

Face Authentication in the Deepfake Era: Strengthening Verification Against Spoofing Attacks

Dissertation

submitted to

Sorbonne Université

*in partial fulfillment of the requirements for the degree of
Doctor of Philosophy*

Author:

Sahar Hussein

<i>President/Examiner</i>	Dr. Marc Antonini	CNRS Research Center, FR
<i>Examiner/Reviewer</i>	Prof. Patrick Le Callet	University of Nantes, FR
<i>Examiner/Reviewer</i>	Dr. Antitza Dantcheva	Inria Research Center, FR
<i>Co-supervisor</i>	Mr. Fabien Aili	Docaposte Biometrics Lab, FR
<i>Thesis Director</i>	Prof. Jean-Luc Dugelay	Eurecom, FR

The research was conducted in the Digital Security Department at EURECOM and Docaposte R&D Center in Sophia Antipolis, France, from December 2021 to March 2025.

Abstract

Biometric face authentication leverages the unique biological features of an individual’s face, providing a secure and convenient alternative to traditional password-based authentication. With the widespread adoption of face verification in remote authentication services and portable devices, ensuring the robustness of these systems against spoofing attacks has become increasingly critical. While traditional biometric threat models primarily focus on vulnerabilities within verification pipelines, the rise of AI-generated deepfake technology introduces a new and sophisticated attack vector. Deepfakes enable real-time manipulation of facial images, posing a significant challenge to authentication security by spoofing verification systems.

This thesis addresses multiple aspects of face authentication, including face verification and attacks such as deepfake and injection attacks. It contributes to improving both the accuracy of biometric authentication systems and the robustness of deepfake detection algorithms, enhancing overall security.

The first contribution of this thesis is the introduction of an advanced face alignment method designed to improve verification accuracy by mitigating the effects of variations in head pose, facial expression, and illumination.

The second contribution focuses on understanding the threats posed by deepfake attacks. We analyze the quality of deepfakes generated by face reenactment methods and introduce a novel deepfake quality assessment protocol. This protocol systematically evaluates the video frame quality of face-reenactment techniques. Given the lack of standardized datasets for such assessments, we propose two video generation approaches utilizing 3D head models to create diverse and controlled evaluation scenarios.

Furthermore, we analyze the impact of beautification filters on deepfake detection systems, revealing significant vulnerabilities in state-of-the-art classifiers when subjected to such modifications.

To improve deepfake detection performance, we propose leveraging raw domain data as input, thereby reducing the impact of common image processing techniques such as compression and beautification filters. By constraining the distribution of real images, our approach enhances the model’s ability to

differentiate between genuine and manipulated content, improving detection accuracy in challenging scenarios.

Lastly, we investigate the role of compression artifacts in detecting digital replay attacks, where adversaries inject authentic video footage into the system via virtual camera software. We explore a novel strategy of bypassing the compression pipeline and directly capturing uncompressed image data from the user's device. This approach strengthens anti-spoofing mechanisms by exploiting the differences between uncompressed sensor data and compressed media typically used in injected attacks.

The findings and methodologies presented in this thesis contribute to the ongoing efforts to secure biometric authentication systems against evolving threats, advancing the field of deepfake detection and face verification security.

Résumé

L'authentification biométrique par reconnaissance faciale exploite les caractéristiques biologiques uniques du visage d'un individu, offrant une alternative sécurisée et pratique à l'authentification traditionnelle par mot de passe. Avec l'adoption généralisée de la vérification faciale dans les services d'authentification à distance et les dispositifs portables, il est devenu de plus en plus crucial de garantir la robustesse de ces systèmes face aux attaques par usurpation. Tandis que les modèles traditionnels de menaces biométriques se concentrent principalement sur les vulnérabilités des pipelines de vérification, l'essor de la technologie des vidéos hyper-truquées générées par IA introduit un nouveau vecteur d'attaque sophistiqué. Les vidéos hyper-truquées permettent la manipulation en temps réel des images faciales, posant un défi majeur pour la sécurité de l'authentification en dupant les systèmes de vérification.

Cette thèse aborde plusieurs aspects de l'authentification faciale, notamment la vérification faciale et les attaques telles que les attaques par des vidéos hyper-truquées et par injection. Elle contribue à améliorer à la fois la précision des systèmes d'authentification biométrique et la robustesse des algorithmes de détection des vidéos hyper-truquées, renforçant ainsi la sécurité globale.

La première contribution de cette thèse est l'introduction d'une méthode avancée d'alignement du visage, conçue pour améliorer la précision de la vérification en atténuant les effets des variations de la position de la tête, de l'expression faciale et de l'éclairage.

La deuxième contribution se concentre sur la compréhension des menaces posées par les attaques par les vidéos hyper-truquées. Nous analysons la qualité des vidéos hyper-truquées générés par des méthodes de réenactement facial et proposons un protocole novateur d'évaluation de la qualité des vidéos hyper-truquées. Ce protocole évalue systématiquement la qualité des images vidéo générées par ces techniques de réenactement facial. En raison du manque de jeux de données standardisés pour de telles évaluations, nous proposons deux approches de génération de vidéos utilisant des modèles 3D de tête afin de créer des scénarios d'évaluation diversifiés et contrôlés.

De plus, nous analysons l'impact des filtres d'embellissement sur les

systèmes de détection de vidéos hyper-truquées, mettant en évidence des vulnérabilités significatives dans les classificateurs de pointe lorsqu'ils sont soumis à de telles modifications.

Pour améliorer les performances de détection des vidéos hyper-truquées, nous proposons d'utiliser des données brutes comme entrée, réduisant ainsi l'impact des techniques de traitement d'image courantes telles que la compression et les filtres d'embellissement. En contraignant la distribution des images réelles, notre approche améliore la capacité du modèle à différencier le contenu authentique du contenu manipulé, augmentant ainsi la précision de détection dans des scénarios complexes.

Enfin, nous examinons le rôle des artefacts de compression dans la détection des attaques par replay numérique, où des attaquants injectent des vidéos authentiques dans le système via des logiciels de caméra virtuelle. Nous explorons une stratégie innovante consistant à contourner le pipeline de compression et à capturer directement les données d'image non compressées à partir de l'appareil de l'utilisateur. Cette approche renforce les mécanismes de lutte contre la fraude en exploitant les différences entre les données brutes des capteurs et les médias compressés généralement utilisés dans les attaques injectées.

Les résultats et méthodologies présentés dans cette thèse contribuent aux efforts en cours pour sécuriser les systèmes d'authentification biométrique contre les menaces évolutives et faire progresser le domaine de la détection de deepfake et de la sécurité de la vérification faciale.

Acknowledgements

This thesis owes its existence to the help, support, and inspiration of many people. First and foremost, I would like to express my deepest gratitude to my supervisor, Prof. Jean-Luc Dugelay, for his guidance, support, and patience over the years in helping make this thesis possible. I am especially thankful for his detailed feedback on my papers. Without him, this thesis would not exist, or at the very least, it would have been very different. I am also very grateful to Mr. Fabien Aili, my co-supervisor, for his guidance and encouragement throughout my research. I had the opportunity to work at the Docaposte R&D Center in France, and I want to thank all my colleagues there for their advice and support, which helped me grow as a researcher.

