methods to detect morphing attacks in real-world scenarios remains a concern, especially when dealing with low-quality or compressed images commonly encountered in border control systems and document verification applications.
- **Reliance on Digital Images**: The reliance on digital images alone is another gap, as many real-world applications, such as passport verification or border control, involve scanned or printed images. These images often suffer from low resolution or compression artifacts, which can reduce the ACC of morph detection techniques designed primarily for high-quality digital images. The performance of existing methods under these conditions is often untested and, when tested, unsatisfactory, leaving a critical gap in their real-world applicability.
- **Privacy Concerns with Synthetic Datasets**: While synthetic datasets like SMDD [59], MONOT [60], and AMONOT [61] are essential for protecting privacy by avoiding the use of real biometric data, they come with their own set of challenges. Ensuring that synthetic data accurately mimics real-world biometric data while maintaining privacy is difficult. There is always a risk that synthetic data could be reverse-engineered or traced back to real individuals, potentially compromising privacy. Additionally, synthetic datasets may not fully capture the diversity and variability of real-world data, which could limit the effectiveness and fairness of MAD systems across different populations and morphing techniques [116].
- **Reliance on Controlled Datasets**: Another major limitation is the reliance on controlled datasets. While datasets with a variety of morphing techniques are useful, there is still a lack of diversity in the subjects used for training. Most studies focus on limited demographic groups, with insufficient representation of subjects of different ages, genders, and ethnicities. Furthermore, the majority of datasets are collected under controlled conditions that often lack variability in environmental factors such as lighting, angle, and distance. This lack of diversity in training data can affect the model's ability to generalize to a broader range of real-world situations.

- **Handling Sophisticated Morphing Methods**: Many of the advanced detection techniques, while promising, face difficulties in handling more sophisticated morphing methods that produce fewer detectable artifacts. As morphing techniques evolve and become more advanced, the artifacts produced are often subtle and harder to detect, which can reduce the effectiveness of existing detection systems. In addition, several methods rely on specific image properties, such as wavelet domain features or pixel-wise analysis, which may not generalize well to all types of morphs, particularly those with minimal detectable distortions. Furthermore, recent studies have shown that attackers are employing face restoration techniques, such as CodeFormer [117], to automatically retouch morphed images, removing artifacts and enhancing visual quality. This process results in high-quality morphed images that are more challenging for detection systems to identify [118].

Overall, while the reviewed studies have made important strides in detecting face morphing attacks, future research must focus on addressing these challenges, in particular, improving the robustness and generalization of MAD techniques, increasing the diversity of datasets, and improving detection capabilities for sophisticated, low-artifact morphs.

2.6 Future Research Directions

Despite advances in MAD, several challenges remain. The following areas offer promising opportunities for future research:

- **Generalization to Unseen Morphing Techniques**: Future research should focus on improving model generalization across different morphing techniques and expanding training datasets to include a wider variety of morphing methods.
- **P&S Images**: Research should aim to develop techniques that perform well under conditions of resolution loss, compression artifacts, and distortion commonly found in real-world applications such as passport control.
- **Expanding Dataset Diversity**: Future studies should aim to include subjects of different ages, genders, and ethnicities, as well as high-quality and P&S images, to improve model generalization and real-world applicability.
- **Addressing Dataset Bias and Domain Adaptation**: Future work should explore unsupervised domain adaptation techniques (e.g., Domain-Adversarial Neural Networks [119]) and image translation methods (e.g., CycleGAN [120]) to reduce performance gaps between training on high-quality images and testing on compressed, scanned, or low-resolution samples. Domain generalization strategies that do not require access to target domain data can further improve adapt-

ability. Additionally, integrating bias mitigation techniques—such as sample reweighting or fairness-aware augmentation—can help counter demographic imbalances and improve real-world performance [121, 122].

- **Robustness to Subtle Artifacts**: Future research should focus on developing models capable of detecting subtle morphing traces using more advanced features, possibly combining pixel-level, frequency-domain, and texture-based methods. The challenge is further exacerbated by the use of face restoration techniques, which reduce or mask typical morphing artifacts, making them less distinguishable to conventional detection systems. As a result, detection methods must become increasingly sensitive to these imperceptible alterations introduced by both sophisticated morphing and post-processing strategies.
- **Real-Time and Cross-Dataset Evaluation**: To improve real-world applicability, future studies should evaluate MAD systems under real-time conditions with varying image resolutions, lighting, and environmental factors. In addition, models should be evaluated across different datasets to ensure that they perform efficiently and accurately in real-world applications such as border control and identity verification.
- **Unsupervised and Semi-Supervised Learning**: Future research could explore unsupervised and semi-supervised approaches, which can help overcome the challenges of acquiring large labeled datasets.
- **Privacy-Oriented and Adaptive Training Strategies**: In addition to synthetic datasets for privacy preservation, future research should explore training strategies that protect sensitive biometric data. Techniques like FL allow models to be trained across multiple institutions without the need to share raw data. Continual and incremental learning approaches enable systems to adapt to emerging morphing techniques over time, maintaining performance without the need for complete retraining or access to previous data. These strategies enhance both privacy and the adaptability of MAD systems [123].
- **Video-Based MAD in Operational Scenarios**: Recent work has explored the use of video data for MAD, introducing methods that leverage temporal and dynamic facial features. This direction, exemplified by the Video-based Morphing Attack Detection (V-MAD) approach, opens new avenues for detecting morphs in realistic settings such as border control, where video sequences are naturally available and can provide richer information than still images [124].
- **Explainability and Interpretability of MAD Models**: Future research should explore explainability techniques, such as Grad-CAM, SHapley Additive exPlanations (SHAP), and feature attribution methods, to improve the transparency and trustworthiness of deep learning-based MAD systems. Although most cur-

rent approaches focus on detection performance, understanding the underlying decision-making processes can enhance reliability in critical applications like border control. The broader machine learning community has also emphasized the importance of interpretability in complex models [125], highlighting its relevance as a complementary direction for advancing MAD research.

By focusing on these areas, future research can address existing gaps and pave the way for more robust and effective MAD systems.

2.7 Conclusions

This literature review analyzes the current state of deep learning techniques for detecting morphed face images, focusing on the effectiveness of different methods and models used in recent research. The results highlight the progress made in MAD, particularly through the use of CNNs, GANs, and autoencoders. These deep learning models have proven effective in automating feature extraction, identifying subtle traces of manipulation, and improving detection performance across different morphing techniques.

Despite these advances, the detection of morphed images still faces several challenges. As morphing techniques evolve and become more sophisticated, the artifacts created by these attacks become less detectable, particularly in high-quality morphs. In addition, while current methods show promising results in controlled environments, their performance often degrades when applied to real-world scenarios involving low-resolution, compressed, or printed images.

Furthermore, issues related to dataset diversity and model generalization across different morphing techniques remain relevant barriers. Many models perform well on specific datasets but struggle when faced with new, unseen morphing attacks, highlighting the need for more robust and generalized detection systems. In addition, reliance on labeled datasets poses limitations, and some methods are exploring unsupervised learning to fill this gap.

In summary, deep learning techniques have proven valuable in improving MAD systems. However, additional research is needed to improve the robustness and generalizability capability of these systems, particularly in detecting subtle morphing artifacts and applying them to practical applications such as identity verification and border control. Future efforts should focus on expanding datasets, improving model adaptability, and developing methods capable of detecting increasingly advanced morphing techniques. Overcoming these challenges will make MAD systems more reliable in protecting against morphing-based identity fraud.

The experimental work presented in this document directly addresses several of these

key challenges through systematic investigation of cross-dataset generalization (see Section 4.1 and Section 4.2), preprocessing strategies and their impact on real-world applicability (see Section 4.1 and Section 4.2), class balancing techniques for handling imbalanced datasets (see Section 4.2), and model interpretability through the development of an occlusion-based framework that reveals which facial regions contribute most to detection decisions (see Section 4.2 and Chapter 5).

3 Methodology

This chapter presents the methodology that was applied to investigate MAD using deep learning techniques. The methodology includes the selection of datasets, the replication and evaluation of existing methods, the implementation of ViTs, and the application of occlusion mapping to enhance model interpretability. It was designed to address the gaps identified in the literature review, with a particular focus on evaluating ViT-based approaches for detecting morphing attacks in 2D face images.

Figure 3.1 provides a high-level overview of the research methodology, outlining the progression from the initial literature review to the experimental phase and concluding with interpretability analysis.

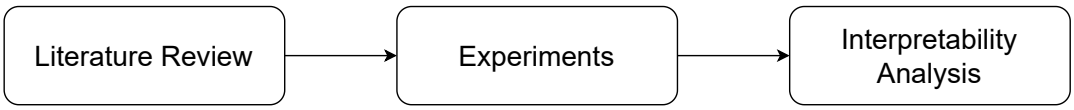


Figure 3.1: Methodology workflow: overview.

3.1 Use of Publicly Available Code

To replicate prior experiments described in the literature and facilitate performance benchmarking, an extensive search was conducted to identify publicly available implementations. Figure 3.2 illustrates the systematic process followed to identify, assess, and select publicly available codebases for replication. Among the studies reviewed, source code or experimental platforms were made available in only four cases:

- **Fang et al. (2022) [75]**: A complete and functional implementation was provided. Their pipeline was successfully used to replicate experiments, with SMDD employed as the training dataset and evaluations performed on both SMDD and FRLL-Morphs.
- **Medvedev et al. (2023) [22]**: A benchmarking platform was released; however, since the modification and extension of model architectures was a key objective of the current research, the platform was deemed unsuitable and was not used.
- **Colbois et al. (2024) [51]**: Although the source code was made available, persistent technical issues prevented successful reproduction of the experiments. The authors were contacted, but the problems could not be resolved, and therefore the work was excluded from this study.

- **Singh and Ramachandra (2024) [28]**: Their work focused on morphing detection using 3D face images. As the current research is limited to 2D image analysis, the approach was not further considered.

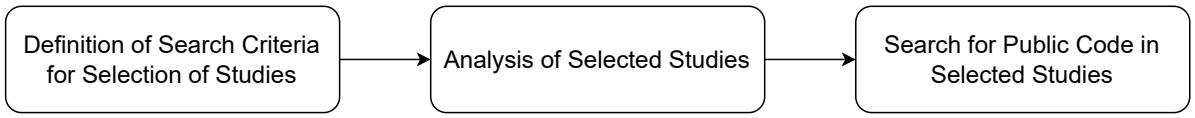


Figure 3.2: Methodology workflow: literature review flowchart.

Ultimately, only the implementation by Fang et al. (2022) [75] was successfully used to conduct experiments. This influenced the selection of datasets, as detailed in Section 3.3.

3.2 Vision Transformer Exploration

While publicly available implementations were being assessed, ViTs were identified during the literature review as a promising yet underexplored architecture in the context of MAD. Among the reviewed studies, ViTs were only investigated by Zhang et al. (2024) [24], whose results indicated competitive performance.

Although their code was not publicly accessible, the authors were contacted, and access to their ViT-based implementation was obtained. This enabled further experimentation to be carried out, specifically using the following dataset combinations:

- FRLL vs. FRLL-Morphs.
- FFHQ vs. FFHQ-Morphs.

The FRGC dataset, originally used in Zhang et al. (2024) [24], was not available for distribution and could not be used. Nonetheless, the experiments conducted on FRLL and FFHQ datasets allowed the generalization performance of ViT-based models to be evaluated in a cross-domain setting.

3.3 Dataset Selection

The selection of datasets was influenced by the experiments conducted using both the publicly available implementation of Fang et al. (2022) [75] and the privately shared implementation of Zhang et al. (2024) [24]. Figure 3.3 illustrates the experimental workflow applied to each selected dataset. Datasets were chosen based on the following criteria:

- Their prior use in experiments described in the reviewed literature.
- The presence of both *bona fide* and morphed face images.

- Their suitability for replicating and benchmarking key experimental configurations.

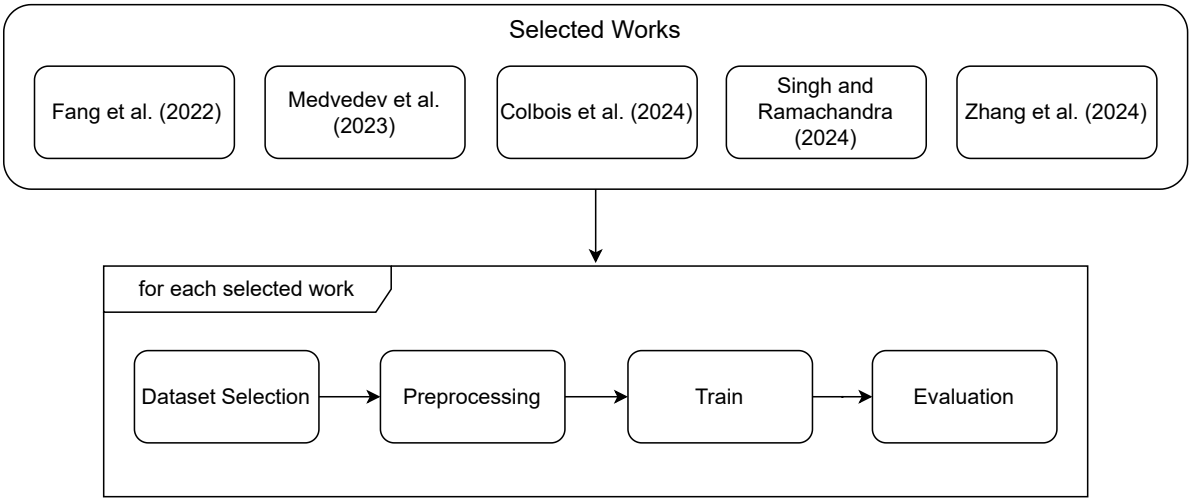


Figure 3.3: Methodology workflow: experiments flowchart.

The datasets used are summarized in Table 3.1, which details the number of images, image resolution, overall dataset size, and principal characteristics of the dataset, with representative sample images provided in Annex E.

In addition, attempts were made to obtain access to other relevant datasets:

- **MorGAN**: Contact was initiated with the authors, but no response was received. As a result, the dataset was not used.
- **FRGC**: Although this dataset was used by Zhang et al. (2024) [24], it is no longer distributed and is therefore unavailable.

In summary, the dataset selection was guided by practical availability and alignment with the experimental pipelines that could be successfully replicated or accessed. This ensured that the conducted experiments were grounded in the literature and could be reproduced under the existing constraints.

3.4 Dataset Preprocessing

The preprocessing stage focused on face cropping, aiming to refine the facial region in each image by removing background distractions and ensuring proper alignment of facial features.

Each image underwent an automatic face detection and cropping procedure to extract a fixed region of interest centered on the face, ensuring consistent facial framing across datasets. The procedure employed Multi-task Cascaded Convolutional Networks (MTCNN) [126] for face detection. For each input image, MTCNN localized the facial Bounding Box (bbox), and the image was cropped to this region to isolate

Table 3.1: Summary of datasets used.

Dataset	Number of Images	Image Resolution	Dataset Size	Description
FRLL [72]	~1,000	1350×1350 pixels	~290 MB	High-quality 2D *bona fide* face images captured under controlled conditions, featuring variations in pose, age, gender, and ethnicity.
FRLL-Morphs [52, 56, 39]	~7,000	Varies by morphing technique: 413×531, 431×513, or 1024×1024 pixels	~240 MB	Morphed images generated from FRLL using GANs and traditional landmark-based morphing techniques.
FFHQ [92]	~70,000	1024×1024 pixels	~90 GB	Large-scale dataset of high-quality face images sourced from Flickr, emphasizing demographic diversity.
FFHQ-Morphs [51]	~25,000	Varies by morphing technique: 256×256 or 1024×1024 pixels	~13 GB	Morphed counterparts of FFHQ images, generated using diffusion-based and other advanced morphing algorithms.
SMDD [59]	~118,000	256×256 for uncropped images and 224×224 for cropped images	~12 GB	Privacy-friendly synthetic dataset consisting of *bona fide* and morphed images, pre-split into training and test sets.

the face from background distractions. An exception was made for the *FFHQ* dataset, which was not cropped due to its large size (approximately 90 GB).

Figure 3.4 illustrates an example of this face cropping workflow, showing a raw image before processing and its corresponding cropped version.

This step is essential to provide consistent and focused input data for the subsequent analysis and model training, ensuring consistency across all datasets and enabling a more fair comparison between the evaluated morphing detection approaches.

3.5 Model Explainability and Interpretability

In recent years, the explainability and interpretability of deep learning models has received increasing attention, particularly in security-critical applications such as biometric recognition and MAD. Deep neural networks, while highly accurate, often behave as black boxes, making it difficult to understand the rationale behind their decisions. In safety-sensitive domains, this lack of transparency can hinder trust, auditability, and

Figure 3.4: Example of dataset preprocessing applied to the image `32950_05186` from the *FFHQ-Morphs* dataset, which was generated using a diffusion-based morphing algorithm.

the ability to detect systematic biases or failure modes.

Several studies have emphasized the value of interpretability techniques for explaining model behavior and ensuring responsible deployment [127, 128, 129]. In the context of MAD, Borghi et al. (2021) [33] demonstrated the use of Grad-CAM [129] to visualize the regions of the face that contributed the most to classification decisions, highlighting how such tools can be used to assess both the focus of the model and bias.

Other popular techniques for model interpretability include gradient-based methods such as SmoothGrad [130], feature attribution frameworks such as SHAP [131], and perturbation-based methods that assess the influence of individual input components [132]. Each of these techniques offers complementary insights into model behavior and decision-making processes.

To address this need for transparency in MAD systems, occlusion-based techniques were employed in this work to interpret the outputs of the model used in Experiment 5 (see Section 4.2). Figure 3.5 illustrates the interpretability analysis workflow, showing how occlusion mapping techniques are applied to generate heatmap visualizations.

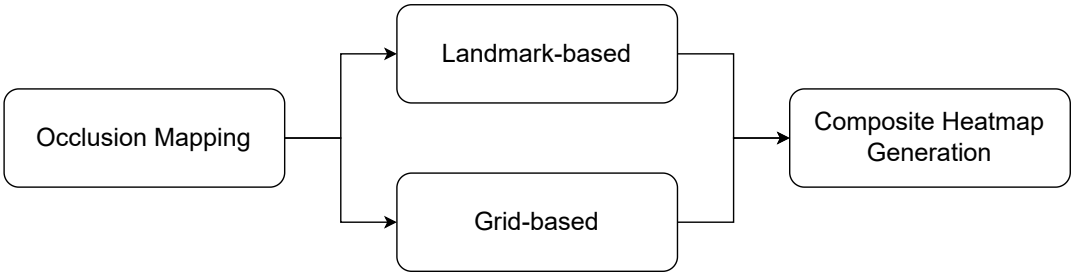


Figure 3.5: Methodology workflow: interpretability flowchart.

3.5.1 Occlusion Mapping Methodology

Occlusion mapping is a perturbation-based interpretability method that provides insights into which regions of an input image are most important for a model's predictions. The technique works by systematically masking parts of the input image and measuring the resulting change in the model output. A drop in confidence following the occlusion of a region suggests that the model relied heavily on that area to make its decision.

Two occlusion strategies, illustrated in Figure 3.6 and described below, were implemented. Each strategy produces its own heatmap visualization representing the importance of different regions. These individual heatmaps are then combined to generate a composite heatmap that integrates insights from both occlusion approaches.

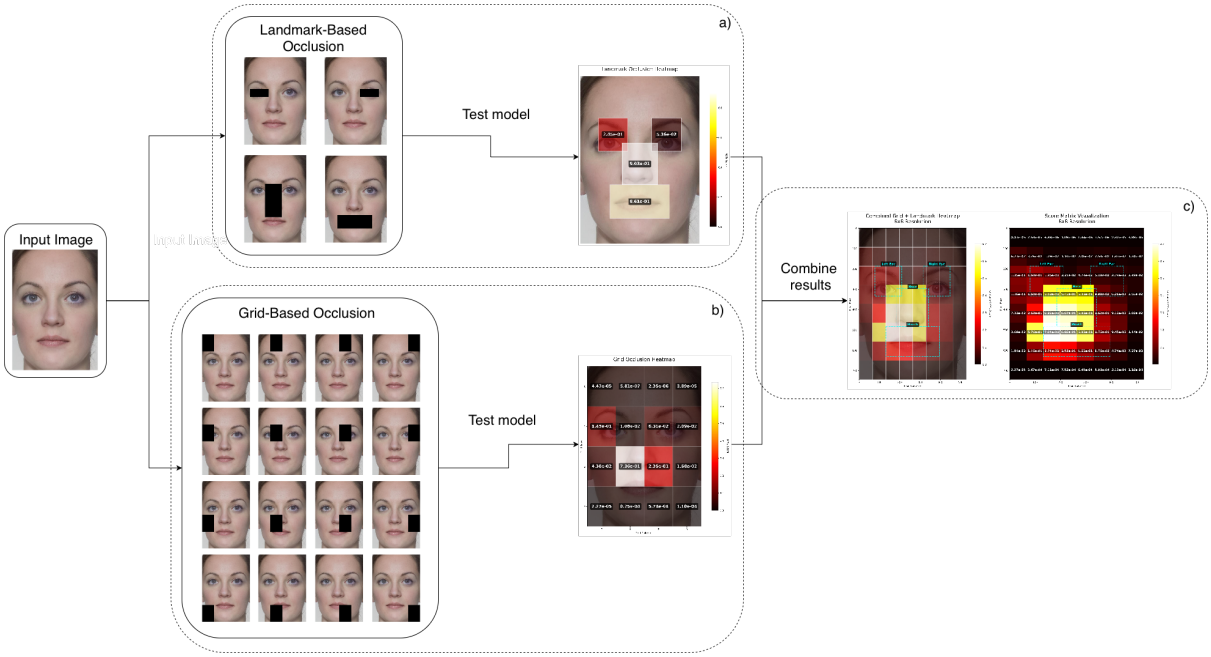


Figure 3.6: Overview of the occlusion mapping methodology. The original cropped image is used to generate multiple occluded variants via (a) landmark-based occlusion and (b) grid-based occlusion. Each occluded image is processed by the trained MAD model, and the resulting prediction scores are recorded. The importance maps from both strategies are fused (c), producing the final *composite heatmap*, which is overlayed on the original cropped image for easier visualization.

Landmark-Based Occlusion

In this approach, occlusions were applied to semantically meaningful regions based on facial landmarks. For each test image, four key areas were occluded with a black rectangle, one at a time:

- Left eye.
- Right eye.

- Nose.
- Mouth.

Each occluded image was processed by the trained model and the resulting changes in prediction scores were recorded. This allowed an estimation of the relative importance of these facial components in the model classification process.

Grid-Based Occlusion

A second, more systematic strategy was based on a spatial grid. Each image was divided into a 4×4 grid, resulting in 16 regions of equal size. One region was occluded at a time with a black square, generating 16 occluded versions of each test image.

By analyzing the change in prediction scores for each occluded version, a spatial importance map was constructed. This map reflects which areas of the image contributed most to the model's output, enabling a finer-grained understanding of spatial dependencies beyond semantic regions.

3.5.2 Heatmap Visualization

To enhance interpretability, three distinct heatmap visualizations were generated for each test image:

- **Landmark-based heatmap**: Derived from the landmark-based occlusion strategy, highlighting the importance of semantically meaningful facial regions (eyes, nose, and mouth).
- **Grid-based heatmap**: Derived from the grid-based occlusion strategy, providing a spatial importance map across the entire face based on a uniform grid division.
- **Composite heatmap**: Created by fusing the outputs of both occlusion techniques, combining semantic and spatial information to provide a comprehensive view of which regions most influenced the model's decision.

In all three heatmap visualizations, the score values are normalized on a scale from 0 to 1, where 0 corresponds to the morph class and 1 corresponds to the *bona fide* class. High values, shown in bright yellow to white, indicate regions that, when occluded, result in the image being classified as *bona fide*, while low values, depicted in dark red to black, correspond to regions that, when occluded, result in the image being classified as morphed. For morphed images, regions highlighted with high values denote critical areas whose occlusion drives the model toward a *bona fide* decision, thereby identifying the regions essential for correct morph classification.

To increase resolution and semantic richness in the composite heatmap, the original 4×4 grid was refined to 8×8, and its scores were combined with landmark-based occlusion results via averaging.

Examples of all three heatmap visualizations are shown in Figure 3.7. The landmark-based and grid-based heatmaps display the heatmap and score matrix overlaid directly on top of the original face image, while the composite heatmap is shown with the heatmap overlaid onto the face image in the left panel and the corresponding score matrix for each grid cell in the right panel. It can be observed from these visualizations that the mouth and central nose area were identified as the most critical regions for the model's prediction.

3.6 Evaluation Metrics

Following the definitions introduced in Subsection 2.1.3, model performance was evaluated using two categories of metrics, combining traditional performance measures with interpretability tools.

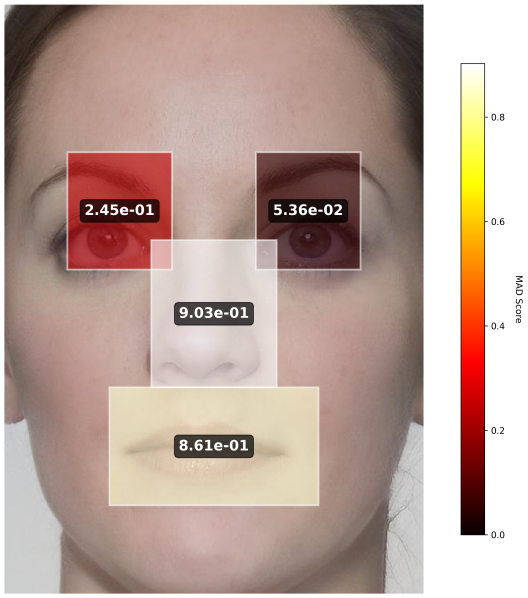
3.6.1 Error Rates

- **D-EER**: The point at which the APCER and BPCER are equal. This threshold balances false acceptances and false rejections, commonly used in biometric evaluations (see Subsection 2.1.3).
- **APCER @ BPCER (1%, 5%, 10%)**: These values indicate the percentage of attack images incorrectly classified as *bona fide* when the system is tuned to reject a fixed proportion of *bona fide* samples (see Subsection 2.1.3). These metrics provide insight into system performance under different security thresholds.
- **ACC**: The percentage of correctly classified samples (see Subsection 2.1.3). Although ACC is less informative in imbalanced datasets, it is included for comparison purposes with other studies.

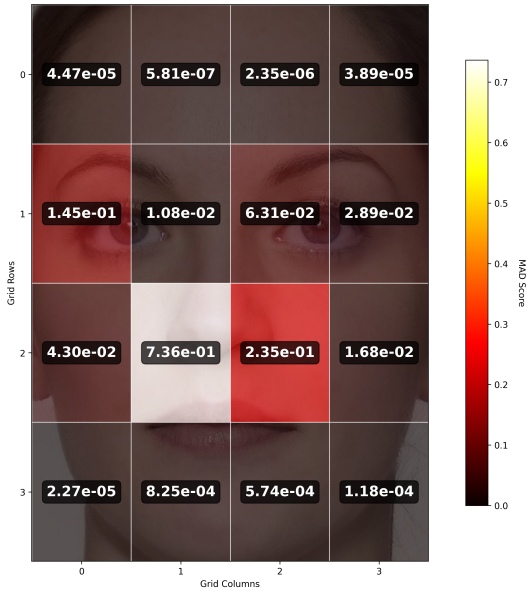
3.6.2 Explainability Tools

- **Occlusion Mapping**: Used as a diagnostic tool to understand which facial regions the model relies on. Landmark-based and grid-based occlusions, has explained in Subsection 3.5.1, were applied to generate qualitative insights and regional importance heatmaps.

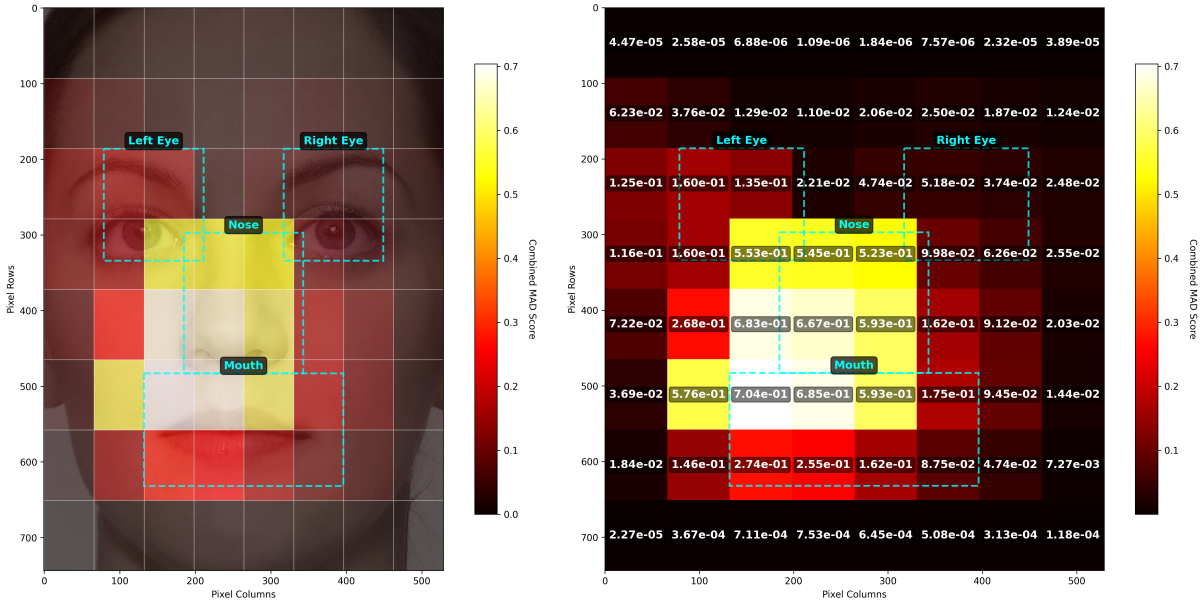
These metrics and tools together provided a comprehensive evaluation of the system's performance and interpretability across both known and cross-domain conditions.



(a) Landmark-based heatmap.



(b) Grid-based heatmap.



(c) Composite heatmap (8×8): left, without MAD scores; right, with MAD scores superimposed.

Figure 3.7: Three heatmap visualizations generated from the occlusion mapping methodology: (a) landmark-based, (b) grid-based, and (c) composite heatmap showing the importance of facial regions to the model's predictions.

3.7 Conclusions

This chapter presented the methodology used to investigate MAD using deep learning techniques, with particular emphasis on ViT-based architectures and model interpretability. The methodology was structured to address the research gaps identified in the literature review, particularly the limited exploration of ViTs in MAD and the need

for enhanced transparency in model decision making.

The approach was grounded in the replication and extension of existing research, leveraging publicly available implementations from Fang et al. (2022) [75] and the privately shared ViT-based implementation from Zhang et al. (2024) [24]. This strategy facilitated performance benchmarking against established baselines while enabling exploration of novel architectural approaches.

Dataset selection was guided by practical availability and alignment with reproducible experimental pipelines. The use of FRLL, FRLL-Morphs, FFHQ, FFHQ-Morphs, and SMDD ensured that experiments were grounded in datasets commonly used in the literature, enabling meaningful comparisons with prior work. The preprocessing pipeline, centered on face detection and cropping using MTCNN, ensured consistency across datasets and enabled fair comparisons between different detection approaches.

A key contribution of this methodology lies in the integration of interpretability techniques through occlusion mapping. Using both landmark-based and grid-based strategies, this approach provided valuable insights into which facial regions the models relied upon for classification, addressing the black box nature of deep learning models in security-critical applications.

The evaluation framework employed traditional error-rate metrics (D-EER, APCER at fixed BPCER levels, and ACC), ensuring assessment of both operational effectiveness and interpretability. The following chapters present the experimental results obtained using this methodology and discuss their implications for face MAD.

4 Experiments

The literature review presented in Chapter 2 identified a diverse range of deep learning approaches for MAD, including CNNs, Siamese Networks, Transformer-based models, and hybrid architectures. While these studies demonstrate promising detection capabilities across various experimental settings, a critical limitation emerged during the review: the majority of published research does not provide access to the source code or trained models, making independent validation and reproducibility challenging.

To address this gap and gain practical insights into the real-world performance of SOTA MAD techniques, this chapter focuses on the empirical evaluation of methods for which implementations were available. The objective is twofold: first, to replicate and validate the reported results under controlled conditions; and second, to extend the original experiments by exploring additional datasets, preprocessing strategies, and evaluation protocols that were not covered in the original studies.

Among the 23 papers reviewed in Chapter 2, only a limited subset made their implementations publicly accessible. After a thorough search of code repositories and supplementary materials, the following studies were identified as having available source code:

- Fang et al. (2022) [75].
- Medvedev et al. (2023) [22].
- Colbois et al. (2024) [51].
- Singh and Ramachandra (2024) [28].

In addition to these publicly available implementations, a fifth study was included in the experimental evaluation. Zhang et al. (2024) [24] proposed a ViT-based approach combined with SVM classification for MAD. Although the implementation was not publicly released, the authors were contacted directly and generously provided access to the source code, along with instructions and clarifications regarding the experimental setup. This work was particularly relevant to include, as it represents the only Transformer-based architecture among the reviewed literature.

It is important to note that not all replication attempts were successful. As documented in Annex F, some implementations could not be reproduced due to incomplete or non-functioning code, missing dependencies, insufficient documentation, or incompatibility with the scope of this research. These challenges underscore the importance of code availability and reproducibility standards in advancing the field.

The following sections describe the experiments conducted with success (Experiment 1 and 5 in Table 4.1), outlining the adaptations made, the training configurations used, and the results obtained.

Table 4.1: Summary of experiments performed.

	Experiment 1	**Experiment 2**	**Experiment 3**	**Experiment 4**	**Experiment 5**
MAD Type	S-MAD	S-MAD	S-MAD	D-MAD	S-MAD
Goal	Self-Paced Anomaly Detection for MAD	Framework for evaluating MAD methods	Explore the use of attack-agnostic features	Create and identify 3D face morphing attacks	Generalized MAD
Model	CAE	CNNs	Pre-trained models (CLIP, DINOv2) + SVM and GMM	Pre-trained CNN models + SVM	ViT + SVM
Source Paper	Fang et al. (2022) [75]	Medvedev et al. (2023) [22]	Colbois et al. (2024) [51]	Singh and Ramachandra (2024) [28]	Zhang et al. (2024) [24]
Section	Section 4.1, Annex G	Annex F	Annex F	Annex F	Section 4.2, Annex H, Annex I
Result	Successful	Unsuccessful	Unsuccessful	Unsuccessful	Successful

4.1 Experiment 1 — Fang et al. (2022)

The goal of this experiment was to evaluate the MAD method proposed by Fang et al. (2022) [75], which incorporates SPL into the training process to enhance model robustness. This technique emulates the human learning process by starting with easier samples and progressively introducing more challenging ones, allowing the model to filter out uncertain examples and refine its ability to distinguish between *bona fide* and morphed images. The implementation used in this experiment corresponds to the publicly available version released by the authors.

Figure 4.1 provides an overview of the workflow followed in this experiment, illustrating the main stages from dataset preparation to model training and evaluation.

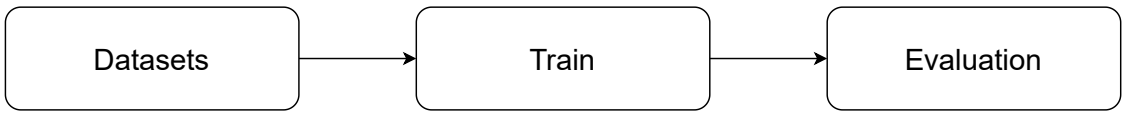


Figure 4.1: Experiment 1: methodology overview.

4.1.1 Datasets

This experiment used two publicly available datasets: **SMDD** [59] and **FRLL-Morphs** [52, 56, 39]. The SMDD dataset [59] was used to train the model for a total of 25 epochs. For testing, the model was evaluated under both intra-dataset conditions using SMDD [59] and cross-dataset conditions using FRLL-Morphs [52, 56, 39], allowing the as-

sessment of detection performance and generalization ability across different dataset scenarios. Figure 4.2 shows the datasets used in this experiment.

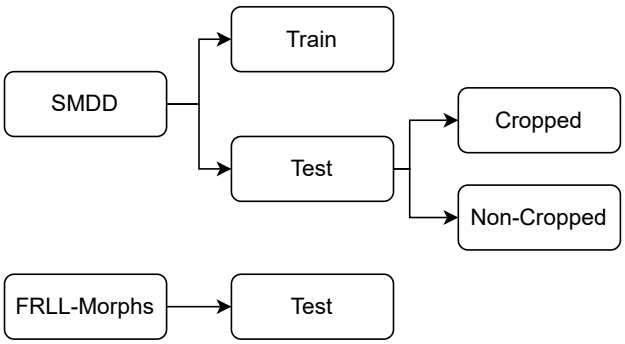


Figure 4.2: Experiment 1: datasets.

4.1.2 Class Balancing

No class balancing was performed in this experiment, in order to maintain consistency with the original study. The training split of the SMDD dataset [59] comprises 15,000 morphed images and 25,000 *bona fide* images.

4.1.3 Evaluation Metrics

Performance was assessed using the **D-EER** metric (see Subsection 2.1.3).

4.1.4 Experimental Setup

This experiment was conducted using *Google Colab* as the computational environment, using an *NVIDIA A100 GPU* to accelerate both model training and evaluation.

4.1.5 Experimental Results

The evaluation was designed to replicate and validate the results reported by Fang et al. (2022) [75]. A more thorough explanation of the experiment is available in Annex G through analysis of the developed code.

During training, the model was saved at the end of each of the 25 epochs, as depicted in Figure 4.3. Each saved model was then evaluated on the test set; however, only the performance obtained at the final epoch is reported, as shown in Figure 4.4. To provide additional insight into the model's learning behaviour, Figures 4.5 and 4.6 illustrate the evolution of the D-EER across training epochs for the SMDD [59] and FRLL-Morphs [52, 56, 39] test datasets, respectively.

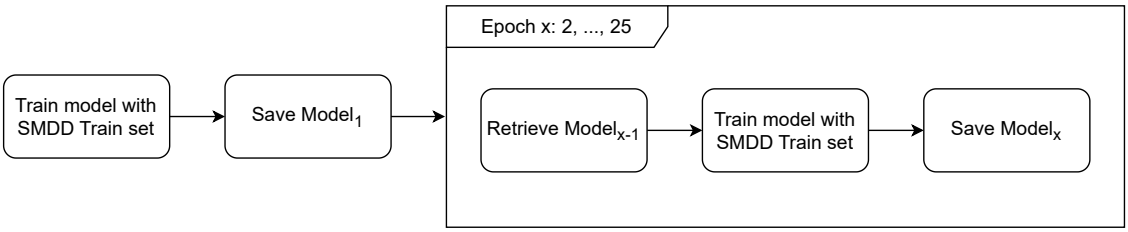


Figure 4.3: Experiment 1: train flowchart.

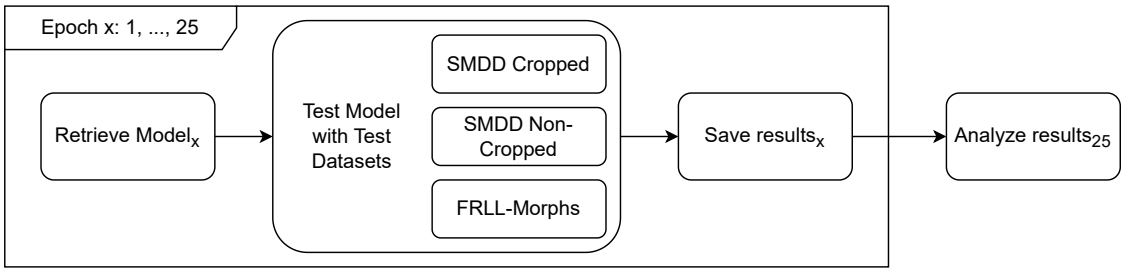


Figure 4.4: Experiment 1: evaluation flowchart.

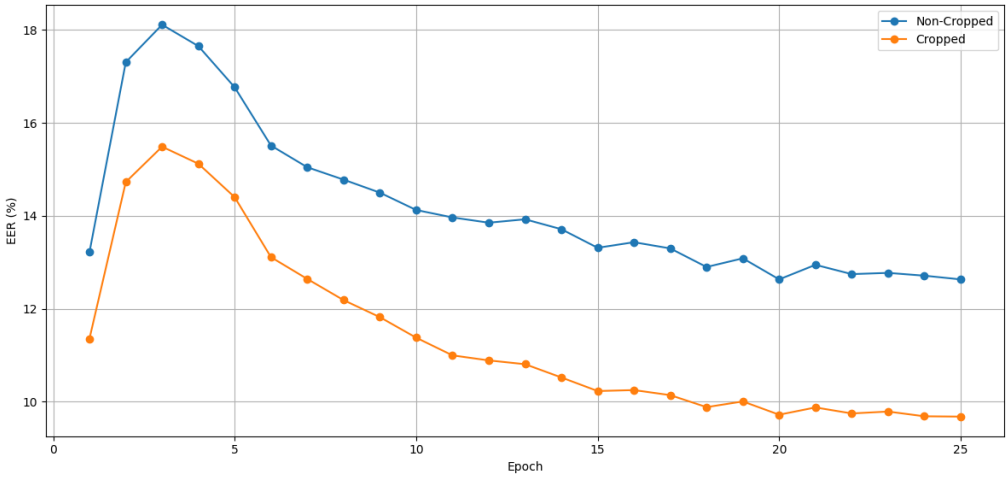


Figure 4.5: Evolution of D-EER across training epochs of the model trained and evaluated on the SMDD dataset [59].

As shown in Figure 4.5, the model exhibits steady improvement when evaluated on the SMDD test set. After a slight increase during the initial epochs, the D-EER decreases consistently as training progresses, reaching 9.68% for cropped images and 12.63% for non-cropped images at the final epoch (see Table 4.2). This behaviour is consistent with an intra-dataset configuration, where the model is trained and tested on data from the same distribution, confirming that it effectively learns discriminative features.

The learning curve for FRLL-Morphs [52, 56, 39], however, reveals a contrasting outcome. As illustrated in Figure 4.6, the D-EER increases steadily throughout all epochs, indicating a progressive degradation in performance. This suggests that the model is unable to generalize to the morphing techniques used in the FRLL-Morphs dataset

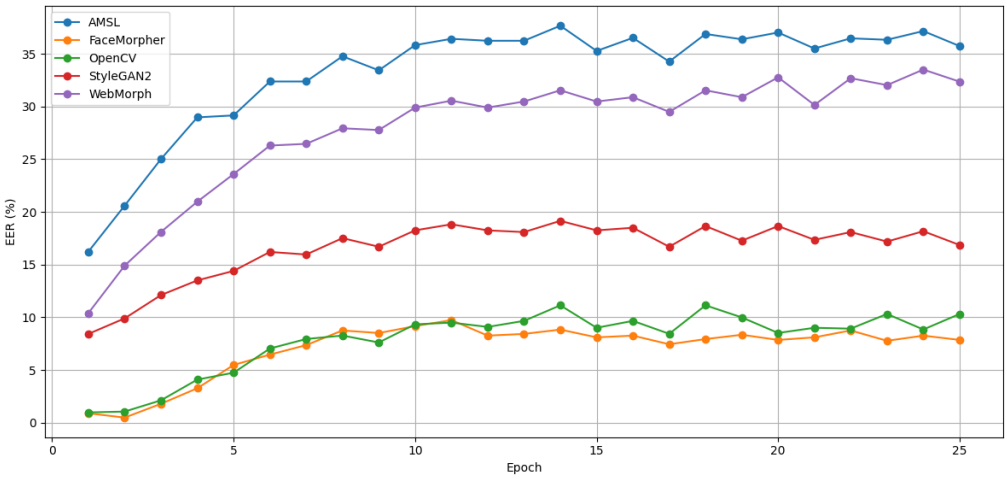


Figure 4.6: Evolution of D-EER across training epochs of the model trained on the SMDD dataset [59] and evaluated on the FRLL-Morphs dataset [52, 56, 39].

[52, 56, 39]. Since training relied exclusively on SMDD [59], the model was never exposed to the visual characteristics and artifacts that are typical of the FRLL-Morphs [52, 56, 39] morphing techniques. This lack of diversity in the training data explains the poor cross-technique generalization, with final D-EER values ranging from 7.86% (FaceMorpher) to 35.72% (AMSL), as detailed in Table 4.2.

As shown in Table 4.2, the replicated results exhibit slightly higher D-EER values compared to those reported in the original study. This performance gap can be attributed to differences in the training data: while the original study trained the model on a combination of CASIA-WebFace [133, 134] and SMDD [59], this experiment used only the SMDD dataset [59]. The reduced diversity in training data likely limited the model's ability to generalize across different morphing techniques, reflected in the increase of the average D-EER from 10.06% to 17.92%.

Table 4.2: Performance, presented in terms of D-EER (%). All experiments used the SPL paradigm.

Test Data \ **Train Data**		**Fang et al. (2022) [75]**	**Experiments**
		CASIA-Web + SMDD	**SMDD**
FRLL-Morphs	**OpenCV**	5.78	10.32
	FaceMorpher	4.67	7.86
	StyleGAN2	12.92	16.86
	WebMorph	15.72	32.35
	AMSL	12.09	35.72
SMDD	**Cropped Images**	9.19	9.68
	Non-Cropped Images	-	12.63
Average performance		10.06	17.92

Beyond the training data differences, two additional factors influence detection performance. Image preprocessing plays a crucial role, as demonstrated by the SMDD dataset [59] results: cropped images yield lower D-EER (9.68%) compared to non-cropped images (12.63%). This suggests that tighter facial crops, which remove background information and focus on the facial region, facilitate more effective morphing detection. This effect is illustrated in Figure 4.7.

(a) Non-cropped: 12.63. (b) Cropped: 9.68.

Figure 4.7: Example morphed image from the SMDD dataset [59], showing both the non-cropped and cropped versions of the same morph. The D-EER (%) shown for each version represents the overall detection performance across all images in that preprocessing condition (non-cropped vs. cropped).

Additionally, detection difficulty varies significantly across morphing techniques. As shown in Figure 4.8, D-EER ranges from 7.86% to 32.35% depending on the morphing technique, indicating that some techniques introduce more detectable artifacts than others. These variations are further influenced by image preprocessing: as noted with the SMDD dataset [59], tighter facial crops consistently improve detection performance. Together, these trends suggest that the model primarily learns artifacts specific to the SMDD dataset [59], which limits its robustness and generalization to other morphing methods.

4.1.6 Conclusions

From this experiment, it can be concluded that the datasets, morphing techniques employed, and preprocessing applied to the images have an impact on the model's performance. The results show that models trained exclusively on the SMDD dataset [59] achieve slightly lower performance compared to those trained on a combination of CASIA-WebFace [133, 134] and SMDD [59], as done by Fang et al. (2022) [75]. This outcome highlights the importance of using diverse and complementary datasets for training to improve the model's generalization capability.

Moreover, the analysis of the impact of image cropping demonstrates that image preprocessing, specifically, focusing on the facial region through tighter cropping, can lead to improved detection performance. This effect was consistently observed across both

(a) FaceMorpher: 7.86. (b) OpenCV: 10.32. (c) StyleGAN2: 16.86. (d) WebMorph: 32.35 (e) AMSL [52, 53, 54]: 35.72.

Figure 4.8: Example morphed images from the FRLL-Morphs dataset [52, 56, 39], showing one representative sample per morphing technique. The D-EER (%) shown for each technique represents the overall detection performance across all images generated with that morphing method. Techniques are ordered by increasing D-EER, from best (lowest error) to worst (highest error) detection performance.

the SMDD [59] and FRLL-Morphs [52, 56, 39] datasets, suggesting that removing irrelevant background information helps the model concentrate on the most discriminative facial features.

Overall, these findings reinforce the sensitivity of MAD systems to dataset composition, morphing techniques, and image preprocessing. Ensuring consistent and well-targeted preprocessing steps, along with the inclusion of varied data sources for training, appears essential for achieving robust and generalizable performance in real-world morph detection scenarios.

4.2 Experiment 5 — Zhang et al. (2024)

This experiment evaluates the MAD method proposed by Zhang et al. (2024) [24], which employs a ViT backbone combined with a SVM classifier. Although the original code was not publicly released, it was obtained directly from the authors upon request.

Figure 4.9 provides an overview of the workflow followed in this experiment, highlighting the three main stages of dataset preparation, model training and evaluation, and interpretability analysis via occlusion mapping.

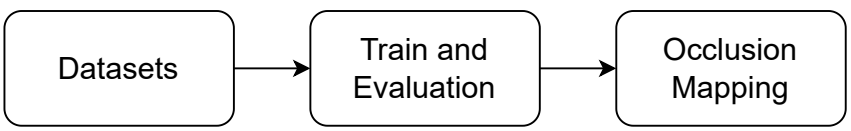


Figure 4.9: Experiment 5: methodology overview.

4.2.1 Datasets

Two publicly available datasets were used: **FRLL-Morphs** [52, 56, 39] and **FFHQ-Morphs** [51]. Beyond their original forms, this experiment also employed **cropped versions** of both datasets, referred to as *FRLL-Cropped* and *FFHQ-Cropped*. These cropped datasets were created as part of this work based on findings from Experiment 1, which showed that preprocessing via face cropping led to improved classification performance. The cropping process aligned and centered the facial region while removing background content and other distracting elements. Figure 4.10 shows the original and cropped datasets used in this experiment.

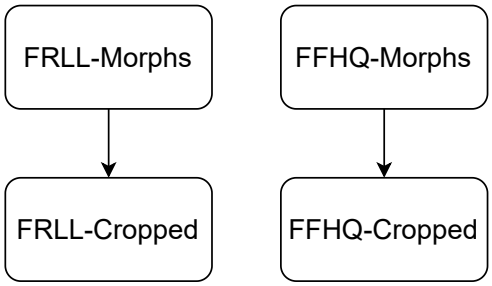


Figure 4.10: Experiment 5: datasets.

The model was trained and evaluated using four dataset configurations: original and cropped versions of both FRLL-Morphs and FFHQ-Morphs. Cross-dataset generalization was assessed by testing each trained model on all other datasets, resulting in a total of 16 train-test combinations. Figure 4.11 provides an overview of the training and evaluation workflow for this experiment.

4.2.2 Class Balancing

To address class imbalance between *bona fide* and morphed samples, an **undersampling** strategy was applied. This technique reduces the number of samples in the majority class by randomly removing them to match the size of the minority class. Experiments were conducted with both unbalanced and balanced dataset versions to examine the impact of class distribution.

4.2.3 Occlusion Mapping

To interpret the model's behavior and assess which facial regions contribute most to classification, two occlusion strategies were used:

- **Landmark-based occlusion**: Key facial features (left eye, right eye, nose, and mouth) were identified using facial landmarks and occluded individually.

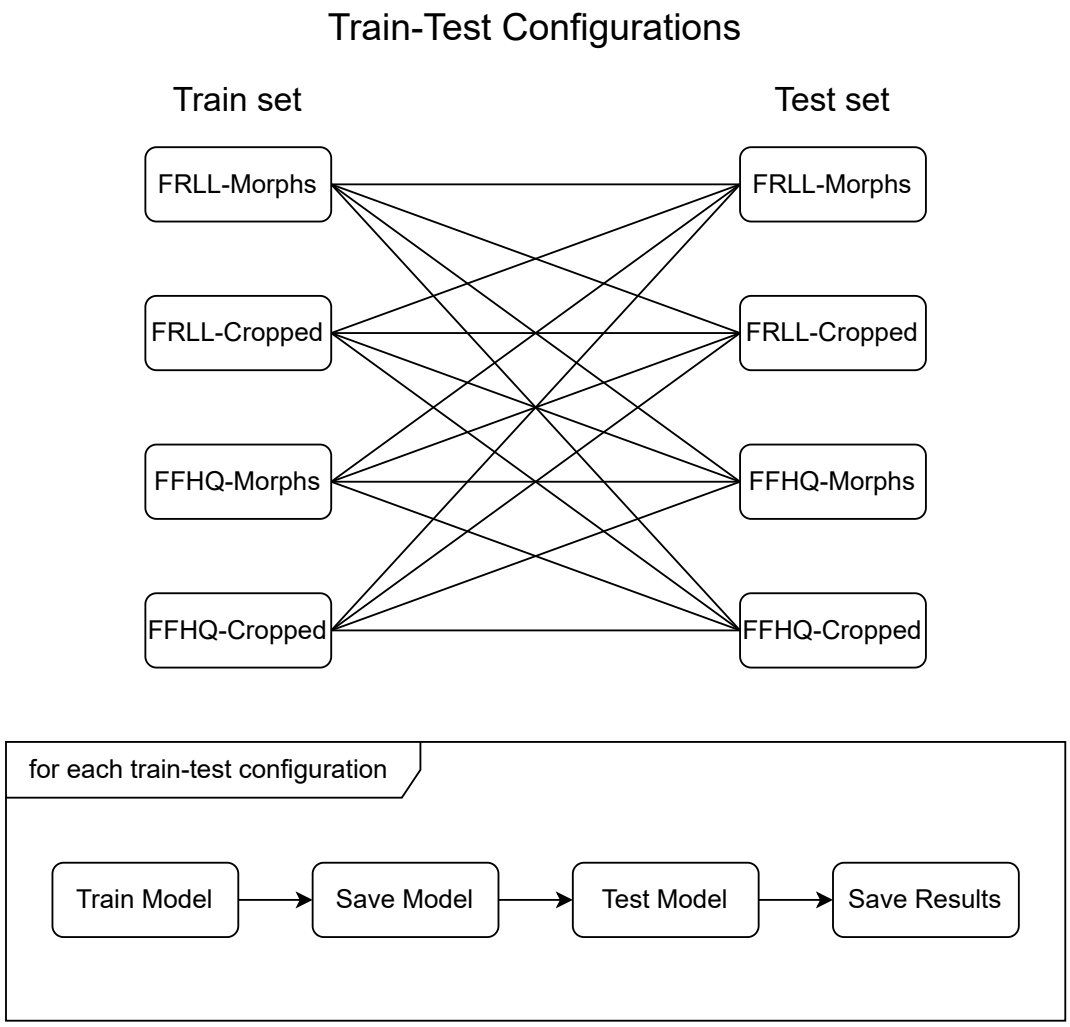


Figure 4.11: Experiment 5: train and evaluation flowchart.

- **Grid-based occlusion**: Each image was divided into a 4×4 grid, with one cell occluded at a time to evaluate regional importance.

Occlusion analysis was performed exclusively on the cropped datasets. This decision was based on the more consistent alignment and reduced background variation provided by cropped images, which made it easier to isolate the impact of occluding specific facial regions.

4.2.4 Evaluation Metrics

Performance was assessed using standard MAD metrics (see Subsection 2.1.3): **APCER**; **BPCER**; **D-EER**; **ACC**. In addition to these quantitative metrics, an occlusion-based evaluation was conducted to enhance model interpretability by identifying which facial regions contributed most to the classification decisions.

4.2.5 Experimental Setup

This experiment was conducted using *Google Colab* as the computational environment. Model training was performed on an *NVIDIA A100 GPU*, which provided the necessary computational power for efficient learning. In contrast, the *testing* and *heatmap generation* stages were executed on the CPU, as these processes required fewer computational resources and did not benefit greatly from GPU acceleration.

4.2.6 Experimental Results

The evaluation was structured to assess both the detection performance and the generalization ability of the model across varying dataset conditions. Cross-dataset evaluation was performed by training and testing the model on different combinations of datasets. This process was repeated using unbalanced and balanced training data to examine the impact of class distribution.

To further analyze model behavior, occlusion mapping tests were conducted using the cropped datasets. These tests aimed to identify the most relevant facial regions by measuring performance changes when specific areas were blocked. The code used to obtain these results is explained in Annex H.

Evaluation Results

The results obtained during the evaluation of the model under different dataset conditions are presented in Tables 4.3 and 4.4. Performance was assessed using both intra- and cross-dataset scenarios, with and without class balancing, and considering different preprocessing states (cropped and not cropped). Detailed numerical results are provided in Annex I.

The intra-dataset results presented in Table 4.3 show that classification performance is highly dependent on the consistency of preprocessing between training and testing. When both the training and testing images share the same preprocessing type, either both cropped or both not cropped, the model achieves strong and stable results. However, performance drops when the train and test sets differ in cropping, highlighting the model's sensitivity to changes in facial alignment and background content.

This effect is more noticeable for the FRLL dataset, which benefits more from cropping due to its greater variability in image composition. On the other hand, FFHQ demonstrates stronger performance even in the uncropped setting, though it still sees clear gains when cropping is consistently applied.

As for class balancing, its impact in intra-dataset scenarios appears to be modest. Some improvements are observed, particularly for FRLL, but the overall trend suggests that ensuring consistent image preprocessing between training and testing is more critical

Table 4.3: Intra-dataset testing results using Zhang et al. (2024) [24] code.

Training Set Balance	Training Set Cropping	Dataset used	Test Set Cropping	APCER@BPCER			D-EER (%)	AUC (%)
				1%	5%	10%		
Unbalanced	Not cropped	FFHQ	Not cropped	6.22	1.48	0.74	2.85	99.63
			Cropped	4.74	0.82	0.22	2.34	99.78
		FRLL	Not cropped	4.89	3.37	2.75	4.30	97.16
			Cropped	85.35	81.10	75.30	39.01	66.85
	Cropped	FFHQ	Not cropped	45.39	23.23	14.13	12.11	94.71
			Cropped	**0.00**	**0.00**	**0.00**	0.12	**100.00**
		FRLL	Not cropped	73.30	60.15	56.37	26.84	80.78
			Cropped	0.21	0.21	0.14	0.11	99.96
Balanced	Not cropped	FFHQ	Not cropped	6.88	1.10	0.42	2.14	99.71
			Cropped	5.42	1.02	0.26	2.34	99.75
		FRLL	Not cropped	2.75	2.75	2.68	2.60	98.36
			Cropped	79.48	34.11	29.79	17.06	90.34
	Cropped	FFHQ	Not cropped	41.66	25.06	15.41	12.61	94.69
			Cropped	**0.00**	**0.00**	**0.00**	**0.04**	**100.00**
		FRLL	Not cropped	77.98	54.30	48.52	31.72	78.48
			Cropped	0.21	0.14	0.14	0.11	99.93

than balancing the dataset in these settings.

The cross-dataset results in Table 4.4 highlight the challenges of generalization when the training and testing datasets differ in domain. In general, the model's performance is notably lower compared to intra-dataset testing, especially in unbalanced scenarios. This drop is more pronounced when the train and test sets differ greatly in structure or quality, such as between FRLL and FFHQ.

Table 4.4: Cross-dataset testing results using Zhang et al. (2024) [24] code.

Training Set Balance	Training Set Cropping	Train dataset	Test Set Cropping	Test dataset	APCER@BPCER			D-EER (%)	AUC (%)
					1%	5%	10%		
Unbalanced	Not cropped	FFHQ	Not cropped	FRLL	97.59	83.76	77.70	39.02	65.72
			Cropped	FRLL	14.58	8.49	3.89	7.30	98.24
		FRLL	Not cropped	FFHQ	98.96	94.02	87.99	46.53	55.0
			Cropped	FFHQ	98.50	93.86	87.99	45.03	57.02
	Cropped	FFHQ	Not cropped	FRLL	71.44	54.92	43.91	31.13	79.11
			Cropped	FRLL	92.85	78.06	69.85	40.94	65.60
		FRLL	Not cropped	FFHQ	92.86	77.13	63.98	31.52	75.21
			Cropped	FFHQ	91.41	76.65	63.80	30.82	75.64
Balanced	Not cropped	FFHQ	Not cropped	FRLL	95.39	76.53	74.05	36.15	71.46
			Cropped	FRLL	**8.28**	**6.86**	**2.83**	**7.09**	**98.69**
		FRLL	Not cropped	FFHQ	95.92	86.31	76.87	40.57	63.30
			Cropped	FFHQ	93.76	80.13	68.04	33.54	72.64
	Cropped	FFHQ	Not cropped	FRLL	70.89	43.98	33.38	22.02	86.53
			Cropped	FRLL	77.14	55.34	53.43	31.18	78.24
		FRLL	Not cropped	FFHQ	88.57	73.18	59.22	29.68	77.66
			Cropped	FFHQ	88.59	71.72	58.27	29.24	77.83

As seen in the intra-dataset results, the model is sensitive to differences in image alignment, and this sensitivity becomes even more evident in the cross-dataset setting. Using cropped images for both training and testing helps reduce the negative effects of domain shift, offering more consistent input conditions. However, this consistency alone is not enough to ensure high performance. Even when cropping is applied on both sides, results remain noticeably below those of intra-dataset testing, underscoring the inherent difficulty of generalizing across datasets with different distributions. These findings suggest that while consistent preprocessing can support generalization, over-

coming substantial dataset variability requires broader strategies, such as training on more diverse or augmented data.

Class balancing shows some positive effect, particularly when combined with cropped data, helping to stabilize performance and reduce error rates. Still, the overall variability across conditions suggests that dataset alignment and domain similarity are far more critical than class balance in cross-dataset scenarios. These results underline the importance of training on diverse and representative data to achieve robust generalization in real-world settings.

The overall results demonstrate that preprocessing consistency, dataset alignment, and class balance are key factors influencing the model's ability to detect morphing attacks effectively. Ensuring similarity between training and testing conditions leads to better performance, while mismatches, whether in data distribution or preprocessing, introduce challenges.

In intra-dataset testing, consistent cropping boosts performance, while class balancing has only a minor impact. In cross-dataset testing, performance drops due to domain shift, though cropping and balancing offer some mitigation without fully resolving the gap.

Overall, robustness relies on standardized preprocessing as well as diverse and balanced training data, though further improvements will likely require domain adaptation or augmentation techniques.

Heatmap Visualization

To enhance the interpretability of the model, occlusion-based heatmaps were generated to visualize the regions of the face that contributed most to the classification outcome[1]. The workflow for this analysis is illustrated in Figure 4.12. The visualization process focused on samples from the *FFHQ-Cropped* and *FRLL-Cropped* datasets and followed a four-stage approach:

1. The original input image was first displayed without any occlusion, serving as a reference for comparison.
2. A heatmap was overlaid on the original input image to visualize the results of the **landmark-based occlusion analysis**. In this approach, predefined facial landmarks corresponding to key regions (such as the left eye, right eye, nose, and mouth) were selectively occluded. The classification performance degradation

[1]The interpretability analysis conducted in Experiment 5 resulted in two conference papers that were accepted and presented. At the date of submission of this document, only the paper presented at ISPA 2025 has been published, whereas the paper presented at RecPad 2025 has not yet been published. The papers are "Composite Heatmaps: Interpreting Deep Models in Morph Attack Detection" by Pedro Morais and Inês Domingues, presented at the ISPA 2025 (see Annex B), and "Interpreting Morph Attack Detection with Composite Heatmaps" by Pedro Morais and Inês Domingues, presented at the RecPad 2025 (see Annex C).

caused by each occlusion was measured to quantify the importance of these specific semantic facial areas.

3. A heatmap was overlaid on the original input image to represent the results of the **grid-based occlusion analysis**. The image was divided into a 4×4 grid, and classification performance degradation was measured as each grid cell was occluded individually. This method provides a localized, spatial-level view of sensitivity.
4. A **composite heatmap** was then produced by combining both the **landmark-based** and **grid-based occlusion** scores into a novel unified visualization. This composite approach, developed specifically for this work, highlights both semantic and spatial importance, revealing how different facial areas contribute to model predictions.

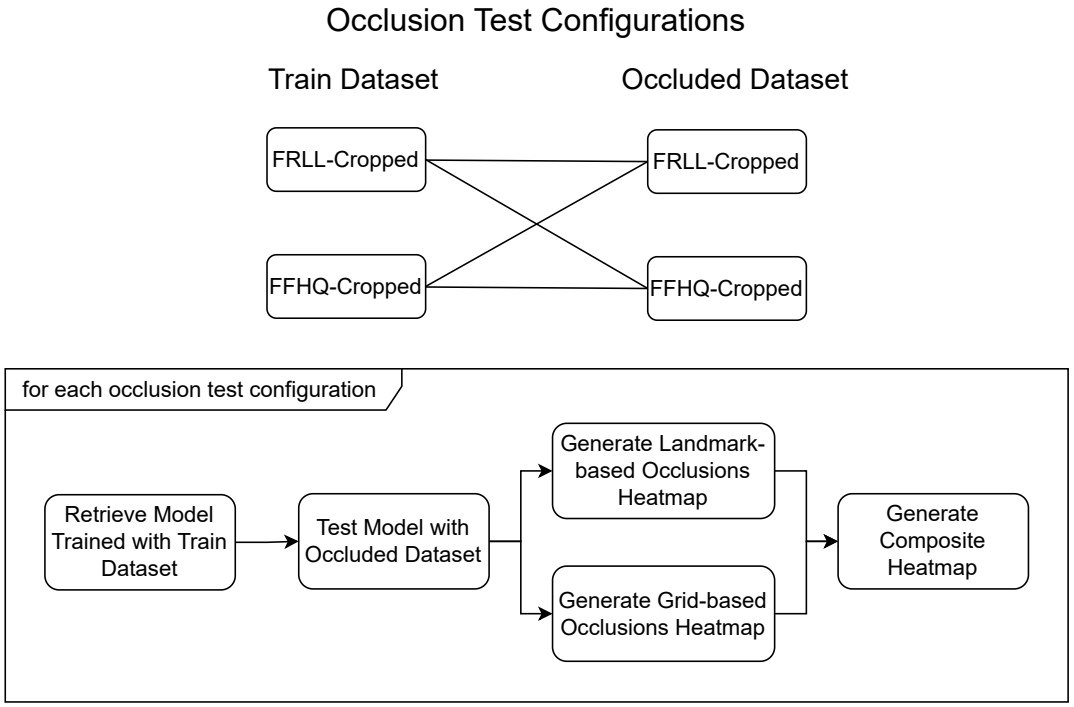
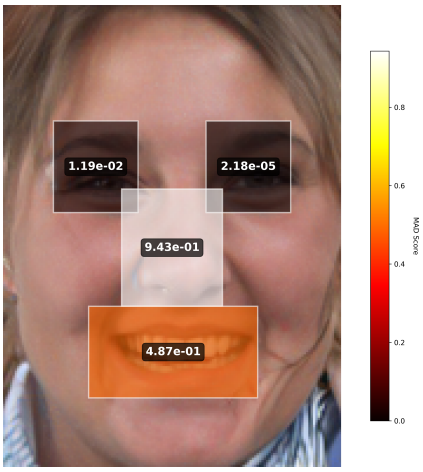


Figure 4.12: Experiment 5: interpretability flowchart.

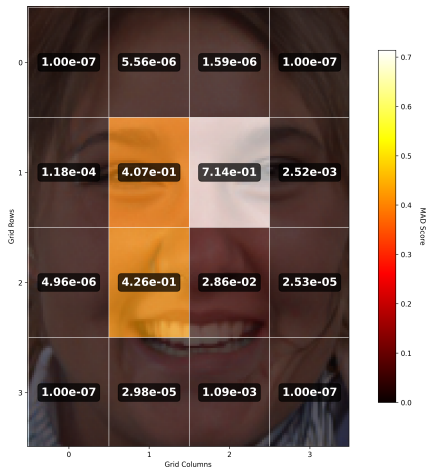
Figure 4.13 illustrates this process using the image `32950_05186` from the *FFHQ-Morphs* dataset, which was generated with a diffusion-based morphing algorithm and pre-processed through face cropping. The visualizations demonstrate the semantic and spatial dependencies and sensitivities of the evaluated model.

Several conclusions can be drawn from the analysis of the resulting heatmaps. Firstly, the model trained on *FFHQ-Cropped* dataset places more emphasis on the **nose**, **mouth**, and to a lesser extent, the **eye** regions. A consistent pattern observed across multiple samples is that the model assigns higher relevance to areas with **depth transitions**, such as the boundary between the nose and cheeks and the corners of the eyes. This

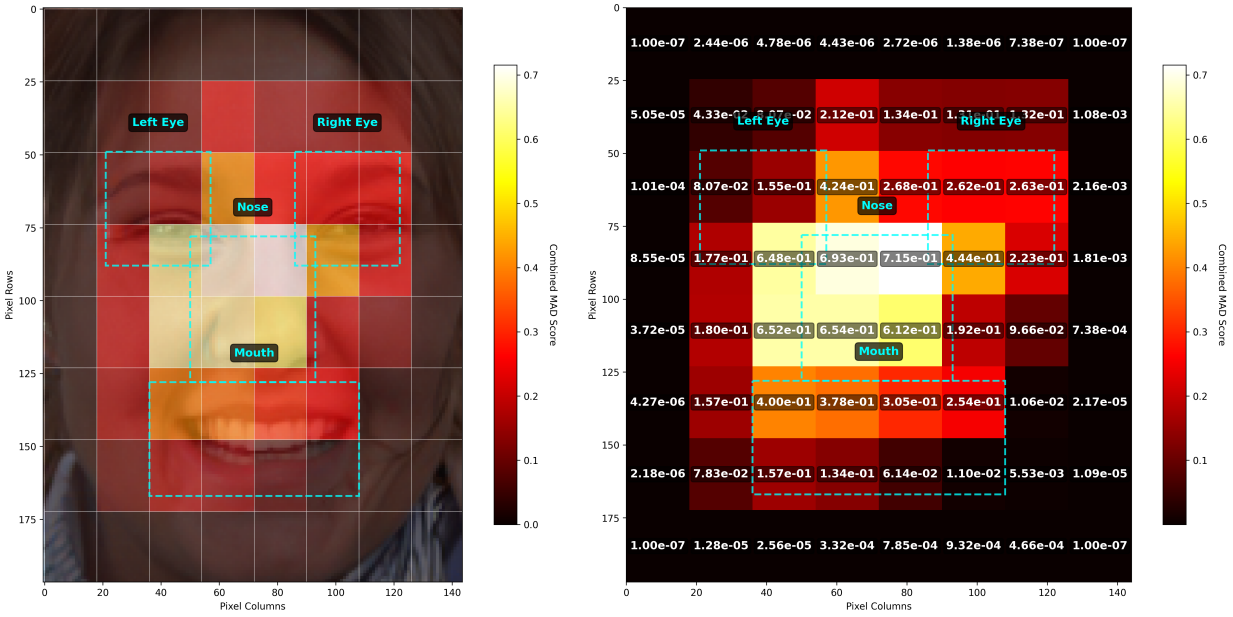
(a) Original cropped image (`32950_05186`).



(b) Landmark occlusion heatmap.



(c) Grid occlusion heatmap (4×4).



(d) Composite heatmap (8×8): left, without MAD scores; right, with MAD scores superimposed.

Figure 4.13: Visualization of occlusion heatmaps for image `32950_05186` from the *FFHQ-Morphs* dataset, generated with a diffusion-based morphing algorithm and preprocessed with face cropping. Lighter regions indicate higher importance in the model's decision.

aligns with the facial regions most affected by morphing artifacts, where blending errors often occur due to geometric inconsistencies. Landmark-based occlusion further confirms this sensitivity, showing that occluding the nose and mouth leads to the largest drops in model confidence.

Moreover, while most visualizations showed consistent and interpretable importance maps, some samples presented irregular patterns that diverged from the general behavior. These *outliers* generally result from the cross-dataset evaluation setup, where

the model was trained and tested on datasets with different visual properties. Variations in morphing complexity, resolution, and lighting can shift the learned feature distributions, leading to less coherent heatmaps and reduced generalization across domains.

This is evident on the model trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*, where no consistent trend is observed in the heatmaps. This irregular behaviour suggests that the model fails to generalize across datasets, as the visual and statistical properties of the training and testing domains differ substantially. The *FRLL-Morphs* dataset employs relatively simple morphing techniques, producing less diversed and more easily distinguishable facial blends, whereas the *FFHQ-Morphs* dataset contains morphs generated with more complex morphing techniques that create subtle and spatially dispersed artifacts. These differences in morphing complexity, combined with variations in resolution and illumination, result in a mismatch between the learned feature representations and the characteristics of the test data. Consequently, the resulting heatmaps exhibit scattered and less interpretable importance regions, reflecting uncertainty in the model's decision process. Figure 4.14 presents an example of this behaviour.

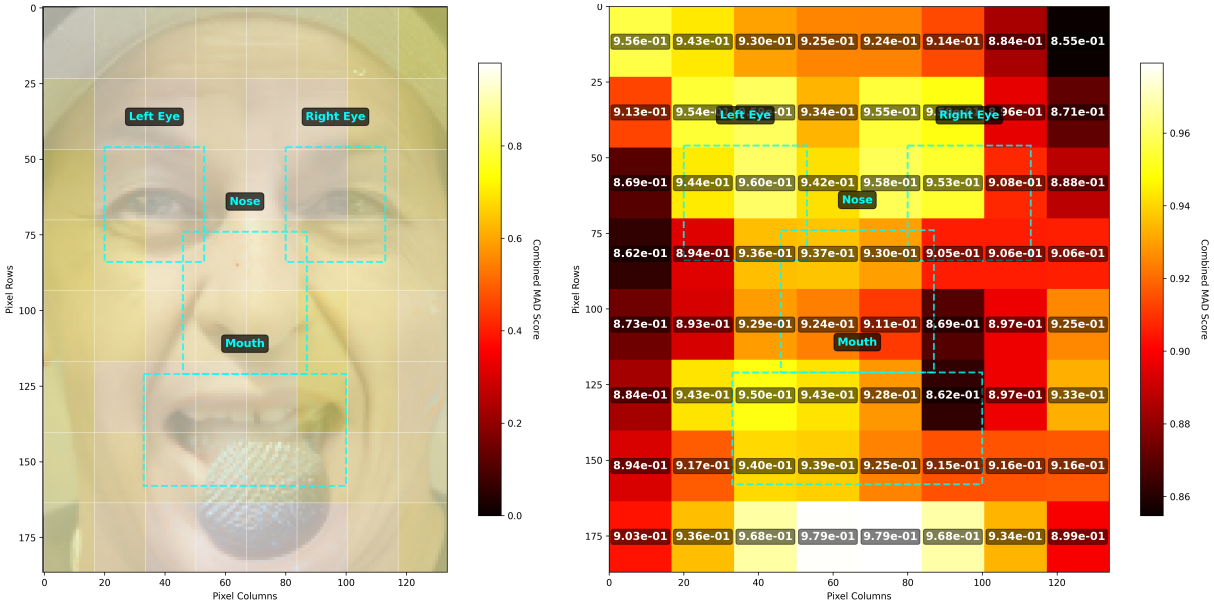


Figure 4.14: Example composite heatmap (8×8), produced by the model trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*: left, without MAD scores; right, with MAD scores superimposed. The visualization shows inconsistent relevance regions, indicating limited generalization across datasets.

Despite these inconsistencies, the cross-domain evaluation provides valuable insights into the model's sensitivity to dataset-specific biases and its capacity to handle unseen morphing techniques. Overall, the analysis demonstrates that landmark-based and grid-based occlusion methods can highlight the regions that drive classification decisions, even in challenging cross-domain conditions, thereby improving the interpretability of the results.

4.2.7 Conclusions

The evaluation of the MAD method proposed by Zhang et al. (2024) [24] confirmed its strong and consistent performance across most intra- and cross-dataset configurations, particularly when preprocessing was applied. The combination of a ViT backbone with an SVM classifier proved highly effective for MAD, achieving near-perfect ACC in several scenarios, especially in intra-dataset configurations.

While generalization across datasets was mostly stable, a specific train-test combination, when the model was trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*, revealed lower detection performance and reduced coherence in the resulting heatmaps. This behavior suggests that interpretability and performance can vary when there are visual or methodological differences between datasets, such as morphing complexity, resolution, or illumination conditions, rather than indicating a fundamental limitation of the model itself.

Occlusion-based analysis further demonstrated the model's ability to focus on semantically meaningful facial regions, particularly the nose, mouth, and eyes, areas commonly associated with morphing artifacts. However, the variation observed in certain cross-dataset visualizations highlights that classification outcomes are influenced by both the characteristics of the datasets and the specific model configuration used, which might render the interpretability framework unusable.

In summary, the Zhang et al. (2024) [24] model exhibits excellent detection performance and interpretability under consistent preprocessing conditions, with only isolated cases of reduced generalization. These findings underscore the importance of ensuring dataset diversity and preprocessing consistency when evaluating MAD methods, as both factors critically influence the overall robustness and reliability of the model's behavior across experimental setups.

5 Occlusion-based Interpretability

This chapter presents the standardized occlusion-based interpretability method developed to analyze and interpret the behavior of MAD models. While the methodology was introduced in Section 3.5 and applied in Experiment 5 (see Section 4.2), this chapter formalizes the approach as a reusable framework and discusses its implementation, capabilities, and insights gained from its application.

The method addresses the critical need for transparency in deep learning-based MAD systems by systematically identifying which facial regions contribute most to model predictions. This is achieved through a three-stage pipeline that generates cropped datasets, applies occlusion strategies, and produces interpretable heatmap visualizations.

5.1 Method Overview

The occlusion-based interpretability method consists of three main stages, each implemented as a standalone Python script with corresponding demonstration notebooks (detailed in Annex J):

1. **Dataset Cropping** (`crop_dataset_builder.py`): Standardizes facial images by detecting, cropping, and resizing face regions to ensure spatial consistency across all samples.

2. **Occlusion Generation** (`occlusion_dataset_builder.py`): Creates systematically occluded variants of the cropped images using two complementary strategies:

   - Grid-based occlusion: Divides each image into a spatial grid and occludes one cell at a time.

   - Landmark-based occlusion: Occludes semantically meaningful facial regions (eyes, nose, mouth) based on detected landmarks.

3. **Heatmap Visualization** (`occlusion_heatmaps_builder.py`): Generates visual heatmaps that quantify the impact of regional occlusions on model predictions, combining both grid-based and landmark-based results.

5.2 Technical Implementation and Features

The standardized framework provides flexible configuration options in order to allow researchers to adapt the occlusion-based analysis to different experimental requirements and model architectures. Three key technical features, described in the following subsections, enable fine-grained control over the interpretability analysis.

5.2.1 Configurable Landmark Precision

The landmark-based occlusion strategy offers two configurable aspects that control how facial landmarks are occluded:

Precision modes define the shape and coverage of the occlusion, as illustrated in Figure 5.1:

- **Pixel precision**: Occludes only the specific pixels that belong to the landmark region, creating a pixel-perfect mask that follows the contours of the facial landmark (e.g., the exact eye region). This enables fine-grained analysis of facial landmarks but may be more sensitive to landmark localization errors.
- **bbox precision**: Occludes a rectangular region around each landmark. This approach provides broader coverage and is more robust to landmark detection variations, though it may occlude areas beyond the actual landmark.

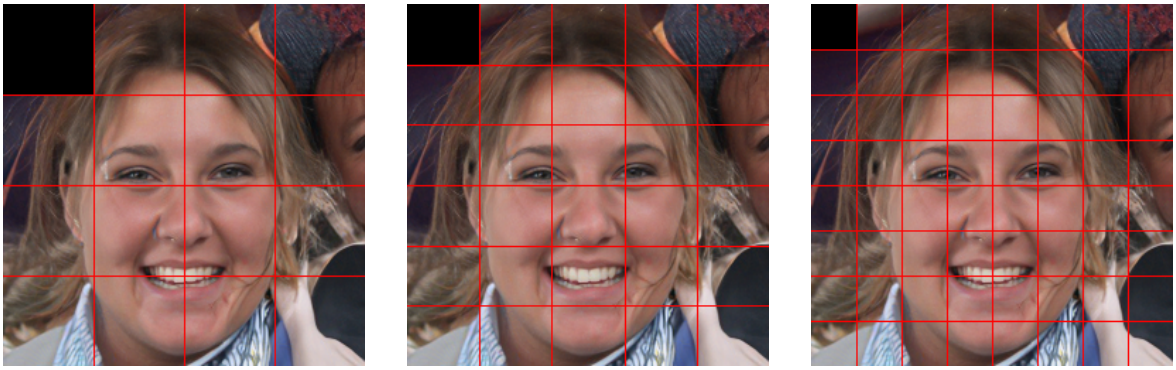
(a) Pixel precision mode. (b) Bbox precision mode.

Figure 5.1: Landmark precision modes example, where the four selected facial regions (left eye, right eye, nose, and mouth) are occluded.

Scaling factor controls the size of the occluded region, expanding or contracting the pixel mask or rectangular bbox depending on the selected precision mode. This allows researchers to adjust the extent of occlusion to balance between fine-grained analysis and robustness to landmark detection errors.

5.2.2 Configurable Grid Resolution

The grid-based occlusion strategy allows researchers to configure the **grid size** ($R \times C$), which defines the spatial resolution by specifying the number of rows and columns in the grid (e.g., 4×4, 6×5, 8×8; see Figure 5.2). This parameter controls the granularity of the spatial analysis, enabling adaptation to different experimental requirements and computational constraints.



(a) 4×4 grid resolution. (b) 6×5 grid resolution. (c) 8×8 grid resolution.

Figure 5.2: Grid resolutions examples, showing how the image is divided into 4×4, 6×5, and 8×8 grids. The top-left cell is occluded in each case, and red lines indicate the grid structure.

5.2.3 Flexible Heatmap Modes

The heatmap generation supports two complementary visualization modes that provide different perspectives on model behavior:

- **Difference mode**: Quantifies the performance degradation caused by each occlusion by computing the difference between the model's prediction on the original image and its prediction on the occluded variant. Higher values indicate regions that are more critical to the model's decision making process.
- **Occluded mode**: Displays the prediction scores obtained for each occluded variant. This mode shows the model's actual output for each occlusion pattern, providing insight into how the model responds to missing information in different facial regions without reference to the original prediction.

These modes serve complementary purposes: **difference mode** emphasizes relative importance by measuring performance degradation, while **occluded mode** reveals absolute model responses under partial information.

5.3 Application to MAD Model Interpretability

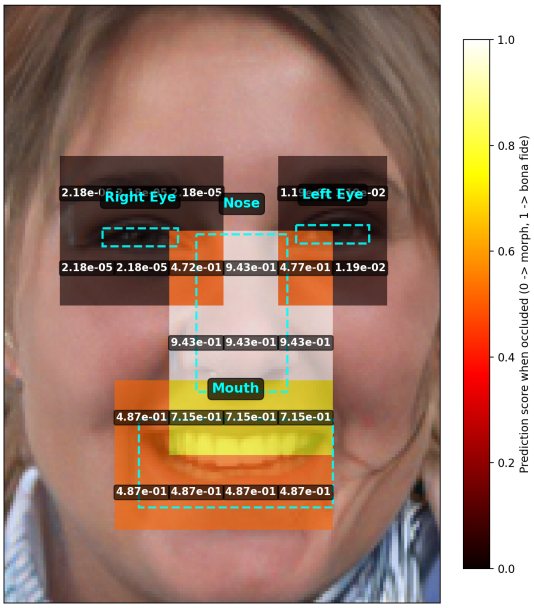
The method was applied to interpret the ViT-SVM model from Experiment 5 (Section 4.2), trained on the *FRLL-Cropped* and *FFHQ-Cropped* datasets. For this analysis, the landmark-based occlusion strategy was employed using bbox precision mode with a scale factor of 1.5, which expands the occluded region around each facial landmark to 150% of its original size. This configuration provides robust coverage of facial landmarks while maintaining sufficient sensitivity to detect regional importance. The heatmap visualizations were generated using occluded mode, which displays the prediction scores obtained for each occluded variant, showing how the model responds when specific facial regions are missing.

5.3.1 Heatmap Examples

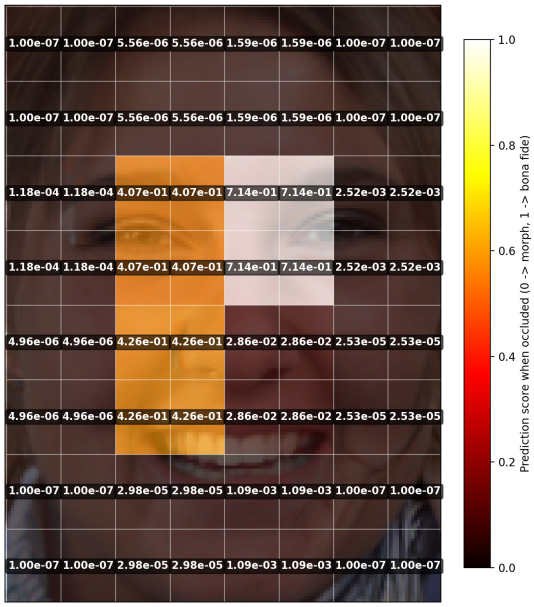
Figure 5.3 illustrates the three complementary types of heatmaps generated by the standardized framework, each providing distinct insights into model behavior. These visualizations correspond to the same analysis presented in Experiment 5 (in Section 4.2 see Figure 4.13), but are generated using the formalized implementation described in this chapter:

- **Landmark-based heatmap**: Reveals the model's reliance on semantically meaningful facial characteristics by systematically occluding individual landmarks (left eye, right eye, nose, mouth). This approach directly maps model predictions to anatomic regions, providing interpretable insights aligned with human understanding of facial structure.
- **Grid-based heatmap**: Provides spatially uniform sensitivity analysis by dividing the image into a 4×4 grid and occluding each cell independently. The results are upscaled to 8×8 resolution for visualization. This strategy captures spatial dependencies without semantic bias, revealing whether the model focuses on specific image regions regardless of their anatomical meaning.
- **Composite heatmap**: Synthesizes both landmark-based and grid-based analyses into a unified 8×8 resolution visualization. The grid-based results are upsampled from 4×4 to 8×8, then averaged with the landmark-based scores to produce a view that combines semantic interpretability with spatial coverage.

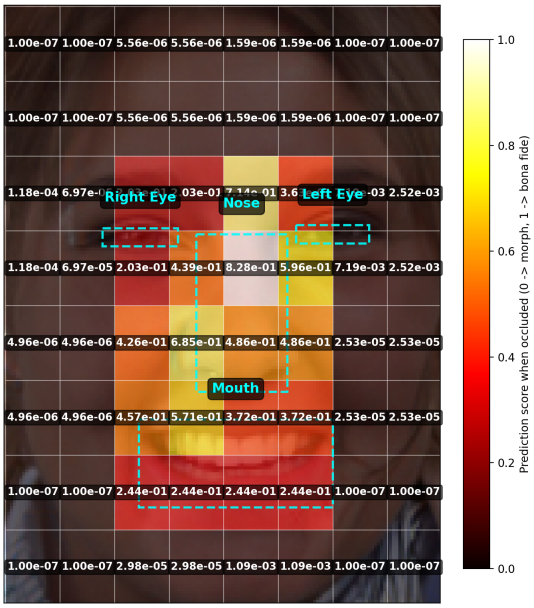
The here proposed composite heatmap provides the most comprehensive interpretability insights by leveraging the strengths of both approaches: the semantic meaningfulness of landmark-based analysis and the spatial completeness of grid-based analysis. This dual perspective enables a more nuanced understanding of which facial regions, both anatomical and spatial, contribute most to the model's decision making process.



(a) Landmark occlusion heatmap.



(b) Grid occlusion heatmap.



(c) Composite heatmap.

Figure 5.3: Occlusion heatmaps generated by the standardized interpretability framework for image `32950_05186` from the *FFHQ-Morphs* dataset (diffusion-based morph). The visualizations show (a) landmark-based, (b) grid-based, and (c) composite analyses using bbox precision (scale factor 1.5) in occluded mode. Lighter colors indicate regions where occlusion causes greater changes in model predictions, revealing areas critical to the model's decision making.

5.3.2 Heatmap Improvements in the Standardized Framework

Comparing Figure 5.3 with the initial results from Experiment 5 (see Figure 4.13) reveals several important refinements introduced by the standardized framework:

- **Refined landmark mapping**: The landmark-based heatmap has been enhanced to adopt anatomically correct mapping of facial landmarks. Building upon Experiment 5 (Section 4.2), the framework now implements a standardized anatomical reference system where landmarks are consistently mapped to the subject's perspective rather than the image perspective. For instance, the right eye refers to the subject's right eye (appearing on the left side of a frontal image). Both standardized scripts (occlusion generation and heatmap visualization) now employ this anatomically correct mapping convention, ensuring precise and consistent interpretations across the entire pipeline.
- **Refined grid resolution**: The grid-based heatmap is now generated at a higher resolution, already upscaled to match the composite resolution.
- **Improved composite averaging**: The composite heatmap now employs a more sophisticated merging strategy that averages landmark-based and grid-based results only in cells where both analyses provide overlapping information. This selective averaging preserves the distinct contributions of each approach while creating a unified visualization that highlights regions identified as important by either or both methods.

These refinements enhance the accuracy and interpretability of the occlusion-based analysis, making the standardized framework more reliable for understanding model behavior in MAD applications.

5.4 Discussion

The application of the occlusion-based interpretability method to the ViT-SVM model from Experiment 5 (Section 4.2) revealed important insights into how the model makes morphing detection decisions. This section discusses the key findings from the interpretability analysis, how preprocessing and class balancing affect model behavior, examining which facial landmarks proved most influential for detection, and what additional facial regions could be explored in future work to deepen the understanding of morphing artifact detection. These insights not only validate the effectiveness of the proposed interpretability framework but also provide actionable guidance for improving MAD systems and understanding the vulnerabilities of morphing algorithms.

5.4.1 Impact of Preprocessing and Class Balancing

The interpretability analysis was conducted exclusively on cropped datasets due to the computational intensity of occlusion mapping, which requires generating and classifying multiple occluded variants for each image. Given the resources constraints, analyzing all preprocessing configurations would have been prohibitive.

The decision to focus on cropped datasets was strategically informed by findings from Experiment 1 (see Section 4.1), which demonstrated that cropping leads to improved detection performance by focusing on the facial region. Furthermore, Experiment 5 results (see Tables 4.3 and 4.4) reinforced this finding, showing that classification performance is dependent on the consistency of preprocessing between training and testing. When both training and testing images share the same preprocessing approach (either both cropped or both not cropped), the model achieves strong and stable results, with near-perfect accuracy in some cases (e.g., D-EER of 0.04% for FFHQ cropped-to-cropped in the balanced scenario). However, performance drops when the train and test sets differ in cropping, highlighting the model's sensitivity to changes in facial alignment and background content.

The impact of class balancing on model behavior presents a more nuanced picture. As shown in Table 4.3, the effect of class balancing in intra-dataset scenarios appears to be modest, with the overall trend suggesting that ensuring consistent image preprocessing between training and testing is more critical than balancing the dataset in these settings. For instance, comparing the FFHQ cropped-to-cropped scenario, both balanced and unbalanced configurations achieve near-perfect performance (D-EER of 0.12% and 0.04%, respectively).

However, class balancing demonstrates more substantial benefits in cross-dataset scenarios, particularly when dealing with domain shift. The cross-dataset results (see Table 4.4) reveal that balanced training data can improve generalization. For example, when training on FFHQ uncropped and testing on FRLL cropped, the balanced configuration achieves a D-EER of 7.09% compared to 7.30% for the unbalanced configuration. More notably, in challenging cross-dataset scenarios such as training on FRLL cropped and testing on FFHQ cropped, balancing reduces the D-EER from 30.82% (unbalanced) to 29.68% (balanced), demonstrating that class balancing helps mitigate overfitting to the majority class and improves the model's ability to generalize across different data distributions.

These findings suggest that while preprocessing consistency remains the primary factor influencing model performance, class balancing plays a complementary role, particularly in scenarios where the model must generalize to unseen datasets or handle domain shift. The interpretability analysis conducted on cropped datasets therefore reflects the model's behavior under optimal preprocessing conditions, where the focus on facial regions is maximized and other distractions are minimized.

5.4.2 Insights from Selected Landmarks

The choice of facial landmarks (left eye, right eye, nose, mouth) was motivated by their semantic importance in face recognition and morphing detection. As detailed in the conclusions of Experiment 5 (Section 4.2), the occlusion-based analysis revealed

that the **nose** and **mouth** regions are the most influential for the model's morphing detection decisions, while the **eye regions** demonstrated somewhat lower importance, though still contributing to the classification process.

The prominence of the nose and mouth can be attributed to the model's sensitivity to **depth transition areas**, such as the boundary between the nose and cheeks and the region around the mouth. These areas are particularly susceptible to morphing artifacts because blending errors often occur at geometric discontinuities where facial depth changes abruptly. The morphing process struggles to seamlessly merge facial structures with different geometric properties, resulting in subtle inconsistencies that the model has learned to detect.

The relatively lower importance of the eye regions may be explained by the fact that the corners of the eyes, despite being depth transition areas where the eye socket meets the surrounding facial structure, are more well-defined and geometrically constrained regions. This geometric regularity may allow morphing algorithms to blend these areas more successfully compared to the more complex three-dimensional structures of the nose and mouth. Consequently, morphing artifacts in the eye regions may be less pronounced or more difficult for the model to detect, leading to a reduced reliance on these landmarks for classification decisions.

The prominence of depth transition areas in the model's decision making process suggests that future morphing detection approaches could benefit from explicitly modeling facial geometry and depth information, rather than relying solely on texture-based characteristics. This observation also motivates the exploration of additional facial landmarks beyond the four analyzed in this study. While the current analysis focused on the eyes, nose, and mouth due to their established importance in face recognition, several other facial regions with similar geometric properties, particularly those involving depth transitions, warrant investigation to provide a more comprehensive understanding of model behavior and morphing artifact distribution:

- **Eyebrows**: The eyebrow region presents challenges for alignment and positioning due to depth transitions and geometric complexity. While morphing algorithms may struggle to properly align eyebrow shapes and positions between different subjects, cosmetic alterations (microblading, shaping, makeup) and high natural variability mean that eyebrow irregularities are often perceived as plausible, making artifacts easier to dismiss as grooming choices rather than manipulation.