I would like to thank the thesis reviewers, Prof. Patrick Le Callet from the University of Nantes, France, and Dr. Antitza Dantcheva from the Inria Research Center, France, for their valuable comments on the thesis. I also wish to thank the thesis examiner, Dr. Marc Antonini, for his time and input.

I also want to thank my colleagues/friends at Eurecom for providing me always a friendly and supportive atmosphere. I am especially thankful to my amazing office mates and friends, Chiara, Nelida, Mira, Sameer, Alexandre, Simone, Abdel, and Andy, for making the journey more enjoyable. I would like to thank Sophie Salmon for her support during every step of the PhD.

I want to give special thanks to my close friends, especially Riikka and Yasmin, for their friendship, emotional support, and constant encouragement. Last but not least, I want to thank my family for always supporting me with love and care. I am especially grateful to my partner, who stood by my side through every step of this journey, not only in academia but also in life. I dedicate this thesis to my beloved family, my partner, Golzhin, and Henry, whose presence fills my life with joy and hope, and who remind me every day that the world is a place of light, love, and endless possibility.

Sahar Husseini Biot, 23 July 2025

Contents

Abstract	i
Résumé	iii
Acknowledgements	v
List of Figures	xiv
List of Tables	xvi
1 Introduction	1
1.1 Background	2
1.2 Motivation	4
1.3 Research Questions	8
1.4 Main Contributions and Thesis Structure	9
2 Literature Review	13
2.1 Face Recognition System	13
2.2 Points of Attack in Biometric Remote Authentication	16
2.3 Deep Generative Models	18
2.4 Face Manipulation in the Era of Deepfakes	19
2.4.1 Overview of Face-Swapping Methods	21
2.4.2 Overview of Face-reenactment Methods	22
2.5 Evaluation Techniques for Face Manipulation Methods	24
2.5.1 Evaluation Techniques for Face-Swapping Methods	24
2.5.2 Evaluation Techniques for Face Manipulation Methods	25
2.6 Deepfake Attacks and Countermeasures	27
2.6.1 Deepfake Detection	28
2.6.2 The Image Signal Processing (ISP) Pipeline and Its Impact on Deepfake Detection	29
2.7 Digital Replay Attack Detection	31

2.8	Compression Detection for Video Forensics: Effectiveness in Replay Attacks and Deepfake Detection	32
3	Enhancing Face Verification Algorithm	34
3.1	Introduction	34
3.2	Methodology	37
3.2.1	3D Face Model Reconstruction	38
3.2.2	Proposed Method	38
3.3	Experimental Evaluation	41
3.3.1	Datasets	41
3.3.2	Face Normalization	41
3.3.3	Face Verification Models	41
3.3.4	Comparisons With State-of-The-Art Methods	42
3.3.5	Ablation Study	43
3.4	Conclusions	43
4	Deepfake Quality Assessemnet	45
4.1	Introduction	45
4.2	Proposed Methodology	47
4.2.1	Protocol	47
4.2.2	Dataset Generation	48
4.2.3	Real Face Dataset	48
4.2.4	Synthesized Dataset of MetaHumans	49
4.3	Subjective Evaluation	51
4.4	Experiment and Results	53
4.5	Future Work	57
4.6	Conclusion	58
5	Effect of Beautification Filters on Deepfake Detectors	59
5.1	Introduction	59
5.2	Dataset	62
5.3	Experimental Setup and Results	63
5.3.1	Deepfake Detectors	63
5.3.2	Experimental Setup	64
5.3.3	Experimental Results	65
5.4	Subjective Evaluation	66
5.4.1	Subjective Evaluation Results	68
5.5	Conclusion and Discussion	69

6	RAW Data: A Key Component for Effective Deepfake Detection	71
6.1	Introduction	71
6.2	Proposed Method	73
6.3	Experiment	75
6.3.1	Experimental Setup	76
6.3.2	Experimental Results and Analysis	77
6.3.3	Limitations and Future Work	79
6.4	Conclusion	80
7	Towards Secure Authentication: Detecting Replay Attacks via Compression Artifacts	82
7.1	Introduction	82
7.2	Proposed Method	85
7.3	Experiment	85
7.3.1	Experimental Setup	85
7.3.2	Experimental Results	87
7.4	Conclusion	89
8	Conclusion and Future Directions	91
8.1	Conclusion	91
8.2	Directions for Future Research	93
	Author's Publications	95
	References	97

List of Figures

1.1	Diagram of a remote identity proofing process, where the user captures identity data and sends it to a trusted server. Various algorithms are employed to verify the authenticity of the data and protect against spoofing before the face comparison algorithm (e.i. face verification) confirms the user's identity. Upon successful verification, the user gains access to the service. Image adapted from [19].	4
2.1	Overview of facial recognition workflow: from registration to the authorization phase, including verification and identification.	14
2.2	Possible attacks on biometric systems (best viewed in color). Image adapted from [30].	17
2.3	Various deepfake generation methods ϕ_G manipulate the input image I_{rgb} based on conditions like audio, video frames, or text. Typically, detection models process ISP-transformed images. The ISP process starts with light focused on the CFA sensor, producing raw pixel values I_{RAW} , which undergo stages ϕ_{ISP} such as white balance and noise removal to yield the final RGB output.	20
2.4	Face-reenactment evaluation protocols: self-reenactment (a), cross-reenactment (b), and our proposed evaluation protocol (c). In this illustration, X and Y represent identities, while S, D, T, and GT correspond to source, driving, target, and ground-truth, respectively. Additionally, m_1 and m_2 indicate the movement of the source and driving, respectively	26

3.1	Overview of Face Normalization by AlignFace: For each image, 3DMM coefficients, including identity (α), expression (β), texture (τ), illumination (γ), and head pose (p) are extracted using the R-Net model. The image x_b is then normalized to produce $x_{b'}$, aligning its expression, head pose, and lighting conditions with those of x_a . During normalization process $x_{b'}$'s identity and texture coefficients are iteratively updated (n iterations) while keeping the parameters of the FR and R-Net models frozen. Although images generated as $x_{b'}$ closely follow the distribution of real images, discrepancies might exist between the distributions of generated $x_{b'}$ and real faces x_a . To ensure accuracy at the face verification phase, the FR model used for extracting face embeddings for $x_{b'}$ is fine-tuned, denoted as FR_* . x_a and $x_{b'}$ represent different identities.	35
3.2	AlignFace Efficacy in Normalizing Pose, Expression, and Illumination. Displayed are the original image x_a , the comparative image x_b , and AlignFace's reconstructed image $x_{b'}$. This demonstrates AlignFace's ability to effectively transfer the extraneous conditions of x_a to x_b while preserving the unique identity features in x_b . Note that in the examples on the left, the identities are the same, whereas in the examples on the right, the identities are different.	39
4.1	Proposed protocol (a). Examples of the source image, driving video frame, generated frame, and corresponding ground-truth provided by our proposed protocol for both the real (b) and synthesized (c) datasets.	45
4.2	Multiview RGB images and their corresponding depth maps utilized to inverse project pixels into point clouds (a). The resulting reconstructed 3D head model (b). Rendered images of 3D models from desired angles(c).	49
4.3	Head (left) and Face (right) Control Rig Boards enabling adjustment of pose and facial expression (a). Two MetaHumans with identical facial expressions and head poses (b).	50
4.4	Pose transferability evaluation using our proposed protocol. The figure presents the results of the image-based overall satisfaction subjective test scores (IS_{JOD}) for different head degrees, along with the corresponding quantitative scores such as SSIM, CSIM, and AKD, computed using ground-truth data following our proposed protocol.	57

4.5	Confusion matrix depicting the correlation of metrics within Real (left) and synthesized (right) datasets	58
5.1	Pipeline of the proposed method. A subset of 464 videos (50% Real and 50% Fake) are selected. Each video is uploaded to the social network Instagram, where one of the four different filters is randomly selected and applied to it. The four filters uniformly appear in the Celeb-DF-B database. The final database has a size of 928 videos and it is used to perform a human-based deepfake detection and to evaluate the robustness of three SotA AI-based detectors.	61
5.2	Frames extracted from four distinct videos within the Celeb-DF-B database are depicted here.	63
5.3	Result of the evaluation on Celeb-DF-B with the the 3 detectors. a) The video-level AUC of the ROC curve and b) The False Negative Rate for different classification score thresholds . . .	66
5.4	The histograms of the classification score for each deepfake detector on real and deepfake videos. In blue (resp. orange) the non-beautified (resp. beautified) subset of Celeb-DF-B. CADDM and FTCN tend to see beautification as additional face manipulation (histogram shifted to the right) whereas RECCE finds fake videos more realistic after the beautification process (histogram shifted to the left)	67
5.5	Each user's accuracy before and after applying filters.	69
6.1	Overview of generating a RAW Self-Blended Image ($I_{\text{RAW_SB}}$). A base image I_{base} is fed into the Source-Target Generator (STG) and the Mask Generator (MG). The STG produces pseudo source and target images from the base image using various image augmentations, while the MG creates a blending mask from facial landmarks and deforms it to enhance mask diversity. The source and target images are then blended with the mask and input into the Inverse ISP pipeline to reconstruct the raw format of the RGB input image.	74
6.2	Example images of I_{base} , I_{RAW} , I_{SB} , and $I_{\text{RAW_SB}}$. The I_{base} samples are sourced from FF++ dataset. I_{RAW} and $I_{\text{RAW_SB}}$ are transformed from I_{base} and I_{SB} respectively, using inverse ISP model.	80

7.1	Illustration of various input streams to a remote face authentication service. The input can originate from different scenarios. In the first scenario, the face of a genuine user is provided to the service, granting access to the application upon successful authentication. In the second scenario, a deepfake injection attack is performed. Here, the attacker generates a real-time video mimicking the victim's expressions and head movements using a single image. This video is streamed via virtual camera software to imitate a legitimate webcam feed, deceiving the authentication system. In the third scenario, the attacker uses either a single image or a pre-recorded video of the victim. The virtual camera streams a genuine video of the victim that lacks visible artifacts. Our main goal is to exploit compression artifacts for detecting digital replay attack.	83
7.2	ROC curves for different codecs	87
7.3	ROC curves for H.264 codec with different QP levels	89
7.4	Guided Backpropagation and Grad-CAM visualizations for uncompressed and compressed video frames, highlighting the areas of support for the compressed category.	90

List of Tables

3.1	Comparative analysis on benchmark datasets: Accuracy metrics for 1:1 verification are presented for LFW, CFP, and AgeDB datasets. For the IJB-B dataset, we report the TAR@FAR=0.01%. Red: best, blue: second-best.	42
4.1	Summary of subjective evaluation methods.	51
4.2	Evaluation results for cross-identity reenactment for real dataset.	53
4.3	Evaluation Results for Cross-identity Reenactment for synthesized MetaHuman Dataset.	54
5.1	Characteristics of the selected Instagram filters. Traits modifications were assessed by visual inspection of pixel differences between original and filtered images.	63
5.2	Type of data from FF++ seen by each deepfake detectors	65
5.3	AUC score of each detector w/o and w/ beautification on Celeb-DF-B	66
5.4	Subjective evaluation results on Celeb-DF-B dataset for Beautified and Non-beautified videos	68
6.1	Cross-Dataset Evaluation on CDF, DFD, and DFDC Datasets Using raw data for deepfake detection is a novel approach with no direct comparisons in the field. We compare our method's performance against established RGB domain techniques. Our methodology includes an inverse ISP model to convert RGB images to raw format for analysis. The model is trained exclusively on the high-quality FF++ dataset using only real data. Results from previous methods are cited from their original papers. Bold values indicate the best performance, while underlined values denote the second-best performance.	78
6.2	Cross-manipulation evaluation on FF++.	79

7.1	Performance metrics of the model on trained codecs (H.264, H.265, VP8, VP9) and an unseen codec (MPEG-4), demonstrating its detection accuracy and generalization capability. . .	86
7.2	Performance metrics of the model for H.264 compression across varying quantization parameters (QP).	89

List of Abbreviations

Abbreviation	Definition
2FA	Two-Factor Authentication
3DMM	3D Morphable Models
AED	Average Euclidean Distance
AIGC	Artificial Intelligence Generated Content
AKD	Average Keypoint Distance
AUC	Area Under the Curve
AUs	Action Units
CDF	Celeb-DF v2 dataset
CFA	Color Filter Array
CNNs	Convolutional Neural Networks
CSIM	Cosine Similarity
DaGAN	Depth-Aware Generative Adversarial Network
DFD	Deep-Fake Detection dataset
DFDC	DeepFake Detection Challenge public test set
ED	Encoder-Decoder
EFNB4	EfficientNet-b4
FAU	Facial Action Unit
FID	Fréchet Inception Distance
FF++	FaceForensics++
FNR	False Negative Rate
FR	Facial Recognition
FV	Face Verification
FVD	Fréchet Video Distance
GANs	Generative Adversarial Networks
GOP	Group of Pictures
HSV	Hue Saturation Value
ISP	Image Signal Processor
JOD	Just-Objectionable-Difference
LIA	Latent Image Animator
LMD	Linear Motion Decomposition

Abbreviation	Definition
LPIPS	Learned Perceptual Image Patch Similarity
LSC	Lens Shading Correction
MLE	Maximum Likelihood Estimation
MKR	Missing Keypoint Rate
PSNR	Signal-to-Noise Ratio
ROC	Receiver Operating Characteristic
ROIs	Regions Of Interest
SoTA	State-of-the-Art
SSIM	Structural Similarity Index Measure
sRGB	Standard RGB
SVMs	Support Vector Machines
TEE	Trusted Execution Environment
VAEs	Variational Autoencoders
WB	White Balance

Chapter 1

Introduction

In today's fast-paced world, the way we interact with technology has transformed many aspects of our daily lives. Tasks that once required physical presence and effort, such as transferring money, scheduling a doctor's appointment, or purchasing a ticket, can now be accomplished effortlessly with just a few taps on a smartphone. This shift is part of a broader digital revolution that has redefined convenience and accessibility, saving countless hours and streamlining everyday processes. However, with this increased reliance on digital systems, there arises a pressing need to address the security challenges associated with these conveniences. Whether it's safeguarding financial transactions, protecting personal data, or ensuring the integrity of remote interactions, robust authentication mechanisms have become a cornerstone of modern digital infrastructure. Without these measures, the very technologies that empower us could become sources of vulnerability, exposing users to risks like fraud, identity theft, and unauthorized access.