- **Chin and jawline**: The lower face contour, including the chin and jawline, involves substantial geometric variation across individuals. These regions may exhibit morphing artifacts related to facial shape blending, particularly when morphing faces with different jaw structures or chin prominence.

- **Forehead**: While often considered a relatively smooth facial region, the forehead includes subtle characteristics such as wrinkles, hairline transitions, and the brow ridge. The analysis of this region could reveal whether the model relies on these subtle textural characteristics or primarily focuses on more prominent facial structures.
- **Face contour and cheeks**: The outer facial boundary and cheek regions represent large areas where morphing blending occurs. Analyzing these regions could help determine whether the model detects artifacts in peripheral facial areas or concentrates primarily on central landmarks.
- **Ears**: Although often partially visible or occluded by hair, ears are highly distinctive anatomical structures that vary widely across individuals. Their geometric complexity might make them difficult to morph convincingly, potentially creating detectable inconsistencies.

Extending the occlusion-based analysis to include these additional landmarks would enable a more complete characterization of which facial regions contribute to morphing detection. This analysis could inform the development of more targeted detection strategies and provide insights into the vulnerabilities of current morphing algorithms. Furthermore, it would help validate whether the observed importance of depth transition areas generalizes to other facial regions with similar geometric properties, strengthening the theoretical understanding of morphing artifact formation and detection.

5.4.3 Methodological Considerations

While the occlusion-based interpretability method offers valuable insights into model behavior, several methodological considerations and limitations must be acknowledged to ensure appropriate application and interpretation of results.

Grid Resolution Trade-offs

The choice of grid resolution for grid-based occlusion analysis involves a critical trade-off between spatial granularity and interpretability. If the grid resolution is too coarse (e.g., 2×2), each cell covers too large an area, providing insufficient spatial detail to identify specific regions of importance. The occluded regions become so large that they may encompass multiple distinct facial characteristics simultaneously, making it difficult to isolate which specific areas contribute to model predictions. On the other hand, if the grid resolution is too fine (e.g., 16×16 or higher), each cell occludes such a small area that it may not have sufficient impact on the model's prediction to produce meaningful interpretability information. The model may be robust to such small occlusions, resulting in minimal performance changes that fail to reveal which regions are truly

important.

Computational Cost of Occlusion Mapping

The occlusion-based interpretability method is inherently computationally intensive, requiring the generation and classification of multiple occluded variants for each image in the dataset. For landmark-based occlusion with k landmarks, each image requires k additional variants. For grid-based occlusion with an $n \times m$ grid, each image requires $n \times m$ variants. When combining both strategies, the total number of variants per image becomes $k + n \times m$. The time and space complexity therefore grows linearly with the number of cells in the grid, for a grid of size $n \times m$, the method must generate, store, and classify $n \times m$ occluded variants of each image. This means that doubling the grid resolution (e.g., from 4×4 to 8×8) quadruples the number of images that must be generated, stored, and processed through the model. For a dataset with thousands of images, this results in a substantial increase in the total number of images that must be processed, which can quickly become prohibitive in terms of both disk storage requirements and computational time.

This computational burden has several practical implications. First, it limits the feasibility of applying the method to very large datasets or using very fine-grained occlusion strategies without access to large computational resources. Second, it requires careful planning of the experimental design to ensure that the chosen occlusion parameters (grid resolution, landmark selection, precision mode, scale factor) provide meaningful insights while remaining computationally tractable. Third, it explains why the interpretability analysis in this study was conducted exclusively on cropped datasets rather than analyzing all preprocessing configurations exhaustively, as the computational cost would have been prohibitive for a comprehensive analysis across all experimental conditions. The 4×4 grid resolution used in this study was selected as a practical balance between spatial detail and computational feasibility, providing sufficient granularity to identify regional importance patterns while maintaining manageable resource requirements.

Landmark Detection Accuracy

The landmark-based occlusion strategy relies on accurate facial landmark detection to identify the regions to be occluded. The quality of the interpretability analysis is therefore dependent on the accuracy and consistency of the landmark detection algorithm. Errors in landmark localization can lead to occlusions that do not correspond to the intended facial characteristics, potentially introducing noise into the interpretability results.

Several factors can affect landmark detection accuracy, including image quality, facial pose variations, occlusions in the original image, and the inherent limitations of the

landmark detection model. In the context of morphing detection, landmark detection may be further complicated by the fact that morphed images contain blended facial characteristics that may not correspond perfectly to natural facial geometry. This can potentially lead to landmark detection errors that are more pronounced for morphed images than for *bona fide* images.

To mitigate these concerns, the standardized framework employs robust landmark detection using established libraries and provides configurable precision modes (bbox and pixel precision) with adjustable scale factors. The bbox precision mode with a scale factor of 1.5, used in this study, provides some tolerance for landmark localization errors by expanding the occluded region beyond the exact landmark boundaries. However, users of the framework should be aware that landmark detection accuracy can influence the interpretability results and should manually validate landmark detection quality on their specific datasets when applying the method.

Generalizability to Other MAD Architectures

The occlusion-based interpretability method was developed and validated using the ViT-SVM architecture from Experiment 5 (described in Section 4.2). While the method is designed to be model-agnostic and applicable to any MAD architecture that produces prediction scores, the specific interpretability insights obtained (e.g., the importance of nose and mouth regions, sensitivity to depth transitions) may vary depending on the model architecture and training procedure.

Different MAD models may learn to focus on different facial regions or rely on different types of characteristics for morphing detection. For example, models based on CNNs may exhibit different spatial attention patterns compared to ViT-based models due to differences in their receptive fields and feature extraction mechanisms. Similarly, models trained with different loss functions, data augmentation strategies, or training datasets may develop different decision making strategies that would be reflected in different occlusion heatmap patterns.

Therefore, while the occlusion-based interpretability framework itself is generalizable and can be applied to any MAD model, the specific findings regarding which facial regions are most important should be interpreted in the context of the particular model being analyzed. Applying the method to different MAD architectures would provide valuable comparative insights into how different models approach the morphing detection task and could reveal whether certain architectural choices lead to more robust or interpretable detection strategies. This represents an important direction for future research using the standardized framework developed in this work.

5.5 Conclusions

This chapter presented a standardized, reusable framework for occlusion-based interpretability analysis in MAD systems. The main contributions include:

- **Methodological formalization**: A dual-strategy occlusion approach combining landmark-based (semantic) and grid-based (spatial) analysis, providing comprehensive interpretability through anatomical and spatial perspectives on model behavior.
- **Flexible technical implementation**: A modular three-stage pipeline with configurable occlusion strategies (pixel/bbox precision, adjustable scale factors), dual visualization modes (difference/occluded), and model-agnostic design applicable to any MAD system producing prediction scores.
- **Empirical insights**: Application to the ViT-SVM model revealed that nose and mouth regions are most influential for detection, with particular sensitivity to depth transition areas where morphing artifacts are pronounced.
- **Practical guidance**: Discussion of methodological considerations (grid resolution trade-offs, computational costs, landmark detection accuracy, architectural generalizability) and identification of additional landmarks for future investigation (eyebrows, chin, jawline, forehead, face contour, ears, nasolabial folds).

The framework is fully documented with demonstration notebooks (see Annex J) and establishes a foundation for systematic comparative studies across different MAD architectures, training strategies, and datasets. By enabling researchers to understand which facial regions their models rely on for detection decisions, this work contributes to developing transparent and trustworthy MAD systems for security-critical applications, while also facilitating iterative model refinement and performance optimization.

6 Conclusions

This work investigated the application of deep learning techniques for MAD, with a particular emphasis on evaluating a ViT-based approach and developing an interpretability method to understand the model decision-making process. The work was motivated by the critical security threat posed by morphing attacks to biometric authentication systems, particularly in border control and identity verification scenarios. While morphing attacks represent one category within the broader landscape of facial manipulation techniques (including deepfakes and face swapping, as discussed in Chapter 1), they pose unique challenges for biometric security due to their specific objective of enabling multiple individuals to use a single identity document. Through a comprehensive literature review, systematic experimentation, and the development of a novel interpretability framework, this research contributes to advancing both the effectiveness and transparency of MAD systems.

6.1 Summary of Contributions

The main contributions can be organized into three interconnected areas: literature analysis, experimental validation, and methodological innovation.

6.1.1 Literature Review and Research Gap Identification

The literature review (Chapter 2) analyzed 23 recent studies on deep learning-based MAD, covering a wide range of approaches, including CNNs, GANs, Siamese Networks, Transformer-based models, and hybrid architectures. This systematic analysis revealed several critical findings:

- **Limited code availability**: Only four of 23 reviewed papers provided public implementations, hindering reproducibility.
- **Underexplored architectures**: ViTs remain underutilized for MAD, with only one prior study [24], which achieved SOTA results.
- **Lack of interpretability**: Existing research prioritizes performance metrics over model explainability, essential for trustworthy deployment.
- **Dataset and generalization challenges**: Persistent issues with dataset diversity and cross-dataset generalization require standardized evaluation protocols.

These findings directly influenced the experimental design and methodological choices, ensuring that the research addressed genuine gaps in the current SOTA.

6.1.2 Experimental Findings

The experimental work systematically investigated multiple aspects of deep learning-based MAD, from preprocessing strategies to model architectures and interpretability methods.

Impact of Preprocessing and Dataset Composition

The investigation into image preprocessing and dataset composition revealed critical insights for effective morphing detection:

- **Face alignment consistency**: Face cropping consistently improved detection performance across multiple datasets, demonstrating the importance of standardized preprocessing.
- **Class balancing**: Class balancing showed modest effects in intra-dataset testing, where face cropping dominated. In cross-dataset scenarios, balancing provided more substantial benefits, suggesting a complementary role in enhancing robustness against domain shift.
- **Dataset diversity**: Combined training on multiple datasets outperformed single-dataset training, demonstrating the value of diverse training data for improving generalization.
- **Morphing technique sensitivity**: Landmark-based morphing techniques were easier to detect than GAN-based morphs and dedicated morphing pipelines designed for higher realism.

These findings established preprocessing consistency as a fundamental requirement for effective MAD and informed the design of subsequent investigations, particularly the importance of maintaining consistent preprocessing across training and testing phases.

Vision Transformer Performance

The evaluation of the ViT-SVM architecture provided insights into model performance, generalization capabilities, and interpretability:

Intra-dataset performance: Under consistent training and testing conditions, the model achieved near-perfect performance, demonstrating that ViT architectures can achieve SOTA performance for MAD.

However, preprocessing inconsistencies caused performance degradation (e.g. training on cropped images and testing on uncropped images), reinforcing the critical im-

portance of preprocessing consistency.

Cross-dataset generalization: Cross-dataset performance dropped substantially relative to intra-dataset results, revealing domain shift limitations.

Results reveal an asymmetry in generalization: exposure to advanced morphing techniques improves detection of simpler methods, whereas training on basic techniques provides limited transfer to more complex attacks.

Model interpretability

Occlusion mapping revealed key findings about facial regions contributing to morphing detection:

- **Regional importance hierarchy**: **Nose** and **mouth** regions were most influential, while **eye regions** showed lower but still meaningful importance.
- **Sensitivity to depth transitions**: The model focused on depth transition areas (nose-cheek boundaries, eye corners), suggesting implicit learning of geometric inconsistencies characteristic of morphing artifacts.
- **Visualization mode complementarity**: Landmark-based (semantic) and grid-based (spatial) occlusion provided complementary insights; composite heatmaps offer the most comprehensive interpretability.
- **Cross-dataset interpretability challenges**: Heatmaps were consistent for intra-dataset scenarios but showed irregular patterns in cross-dataset evaluation, providing visual diagnostics for generalization failures.

These interpretability findings not only validate the model's learned behavior but also provide actionable insights for improving morphing detection systems and understanding the vulnerabilities of morphing algorithms.

Novel Composite Heatmap Visualization

A key methodological innovation of this work is the development of a novel visualization technique that unifies landmark-based (semantic) and grid-based (spatial) occlusion analyses into a single representation: composite heatmaps. While occlusion-based interpretability methods have been applied in various computer vision domains, the composite heatmap approach developed here addresses the unique requirements of MAD by simultaneously capturing both the semantic importance of facial landmarks (left eye, right eye, nose, mouth) and the fine-grained spatial sensitivity across the entire face region.

This dual-perspective visualization provides interpretability advantages that neither approach achieves independently: semantic landmarks reveal which facial landmarks are most important for detection, while spatial grids expose localized artifacts and tran-

sition zones that may not align with predefined anatomical structures. The composite heatmap synthesizes these complementary views, enabling researchers to understand model behavior at multiple levels of granularity simultaneously.

6.1.3 Standardized Interpretability Framework

This work formalized the occlusion-based interpretability method into a reusable, standardized framework that represents a methodological contribution to the MAD field. The framework addresses the critical need for transparency in deep learning-based security systems by providing researchers with tools to systematically analyze and interpret model behavior.

The framework's key technical features include:

- **Model-agnostic design**: Applicable to any MAD architecture producing prediction scores.
- **Configurable occlusion strategies**: Flexible precision modes and scale factors adaptable to different image resolutions and landmark detection accuracies.

Practical Guidance and Limitations

The framework documentation includes a discussion of methodological considerations that researchers must understand for appropriate application:

- **Grid resolution trade-offs**: Interpretability depends on the chosen spatial granularity; coarse grids reduce detail, while fine grids may yield less meaningful perturbations.
- **Computational cost considerations**: Computational requirements scale linearly with the number of grid cells.
- **Landmark detection accuracy dependencies**: Interpretability quality is conditioned on the accuracy of facial landmark detection.
- **Architectural generalizability**: Although the approach is model-agnostic, the resulting interpretability patterns may differ across MAD architectures, training procedures, and datasets.

Research Impact and Reusability

The standardized framework, fully documented with demonstration notebooks, establishes a foundation for systematic comparative studies across different MAD architectures, training strategies, and datasets. By enabling researchers to understand which facial regions their models rely on for detection decisions, this work contributes to developing transparent and trustworthy MAD systems for security-critical applications.

The framework's model-agnostic design means it can be applied to any MAD system. This versatility enables comparative interpretability studies that could reveal whether certain architectural choices lead to more robust or interpretable detection strategies, representing an important direction for future research.

6.2 Future Work

While the work presented here has advanced the understanding and effectiveness of deep learning-based MAD, several promising directions for future research have been identified.

6.2.1 Extended Landmark Analysis

The occlusion-based interpretability analysis focused on four key facial landmarks: left eye, right eye, nose, and mouth. A more comprehensive analysis including additional facial regions would provide deeper insights into model behavior and morphing artifact distribution. Recommended landmarks for future investigation include: eyebrows, chin and jawline, forehead, face contour and cheeks, and ears.

Extending the occlusion-based analysis to include these additional landmarks would enable a more complete characterization of which facial regions contribute to morphing detection, inform the development of more targeted detection strategies, and validate whether the observed importance of depth transition areas generalizes to other facial regions with similar geometric properties.

6.2.2 Comparative Interpretability Studies Across Architectures

The standardized interpretability framework was developed and validated using the ViT-SVM architecture. Applying the framework to other MAD architectures would provide valuable comparative insights: CNN-based models, siamese networks, hybrid architectures, and different training strategies.

Such comparative studies could reveal whether certain architectural choices lead to more robust or interpretable detection strategies and whether there are universal patterns in how effective MAD models approach the detection task.

6.2.3 Geometry-Aware Detection Methods

The interpretability analysis revealed that effective models focus on depth transition areas where geometric blending challenges create morphing artifacts. This insight motivates the development of detection methods that explicitly model facial geometry:

3D face reconstruction, depth map analysis, multi-modal fusion, and geometric consistency constraints.

6.2.4 Domain Adaptation and Generalization

Cross-dataset evaluation revealed substantial performance degradation due to domain shift. Future research should explore techniques to improve generalization:

- **Domain adaptation methods**: Apply techniques such as domain adversarial training, self-training, or meta-learning to improve cross-dataset generalization.
- **Data augmentation strategies**: Develop augmentation techniques specifically designed to simulate the variability encountered across different datasets and morphing techniques.
- **Multi-dataset training**: Systematically investigate the benefits of training on multiple diverse datasets simultaneously, including optimal strategies for dataset combination and balancing.
- **Transfer learning approaches**: Explore whether pre-training on large-scale face recognition datasets or other related tasks improves generalization to unseen morphing techniques.

6.2.5 Advanced Morphing Techniques and Adversarial Robustness

As morphing techniques continue to evolve, MAD systems must adapt to detect increasingly sophisticated attacks:

- **Diffusion model morphs**: Investigate detection of morphs generated using SOTA diffusion models, which may produce more realistic blends than traditional techniques.
- **Continual learning**: Develop systems that can adapt to emerging morphing techniques over time without catastrophic forgetting of previously learned detection capabilities.
- **Zero-shot detection**: Investigate whether models can be trained to recognize general morphing artifacts (blending inconsistencies, geometric distortions, unnatural transitions) rather than signatures of specific morphing techniques, enabling detection of novel morphing techniques without requiring examples during training.

6.2.6 Cross-Domain Detection: Morphing Attacks and Deepfakes

While this work focused specifically on MAD, the relationship between morphing detection and other facial manipulation detection tasks, particularly deepfake detection,

presents promising research opportunities. The interpretability framework developed here could be applied to identify which facial regions different manipulation detectors rely on, potentially revealing whether unified detection approaches or transfer learning between manipulation types is feasible. Additionally, as both morphing and deepfakes increasingly rely on advanced generative models, investigating detection methods that identify general generative artifacts rather than manipulation-specific characteristics could improve robustness to novel techniques.

6.2.7 Operational Deployment Considerations

Transitioning MAD systems from research prototypes to operational deployment requires addressing several practical challenges:

- **Real-world image quality**: Evaluate performance on low-resolution, compressed, or printed images typical of operational scenarios, and develop preprocessing or model adaptations to maintain robustness.
- **Video-based MAD**: Extend detection methods to leverage temporal information from video sequences, as explored in recent V-MAD approaches [124], which may provide richer information than still images.
- **Computational efficiency**: Optimize models for real-time inference in resource-constrained environments such as border control checkpoints, potentially through model compression, quantization, or efficient architecture design.
- **Privacy preserving training**: Explore federated learning and other privacy preserving techniques to enable training on sensitive biometric data across multiple institutions without sharing raw data.
- **Explainability for operators**: Develop user interfaces and visualization tools that present interpretability information in actionable formats for border control officers and other operational users.

6.2.8 Standardization and Benchmarking

The field would benefit from community-wide efforts to establish standardized evaluation protocols:

- **Reproducibility standards**: Establish community norms for code and model sharing, experimental documentation, and reproducibility verification.
- **Fairness and bias evaluation**: Systematically evaluate MAD systems for demographic biases and develop techniques to ensure fair performance across different population groups.
- **Interpretability benchmarks**: Extend the standardized interpretability frame-

work to create benchmark interpretability analyses for common MAD architectures, enabling systematic comparison of model behavior.

6.3 Final Remarks

The research has demonstrated that deep learning-based MAD systems can achieve exceptional performance when properly designed and configured, with ViT-based architectures showing particular promise. However, the work has also revealed that achieving robust, generalizable, and interpretable detection remains challenging, requiring careful attention to preprocessing, dataset diversity, and model design.

The development of a standardized interpretability framework represents a step toward transparent and trustworthy MAD systems. By enabling systematic analysis of model behavior, the work contributes to building security systems that are not only accurate but also understandable and auditable, essential properties for deployment in critical applications.

As morphing techniques continue to evolve and become more sophisticated, the MAD field must continue to advance through rigorous experimentation, methodological innovation, and collaborative efforts to establish reproducible and standardized evaluation practices. The findings and tools presented here provide a foundation for these ongoing efforts, contributing to the development of more robust and trustworthy biometric security systems.

Bibliography

[1] M. Ibsen, L. J. Gonzalez-Soler, C. Rathgeb, and C. Busch, "Tetraloss: Improving the robustness of face recognition against morphing attacks," in *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, 2024, pp. 1–10.

[2] S. Baechler, "Document fraud: Will your identity be secure in the twenty-first century?" *European Journal on Criminal Policy and Research*, vol. 26, no. 3, pp. 379–398, 2020.

[3] R. Kessler, K. Raja, J. Tapia, and C. Busch, "Towards minimizing efforts for morphing attacks—deep embeddings for morphing pair selection and improved morphing attack detection," *PLOS ONE*, vol. 19, p. e0304610, 2024.

[4] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23 012–23 026, 2019.

[5] FaceShape, "Face morph tool," 2025, accessed: 2025-01-19. [Online]. Available: https://www.faceshape.com/face-morph

[6] @afroaira, "Social media post on threads," 2025, accessed: 2025-01-19. [Online]. Available: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRKSK153VgujNLAsMhEq75L_6GpbpQhENgIi2490heM_92CrAbZ

[7] Pinterest, "Rihanna at the fenty beauty launch in paris," 2025, accessed: 2025-01-19. [Online]. Available: https://www.pinterest.com/pin/rihanna-at-the-fenty-beauty-launch-in-paris--522136150566546614/

[8] A. Mujahed, "Beyoncé at the 55th annual grammy awards," 2013, accessed: 2025-01-19. [Online]. Available: https://whatsinfashioncwb.wordpress.com/tag/grammy-2013/

[9] Pinterest, "Rihanna makeup at christian dior front row, paris fashion week," 2015, accessed: 2025-01-19. [Online]. Available: https://www.pinterest.com/pin/532902568386886802/

[10] FaceShape, "Face compare tool," 2025, accessed: 2025-01-19. [Online]. Available: https://www.faceshape.com/face-compare

[11] Freepik, "Check icon," 2025, accessed: 2025-01-19. [Online]. Available: https://www.freepik.com/icon/check_5610944

[12] ——, "Delete icon," 2025, accessed: 2025-01-19. [Online]. Available: https://www.freepik.com/icon/delete_5610967

[13] S. R. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, R. Ramachandra, and C. Busch, "Analyzing human observer ability in morphing attack detection—where do we stand?" *IEEE Transactions on Technology and Society*, vol. 4, pp. 125–145, 2023.

[14] R. Raghavendra, K. B. Raja, and C. Busch, "Detecting morphed face images," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2016, pp. 1–7.

[15] C. Busch, "Morphing attack detection - state of the art and challenges," Presentation at the 20th IAPR/IEEE International Summer School for Advanced Studies on Biometrics, pp. 14–22, 2023. [Online]. Available: https://www.christoph-busch.de/files/Busch-MAD-230606.pdf

[16] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and beyond: A survey of face manipulation and fake detection," *Information Fusion*, vol. 64, pp. 131–148, 2020.

[17] F.-A. Croitoru, A.-I. Hiji, V. Hondru, N. C. Ristea, P. Irofti, M. Popescu, C. Rusu, R. T. Ionescu, F. S. Khan, and M. Shah, "Deepfake media generation and detection in the generative ai era: A survey and outlook," 2024, arXiv:2411.19537.

[18] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Face morphing attack generation & detection: A comprehensive survey," *IEEE Transactions on Technology and Society*, vol. PP, pp. 1–1, 2021.

[19] N. Damer, V. Boller, Y. Wainakh, F. Boutros, P. Terhörst, A. Braun, and A. Kuijper, "Detecting face morphing attacks by analyzing the directed distances of facial landmarks shifts," in *Pattern Recognition*, T. Brox, A. Bruhn, and M. Fritz, Eds. Springer International Publishing, 2019, pp. 518–534.

[20] P. Morais, I. Domingues, and J. Bernardino, "Deep learning techniques for detecting morphed face images: A literature review," *IEEE Access*, vol. 13, pp. 105 952–105 981, 2025.

[21] P. Morais and I. Domingues, "Composite heatmaps: Interpreting deep models in morph attack detection," in *14th International Symposium on Image and Signal Processing and Analysis (ISPA)*, 2025, pp. 1–6.

[22] I. Medvedev, F. Shadmand, and N. Goncalves, "Mordeephy: Face morphing detection via fused classification," *International Conference on Pattern Recognition Applications and Methods*, vol. 1, pp. 193–204, 2023.

[23] I. Medvedev, J. A. Pimenta, and N. Goncalves, "Fused classification for differential face morphing detection," *IEEE Winter Conference on Applications of Computer Vision Workshops, WACVW*, pp. 1043–1050, 2024.

[24] H. Zhang, R. Ramachandra, K. Raja, and C. Busch, "Generalized single-image-based morphing attack detection using deep representations from vision transformer," *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1510–1518, 2024.

[25] N. D. Domenico, G. Borghi, A. Franco, and D. Maltoni, "Combining identity features and artifact analysis for differential morphing attack detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14233 LNCS. Springer Science and Business Media Deutschland GmbH, 2023, pp. 100–111.

[26] C. K. Jia, Y. C. Liu, and Y. L. Chen, "Face morphing attack detection based on high-frequency features and progressive enhancement learning," *Frontiers in Neurorobotics*, vol. 17, pp. 1–11, 2023.

[27] I. S. Razaq and B. K. Shukur, "Combining wavelet transforms features and high-level features using CNN for face morphing attack detection," *International Journal of Information Technology (Singapore)*, vol. 15, pp. 3957–3966, 2023.

[28] J. M. Singh and R. Ramachandra, "3-D face morphing attacks: Generation, vulnerability and detection," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, pp. 103–117, 2024.

[29] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *IEEE International Joint Conference on Biometrics*, 2014, pp. 1–7.

[30] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, "Detection of face morphing attacks based on PRNU analysis," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 1, pp. 302–317, 2019.

[31] P. Sharma, M. Kumar, and H. Sharma, "Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation," *Multimedia Tools and Applications*, vol. 82, no. 12, pp. 18 117–18 150, 2023.

[32] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 864–879, 2015.

[33] G. Borghi, E. Pancisi, M. Ferrara, and D. Maltoni, "A double siamese framework for differential morphing attack detection," *Sensors*, vol. 21, 2021.

[34] J. M. Singh and R. Ramachandra, "Reliable face morphing attack detection in on-the-fly border control scenario with variation in image resolution and capture distance," *IEEE International Joint Conference on Biometrics, IJCB*, 2022.

[35] M. Ferrara, A. Franco, and D. Maltoni, "Face morphing detection in the presence of printing/scanning and heterogeneous image sources," *IET Biometrics*, vol. 10, pp. 290–303, 2021.

[36] U. Scherhag, C. Rathgeb, and C. Busch, *Face Morphing Attack Detection Methods*. Cham: Springer International Publishing, 2022, pp. 331–349.

[37] J. M. Singh, R. Ramachandra, K. B. Raja, and C. Busch, "Robust morph-detection at automated border control gate using deep decomposed 3D shape & diffuse reflectance," in *15th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, 2019, pp. 106–112.

[38] A. Makrushin, T. Neubert, and J. Dittmann, "Automatic generation and detection of visually faultless facial morphs," *VISIGRAPP - 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, vol. 6, pp. 39–50, 2017.

[39] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Are GAN-based morphs threatening face recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2959–2963, 2022.

[40] International Organization for Standardization, "ISO/IEC 30107-3:2023 Information technology — Biometric presentation attack detection — Part 3: Testing and reporting," Geneva, Switzerland, 2023.

[41] K. Raja, M. Ferrara, A. Franco, L. Spreeuwers, I. Batskos, F. D. Wit, M. Gomez-Barrero, U. Scherhag, D. Fischer, S. K. Venkatesh, J. M. Singh, G. Li, L. Bergeron, S. Isadskiy, R. Ramachandra, C. Rathgeb, D. Frings, U. Seidel, F. Knopjes, R. Veldhuis, D. Maltoni, and C. Busch, "Morphing attack detection-database, evaluation platform, and benchmarking," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 4336–4351, 2021.

[42] A. Sequeira, "Liveness detection and robust recognition in iris and fingerprint biometric systems," Ph.D. dissertation, Faculdade de Engenharia da Universidade do Porto, 2015.

[43] S. Gonzalez and J. Tapia, "Improving presentation attack detection for ID cards on remote verification systems," 2023, arXiv:2301.09542.

[44] J. Kolberg, D. Gläsner, R. Breithaupt, M. Gomez-Barrero, J. Reinhold, A. von Twickel, and C. Busch, "On the effectiveness of impedance-based fingerprint presentation attack detection," *Sensors*, vol. 21, no. 17, 2021.

[45] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 2011.

[46] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, and C. Busch, "Detecting morphed face attacks using residual noise from deep multi-

scale context aggregation network," *IEEE Winter Conference on Applications of Computer Vision, WACV*, pp. 269–278, 2020.

[47] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello, "Border control morphing attack detection with a convolutional neural network de-morphing approach," *IEEE Access*, vol. 8, pp. 92 301–92 313, 2020.

[48] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper, "PW-MAD: Pixel-wise supervision for generalized face morphing attack detection," *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 13017 LNCS, pp. 291–304, 2021.

[49] B. Chaudhary, P. Aghdaie, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Differential morph face detection using discriminative wavelet sub-bands," in *Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE Computer Society, 2021, pp. 1425–1434.

[50] C. Liu, M. Ferrara, A. Franco, G. Borghi, and D. Zhong, "Differential morphing attack detection via triplet-based metric learning and artifact extraction," in *International Conference of the Biometrics Special Interest Group* (*BIOSIG*), 2024, pp. 1–7.

[51] L. Colbois and S. Marcel, "Evaluating the effectiveness of attack-agnostic features for morphing attack detection," in *IEEE International Joint Conference on Biometrics* (*IJCB*), 2024, pp. 1–9.

[52] I. R. Institute, "FRLL-Morphs dataset," accessed: 2024-12-26. [Online]. Available: https://www.idiap.ch/en/scientific-research/data/frll-morphs

[53] A. M. S. L. (AMSL), "AMSL face morph image dataset," accessed: 2025-01-20. [Online]. Available: https://omen.cs.uni-magdeburg.de/disclaimer/index.php

[54] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, "Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biometrics*, vol. 7, no. 4, pp. 325–332, 2018.

[55] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman, "VGGFace2 dataset," accessed: 2024-12-26. [Online]. Available: https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/

[56] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," 2020, arXiv:2012.05344.

[57] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Transactions on Information Forensics and Security*, vol. 13, pp. 1008–1017, 2018.

[58] N. Damer, A. M. Saladie, A. Braun, and A. Kuijper, "Morgan: Recognition vulnerability and attack detectability of face morphing attacks created by generative

adversarial network," *IEEE 9th International Conference on Biometrics Theory, Applications and Systems, BTAS*, 2018.

[59] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros, "Privacy-friendly synthetic data for the development of face morphing attack detectors," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2022, pp. 1605–1616.

[60] G. Borghi, N. D. Domenico, M. Ferrara, A. Franco, U. Latif, and D. Maltoni, "Monot: High-quality privacy-compliant morphed synthetic images for everyone," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2024, pp. 1–8.

[61] G. Spathis, N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Amonot: Synthetic aging for differential morphing attack detection systems," in *International Conference on Pattern Recognition Workshops (ICPRW)*, 2024.

[62] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, p. 84–90, 2017.

[63] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[64] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4685–4694.

[65] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 77–85.

[66] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015.

[67] V. H. Phung and E. J. Rhee, "A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets," *Applied Sciences*, vol. 9, no. 21, 2019.

[68] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 2261–2269.

[69] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, arXiv:1409.1556.

[70] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *British Machine Vision Conference (BMVC)*, X. Xie, M. W. Jones, and G. K. L. Tam, Eds. BMVA Press, 2015, pp. 41.1–41.12.

[71] L. Qin, F. Peng, and M. Long, "Face morphing attack detection and localization based on feature-wise supervision," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3649–3662, 2022.

[72] F. R. L. London, "Face research lab London set," 2017, accessed: 2025-01-18. [Online]. Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666

[73] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[74] J. Cunningham, D. Shu, T. Simpson, and C. Tucker, "A sparsity preserving genetic algorithm for extracting diverse functional 3D designs from deep generative neural networks," *Design Science*, vol. 6, 2020.

[75] M. Fang, F. Boutros, and N. Damer, "Unsupervised face morphing attack detection via self-paced anomaly detection," in *IEEE International Joint Conference on Biometrics, IJCB*. Institute of Electrical and Electronics Engineers Inc., 2022.

[76] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 3, 2014.

[77] B. Sabiri, B. E. Asri, and M. Rhanoui, "Impact of hyperparameters on the generative adversarial networks behavior," *International Conference on Enterprise Information Systems, ICEIS*, vol. 1, pp. 428–438, 2022.

[78] S. Banerjee and A. Ross, "Conditional identity disentanglement for differential face morph detection," *IEEE International Joint Conference on Biometrics, IJCB*, 2021.

[79] N. Damer, F. Boutros, A. M. Saladié, F. Kirchbuchner, and A. Kuijper, "Realistic dreams: Cascaded enhancement of GAN-generated images with an example in face morphing attacks," in *IEEE 10th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, 2019, pp. 1–10.

[80] Z. Blasingame and C. Liu, "Leveraging adversarial learning for the detection of morphing attacks," in *IEEE International Joint Conference on Biometrics, IJCB*. Institute of Electrical and Electronics Engineers Inc., 2021.

[81] G. H. de Rosa and J. P. Papa, "Chapter 7 - learning to weight similarity measures with siamese networks: a case study on optimum-path forest," in *Optimum-Path Forest*, A. X. Falcão and J. P. Papa, Eds. Academic Press, 2022, pp. 155–173.

[82] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[83] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First*

AAAI Conference on Artificial Intelligence, ser. AAAI'17. AAAI Press, 2017, p. 4278–4284.

[84] S. Soleymani, B. Chaudhary, A. Dabouei, J. Dawson, and N. M. Nasrabadi, "Differential morphed face detection using deep siamese networks," *Lecture Notes in Computer Science* (*including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics*), vol. 12666 LNCS, pp. 560–572, 2021.

[85] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, arXiv:2010.11929.

[86] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[87] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Morph detection enhanced by structured group sparsity," *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops, WACVW*, pp. 311–320, 2022.

[88] A. I. L. of FEI, "FEI face database," accessed: 2024-12-26. [Online]. Available: https://fei.edu.br/~cet/facedatabase.html

[89] M.-T. Lab, "FEI-Morph dataset," 2023, accessed: 2025-01-18. [Online]. Available: https://miatbiolab.csr.unibo.it/fei-morph-dataset/

[90] F. Peng, L. B. Zhang, and M. Long, "FD-GAN: Face de-morphing generative adversarial network for restoring accomplice's facial image," *IEEE Access*, vol. 7, pp. 75 122–75 131, 2019.

[91] N. I. of Standards and T. (NIST), "Face recognition grand challenge (FRGC)," accessed: 2024-12-26. [Online]. Available: https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc

[92] T. Karras, S. Laine, and T. Aila, "Flickr-Faces-HQ Dataset (FFHQ)," accessed: 2024-12-26. [Online]. Available: https://github.com/NVlabs/ffhq-dataset

[93] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), 2017, pp. 1800–1807.

[94] M. Grgic, K. Delac, S. Grgic, and B. Klimpak, "Scface - surveillance cameras face database," University of Zagreb, Faculty of Electrical Engineering and Computing, accessed: 2024-12-26. [Online]. Available: https://www.scface.org/

[95] M. Robledo-Moreno, G. Borghi, N. Di Domenico, A. Franco, K. Raja, and D. Maltoni, "Towards federated learning for morphing attack detection," in *IEEE Inter-*

national Joint Conference on Biometrics (IJCB), 2024, pp. 1–10.

[96] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[97] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: deep hierarchical feature learning on point sets in a metric space," in *31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5105–5114.

[98] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting Point Cloud Classification: A New Benchmark Dataset and Classification Model on Real-World Data ," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. Los Alamitos, CA, USA: IEEE Computer Society, 2019, pp. 1588–1597.

[99] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: a large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[100] H. Zhu, H. Yang, L. Guo, Y. Zhang, Y. Wang, M. Huang, M. Wu, Q. Shen, R. Yang, and X. Cao, "FaceScape: 3D facial dataset and benchmark for single-view 3D face reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2023.

[101] H. Zhu, H. Yang, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape," accessed: 2024-12-26. [Online]. Available: https://facescape.nju.edu.cn/

[102] NVIDIA, "StyleGAN2," accessed: 2025-01-05. [Online]. Available: https://github.com/NVlabs/stylegan2

[103] FaceMorpher, "FaceMorpher," accessed: 2025-01-05. [Online]. Available: http://www.facemorpher.com/

[104] O. Team, "OpenCV," accessed: 2025-01-05. [Online]. Available: https://opencv.org/

[105] M. Ngan, P. Grother, K. Hanaoka, and J. Kuo, "Face recognition vendor test (FRVT) part 4: Morph - performance of automated face morph detection," National Institute of Standards and Technology (NIST), Gaithersburg, MD, USA, Tech. Rep. NISTIR 8292, 2020.

[106] L. Omelina, J. Goga, J. Pavlovicova, M. Oravec, and B. Jansen, "A survey of iris datasets," *Image and Vision Computing*, vol. 108, p. 104109, 2021.

[107] W. V. University, "Twins day dataset 2010–2015," accessed: 2025-01-05. [Online]. Available: https://biic.wvu.edu/data-sets/twins-day-dataset-2010-1015

[108] K. Hollingsworth, K. W. Bowyer, and P. J. Flynn, "Similarity of iris texture between identical twins," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, 2010, pp. 22–29.

[109] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt, "Distinguishing identical twins by face recognition," in *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2011, pp. 185–192.

[110] B. S. Laboratory, "FVC-onGoing - BOEP," accessed: 2025-05-03. [Online]. Available: https://biolab.csr.unibo.it/fvcongoing/UI/Form/BOEP.aspx

[111] ——, "FVC-onGoing," accessed: 2025-05-03. [Online]. Available: https://biolab.csr.unibo.it/FvcOnGoing/UI/Form/Home.aspx

[112] B. Dorizzi, R. Cappelli, M. Ferrara, D. Maio, D. Maltoni, N. Houmani, S. Garcia-Salicetti, and A. Mayoue, "Fingerprint and on-line signature verification competitions at ICB 2009," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 725–732.

[113] N. I. of Standards and Technology, "Face Analysis Technology Evaluation (FATE) MORPH," accessed: 2025-05-03. [Online]. Available: https://pages.nist.gov/frvt/html/frvt_morph.html

[114] N. U. of Science and Technology, "State of the art of Morphing Detection (SOTAMD)," accessed: 2025-05-03. [Online]. Available: https://www.ntnu.edu/iik/sotamd

[115] B. S. Laboratory, "FVC-onGoing - Background," accessed: 2025-05-03. [Online]. Available: https://biolab.csr.unibo.it/FvcOnGoing/UI/Form/Background.aspx

[116] H. O. Shahreza and S. Marcel, "Unveiling synthetic faces: How synthetic datasets can expose real identities," arXiv:2410.24015.

[117] S. Zhou, K. C. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," in *NeurIPS*, 2022.

[118] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Face restoration for morphed images retouching," in *12th International Workshop on Biometrics and Forensics (IWBF)*, 2024, pp. 1–6.

[119] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2096–2030, 2016.

[120] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[121] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 8916–8925.

[122] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, ser. EAAMO '21. New York, NY, USA: Association for Computing Machinery, 2021.

[123] Y. Li, Q. Li, H. Wang, R. Li, W. Zhong, and G. Zhang, "Towards efficient replay in federated incremental learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 12 820–12 829.

[124] G. Borghi, A. Franco, N. di Domenico, M. Ferrara, and D. Maltoni, "V-MAD: Video-based morphing attack detection in operational scenarios," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–10.

[125] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Munyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the opportunities and risks of foundation models," 2022, arXiv:2108.07258.

[126] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, Oct 2016.

[127] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Com-*

munication Networks and Services, vol. 1, pp. 39–48, 2018.

[128] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, arXiv:1702.08608.

[129] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[130] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," 2017, arXiv:1706.03825.

[131] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st International Conference on Neural Information Processing Systems*, ser. NIPS. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[132] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[133] D. Yi, Z. Lei, S. Liao, and S. Li, "Learning face representation from scratch," 2014, arXiv:1411.7923.

[134] D. T. Roy, "CASIA-WebFace dataset," 2023, accessed: 2025-11-03. [Online]. Available: https://www.kaggle.com/datasets/debarghamitraroy/casia-webface

[135] P. I. C. at Stirling (PICS), "Psychological image collection at stirling (PICS)," accessed: 2024-12-26. [Online]. Available: https://pics.stir.ac.uk/

[136] G. Borghi, N. D. Domenico, A. Franco, M. Ferrara, and D. Maltoni, "Revelio: A modular and effective framework for reproducible training and evaluation of morphing attack detectors," *IEEE Access*, vol. 11, pp. 120 419–120 437, 2023.

[137] M.-T. Lab, "ChiMo dataset," 2023, accessed: 2025-01-18. [Online]. Available: https://miatbiolab.csr.unibo.it/chimo-dataset/

[138] N. I. of Standards and T. (NIST), "Face recognition technology (FERET)," accessed: 2024-12-26. [Online]. Available: https://www.nist.gov/programs-projects/face-recognition-technology-feret

[139] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer, "Detecting facial retouching using supervised deep learning," *IEEE Transactions on Information Forensics and Security*, vol. 11, pp. 1903–1913, 2016.

[140] C. V. R. L. (CVRL), "ND-IIITD retouched face database," accessed: 2024-12-28. [Online]. Available: https://cvrl.nd.edu/projects/data/#nd-iiitd-retouched-face-database

[141] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2020, arXiv:1905.11946.

[142] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 6848–6856.

[143] A. Martinez and R. Benavente, "The AR face database," *Tech. Rep. 24 CVC Technical Report*, 1998.

[144] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[145] D. E. King, "dlib c++ library - face landmark detection example," 2009, accessed: 2025-07-09. [Online]. Available: https://dlib.net/face_landmark_detection.py.html

Annexes

Annex A - IEEE Access Journal Paper

IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 11 May 2025, accepted 29 May 2025, date of publication 9 June 2025, date of current version 25 June 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3578199

SURVEY

Deep Learning Techniques for Detecting Morphed Face Images: A Literature Review

PEDRO MORAIS[1], INÊS DOMINGUES[1,2], AND JORGE BERNARDINO[1,3], (Member, IEEE)
[1]Coimbra Institute of Engineering, Polytechnic University of Coimbra, 3045-093 Coimbra, Portugal
[2]Research Center of the Portuguese Institute of Oncology of Porto (CI-IPOP), Medical Physics, Radiobiology, and Radiological Protection Group, 4200-072 Porto, Portugal
[3]CISUC/SSE-Centre for Informatics and Systems, University of Coimbra, Polo II, 3030-290 Coimbra, Portugal

Corresponding author: Pedro Morais (a21280686@isec.pt)

This work was supported by IEEE Access by waiving the APC.

ABSTRACT Face morphing, as a digital image manipulation technique, is becoming increasingly prevalent, posing challenges in various fields, including biometric security, forensics, and identity verification systems. This paper provides a comprehensive review of the state-of-the-art of deep learning techniques employed in morphed image detection, including Convolutional Neural Networks (CNNs), Autoencoders, and Generative Adversarial Networks (GANs). These models have shown considerable potential in automating feature extraction, identifying subtle manipulation traces, and improving detection accuracy across different morphing techniques. However, as morphing techniques continue to evolve and produce more sophisticated attacks, current methods face challenges in detecting subtle artifacts, particularly in high-quality morphs. The review also highlights the issues related to dataset diversity and model generalization, noting that many approaches struggle with unseen morphing attacks and real-world scenarios, such as low-resolution, compressed, or printed & scanned images. Despite these challenges, deep learning techniques have made significant progress in improving Morph Attack Detection (MAD) systems. The paper suggests future directions for enhancing the robustness and generalization of these systems, focusing on expanding datasets, improving model adaptability, and developing methods that can detect increasingly sophisticated morphing attacks in real-world applications.

INDEX TERMS Face morphing, morphed image detection, deep learning, generative adversarial networks, convolutional neural networks.

I. INTRODUCTION

The falsification of identification documents presents a remarkable threat to public security, personal data protection, and fraud prevention. A growing concern within this domain is the use of morphed facial images – a deceptive technique that has become more prevalent with the rapid advancement of artificial intelligence and image processing technologies [1], [2].

Morphed images are created by blending the facial features of two or more individuals to create a composite image that resembles all contributors, as illustrated in Figure 1. This manipulation enables multiple individuals to use a single identification document, for example, to gain unauthorized access across international borders [3], [4].

Detecting these morphed images poses a considerable challenge. Humans are often unable to detect the manipulation, and even advanced facial recognition systems struggle [13]. As the sophistication of morphing techniques continues to evolve, attackers are able to produce more convincing morphs, making detection increasingly difficult. This requires ongoing research and innovation to develop more effective detection methods and keep up with the rapidly advancing techniques used in morphing attacks.

In response to this challenge, numerous studies have explored the field of Morphing Attack Detection (MAD) [14], [15], [16], [17], [18], [19], [20]. These approaches offer varying degrees of success in detecting morphed images.

The associate editor coordinating the review of this manuscript and approving it for publication was Abedalrhman Alkhateeb.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

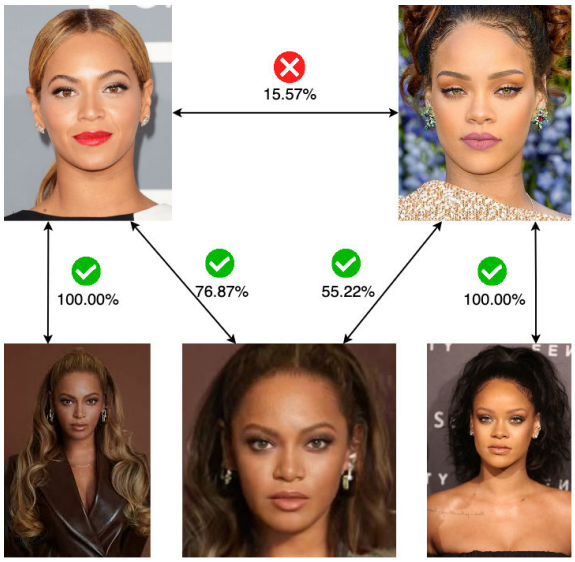


FIGURE 1. Example of a morphed image generated using the online FaceShape face morph tool available at [5] with source images [6], [7]. The resulting morph was compared to reference images [8], [9] using the FaceShape compare faces tool available at [10]. Additional icons used in the diagram are credited to [11] and [12]. The methodology was inspired by the survey presented in [4].

Importantly, this review also emphasizes the practical implications of MAD techniques, particularly in the context of biometric security systems used at border checkpoints, automated identity verification in e-government services, and forensic analysis for criminal investigations. As morphed images undermine the reliability of facial recognition systems, robust MAD solutions are critical for preventing identity fraud, enforcing immigration policies, and supporting legal proceedings that rely on the integrity of biometric evidence. In border control scenarios, for example, the inability to detect a morphing attack can lead to unauthorized individuals crossing borders undetected, posing a risk to national and international security. In forensic analysis, undetected morphs may compromise the attribution of digital identities in criminal cases, which could affect both investigations and judicial outcomes. Therefore, the development and deployment of accurate and generalizable MAD techniques have tangible consequences for public safety, legal accountability, and the trustworthiness of modern identity systems [21], [22].

The objective of this paper is to provide a comprehensive review of recent advances in MAD, with a particular focus on deep learning techniques. By critically analyzing current approaches and results, this literature review aims to identify the existing research gaps and highlight areas where further investigation is needed.

While this review focuses on deep learning-based MAD, related methods have also addressed broader biometric security tasks such as spoofing detection and template protection using deep learning techniques [23]. Before the adoption of these models, researchers explored forensic approaches such as symmetry analysis and compression artifacts to detect morphs, though these proved limited in practical scenarios [24]. Handcrafted features like Local Binary Patterns (LBP) and Binarized Statistical Image Features (BSIF) were also used, but their performance degraded under compression and cross-domain variation [25], [26].

Given the severity of the threat posed by morphing attacks to public security, this paper seeks to address the following research question:

RQ: How can deep learning techniques be used to improve MAD?

To answer this question, a comprehensive literature review was conducted, evaluating the effectiveness of deep learning-based approaches in MAD. This review assesses their strengths and limitations, highlighting areas where further research is needed to advance the field.

The key contributions of this paper are as follows:

- Understanding the role of deep learning techniques in MAD.
- Identifying key areas for future research to improve detection capabilities.

The structure of the paper is as follows: Section II explains the background concepts. Section III outlines the methodologies used for the literature selection. Section IV reviews recent studies on deep learning techniques for MAD. Section V summarises the key findings from the literature. Section VI identifies the major research gaps in the field. Section VII proposes future research directions. Finally, Section VIII presents the main conclusions of the paper.

II. BACKGROUND CONCEPTS

In this section, we introduce the key concepts necessary for understanding MAD, particularly the approaches and metrics discussed in the literature review.

A. D-MAD AND S-MAD

MAD can be broadly categorized into two types: Differential Morphing Attack Detection (D-MAD) and Single-image Morphing Attack Detection (S-MAD).

D-MAD involves comparing a suspected morphed image with a trusted reference image, typically obtained from a live capture. The goal of D-MAD is to identify discrepancies between the two images, which could indicate that the suspected image has been morphed. This approach is widely explored in studies, such as those by Borghi et al. [27] and Singh and Ramachandra [28].

S-MAD, on the other hand, focuses on detecting whether a given single facial image has been morphed without any reference image available for comparison. It is a more challenging task due to the lack of a ground-truth reference, relying solely on inherent features of the image itself. Methods like the one proposed by Ferrara et al. [29] and Medvedev et al. [14] exemplify S-MAD approaches.

While both D-MAD and S-MAD aim to detect morphing attacks, their effectiveness and applicability differ depending on the operational context. D-MAD generally achieves higher

detection accuracy by leveraging a trusted reference image to spot subtle inconsistencies. However, it is only feasible in verification scenarios–such as border control–where a live capture or previously enrolled image is available. In contrast, S-MAD is crucial in standalone settings, like passport issuance, where only the submitted image can be analyzed. Despite being more widely applicable, S-MAD methods face greater challenges, especially when dealing with high-quality morphs or images affected by compression and resolution loss [30], [31]. Thus, the choice between D-MAD and S-MAD should be guided by the availability of reference data and the specific requirements of the biometric application.

Figure 2 illustrates the adoption of S-MAD and D-MAD techniques by the reviewed literature, with some papers exploring both approaches.

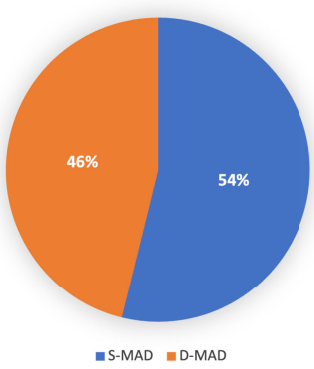


FIGURE 2. Proportion of papers employing S-MAD and D-MAD techniques.

B. BONA FIDE AND ATTACK IMAGES

In the context of MAD, two types of images are commonly referenced:

- **Bona Fide Image**: A bona fide image is a genuine, unaltered facial photograph used in identification or verification systems. It represents the actual appearance of an individual without any manipulation or morphing. The goal of the detection system is to correctly identify these images as authentic [32].
- **Attack Image**: An attack image refers to a manipulated or morphed facial image used with the intent to deceive identification systems. In the context of MAD, these are the images created by blending features from multiple individuals to create a composite that resembles each contributor. Detecting these morphs is critical to prevent fraudulent access or identity misuse. Studies such as those by Makrushin et al. [33] and Sarkar et al. [34] investigate the characteristics and challenges of detecting these attack images.

C. EVALUATION METRICS

Various metrics are used to evaluate the performance of MAD systems, including those specified by ISO/IEC 30107-3, which provides guidelines for the testing and performance evaluation of biometric systems in detecting morphing attacks [35]. Key metrics commonly referenced include:

- **Attack Presentation Classification Error Rate (APCER)**: APCER represents the proportion of attack samples (i.e., morphed images) that are incorrectly classified as bona fide by the MAD system. It is conceptually similar to the False Acceptance Rate (FAR), as both metrics indicate the system's tendency to erroneously accept attack samples [35], [36], [37]. The formula for APCER is:

$$\text{APCER} = \frac{\text{Number of false acceptances}}{\text{Total number of attack presentations}}.$$

  In the context of morphed images, APCER and Morphing Attack Classification Error Rate (MACER) are often treated as equivalent, as both measure the FAR for morphed faces. In this way, note that the name of the metric has been updated to MACER in some contexts, in accordance with a new ISO standard currently under development.
- **Bona Fide Presentation Classification Error Rate (BPCER)**: BPCER measures the proportion of bona fide samples that are incorrectly classified as attacks by the system. It is analogous to the False Rejection Rate (FRR), indicating the system's propensity to falsely reject genuine samples [35], [36], [37]. The formula for BPCER is:

$$\text{BPCER} = \frac{\text{Number of false rejections}}{\text{Total number of bona fide presentations}}.$$

- **BPCER @ APCER**: In MAD tasks, BPCER is typically assessed at specific APCER values that reflect common operating points in real-world applications. These thresholds are chosen to balance the trade-off between false acceptance and false rejection. This approach emphasizes security, as it prioritizes minimizing the acceptance of attacks by fixing APCER and evaluating the impact on BPCER. The relationship between APCER and the corresponding BPCER is often illustrated using a Detection Error Tradeoff (DET) curve, which also facilitates the estimation of the Equal Error Rate (EER) [38].
- **APCER @ BPCER**: Although less commonly used than BPCER @ APCER, this metric evaluates APCER while maintaining a fixed BPCER, focusing on usability by minimizing the impact on bona fide users. It measures how effectively the system detects attacks while ensuring an acceptable rate of false rejections [39].
- **Equal Error Rate (EER) / Detection Equal Error Rate (D-EER)**: EER and D-EER refer to the operational point at which APCER equals BPCER. While EER is commonly used across biometric evaluation tasks, D-EER is specifically applicable to MAD, representing the point where the attack and bona fide errors are balanced. A lower D-EER signifies better overall system performance, as it shows a reduced error rate across

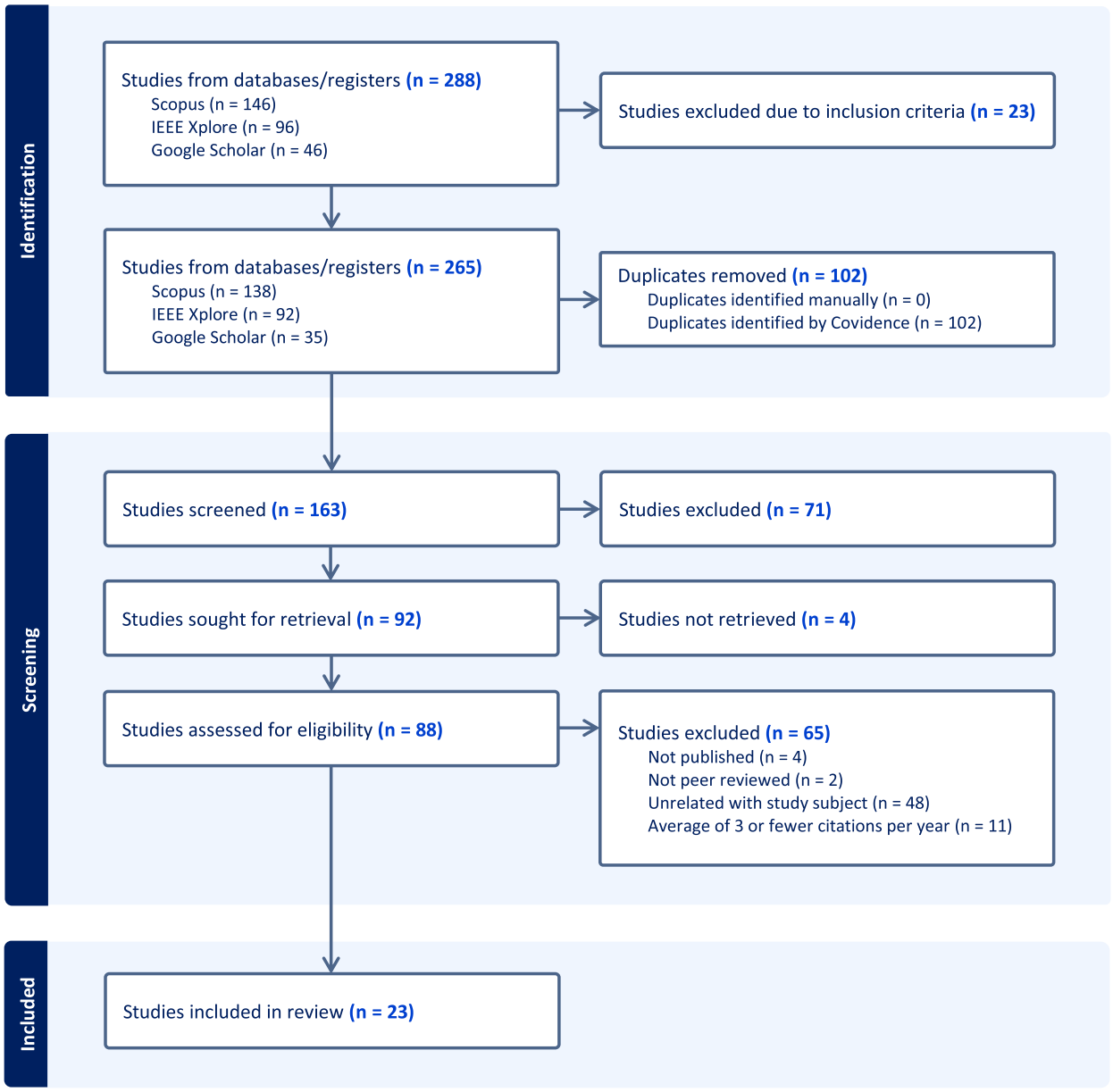


FIGURE 3. Methodology applied in the literature review.

both attack and bona fide classifications [25], [36]. The D-EER value is determined as the point where:

$$\text{APCER} = \text{BPCER}.$$

- **Average Classification Error Rate (ACER)**: ACER is the average of APCER and BPCER, providing a balanced measure of the system's performance. Although ACER provides a single performance metric for MAD systems, it assumes that both types of errors are equally significant. However, in practical applications, APCER and BPCER may have different priorities, and ACER may not fully capture the trade-offs between these error rates [37]. The formula for ACER is:

$$\text{ACER} = \frac{\text{APCER} + \text{BPCER}}{2}.$$

- **Accuracy (ACC)**: Accuracy measures the overall performance by calculating the ratio of correctly classified samples to the total number of samples [19]. While accuracy provides an overview of performance, it is not commonly used in MAD evaluations. In imbalanced datasets, accuracy can be misleading, as a high accuracy may occur by simply favoring the majority class (bona fide samples), neglecting the importance of detecting attacks [40]. The formula for accuracy is:

$$\text{ACC} = \frac{\text{Number of correctly classified samples}}{\text{Total number of samples}}.$$

III. LITERATURE REVIEW METHODOLOGY

The current section outlines the process used for the literature review, as shown in Figure 3. It includes the selection of

Data Sources, the construction of the Search Query, and the Inclusion and Exclusion Criteria applied to filter relevant studies. Finally, the Characterization of Selected Papers presents an analysis of the selected papers.

A. DATA SOURCES

Three data sources were used to gather literature for this review: Google Scholar, IEEE Xplore, and Scopus. Google Scholar (https://scholar.google.com/) provides broad coverage of peer-reviewed papers, while IEEE Xplore (https://ieeexplore.ieee.org/) offers specialized content in technical fields like deep learning and image processing. Scopus (https://www.scopus.com) ensures access to a wide range of scientific research, allowing for a comprehensive literature review across relevant disciplines.

B. SEARCH QUERY

On 12 January 2025, the search query ""morph* attack* detection'' OR ''face morph* detection"" was performed in Google Scholar and returned 46 results. The same query applied in IEEE Xplore returned 96 results, while Scopus returned a total of 146 results.

C. INCLUSION CRITERIA

The inclusion criteria for this review included papers published between 2019 and 2024. This timeframe was chosen to ensure that the studies reviewed were relevant to the current landscape of the field. After applying these criteria, the filtered results included 35 papers from Google Scholar, 92 from IEEE Xplore, and 138 from Scopus.

D. EXCLUSION CRITERIA

Several exclusion criteria were applied in this review. First, publications not written in English were excluded, as were those not peer-reviewed and for which full-text was not available. Literature reviews were also excluded. Finally, papers deemed irrelevant to the research question were eliminated, as were those with an average of three or fewer citations per year, as determined from Google Scholar. This last criterion was not applied to publications from 2024 onwards due to the limited period for citation acquisition.

E. CHARACTERIZATION OF SELECTED PAPERS

Figure 3 also shows the total number of papers retrieved from Scopus (146), IEEE Xplore (96), and Google Scholar (46) as data sources. A total of 265 papers were initially identified based on the inclusion criteria, which required articles to be published between 2019 and 2024.

After removing duplicates, the total number of papers was reduced to 163. As shown in Figure 4, face MAD research has experienced steady growth, with a moderate number of publications in 2019 and 2020, and a significant increase starting in 2022. This trend highlights the increasing recognition of face morphing as a serious security threat, and the continued relevance of MAD techniques.

The review process narrowed down the selection to include only those papers that were most relevant to the study, focusing on their contribution to the detection of face morphing attacks using deep learning techniques.

Using the exclusion criteria outlined in Section III-D, the number of papers was further reduced to 23. One of the exclusion criteria was to remove papers with an average of three or fewer citations per year. The distribution of papers by average number of citations per year is shown in Figure 5, where six papers with [0-1] citations per year correspond to those published in 2024.

The distribution per year for each of the final 23 papers selected is illustrated in Figure 6. In this graph, the papers are ordered by their total citation count, from highest to lowest, to highlight the relative impact of each work over time. The citation counts were collected from Google Scholar as of 27 April 2025. This visualization provides an overview of how the relevance and influence of each paper has progressed, reflecting both the maturity of the research and the growing interest in face morphing detection techniques using deep learning.

The citation analysis presented in Figure 6 reveals that Ferrara et al. [29] stands out as the most cited work across the reviewed period (2019–2025), consistently accumulating a high number of citations and peaking at around 20 citations in 2022, followed by sustained influence through 2024.

Other highly cited papers include Venkatesh et al. [41], which exhibited a sharp rise and a strong peak in 2022, as well as Ortega-Delcampo et al. [42] and Damer et al. [43], both of which showed significant citation growth between 2021 and 2023. Similarly, Borghi et al. [27] and Chaudhary et al. [44] demonstrated steady citation increases over the years, indicating ongoing relevance within the community.

These highly cited papers dominate the field, suggesting that they have had a foundational impact on advancing research in MAD using deep learning approaches.

Looking at the overall citation evolution, some general trends can be observed:

- There is a clear peak around 2022, where the majority of the papers reached their highest yearly citation counts.
- After 2022, a gradual decline in yearly citations becomes apparent, although several papers continued to attract moderate attention through 2024 and into 2025.
- Recent papers published in 2023 and 2024 (e.g., Medvedev et al. [14], Singh and Ramachandra [20]) are starting to accumulate citations but have not yet reached the impact levels of earlier publications, which is expected given their limited exposure time.
- Papers from 2024, including Liu et al. [45] and Colbois et al. [46], currently show low citation counts, consistent with their recent release.

The overall trends suggests that the research area experienced a surge in activity and visibility around 2022, likely reflecting an intensified research interest during that period, followed by a phase of stabilization. This pattern is typical of maturing fields where initial breakthroughs drive a surge

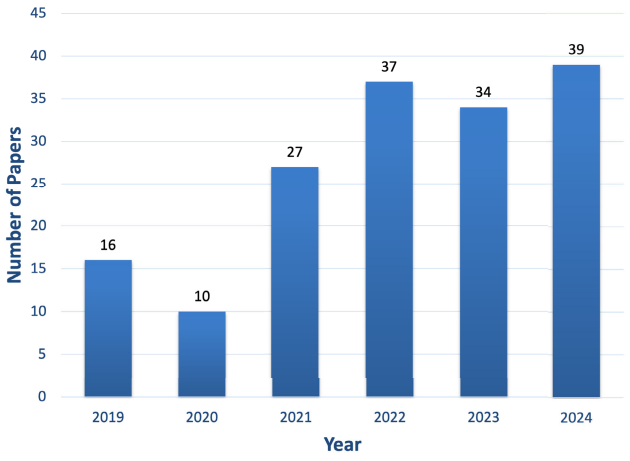


FIGURE 4. Distribution of papers by year of publication.

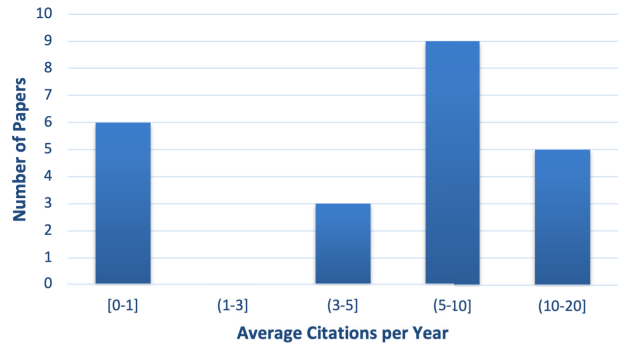


FIGURE 5. Distribution of selected papers by average citations per year.

in publications and citations, subsequently leading to a more steady pace of contributions.

IV. LITERATURE REVIEW ANALYSIS

This section presents an analysis of the major deep learning approaches to face MAD as identified in the reviewed literature. These approaches include Convolutional Neural Networks (CNNs), Siamese Networks, Transformer-Based Models, and Hybrid Architectures. Each method offers unique capabilities and addresses different aspects of MAD, such as identity verification, artifact detection, and feature representation. The review focuses on evaluating their strengths, limitations, and contributions to the advancement of MAD techniques in various experimental settings. Through this analysis, this section aims to highlight the diversity of strategies and the potential for combining approaches to enhance robustness and generalization in addressing morphing attacks.

A. DATASETS USED

The papers reviewed in the following subsections refer to a variety of datasets, many of which are critical to experiments or serve as the basis for creating new datasets. Table 1 summarizes the major datasets mentioned in the literature, highlighting those used for experimental evaluation or dataset creation. Note that not all datasets are publicly available; in some cases, it may be necessary to contact the authors of the dataset to gain access. Figure 7 presents a Venn diagram illustrating the distribution of datasets by their typologies (Public, Restricted, and Created) across the reviewed literature. It highlights the number of papers using each dataset type and visualizes the overlap between them.

Regarding dataset availability, the datasets can be categorized into several types:

- **Publicly Available**: Datasets such as the AMSL Face Morph [47], [48], [49], VGGFace2 [50], and FRLL-Morphs [34], [47], [51] are made accessible to the public, often under licenses like Creative Commons. These datasets are generally free to use, although they may have conditions such as requiring citation or restricting commercial use.
- **Requires Authorization**: Some datasets, including MorphDB [52], MorGAN [53], and the LMA Dataset [53], are only accessible after receiving explicit authorization from the dataset creators or associated organizations. Access to these datasets typically involves submitting a formal request and providing a justification for their use, along with possible non-disclosure agreements.
- **Sequestered Evaluation Platform**: Datasets such as the SOTAMD Dataset [36] are accessible exclusively through a controlled evaluation platform, where researchers submit their models for assessment, ensuring the data remains secure and is used within specified parameters.
- **Suspended**: Certain datasets, such as the FRGC-Morphs [51], have been withdrawn from distribution due to concerns regarding their use or changes in distribution policies.

In response to these access constraints and growing privacy concerns, recent research has increasingly relied on synthetic datasets. These datasets are generated to resemble real-world biometric data while preserving privacy, and have become an essential resource for both the creation and evaluation of MAD systems. Notable examples include the SMDD [54], MONOT [55], and AMONOT [56] datasets. These synthetic datasets offer high-quality, privacy-compliant biometric images and have been instrumental in advancing MAD techniques. By enabling experimentation without compromising real individuals' privacy, they present a valuable alternative in contexts where real data is restricted or unavailable [54], [55], [56].

B. PRE-TRAINED MODELS USED

The pre-trained models discussed in the reviewed literature are commonly used for feature extraction, transfer learning, and fine-tuning in MAD tasks. Table 2 provides an overview of the major pre-trained models referenced, summarizing

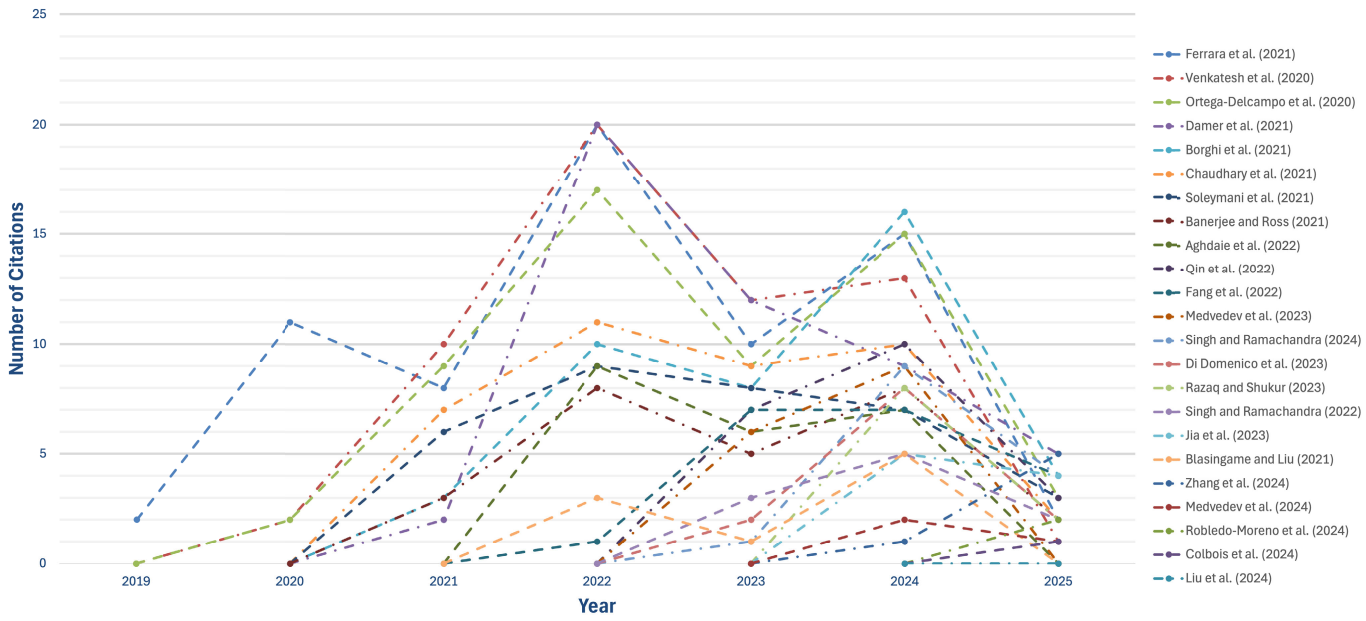


FIGURE 6. Selected papers citation count per year, ordered by total number of citations.

FIGURE 7. Number of papers in the literature review using each dataset type (Public, Restricted, and Created) and their overlap.

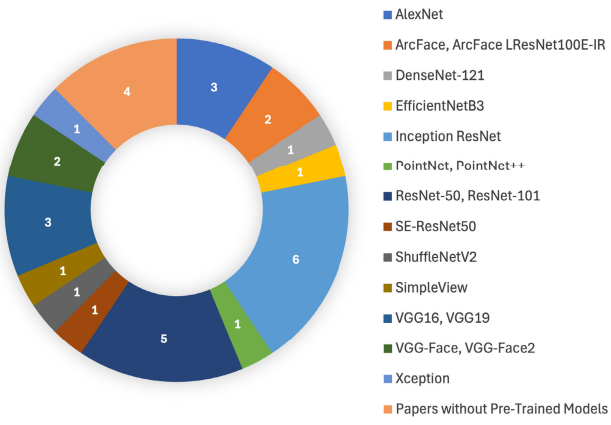


FIGURE 8. Distribution of pre-trained models used in the reviewed literature, highlighting the number of papers using each model.

their characteristics and applications. These models include widely used architectures such as AlexNet [81] and ResNet [82], as well as more specialized frameworks such as ArcFace [83] and PointNet [84]. Figure 8 provides a visualization of the pre-trained models used in the literature review, illustrating the distribution of papers using each pre-trained model.

C. ARCHITECTURE

The architecture of MAD models is critical to their effectiveness and adaptability. This subsection examines key approaches, including CNNs, Autoencoders, GANs, Siamese Networks, Transformer-Based Models, and Hybrid Architectures. Each architecture offers unique strategies for dealing with morphing attacks, from pairwise comparisons to the use of different features and representations. Figure 9 illustrates the distribution of papers in the literature review by architecture type, showing the prevalence of CNN-based models, followed by hybrid models, siamese networks, autoencoders, GAN-based models, and transformer-based models. The following subsections explore these innovations and their contributions to the advancement of MAD.

1) CONVOLUTIONAL NEURAL NETWORKS

CNNs are a type of deep learning model that is widely used in computer vision for its ability to learn features from images. A CNN typically consists of convolutional layers that apply filters to input images to detect features such as edges, textures, and patterns. These features are then passed

TABLE 1. Major datasets referenced in the literature reviewed, focusing on those used for experimental evaluations or as a basis for creating new datasets.

Dataset	Number of Images	Mode of Obtention	Availability	Summary	Reviewed Literature Using the Dataset
AMSL Face Morph Dataset [47]–[49]	~2,000	Morphed images using advanced techniques	Publicly available under the Creative Commons Attribution-ShareAlike 3.0 License	High-quality dataset of morphed images, designed to evaluate vulnerabilities in biometric systems	[14], [27], [44], [57]
Aberdeen Dataset [58]	~700	Controlled photography sessions	Publicly available	Controlled dataset capturing variations in lighting and pose	[15]
ChiMo Dataset [59], [60]	~24,000	Morphed images from the Chicago Faces Database using various morphing algorithms	Publicly available	Dataset of morphed faces designed to evaluate the vulnerability of facial recognition systems to morphing attacks, with raw and compressed versions	[61]
FEI Dataset [62]	~2800	Photographs under controlled conditions	Publicly available for research purposes only	Simple dataset of frontal and profile face images, designed for basic face recognition tasks	[15], [17], [18], [45]
FEI-Morph Dataset [17], [63]	~6000	Morphed images created from the FEI Dataset using various morphing techniques	Publicly available	Dataset of morphed faces designed to test facial recognition systems against morphing attacks	[17], [61]
FERET Dataset [64]	~14,000	Controlled and semi-controlled photography	Publicly available	Benchmark dataset for face recognition research, featuring diverse variations in facial characteristics	[65], [66]
FRGC Dataset [67]	~50,000	Controlled 2D and 3D facial images	Publicly available	Comprehensive dataset including 2D and 3D images, designed for advanced facial recognition studies	[16], [46]
FRGC-Morphs Dataset [51]	~6,000	Morphed using Generative Adversarial Networks (GANs) and traditional methods	Distribution has been suspended	Large-scale dataset of morphed images, created to test MAD system robustness against morphing techniques	[14], [65]
FRLL Dataset [68]	~1,020	Photographs under controlled conditions	Publicly available	High-resolution 2D facial images with variations in pose, age, gender, and ethnicity	[15], [46]
FRLL-Morphs Dataset [34], [47], [51]	~1,000	Morphed using GANs and traditional methods	Publicly available	Dataset of high-quality morphed images, used to evaluate morphing attack vulnerabilities	[14], [15], [65], [66]
Facescape Dataset [69]–[71]	~17,000	3D scans of faces	Publicly available	Large-scale dataset of 3D facial scans, designed to capture diverse expressions and identities	[20]
Flickr-Faces-HQ (FFHQ) Dataset [72]	~70,000	Collected from Flickr	Publicly available	High-resolution dataset emphasizing demographic diversity, often used for GAN training	[46], [66]
HNU Dataset [73]	~1,000	Photographs and morphed images using diverse techniques	Requires direct authorization from the authors for research use	Diverse dataset of photographs and morphed images, emphasizing demographic diversity for morphing research	[18]
Idiap Morph Dataset (AMSL + FRLL-Morphs Datasets) [34], [51]	~3,000	Generated morphs	Publicly available	Dataset of morphed images, created to test the robustness of detection systems	[17], [45], [61]
LMA Dataset [53]	~500	Light morphing techniques	Requires approval for research use	Dataset focusing on subtle morphing effects, designed to study vulnerabilities in facial recognition systems	[44], [74]
LMA-DRD Dataset [43]	~54,000	Digital and printed morphs using diverse techniques	Requires approval for research use	Dataset of digital and scanned morphed images, designed to test detection systems in realistic scenarios	[43], [65]
MorGAN Dataset [53]	~3,000	GAN-generated morphs	Requires approval for research use	Synthetic dataset of GAN-generated morphs, exploring vulnerabilities in automated systems	[44], [57], [65], [74], [75]
MorphDB Dataset [52]	~2,000	Morphed and real images	Requires direct authorization from the authors for research use	Comprehensive database of real and morphed images, designed to analyze morphing detection methods	[17], [27], [29], [45]
ND-IIITD Dataset [76], [77]	~5,000	Controlled and natural images	Requires direct authorization from researchers and organizations through a formal request	Collaborative dataset combining natural and controlled face images, used for morphing research	[78]
NIST FRVT Morph Dataset [79]	~5,000	Controlled morphed images	Requires direct authorization from researchers and organizations through a formal request	Benchmark dataset by NIST, created to evaluate face recognition systems against morphing attacks	[29]
Progressive Morphing Database (PMDB) [52]	~1,000	Progressive morphing techniques	Requires direct authorization from the authors for research use	Dataset featuring gradual morphing effects, designed to study detection sensitivity	[17], [27], [29], [45], [61]
SCFace Dataset [80]	~4,000	Surveillance-like imagery	Publicly available	Low-resolution dataset of surveillance-like imagery, often used for challenging conditions in research	[28]
SOTAMD Dataset [36]	~5,000	Morphed and real images	Restricted through a sequestered evaluation platform	Large dataset of morphed and real images, designed for evaluating detection methods across techniques	[29]
Utrecht ECVP Dataset [58]	~130	Controlled photography	Publicly available	Controlled dataset focused on facial expressions and poses, used in biometric and psychological studies	[15]
VGGFace2 Dataset [50]	~3.3 million	Crawled from the web	Publicly available	Large-scale dataset with diverse face images, widely used for training deep learning models	[14], [15]
VISAPP Dataset [33]	~1,000	Morphed images	Publicly available	Dataset of morphed images, designed for evaluating vulnerability to morphing attacks	[44], [74], [75]

TABLE 2. Summary of the major pre-trained models referenced in the literature reviewed, including details on their architecture, key features, and relevance to MAD.

Pre-trained Model	Summary	Literature Reviewed Using the Dataset
AlexNet [81]	AlexNet is a landmark deep learning model introduced in 2012, which significantly advanced image classification by leveraging CNNs. It used innovative techniques like ReLU activation, dropout regularization, and GPU acceleration to achieve a top-5 error rate of 15.3% on ImageNet [85], far surpassing previous methods. AlexNet's success revolutionized computer vision and popularized deep learning for large-scale image tasks.	[28], [29], [41]
ArcFace, ArcFace LResNet100E-IR [83]	ArcFace is a deep learning framework for face recognition introduced in 2019, that uses additive angular margin loss to improve the discriminative power of face embeddings. ArcFace LResNet100E-IR, a variant of ArcFace, is built upon a ResNet architecture and is fine-tuned using the ArcFace framework.	**ArcFace:** [61] **ArcFace LResNet100E-IR:** [78]
DenseNet-121 [86]	DenseNet-121 is a deep learning model introduced in 2017 as part of the DenseNet family, which focuses on efficient feature propagation by densely connecting each layer to every other layer. This architecture reduces the number of parameters while improving gradient flow and feature reuse, making it computationally efficient. DenseNet-121, with 121 layers, has been widely used for image recognition and transfer learning tasks due to its compactness and strong performance across diverse datasets.	[43]
EfficientNetB3 [87]	EfficientNetB3 is part of the EfficientNet family introduced in 2019, which employs a compound scaling method to balance network width, depth, and resolution for optimal performance. EfficientNetB3 achieves state-of-the-art accuracy on image recognition tasks while being computationally efficient, making it a popular choice for resource-constrained environments. Its design emphasizes achieving high accuracy with fewer parameters compared to traditional architectures.	[15]
Inception-ResNet [88], [89]	Inception-ResNet combines the strengths of the Inception architecture with residual connections to improve training stability and efficiency. First introduced in 2015, it merges the feature extraction capabilities of Inception modules with the optimization benefits of residual learning. The model helps mitigate the vanishing gradient problem and allows for deeper and more accurate networks. Inception-ResNet v1, an early version, incorporated residual connections to the Inception modules.	**Inception-ResNet (unspecified version):** [61] **Inception-ResNet v1:** [17], [44], [45], [74], [75]
PointNet [84], PointNet++ [90]	PointNet is a deep learning framework introduced in 2017 for directly processing unordered 3D point clouds, enabling tasks like classification and segmentation without requiring voxelization. PointNet++ builds on this by incorporating a hierarchical structure to capture local geometric features, improving performance on complex 3D data with varying densities and scales. These models are foundational for 3D data processing.	[20]
ResNet-50, ResNet-101 [82]	ResNet-50 and ResNet-101, introduced in 2016 as part of the ResNet family, are deep CNNs designed to address the vanishing gradient problem through residual connections. ResNet-50, with 50 layers, offers a balance of depth and efficiency, making it a standard for image recognition and transfer learning. ResNet-101, a deeper variant with 101 layers, provides improved feature representation and accuracy, particularly for high-resolution and detailed image analysis. Both models maintain computational efficiency and are widely used for diverse computer vision tasks.	**ResNet-50:** [14], [17], [27], [28], [46] **ResNet-101:** [28]
SE-ResNet50 [91]	SE-ResNet50 is an enhanced version of ResNet50 introduced as part of the Squeeze-and-Excitation Networks (SENets) in 2018. It incorporates Squeeze-and-Excitation (SE) blocks, which adaptively recalibrate channel-wise feature responses by modeling interdependencies between feature channels. This design improves the network's representational power while maintaining computational efficiency, making SE-ResNet50 particularly effective for tasks requiring nuanced feature extraction.	[27]
ShuffleNetV2 [92]	ShuffleNetV2, introduced in 2018, is a lightweight CNN designed for mobile and embedded applications. It improves efficiency and accuracy by optimizing operations such as channel splitting, pointwise group convolutions, and channel shuffling. ShuffleNetV2 achieves a better balance of speed and performance compared to its predecessors, making it ideal for resource-constrained environments.	[18]
SimpleView [93]	SimpleView, introduced in 2021, is a lightweight framework for processing 3D point clouds by projecting them into multiple 2D views. This approach leverages standard 2D CNNs to extract features, simplifying point cloud classification and achieving competitive performance with reduced computational complexity. It is particularly effective for tasks requiring efficient and scalable 3D data processing.	[20]
VGG16, VGG19 [94]	VGG16 and VGG19 are deep CNN architectures introduced in 2015, known for their simplicity and depth, using small 3x3 convolutional filters in a sequential design. With 16 and 19 layers respectively, they demonstrated that increased network depth improves performance, achieving state-of-the-art results on ImageNet [85]. These models have become widely used for transfer learning and influenced the design of subsequent deep learning architectures.	[28], [29], [66]
VGG-Face [95], VGG-Face2 [50]	VGG-Face and VGG-Face2 are large-scale datasets designed for face recognition tasks, emphasizing diversity across poses, expressions, and demographics. These datasets are commonly used to train deep learning models, such as VGG16 or ResNet architectures, resulting in widely adopted pre-trained models often referred to as "VGG-Face models".	[29], [42]
Xception [96]	Xception, introduced in 2017, is an extension of the Inception architecture that uses depthwise separable convolutions to improve efficiency and performance. By replacing standard convolutions with separable ones, Xception reduces computational complexity while maintaining high accuracy, making it well-suited for tasks requiring efficient deep learning models.	[28]

through pooling layers, which reduce spatial dimensions while preserving essential information, and fully connected layers for classification or regression tasks [97], [98]. The structure and flow of a basic CNN are illustrated in Figure 10, which show how convolutional, pooling, and fully connected layers work together to process and classify input data. This ability to automatically extract features from raw image data makes CNNs effective for MAD, where detection of subtle variations is essential. Several studies have adapted CNN architectures to improve their robustness and generalization in MAD.

Venkatesh et al. [41] propose a method to detect morphed face attacks using residual noise extracted with a deep Multi-scale Context Aggregation Network (MS-CAN). The approach involves the aggregation of multiple denoising techniques – Wavelet Denoising (WD), Block Matching and 3D filtering (BM3D), Multiresolution Bilateral Filtering (MBF), and Denoising Convolutional Neural Networks (DnCNN) – to create a fused denoised image. The residual noise from the aggregated denoising process is then analyzed using a pre-trained AlexNet [81] to extract discriminative texture features, which are classified using a Probabilistic Collaborative Representation Classifier (P-CRC). In addition to using CNN-based feature extraction, the authors also incorporate texture analysis by employing advanced descriptors such as BSIF, Histogram of Oriented Gradients (HOG), and LBP.

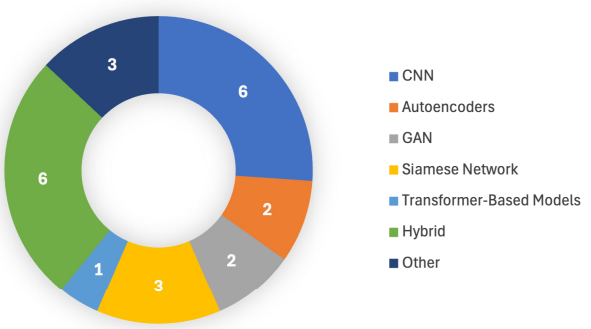


FIGURE 9. Distribution of papers in the reviewed literature by architecture type.

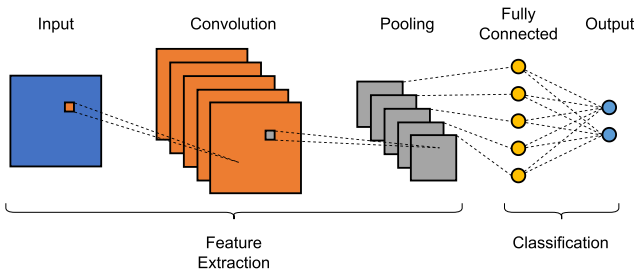


FIGURE 10. CNN architecture, inspired by [98].

These texture descriptors effectively capture discriminative local features, enhancing the robustness of MAD when combined with residual noise features. The evaluation of the method was conducted on three distinct morphed face datasets, each constructed using unique morphing techniques and compositions. The datasets included both bona fide and morphed images and were divided into training and testing partitions to ensure robust validation. The evaluation followed ISO-IEC 30107-3 standards, using metrics such as D-EER, BPCER, and APCER. Performance was evaluated in three experimental setups: a) training and testing on the same dataset; b) merging all datasets to evaluate scalability; c) cross-dataset evaluation to test generalizability. This comprehensive evaluation strategy assessed the adaptability of the proposed MS-CAN framework across different morphing scenarios and varying image quality conditions.

Damer et al. [43] present a pixel-wise supervision approach for S-MAD, called Pixel-Wise Morphing Attack Detection (PW-MAD). Unlike traditional methods that provide a single binary classification label for an entire image, PW-MAD employs pixel-wise labels to guide the network during training, thereby focusing on more subtle morphing artifacts that may be distributed throughout the image. The proposed method uses a modified DenseNet-121 [86] architecture, adapted to include two dense blocks, two transition blocks, and a convolutional layer that outputs a feature map for pixel-wise supervision. The evaluation of the PW-MAD framework was performed using the LMA-DRD [43] dataset. Experiments were performed under different training and testing configurations, including: a) intra-dataset settings (training and testing on the same data type); b) cross-dataset settings (training on digital images and testing on re-digitized images, and vice versa). Metrics such as the APCER and the BPCER, defined by ISO/IEC 30107-3 standards, were used to evaluate performance. Receiver Operating Characteristic (ROC) curves were also plotted to visualize the trade-off between APCER and BPCER at different operating points. This evaluation strategy aimed to measure the ability of the PW-MAD method to generalize across different types of morphing attacks and image acquisition scenarios.

Ferrara et al. [29] present an approach for face MAD that focuses on the challenges posed by print-scanned (P&S) images and heterogeneous image sources. Their method uses CNNs, specifically pre-trained models such as AlexNet [81], VGG19 [94], VGG-Face [95], and VGG-Face2 [50], which are fine-tuned to adapt to the morphing detection task. To address the difficulties of training on real P&S images, they simulate the P&S process during data augmentation, which involves modeling the image degradation typically caused by the printing and scanning stages. The evaluation of the proposed method was conducted using multiple datasets, including MorphDBD [52], MorphDBP&S [52], and the SOTAMD [36] benchmark. These datasets represent different morphing scenarios, from digitally morphed images to those processed through printing and scanning pipelines. The evaluation metrics included accuracy, EER, and error trade-offs between APCER and BPCER, according to ISO/IEC 30107-3 standards. The method was tested under both intra-dataset and cross-dataset conditions, emphasizing generalizability and robustness. In addition, simulations of the P&S process were validated by comparing real and synthetic P&S images to ensure the reliability of the augmentation strategy. These extensive experiments provided insight into the performance of the method under different acquisition and morphing scenarios.

Qin et al. [78] introduce a face MAD and localization approach based on feature-wise supervision, designed for both S-MAD and D-MAD. Unlike traditional score-level losses, their method employs a CNN with feature-wise loss to better capture the morphing patterns present in different face morphing attacks. For S-MAD, the approach uses a fine-grained classification loss that allows both the detection of morphing attacks and the precise localization of morphed areas within an image. For D-MAD, the method constructs differential loss functions based on similarity and distance, which improves generalization to unseen morphing attacks and increases robustness to low-resolution and non-frontal probe images. The evaluation of this approach was conducted using a combination of self-collected and publicly available datasets, with images processed to simulate morphing techniques such as complete, splicing, retouching, and combined morphing. Several protocols were designed to test different aspects of the system, including: a) overall detection performance to intra-technique (same morphing technique) morphing types; b) generalization to unseen morphing types; c) cross-dataset performance; d) robustness

to varying probe image qualities; e) robustness to non-frontal probe face images. Metrics such as EER, BPCER, and APCER were used following ISO/IEC 30107-3 standards. Furthermore, visualization techniques were used to validate the localization of morphed areas, providing interpretable results for potential human verification.

Medvedev et al. [14] present MorDeephy, a MAD method that employs a fused classification approach using two CNNs trained in parallel. The proposed method is designed for S-MAD and aims to discriminate between bona fide and morphed images by learning deep facial features that encode authenticity information. The architecture uses two ResNet-50 [82] networks, each of which is trained with different identity labels for the morphed images. During training, both networks learn high-level identity-specific features, and their outputs are combined through a dot product, followed by a sigmoid activation for morph detection. The authors also employ a unique dataset filtering strategy using VGGFace2 [50], generating selfmorphs to reduce the bias towards detecting superficial morph artifacts. The evaluation of MorDeephy was conducted using multiple publicly available benchmark protocols, such as FRGC-Morphs [51] and AMSL [47], [48], [49]. These benchmarks incorporate different morphing techniques, including landmark-based and GAN-generated morphs. Metrics such as APCER and BPCER were used to evaluate performance at different operational thresholds. DET curves were plotted to analyze the trade-offs between detection accuracy and error rates across different datasets. The evaluation covered intra-dataset and cross-dataset scenarios, testing MorDeephy's robustness to: a) unseen morphing attacks; b) varying image acquisition conditions, such as P&S images. In addition, the experiments included comparisons with state-of-the-art methods to assess generalizability and scalability.

Medvedev et al. [15] propose an extension of the MorDeephy approach [14] to D-MAD using a fused classification strategy. The method employs a dual CNN-based setup, similar to MorDeephy, where two distinct networks process input image pairs and perform identity classification independently. This extension adapts the previous no-reference setup to handle D-MAD scenarios. To improve detection performance, the authors introduce a data mining strategy leveraging GAN-based and landmark-based morphs and employ a class allocation method for morphed samples to improve feature discriminability. The evaluation was conducted on multiple benchmark datasets, including custom protocols designed for D-MAD scenarios. These protocols used a mix of bona fide and morphed image pairs, generated from datasets such as FRLL [68] and VGGFace2 [50], and included various morphing techniques, including StyleGAN and landmark-based approaches. Metrics such as APCER and BPCER, in accordance with ISO/IEC 30107-3 standards, were used for evaluation. DET curves were plotted to visualize performance trade-offs across different thresholds and morphing scenarios. Additionally, the study included comparative experiments with baseline binary classification methods and alternative fused strategies to assess generalization capabilities and performance across unseen datasets. This comprehensive evaluation was designed to examine the adaptability of the method to different morphing attack techniques and operating conditions.

The CNN-based approaches reviewed here illustrate different ways in which convolutional architectures can be effectively tailored for MAD. Venkatesh et al. [41] use a MS-CAN and residual noise analysis to improve robustness by aggregating denoising techniques, while Damer et al. [43] focus on pixel-wise supervision to detect subtle artifacts, using a modified DenseNet-121 architecture for PW-MAD. Ferrara et al. [29] adapt pre-trained CNNs to detect morphing in P&S images, addressing the unique challenges of image degradation in real-world conditions. Qin et al. [78] propose a feature-wise supervision approach for both S-MAD and D-MAD, emphasizing fine-grained classification and morphing localization. Medvedev et al. [14] and their extension Medvedev et al. [15] use parallel CNNs for both S-MAD and D-MAD, with an emphasis on cross-protocol generalizability by training on identity-specific features.

Together, these CNN-based models showcase a wide range of techniques, from adversarial learning to pixel-wise supervision and feature-wise loss, each offering a distinct approach to handling different aspects of MAD. These methods reflect the versatility and adaptability of CNN architectures in addressing different challenges in morphing attack scenarios.