When designing an authentication system, it is crucial to consider both the security of the system and the convenience of the user. Various methods can be employed to secure authentication systems. The most basic method is password-based authentication, where a user employs a password to gain access to the system. Passwords should be strong, which can make them difficult to remember, and should be changed regularly to ensure security. Simple passwords are at risk of being stolen and used by impostors.

Another method is token-based authentication, where users receive a token that serves as proof of identity. This token can be a physical device, like a smart card, or a digital token generated by an authentication application. While token-based authentication offers a higher level of security, it requires additional actions from the user which can be inconvenient; users need to carry a physical device, like a smart card, or use a digital token generated by an authentication application, which may not always be user-friendly.

Two-Factor Authentication (2FA) provides a balance of convenience and security. It typically combines something the user knows, like a password, with something the user has, such as a code sent to a mobile device. However, the additional step can be cumbersome. Users must have access to the second factor, such as a mobile device to receive the code, which can be inconvenient if the device is unavailable or if the user is in an area with poor reception.

Finally, biometric authentication leverages unique biological characteristics of an individual, such as fingerprints, face, or iris scans, eliminating the need to remember anything. This method offers several significant benefits, including enhanced security due to the difficulty of replicating biometric traits, improved user convenience as it bypasses the need for passwords or tokens, and a seamless user experience. As a result, biometric authentication is increasingly integrated into remote authentication services and portable devices, including laptops and smartphones, to provide a secure and user-friendly solution for access control.

In this thesis, we focus on biometric authentication, with a particular emphasis on face authentication services. Our study examines the vulnerabilities of face recognition systems to various external factors and attacks, with a specific focus on deepfake attacks.

1.1 Background

The human face has served as the primary means of recognition among individuals since the dawn of humanity. Beyond its pivotal role in social interaction, it encodes a wealth of unique biometric information, including facial structure, texture, and expressions, which together distinguish one person from another. Facial authentication services utilize these distinctive characteristics to perform both verification and identification tasks. Identity verification seeks to answer the question, ‘Is this person who they claim to be?’ It involves confirming whether a face corresponds to the claimed identity. In contrast, identity identification addresses the question, ‘Who is this person?’ by determining the identity of an unknown individual through comparison with a database of known faces.

Early face verification techniques relied on geometric measurements of facial features, such as the distances between the eyes, nose, and mouth. These methods were later enhanced by statistical and texture-based approaches, which relied on manually extracting features to enhance accuracy. The advent of machine learning, particularly deep learning, revolutionized face recognition by achieving exceptional accuracy and robustness.

The modern face recognition system utilizes Convolutional Neural Net-

works (CNNs), converting face images into compact latent space representations, known as embeddings, that cluster similar identities together. This powerful encoding effectively encapsulates the core characteristics of an individual's identity, enabling highly reliable recognition systems.

Despite significant advancements, challenges persist in face recognition when images are affected by uncertainties arising from variations in head pose, expression, illumination, and other external factors. These variations can obscure critical identity-related features, resulting in unreliable and error-prone representations. To address these challenges, recent approaches have focused on enhancing the preprocessing of input images before they are passed to face recognition models [1], [2]. Furthermore techniques such as optimizing loss functions [3]–[5] have been explored to enhance the accuracy of the models.

Security is a key consideration for biometric authentication systems. Authentication service providers must ensure that the individual on the client side is indeed the legitimate user and not an impostor. Without robust security measures, sensitive client information can be compromised. As face verification technology becomes increasingly integrated into authentication systems, it has also become a more attractive target for malicious actors aiming to impersonate legitimate users and bypass biometric authentication mechanisms. One common form of attack is a presentation attack, where attackers use physical media, such as printed photos or videos, to deceive the system and gain unauthorized access [6]. Similarly, injection attacks leverage authentic video footage, often sourced from social media, and use virtual cameras or sophisticated software to inject it into the system, further complicating detection as the video appears genuine and lacks typical manipulation artifacts [7].

Alongside traditional attacks, the rise of deepfake technology has introduced a more sophisticated and rapidly evolving threat. Deepfake attacks utilize advanced deep learning methods, including adversarial training [8] and, more recently, diffusion models [9], to create hyper-realistic synthetic media capable of convincingly impersonating individuals. Modern deepfake algorithms have advanced to the point where they can manipulate existing images in real-time, further complicating the ability of remote authentication systems to differentiate between genuine and fabricated content.

To effectively mitigate the risks posed by deepfakes, it is good practice to first analyze deepfake content and evaluate its quality based on various factors, such as face movements or lightening conditions. Additionally, it is essential to examine how image post-processing techniques influence the quality and characteristics of deepfakes.

In the recent years, a range of detection methods have been developed,

evolving from early handcrafted feature-based techniques [10], [11] to modern deep learning approaches [12]–[16], and more recently, hybrid models [17], [18]. Existing methodologies in deepfake detection typically rely on a supervised approach. This involves developing a real vs fake image classifier by assembling a large dataset of generated images from multiple generative models and training a binary classifier. However, in practical scenarios, the specific techniques used for facial manipulation are unknown beforehand, and access to the attacker’s model is typically unavailable. Despite achieving high detection accuracies, approaching 98%, these classifiers are prone to overfitting.

1.2 Motivation

In a remote face verification system, users capture an image or short video of themselves, along with an image of their identity proof card, using potentially untrusted devices and upload them to a secure, centralized server for verification. This trusted server processes, stores, and validates biometric data using encryption protocols and robust access controls to ensure privacy and data integrity. Unlike traditional local verification methods, such as iPhone’s FaceID, which operates within a Trusted Execution Environment (TEE) on the user’s device, this approach leverages a centralized infrastructure to handle authentication securely. Figure 1.1 illustrates the generic remote verification process.

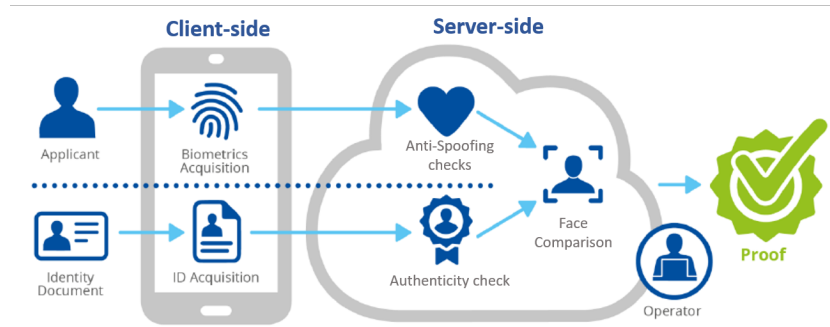


Figure 1.1: Diagram of a remote identity proofing process, where the user captures identity data and sends it to a trusted server. Various algorithms are employed to verify the authenticity of the data and protect against spoofing before the face comparison algorithm (e.i. face verification) confirms the user’s identity. Upon successful verification, the user gains access to the service. Image adapted from [19].