2) AUTOENCODERS

Autoencoders, introduced by Hinton and Salakhutdinov [99], are a type of neural network designed for unsupervised learning, with a primary focus on dimensionality reduction. The network is composed of two main components: an encoder and a decoder. The encoder compresses the input data into a lower-dimensional latent representation, while the decoder reconstructs the original input from this compressed form. The primary objective during training is to minimize the reconstruction error, enabling the model to learn an efficient and compact representation that retains the most important features of the data [99]. As illustrated in Figure 11, the autoencoder architecture facilitates this process by using the encoder to map the input into a latent space and the decoder to reconstruct the input, with both parts working together to reduce the reconstruction error.

Ortega-Delcampo et al. [42] propose a de-morphing approach for MAD at Automated Border Control (ABC) systems, using CNNs. The method is based on the comparison of two images: the potentially morphed chip image from an electronic passport and an *in vivo* image captured at the ABC system. The de-morphing process, facilitated by an autoencoder-based CNN, attempts to unravel the morphing and isolate the original identity by reconstructing an image that highlights any morphing artifacts. The evaluation of the proposed system has been performed using the FRAV-ABC dataset, which includes images captured under real

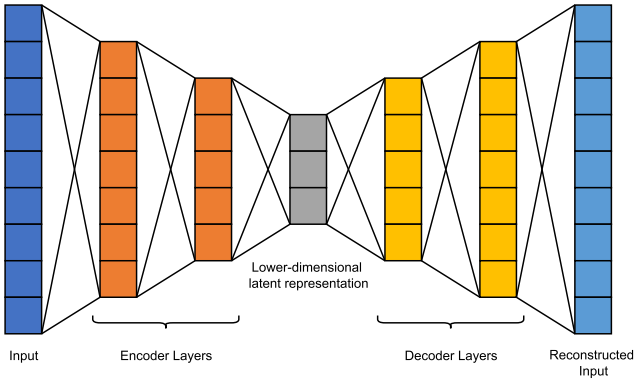


FIGURE 11. Autoencoder architecture, inspired by [100].

ABC conditions. The dataset consists of chip images, *in vivo* images, and P&S versions of the morphed images. The authors used various evaluation metrics, including the EER and DET curves, to assess the effectiveness of their approach. The evaluation was performed on both digital and P&S images to assess the system's ability to operate under different image quality conditions and real-world scenarios.

Fang et al. [65] propose an unsupervised MAD method called Self-Paced Anomaly Detection for Morphing Attack Detection (SPL-MAD). This method uses a Convolutional Autoencoder (CAE) to analyze reconstruction errors as an indicator of morphing artifacts. The CAE is trained on large-scale, general face recognition datasets that may contain unlabeled and unintentionally manipulated samples–referred to as "contaminated data". These contaminated samples, which are not intentionally labeled as morphed, may contain changes or manipulations that are not immediately apparent. The model learns to distinguish between bona fide and morphed images by adapting to these subtle differences in reconstruction. The Self-Paced Learning (SPL) paradigm is incorporated to iteratively filter out suspicious samples during training, thereby widening the reconstruction gap between bona fide and attack images. The evaluation of SPL-MAD was conducted on eight public MAD datasets, including both landmark- and GAN-based morphing techniques, to test its generalizability. Metrics such as APCER and BPCER were used, along with EER for overall performance comparison. The model's reconstruction error was analyzed using ROC curves to visualize performance across various operating points. Cross-dataset evaluations sought to assess the robustness of SPL-MAD against unknown attacks, while ablation studies aimed to demonstrate the contribution of the SPL paradigm in improving detection accuracy by increasing the reconstruction error gap between bona fide and attack samples.

The autoencoder-based approaches reviewed here illustrate different strategies to address the challenges of MAD. Ortega-Delcampo et al. [42] present a de-morphing approach that uses an autoencoder-based CNN to isolate and reveal hidden identities within morphed images. Fang et al. [65] present an unsupervised anomaly detection approach using a CAE and SPL, emphasizing the importance of adapting to unknown morphing attacks without the need for extensive labeled data.

Together, these autoencoder-based methods emphasize the effectiveness of this architecture in unsupervised learning for MAD.

3) GENERATIVE ADVERSARIAL NETWORKS

GANs are a type of deep learning model introduced by Goodfellow et al. [101]. GANs consist of two neural networks: a generator and a discriminator, which are trained together in an adversarial manner. The generator learns to create synthetic data, such as images, by trying to mimic the distribution of real data. The discriminator's role is to distinguish between real and fake data generated by the model. The generator and discriminator are in a continuous feedback loop, with the generator trying to fool the discriminator into classifying fake data as real, and the discriminator improving its ability to correctly identify real and generated data. This adversarial training process results in the generator producing increasingly realistic data over time. GANs have been widely used in image generation and enhancement tasks, demonstrating their ability to produce high-quality, realistic results [101]. The architecture of GANs, including the interactions between the generator and the discriminator, is shown in Figure 12.

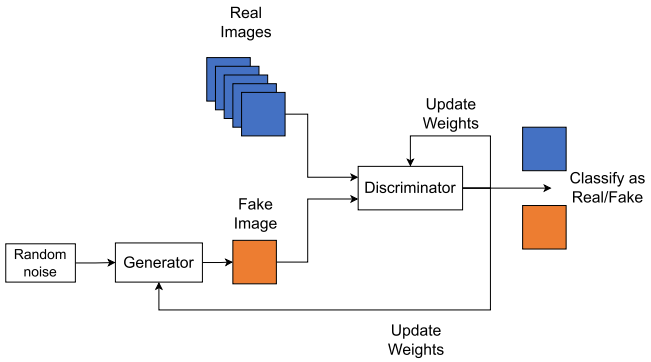


FIGURE 12. GAN architecture, inspired by [102].

Banerjee and Ross [57] propose a D-MAD method that uses a Conditional Generative Adversarial Network (cGAN) for identity disentanglement. The approach aims to determine whether a document image is morphed by learning to disentangle identities from the morphed image, conditioned on a trusted reference image. Using a cGAN, the model is trained to reconstruct the reference image from the morphed document image, effectively removing the contribution of the second identity present in the morph. The disentangled output is then compared with the reference using a biometric comparator to generate a similarity score, which is used for morph detection. Evaluation was performed on the AMSL [47], [48], [49], MorGAN [53], and EMorGAN [103] datasets, which include both landmark-based and GAN-based morphing techniques.

Intra-dataset evaluations measured the effectiveness of the method under controlled conditions, while cross-dataset tests assessed its robustness and generalizability to unseen morphing scenarios, including high-quality morphs, cross-domain variations, and hybrid morphing techniques. Metrics such as APCER and BPCER were used, consistent with ISO/IEC 30107-3 standards. Additionally, the system's ability to handle identity disentanglement in scenarios where document and reference image pairs are highly correlated was analyzed. These experiments evaluate the framework's strength in detecting morphing attacks without requiring explicit labels during training.

Blasingame and Liu [66] propose a S-MAD method using an adversarial learning approach. The architecture leverages a GAN to enhance the training process of the detector, which is designed to be more robust to various morphing attacks. The GAN architecture includes an encoder, a generator, and a critic, with an additional component called the "combinator" that combines latent codes to generate morphs. Latent codes are vector representations in a compressed latent space that encapsulate essential features of the original images, facilitating morph generation by mixing these features in the latent space. The critic, trained using adversarial techniques, is then repurposed as the morphing attack detector, with transfer learning used to fine-tune it for the S-MAD task. The evaluation was performed on two datasets containing both landmark-based and deep-learning-based morphs. Several experiments were conducted to assess the robustness and generalization of the proposed method. These experiments included: a) training and testing on the same dataset; b) evaluating the system across different morphing techniques and datasets. To ensure reliable evaluation metrics such as D-EER and DET curves were used as recommended by ISO/IEC 30107-3 standards.

The GAN-based methods discussed here highlight the flexibility and creativity of GANs for MAD, with each approach incorporating different strategies and improvements. Banerjee and Ross [57] use a cGAN for identity disentanglement, allowing the system to reconstruct and compare identities to detect morphing, while Blasingame and Liu [66] apply adversarial learning with GANs to improve the model's resistance to various morphing attacks.

Together, these GAN-based models demonstrate how adversarial learning and identity disentanglement can be effectively used to improve MAD performance.

4) SIAMESE NETWORKS

Siamese Networks are a popular architecture for D-MAD because they are well-suited for comparing pairs of images to find similarities or discrepancies. A Siamese Network consists of two identical neural network branches that process each input image separately but share the same weights and parameters. This design allows the network to learn comparable feature representations for each image [104]. By evaluating the distance or similarity between these features, the network can determine whether the images are genuine or manipulated. The architecture of a Siamese Network, as shown in Figure 13, highlights the dual-branch structure and the feature comparison process.

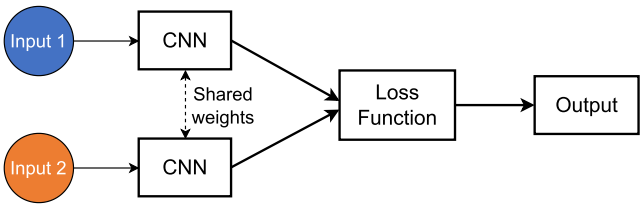


FIGURE 13. Siamese network architecture, inspired by [104].

Borghi et al. [27] present a framework for D-MAD using a double Siamese architecture. The proposed method consists of two parallel deep learning-based modules: the Identity block, which uses a ResNet50 [82] model pre-trained on the MS-Celeb-1M and VGGFace2 [50] datasets to extract identity-specific information from input face images, and the Artifact block, which employs a SE-ResNet50 [91] model to identify artifacts caused by morphing. The output features of both blocks are merged using a final fully connected layer, resulting in a prediction score that indicates whether the probe image is bona fide or morphed. The evaluation was conducted on three datasets–PMDB [52], MorphDB [52], and AMSL [47], [48], [49]–using a cross-dataset validation strategy. The PMDB [52] dataset was used for training, while MorphDB [52] and AMSL [47], [48], [49] were employed for testing to assess the framework's generalization capability. Performance was measured using metrics such as EER, BPCER, APCER, with DET curves plotted for detailed performance analysis. The evaluation included different test scenarios involving variations in morphing factors and types of probe images (criminal or accomplice). These scenarios provided insight into the robustness of the framework against different morphing attack conditions and its ability to effectively handle artifacts and identity inconsistencies.

Chaudhary et al. [44] present a D-MAD method using an undecimated 2D Discrete Wavelet Transform (DWT) and a deep Siamese network. The approach focuses on identifying morphing artifacts in the wavelet domain, where they are easier to detect than in the spatial image domain. The method involves decomposing images into 48 mid- and high-frequency wavelet sub-bands and applying Kullback-Liebler Divergence (KLD) to select the 22 most discriminative sub-bands for training. A Siamese network, using a pre-trained Inception ResNet v1 architecture [89] is then trained on these selected sub-bands, with contrastive loss applied to maximize the dissimilarity between morph and genuine images. In addition to using Siamese Networks, the authors used advanced texture descriptors, such as BSIF and LBP, to extract highly discriminative features for MAD, which were then combined with Support Vector Machines (SVMs) for classification. These texture

descriptors effectively capture subtle variations in local textures, which are indicative of morphing artifacts, and complement the deep learning-based approach to improve robustness. By combining these texture-based features with wavelet domain analysis, the method ensures the enhanced detection of localized and frequency-based morphing inconsistencies. The evaluation of the proposed framework includes two primary approaches: BW-22 and RGB-66. BW-22 works on grayscale images, where the image is converted to black-and-white, and wavelet decomposition is applied to the grayscale channels. A total of 22 highly discriminative sub-bands are selected from the mid- and high-frequency components of the wavelet decomposition. This approach emphasizes the texture and structural features that emerge in the wavelet domain, making it suitable for scenarios where color information is less critical. RGB-66, on the other hand, decomposes each of the three color channels (Red, Green and Blue - RGB) of the image separately, resulting in 144 wavelet sub-bands. After selecting the 22 most discriminative sub-bands from each channel, a total of 66 sub-bands are used for training. This method preserves the color information and provides a basis for evaluating whether color features contribute to morph detection. The evaluation of both methods was performed on three datasets–VISAPP [33], MorGAN [53], and LMA [53]–each containing different morphing techniques, such as splicing and GAN-based generation. The model was trained on a universal dataset, which combined the training sets from all three datasets, and then tested individually on each dataset as well as on the universal dataset. In addition, the method was tested on AMSL [47], [48], [49] to validate its ability to generalize to unseen morphing scenarios. Metrics such as APCER and BPCER were used, following ISO/IEC 30107-3 standards, with DET curves plotted to illustrate performance over different thresholds.

Soleymani et al. [75] propose a D-MAD approach using a Siamese network architecture with a pre-trained Inception ResNet v1 [89] as the base model. The model is fine-tuned to compare image pairs by contrastive loss, which emphasizes the similarity or dissimilarity between images in a latent embedding space. The evaluation was performed on two morphing datasets, VISAPP17 [33] and MorGAN [53], each containing images generated using different morphing techniques. The evaluation included three decision-making strategies: a) Euclidean distance between the embeddings; b) feature difference combined with a SVM; c) feature concatenation with an SVM. Metrics such as D-EER, APCER, and BPCER were used to evaluate the performance of the framework according to ISO/IEC 30107-3 standards. Additionally, cross-dataset validation was performed to evaluate the robustness and generalizability of the approach across different morphing techniques and datasets. The experiments also compared the proposed method against baseline techniques, including traditional handcrafted features such as BSIF and LBP, as well as other deep learning-based approaches.

The approaches described in [27], [44], and [75] demonstrate the versatility and innovation of Siamese networks for D-MAD, each employing different strategies and extensions. These different strategies highlight distinct contributions to improving MAD, with each approach excelling in different aspects. Borghi et al. [27] combine identity verification and artifact detection through a double Siamese framework with specialized modules–an Identity block and an Artifact block. This combination enables the model to simultaneously leverage the strengths of identity matching and artifact analysis, making it particularly robust in cases where high-quality morphed images contain minimal visible artifacts. On the other hand, Chaudhary et al. [44] adopt a unique preprocessing strategy by analyzing morphing artifacts in the wavelet domain, using 2D DWT to focus on specific frequency-based characteristics. This approach provides an advantage in identifying subtle morphing artifacts that may be less distinguishable in the spatial domain, thus increasing the robustness of the model to different morphing techniques. Finally, Soleymani et al. [75] take a different path by focusing on embedding analysis, using a simple Siamese architecture with a pre-trained Inception ResNet v1 [89] to produce high-quality latent representations that enable effective classification using multiple decision strategies.

Together, these approaches provide complementary perspectives on the challenges of MAD, emphasizing embedding quality, artifact-specific feature extraction, and multi-faceted analysis to advance the field.

5) TRANSFORMER-BASED MODELS

Transformer-based models for image processing, such as Vision Transformers (ViTs), are a type of deep learning architecture introduced by Dosovitskiy et al. [105]. These models are based on the Transformer architecture originally developed for natural language processing, but adapted for image classification and other vision tasks. Instead of using convolutional layers, ViTs divide an image into fixed-size patches, which are then linearly embedded into a sequence of tokens. These tokens are processed by the Transformer's self-attention mechanism, which allows the model to capture global relationships between image patches. Like traditional Transformers, ViTs consist of an encoder-decoder structure, where the encoder processes the sequence of image tokens, and the decoder generates the output, typically the class label for image classification tasks. The self-attention mechanism in the encoder helps the model focus on relevant parts of the image, while the decoder performs the final classification. ViTs have shown significant success in image processing tasks, outperforming traditional CNNs in various benchmarks, particularly when trained on large datasets [105]. The architecture of ViTs, including the partitioning of images into patches and the self-attention layers, is illustrated in Figure 14.

Zhang et al. [16] propose a generalized S-MAD algorithm leveraging deep representations from a ViT. Unlike

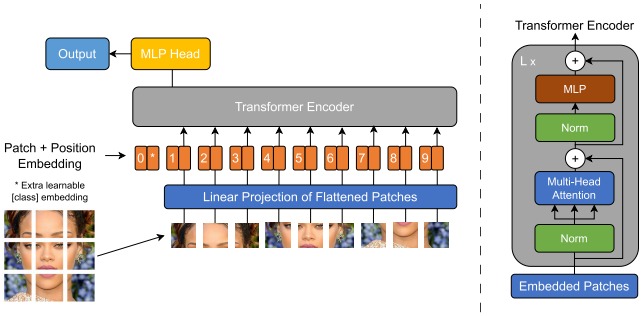


FIGURE 14. ViT architecture, inspired by [105].

traditional CNN-based methods, the ViT-based approach integrates local and global information using self-attention, which is well-suited for detecting subtle and distributed morphing artifacts. The ViT-L model, pre-trained on ImageNet [85], is used to extract features from face images, which are classified using a linear SVM for morph detection. The evaluation was conducted on datasets containing different morphing algorithms, including GAN-based and landmark-based methods, with different image types: digital, P&S, and print-scanned-compressed. The study used cross-dataset testing to evaluate the generalizability of the method to unseen scenarios, and intra-dataset testing to evaluate performance in known contexts. Metrics such as D-EER and MACER – equivalent to the ISO/IEC-defined APCER – at fixed BPCER were used. The method was benchmarked against state-of-the-art algorithms, including both feature-based and deep-learning-based approaches. Experiments were conducted to evaluate the system's ability to generalize to different morphing techniques and operational scenarios, especially in digital contexts. Additionally, statistical analysis, including mean and standard deviation of D-EER values, were performed to quantify variability and robustness across datasets and morphing techniques.

The model proposed by Zhang et al. [16] demonstrates the potential of transformer-based architectures for MAD. Using a ViT, the method integrates both local and global information through self-attention, which is well-suited for detecting subtle and distributed morphing artifacts. This approach explores how transformer models can improve the robustness and adaptability of MAD systems, especially in the context of digital images.

6) HYBRID ARCHITECTURES

Hybrid architectures for MAD combine multiple complementary methods to leverage the strengths of each, thereby increasing the overall robustness and effectiveness of detection models.

Aghdaie et al. [74] present a face MAD framework that aims to improve detection by leveraging structured group sparsity in conjunction with wavelet domain analysis. This method decomposes face images into 48 spatial-frequency wavelet sub-bands and isolates morphing artifacts in the high-frequency components. A Deep Neural Network (DNN) equipped with structured group sparsity selects the most discriminative wavelet sub-bands during training. By imposing a group sparsity constraint on the weights of the first convolutional layer, irrelevant sub-bands are effectively discarded, leaving only the most informative features for morph detection. The remaining sub-bands are retrained to fine-tune the network, optimizing its ability to distinguish bona fide from morphed images. These textural features, combined with the DNN's high-level spatial feature extraction, aim to provide a comprehensive representation of morphing inconsistencies. The evaluation of this framework was performed using three datasets: VISAPP17 [33], LMA [53], and MorGAN [53]. Two experimental setups were used: a) using individual datasets for training and testing; b) using a universal dataset that combines samples from all three datasets for training, and each of the aforementioned datasets for testing. Metrics such as D-EER, APCER, and BPCER were used in accordance with ISO/IEC 30107-3 standards. Additionally, DET curves were plotted to illustrate the model's performance at various operational thresholds. These evaluations were designed to assess the framework's ability to generalize across datasets and adapt to different morphing techniques and image types.

Di Domenico et al. [17] propose a combined approach for D-MAD that fuses identity features and artifact analysis, using both D-MAD and S-MAD features. The proposed system uses an Inception ResNet v1 [89] as the S-MAD module, fine-tuned on various morphing datasets, and a ResNet-50 [82] network as the D-MAD module using ArcFace loss for identity verification. The features extracted by both modules are fused, and a Multi-Layer Perceptron (MLP) classifier is used to produce a final morphing detection score. This work builds on the concept of combining identity and artifact features for morphing detection, a direction also explored by Borghi et al. [27], but extends it by explicitly fusing S-MAD and D-MAD modules to address the limitations of standalone approaches. The evaluation was conducted using several datasets, including PMDB [52], MorphDB [52], Idiap Morph [34], [51], and FEI [62], which cover a variety of morphing techniques and quality levels. They also developed the FEI-Morphs dataset [63], which is based on the FEI dataset [62]. The dataset contains morphed images generated using three morphing algorithms (FaceFusion, UTW, NTNU) and two morphing factors (0.3 and 0.5). Three experimental scenarios were used: a) assessing performance with the criminal subject; b) with the accomplice subject; c) across both. Metrics such as EER, BPCER, and APCER were used in accordance with ISO/IEC 30107-3 standards. Feature fusion strategies, including Min-Max normalization, were explored to align the contributions of the S-MAD and D-MAD features. Additionally, the cosine similarity of embeddings from the D-MAD module was incorporated to further refine the classification. The purpose of these evaluations was to assess the framework's ability to handle challenging morphing scenarios, especially when the accomplice's image was involved in the comparison.

Jia et al. [18] propose a MAD method based on a hybrid architecture that combines high-frequency feature extraction with a progressive enhancement learning framework. This hybrid approach merges texture-based and deep learning methodologies to address the challenges of detecting subtle morphing artifacts. High-frequency features are first extracted from the three color channels (RGB) of an image using a Discrete Fourier Transform (DFT) and high-pass filtering. These features, which capture fine-grained texture variations indicative of morphing artifacts, are subsequently fused with standard RGB image data through a two-stream neural network architecture. The progressive enhancement learning framework consists of two modules: a self-enhancement module (SEM) and an interactive-enhancement module (IEM). The SEM enhances the individual features of each stream, while the IEM facilitates cross-stream interaction between RGB and high-frequency data, progressively refining feature representations. This combination allows the model to effectively capture both global and local morphing traces. The evaluation of the proposed method was conducted using the FEI [62] and HNU [73] datasets, which contain different morphing scenarios with varying pixel fusion and position fusion factors. The experiments included a single dataset and cross-dataset tests to assess the generalization ability of the model. Standard metrics such as EER, APCER, BPCER, and ACER were used in accordance with ISO/IEC 30107-3 standards. Ablation studies were also performed to analyze the contributions of the SEM and IEM, in order to investigate their role in improving feature representation and interaction. The evaluations aim to determine the robustness of the framework in capturing subtle morphing traces and its adaptability to different dataset conditions.

Razaq and Shukur [19] propose a face MAD framework that combines wavelet scatter transform with CNN features. This dual feature extraction approach provides a comprehensive analysis of morphing artifacts by capturing both subtle textural variations and complex spatial patterns. In the feature extraction phase, the wavelet scatter transform is used to capture frequency-domain features, while a CNN extracts high-level spatial features. These features are fused using Canonical Correlation Analysis (CCA) to provide an enhanced feature set, which is then classified using multiple classifiers, including a DNN and SVM. The evaluation was conducted on a custom dataset created using a combination of publicly available and self-generated morph images, as well as the AMSL dataset [47], [48], [49]. These datasets included different morphing techniques, such as automatic and manual landmark selection, and StyleGAN-based morphing. The model was evaluated using metrics such as accuracy, FAR, and FRR, with a focus on minimizing errors in both bona fide and morphing classifications. These evaluations aimed to measure the robustness of the combined wavelet-CNN approach, particularly in detecting morphing artifacts in high-resolution images and scenarios with minimal visible artifacts, and to evaluate its suitability for border control applications.

Colbois et al. [46] propose a hybrid approach for MAD that combines attack-agnostic feature extraction with supervised and one-class classification. The attack-agnostic feature extraction uses pre-trained vision models such as CLIP, DINOv2, and AIM to extract features from bona fide images. These attack-agnostic features, which are not specific to any particular attack, are used to detect morphing attacks across different techniques, including landmark-based, GAN-based, and diffusion-based morphs. By using pre-trained models trained exclusively on bona fide data, the system can generalize across a variety of morphing methods. For classification, the authors employ two methods: a supervised binary SVM classifier trained on the extracted features and a one-class Gaussian Mixture Model (GMM) trained only on bona fide data to detect out-of-distribution morphs. The approach was evaluated on several types of morphing attacks and datasets, including FRGC [67], FFHQ [72], and FRLL [68], with a focus on generalization to unseen attacks and datasets, as well as performance on P&S data. The evaluation used D-EER to assess the system's ability to distinguish between real and morphed images in various scenarios.

Liu et al. [45] propose a D-MAD approach that combines triplet-based metric learning with artifact analysis. This hybrid approach integrates two complementary techniques: identity-based feature extraction and artifact detection. The identity-based feature extraction uses metric learning to separate morphed images from their contributing subjects in the feature space. This is achieved through a triplet loss function, which increases the distance between the morphed image and its source subjects, making it easier to distinguish between genuine and morphed images. In parallel, the artifact detection module focuses on identifying the latent artifacts introduced during the morphing process. These artifacts–such as inconsistencies in lighting, texture, or facial features–are captured and analyzed to further validate whether an image has been morphed. The outputs of the two modules are processed independently before being combined at the score level to make the final decision. The evaluation of this approach involved training the model on datasets such as PMDB [52], Idiap Morph [34], [51], and MorphDB [52], while FEI Morph [17], [63] was used for testing. The performance of the system was tested in three different experimental scenarios: a) using a trusted live image of the accomplice subject, b) using an image of a criminal subject, and c) a combined (''Both'') scenario involving both accomplice and criminal subjects. The evaluation used standard metrics such as BPCER and EER, in accordance with ISO/IEC 30107-3 standards.

The hybrid architectures discussed in [17], [18], [19], [45], [46], and [74] are distinguished by their unique approaches to combining different techniques for MAD. Aghdaie et al. [74] integrate structured group sparsity with wavelet domain analysis, which allows the method to focus on high-frequency

components in the wavelet domain, enabling a more targeted detection of morphing artifacts. Di Domenico et al. [17] adopt a fusion strategy combining S-MAD and D-MAD features, using different neural networks (Inception ResNet v1 [89] for S-MAD and ResNet-50 [82] for D-MAD) and incorporating ArcFace loss for identity verification, creating a more refined approach for morphing detection in difficult scenarios. Jia et al. [18] propose a two-stream architecture that combines high-frequency feature extraction with RGB image data through a progressive enhancement learning framework, allowing for both individual feature enhancement and cross-stream interaction. Razaq and Shukur [19] use a dual feature extraction strategy by combining wavelet scattering transform and CNNs, which allows them to capture both frequency-domain and spatial features, with the features fused using CCA for better representation. Colbois et al. [46] use attack-agnostic features extracted using pre-trained models such as CLIP, DINOv2, and AIM, which are then used for supervised and one-class detection to identify morphing attacks without being biased by specific attack types. Liu et al. [45] combine triplet-based metric learning with artifact analysis. They integrate identity-based feature extraction, which uses triplet loss to separate morphed images from their contributing subjects, with artifact detection to identify inconsistencies introduced by morphing.

Taken together, these hybrid models highlight the potential of combining different techniques and capabilities to advance MAD. Each study presents a distinct approach to integrating identity and artifact information, underscoring the value of hybrid strategies in addressing the complexities of MAD across diverse datasets and challenging scenarios.

7) OTHER APPROACHES

Other approaches include noise analysis, deep learning-based artifact analysis, shape-based detection methods, and additional deep learning strategies that do not fit neatly into the previously discussed categories.

Singh and Ramachandra [28] present a D-MAD method tailored for On-The-Fly (OTF) border control scenarios with varying image resolutions and capture distances. The proposed algorithm integrates Spherical Linear Interpolation (SLERP) with hierarchical fusion techniques to enhance the detection of morphing attacks. Features are extracted from six pre-trained CNN models, including AlexNet [81], VGG16 [94], VGG19 [94], Xception [96], ResNet50 [82], and ResNet101 [82]. These features undergo hierarchical fusion at both score and feature levels to effectively combine complementary information. The evaluation was performed using the SCFace-Morph dataset, derived from the SCFace [80] dataset, which simulates real-world border control conditions with variations in camera resolutions and capture distances. Separate evaluation protocols were developed for intra- and cross-medium scenarios, comparing digital and P&S images. Metrics such as EER, APCER, and BPCER were used, following the ISO/IEC 30107-3 standards.

Robledo-Moreno et al. [61] propose a MAD approach based on Federated Learning (FL), a privacy-preserving technique that enables the training of models across distributed locations without centralizing sensitive biometric data. This approach addresses privacy concerns associated with biometric data sharing, a key challenge in the field. The authors evaluate FL for both S-MAD and D-MAD tasks and compare its performance to traditional centralized training methods. For the S-MAD task, an Inception-ResNet [88], [89] model is trained on images generated by several morphing algorithms. In the D-MAD task, the authors use ArcFace [83] embeddings to compare trusted and suspect images for detecting morphing attacks. The evaluation process involves training the models across multiple locations in a federated environment and evaluating their ability to detect morphing attacks. The evaluation uses standard metrics for MAD, such as BPCER and EER, to measure the performance of the FL-based models in comparison to centralized models.

Singh and Ramachandra [20] propose three different 3D Morphing Attack Detection (3D-MAD) algorithms to detect 3D face morphing attacks. These algorithms are based on deep features extracted from pre-trained 3D point-based CNN models: PointNet [84], PointNet++ [90], and SimpleView [93]. PointNet [84] works directly on 3D point clouds, extracting features from the classification layer. PointNet++ [90] builds upon this approach by using hierarchical networks to capture more detailed geometric structures, while SimpleView [93] uses multi-view projections of point clouds, extracting a 40-dimensional feature vector from each depth map for detection analysis. These detection methods take advantage of depth information to better distinguish between bona fide and morphed face models. To evaluate the proposed 3D-MAD algorithms, the authors developed a dataset (3DMD) containing both bona fide and morphed 3D face models. This dataset includes 3D point clouds of 41 individuals, which were processed using various 3D morphing techniques to create morphed models. Additionally, the publicly available Facescape [69], [70], [71] dataset was used to extend the evaluation with a broader set of subjects and 3D face models. The evaluation used APCER, BPCER, and D-EER as performance metrics, following the ISO/IEC 30107-3 standards. The features extracted by PointNet [84], PointNet++ [90], and SimpleView [93] were classified using a linear SVM. DET curves were also plotted to analyze the trade-off between APCER and BPCER at different thresholds.

The approaches reviewed here highlight a variety of strategies for MAD, each offering a unique perspective. Singh and Ramachandra [28] integrate deep features from multiple CNNs using hierarchical fusion, which enables effective detection across different image resolutions and challenging OTF border control settings. Robledo-Moreno et al. [61] introduce a FL-based approach for MAD that addresses privacy concerns by enabling model training across distributed locations without centralizing sensitive biometric

data. Singh and Ramachandra [20] focus on 3D morph detection, providing insights into the use of depth information to improve detection robustness in both 2D and 3D scenarios.

Taken together, these studies illustrate the diversity of approaches that can be employed to address MAD, ranging from GAN-based identity reconstruction to unsupervised learning techniques, demonstrating the potential for innovative solutions beyond traditional supervised learning paradigms.

D. EXPERIMENTAL SETUP

This section presents the hardware configurations and training procedures adopted in different studies. Table 3 summarizes the key specifications, including GPU models, batch sizes, learning rates, training strategies, number of epochs, and training times, for the papers that reported them.

The training setups reported across the reviewed studies exhibit notable diversity in architecture types, hardware configurations, hyperparameters, and methodological strategies. The key findings are summarized below:

- **Hardware configurations** range from high-end setups to more modest consumer-grade GPUs:
  - High-performance GPUs were used by Blasingame and Liu [66] (4x NVIDIA V100), Aghdaie et al. [74] (2x TITAN X (Pascal)), and Chaudhary et al. [44] (3x TITAN X (Pascal)).
  - In contrast, Jia et al. [18] and Robledo-Moreno et al. [61] used consumer-grade GPUs (GTX 1060Ti and GTX 1070, respectively).
- **Architectures** varied across studies:
  - Blasingame and Liu [66] evaluated two models: a GAN, trained for 200 epochs, and a combinator, trained for 150 epochs.
  - Aghdaie et al. [74] and Jia et al. [18] used hybrid architectures.
  - Chaudhary et al. [44] implemented a siamese network, while Robledo-Moreno et al. [61] explored other approaches under a federated learning framework.
- **Learning rates** and their adaptation strategies were also varied:
  - Blasingame and Liu [66] applied 0.00001 for the GAN and 0.001 for the Combinator model.
  - Aghdaie et al. [74] began with 0.001, reducing it by a factor of 10 every 20 epochs.
  - Chaudhary et al. [44] and Jia et al. [18] both used a fixed learning rate of 0.0001.
  - Robledo-Moreno et al. [61] also used 0.001, with no further strategy specified.
- **Batch sizes** were reported as follows:
  - Blasingame and Liu [66] used a batch size of 256 across both models.
  - Chaudhary et al. [44] reported batch sizes of 64 for BW-22 and 32 for RGB-66.
  - Jia et al. [18] used a small batch size of 16.
  - Aghdaie et al. [74] and Robledo-Moreno et al. [61] did not report batch size information.
- **Training durations and epoch counts** varied significantly:
  - Blasingame and Liu [66] trained their GAN for 200 epochs (17 hours) and their Combinator model for 150 epochs (6 hours).
  - Aghdaie et al. [74] and Chaudhary et al. [44] both trained for 150 epochs, though only the latter included early stopping after 35 epochs without improvement.
  - Jia et al. [18] trained for only 20 epochs, likely constrained by hardware.
  - Robledo-Moreno et al. [61] did not specify training duration or number of epochs.
- **Training strategies** reflected the experimental goals and limitations of each study:
  - Aghdaie et al. [74] and Chaudhary et al. [44] incorporated learning rate scheduling based on validation performance.
  - Blasingame and Liu [66] performed full supervised training for both architectures.
  - Jia et al. [18] used standard training with a small batch size.
  - Robledo-Moreno et al. [61] applied a federated learning framework (FLARE), aggregating model updates after each epoch across decentralized sites.

In summary, the reviewed studies showcase diverse training setups, varying in hardware, architecture, learning rates, and strategies. These differences highlight the flexibility of approaches used to meet specific research goals and constraints, offering insights into optimizing future training configurations.

E. RESULTS

This section outlines the key findings reported in the reviewed literature on MAD. By focusing on the results, it provides a concise summary of the performance metrics and conclusions drawn from various studies. These findings will serve as a foundation for the discussion that follows, where trends, strengths, and limitations of MAD techniques are analyzed in more detail.

1) REPORTED RESULTS FROM REVIEWED LITERATURE

Table 4 presents the best performing results obtained in each reviewed paper for easier comparison. Note that results are not directly comparable, as they are obtained on different test sets. For results from standard benchmarks, we refer to Table 5.

The study by Aghdaie et al. [74] highlights the effectiveness of structured group sparsity in selecting discriminative features within the wavelet domain, facilitating robust morph detection in diverse datasets. Texture analysis plays a pivotal role in this approach, as the decomposition into wavelet sub-bands captures subtle, localized texture variations that

TABLE 3. Training setups and strategies.

Author	Architecture	Hardware	Batch Size	Learning Rate	Training Strategy	Epochs	Training Time
Aghdaie et al. [74]	Hybrid	2x 12 GB TITAN X (Pascal) GPUs	-	0.001[1]	Learning rate scheduling, no batch size or training time details	150	-
Blasingame and Liu [66]	GAN	4x NVIDIA V100 GPUs	256	0.00001	Full training with 200 epochs	200	17 hours
	Combinator			0.001	Network trained with 150 epochs	150	6 hours
Chaudhary et al. [44]	Siamese Network	3x 12 GB TITAN X (Pascal) GPUs	64[2]	0.0001	Learning rate reduction on validation loss plateau; early stopping after 35 epochs without improvement	150	-
			32[3]				
Jia et al. [18]	Hybrid	GeForce GTX 1060Ti	16	0.0001	Standard training with small batch size	20	-
Robledo-Moreno et al. [61]	Other Approaches	Nvidia GeForce GTX 1070 6GB (per site)	-	0.001	Federated learning using FLARE; model updates and aggregation each epoch	-	-

[1] Learning rate divided by 10 every 20 epochs.
[2] For BW-22.
[3] For RGB-66.

TABLE 4. Summary of the best results reported in each study included in the literature review. Only the best results were selected for this comparison. *Universal dataset* refers to a combination of all datasets used in the study. *Intra-dataset evaluation* indicates that both training and testing were performed on samples from the same dataset. *Intra-technique evaluation* indicates that training and testing were performed on samples generated using the same morphing technique, while *cross-technique evaluation* refers to training and testing on samples generated using different morphing techniques. *Values marked with an asterisk represent the average performance across all reported results.

Reviewed Literature	Experiment Description	BPCER (%) @ APCER			APCER (%) @ BPCER			D-EER (%)	Acc (%)
		1%	5%	10%	1%	5%	10%		
Aghdaie et al. [74]	Intra-dataset evaluation on VISAPP [33]	-	0.00	0.00	-	-	-	0.00	-
	Training on universal dataset and testing on VISAPP [33]	-	0.00	0.00	-	-	-	0.00	-
Banerjee and Ross [57]	Intra-dataset evaluation on AMSL [47]–[49]	-	-	6.1	-	-	5.2	-	-
	Training on AMSL [47]–[49] and testing on EMorGAN [103]	-	-	4.4	-	-	0.8	-	-
Blasingame and Liu [66]	Cross-technique and intra-dataset evaluation on FRLL [34], [47], [51]	0.00	-	0.00	-	-	-	0.47	-
	Intra-technique and cross-dataset evaluation; training on FERET [64] and testing on FRLL [34], [47], [51]	0.00	-	0.00	-	-	-	0.04	-
Borghi et al. [27]	Training on PMDB [52] and testing on MorphDB [52] criminal images	0.00	-	0.00	0.00	-	-	-	-
Chaudhary et al. [44]	Training on universal dataset (BW-22) and testing on VISAPP [33]	-	0.00	0.00	-	0.00	0.00	0.00	-
	Training on universal dataset (BW-22) and testing on AMSL [47]–[49]	-	-	-	-	33.78	23.61	16.4	-
Colbois et al. [46]	Cross-technique evaluation (different combinations)	-	-	-	-	-	-	0.00	-
	Cross-dataset evaluation (different combinations with training on FFHQ [72])	-	-	-	-	-	-	0.00	-
	Training on FRGC [67] digital images and testing on FRGC [67]-derived P&S images	-	-	-	-	-	-	3.99	-
	One-class detector trained on FRGC [67] and tested on diffusion morphs	-	-	-	-	-	-	0.00	-
Damer et al. [43]	Training and testing on digital images	-	16.12	6.45	-	-	-	-	-
Di Domenico et al. [17]	Training on universal dataset and testing on FEI [62] accomplice images with Min-Max feature fusion and cosine similarity	46.8	23.7	-	-	-	-	12.5	-
Fang et al. [65]	Training on CASIA-Web and testing on FRLL [34], [47], [51]	-	-	-	-	-	-	2.98	-
Ferrara et al. [29]	Fine-tuning of VGG-Face16 on $Digital_{Mc}$ and testing on MorphDB [52] digital images	0.0	0.0	0.0	-	-	-	0.9	97.4
	Fine-tuning of VGG-Face16 on $Digital_{Mc} + P\widetilde{\&}S_{MC}$ and testing on MorphDB [52] P&S images	43.8	6.9	2.3	-	-	-	6.1	93.5
Jia et al. [18]	Intra-dataset evaluation on FEI [62]	-	0.66	0.17	-	-	-	0.12	-
	Training on HNU (MDB1) [73] and testing on FEI [62]	-	3.03	1.47	-	-	-	3.26	-
Liu et al. [45]	Criminal scenario	11.5	2.0	-	-	-	-	3.7	-
Medvedev et al. [14]	Testing on FaceMorpher [107] (composed of simple morphs with foreground and background artifacts)	84.8	-	56.2	78.6	-	43.1	-	-
Medvedev et al. [15]	Testing of Fused Classification model on AMSL [47]–[49]	-	-	-	27.5	-	3.9	-	-
Ortega-Delcampo et al. [42]	Testing on digital images	-	-	-	-	-	-	0.78	98.7
Qin et al. [78]	D-MAD: Training on complete morphing images and testing on combined morphing images	-	0.00	-	-	-	-	0.48	-
	S-MAD: Training on combined morphing images and testing on splicing morphing images	-	0.00	-	-	-	-	0.63	-
Razaq and Shukur [19]	Intra-dataset evaluation of DNN on universal dataset	-	-	-	-	-	-	-	99.69
	Training DNN on universal dataset and testing on AMSL [47]–[49] (100 epochs)	-	-	-	-	-	-	-	95.8
Robledo-Moreno et al. [61]	Cross-technique evaluation of FL for S-MAD	0.00	0.00	-	-	-	-	0.00	-
	Cross-dataset evaluation of FL for D-MAD	34.3	9.7	-	-	-	-	7.1	-
Singh and Ramachandra [28]	Training and testing on digital images	-	-	-	-	-	-	0.00	-
Singh and Ramachandra [20]	Feature extraction using SimpleView [93]	-	2.5	0.00	-	-	-	1.59	-
Soleymani et al. [75]	Classification by feature difference combined with a SVM; training and testing on VISAPP [33]	-	6.12	3.71	-	5.29	3.17	5.2	-
Venkatesh et al. [41]	Intra-dataset evaluation on a derivative of FRGC [67]	-	1.11	1.11	-	-	-	2.63	-
	Intra-dataset evaluation on the universal dataset	-	5.01	3.05	-	-	-	4.96	-
	Training on a derivative of FRGC [67] and testing on a combination of a subset of FRGC [67] and a private dataset	-	4.66	2.88	-	-	-	4.66	-
Zhang et al. [16]	Cross-technique evaluation on morphs obtained from FRGC [67] digital images	-	-	-	-	-	-	13.63*	-

are characteristic of morphing artifacts. While the individual dataset training setup excels at detecting intra-dataset morphs, especially in VISAPP [33] and MorGAN [53], the universal dataset training also shows strong generalization to unseen morphing techniques in the aforementioned datasets, with slightly less satisfactory results on LMA [53].

The method proposed by Banerjee and Ross [57] achieves comparable results to state-of-the-art approaches despite not requiring explicit supervision or labeled training data to discriminate between morphed and non-morphed images. In particular, it outperforms existing methods in cross-dataset evaluations when trained on the AMSL [47], [48], [49] dataset and tested on the MorGAN [53] dataset, surpassing previous benchmarks. However, the method exhibits reduced performance when trained on the MorGAN [53] dataset and tested on AMSL [47], [48], [49] dataset, which may be due to the low correlation and overall biometric utility of the images in the MorGAN [53] dataset. The method also shows strong potential for decoding the second identity in morphed images.

The approach proposed by Blasingame and Liu [66] showed strong performance across different experimental setups, excelling in 35 out of 36 experiments. It showed remarkable resilience to changes in datasets and morphing algorithms, although performance decreased when both variables were changed simultaneously. The incorporation of the combinator into the training process improved the effectiveness of the method over alternatives without it. However, the contribution of the combinator was ultimately overshadowed by the advantages provided by the GAN framework and triplet loss, which played a more central role in the success of the method. This suggests that the performance improvements are less about the combinator itself and more about the unique training approach used. While the method showed reduced generalization to high-quality StyleGAN2 [106] morphs, this was not attributed to a failure of the model but rather to the inherent strength of these morphs, which limits detectable artifacts. Interestingly, when trained on StyleGAN2 [106] morphs, the method performed exceptionally well against weaker morphs such as FaceMorpher [107] and OpenCV [108], supporting the hypothesis that models trained on stronger morphs can generalize to weaker ones.

The framework proposed by Borghi et al. [27] achieved strong performance in D-MAD, particularly in the ''criminal'' scenario, where identity features were more discriminative. However, its effectiveness varied between datasets due to differences in the composition of morphs. In the MorphDB [52] dataset, morphed images closely resembled the accomplice, whereas in the AMSL [47], [48], [49] dataset, morphs exhibited a balanced mix of criminal and accomplice features. This highlights the adaptability of the framework to different morphing scenarios. Furthermore, the results emphasize that the performance of the framework is closely related to the quality of the training dataset. When artifacts are less prominent, such as in JPEG2000-compressed images, or when the framework encounters unseen image generation techniques, its performance declines, thus highlighting the need for diverse datasets that include different morphing tools to improve generalization. The study demonstrated the importance of combining identity- and artifact-related features for robust detection. While the individual modules of the framework, focusing on identity and artifacts respectively, performed worse in isolation, the integration through the final fully connected layer was critical. This layer balanced the contributions of the modules, allowing the framework to achieve high accuracy across datasets.

Chaudhary et al. [44] obtained a particularly strong performance of the RGB-66 on the MorGAN [53] dataset, surpassing that of BW-22. This highlights the role of color information in detecting GAN-based morphs, which inherently contain more color details than landmark-based morphs such as LMA [53] and VISAPP [33]. However, the wavelet analysis reveals that color information is less critical in this domain, as evidenced by the strong performance of BW-22, especially on challenging morphs from the LMA dataset [53]. The VISAPP dataset [33] showed consistent performance across all frameworks, likely due to its smaller size and uniform morphing technique, making it easier for models to learn. In contrast, LMA [53] morphs were harder to detect, consistent with previous findings that landmark manipulation morphs pose greater detection challenges than GAN-generated morphs. The framework also performed well on an unseen dataset, AMSL [47], [48], [49], demonstrating the robustness of wavelet decomposition with selective sub-bands in detecting morphs across different techniques.

In the study by Colbois et al. [46], the baseline experiment showed that all methods achieved near-perfect results on the constrained FRGC [67] dataset, although their effectiveness declined when applied to the more diverse FFHQ [72] dataset. Generalization to unseen attacks confirmed the adaptability of the model, while detectors trained on the FFHQ [72] dataset generally outperformed those trained on FRGC [67] when tested on different source datasets. However, the P&S scenario proved challenging, as artifacts were degraded during the printing and scanning process, which affected the detection performance. One-class detection was highly effective on the FRGC [67] dataset but underperformed on FFHQ [72] data. Among the features tested, DNADet excelled at one-class digital modeling, but struggled with P&S generalization due to its pretraining. AIM features performed well for unseen attacks but were inconsistent in other scenarios, limiting their broader applicability. DINOv2 excelled at P&S generalization, making it suitable for scenarios where training on P&S data is impractical. CLIP features, while rarely the best, performed consistently well across all scenarios, making them very versatile for complex, multi-challenge environments.

Damer et al. [43] demonstrate that the proposed PW-MAD framework outperforms baseline methods in both intra-dataset and cross-dataset evaluations, particularly in scenarios involving re-digitized morphing attacks. The pixel-wise supervision approach allowed the framework to focus on localized manipulation effects, improving generalizability to unseen variations, including re-digitized morphs. In the most challenging settings, such as training on digital morphs and testing on re-digitized ones, PW-MAD achieved lower error rates compared to existing methods, demonstrating its robustness.

The results obtained by Di Domenico et al. [17] show that while D-MAD generally outperforms S-MAD when the live image contains the criminal, its performance deteriorates when the accomplice is present due to greater identity similarity. The proposed fusion method effectively addresses this limitation by integrating S-MAD's artifact-based features and D-MAD's identity-based features. However, combining these features proved to be non-trivial. The study showed that features from S-MAD and D-MAD occupy different regions in the feature space, and normalization using the Min-Max strategy improved performance. Incorporating cosine similarity as an additional feature further improved detection, provided it was kept in its original domain.