Studies have highlighted the sensitivity of face verification systems to variations in head pose, facial expressions, illumination, and other external

factors. This vulnerability must be carefully considered in remote authentication systems, as data collected by the end users often exhibit different lighting conditions, head pose, or expressions compared to their ID card images, presenting significant challenges for accurate matching and verification

To address these challenges, an effective strategy is to optimize input images before processing them with face recognition models. Among the available preprocessing techniques, input image normalization stands out as a robust solution, targeting critical issues such as lighting inconsistencies, variations in facial expressions, and differences in head poses between input image pairs. Illumination normalization reduces the impact of lighting conditions on facial appearance, ensuring consistent texture and color [1], [2], while head pose normalization aims to frontalize the face, and expression normalization seeks to neutralize facial expressions. In recent years, deep learning-based solutions [20], [21] have addressed both face frontalization and neutralization of facial expressions, leveraging the capabilities of neural networks. Despite showcasing promising synthesis quality, these methods encounter challenges in preserving face identity details, especially in scenarios with substantial pose variations.

In this thesis we propose an innovative normalization algorithm designed for preprocessing input images in the context of face verification. Diverging from conventional methods, our approach places a distinctive emphasis on achieving consistency in head pose, expression, and illumination conditions between two images, avoiding an exclusive focus on the normalization of extraneous elements at specific values. Specifically, our methodology involves estimating the head pose, expression, and illumination conditions in one image, followed by the reconstruction of the second image to align with the same head pose, expression, and illumination conditions while preserving its own unique identity features. By adopting this approach, our algorithm allows the verification process to concentrate solely on identity evaluation, unaffected by variations in non-essential extraneous and synthesized features.

In addition to ensuring the accuracy of the face verification algorithm, it is crucial to prioritize the security of the authentication system. This means that, before applying face verification, input images should be inspected by anti-spoofing algorithms to ensure that the image sent by the end user is authentic and unmodified. In other words, the system should verify that no spoofing attack is being attempted from the client side.

Recent advancements in deepfake algorithms have enabled attackers to create highly realistic images and videos, facilitating impersonation attacks and posing significant challenges to authentication systems. By obtaining a single image of a victim, often sourced from publicly available platforms like social media, and leveraging deepfake technologies such as face swap

and face-reenactment, attackers can generate real-time manipulated videos. These videos mimic the facial expressions and head movements required by authentication services and can be used for presentation or injection attack.

In a presentation attack scenario, the deepfake video is typically generated on a separate device, such as a laptop, and then displayed to the authentication sensor. This allows the attacker to deceive the system by presenting artificial biometric traits, effectively simulating the victim’s appearance. Similarly, injection attacks take this a step further by streaming these manipulated videos directly to authentication systems via virtual camera software (e.g., OBS), bypassing physical camera sensors altogether.

These attack strategies exploit the vulnerabilities in traditional security measures, making deepfake detection a critical and urgent area of research to safeguard biometric authentication systems.

Effective deepfake detection begins with a thorough understanding of deepfake generators. Analyzing the image quality of deepfakes and distinguishing these from genuine images is crucial, as it can significantly enhance the ability to identify manipulation techniques.

Since deepfakes are entirely or partially synthetic images that did not exist prior to their creation, evaluating their quality is inherently challenging. The challenge lies in the lack of a ground truth, making direct comparison to real images difficult. In recent research, efforts have been made to address this by employing metrics that do not depend on explicit ground-truth comparisons. These metrics often utilize pretrained network feature extraction and aim to provide reliable assessments. However, despite their usefulness, these metrics may not capture all relevant aspects of quality assessment. For instance, they may overlook finer details, such as pixel-level quality, which are essential for a comprehensive evaluation and effective detection of deepfakes artifacts, including warping and blending irregularities.

To address this issue, this thesis introduces a novel protocol for the quantitative evaluation of images generated by face-reenactment techniques. This protocol enable us to objectively assess the image quality of various face-reenactment methods across different head poses and facial expressions. Our analysis reveals that the quality of deepfake images has significantly improved over time, with these images increasingly resembling real ones. Older deepfakes struggled to accurately replicate extreme head movements and facial expressions. However, with recent advancements in this technology, our quantitative results demonstrate that modern deepfakes can now replicate facial and head movements with remarkable accuracy, making them increasingly difficult to distinguish from genuine images.

Another important aspect is assessing the impact of various image processing steps on deepfake quality and characteristics. Specifically, it is crucial to

analyze whether deepfake artifacts become more visible or are concealed when applying processes such as compression, contrast adjustment, or beautification filters. Understanding this relationship can provide valuable insights into how post-processing affects deepfake detection. Recent studies have highlighted vulnerabilities in deepfake detectors when exposed to certain post-processing methods [22]–[25]. Building on these findings, this thesis focuses specifically on examining the impact of beautification filters on the accuracy and robustness of deepfake detection systems. Our findings reveal a significant decline in detection accuracy and indicate that various image processing steps can obscure key indicators of forgery, resulting in inaccurate decisions.

Several publicly available deepfake detection algorithms have been developed, spanning from early handcrafted feature-based techniques to advanced deep learning approaches. While these methods have shown effectiveness in identifying artifacts within the training data, they often suffer from overfitting, which limits their generalizability to unseen scenarios.

Given the vulnerabilities of deepfake detectors to image processing techniques and their tendency to overfit, coupled with the continuous advancements in deepfake quality, this thesis proposes redefining the boundary between real and fake images. Specifically, we suggest narrowing the definition of authentic samples to a stage closer to the raw radiance of the scene as captured by the camera sensor, prior to any transformations introduced by the Image Signal Processor (ISP).

Current detection models often struggle to differentiate between real and fake images because the existing definition of real images includes both raw content captured by camera sensors and content processed through various stages of image and video enhancement, including both linear and nonlinear adjustments. As a result, images are considered real even after undergoing multiple processing operations such as denoising, compression, deblurring, and white balance adjustments. These ISP processing steps are designed to produce aesthetically pleasing images for human viewers, but they pose significant challenges for deepfake detection. The issue arises from the fact that each device employs a unique ISP pipeline with distinct enhancement blocks, which obscure subtle cues that are crucial for detecting deepfakes. This variability forces detection models to adapt to unseen ISP configurations, making it more difficult to accurately identify real images.

In this thesis, we focus on deepfake detection in the context of injection attacks. In such scenarios, an attacker may employ an embedded sensor and a deepfake algorithm to generate manipulated video content in real-time. Ideally, if the device’s sensor is genuinely used by the end user, the captured images or videos should be free from any post-processing steps and retain their original, unaltered form.

We also analyzed the feasibility of detecting compression artifacts and distinguishing between compressed and non-compressed frames. This investigation stems from the hypothesis that if the end user’s device captures uncompressed frames, the absence of compression artifacts would facilitate a clearer and more accurate distinction between authentic and injected frames. Since most attacker-sourced videos—often obtained from the internet—are typically compressed and inherently exhibit compression artifacts, detecting these artifacts serves as a key indicator for identifying injection attacks.

1.3 Research Questions

The main contributions of this thesis focus on enhancing and fortifying the performance of face verification systems in authentication applications. This research aims to improve the robustness of these systems under diverse conditions, including variations in head pose, facial expressions, and lighting environments. Ensuring robustness under such challenges is critical for maintaining the reliability of face verification systems in real-world scenarios.

In parallel, the accuracy of face verification systems is increasingly threatened by the emergence of deepfakes. To address this growing concern, this thesis also examines deepfake image quality from multiple perspectives, such as variations in head pose and expressions. This analysis provides valuable insights into the capabilities of deepfake generators, enabling a better understanding of their mechanisms and aiding in the development of robust defenses for authentication systems. Furthermore, the impact of beautification filters on the accuracy of deepfake detectors is investigated, highlighting how such post-processing techniques can affect detection performance.

Building on these findings, this thesis proposes a novel deepfake detection algorithm tailored for injection attack scenarios, addressing a critical vulnerability in authentication systems.

Comprehensive experiments are conducted, and the results are analyzed, utilizing diverse evaluation metrics to align with international standards and facilitate meaningful comparisons with prior work.

The central objectives of this thesis are outlined in the following Research Questions(RQ):

- RQ1: Does aligning the head pose, expression, and lighting conditions of image pairs improve the face verification model’s accuracy under diverse scenarios?
- RQ2: How can we leverage 3D environments to address the lack of ground truth in deepfake detection and facilitate the evaluation of fine

details in deepfake images?

- RQ3: Do beautification filters undermine the effectiveness of existing deepfake detection systems?
- RQ4: Can the use of RAW image data help in developing a more robust deepfake detector that remains resilient to various image processing techniques, such as compression, beautification, and stylization?
- RQ5: Can we distinguish between raw uncompressed video frames and compressed video frames by analyzing compression artifacts, to enhance the detection of digital replay attacks in face authentication?

1.4 Main Contributions and Thesis Structure

Thanks to the growing interest in biometric authentication systems, significant research efforts have been devoted to enhancing the accuracy and security of face verification systems. In this thesis, we propose a novel alignment algorithm aimed at improving the accuracy of face verification models under varying head poses, facial expressions, and lighting conditions.

Given the growing threat posed by deepfakes, this thesis also evaluates the quality of deepfakes from multiple perspectives, including head pose and facial expression, to better understand their characteristics and the capabilities of modern deepfake generation algorithms. Furthermore, we analyze the vulnerabilities of current deepfake detectors, with a specific focus on their susceptibility to post-processing techniques such as beautification filters, which can obscure deepfake artifacts. Based on our findings, we introduce a new deepfake detection approach that demonstrates robustness against various image processing techniques, addressing critical challenges in distinguishing between authentic and manipulated content.

The thesis is organized into eight chapters, with a brief summary of each provided below:

- In **Chapter 1**, we introduced the thesis and presented the research questions that motivated this work.
- In **Chapter 2**, we provide a comprehensive literature review on advancements in face recognition systems and the potential attacks that can occur during remote authentication. Special emphasis is placed on deepfake attacks, including their generation, quality assessment, and detection methods. Additionally, we explore the various stages of the

Image Signal Processing (ISP) pipeline and position our contributions in the context of the current state-of-the-art.

- In **Chapter 3**, We present our proposed face alignment algorithm designed for preprocessing input images in the context of face verification. our methodology involves estimating the head pose, expression, and illumination conditions in one image, followed by the reconstruction of the second image to align with the same head pose, expression, and illumination conditions while preserving its own unique identity features. By adopting this approach, our algorithm allows the verification process to concentrate solely on identity evaluation, unaffected by variations in non-essential extraneous and synthesized features.
 - [P1] **S. Hussein** and J. -L. Dugelay, “**Alignface: Enhancing face verification models through adaptive alignment of pose, expression, and illumination,**” in 2024 IEEE International Conference on Image Processing (ICIP), IEEE, 2024, pp. 3243–3249.
- In **Chapter 4**, we introduce our proposed dataset and protocol for assessing the quality of face-reenactment deepfakes. Using this protocol, we evaluate various deepfake generation methods with well-established metrics such as SSIM and LPIPS. Additionally, we conduct a subjective evaluation to determine whether quantitative results align with qualitative assessments. To gain a comprehensive understanding of deepfake image quality, we further analyze the performance of these generators under varying movements and lighting conditions.
 - [P2] **S. Hussein**, J. -L. Dugelay, F. Aili, and E. Nars, “**A 3d-assisted frame-work to evaluate the quality of head motion replication by reenactment deepfake generators,**” in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2023, pp. 1–5.
 - [P3] **S. Hussein** and J. -L. Dugelay, “**Metahumans help to evaluate deepfake generators,**” in 2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP), IEEE, 2023, pp. 1–6.
 - [P4] **S. Hussein** and J.-L. Dugelay, “**A comprehensive frame-work for evaluating deepfake generators: Dataset, metrics performance, and comparative analysis,**” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 372–381.

- In **Chapter 5**, we present our study analyzing the impact of beautification filters on three deepfake detectors. Our findings reveal a notable drop in video-level AUC performance, demonstrating how social media beautification filters can either enhance the perceived authenticity of fake videos or make real videos appear fake based on these detectors. Additionally, we conducted a user study to assess whether beautification filters pose challenges for human observers in distinguishing between real and deepfake videos, further highlighting the complexities introduced by such post-processing techniques.
 - [P5] A. Libourel, **S. Hussein**, N. Mirabet-Herranz, and J. -L. Dugelay, “**A case study on how beautification filters can fool deepfake detectors,**” in IWBF 2024, 12th IEEE International Workshop on Biometrics and Forensics, 2024.
- In **Chapter 6**, we present our proposed method for deepfake detection in the context of injection attacks. This approach builds on the evaluation results and experiments conducted on deepfake generation and detection. Our method redefines the boundary between real and fake images by narrowing the definition of authentic samples. Specifically, we advocate for focusing on a stage closer to the raw radiance captured by the camera sensor, prior to any transformations introduced by the Image Signal Processor (ISP).
 - [P6] **S. Hussein** and J. -L. Dugelay, “**Raw data: A key component for effective deepfake detection,**” in ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.
- In **Chapter 7**, we present our proposed method which investigates whether providing uncompressed video access to face anti-spoofing service providers can enhance the detection of injected versus authentic video streams. Building upon this, we propose bypassing the compression step and directly capturing uncompressed image data from the user’s device during authentication.
 - [P7] **S. Hussein** and J. -L. Dugelay, “**Towards secure authentication: Detecting replay attacks via compression artifacts,**” Submitted to IWBF, 2025.

Finally, in **Chapter 8** we conclude the thesis by summarizing the key findings and contributions of this work. We also discuss potential future directions, including improving alignment algorithms, refining datasets and protocols

for deepfake quality assessment, and addressing emerging threats posed by advancements in deepfake generation techniques. These directions aim to provide a solid foundation for further research in improving the security and reliability of face verification and deepfake detection systems.