Experimental results indicate that the proposed method outperformed competitors in scenarios involving the accomplice, highlighting the value of S-MAD features in improving D-MAD performance. Conversely, performance with the criminal remains an area for improvement, suggesting the need for tailored fusion methods and separate handling of criminal and accomplice scenarios.

Fang et al. [65] conclude that the inclusion of the SPL paradigm improves the performance of the unsupervised SPL-MAD approach. Models trained with SPL consistently outperformed those trained without SPL in various scenarios, including cases with uncontaminated and contaminated training data. While the benefit of SPL was less evident for certain types of attacks, such as FRLL-based [34], [47], [51] attacks with contaminated training data, the overall result underscores its important role in improving MAD performance. The SPL-MAD approach demonstrated comparable results to supervised methods while maintaining greater consistency and avoiding extreme variations in performance. Additionally, its unsupervised nature allowed for effective treatment of unknown morphing attacks, demonstrating higher generalizability compared to supervised methods.

Ferrara et al. [29] show that the effectiveness of MAD varies depending on the type of images and pre-training of the networks. For digital morphs, networks pre-trained on natural images, such as AlexNet [81] and VGG-19 [94], perform well, using general filters that effectively detect digital artifacts and manipulation traces. However, these networks struggle with P&S images because the P&S process removes most digital artifacts, rendering bona fide and morphed images indistinguishable. In contrast, networks pre-trained on large face datasets, such as the VGG-Face [95] models, perform better on P&S images due to their ability to detect fine, face-specific features that survive the printing and scanning process. Incorporating simulated P&S images into training further improves performance, allowing networks like VGG-Face16 [95] to adapt to the specific challenges of P&S morph detection, although AlexNet [81] does not benefit from this additional training. Finally, while the proposed approach has shown superiority over existing methods in specific benchmarks such as NIST FRVT MORPH [79] and SOTAMD [36], overall performance remains modest.

Jia et al. [18] conclude that the proposed approach effectively enhances MAD by leveraging high-frequency features and a progressive enhancement learning framework. High-frequency features proved superior to RGB-based features in distinguishing between bona fide and morphed images, capturing subtle morphing traces that RGB color space has difficulty identifying. The fusion of high-frequency and RGB features in a two-stream network outperformed single-branch networks, highlighting the benefits of combining different feature types for comprehensive representation. The inclusion of self-enhancement and interactive-enhancement modules further improved feature interaction and representation, enabling the detection of subtle morphing artifacts. Experiments demonstrated robust performance across datasets with varying pixel and position fusion factors, highlighting the adaptability of the method. Cross-dataset evaluations confirmed the strong generalizability of the approach, which maintained high effectiveness even when tested on unseen data.

The model proposed by Liu et al. [45] shows superior performance in the identification of morphed images, particularly in scenarios involving accomplices, while maintaining a slightly lower but competitive effectiveness in criminal cases. It excels in the comprehensive ‘‘Both’’ scenario and achieves superior results that confirm its overall reliability and applicability in practical contexts. The authors attribute its success to the innovative use of metric learning to disrupt identity pathways, and the integration of artifact features, which increases the stability of the model and ensures consistent performance across different evaluation scenarios.

Medvedev et al. [14] conclude that the proposed MorDeephy framework demonstrates robust performance and generalizability in detecting face morphing attacks. It effectively generalizes to unseen scenarios, including GAN-generated morphs and P&S morphs, highlighting its adaptability. The use of selfmorphs during training reduces the reliance on detecting blending artifacts and shifts the focus to deep feature behavior, which enhances its ability to handle diverse and previously unseen attack types. The method outperforms other approaches in detecting landmark-based and manually generated morphs while maintaining stable performance across protocols, without bias toward specific generative morphing techniques. Importantly, the method achieves these results without extensive modifications for realistic scenarios, highlighting the effectiveness of focusing on the authenticity of deep facial features for morph detection.

Medvedev et al. [15] build on the previous study. The approach outperforms baseline methods and demonstrates strong generalization capabilities, achieving robust performance across diverse datasets and unseen scenarios without bias towards specific morphing generative strategies. The V2 strategy, which separates morphs from their original classes, further improves performance compared to the V1 strategy, where morphs are linked to their original classes. The integration of face recognition signals with the fused classification detector improves detection, achieving superior results in all benchmarks tested.

The results reported by Ortega-Delcampo et al. [42] indicate that superior MAD results are consistently obtained with digital images compared to P&S images, which perform similarly regardless of resolution. Furthermore, the results surpass those reported in other studies in the literature for both digital and P&S images. The research also confirms the suitability of the CNN-based approach for real-world scenarios, providing robust and accurate detection capabilities. Finally, the study highlights the adaptability of the method to ABC system requirements and its potential for broader applications, such as revealing hidden imposter identities in morphing attacks.

In the experiments conducted by Qin et al. [78], the proposed method achieves reliable detection for complete, spliced and combined morphs when training and testing on the same morphing technique, with slightly higher error rates observed for retouched morphs. The method demonstrates effective localization of morphed regions, although it struggles with complete morphs due to their fully manipulated nature. D-MAD consistently outperforms S-MAD by using probe image information to enhance detection features. The method generalizes well to unseen morphing attacks, particularly in D-MAD, while S-MAD encounters higher error rates for complete morphs, again due to their uniquely tampered nature. It also exhibits stable performance in cross-dataset scenarios, with little degradation caused by variations in datasets, imaging conditions, and morphing parameters. Additionally, the model shows resilience to low-resolution probe images but experiences reduced performance with non-frontal images due to reduced biometric information. Ablation studies demonstrate the importance of components such as feature-wise differential loss that help improve detection accuracy and enhance generalization to unseen morphing attacks and varied datasets.

Razaq and Shukur [19] demonstrate the effectiveness of combining wavelet transform features and high-level features from a CNN for face MAD. DNN outperformed all other classifiers, closely followed by SVM. Other classifiers, such as K-Nearest Neighbors (KNN), Decision Trees, and Naïve Bayes, showed lower performance. The proposed method also performed effectively on the AMSL [47], [48], [49] dataset, identifying morphs with high accuracy, even when the changes targeted specific facial landmarks such as eyes, nose, or mouth.

Robledo-Moreno et al. [61] showed that FL performs comparably to centralized approaches, with significantly better performance when the same morphing technique is used for both training and testing. In S-MAD, FL effectively handles unbalanced datasets and benefits from the inclusion of low-quality morphs, but struggles with high-quality morphs. In D-MAD, FL often outperforms centralized methods, particularly when artifacts in the training data degrade the performance of centralized models. Nevertheless, centralized methods have lower maximum error rates and benefit from augmentation techniques not explored in this study.

The approach proposed by Singh and Ramachandra [28] achieves optimal results in intra-evaluation protocols where the same medium (digital or P&S) is used for training and testing, demonstrating robustness across different cameras and capture distances, illustrating its robustness in real-world scenarios. In cross-evaluation protocols, the method performs better when trained on digital data and tested on P&S images but struggles when the training and testing mediums are reversed, likely due to differences in image quality between digital and P&S data.

Singh and Ramachandra [20] demonstrate the feasibility of MAD in 3D images, with the SimpleView [93] network achieving the best performance, highlighting its strength in using depth-based features to detect 3D morphing attacks.

Soleymani et al. [75] demonstrated that fine-tuning their proposed Siamese network on the training portions of the VISAPP17 [33] and MorGAN [53] datasets improved its detection capabilities, significantly reducing error rates. However, error rates remained higher for the MorGAN [53] dataset, reflecting the inherent difficulty of the dataset due to its lower resolution and generative properties. Texture descriptors outperformed key-point-based approaches in capturing morphing artifacts. The Siamese network's ability to detect subtle differences between images was further strengthened through pre-training on the WVU Twins Day dataset[1] [110], [111], [112], resulting in improved detection accuracy. Additionally, the integration of SVM-based decision-making frameworks, using feature concatenation and difference further enhanced the model's performance.

The model proposed by Venkatesh et al. [41] achieves the best detection performance across three datasets in the intra-dataset experiment. This superior performance is attributed to its innovative aggregated denoising and fusion scheme, which uses sub-band selection for optimal results, and the robustness of MS-CAN in extracting noise features from natural images. The research highlights the advantages of deep features over non-deep feature-based techniques, with AlexNet [81] and ResNet101 [82] achieving the best results among the deep learning models. Moreover, the method shows robustness to increasing sample sizes and varying image characteristics, confirming its adaptability to diverse datasets. It also improves computational efficiency by consolidating multiple denoising operations into a single coherent framework.

The method proposed by Zhang et al. [16] achieves superior performance for digital images with the lowest mean and standard deviation of D-EER among the tested methods, showing an improvement in robustness. However, performance degrades for P&S and print-scan-compression scenarios due to the pre-training of the model on digital images and the lower resolution of P&S data. The approach demonstrates strong generalization capabilities in cross-dataset testing, particularly for GAN-based morphs, where features are well-separated from bona fide samples. In contrast, landmark-based morphs are more difficult to discriminate due to their less separable features. They also conclude that post-processing of morphs increases the strength of attacks, making them difficult to detect.

2) BENCHMARK EVALUATIONS FROM THE REVIEWED LITERATURE

This subsection presents the performance of selected approaches when evaluated on common benchmarks, following consistent evaluation protocols. These benchmark results offer direct comparability between methods under controlled

[1]The WVU Twins Day Dataset 2010-2015 is only available to US nationals and researchers based in the USA [109].

conditions and complement the reported results extracted from the literature.

There are two main platforms: BOEP [113], which is part of the FVC-OnGoing framework [114], [115], and NIST FATE MORPH [116]. Each platform offers multiple benchmarks (e.g., SOTAMD [117] or iMARS for BOEP, and different evaluation tiers for NIST).

FVC-onGoing [114], [115] is a long-standing benchmarking platform developed by the Biometric System Laboratory at the University of Bologna. Originally focused on fingerprint recognition, it has expanded to encompass broader biometric challenges, including face morphing detection [118].

Building upon this infrastructure, the Bologna Online Evaluation Platform (BOEP) [113] was developed to offer a standardized evaluation environment specifically tailored for MAD algorithms. BOEP [113] facilitates the submission and automated evaluation of algorithms, ensuring consistent and reproducible assessments across different MAD approaches [113].

Within this ecosystem, the State-of-the-Art Morphing Detection (SOTAMD) benchmark [117] was introduced to evaluate MAD systems under realistic and challenging conditions, such as varying image qualities and morphing techniques. SOTAMD leverages the BOEP framework to provide a rigorous testing environment that reflects operational scenarios encountered in border control and identity verification processes [113].

In parallel, NIST FATE Morph [116] is an independent benchmark developed by the United States (U.S.) National Institute of Standards and Technology (NIST) as part of the Face Analysis Technology Evaluation (FATE) program, formerly part of the FRVT (Face Recognition Vendor Test) framework. This benchmark focuses on evaluating face morphing detection algorithms in high-security identity verification contexts, such as border control and credential issuance. It employs proprietary datasets and rigorously defined evaluation protocols, emphasizing operational effectiveness, robustness to image quality variations, and resistance to diverse morphing techniques. Although developed independently from European frameworks like BOEP and SOTAMD, NIST FATE Morph serves as a complementary international benchmark, enabling broader cross-national comparisons and fostering standardization across institutions engaged in MAD research [116].

Table 5 presents the best performing results from the reviewed papers that report performance on standardized MAD benchmarks.

Ferrara et al. [29] show strong performance with VGG-Face16 fine-tuned on digital images from MorphDB, achieving high accuracy (97.4%) and low error rates (0.0% BPCER). However, performance drops significantly on P&S images, with BPCER rising to 43.8% at 1% APCER, and accuracy falling to 93.5%. While the model performs well on the NIST S-MAD benchmark using the UNIBO Automatic Morphed Face Generation Tool v1.0–achieving an APCER of 5% at a BPCER of 1%–this result may be misleading. On four other benchmarks within the same NIST S-MAD evaluation, the model records an APCER as high as 91% at the same BPCER threshold. Furthermore, in the FVC S-MAD SMAD-SOTAMD_P&S-1.0 test, BPCER hits 100% at 5% and 10% APCER, reflecting a complete failure in attack detection.

Zhang et al. [16] present moderate performance in controlled scenarios but struggle under standardized evaluation. In a cross-technique evaluation on digital morphs from the FRGC dataset, their method achieves a D-ERR of 13.63%, indicating a fair ability to detect morphs generated through varying techniques. However, in the FVC S-MAD SMAD-BIOLAB-1.0 benchmark, which simulates more realistic and diverse conditions, performance degrades substantially, with a BPCER of 78.41% at 10% APCER and a D-ERR of 40.45%. These results highlight a limited generalization capability when moving from controlled datasets to real-world applications.

Di Domenico et al. [17] report consistent but moderate performance across both controlled and standardized settings. When trained on a universal dataset and tested on FEI accomplice images using Min-Max feature fusion and cosine similarity, their method achieves a D-ERR of 12.5% and a BPCER of 46.8% at 1% APCER, improving to 23.7% at 5% APCER, indicating challenges in correctly classifying bona fide images under stricter thresholds. In the standardized FVC D-MAD DMAD-SOTAMD D-1.0 benchmark, the model performs slightly better with a D-ERR of approximately 10%, suggesting reasonable generalization and reliability in more diverse and realistic conditions.

Liu et al. [45] achieve strong results in a controlled criminal scenario, with a D-ERR of 3.7% and a low BPCER of 11.5% at 1% APCER, improving to just 2% at 5% APCER, demonstrating effective morph detection and minimal false rejections. However, in the standardized FVC D-MAD DMAD-SOTAMD D-1.0 benchmark, performance drops, with BPCER increasing to 34.1% at 1% APCER and a D-ERR of 9.2%, indicating reduced robustness and difficulty in maintaining accuracy under more realistic and varied conditions.

Robledo-Moreno et al. [61] demonstrate outstanding performance in a cross-technique evaluation for S-MAD, achieving 0.00% BPCER and D-ERR at both 1% and 5% APCER, suggesting highly effective morph detection under controlled technique variation. However, such perfect detection may also indicate possible overfitting, especially given the contrast with other evaluations. In a cross-dataset evaluation for D-MAD, performance decreases notably, with BPCER rising to 34.3% at 1% APCER and a D-ERR of 7.1%, pointing to limited generalization across different data sources. In the standardized FVC D-MAD DMAD-SOTAMD benchmark–where the type of images used (digital, P&S, or others) is not specified–performance drops further, with BPCER reaching 28.8% at 10% APCER and D-ERR

TABLE 5. Summary of the best results reported from each study included in the literature review on standardized MAD benchmarks. For comparison purposes, only the best results per evaluation system were selected.

Evaluation System	MAD Type	Benchmark	Reviewed Literature	BPCER (%) @ APCER 1%	BPCER (%) @ APCER 5%	BPCER (%) @ APCER 10%	APCER (%) @ BPCER 1%	D-EER (%)
NIST	S-MAD	UNIBO Automatic Morphed Face Generation Tool v1.0	Ferrara et al. [29]	-	-	-	5	-
FVC	S-MAD	SMAD-SOTAMD_P&S-1.0		-	100.00	100.00	-	37.10
		SMAD-BIOLAB-1.0	Zhang et al. [16]	-	-	78.41	-	40.45
	D-MAD	DMAD-SOTAMD_D-1.0	Di Domenico et al. [17]	-	-	-	-	∼10
			Liu et al. [45]	34.1	16.8	8.0	-	9.2
		DMAD-SOTAMD[1]	Robledo-Moreno et al. [61]	-	-	28.8	-	18.2

[1] Not further specified.

increasing to 18.2%, suggesting reduced robustness under more realistic and varied conditions.

Moving on to direct comparisons between studies using the same benchmark, the FVC D-MAD DMAD-SOTAMD D-1.0 benchmark, conducted on high-resolution digital images, provides an interesting contrast. Di Domenico et al. [17] achieves a D-ERR of approximately 10%, while Liu et al. [45] slightly outperforms with a D-ERR of 9.2%. Liu's results also provide more detailed information, with BPCER values of 34.1% at 1% APCER, 16.8% at 5%, and 8% at 10%, showing a notable improvement in performance as more attack errors are tolerated. Di Domenico et al. [17], while similar in overall detection performance, does not report BPCER values, limiting the ability to assess its handling of bona fide samples. Both studies exhibit moderate robustness on high-resolution digital images, with Liu et al. [45] offering a slightly better balance between attack detection and user convenience.

V. FINDINGS AND TRENDS

The literature review uncovers several findings and emerging trends in the application of deep learning techniques to MAD, providing valuable insights into how these techniques can improve MAD performance. Building on these findings, this section directly addresses the Research Question: "How can deep learning techniques be used to improve MAD?" The following key points summarize the ways in which deep learning techniques can improve MAD:

- **Feature Extraction and Hybrid Approaches:** Feature extraction is a critical step in improving MAD. Deep learning models, particularly CNNs, automate this process, allowing the identification of subtle traces of manipulation that traditional methods may miss, thus improving the detection accuracy for various morphing techniques. According to Venkatesh et al. [41], deep learning techniques provide features that yield better performance, allowing models to better distinguish between morphed and genuine images. This highlights the importance of using advanced methods that capture subtle patterns or inconsistencies in morphed images. Several studies - Aghdaie et al. [74], Chaudary et al. [44], and Razaq and Shukur [19] - have shown how models use wavelet domain analysis to identify discriminative features. In addition, the combination of identity and artifact-based features, as demonstrated by Borghi et al. [27], proves effective for robust detection. The trend of incorporating multiple feature types, such as high-frequency and RGB image data, as seen in Jia et al. [18], is gaining traction for its ability to improve detection accuracy across different morphing techniques.
- **Cross-Dataset Generalization:** One of the key trends is the ability of deep learning models to generalize across different morphing techniques and datasets. Damer et al. [43] highlight the importance of generalization for real-world applications, where unseen morphing techniques frequently occur. However, Banerjee and Ross [57] and Blasingame and Liu [66] showed that many models trained on one dataset often struggle when tested on another, especially when different morphing techniques are used. In addition, variations in environmental factors such as lighting, angle, and distance across datasets can further hinder the generalizability capability of these models, as they are often trained on controlled datasets that do not adequately capture such real-world conditions.
- **Unsupervised and Semi-Supervised Learning:** Unsupervised learning methods have emerged as a critical trend in MAD, enabling the detection of unknown morphing techniques without requiring large labeled datasets. Studies such as those by Fang et al. [65] and Banerjee and Ross [57] highlight the benefits of unsupervised approaches that improve flexibility and detection of new morphing attacks.
- **GAN-Based and Landmark-Based Morphs:** A notable trend in the field is the challenge of detecting high-quality landmark-generated morphs compared to GAN-based morphs. Chaudhary et al. [44] and Zhang et al. [16] found that landmark-based morphs are harder to detect due to more subtle artifacts created by manipulating facial features. On the other hand, GAN-based morphs produce more pronounced features that are easier to detect. This points to the need for deep

learning models that are capable of detecting both strong and weak morphs across different types of attacks.

- **Performance Variations with Image Types:** The performance of MAD techniques varies considerably depending on the image type. Studies by Ferrara et al. [29] and Ortega-Delcampo et al. [42] demonstrate that digital images typically yield better results than P&S images, which are often used in real-world scenarios such as passport verification. The challenge posed by P&S morphs, which may lack detectable artifacts, highlights the importance of training models on different image types to ensure robustness and improve detection in different scenarios.
- **Advanced Morphing Techniques:** With the increasing sophistication of morphing techniques, the development of MAD models that can keep up with these advances has become a major concern. Techniques such as enhanced post-processing aim to enhance the quality of morphed images, making them more difficult to detect. Zhang et al. [16] discuss how these attacks complicate the detection process, requiring more advanced methods capable of detecting subtle traces of morphing.

In summary, deep learning techniques have played a critical role in advancing MAD by automating feature extraction, improving generalization across different datasets, and enabling the identification of images created by advanced morphing methods. Approaches that incorporate hybrid models, unsupervised learning, and multi-modal features continue to evolve and provide solutions to the growing complexity of morphing attacks.

VI. CHALLENGES AND GAPS

This section discusses the challenges and gaps identified in the reviewed studies. Despite significant advances in MAD, several limitations remain in the proposed methods. The key challenges that hinder the effectiveness and applicability of current MAD techniques are:

- **Generalization to Unseen Morphing Techniques:** Many existing methods perform well on specific datasets or types of attacks but struggle when faced with novel morphing techniques, particularly those that involve subtle or complex image manipulation. Furthermore, the ability of these methods to detect morphing attacks in real-world scenarios remains a concern, especially when dealing with low-quality or compressed images commonly encountered in border control systems and document verification applications.
- **Reliance on Digital Images:** The reliance on digital images alone is another gap, as many real-world applications, such as passport verification or border control, involve scanned or printed images. These images often suffer from low resolution or compression artifacts, which can reduce the accuracy of morph detection techniques designed primarily for high-quality digital images. The performance of existing methods under these conditions is often untested and, when tested, unsatisfactory, leaving a critical gap in their real-world applicability.
- **Privacy Concerns with Synthetic Datasets:** While synthetic datasets like SMDD [54], MONOT [55], and AMONOT [56] are essential for protecting privacy by avoiding the use of real biometric data, they come with their own set of challenges. Ensuring that synthetic data accurately mimics real-world biometric data while maintaining privacy is difficult. There is always a risk that synthetic data could be reverse-engineered or traced back to real individuals, potentially compromising privacy. Additionally, synthetic datasets may not fully capture the diversity and variability of real-world data, which could limit the effectiveness and fairness of MAD systems across different populations and morphing techniques [119].
- **Reliance on Controlled Datasets:** Another major limitation is the reliance on controlled datasets. While datasets with a variety of morphing techniques are useful, there is still a lack of diversity in the subjects used for training. Most studies focus on limited demographic groups, with insufficient representation of subjects of different ages, genders, and ethnicities. Furthermore, the majority of datasets are collected under controlled conditions that often lack variability in environmental factors such as lighting, angle, and distance. This lack of diversity in training data can affect the model's ability to generalize to a broader range of real-world situations.
- **Handling Sophisticated Morphing Methods:** Many of the advanced detection techniques, while promising, face difficulties in handling more sophisticated morphing methods that produce fewer detectable artifacts. As morphing techniques evolve and become more advanced, the artifacts produced are often subtle and harder to detect, which can reduce the effectiveness of existing detection systems. In addition, several methods rely on specific image properties, such as wavelet domain features or pixel-wise analysis, which may not generalize well to all types of morphs, particularly those with minimal detectable distortions. Furthermore, recent studies have shown that attackers are employing face restoration techniques, such as CodeFormer [120], to automatically retouch morphed images, removing artifacts and enhancing visual quality. This process results in high-quality morphed images that are more challenging for detection systems to identify [121].

Overall, while the reviewed studies have made important strides in detecting face morphing attacks, future research must focus on addressing these challenges, in particular, improving the robustness and generalization of MAD techniques, increasing the diversity of datasets, and improving detection capabilities for sophisticated, low-artifact morphs.

VII. FUTURE RESEARCH DIRECTIONS

Despite advances in MAD, several challenges remain. The following areas offer promising opportunities for future research:

- **Generalization to Unseen Morphing Techniques:** Future research should focus on improving model generalization across different morphing techniques and expanding training datasets to include a wider variety of morphing methods.
- **P&S Images:** Research should aim to develop techniques that perform well under conditions of resolution loss, compression artifacts, and distortion commonly found in real-world applications such as passport control.
- **Expanding Dataset Diversity:** Future studies should aim to include subjects of different ages, genders, and ethnicities, as well as high-quality and P&S images, to improve model generalization and real-world applicability.
- **Addressing Dataset Bias and Domain Adaptation:** Future work should explore unsupervised domain adaptation techniques (e.g., Domain-Adversarial Neural Networks [122]) and image translation methods (e.g., CycleGAN [123]) to reduce performance gaps between training on high-quality images and testing on compressed, scanned, or low-resolution samples. Domain generalization strategies that do not require access to target domain data can further improve adaptability. Additionally, integrating bias mitigation techniques–such as sample re-weighting or fairness-aware augmentation–can help counter demographic imbalances and improve real-world performance [124], [125].
- **Robustness to Subtle Artifacts:** Future research should focus on developing models capable of detecting subtle morphing traces using more advanced features, possibly combining pixel-level, frequency-domain, and texture-based methods. The challenge is further exacerbated by the use of face restoration techniques, which reduce or mask typical morphing artifacts, making them less distinguishable to conventional detection systems. As a result, detection methods must become increasingly sensitive to these imperceptible alterations introduced by both sophisticated morphing and post-processing strategies.
- **Real-Time and Cross-Dataset Evaluation:** To improve real-world applicability, future studies should evaluate MAD systems under real-time conditions with varying image resolutions, lighting, and environmental factors. In addition, models should be evaluated across different datasets to ensure that they perform efficiently and accurately in real-world applications such as border control and identity verification.
- **Unsupervised and Semi-Supervised Learning:** Future research could explore unsupervised and semi-supervised approaches, which can help overcome the challenges of acquiring large labeled datasets.
- **Privacy-Oriented and Adaptive Training Strategies:** In addition to synthetic datasets for privacy preservation, future research should explore training strategies that protect sensitive biometric data. Techniques like Federated learning allow models to be trained across multiple institutions without the need to share raw data. Continual and incremental learning approaches enable systems to adapt to emerging morphing techniques over time, maintaining performance without the need for complete retraining or access to previous data. These strategies enhance both privacy and the adaptability of MAD systems [126].
- **Video-Based MAD in Operational Scenarios:** Recent work has explored the use of video data for MAD, introducing methods that leverage temporal and dynamic facial features. This direction, exemplified by the Video-based MAD (V-MAD) approach, opens new avenues for detecting morphs in realistic settings such as border control, where video sequences are naturally available and can provide richer information than still images [127].
- **Explainability and Interpretability of MAD Models:** Future research should explore explainability techniques, such as Grad-CAM, SHAP, and feature attribution methods, to improve the transparency and trustworthiness of deep learning-based MAD systems. While most current approaches emphasize detection performance, understanding the underlying decision-making processes can enhance reliability in critical applications like border control. The broader machine learning community has also emphasized the importance of interpretability in complex models [128], highlighting its relevance as a complementary direction for advancing MAD research.

By focusing on these areas, future research can address existing gaps and pave the way for more robust and effective MAD systems.

VIII. CONCLUSION

This literature review analyzes the current state of deep learning techniques for detecting morphed face images, focusing on the effectiveness of different methods and models used in recent research. The results highlight the progress made in MAD, particularly through the use of CNNs, GANs, and autoencoders. These deep learning models have proven effective in automating feature extraction, identifying subtle traces of manipulation, and improving detection accuracy across different morphing techniques.

However, despite these advances, the detection of morphed images still faces several challenges. As morphing techniques evolve and become more sophisticated, the artifacts created by these attacks become less detectable, particularly in high-quality morphs. In addition, while current methods

show promising results in controlled environments, their performance often degrades when applied to real-world scenarios involving low-resolution, compressed, or printed images.

Furthermore, issues related to dataset diversity and model generalization across different morphing techniques remain relevant barriers. Many models perform well on specific datasets but struggle when faced with new, unseen morphing attacks, highlighting the need for more robust and generalized detection systems. In addition, reliance on labeled datasets poses limitations, and some methods are exploring unsupervised learning to fill this gap.

In summary, deep learning techniques have proven valuable in improving MAD systems. However, additional research is needed to improve the robustness and generalizability capability of these systems, particularly in detecting subtle morphing artifacts and applying them to practical applications such as identity verification and border control. Future efforts should focus on expanding datasets, improving model adaptability, and developing methods capable of detecting increasingly advanced morphing techniques. Overcoming these challenges will make MAD systems more reliable in protecting against morphing-based identity fraud.

ACKNOWLEDGMENT

The authors acknowledge the use of OpenAI's ChatGPT to assist with grammar and language refinement during the preparation of this manuscript. No AI-generated content was used for technical analysis, literature synthesis, and figure creation. All scientific interpretations and conclusions are the sole responsibility of them.

REFERENCES

[1] M. Ibsen, L. J. Gonzalez-Soler, C. Rathgeb, and C. Busch, "TetraLoss: Improving the robustness of face recognition against morphing attacks," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2024, pp. 1–10.

[2] S. Baechler, "Document fraud: Will your identity be secure in the twenty-first century?" *Eur. J. Criminal Policy Res.*, vol. 26, no. 3, pp. 379–398, Sep. 2020, doi: 10.1007/S10610-020-09441-8.

[3] R. Kessler, K. Raja, J. Tapia, and C. Busch, "Towards minimizing efforts for morphing attacks—Deep embeddings for morphing pair selection and improved morphing attack detection," *PLoS ONE*, vol. 19, no. 5, May 2024, Art. no. e0304610.

[4] U. Scherhag, C. Rathgeb, J. Merkle, R. Breithaupt, and C. Busch, "Face recognition systems under morphing attacks: A survey," *IEEE Access*, vol. 7, pp. 23012–23026, 2019.

[5] FaceShape. (2025). *Face Morph Tool*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.faceshape.com/face-morph

[6] Afroaira. (2025). *Social Media Post on Threads*. Accessed: Jan. 19, 2025. [Online]. Available: https://encrypted-tbn0.gstatic.com/images?q=tbn:ANd9GcRKSK153VgujNLAsMhEq75L_6GpbpQhENgIi2490heM_92CrAbZ

[7] Pinterest. (2025). *Rihanna at the Fenty Beauty Launch in Paris*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.pinterest.com/pin/rihanna-at-the-fenty-beauty-launch-in-paris–522136150566546614/

[8] A. Mujahed. (2013). *Beyonce at the 55th Annual Grammy Awards*. Accessed: Jan. 19, 2025. [Online]. Available: https://whatsinfashioncwb.wordpress.com/tag/grammy-2013/

[9] Pinterest. (2015). *Rihanna Makeup at Christian Dior Front Row, Paris Fashion Week*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.pinterest.com/pin/532902568386886802/

[10] FaceShape. (2025). *Face Compare Tool*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.faceshape.com/face-compare

[11] Freepik. (2025). *Check Icon*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.freepik.com/icon/check_5610944

[12] Freepik. (2025). *Delete Icon*. Accessed: Jan. 19, 2025. [Online]. Available: https://www.freepik.com/icon/delete_5610967

[13] S. R. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, R. Ramachandra, and C. Busch, "Analyzing human observer ability in morphing attack detection—Where do we stand?" *IEEE Trans. Technol. Soc.*, vol. 4, no. 2, pp. 125–145, Jun. 2023.

[14] I. Medvedev, F. Shadmand, and N. Gonçalves, "MorDeephy: Face morphing detection via fused classification," in *Proc. 12th Int. Conf. Pattern Recognit. Appl. Methods*, Sep. 2023, pp. 193–204.

[15] I. Medvedev, J. A. Pimenta, and N. Gonçalves, "Fused classification for differential face morphing detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2024, pp. 1043–1050.

[16] H. Zhang, R. Ramachandra, K. Raja, and C. Busch, "Generalized single-image-based morphing attack detection using deep representations from vision transformer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2024, pp. 1510–1518.

[17] N. D. Domenico, G. Borghi, A. Franco, and D. Maltoni, "Combining identity features and artifact analysis for differential morphing attack detection," in *Image Analysis and Processing—ICIAP 2023*, vol. 14233. Cham, Switzerland: Springer, 2023, pp. 100–111.

[18] C.-K. Jia, Y.-C. Liu, and Y.-L. Chen, "Face morphing attack detection based on high-frequency features and progressive enhancement learning," *Frontiers Neurorobotics*, vol. 17, pp. 1–11, Jun. 2023.

[19] I. S. Razaq and B. K. Shukur, "Combining wavelet transforms features and high-level features using CNN for face morphing attack detection," *Int. J. Inf. Technol.*, vol. 15, no. 7, pp. 3957–3966, Oct. 2023.

[20] J. M. Singh and R. Ramachandra, "3-D face morphing attacks: Generation, vulnerability and detection," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 6, no. 1, pp. 103–117, Jan. 2024.

[21] R. Raghavendra, K. B. Raja, and C. Busch, "Detecting morphed face images," in *Proc. IEEE 8th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2016, pp. 1–7.

[22] C. Busch, "Morphing attack detection—State of the art and challenges," Presentation at the 20th IAPR/IEEE Int. Summer School Adv. Stud. Biometrics, Jun. 2023, pp. 14–20. [Online]. Available: https://www.christoph-busch.de/files/Busch-MAD-230606.pdf

[23] D. Menotti, G. Chiachia, A. Pinto, W. R. Schwartz, H. Pedrini, A. X. Falcão, and A. Rocha, "Deep representations for iris, face, and fingerprint spoofing detection," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 4, pp. 864–879, Apr. 2015.

[24] M. Ferrara, A. Franco, and D. Maltoni, "The magic passport," in *Proc. IEEE Int. Joint Conf. Biometrics*, Sep. 2014, pp. 1–7.

[25] U. Scherhag, L. Debiasi, C. Rathgeb, C. Busch, and A. Uhl, "Detection of face morphing attacks based on PRNU analysis," *IEEE Trans. Biometrics, Behav., Identity Sci.*, vol. 1, no. 4, pp. 302–317, Oct. 2019.

[26] P. Sharma, M. Kumar, and H. Sharma, "Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: An evaluation," *Multimedia Tools Appl.*, vol. 82, no. 12, pp. 18117–18150, May 2023, doi: 10.1007/s11042-022-13808-w.

[27] G. Borghi, E. Pancisi, M. Ferrara, and D. Maltoni, "A double Siamese framework for differential morphing attack detection," *Sensors*, vol. 21, no. 10, p. 3466, May 2021.

[28] J. M. Singh and R. Ramachandra, "Reliable face morphing attack detection in on-the-fly border control scenario with variation in image resolution and capture distance," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–10.

[29] M. Ferrara, A. Franco, and D. Maltoni, "Face morphing detection in the presence of printing/scanning and heterogeneous image sources," *IET Biometrics*, vol. 10, no. 3, pp. 290–303, May 2021.

[30] U. Scherhag, C. Rathgeb, and C. Busch, *Face Morphing Attack Detection Methods*. Cham, Switzerland: Springer, 2022, pp. 331–349, doi: 10.1007/978-3-030-87664-7_15.

[31] S. Venkatesh, R. Ramachandra, K. Raja, and C. Busch, "Face morphing attack generation and detection: A comprehensive survey," *IEEE Trans. Technol. Soc.*, vol. 2, no. 3, pp. 128–145, Sep. 2021.

[32] J. M. Singh, R. Ramachandra, K. B. Raja, and C. Busch, "Robust morph-detection at automated border control gate using deep decomposed 3D shape & diffuse reflectance," in *Proc. 15th Int. Conf. Signal-Image Technol. Internet-Based Syst. (SITIS)*, Nov. 2019, pp. 106–112.

[33] A. Makrushin, T. Neubert, and J. Dittmann, "Automatic generation and detection of visually faultless facial morphs," in *Proc. 12th Int. Joint Conf. Comput. Vis., Imag. Comput. Graph. Theory Appl. (VISIGRAPP)*, Aug. 2017, pp. 39–50.

[34] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Are GAN-based morphs threatening face recognition?" in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2022, pp. 2959–2963.

[35] *Information Technology—Biometric Presentation Attack Detection—Part 3: Testing and Reporting*, Standard 30107-3:2023, IOS, Geneva, Switzerland, 2023.

[36] K. Raja, M. Ferrara, A. Franco, L. Spreeuwers, I. Batskos, and F. de Wit, "Morphing attack detection-database, evaluation platform, and benchmarking," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 4336–4351, 2021.

[37] A. Sequeira, "Liveness detection and robust recognition in iris and fingerprint biometric systems," Ph.D. dissertation, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal, 2015.

[38] S. Gonzalez and J. Tapia, "Improving presentation attack detection for ID cards on remote verification systems," 2023, *arXiv:2301.09542*.

[39] J. Kolberg, D. Gläsner, R. Breithaupt, M. Gomez-Barrero, J. Reinhold, A. von Twickel, and C. Busch, "On the effectiveness of impedance-based fingerprint presentation attack detection," *Sensors*, vol. 21, no. 17, p. 5686, Aug. 2021. [Online]. Available: https://www.mdpi.com/1424-8220/21/17/5686

[40] N. Japkowicz and M. Shah, *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge, U.K.: Cambridge Univ. Press, 2011.

[41] S. Venkatesh, R. Ramachandra, K. Raja, L. Spreeuwers, R. Veldhuis, and C. Busch, "Detecting morphed face attacks using residual noise from deep multi-scale context aggregation network," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 269–278.

[42] D. Ortega-Delcampo, C. Conde, D. Palacios-Alonso, and E. Cabello, "Border control morphing attack detection with a convolutional neural network de-morphing approach," *IEEE Access*, vol. 8, pp. 92301–92313, 2020.

[43] N. Damer, N. Spiller, M. Fang, F. Boutros, F. Kirchbuchner, and A. Kuijper, "PW-MAD: Pixel-wise supervision for generalized face morphing attack detection," in *Proc. Int. Symp. Vis. Comput.*, vol. 13017, 2021, pp. 291–304.

[44] B. Chaudhary, P. Aghdaie, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Differential morph face detection using discriminative wavelet sub-bands," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 1425–1434.

[45] C. Liu, M. Ferrara, A. Franco, G. Borghi, and D. Zhong, "Differential morphing attack detection via triplet-based metric learning and artifact extraction," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Sep. 2024, pp. 1–7.

[46] L. Colbois and S. Marcel, "Evaluating the effectiveness of attack-agnostic features for morphing attack detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2024, pp. 1–9.

[47] I. R. Inst. (2024). *FRLL-Morphs Dataset*. Accessed: Dec. 26, 2024. [Online]. Available: https://www.idiap.ch/en/scientific-research/data/frll-morphs

[48] AMSL. *AMSL Face Morph Image Dataset*. Accessed: Jan. 20, 2025. [Online]. Available: https://omen.cs.uni-magdeburg.de/disclaimer/index.php

[49] T. Neubert, A. Makrushin, M. Hildebrandt, C. Kraetzer, and J. Dittmann, "Extended stirtrace benchmarking of biometric and forensic qualities of morphed face images," *IET Biometrics*, vol. 7, no. 4, pp. 325–332, Jul. 2018. [Online]. Available: https://ietresearch.onlinelibrary.wiley.com/doi/abs/10.1049/iet-bmt.2017.0147

[50] Q. Cao, L. Shen, W. Xie, O. Parkhi, and A. Zisserman. (2024). *VGGFace2 Dataset*. Accessed: Dec. 26, 2024. [Online]. Available: https://www.robots.ox.ac.uk/

[51] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," 2020, *arXiv:2012.05344*.

[52] M. Ferrara, A. Franco, and D. Maltoni, "Face demorphing," *IEEE Trans. Inf. Forensics Security*, vol. 13, no. 4, pp. 1008–1017, Apr. 2018.

[53] N. Damer, A. M. Saladié, A. Braun, and A. Kuijper, "MorGAN: Recognition vulnerability and attack detectability of face morphing attacks created by generative adversarial network," in *Proc. IEEE 9th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Oct. 2018, pp. 1–10.

[54] N. Damer, C. A. F. López, M. Fang, N. Spiller, M. V. Pham, and F. Boutros, "Privacy-friendly synthetic data for the development of face morphing attack detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1605–1616.

[55] G. Borghi, N. D. Domenico, M. Ferrara, A. Franco, U. Latif, and D. Maltoni, "MONOT: High-quality privacy-compliant morphed synthetic images for everyone," in *Proc. IEEE Int. Workshop Inf. Forensics Secur. (WIFS)*, Dec. 2024, pp. 1–8.

[56] G. Spathis, N. D. Domenico, G. Borghi, A. Franco, and D. Maltoni, "AMONOT: Synthetic aging for differential morphing attack detection systems," in *Proc. Int. Conf. Pattern Recognit. Workshops (ICPRW)*, Jan. 2025, pp. 353–367.

[57] S. Banerjee and A. Ross, "Conditional identity disentanglement for differential face morph detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.

[58] PICS. *Psychological Image Collection at Stirling (PICS)*. Accessed: Dec. 26, 2024. [Online]. Available: https://pics.stir.ac.uk/

[59] G. Borghi, N. D. Domenico, A. Franco, M. Ferrara, and D. Maltoni, "Revelio: A modular and effective framework for reproducible training and evaluation of morphing attack detectors," *IEEE Access*, vol. 11, pp. 120419–120437, 2023.

[60] M.-T. Lab. (2023). *ChiMo Dataset*. Accessed: Jan. 18, 2025. [Online]. Available: https://miatbiolab.csr.unibo.it/chimo-dataset/

[61] M. Robledo-Moreno, G. Borghi, N. Di Domenico, A. Franco, K. Raja, and D. Maltoni, "Towards federated learning for morphing attack detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2024, pp. 1–10.

[62] *FEI Face Database*. Accessed: Dec. 26, 2024. [Online]. Available: https://fei.edu.br/

[63] M.-T. Lab. (2023). *FEI-Morph Dataset*. Accessed: Jan. 18, 2025. [Online]. Available: https://miatbiolab.csr.unibo.it/fei-morph-dataset/

[64] NIST. (2024). *Face Recognition Technology (FERET)*. Accessed: Dec. 26, 2024. [Online]. Available: https://www.nist.gov/programs-projects/face-recognition-technology-feret

[65] M. Fang, F. Boutros, and N. Damer, "Unsupervised face morphing attack detection via self-paced anomaly detection," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Oct. 2022, pp. 1–11.

[66] Z. Blasingame and C. Liu, "Leveraging adversarial learning for the detection of morphing attacks," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Aug. 2021, pp. 1–8.

[67] NIST. (2024). *Face Recognition Grand Challenge (FRGC)*. Accessed: Dec. 26, 2024. [Online]. Available: https://www.nist.gov/programs-projects/face-recognition-grand-challenge-frgc

[68] F. R. L. London. (2017). *Face Research Lab London Set*. Accessed: Jan. 18, 2025. [Online]. Available: https://figshare.com/articles/dataset/Face_Research_Lab_London_Set/5047666

[69] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao, "FaceScape: A large-scale high quality 3D face dataset and detailed riggable 3D face prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 598–607. [Online]. Available: https://facescape.nju.edu.cn/

[70] H. Zhu, H. Yang, L. Guo, Y. Zhang, Y. Wang, M. Huang, M. Wu, Q. Shen, R. Yang, and X. Cao, "FaceScape: 3D facial dataset and benchmark for single-view 3D face reconstruction," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 14528–14545, Dec. 2023. [Online]. Available: https://facescape.nju.edu.cn/

[71] H. Zhu, H. Yang, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. (2024). *FaceScape*. Accessed: Dec. 25, 2024. [Online]. Available: https://facescape.nju.edu.cn/

[72] T. Karras, S. Laine, and T. Aila. (2024). *Flickr-Faces-HQ Dataset (FFHQ)*. Accessed: Dec. 25, 2024. [Online]. Available: https://github.com/NVlabs/ffhq-dataset

[73] F. Peng, L.-B. Zhang, and M. Long, "FD-GAN: Face de-morphing generative adversarial network for restoring accomplice's facial image," *IEEE Access*, vol. 7, pp. 75122–75131, 2019.

[74] P. Aghdaie, B. Chaudhary, S. Soleymani, J. Dawson, and N. M. Nasrabadi, "Morph detection enhanced by structured group sparsity," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. Workshops (WACVW)*, Jan. 2022, pp. 311–320.

[75] S. Soleymani, B. Chaudhary, A. Dabouei, J. Dawson, and N. M. Nasrabadi, "Differential morphed face detection using deep Siamese networks," in *Proc. Int. Conf. Pattern Recognit.*, vol. 12666, 2021, pp. 560–572.

[76] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer, "Detecting facial retouching using supervised deep learning," *IEEE Trans. Inf. Forensics Security*, vol. 11, no. 9, pp. 1903–1913, Sep. 2016.
[77] CVRL. *ND-IIITD Retouched Face Database*. Accessed: Dec. 28, 2024. [Online]. Available: https://cvrl.nd.edu/projects/data/#nd-iiitd-retouched-face-database
[78] L. Qin, F. Peng, and M. Long, "Face morphing attack detection and localization based on feature-wise supervision," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 3649–3662, 2022.
[79] M. Ngan, P. Grother, K. Hanaoka, and J. Kuo, "Face recognition vendor test (FRVT) part 4: Morph—Performance of automated face morph detection," Nat. Institute Standards Technol. (NIST), Gaithersburg, MD, USA, Tech. Rep. NISTIR 8292, 2020.
[80] M. Grgic, K. Delac, S. Grgic, and B. Klimpak. (2024). *Scface—Surveillance Cameras Face Database*. Accessed: Dec. 25, 2024. [Online]. Available: https://www.scface.org/
[81] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017, doi: 10.1145/3065386.
[82] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[83] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.
[84] R. Q. Charles, H. Su, M. Kaichun, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 77–85.
[85] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[86] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2261–2269.
[87] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[88] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9.
[89] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, Feb. 2017, pp. 4278–4284.
[90] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.* Red Hook, NY, USA: Curran Associates Inc., 2017, p. 5105.
[91] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[92] X. Zhang, X. Zhou, M. Lin, and J. Sun, "ShuffleNet: An extremely efficient convolutional neural network for mobile devices," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.
[93] M. A. Uy, Q.-H. Pham, B.-S. Hua, T. Nguyen, and S.-K. Yeung, "Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1588–1597, doi: 10.1109/ICCV.2019.00167.
[94] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[95] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.*, X. Xie, M. W. Jones, and G. K. L. Tam, Eds., Sep. 2015, pp. 41.1–41.12, doi: 10.5244/C.29.41.
[96] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1800–1807.
[97] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 44–436, May 2015.
[98] V. H. Phung and E. J. Rhee, "A high-accuracy model average ensemble of convolutional neural networks for classification of cloud image patches on small datasets," *Appl. Sci.*, vol. 9, no. 21, p. 4500, Oct. 2019.
[99] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006, doi: 10.1126/science.1127647.
[100] J. D. Cunningham, D. Shu, T. W. Simpson, and C. S. Tucker, "A sparsity preserving genetic algorithm for extracting diverse functional 3D designs from deep generative neural networks," *Design Sci.*, vol. 6, p. e11, Jun. 2020.
[101] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 3, 2014, pp. 1–11.
[102] B. Sabiri, B. E. Asri, and M. Rhanoui, "Impact of hyperparameters on the generative adversarial networks behavior," in *Proc. 24th Int. Conf. Enterprise Inf. Syst.*, vol. 1, Jul. 2022, pp. 428–438.
[103] N. Damer, F. Boutros, A. M. Saladié, F. Kirchbuchner, and A. Kuijper, "Realistic dreams: Cascaded enhancement of GAN-generated images with an example in face morphing attacks," in *Proc. IEEE 10th Int. Conf. Biometrics Theory, Appl. Syst. (BTAS)*, Sep. 2019, pp. 1–10.
[104] G. H. de Rosa and J. P. Papa, "Learning to weight similarity measures with Siamese networks: A case study on optimum-path forest," in *Optimum-Path Forest*, A. X. Falcao and J. P. Papa, Eds., New York, NY, USA: Academic, 2022, ch. 7, pp. 155–173. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128226889000153
[105] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
[106] NVIDIA. (2024). *StyleGAN2*. Accessed: Jan. 5, 2024. [Online]. Available: https://github.com/NVlabs/stylegan2
[107] (2024). *FaceMorpher*. Accessed: Jan. 5, 2024. [Online]. Available: http://www.facemorpher.com/
[108] O. Team. (2024). *OpenCV*. Accessed: Jan. 5, 2024. [Online]. Available: https://opencv.org/
[109] L. Omelina, J. Goga, J. Pavlovicova, M. Oravec, and B. Jansen, "A survey of iris datasets," *Image Vis. Comput.*, vol. 108, Apr. 2021, Art. no. 104109. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0262885621000147
[110] W. V. Univ. (2019). *Twins Day Dataset 2010–2015*. Accessed: Jan. 5, 2024. [Online]. Available: https://biic.wvu.edu/data-sets/twins-day-dataset-2010-1015
[111] K. Hollingsworth, K. W. Bowyer, and P. J. Flynn, "Similarity of iris texture between identical twins," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, Jun. 2010, pp. 22–29.
[112] P. J. Phillips, P. J. Flynn, K. W. Bowyer, R. W. V. Bruegge, P. J. Grother, G. W. Quinn, and M. Pruitt, "Distinguishing identical twins by face recognition," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Mar. 2011, pp. 185–192.
[113] B. S. Laboratory. (2025). *FVC-onGoing—BOEP*. Accessed: May 3, 2025. [Online]. Available: https://biolab.csr.unibo.it/fvcongoing/UI/Form/BOEP.aspx
[114] B. S. Lab. *FVC-onGoing*. Accessed: May 3, 2025. [Online]. Available: https://biolab.csr.unibo.it/FvcOnGoing/UI/Form/Home.aspx
[115] B. Dorizzi, R. Cappelli, M. Ferrara, D. Maio, D. Maltoni, N. Houmani, S. Garcia-Salicetti, and A. Mayoue, "Fingerprint and on-line signature verification competitions at ICB 2009," in *Advances in Biometrics*, M. Tistarelli and M. S. Nixon, Eds., Cham, Germany: Springer, 2009, pp. 725–732.
[116] NIST. *Face Analysis Technology Evaluation (FATE) MORPH*. Accessed: May 3, 2025. [Online]. Available: https://pages.nist.gov/frvt/html/frvt_morph.html
[117] NIST. *State of the Art of Morphing Detection (SOTAMD)*. Accessed: May 3, 2025. [Online]. Available: https://www.ntnu.edu/iik/sotamd
[118] B. S. Laboratory. (2025). *FVC-onGoing—Background*. Accessed: May 3, 2025. [Online]. Available: https://biolab.csr.unibo.it/FvcOnGoing/UI/Form/Background.aspx
[119] H. O. Shahreza and S. Marcel, "Unveiling synthetic faces: How synthetic datasets can expose real identities," 2024, *arXiv:2410.24015*.
[120] S. Zhou, K. C. Chan, C. Li, and C. C. Loy, "Towards robust blind face restoration with codebook lookup transformer," in *Proc. NeurIPS*, Aug. 2022, pp. 1–12.
[121] N. Di Domenico, G. Borghi, A. Franco, and D. Maltoni, "Face restoration for morphed images retouching," in *Proc. 12th Int. Workshop Biometrics Forensics (IWBF)*, Apr. 2024, pp. 1–6.
[122] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2030–2096, Jan. 2016.