Chapter 2

Literature Review

In this chapter, we describe the background topics relevant to this thesis. We begin with an overview of face recognition systems, followed by a discussion on various types of spoofing attacks targeting facial authentication systems, with a particular emphasis on deepfake generation, quality assessment, and detection.

2.1 Face Recognition System

Facial Recognition (FR) technology leverages images or video frames to identify or verify a person's identity. This technology facilitates two primary functions: verification and identification, which are distinguished by their matching techniques. Verification employs a one-to-one matching process to confirm if the query face, known as the probe, matches a claimed identity. On the other hand, identification employs a one-to-many matching process to define the actual identity of the probe by comparing it against multiple identities enrolled in the system.

The face recognition process is composed of several steps, categorized into two phases: the enrollment phase and the recognition phase. During the enrollment phase, biometric data is captured and securely stored. In the recognition phase, newly captured biometric data is compared against the previously stored data to either verify or identify the individual. The operational steps of FR systems are detailed below and visualized in Figure 2.1:

1. **Data capture:** A camera sensor captures an image or video of an individual, referred to as the probe.
2. **Face detection:** A face detector locates the face within the captured

image.

3. **Image preprocessing:** The detected face is adjusted—typically cropped and aligned—to accommodate variations in head pose, facial expressions, and lighting conditions, making it suitable for recognition.
4. **Feature extraction:** Distinctive facial features are extracted from the preprocessed image to create a template which is, also known as a feature embedding or latent space representation. The recent FR systems commonly employ Convolutional Neural Networks for this task.
5. **Template matching:** In the verification phase, the FR system compares the probe's template to the corresponding target template based on a predefined similarity metric and a threshold. For identification, the system compares the probe's template against all stored templates and selects the best match based on similarity scores.

Face recognition technology has advanced significantly in recent years, with various methods proposed to enhance different aspects of the process. Additionally, numerous feature extraction techniques have been developed. Among these, State-of-the-Art (SoTA) approaches stand out by focusing on mapping

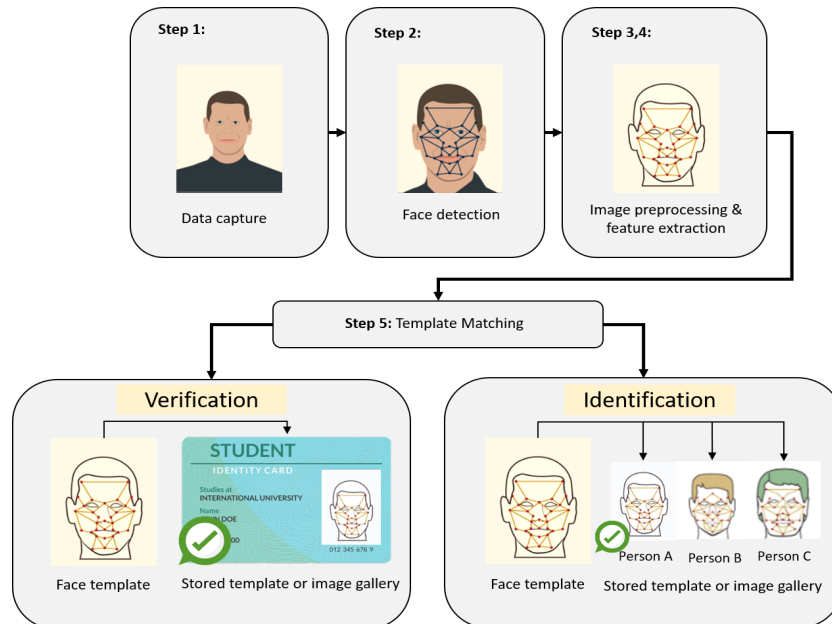


Figure 2.1: Overview of facial recognition workflow: from registration to the authorization phase, including verification and identification.

facial images to latent space representations that accurately capture an individual's identity, effectively clustering the representations of the same person together. Challenges arise when face images contain uncertainty, arising from variations in extraneous elements such as head pose, expression, and illumination between image pairs, which can obscure crucial identity information, resulting in learned representations unreliable and error-prone. To enhance face recognition algorithms and tackle challenges posed by these factors, recent approaches have adopted a variety of strategies, which generally fall into two main categories:

- Incorporating image quality factors, such as head pose and illumination, into the loss function during the training of the feature extractor.
- Implementing preprocessing techniques before feature extraction to normalize elements such as head pose, facial expression, and illumination conditions in image pairs.

In the first approach, various loss functions, including contrastive loss [26], triplet loss [27] are investigated, to enhance the discriminative power of the algorithms under varying image conditions. However, a notable transformation has taken place more recently, as researchers have shifted their focus towards optimizing loss functions to reduce the demand for extensive training data. Central to these innovative methods is the adoption of margin-based softmax loss functions for training Face Recognition models. The incorporation of a margin is crucial in these loss functions, as it empowers the learned features to become more discerning and discriminative. Pioneering contributions to this field include SphereFace [3], CosFace [4], and ArcFace [5], each introducing distinct variations of margin functions. However, these loss functions share a common limitation: they rely on fixed margin values that do not account for inherent variations, such as differences in image quality, within the same class. This limitation has prompted the development of solutions based on adaptive margin loss. MagFace [28] integrates the quality attributes of a face image sample—such as head pose, sharpness, and illumination—directly into the margin calculation. This approach aims to cluster high-quality samples (e.g., frontal faces) in a compact region around the class centers, while positioning low-quality samples farther from these centers. This approach helps prevent the algorithm from overly emphasizing noisy or difficult samples, which could otherwise compromise its effectiveness and lead to overfitting.

In the second approach the focus has been on optimizing input images before they are fed into the FR model or subjected to feature extraction. A crucial preprocessing step in this regard is **face normalization**, which addresses various aspects, including illumination, expression, and head pose

normalization. Illumination normalization seeks to reduce the impact of lighting conditions on facial appearance, ensuring that the texture and color of the face remain consistent [1], [2]. On the other hand face frontalization is aimed at transforming facial images into a frontal view, even in the presence of potential occlusions. In recent years, deep learning-based solutions [20], [21] have addressed both face frontalization and neutralization of facial expressions, leveraging the capabilities of neural networks. Despite showcasing promising synthesis quality, these methods encounter challenges in preserving face identity details, especially in scenarios with substantial pose variations.

In **Chapter 3**, we present the results from one of the author’s publications [P1], which introduces a normalization algorithm for preprocessing input images prior to face verification. This approach emphasizes achieving consistency in head pose, facial expression, and illumination conditions between two images. Instead of merely normalizing extraneous elements to predefined values, the head pose, illumination, and expression of the first image are normalized to match those of the second image.

2.2 Points of Attack in Biometric Remote Authentication

The integration of biometric authentication, particularly face verification algorithms, into remote authentication systems has gained significant popularity due to their user-friendliness and ease of use. However, the growing adoption of such systems has also made them an appealing target for attackers, underscoring the importance of implementing robust security measures.