[123] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.
[124] Z. Wang, K. Qinami, I. C. Karakozis, K. Genova, P. Nair, K. Hata, and O. Russakovsky, "Towards fairness in visual recognition: Effective strategies for bias mitigation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8916–8925.
[125] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Proc. Equity Access Algorithms, Mech., Optim.* New York, NY, USA: ACM, Oct. 2021, pp. 1–9, doi: 10.1145/3465416.3483305.
[126] Y. Li, Q. Li, H. Wang, R. Li, W. Zhong, and G. Zhang, "Towards efficient replay in federated incremental learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 12820–12829.
[127] G. Borghi, A. Franco, N. di Domenico, M. Ferrara, and D. Maltoni, "V-MAD: Video-based morphing attack detection in operational scenarios," in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2024, pp. 1–10.
[128] R. Bommasani et al., "On the opportunities and risks of foundation models," 2021, *arXiv:2108.07258*.

PEDRO MORAIS received the B.S. degree in informatics engineering from the Polytechnic University of Coimbra, in 2023, where he is currently pursuing the M.S. degree. His research interests include biometric systems vulnerabilities to morphing attacks and the exploration of cross-domain detection techniques to enhance biometric data integrity and security.

INÊS DOMINGUES received the bachelor's degree in applied mathematics from the University of Porto, in 2004, the master's degree in electrical and telecommunications engineering from Aveiro University, in 2008, and the Ph.D. degree (cum laude) in electrical and computer engineering from the University of Porto, in 2015. She is currently an Assistant Professor with ISEC and a Researcher with CI-IPOP. Her research interests include oncologically through image processing, pattern recognition, and machine learning.

JORGE BERNARDINO (Member, IEEE) received the Ph.D. degree from the University of Coimbra, in 2002. From 2005 to 2010, he was the President of the Coimbra Engineering Institute (ISEC). From 2017 to 2019, he was also the President of the ISEC Scientific Council. In 2014, he was a Visiting Professor with CMU. He was the Director of the Applied Research Institute (i2A), Polytechnic University of Coimbra, from 2019 to 2021. He is currently a Coordinator and a Professor with the Polytechnic University of Coimbra-ISEC, Portugal. He has authored more than 200 publications in refereed conferences and journals and participated in several national and international projects. His research interests include big data, NoSQL, data warehousing, dependability, the Internet of Things, and software engineering.

Annex B - ISPA International Conference Paper

Composite Heatmaps: Interpreting Deep Models in Morph Attack Detection

Pedro Morais
Coimbra Institute of Engineering,
Polytechnic University of Coimbra
Rua da Misericórdia, Lagar dos Cortiços,
S. Martinho do Bispo, 3045-093 Coimbra, Portugal
a21280686@isec.pt

Inês Domingues
Coimbra Institute of Engineering,
Polytechnic University of Coimbra
Research Center of the Portuguese Institute of Oncology of Porto
Portugal
inesdomingues@gmail.com

2025 14th International Symposium on Image and Signal Processing and Analysis (ISPA) | 979-8-3315-7755-1/25/$31.00 ©2025 IEEE | DOI: 10.1109/ISPA66905.2025.11259485

***Abstract*—Interpretability is essential in security-critical applications such as Morph Attack Detection (MAD), where model transparency supports trust, auditability, and bias assessment. This work investigates the use of occlusion mapping to explain the decision process of a Vision Transformer (ViT)–based MAD system combined with a Support Vector Machine (SVM) classifier. Two complementary occlusion strategies were explored: landmark-based occlusion, targeting semantically relevant facial regions (eyes, nose, mouth), and grid-based occlusion, dividing images into spatial cells. Experiments were conducted on cropped and aligned versions of the FRLL-Morphs and FFHQ-Morphs datasets. The results, visualized through heatmaps, reveal that the model consistently assigns highest importance to the nose and mouth regions, with strong sensitivity to depth transition areas prone to morphing artifacts. These findings provide actionable insights for improving the robustness and bias mitigation of MAD systems.**

***Index Terms*—Model interpretability, occlusion mapping, heatmap visualization, morph attack detection, biometric security**

I. Introduction

The interpretability and explainability of deep learning models have become essential research areas [1], [2], especially in security-critical domains such as biometric recognition and Morph Attack Detection (MAD). Deep neural networks achieve high accuracy but often behave as ”black boxes”, making it difficult to understand the rationale behind their decisions. This opacity limits trust, auditability, and the identification of systematic biases or failure modes in safety-sensitive applications [3], [4].

Understanding which facial features influence a model's decision is particularly relevant in MAD, where subtle manipulations can fool both humans and automated systems [5].

Several interpretability methods have been proposed to better understand deep learning models, including gradient-based approaches (e.g., SmoothGrad [6]), feature attribution frameworks (e.g., SHAP [4]), and perturbation-based methods [7]. Among the latter, occlusion-based methods are well-suited for MAD because they systematically mask input regions, allowing the identification of facial features that are critical for classification. Landmark-based occlusion focuses on semantically meaningful facial points, while grid-based occlusion evaluates the contribution of uniform spatial patches across the face. Together, these approaches provide complementary insights into model behavior.

Recent research demonstrates the importance of interpretability techniques in MAD. Borghi et al. (2021) [3] applied Grad-CAM [8] to visualize relevant face regions, while more recent works explore pixel-level explanations and multimodal approaches [9], [10]. Occlusion-based analysis extends these insights by quantitatively assessing which areas of the face drive the model's decision, improving transparency and supporting more robust, bias-aware morph detection.

The objective of this paper is to demonstrate the use of landmark-based and grid-based occlusion techniques to interpret deep learning models for MAD. By systematically perturbing input images, this work identifies the facial characteristics that allow the model to distinguish between bona fide and morphed images, providing both visual and quantitative explanations.

Given the increasing reliance on automated biometric systems, this paper addresses the following research question:

RQ: How can landmark-based and grid-based occlusion techniques improve understanding of the facial characteristics that drive deep learning-based MAD?

The key contributions of this paper are as follows:

- Demonstrating the application of landmark-based and grid-based occlusion methods for MAD interpretability.
- Proposing a novel strategy that combines the results of landmark-based and grid-based occlusion approaches into *composite heatmaps*.
- Identifying facial characteristics that are most influential for distinguishing bona fide images from morphs.
- Providing insights that can guide the development of more robust and explainable MAD models.

The structure of the paper is as follows: Section II reviews recent studies on MAD interpretability. Section III details the methodology used for applying landmark-based and grid-based occlusion techniques. Section IV presents the experimental setup and evaluation protocol. Section V presents and discusses the results of the occlusion analyses. Finally, Section VI presents the main conclusions of the paper.

Authorized licensed use limited to: b.on Instituto Politecnico de Coimbra. Downloaded on December 07,2025 at 16:20:36 UTC from IEEE Xplore. Restrictions apply.

II. Related Work

Several studies have emphasized the value of interpretability techniques for explaining model behavior and ensuring responsible deployment [1], [2], [8]. Widely used methods include gradient-based techniques such as SmoothGrad [6], feature attribution frameworks such as SHAP [4], and perturbation-based analysis [7], each offering complementary insights into model decision-making.

An early and influential contribution to explainable MAD was presented by Seibold et al. (2021) [9], who introduced the Focused Layer-wise Relevance Propagation (FLRP) method. This approach produced pixel-level attribution maps that highlight the regions most critical to a model's decision, enabling researchers to assess whether the network relies on meaningful facial characteristics or non-informative image areas. This study set the foundation for explainable MAD by showing that interpretability can be systematically integrated into model evaluation.

Caldeira et al. (2023) [11] introduced IDistill, a domain-informed autoencoder that separates individual identity contributions in a morph. Their approach not only achieved competitive detection metrics but also provided interpretable scores representing the degree of identity separation, offering an inherently explainable decision process.

Dwivedi et al. (2024) [12] proposed an ensemble explainable AI (XAI) approach combining saliency maps, Class Activation Maps (CAM) [13], and Grad-CAM [8]. This integration provided more comprehensive and consistent explanations compared to single-method visualizations. This approach highlighted the benefits of combining multiple interpretability methods, demonstrating how ensemble XAI can offer deeper insights into MAD decisions.

A more recent trend explores multimodal interpretability. Patwardhan et al. (2026) [10] employed CLIP-based models [14] to generate not only visual but also textual explanations for MAD decisions. This allows the system to justify predictions in natural language, improving usability for non-expert stakeholders such as border security officers. By moving beyond purely visual heatmaps, this work broadened the scope of interpretability to human-centered explanations.

In contrast to prior works, the methodology presented in this paper prioritizes simplicity and practicality, emphasizing techniques that are straightforward to implement and interpret while retaining relevance for operational deployment. Rather than relying on complex attribution models, ensemble strategies, or multimodal frameworks, it adopts a lightweight yet effective approach that integrates grid-based and landmark-based occlusion into composite heatmaps. This design minimizes computational overhead, facilitates seamless integration into existing MAD pipelines, and produces intuitive visual representations that can be readily understood by practitioners and decision-makers, thereby bridging the gap between advanced interpretability methods and practical biometric security requirements.

For more details, the review by Morais et al. (2025) [15] consolidates existing deep learning techniques for detecting morphed face images, providing a comprehensive overview of methodologies, challenges, and evaluation practices.

III. Methodology

This study employs occlusion mapping to interpret the decision-making process of a Vision Transformer (ViT)–based MAD model combined with a Support Vector Machine (SVM) classifier. Occlusion mapping is a perturbation-based interpretability method in which regions of the input image are systematically masked, and the resulting changes in model output are measured. A substantial reduction in prediction score after occluding a region indicates that the model relied on that area for its decision [7].

Two complementary occlusion strategies were implemented: **landmark-based occlusion**, targeting semantically meaningful facial regions (specified in Section III-A), and **grid-based occlusion**, enabling fine-grained spatial analysis. The combination of these methods allows both semantic and spatial importance to be assessed.

Figure 1 summarizes the entire occlusion analysis pipeline, from the generation of occluded image variants to the fusion of their results into a final interpretability heatmap.

A. Landmark-Based Occlusion

Facial landmarks were automatically detected for each test image, with four key regions selected for analysis: (1) Left eye, (2) Right eye, (3) Nose, and (4) Mouth.

Each region was occluded individually using a black rectangle while preserving the rest of the image. This was implemented with dlib's 68-point face landmark predictor [16], [17], which identifies the four key facial regions. To ensure complete coverage, the occlusion area was expanded by a factor of 1.5 relative to the detected landmark boundaries. The occluded image was then processed by the trained model, and its prediction score was recorded. This process produced a prediction score for each facial component, enabling the estimation of its relative contribution to the classification outcome. These prediction scores were aggregated into a heatmap across all landmarks, providing a spatial importance map that offers a fine-grained view of the regions most influential to the classifier.

B. Grid-Based Occlusion

To capture spatial dependencies beyond predefined facial features, a 4×4 uniform grid was overlaid on each image, dividing it into 16 non-overlapping cells of equal size. Each cell was occluded in turn, and the corresponding prediction score was measured. The resulting prediction scores were assembled into a spatial importance map, providing a fine-grained view of the regions most influential to the classifier.

C. Composite Heatmaps

For interpretability, a novel strategy was developed that combines the results of both occlusion approaches into *composite heatmaps*:

- For the grid-based occlusion, the initial 4×4 grid was subdivided into an 8×8 grid to increase the resolution

Authorized licensed use limited to: b:on Instituto Politecnico de Coimbra. Downloaded on December 07,2025 at 16:20:36 UTC from IEEE Xplore. Restrictions apply.

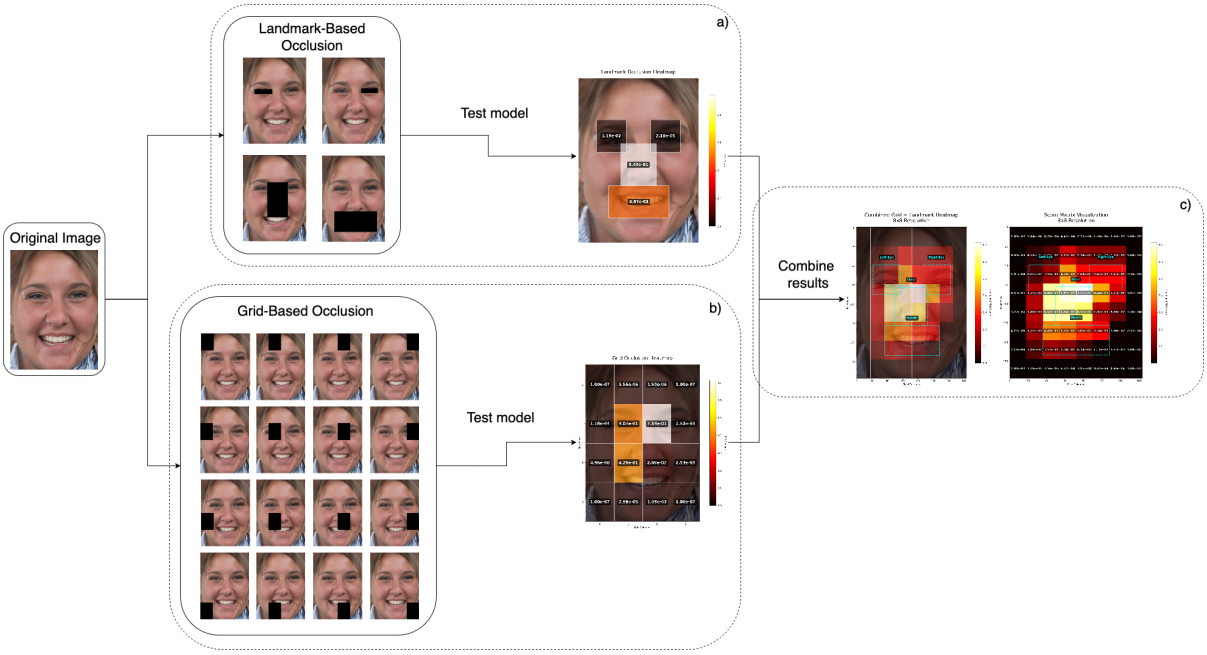


Fig. 1: Overview of the occlusion mapping process. The original cropped image is used to generate multiple occluded variants via (a) landmark-based occlusion and (b) grid-based occlusion. Each occluded image is processed by the trained MAD model, and the resulting prediction scores are recorded. The importance maps from both strategies are fused (c), producing the final *composite heatmap*, which is overlayed on the original cropped image for easier visualization.

of the occlusion map. While the finer grid alone does not add new information, since each cell initially inherits the value from the coarser grid, the higher resolution becomes meaningful after fusion with landmark-based prediction score, allowing the heatmap to capture fine-grained spatial details corresponding to semantically important regions.

- Grid-based prediction scores were then normalized and combined with landmark-based prediction scores using averaging, preserving both semantic and spatial relevance in the resulting heatmap.

In the resulting heatmaps, regions with higher importance are shown in lighter colors, while areas with minimal influence appear darker. For visual inspection, these heatmaps were overlaid on the original cropped images. In addition, the underlying score matrices, which contain the numerical prediction values for each grid cell or landmark, were retained for quantitative analysis and comparison.

IV. Experiments

This section describes the experiments perfomed to assess the proposed occlusion mapping approach for MAD. The experiments aim to interpret and explain the model's decision-making process by revealing which facial characteristics contribute most to morph detection. To this end, occlusion mapping and heatmap visualization were employed to investigate: (i) the model's robustness to partial facial occlusions, and (ii) the specific facial regions and characteristics that the model considers most important.

A. Datasets

Two publicly available datasets were used in the experiments: **FFHQ-Morphs** [18] and **FRLL-Morphs** [19]–[21]. In addition to the original images, **cropped variants**, *FFHQ-Cropped* and *FRLL-Cropped*, were created by aligning, centering, and cropping the facial region while removing background elements. This preprocessing step aimed to mitigate background bias and focus the analysis exclusively on facial content, as illustrated in Figure 2.

Fig. 2: Example of the cropping procedure.

Occlusion analysis was performed exclusively on the cropped datasets. This decision was based on the more consistent alignment and reduced background variation provided by cropped images, which made it easier to isolate the impact of occluding specific facial regions.

B. Model and Training

The MAD system consisted of a ViT backbone pre-trained on the ImageNet [22] dataset for feature extraction and a SVM

Authorized licensed use limited to: b:on Instituto Politecnico de Coimbra. Downloaded on December 07,2025 at 16:20:36 UTC from IEEE Xplore. Restrictions apply.

classifier for final decision-making. The ViT was fine-tuned on the training data, and the resulting feature embeddings were subsequently used to train the SVM.

A four-configuration cross-dataset evaluation protocol was employed to assess both intra-dataset and cross-dataset generalization. The configurations were as follows:

1) Train on *FFHQ-Cropped*.
   a) Test on *FFHQ-Cropped* (intra-dataset).
   b) Test on *FRLL-Cropped* (cross-dataset).
2) Train on *FRLL-Cropped*.
   a) Test on *FFHQ-Cropped* (cross-dataset).
   b) Test on *FRLL-Cropped* (intra-dataset).

This setup enables the assessment of the model's generalization capability to previously unseen data, both within the same dataset and across distinct datasets.

All experiments were conducted on Google Colab with an NVIDIA A100 GPU. The full implementation is publicly available online on GitHub[1] to facilitate reproducibility.

C. Occlusion-Based Evaluation

To interpret the model's decision-making and assess its robustness to partial occlusions, a random sample of 25 test images was drawn from the corresponding test dataset. For every image, occlusion perturbations were applied (Section III), and the previously trained model was then used to classify the perturbed images and generate prediction scores. Once the heatmaps for the individual occlusion strategies were obtained, they were merged into *composite heatmaps* by averaging the prediction scores. The resulting visualizations exposed both spatial and semantic patterns of feature importance, highlighting differences in model focus across datasets and training/testing scenarios.

D. Evaluation Metrics

Performance was assessed using the following metrics:

- **Heatmap Visualization:** Visual representation of occlusion sensitivity. Lighter colors indicate regions that have a stronger influence on the model's classification decision.
- **Prediction Score:** The model's confidence in classifying an image, represented as a continuous value between 0 and 1, where values closer to 0 indicate higher confidence in the image being a morph, and values closer to 1 indicate higher confidence in it being bona fide.

V. Results

This section presents the findings of the occlusion-based analysis, which was conducted to interpret the model's focus on different facial regions during MAD. Heatmaps were generated to visualize the areas that most influenced the model's decisions. The visualization followed a four-stage procedure:

1) The original input image was first displayed without any occlusion, serving as a reference for comparison.
2) A heatmap based on **landmark-based occlusion** was generated. Key facial regions were occluded individually, and the resulting classification performance was recorded.
3) A heatmap based on **grid-based occlusion** was produced. Each image was divided into a 4×4 grid, and classification performance was measured as each cell was occluded individually, producing a spatial importance map.
4) A combined heatmap incorporating both grid-based and landmark-based prediction scores was generated. This visualization captures both spatial and semantic feature importance in a single map. Two different heatmap visualizations are created: one overlaid on the original cropped image, and another with the value labels.

Figure 3 illustrates this process using image `00607_37173` from the *FFHQ-Morphs* dataset, generated with a diffusion-based morphing algorithm and pre-processed via face cropping. The visualizations expose the model's learned spatial patterns and the importance of individual regions.

Across all four configurations, with the exception of the model trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*, heatmap analysis reveals that the model focuses primarily on the **nose** and **mouth** regions, with moderate attention to the **eyes** (see Figure 3d for an example). Across multiple samples, areas exhibiting **depth transitions**, such as the boundary between the nose and cheeks and the corners of the eyes, were consistently highlighted, reflecting regions prone to morphing artifacts caused by geometric blending errors. Landmark-based occlusion further corroborates these findings: occluding the nose or mouth produced the largest drops in model confidence, confirming the semantic significance of these facial features in morph detection.

For the model trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*, no consistent trend is observed in the heatmaps. This may be due to differences in the types of morphs, image characteristics (e.g., resolution, lighting, cropping), or subject diversity between the training and test datasets. Figure 4 presents an example of the heatmap produced by this model.

These results demonstrate that, depending on the model and datasets used, it is possible to understand the facial features that contribute most to the prediction results through landmark-based and grid-based occlusion techniques, thus improving interpretability.

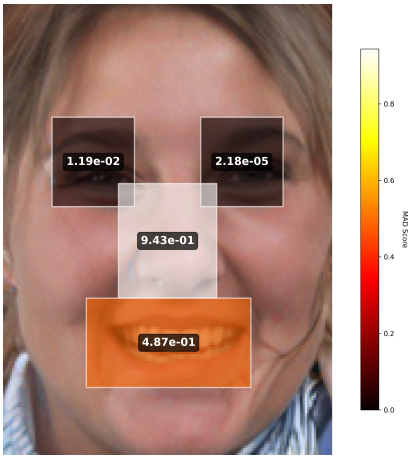
VI. Conclusions

This paper investigated the use of occlusion-based interpretability techniques to analyze the decision-making process of a ViT–based MAD system with an SVM classifier. By employing both landmark-based and grid-based occlusion strategies, this study provided complementary insights into the spatial and semantic relevance of facial characteristics for morph detection.

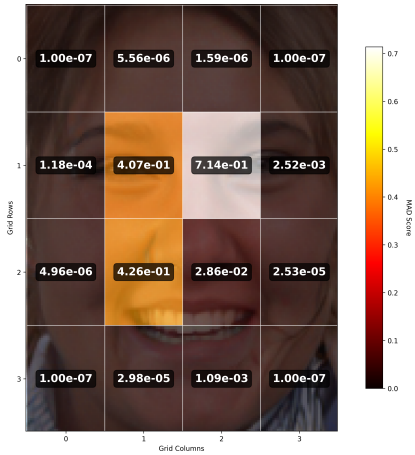
[1] https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Experiments/5-ViT_SVM-OcclusionMapping

Authorized licensed use limited to: b:on Instituto Politecnico de Coimbra. Downloaded on December 07,2025 at 16:20:36 UTC from IEEE Xplore. Restrictions apply.

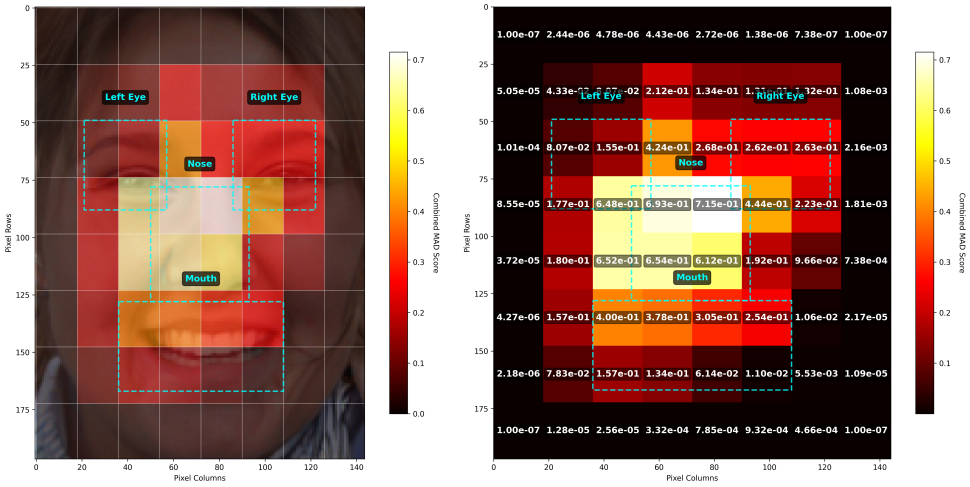
(a) Original cropped image.



(b) Landmark occlusion heatmap.



(c) Grid occlusion heatmap (4×4).



(d) Combined grid + landmark heatmap (8×8). Left: heatmap overlaid on the original cropped image for intuitive visualization; Right: standalone heatmap with numerical prediction values, providing a precise quantification of regional importance.

Fig. 3: Occlusion heatmaps for image `00607_37173` (diffusion-based morph, face-cropped), produced by the model trained on *FFHQ-Cropped* and tested on *FFHQ-Cropped*. Lighter regions indicate greater importance in the model's decision-making.

The experimental results demonstrated that the MAD model consistently assigns high importance to the nose and mouth regions, while eye regions were moderately influential. Grid-based occlusion revealed additional spatial patterns, highlighting areas around depth transitions and morphing artifacts that are critical for classification.

When comparing the occlusion strategies, landmark-only perturbations provided a focused assessment of key facial landmarks, highlighting regions such as the nose and mouth but offering limited spatial coverage. Grid-only occlusion, implemented with a 4×4 grid, captured broader spatial patterns, revealing additional areas associated with morphing artifacts and depth transitions that landmarks alone might miss. The fusion of landmark- and grid-based heatmaps yielded the most comprehensive interpretation, combining the precision of landmark analysis with the extended coverage of grid occlusion to produce high-resolution visualizations of both local and distributed features. While grid occlusion could be further refined by using a finer grid, this would substantially increase disk space requirements, as storage demands grow linearly with the number of grid cells.

These findings have several implications for future MAD research and deployment. First, they provide evidence that targeted occlusion mapping can identify the facial regions most relied upon by deep learning models, offering a means to audit and improve robustness. Second, the approach can guide the design of more explainable and reliable MAD systems by focusing on critical facial features while mitigating potential biases introduced by less relevant regions. Finally, the methodology can be extended to other biometric security applications where interpretability is essential for trust and accountability.

Authorized licensed use limited to: b:on Instituto Politecnico de Coimbra. Downloaded on December 07,2025 at 16:20:36 UTC from IEEE Xplore. Restrictions apply.

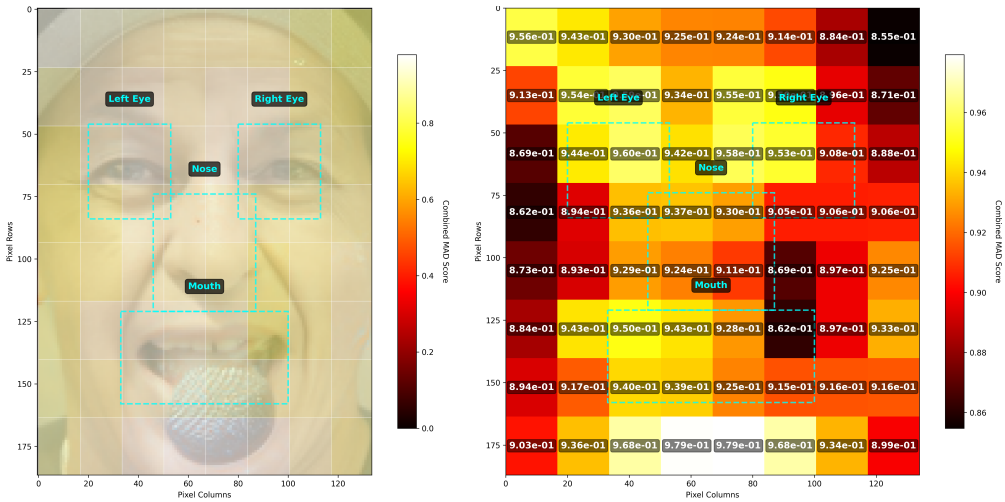


Fig. 4: Example heatmap produced by the model trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*.

In summary, this work confirms that occlusion-based analysis is an effective tool for explaining ViT–SVM morph detection models, providing both visual and quantitative insights that support robust, transparent, and trustworthy biometric security systems.

Acknowledgment

The authors acknowledge the use of OpenAI's ChatGPT to assist with grammar and language refinement during the preparation of this manuscript. No AI-generated content was used for technical analysis, literature synthesis, or figure creation. All scientific interpretations and conclusions are the sole responsibility of the authors.

References

[1] W. Samek, T. Wiegand, and K.-R. Müller, "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models," *ITU Journal: ICT Discoveries - Special Issue 1 - The Impact of Artificial Intelligence (AI) on Communication Networks and Services*, vol. 1, pp. 39–48, 2018.

[2] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017.

[3] G. Borghi, E. Pancisi, M. Ferrara, and D. Maltoni, "A double siamese framework for differential morphing attack detection," *Sensors*, vol. 21, 2021.

[4] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *31st International Conference on Neural Information Processing Systems*, ser. NIPS. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.

[5] S. R. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, R. Ramachandra, and C. Busch, "Analyzing human observer ability in morphing attack detection—where do we stand?" *IEEE Transactions on Technology and Society*, vol. 4, no. 2, pp. 125–145, 2023.

[6] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg, "SmoothGrad: removing noise by adding noise," *CoRR*, vol. abs/1706.03825, 2017.

[7] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Müller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 28, no. 11, pp. 2660–2673, 2017.

[8] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626.

[9] C. Seibold, A. Hilsmann, and P. Eisert, "Focused LRP: Explainable AI for face morphing attack detection," in *IEEE Winter Conference on Applications of Computer Vision Workshops (WACVW)*, 2021, pp. 88–96.

[10] S. Patwardhan, R. Ramachandra, and S. Venkatesh, "Empowering morphing attack detection using interpretable image-text foundation model," in *Computer Vision and Image Processing*, J. Kakarla, R. Balasubramanian, S. Murala, S. K. Vipparthi, and D. Gupta, Eds. Cham: Springer Nature Switzerland, 2026, pp. 190–205.

[11] E. Caldeira, P. C. Neto, T. Gonçalves, N. Damer, A. F. Sequeira, and J. S. Cardoso, "Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection," in *31st European Signal Processing Conference (EUSIPCO)*, 2023, pp. 955–959.

[12] R. Dwivedi, P. Kothari, D. Chopra, M. Singh, and R. Kumar, "An efficient ensemble explainable AI (XAI) approach for morphed face detection," *Pattern Recognition Letters*, vol. 184, pp. 197–204, 2024.

[13] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 2921–2929.

[14] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *38th International Conference on Machine Learning*, ser. Machine Learning Research, vol. 139. PMLR, 2021, pp. 8748–8763.

[15] P. Morais, I. Domingues, and J. Bernardino, "Deep learning techniques for detecting morphed face images: A literature review," *IEEE Access*, vol. 13, pp. 105 952–105 981, 2025.

[16] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[17] D. E. King, "dlib c++ library - face landmark detection example," 2009, accessed: 2025-07-09. [Online]. Available: https://dlib.net/face_landmark_detection.py.html

[18] L. Colbois and S. Marcel, "Evaluating the effectiveness of attack-agnostic features for morphing attack detection," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2024, pp. 1–9.

[19] I. R. Institute, "FRLL-Morphs dataset," accessed: 2024-12-26. [Online]. Available: https://www.idiap.ch/en/scientific-research/data/frll-morphs

[20] E. Sarkar, P. Korshunov, L. Colbois, and S. Marcel, "Vulnerability analysis of face morphing attacks from landmarks and generative adversarial networks," 2020, arxiv:2012.05344.

[21] ——, "Are GAN-based morphs threatening face recognition," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2959–2963, 2022.

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

Authorized licensed use limited to: b:on Instituto Politecnico de Coimbra. Downloaded on December 07,2025 at 16:20:36 UTC from IEEE Xplore. Restrictions apply.

Annex C - RecPad Paper

Interpreting Morph Attack Detection with Composite Heatmaps

Pedro Morais
a21280686@isec.pt

Inês Domingues
inesdomingues@gmail.com

Polytechnic University of Coimbra, Rua da Misericórdia, Lagar dos Cortiços, S. Martinho do Bispo, 3045-093 Coimbra, Portugal.

Research Center of the Portuguese Institute of Oncology of Porto (CI-IPOP), Medical Physics, Radiobiology and Radiological Protection Group,

Abstract

Interpretability is essential in security-critical applications such as Morph Attack Detection (MAD), where model transparency supports trust, auditability, and bias assessment. This work investigates the use of occlusion mapping to explain the decision process of a Vision Transformer (ViT)–based MAD system combined with a Support Vector Machine classifier. Two complementary occlusion strategies were explored: landmark-based occlusion, targeting semantically relevant facial regions (eyes, nose, mouth), and grid-based occlusion, dividing images into a grid. Experiments were conducted on cropped and aligned versions of the FRLL-Morphs and FFHQ-Morphs datasets. The results, visualized through heatmaps, reveal that the model consistently assigns highest importance to the nose and mouth regions, with strong sensitivity to depth transition areas prone to morphing artifacts. These findings provide actionable insights for improving the robustness and bias mitigation of MAD systems.

1 Introduction

Deep learning models achieve strong performance in biometric security tasks, but their decision-making often remains opaque. In Morph Attack Detection (MAD), this lack of transparency limits trust, restricts audits, and makes it harder to detect biases. Identifying the facial regions most influential in classification is crucial because morphs involve subtle changes that can deceive both humans and automated systems [5].

Existing interpretability techniques for deep learning models include gradient-based approaches (e.g., SmoothGrad [15]), feature attribution (e.g., SHAP [9]), and perturbation-based methods. Among the latter, occlusion mapping is particularly suitable for MAD, as it systematically masks regions to reveal the importance of spatial or semantic facial features. Prior work has explored saliency maps, Grad-CAM [14], and even multimodal explanations, but these often increase complexity and computational cost.

This paper proposes *Composite Heatmaps* [1] a practical and lightweight approach that combines **landmark-based occlusion**, targeting meaningful areas (eyes, nose, mouth), and **grid-based occlusion**, providing broader spatial coverage. The resulting heatmaps offer interpretable visualizations of model focus, supporting the development of more reliable and bias-aware MAD systems.

2 Related Work

Interpretability methods have been widely studied to explain deep learning models, especially in safety-critical domains. Gradient-based techniques such as SmoothGrad [15], attribution methods like SHAP [9], and perturbation-based approaches [12] provide different perspectives on model behavior.

In the context of MAD, Seibold et al. [13] introduced Focused Layer-wise Relevance Propagation (FLRP) for pixel-level explanations. Caldeira et al. [1] proposed IDistill, a domain-informed autoencoder that outputs interpretable identity separation scores. Dwivedi et al. [4] explored ensemble explainable AI (XAI) by combining saliency maps, CAM [16], and Grad-CAM [14]. Patwardhan et al. [10] extended interpretability to natural language explanations using CLIP-based models [11].

While effective, many of these approaches are complex and computationally demanding. This work focuses on a **lightweight and practical strategy** that integrates semantic (landmark) and spatial (grid) occlusion techniques, producing *composite heatmaps* that are intuitive, reproducible, and suitable for operational MAD systems.

[1] An extended version of this work has been submitted to the 14th Int'l Symposium on Image and Signal Processing and Analysis (ISPA 2025) under the title Composite Heatmaps: Interpreting Deep Models in Morph Attack Detection

3 Methodology

This study employs occlusion mapping to interpret the decision-making of a Vision Transformer (ViT)–based MAD system combined with a Support Vector Machine (SVM) classifier. Occlusion mapping is a perturbation-based interpretability technique in which selected regions of an input image are systematically masked, and the resulting changes in the prediction score are measured. A significant drop in confidence after occlusion indicates that the region is important for classification.

Figure 1 summarizes the complete analysis pipeline, from image preprocessing to the generation of *composite heatmaps*. First, input images are cropped and aligned to reduce background variation. Two complementary occlusion strategies are then applied:

- **Landmark-based occlusion:** Facial landmarks (left eye, right eye, nose, and mouth) are detected automatically using dlib's 68-point predictor [7, 8]. Each region is occluded with a black rectangle while preserving the rest of the image. The model's prediction score is recorded for each occlusion, allowing an assessment of the relative importance of these semantically meaningful features.

- **Grid-based occlusion:** To capture information beyond predefined landmarks, a uniform 4×4 grid divides the facial image into 16 equal cells. Each cell is occluded sequentially, and the classification score is measured. This method provides a spatial importance map that can highlight regions not covered by landmarks.

Composite heatmaps: The outputs from both strategies are fused to create a more detailed view. The grid is further subdivided into an 8×8 resolution and combined with landmark-based scores via averaging. The resulting heatmaps highlight both semantic and spatial features, where lighter colors indicate regions of higher importance.

These visualizations, along with the underlying numerical scores, provide interpretable insights into which facial areas most influence the model's decisions, supporting bias analysis and model improvement.

4 Experiments

To evaluate the proposed approach, two public datasets were used: **FFHQ-Morphs** [2] and **FRLL-Morphs** [6]. The MAD system consists of a ViT pre-trained on ImageNet [3] for feature extraction, followed by a SVM classifier. Four configurations were considered: each dataset was used for training and tested both on itself (intra-dataset) and on the other dataset (cross-dataset). This setup allowed assessment of model performance and generalization.

For each configuration, 25 images were randomly selected from the test set. Landmark-based and grid-based occlusions (Section 3) were applied to generate perturbed variants, which were classified by the trained module to generate prediction scores. These results were merged into *composite heatmaps* to visualize the most influential facial regions.

Evaluation relied on two main outputs:

- **Heatmaps** highlighting the regions most critical to the classification.

- **Prediction scores** quantifying the model's confidence in classifying an image, represented as a continuous value between 0 and 1, where values closer to 0 indicate higher confidence in the image

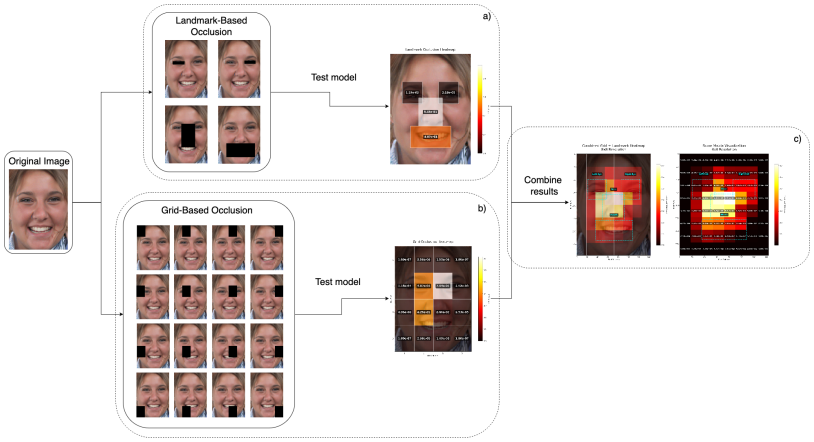


Figure 1: Overview of the occlusion mapping process: (a) landmark-based occlusion, (b) grid-based occlusion, and (c) fusion into a composite heatmap.

being a morph, and values closer to 1 indicate higher confidence in it being bona fide.

All experiments were run on Google Colab with an NVIDIA A100 GPU. The full implementation is available on GitHub[2].

5 Results

Occlusion analysis revealed consistent patterns on all four configurations, with the exception of the model trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*. The ViT–SVM system focused primarily on the **nose** and **mouth**, with moderate attention to the **eyes**. Depth-transition regions (e.g., nose–cheek boundary, eye corners) were frequently highlighted in light colors, reflecting areas prone to morphing artifacts.

Landmark occlusions confirmed the semantic importance of these features, as masking the nose or mouth caused the largest drop in prediction score. Grid-based occlusion complemented this by exposing broader spatial patterns, and their fusion into *composite heatmaps* provided the most comprehensive visualization.

A different behavior was observed when the model was trained on *FRLL-Cropped* and tested on *FFHQ-Cropped*: heatmaps showed no consistent focus, likely due to differences in the types of morphs, image characteristics (e.g., resolution, lighting, cropping), or subject diversity between the training and test datasets.

6 Conclusions

This study explored occlusion-based interpretability techniques to analyze a ViT–SVM MAD system, using both landmark- and grid-based strategies. Results showed that the model consistently prioritizes the nose and mouth, with moderate influence from the eyes, while grid occlusion revealed additional spatial patterns linked to morphing artifacts and depth transitions. Combining landmark- and grid-based heatmaps provided the most comprehensive interpretation, merging precise landmark insights with broader spatial coverage. A finer grid would yield more precise results, but it would also increase storage requirements.

These findings demonstrate that occlusion mapping can identify critical facial regions, supporting the design of more explainable and robust MAD systems. Future research could investigate whether this approach can be extended to other biometric security tasks where interpretability is crucial, potentially providing both visual and quantitative insights that enhance transparency, reliability, and trustworthiness in biometric systems.

References

[1] E. Caldeira, P. C. Neto, T. Gonçalves, N. Damer, A. F. Sequeira, and J. S. Cardoso. Unveiling the two-faced truth: Disentangling morphed identities for face morphing detection. In *31st European Signal Processing Conf (EUSIPCO)*, pages 955–959, 2023.

[2] https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Experiments/5-ViT_SVM-OcclusionMapping

[2] L. Colbois and S. Marcel. Evaluating the effectiveness of attack-agnostic features for morphing attack detection. In *IEEE International Joint Conf on Biometrics (IJCB)*, pages 1–9, 2024.

[3] J. Deng, W. Dong, R. Socher, LJ Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conf on Comput Vision and Pattern Recogn*, pages 248–255, 2009.

[4] R. Dwivedi, P. Kothari, D. Chopra, M. Singh, and R. Kumar. An efficient ensemble explainable AI (XAI) approach for morphed face detection. *Pattern Recogn Letters*, 184:197–204, 2024.

[5] S.R. Godage, F. Løvåsdal, S. Venkatesh, K. Raja, R. Ramachandra, and C. Busch. Analyzing human observer ability in morphing attack detection—where do we stand? *IEEE Trans. Technol. Soc.*, 4(2): 125–145, 2023.

[6] Idiap Research Institute. FRLL-Morphs dataset. URL https://www.idiap.ch/en/scientific-research/data/frll-morphs. Accessed: 2024-12-26.

[7] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *IEEE Conf on Comput Vision and Pattern Recogn*, pages 1867–1874, 2014.

[8] Davis E. King. dlib c++ library - face landmark detection example, 2009. URL https://dlib.net/face_landmark_detection.py.html. Accessed: 2025-07-09.

[9] S. M. Lundberg and SI Lee. A unified approach to interpreting model predictions. In *31st International Conf on Neural Information Processing Systems (NIPS)*, page 4768–4777, 2017.

[10] S. Patwardhan, R. Ramachandra, and S. Venkatesh. Empowering morphing attack detection using interpretable image-text foundation model. In *Comput Vision and Image Processing*, pages 190–205. Springer Nature Switzerland, 2026.

[11] A. Radford, JW Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In *38th International Conf on Machine Learning*, volume 139, pages 8748–8763, 2021.

[12] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and KR Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE Trans Neural Net Learn Syst*, 28(11):2660–73, 2017.

[13] C. Seibold, A. Hilsmann, and P. Eisert. Focused LRP: Explainable AI for face morphing attack detection. In *IEEE Winter Conf on Appl of Comput Vision Workshops (WACVW)*, pages 88–96, 2021.

[14] RR Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual expla.tions from deep networks via gradient-based localization. In *IEEE International Conf on Comput Vision (ICCV)*, pages 618–626, 2017.

[15] D. Smilkov, N. Thorat, B. Kim, FB Viégas, and M. Wattenberg. SmoothGrad: removing noise by adding noise. *CoRR*, abs/1706.03825, 2017.

[16] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *IEEE Conf on Comput Vision and Pattern Recogn (CVPR)*, pages 2921–9, 2016.

Annex D - State of the Art: Tables

Table D.1: Major datasets referenced in the literature reviewed, focusing on those used for experimental evaluations or as a basis for creating new datasets.

Dataset	Number of Images	Mode of Obtention	Availability	Summary	Reviewed Literature Using the Dataset
AMSL Face Morph Dataset [52, 53, 54]	~2,000	Morphed images using advanced techniques	Publicly available under the Creative Commons Attribution-ShareAlike 3.0 License	High-quality dataset of morphed images, designed to evaluate vulnerabilities in biometric systems	[78, 49, 33, 22]
Aberdeen Dataset [135]	~700	Controlled photography sessions	Publicly available	Controlled dataset capturing variations in lighting and pose	[23]
ChiMo Dataset [136, 137]	~24,000	Morphed images from the Chicago Faces Database using various morphing algorithms	Publicly available	Dataset of morphed faces designed to evaluate the vulnerability of facial recognition systems to morphing attacks, with raw and compressed versions	[95]
FEI Dataset [88]	~2800	Photographs under controlled conditions	Publicly available for research purposes only	Simple dataset of frontal and profile face images, designed for basic face recognition tasks	[23, 26, 25, 50]
FEI-Morph Dataset [89, 25]	~6000	Morphed images created from the FEI Dataset using various morphing techniques	Publicly available	Dataset of morphed faces designed to test facial recognition systems against morphing attacks	[25, 95]
FERET Dataset [138]	~14,000	Controlled and semi-controlled photography	Publicly available	Benchmark dataset for face recognition research, featuring diverse variations in facial characteristics	[75, 80]
FRGC Dataset [91]	~50,000	Controlled 2D and 3D facial images	Publicly available	Comprehensive dataset including 2D and 3D images, designed for advanced facial recognition studies	[24, 51]
FRGC-Morphs Dataset [56]	~6,000	Morphed using GANs and traditional methods	Distribution has been suspended	Large-scale dataset of morphed images, created to test MAD system robustness against morphing techniques	[75, 22]
FRLL Dataset [72]	~1,020	Photographs under controlled conditions	Publicly available	High-resolution 2D facial images with variations in pose, age, gender, and ethnicity	[23, 51]
FRLL-Morphs Dataset [52, 56, 39]	~1,000	Morphed using GANs and traditional methods	Publicly available	Dataset of high-quality morphed images, used to evaluate morphing attack vulnerabilities	[23, 75, 22, 80]
Facescape Dataset [99, 100, 101]	~17,000	3D scans of faces	Publicly available	Large-scale dataset of 3D facial scans, designed to capture diverse expressions and identities	[28]
FFHQ Dataset [92]	~70,000	Collected from Flickr	Publicly available	High-resolution dataset emphasizing demographic diversity, often used for GAN training	[80, 51]
HNU Dataset [90]	~1,000	Photographs and morphed images using diverse techniques	Requires direct authorization from the authors for research use	Diverse dataset of photographs and morphed images, emphasizing demographic diversity for morphing research	[26]
Idiap Morph Dataset (AMSL + FRLL-Morphs Datasets) [39, 56]	~3,000	Generated morphs	Publicly available	Dataset of morphed images, created to test the robustness of detection systems	[25, 50, 95]
LMA Dataset [58]	~500	Light morphing techniques	Requires approval for research use	Dataset focusing on subtle morphing effects, designed to study vulnerabilities in facial recognition systems	[49, 87]
LMA-DRD Dataset [48]	~54,000	Digital and printed morphs using diverse techniques	Requires approval for research use	Dataset of digital and scanned morphed images, designed to test detection systems in realistic scenarios	[48, 75]
MorGAN Dataset [58]	~3,000	GAN-generated morphs	Requires approval for research use	Synthetic dataset of GAN-generated morphs, exploring vulnerabilities in automated systems	[78, 75, 49, 87, 84]
MorphDB Dataset [57]	~2,000	Morphed and real images	Requires direct authorization from the authors for research use	Comprehensive database of real and morphed images, designed to analyze morphing detection methods	[35, 33, 25, 50]
ND-IIITD Dataset [139, 140]	~5,000	Controlled and natural images	Requires direct authorization from researchers and organizations through a formal request	Collaborative dataset combining natural and controlled face images, used for morphing research	[71]
NIST FRVT Morph Dataset [105]	~5,000	Controlled morphed images	Requires direct authorization from researchers and organizations through a formal request	Benchmark dataset by NIST, created to evaluate face recognition systems against morphing attacks	[35]
PMDB [57]	~1,000	Progressive morphing techniques	Requires direct authorization from the authors for research use	Dataset featuring gradual morphing effects, designed to study detection sensitivity	[35, 33, 25, 50, 95]
SCFace Dataset [94]	~4,000	Surveillance-like imagery	Publicly available	Low-resolution dataset of surveillance-like imagery, often used for challenging conditions in research	[34]
SOTAMD Dataset [41]	~5,000	Morphed and real images	Restricted through a sequestered evaluation platform	Large dataset of morphed and real images, designed for evaluating detection methods across techniques	[35]
Utrecht ECVP Dataset [135]	~130	Controlled photography	Publicly available	Controlled dataset focused on facial expressions and poses, used in biometric and psychological studies	[23]
VGGFace2 Dataset [55]	~3.3 million	Crawled from the web	Publicly available	Large-scale dataset with diverse face images, widely used for training deep learning models	[23, 22]
VISAPP Dataset [38]	~1,000	Morphed images	Publicly available	Dataset of morphed images, designed for evaluating vulnerability to morphing attacks	[49, 87, 84]

Table D.2: Summary of the major pre-trained models referenced in the literature reviewed, including details on their architecture, key features, and relevance to MAD.

Pre-trained Model	Summary	Literature Reviewed Using the Dataset
AlexNet [62]	AlexNet is a landmark deep learning model introduced in 2012, which substantially advanced image classification by leveraging CNNs. It used innovative techniques like ReLU activation, dropout regularization, and GPU acceleration to achieve a top-5 error rate of 15.3% on ImageNet [86], far surpassing previous methods. AlexNet's success revolutionized computer vision and popularized deep learning for large-scale image tasks.	[34, 46, 35]
ArcFace, ArcFace LResNet100E-IR [64]	ArcFace is a deep learning framework for face recognition introduced in 2019, that uses additive angular margin loss to improve the discriminative power of face embeddings. ArcFace LResNet100E-IR, a variant of ArcFace, is built upon a ResNet architecture and is fine-tuned using the ArcFace framework.	**ArcFace**: [95] **ArcFace LResNet100E-IR**: [71]
DenseNet-121 [68]	DenseNet-121 is a deep learning model introduced in 2017 as part of the DenseNet family, which focuses on efficient feature propagation by densely connecting each layer to every other layer. This architecture reduces the number of parameters while improving gradient flow and feature reuse, making it computationally efficient. DenseNet-121, with 121 layers, has been widely used for image recognition and transfer learning tasks due to its compactness and strong performance across diverse datasets.	[48]
EfficientNetB3 [141]	EfficientNetB3 is part of the EfficientNet family introduced in 2019, which employs a compound scaling method to balance network width, depth, and resolution for optimal performance. EfficientNetB3 achieves SOTA ACC on image recognition tasks while being computationally efficient, making it a popular choice for resource-constrained environments. Its design emphasizes achieving high ACC with fewer parameters compared to traditional architectures.	[23]
Inception-ResNet [96, 83]	Inception-ResNet combines the strengths of the Inception architecture with residual connections to improve training stability and efficiency. First introduced in 2015, it merges the feature extraction capabilities of Inception modules with the optimization benefits of residual learning. The model helps mitigate the vanishing gradient problem and allows for deeper and more accurate networks. Inception-ResNet v1, an early version, incorporated residual connections to the Inception modules.	**Inception-ResNet (unspecified version)**: [95] **Inception-ResNet v1**: [87, 49, 84, 25, 50]
PointNet [65], PointNet++ [97]	PointNet is a deep learning framework introduced in 2017 for directly processing unordered 3D point clouds, enabling tasks like classification and segmentation without requiring voxelization. PointNet++ builds on this by incorporating a hierarchical structure to capture local geometric features, improving performance on complex 3D data with varying densities and scales. These models are foundational for 3D data processing.	[28]
ResNet-50, ResNet-101 [63]	ResNet-50 and ResNet-101, introduced in 2016 as part of the ResNet family, are deep CNNs designed to address the vanishing gradient problem through residual connections. ResNet-50, with 50 layers, offers a balance of depth and efficiency, making it a standard for image recognition and transfer learning. ResNet-101, a deeper variant with 101 layers, provides improved feature representation and ACC, particularly for high-resolution and detailed image analysis. Both models maintain computational efficiency and are widely used for diverse computer vision tasks.	**ResNet-50**: [33, 22, 25, 34, 51] **ResNet-101**: [34]
SE-ResNet50 [82]	SE-ResNet50 is an enhanced version of ResNet50 introduced as part of the Squeeze-and-Excitation Networks (SENets) in 2018. It incorporates SE blocks, which adaptively recalibrate channel-wise feature responses by modeling interdependencies between feature channels. This design improves the network's representational power while maintaining computational efficiency, making SE-ResNet50 particularly effective for tasks requiring nuanced feature extraction.	[33]
ShuffleNetV2 [142]	ShuffleNetV2, introduced in 2018, is a lightweight CNN designed for mobile and embedded applications. It improves efficiency and ACC by optimizing operations such as channel splitting, pointwise group convolutions, and channel shuffling. ShuffleNetV2 achieves a better balance of speed and performance compared to its predecessors, making it ideal for resource-constrained environments.	[26]
SimpleView [98]	SimpleView, introduced in 2021, is a lightweight framework for processing 3D point clouds by projecting them into multiple 2D views. This approach leverages standard 2D CNNs to extract features, simplifying point cloud classification and achieving competitive performance with reduced computational complexity. It is particularly effective for tasks requiring efficient and scalable 3D data processing.	[28]
VGG16, VGG19 [69]	VGG16 and VGG19 are deep CNN architectures introduced in 2015, known for their simplicity and depth, using small 3x3 convolutional filters in a sequential design. With 16 and 19 layers respectively, they demonstrated that increased network depth improves performance, achieving SOTA results on ImageNet [86]. These models have become widely used for transfer learning and influenced the design of subsequent deep learning architectures.	[35, 34, 80]
VGG-Face [70], VGG-Face2 [55]	VGG-Face and VGG-Face2 are large-scale datasets designed for face recognition tasks, emphasizing diversity across poses, expressions, and demographics. These datasets are commonly used to train deep learning models, such as VGG16 or ResNet architectures, resulting in widely adopted pre-trained models often referred to as "VGG-Face models".	[35, 47]
Xception [93]	Xception, introduced in 2017, is an extension of the Inception architecture that uses depthwise separable convolutions to improve efficiency and performance. By replacing standard convolutions with separable ones, Xception reduces computational complexity while maintaining high ACC, making it well-suited for tasks requiring efficient deep learning models.	[34]

Table D.3: Summary of the best results reported in each study included in the literature review. Only the best results were selected for this comparison. *Universal dataset* refers to a combination of all datasets used in the study. *Intra-dataset evaluation* indicates that both training and testing were performed on samples from the same dataset. *Intra-technique evaluation* indicates that training and testing were performed on samples generated using the same morphing technique, while *cross-technique evaluation* refers to training and testing on samples generated using different morphing techniques. *Values marked with an asterisk represent the average performance across all reported results.

Reviewed Literature	Experiment Description	BPCER (%) @ APCER			APCER (%) @ BPCER			D-EER (%)	Acc (%)
		1%	5%	10%	1%	5%	10%		
Aghdaie et al. (2022) [87]	Intra-dataset evaluation on VISAPP [38]	-	0.00	0.00	-	-	-	0.00	-
	Training on universal dataset and testing on VISAPP [38]	-	0.00	0.00	-	-	-	0.00	-
Banerjee and Ross (2021) [78]	Intra-dataset evaluation on AMSL [52, 53, 54]	-	-	6.1	-	-	5.2	-	-
	Training on AMSL [52, 53, 54] and testing on EMorGAN [79]	-	-	4.4	-	-	0.8	-	-
Blasingame and Liu (2021) [80]	Cross-technique and intra-dataset evaluation on FRLL [52, 56, 39]	0.00	-	0.00	-	-	-	0.47	-
	Intra-technique and cross-dataset evaluation; training on FERET [138] and testing on FRLL [52, 56, 39]	0.00	-	0.00	-	-	-	0.04	-
Borghi et al. (2021) [33]	Training on PMDB [57] and testing on MorphDB [57] criminal images	0.00	-	0.00	0.00	-	-	-	-
Chaudhary et al. (2021) [49]	Training on universal dataset (BW-22) and testing on VISAPP [38]	-	0.00	0.00	-	0.00	0.00	0.00	-
	Training on universal dataset (BW-22) and testing on AMSL [52, 53, 54]	-	-	-	-	33.78	23.61	16.4	-
Colbois et al. (2024) [51]	Cross-technique evaluation (different combinations)	-	-	-	-	-	-	0.00	-
	Cross-dataset evaluation (different combinations with training on FFHQ [92])	-	-	-	-	-	-	0.00	-
	Training on FRGC [91] digital images and testing on FRGC [91]-derived P&S images	-	-	-	-	-	-	3.99	-
	One-class detector trained on FRGC [91] and tested on diffusion morphs	-	-	-	-	-	-	0.00	-
Damer et al. (2021) [48]	Training and testing on digital images	-	16.12	6.45	-	-	-	-	-
Di Domenico et al. (2023) [25]	Training on universal dataset and testing on FEI [88] accomplice images with Min-Max feature fusion and cosine similarity	46.8	23.7	-	-	-	-	12.5	-
Fang et al. (2022) [75]	Training on CASIA-Web and testing on FRLL [52, 56, 39]	-	-	-	-	-	-	2.98	-
Ferrara et al. (2021) [35]	Fine-tuning of VGG-Face16 on $Digital_{Mc}$ and testing on MorphDB [57] digital images	0.0	0.0	0.0	-	-	-	0.9	97.4
	Fine-tuning of VGG-Face16 on $Digital_{Mc} + \widetilde{P\&S}_{MC}$ and testing on MorphDB [57] P&S images	43.8	6.9	2.3	-	-	-	6.1	93.5
Jia et al. (2023) [26]	Intra-dataset evaluation on FEI [88]	-	0.66	0.17	-	-	-	0.12	-
	Training on HNU (MDB1) [90] and testing on FEI [88]	-	3.03	1.47	-	-	-	3.26	-
Liu et al. (2024) [50]	Criminal scenario	11.5	2.0	-	-	-	-	3.7	-
Medvedev et al. (2023) [22]	Testing on FaceMorpher [103] (composed of simple morphs with foreground and background artifacts)	84.8	-	56.2	78.6	-	43.1	-	-
Medvedev et al. (2024) [23]	Testing of Fused Classification model on AMSL [52, 53, 54]	-	-	-	27.5	-	3.9	-	-
Ortega-Delcampo et al. (2020) [47]	Testing on digital images	-	-	-	-	-	-	0.78	98.7
Qin et al. (2022) [71]	D-MAD: Training on complete morphing images and testing on combined morphing images	-	0.00	-	-	-	-	0.48	-
	S-MAD: Training on combined morphing images and testing on splicing morphing images	-	0.00	-	-	-	-	0.63	-
Razaq and Shukur (2023) [27]	Intra-dataset evaluation of DNN on universal dataset	-	-	-	-	-	-	-	99.69
	Training DNN on universal dataset and testing on AMSL [52, 53, 54] (100 epochs)	-	-	-	-	-	-	-	95.8
Robledo-Moreno et al. (2024) [95]	Cross-technique evaluation of FL for S-MAD	0.00	0.00	-	-	-	-	0.00	-
	Cross-dataset evaluation of FL for D-MAD	34.3	9.7	-	-	-	-	7.1	-
Singh and Ramachandra (2022) [34]	Training and testing on digital images	-	-	-	-	-	-	0.00	-
Singh and Ramachandra (2024) [28]	Feature extraction using SimpleView [98]	-	2.5	0.00	-	-	-	1.59	-
Soleymani et al. (2021) [84]	Classification by feature difference combined with a SVM; training and testing on VISAPP [38]	-	6.12	3.71	-	5.29	3.17	5.2	-
Venkatesh et al. (2020) [46]	Intra-dataset evaluation on a derivative of FRGC [91]	-	1.11	1.11	-	-	-	2.63	-
	Intra-dataset evaluation on the universal dataset	-	5.01	3.05	-	-	-	4.96	-
	Training on a derivative of FRGC [91] and testing on a combination of a subset of FRGC [91] and a private dataset	-	4.66	2.88	-	-	-	4.66	-
Zhang et al. (2024) [24]	Cross-technique evaluation on morphs obtained from FRGC [91] digital images	-	-	-	-	-	-	13.63*	-

Annex E - Selected Datasets Sample Images

(a) Morphed training image *morphed_img000184_-img282038*.

(b) *Bona fide* training image *img000025*.

(c) Morphed testing cropped image *morphed_-img000030_img279570*.

(d) *Bona fide* testing cropped image *img000146*.

Figure E.1: Sample images from the SMDD dataset [59].

(a) Image *001_03* in pose *neutral_front*.

(b) Image *009_07* in pose *smiling_left_3quarter*.

(c) Image *006_06* in pose *smiling_left_profile*.

Figure E.2: Sample images from the FRLL dataset [72].

(a) Image *008_037* generated using the morphing technique *AMSL*.

(b) Image *008_037* generated using the morphing tool *FaceMorpher*.

(c) Image *008_037* generated using the morphing tool *OpenCV*.

(d) Image *008_037* generated using the morphing tool *StyleGAN*.

(e) Image *008_037* generated using the morphing tool *WebMorph*.

Figure E.3: Sample images from the FRLL-Morphs dataset [52, 56, 39].

(a) Image *00003*. (b) Image *10009*. (c) Image *05000*.

Figure E.4: Sample images from FFHQ dataset [92].

(a) Image *00202_34541* generated using a diffusion-based morphing algorithm (*diffae*).

(b) Image *00202_34541* generated using a landmark-based morphing algorithm (*landmark-combined*).

(c) Image *00202_34541* generated using a landmark-based morphing algorithm (*landmark-complete*).

(d) Image *00202_34541* generated using a GAN-based morphing algorithm (*stylegan-interp-W+*).

(e) Image *00202_34541* generated using a GAN-based morphing algorithm (*stylegan-interp-W*).

Figure E.5: Sample images from the FFHQ-Morphs dataset [51].

Annex F - Unsuccessful Experiments

Detailed below are the replication attempts of three studies whose provided resources did not allow for successful experimental reproduction. These experiments are included to maintain transparency regarding the exploratory process and to document the technical and methodological constraints encountered. In the case of Medvedev et al. (2023) [22], the publicly available repository contained the benchmarking framework without the actual implementation of the proposed detection model, preventing a faithful reproduction of their results. The attempt to replicate Colbois et al. (2024) [51] was hindered by unresolved issues within the provided pipeline, even after communication with the authors, which led to the termination of the replication attempt. Finally, the study by Singh and Ramachandra (2024) [28] focused on 3D face morphing detection, a domain outside the scope of this work, that concentrates on 2D facial image analysis.

Experiment 2 — Medvedev et al. (2023)

The code provided by Medvedev et al. (2023) [22] is publicly available through a GitHub repository[1] and serves primarily as a benchmark framework for evaluating MAD methods. Instead of offering a complete implementation of the detection model described in their publication, the repository provides a structured protocol where users can insert their own models and obtain standardized evaluation metrics.

A custom Jupyter notebook[2] was developed as part of this dissertation to explore the capabilities and limitations of the code available. This notebook replicates the full demonstration pipeline described in the repository and confirms that the code is designed for benchmark integration rather than direct model evaluation.

The workflow demonstrated in the notebook follows the key steps:

1. **Repository Setup and Data Download**: The repository is cloned, and a script is used to download predefined datasets used in the benchmark. These include subsets of publicly available face datasets. However, not all datasets referenced in the benchmark are fully accessible, for example, the AR Face Database [143] could not be downloaded automatically.

[1] `https://github.com/iurii-m/MorDeephy.git`

[2] `https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Experiments/2-MorDeephy-Benchmark`

2. **Dependency Installation and Data Extraction**: Required Python packages are installed, and the downloaded data is extracted and structured in preparation for the following steps.
3. **Face Alignment**: A protocol-specific script is executed to align facial images and organize them according to the benchmark's expected directory structure.
4. **Prediction Simulation**: A demonstration script generates randomized prediction scores. This step does not involve any actual model inference and serves only to illustrate the structure and functionality of the benchmark pipeline.
5. **Benchmark Execution**: The benchmark script processes the simulated predictions and computes evaluation metrics such as APCER, BPCER, and ACER for the selected protocol.

This pipeline highlights that the benchmark is designed to facilitate reproducibility and comparability across different morphing detection approaches. However, Medvedev et al. (2023)'s [22] actual method, described in the paper, was not implemented in the public repository and would need to be manually developed and integrated by the user. As such, no experiments were performed using the benchmark framework beyond verifying the demonstration pipeline. The provided infrastructure was valuable for understanding the protocol format and evaluation procedure but did not include sufficient code to reproduce the detection results reported in the original publication.

Experiment 3 — Colbois et al. (2024)

The work by Colbois et al. (2024) [51] explores the use of attack-agnostic features for MAD and provides an accompanying implementation through a GitLab repository[3]. The code is structured around a Snakemake pipeline designed to reproduce the experiments reported in the publication by orchestrating dataset preparation, feature extraction, model training, and evaluation.

A custom notebook[4] was developed to test and adapt this pipeline using the FFHQ [92] and FFHQ-Morphs dataset [51]. However, several technical challenges were encountered during the setup and execution of the provided workflow. Initially, execution failed due to missing configuration parameters, such as the undefined `weights.root` variable, which was later resolved following communication with the authors. Despite this progress, subsequent stages of the Snakemake pipeline encountered import and dependency errors within the `mada.extractors` module, preventing successful feature extraction and halting the overall workflow.

Through direct correspondence, the author confirmed that the Snakemake pipeline

[3] `https://gitlab.idiap.ch/bob/bob.paper.ijcb2024_agnostic_features_mad`

[4] `https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Experiments/3-Attack-Agnostic-Features`

was primarily intended for strict reproducibility of the original publication and therefore depended on the exact structure and availability of datasets, including the FRGC-Morphs dataset [56], which is no longer accessible. Moreover, the FRLL-Morphs dataset used in the paper had been regenerated with updated attack types (e.g., diffusion-based morphs) not yet publicly released at the time of testing. The author suggested focusing instead on direct script-based execution using the available FFHQ-Morphs dataset [51], bypassing the full pipeline.

Due to the combination of unresolved code malfunctions, dependency errors, and the unavailability of the regenerated FRLL-Morphs dataset, the replication attempt was ultimately discontinued.

Experiment 4 — Singh and Ramachandra (2024)

The code provided by Singh and Ramachandra (2024) [28] is publicly available through a GitHub repository[5] and offers a sample implementation of their method for creating and identifying 3D face morphing attacks from point clouds. Since identification documents rely on 2D facial images, and this research with its accompanying code is specifically designed for processing 3D facial data, it was not directly applicable to the experiments in the current research and was not used for experiments.

[5] https://github.com/jagmohaniiit/3DFaceMorph

Annex G - Experiment 1: Experimental Code

This experiment implements and evaluates the SPL-MAD system proposed by Fang et al. (2022) [75], which employs an unsupervised anomaly detection approach for MAD. The implementation is organized into two Jupyter notebooks[6], each serving a specific role in the experimental pipeline:

- `SPL-MAD.ipynb`: The primary notebook that handles dataset preparation, organization, and preprocessing to ensure compatibility with the SPL-MAD training pipeline.
- `SPL-MAD_ERR-Analysis.ipynb`: An analysis notebook that processes experimental results and computes performance metrics, particularly focusing on EER analysis.

Notebook 1: `SPL-MAD.ipynb` (Dataset Preparation, Train, and Test)

This notebook implements the dataset preparation and preprocessing pipeline required for the SPL-MAD system. Unlike supervised approaches, SPL-MAD operates under an unsupervised learning paradigm, where the model learns to distinguish *bona fide* facial representations from morphed images through a self-paced learning mechanism that gradually incorporates more challenging samples during training.

The experimental workflow comprises the following key stages:

1. **Environment Setup**: The notebook initializes the computational environment by mounting Google Drive for data access and configuring all necessary dataset and script paths. This ensures that all resources are properly accessible for subsequent processing stages.
2. **Dataset Acquisition and Synchronization**: The notebook handles the download and synchronization of required datasets, including:
   - SMDD [59].
   - FRLL-Morphs [52, 56, 39].

   These datasets encompass multiple morphing techniques and provide diverse evaluation scenarios for assessing the robustness of the SPL-MAD approach.
3. **Dataset Structuring**: To organize image data, the notebook executes helper scripts (`frll_morphs_create_csv.py`) and generates CSV files with labels (`attack` or *bona*

[6] `https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Experiments/1-SPL-MAD`

fide), producing one file for each morphing technique in the FRLL-Morphs dataset[52, 56, 39].

4. **Data Formatting for SPL-MAD Pipeline**: The notebook ensures that all datasets are properly formatted and structured according to the requirements of the SPL-MAD training pipeline. This includes organizing images into appropriate directory hierarchies and ensuring that metadata files correctly reference all image paths and labels.
5. **Training and Evaluation Pipeline Integration**: The notebook prepares and organizes all necessary resources to enable model training and evaluation using the authors' official SPL-MAD implementation[7].

The self-paced learning mechanism is a key innovation of this approach, as it allows the model to start with easier samples and gradually incorporate more challenging examples, thereby improving the model's ability to generalize across different morphing techniques.

Notebook 2: `SPL-MAD_ERR-Analysis.ipynb` (Error Rate Analysis)

This notebook performs a post-hoc analysis of experimental results from the SPL-MAD pipeline, with particular emphasis on computing and visualizing EER metrics across different datasets and morphing techniques.

The analysis workflow encompasses:

- **Environment Configuration**: The notebook mounts Google Drive and locates experiment log files generated during the SPL-MAD training and evaluation phases. These log files contain performance metrics for each experimental configuration.
- **Log Parsing and Metric Extraction**: The notebook executes the analysis script `analyze_logs.py` to systematically parse EER results from experiment logs. Performance data is processed across multiple dimensions:
  - Different datasets.
  - Multiple morphing methods.
  - Various training configurations.
- **Performance Aggregation and Comparison**: The notebook aggregates error metrics across different experimental conditions and facilitates both visual and quantitative assessment of SPL-MAD's robustness. Key metrics include:
  - EER comparison between cropped and non-cropped SMDD [59] variants.
  - EER for each morphing technique in FRLL-Morphs [52, 56, 39].

[7] `https://github.com/meilfang/SPL-MAD`

- **Robustness Evaluation**: By analyzing performance across diverse morphing techniques, the notebook enables assessment of the SPL-MAD system's ability to generalize to unseen morphing methods, which is a critical requirement for practical MAD systems deployed in real-world scenarios.

This experimental framework enables the evaluation of the unsupervised SPL-MAD approach across diverse datasets and morphing techniques. The self-paced learning mechanism represents a key advantage over supervised methods, as it allows the model to progressively learn discriminative features that distinguish *bona fide* from morphed images.

Annex H - Experiment 5: Experimental Code

This experiment implements and evaluates the MAD architecture proposed by Zhang et al. (2024) [24], which combines a ViT for feature extraction with a SVM for classification. The implementation is organized into three Jupyter notebooks[8], each serving a specific role in the experimental pipeline:

- `Zhang2024.ipynb`: The primary notebook that handles the main experimental setup, model training, and evaluation across multiple dataset configurations.
- `OcclusionMappingForDatasets.ipynb`: A preprocessing notebook that generates systematically occluded versions of the datasets to enable interpretability analysis.
- `Zhang2024_Analysis.ipynb`: An analysis notebook that processes experimental results and generates visualizations, including occlusion heatmaps.

Notebook 1: `Zhang2024.ipynb` (Main Experiment)

This notebook implements the core experimental pipeline based on the architecture described by Zhang et al. (2024) [24], which employs a pre-trained ViT for deep feature extraction followed by a SVM for binary classification of *bona fide* versus morphed facial images.

The experimental workflow comprises the following key stages:

1. **Environment Setup**: The notebook initializes the computational environment by mounting Google Drive for data access and configuring a Conda environment with Python 3.10 and all required dependencies. This ensures reproducibility and consistency across different execution environments.
2. **Model and SVM Configuration**: Configuration parameters are defined for both the ViT feature extractor and the SVM classifier. Key parameters include dataset paths, target image dimensions for preprocessing, and the SVM kernel type (Radial Basis Function, `'rbf'`). The pre-trained ViT model (e.g., `vit_l_32`) is loaded with its classification head removed to extract embedding vectors.
3. **Data Loading and Preprocessing**: The experiment utilizes four datasets derived from two sources:

[8] `https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Experiments/5-ViT_SVM-OcclusionMapping`

- FRLL-Morphs [52, 56, 39] and its cropped variant (*FRLL-Cropped*).
- FFHQ-Morphs [51] and its cropped variant (*FFHQ-Cropped*).

The cropped datasets are generated by applying face detection, alignment, cropping, resizing, and normalization to the original images. All datasets are partitioned into training and testing subsets in the ratio 80/20.

4. **Feature Extraction and SVM Training**: Deep visual features are extracted from preprocessed images using the pre-trained ViT backbone. These high-dimensional embedding vectors serve as input to the SVM classifier, which is trained to discriminate between *bona fide* and morphed samples. Both the trained SVM model and extracted features are persisted for subsequent analysis.
5. **Cross-Dataset Evaluation**: To assess cross-dataset generalization capabilities, a comprehensive evaluation protocol is implemented. Each of the four datasets serves as a training set exactly once, and each resulting model is evaluated on all four test sets, yielding 16 distinct train-test combinations. This entire experimental protocol is executed under two conditions:
   - Using the original unbalanced datasets.
   - Using balanced datasets (via undersampling).

The class balancing strategy employs random undersampling, where samples from the majority class are randomly removed to match the cardinality of the minority class within each dataset, thereby mitigating potential class imbalance bias.

After generating the occluded datasets using `OcclusionMappingForDatasets.ipynb`, an additional evaluation phase is conducted to assess model interpretability. Since occlusion mapping is computationally intensive and was applied exclusively to the cropped datasets, this analysis is restricted to models trained on *FRLL-Cropped* and *FFHQ-Cropped*. Each model is evaluated on four occluded test sets (grid-based and landmark-based occlusions for both *FRLL-Cropped* and *FFHQ-Cropped*), resulting in eight additional train-test combinations that enable occlusion sensitivity analysis.

Notebook 2: `OcclusionMappingForDatasets.ipynb` (Occlusion Mapping)

This notebook implements a systematic occlusion-based interpretability analysis to identify which facial regions are most critical for the model's classification decisions. The approach is inspired by occlusion sensitivity methods commonly used in deep learning interpretability research. By systematically masking different facial regions and measuring the resulting performance degradation, it is possible to infer which facial characteristics the model relies upon most heavily for distinguishing *bona fide* from morphed images.

The notebook implements two complementary occlusion strategies:

- **Landmark-based Occlusion**:
  1. This semantically-informed approach leverages dlib's 68-point facial landmark detector [144, 145] to localize four key facial regions: left eye, right eye, nose, and mouth.
  2. For each input image, four occluded variants are generated, each with exactly one facial region masked by a black rectangle. The occlusion area is expanded by a scale factor of 1.5 relative to the landmark bbox to ensure complete coverage of the target landmark.
  3. The occluded images are saved in structured directories that preserve the original dataset organization, facilitating systematic evaluation of region-specific importance.
- **Grid-based Occlusion**:
  1. This spatially-uniform approach partitions each image into a 4×4 grid, yielding 16 equally-sized cells.
  2. For each grid cell, an occluded variant is created by masking that specific cell with a black rectangle, resulting in 16 occluded versions per input image.
  3. This method enables fine-grained spatial analysis of model sensitivity without relying on facial landmark detection, providing complementary insights to the landmark-based approach.

The grid-based method offers higher spatial granularity and is agnostic to facial structure, while the landmark-based method provides semantically meaningful occlusions aligned with facial landmarks. It is important to note that both occlusion strategies were not applied to the *bona fide* subset of the FFHQ dataset [92] due to its substantial size (approximately 90GB), which would have resulted in high storage requirements (exceeding 1TB for all occluded variants) and excessive computational overhead.

Notebook 3: `Zhang2024_Analysis.ipynb` (Results and Visualization)

This notebook performs a post-hoc analysis of experimental results and generates visualizations to communicate findings. The analysis pipeline integrates both quantitative performance metrics and qualitative interpretability insights derived from the occlusion mapping experiments.

The analysis workflow encompasses:

- **Performance Metrics Calculation**: The notebook loads prediction outputs from all experimental configurations and computes standard biometric evaluation metrics including APCER, BPCER, EER, and overall classification ACC. These metrics

are computed for each train-test combination and presented in structured tables that facilitate cross-dataset comparison for both FRLL and FFHQ dataset families.

- **Occlusion Heatmaps Visualization**: The notebook generates interpretability visualizations that quantify the impact of regional occlusions on model performance. Two primary visualization types are produced:
  - *Landmark-based occlusion maps*: Heatmaps showing performance degradation when specific facial landmarks (eyes, nose, mouth) are occluded.
  - *Grid-based occlusion maps*: Spatial heatmaps overlaid on face templates showing performance degradation for each grid cell.

  A larger performance drop for a particular region indicates higher model reliance on that area for classification, thereby revealing which facial characteristics are most discriminative for MAD.
- **Comparative Analysis**: The notebook produces comparative visualizations including performance curves, cross-dataset generalization matrices, and statistical comparisons across different experimental conditions (balanced vs. unbalanced datasets, different morphing techniques, etc.).

This three notebook experimental framework enables not only quantitative evaluation of the ViT-SVM architecture for MAD but also provides interpretability insights through systematic occlusion analysis, thereby contributing to a deeper understanding of which facial characteristics are most critical for distinguishing *bona fide* from morphed images.

Annex I - Experiment 5: Results

Table I.1: Results from experiments using Zhang et al. (2024) [24] code.

Training Set Balance	Training Set Cropping	Train dataset	Test Set Cropping	Test dataset	APCER@BPCER			D-EER (%)	AUC (%)
					1%	5%	10%		
Unbalanced	Not cropped	FFHQ	Not cropped	FFHQ	6.22	1.48	0.74	2.85	99.63
				FRLL	97.59	83.76	77.70	39.02	65.72
			Cropped	FFHQ	4.74	0.82	0.22	2.34	99.78
				FRLL	14.58	8.49	3.89	7.30	98.24
		FRLL	Not cropped	FFHQ	98.96	94.02	87.99	46.53	55.0
				FRLL	4.89	3.37	2.75	4.30	97.16
			Cropped	FFHQ	98.50	93.86	87.99	45.03	57.02
				FRLL	85.35	81.10	75.30	39.01	66.85
	Cropped	FFHQ	Not cropped	FFHQ	45.39	23.23	14.13	12.11	94.71
				FRLL	71.44	54.92	43.91	31.13	79.11
			Cropped	FFHQ	**0.00**	**0.00**	**0.00**	0.12	**100.00**
				FRLL	92.85	78.06	69.85	40.94	65.60
		FRLL	Not cropped	FFHQ	92.86	77.13	63.98	31.52	75.21
				FRLL	73.30	60.15	56.37	26.84	80.78
			Cropped	FFHQ	91.41	76.65	63.80	30.82	75.64
				FRLL	0.21	0.21	0.14	0.11	99.96
Balanced	Not cropped	FFHQ	Not cropped	FFHQ	6.88	1.10	0.42	2.14	99.71
				FRLL	95.39	76.53	74.05	36.15	71.46
			Cropped	FFHQ	5.42	1.02	0.26	2.34	99.75
				FRLL	8.28	6.86	2.83	7.09	98.69
		FRLL	Not cropped	FFHQ	95.92	86.31	76.87	40.57	63.30
				FRLL	2.75	2.75	2.68	2.60	98.36
			Cropped	FFHQ	93.76	80.13	68.04	33.54	72.64
				FRLL	79.48	34.11	29.79	17.06	90.34
	Cropped	FFHQ	Not cropped	FFHQ	41.66	25.06	15.41	12.61	94.69
				FRLL	70.89	43.98	33.38	22.02	86.53
			Cropped	FFHQ	**0.00**	**0.00**	**0.00**	**0.04**	**100.00**
				FRLL	77.14	55.34	53.43	31.18	78.24
		FRLL	Not cropped	FFHQ	88.57	73.18	59.22	29.68	77.66
				FRLL	77.98	54.30	48.52	31.72	78.48
			Cropped	FFHQ	88.59	71.72	58.27	29.24	77.83
				FRLL	0.21	0.14	0.14	0.11	99.93

Annex J - Occlusion-based Interpretability Method: Source Code

This contribution presents a complete pipeline for occlusion-based interpretability analysis of MAD models. The approach systematically occludes different facial regions using both grid-based and landmark-based strategies, measuring the impact on model predictions to generate visual heatmaps that highlight the most influential areas for morphing detection. This enhances model transparency and trustworthiness in security-critical applications by revealing which facial characteristics deep learning models rely upon when making detection decisions. The implementation comprises three Python scripts[9] and three demonstration Jupyter notebooks.

Python Scripts:

- `crop_dataset_builder.py`: Prepares standardized datasets by detecting, cropping, and resizing facial images to ensure consistent face regions across all samples.
- `occlusion_dataset_builder.py`: Generates occluded variants of facial images using grid-based and landmark-based occlusion strategies to test model robustness and identify critical facial regions.
- `occlusion_heatmaps_builder.py`: Creates visual heatmaps that highlight which facial regions influence model predictions the most, combining grid and landmark occlusion scores into composite visualizations.

Demonstration Notebooks:

- `Crop-Dataset-Builder.ipynb`: Demonstrates, in a Google Colab environment, the usage of `crop_dataset_builder.py`, processing 7,265 images from the FRLL-Morphs dataset [52, 56, 39].
- `Occlusion-Dataset-Builder.ipynb`: Demonstrates, in a Google Colab environment, the usage of `occlusion_dataset_builder.py`, generating 247,010 occluded variants from 7,265 facial images using both grid and landmark occlusion methods.
- `Occlusion-Heatmaps-Builder.ipynb`: Demonstrates, in a Google Colab environment, the usage of `occlusion_heatmaps_builder.py`, generating interpretability

[9] https://github.com/pedrojfmorais/MEI-Thesis/tree/main/Occlusion-based-Interpretability

heatmaps from model prediction scores on regular, grid-occluded, and landmark-occluded images.

Script 1: `crop_dataset_builder.py` (Face Detection and Cropping)

This script prepares clean, standardized datasets by detecting faces and cropping images, ensuring that every image has a uniform face region. This preprocessing step is essential for subsequent occlusion experiments, as it guarantees spatial consistency across all facial images and enables meaningful comparison of occlusion effects.

The processing pipeline comprises the following key stages:

- **Face Detection and Cropping**: The script employs MTCNN [126] for robust face detection across diverse imaging conditions. For each input image, MTCNN [126] localizes the facial bbox, and the image is cropped to this region to isolate the face from background clutter. This ensures consistent facial framing across all datasets, eliminating variability introduced by different image compositions.
- **Folder Structure Preservation**: Cropped faces are saved in the same relative directory tree under the output folder, maintaining the original dataset organization. This preservation of hierarchical structure facilitates easy tracking of image provenance and ensures compatibility with downstream processing scripts that rely on consistent folder organization.
- **Multiprocessing Support**: The script supports optional parallel processing for large datasets through configurable worker processes. This reduces processing time for datasets containing thousands of images, with performance scaling approximately linearly with the number of CPU cores available.
- **Detailed Logging**: The script records comprehensive processing information, including progress updates, warnings for failed detections, and processing times, to both console and timestamped log files. This logging infrastructure enables post-hoc analysis of processing quality and identification of problematic images that may require manual inspection.

The demonstration notebook `Crop-Dataset-Builder.ipynb` illustrates practical usage in a Google Colab environment, processing 7,265 images from FRLL-Morphs [52, 56, 39] in approximately 11 minutes using 12 parallel workers, producing a 943 MB compressed dataset of standardized face crops.

Script 2: `occlusion_dataset_builder.py` (Occlusion Generation)

This script generates occluded variants of facial images to simulate different types of facial occlusion, enabling systematic evaluation of which facial regions are most critical for MAD model decisions. The script implements two complementary occlusion

strategies that provide both spatial and semantic perspectives on model behavior.

The occlusion generation pipeline encompasses:

- **Grid-Based Occlusion**: This spatially-uniform approach divides each face into an $R \times C$ grid (default 4×4, configurable) and generates $R \times C$ occluded variants per image, each with exactly one grid cell masked by a black rectangle. This method enables fine-grained spatial analysis of model sensitivity without relying on facial landmark detection, providing comprehensive coverage of the entire facial region. The grid-based approach is particularly valuable for identifying unexpected spatial dependencies that may not align with semantic facial landmarks.
- **Landmark-Based Occlusion**: This semantically informed approach uses dlib's 68-point facial landmark detector [144, 145] to localize key facial regions (left eye, right eye, nose, mouth). For each input image, four occluded variants are generated, each with exactly one facial region masked. The script supports two masking strategies:
  - *Pixel-precise polygon masking*: Uses the exact landmark point coordinates to define irregular polygonal regions that follow facial landmarks boundaries.
  - *Bbox masking*: Computes axis-aligned bboxes around landmark regions for simpler rectangular occlusions.

  A configurable scale factor (default 1.5) expands the occlusion area beyond the landmark boundaries to ensure complete coverage of the target landmark. This semantic approach provides interpretable insights into which facial landmarks (e.g., eyes versus mouth) are most discriminative for morphing detection.
- **Flexible Execution Modes**: The script supports three execution modes to accommodate different experimental requirements:
  - `grid`: Generates only grid-based occlusions for spatial analysis.
  - `landmark`: Generates only landmark-based occlusions for semantic analysis.
  - `both`: Generates both types in a single execution for comprehensive analysis.
- **Automatic Model Management**: The script automatically downloads and extracts the dlib's 68-point shape predictor model [144, 145] if not present, eliminating manual dependency management and ensuring reproducibility across different execution environments.
- **Multiprocessing and Logging**: Similar to the cropping script, this tool supports parallel processing with configurable worker counts and maintains timestamped logs tracking processing progress, warnings, and errors.

The demonstration notebook `Occlusion-Dataset-Builder.ipynb` illustrates practical usage, processing 7,265 facial images with a 6×5 grid configuration and landmark-

based occlusion (1.5× scale factor, bbox precision), generating 247,010 occluded variants in approximately 1 hour 52 minutes using 12 workers, producing a 13.6 GB compressed dataset.

Script 3: `occlusion_heatmaps_builder.py` (Heatmap Visualization)

This script creates visual heatmaps that quantify and visualize which facial regions influence model predictions the most. It generates three types of heatmaps: grid-based, landmark-based, and composite heatmaps that combine both approaches, providing comprehensive interpretability insights into MAD model decision making processes.

The heatmap generation pipeline comprises:

- **Flexible Score Parsing**: The script accepts three types of CSV files containing model predictions:
  - *Regular scores*: Baseline predictions on original (cropped or non-cropped) images, providing reference performance.
  - *Grid scores*: Predictions on grid-occluded images, where each row represents a prediction when a specific grid cell is occluded. The script uses flexible path matching and regex patterns to extract grid cell indices from file paths (supporting formats like `r3_c5`, `row3_col5`, `cell_3_5`).
  - *Landmark scores*: Predictions on landmark-occluded images, where each row represents a prediction when a specific facial region is occluded. The script extracts landmark region names from file paths (e.g., `left_eye`, `nose`, `mouth`).

  The script automatically matches occluded predictions back to their corresponding baseline predictions using flexible path matching that accounts for the original dataset organization.
- **Grid-Based Heatmaps**: For each image, the script constructs a spatial heatmap by mapping grid occlusion scores to their corresponding grid cell positions. The heatmap is overlaid on the original image with grid lines and cell values, using color gradients to indicate prediction sensitivity (lighter colors indicate high impact regions where occlusion considerably degrades performance, darker colors indicate low impact regions).
- **Landmark-Based Heatmaps**: The script generates heatmaps using facial landmark occlusion scores, employing dlib's 68-point shape predictor [144, 145] to map landmark regions to precise image coordinates. This enables visualization of semantic landmark importance, showing whether eyes, nose, or mouth are most critical for detection decisions. The heatmap is overlaid on the original image, using the same color gradients as the grid-based heatmaps.
- **Composite Heatmaps**: The script can optionally combine grid and landmark

heatmaps into unified visualizations that provide both spatial granularity and semantic interpretability. This composite view enables comprehensive understanding of model behavior from multiple perspectives. Similar to the previous heatmaps, it is overlaid on the original image and uses the same color gradient for prediction sensitivity.

- **Heatmap Visualization Modes**: The script supports two visualization modes:
  - `occluded`: Displays raw model prediction scores when each region is occluded, showing absolute performance levels.
  - `diff`: Displays absolute difference from baseline scores, highlighting the impact of occlusion by showing performance degradation relative to the non-occluded reference.

  The `diff` mode is particularly valuable for interpretability, as it directly quantifies how much each facial region contributes to the model's decision-making process.
- **Visual Overlay and Customization**: The script generates PNG heatmaps with customizable transparency (alpha blending), colormap selection (default: `hot`), and overlay styles. Grid lines, bboxes, and cell values are rendered on top of the original images to facilitate interpretation. The output preserves the original dataset folder structure, enabling easy comparison across different images and experimental conditions.
- **Multiprocessing and Logging**: The script supports efficient parallel processing of large datasets with configurable worker counts and maintains timestamped logs with adjustable logging levels (`DEBUG`, `INFO`, `WARNING`, `ERROR`).

The demonstration notebook `Occlusion-Heatmaps-Builder.ipynb` illustrates practical usage, generating three types of heatmaps (grid, landmark, and composite) from model prediction CSV files using the `diff` mode to visualize performance degradation caused by regional occlusions.

Suggested Workflow and Practical Considerations

The complete occlusion-based interpretability pipeline follows a four-stage workflow:

1. **Dataset Cropping**: Executes `crop_dataset_builder.py` to obtain clean, standardized face crops from raw datasets, ensuring spatial consistency for subsequent occlusion experiments.
2. **Occlusion Generation**: Executes `occlusion_dataset_builder.py` to generate occluded variants of the cropped images, both grid- and landmark-based, creating test sets for interpretability analysis.
3. **Model Prediction**: Pass all images (regular and occluded) through the MAD model to produce prediction score files in CSV format, capturing model behavior

under different occlusion conditions.

4. **Heatmap Visualization**: Executes `occlusion_heatmaps_builder.py` with the regular, grid, and landmark score files to create visual heatmaps showing which facial regions influence model decisions the most.

This pipeline enables systematic evaluation of which facial regions most influence MAD model performance and ensures a reproducible, automated analysis workflow. Key practical considerations include:

- **Performance Optimization**: Multiprocessing (`--multiprocessing --workers <n>`, with `n` being the number of workers) accelerates processing for large datasets, with near-linear scaling up to the number of available CPU cores.
- **Storage Management**: When working with Google Colab or similar cloud environments, avoid writing directly to cloud storage during processing. Instead, save locally, compress to zip format, then upload to prevent I/O bottlenecks that can severely degrade performance.
- **Dependency Management**: All scripts automatically download required models (MTCNN [126] for face detection, and dlib's shape predictor [144, 145] for landmark detection) on first execution, eliminating manual setup and ensuring reproducibility.
- **Output Organization**: Heatmaps are saved as PNG images with the same folder structure as the input dataset, facilitating easy comparison and integration with existing analysis workflows.

This contribution provides a comprehensive, automated, and reproducible framework for interpretability analysis of MAD models, enabling the systematic investigation of which facial characteristics are most discriminative for morphing detection across different model architectures and datasets.

Instituto Superior de Engenharia
Politécnico de Coimbra