Attacks can occur at various stages of the system’s operation, including biometric registration and authorization processes. Given the similarity between these phases, especially the incorporation of anti-spoofing mechanisms, this analysis primarily focuses on the authorization stage.

Figure 2.2 provides a visual representation of the threat model for a biometric system integrated with cloud computing, highlighting potential points of vulnerability within the system.

During the authorization phase, the user’s device sensors capture biometric data from the physical environment and transform it into digital representations. These representations are subsequently processed by data processing systems, such as machine learning algorithms, which operate within the application and on the server.

Attacks during the data capture process can occur in either the physical or digital domain [29], [30]. In the physical domain, a presentation attack

involves an attacker using physical objects such as printed photos, masks, or 3D models to deceive the sensor and mimic the legitimate user. Conversely, in the digital domain, attackers bypass the device's embedded hardware sensor entirely. This type of attack may involve a virtual camera or advanced manipulation software to inject fabricated or manipulated biometric data directly into the system.

While these attacks predominantly occur on the client side, vulnerabilities also exist on the application or server side. For instance, biometric data stored in the database can be overridden or modified, potentially compromising the system's integrity. Additionally, processors such as anti-spoofing method or face verification algorithms can be overridden or have their effectiveness undermined, leaving the system susceptible to exploitation.

In this thesis, we assume that both the application running on the user's device and the server managing the data are securely protected through established security engineering practices. With a focus on authentication services and the escalating threat posed by deepfakes, this work specifically addresses deepfake digital injection attacks.

To build a comprehensive understanding of deepfakes, the subsequent subsections provide an overview of the deepfake generation process, quality assessment methods, and delve into deepfake attacks alongside detection techniques.

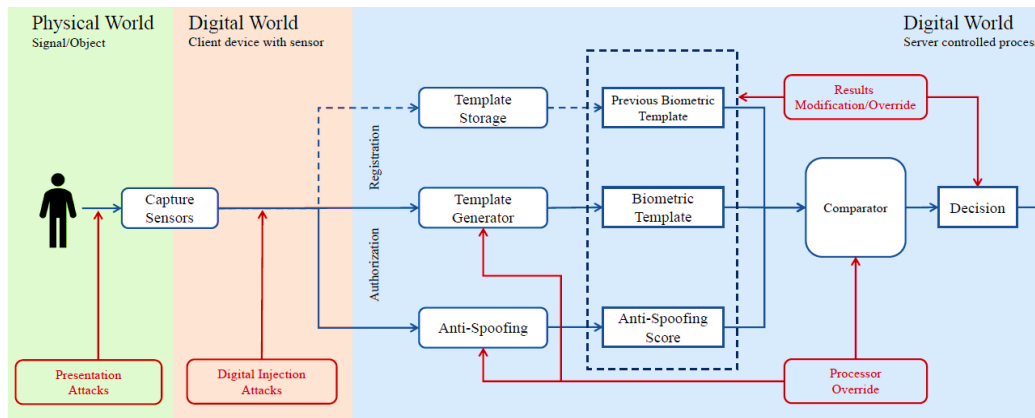


Figure 2.2: Possible attacks on biometric systems (best viewed in color). Image adapted from [30].

2.3 Deep Generative Models

Artificial Intelligence Generated Content (AIGC) has drawn significant attention in both academic and industrial domains in recent years, particularly with the notable advancements in deepfake technology within the generative domain. Generative models possess an extraordinary ability to create highly realistic images, videos, and other forms of visual media, continuously pushing the boundaries of synthetic content creation and redefining what is achievable in this field.

Unlike discriminative models such as Convolutional Neural Networks, logistic regression, or Support Vector Machines (SVMs), which focus on classifying data by learning to identify decision boundaries between different classes, generative models are designed to learn the underlying distribution of the data itself. By modeling this distribution, generative models can create new data points that closely resemble the original dataset, effectively synthesizing realistic samples that capture the same patterns and characteristics as the training data. Among the various approaches to generative modeling, three major techniques stand out: Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Denoising Diffusion Models.

Generative Adversarial Networks introduced by Goodfellow et al. in 2014, which have rapidly become one of the most influential and widely adopted generative models in computer vision [31]. GANs operate based on a dual-network architecture that pits two neural networks against each other in a zero-sum game. The generator network's goal is to produce synthetic images that are indistinguishable from real ones, while the discriminator network attempts to correctly differentiate between real and fake images.

This adversarial training process drives both networks to improve continuously. As the generator learns to create more realistic images, the discriminator becomes more adept at identifying subtle imperfections. This iterative refinement enables GANs to generate highly detailed and realistic images, making them the go-to choice for tasks such as image synthesis, style transfer, and super-resolution.

Variational Autoencoders (VAEs), introduced by Kingma and Welling in 2013, offer a different approach to generative modeling. VAEs are designed to learn a probabilistic latent space representation of the data, which can then be used to generate new samples [32]. The architecture of a VAE consists of two primary components: an encoder and a decoder. The encoder maps input data, such as images, into a lower-dimensional latent space, while the decoder reconstructs the data from this latent representation.

A key feature of VAEs is their ability to generate new, coherent data by sampling from the latent space. Unlike GANs, which focus on adversarial

training, VAEs optimize a variational lower bound, balancing the trade-off between reconstruction accuracy and the smoothness of the latent space. This makes VAEs particularly well-suited for tasks where interpretability and smooth transitions in the latent space are important, such as in anomaly detection, data compression, and generation of variations on a theme.

While VAEs may not always produce images as sharp and realistic as those generated by GANs, they excel in providing a continuous and interpretable latent space. This makes them ideal for applications that require a clear understanding of the underlying data distribution and the ability to generate diverse outputs.

Denoising Diffusion Models represent a newer and increasingly popular class of generative models that have shown exceptional promise in recent years [33]. Unlike GANs and VAEs, diffusion models are based on a stochastic process that involves gradually adding noise to data and then learning to reverse this process to recover the original data. This iterative denoising process allows the model to capture complex, high-dimensional data distributions effectively.

In practice, diffusion models start with a simple, known distribution—such as gaussian noise—and progressively refine it into a realistic image or data point through a series of denoising steps. This approach differs from the direct adversarial training of GANs or the latent space optimization of VAEs. The gradual refinement process allows diffusion models to produce images with high fidelity, capturing fine details and textures that are often challenging for other models.

Recent advancements [9], [34] have demonstrated the impressive capabilities of diffusion models in generating not only images but also audio, videos, and 3D objects that closely resemble real-world data. One significant benefit of diffusion models is their ability to handle missing data. Because they model the entire data generation process as a sequence of small, reversible steps, they can easily be adapted to tasks like inpainting or data completion, where parts of the data are missing or corrupted.

2.4 Face Manipulation in the Era of Deepfakes

Deepfake technology, has demonstrated extraordinary capabilities in producing highly realistic and convincing facial media content, transitioning from traditional graphics-based methods to sophisticated deep learning approaches by initially employing advanced techniques such as variational autoencoders [35], generative adversarial networks [8] and diffusion models [33]. Figure 2.3 illustrates the range of deepfake generation methods; each can be understood

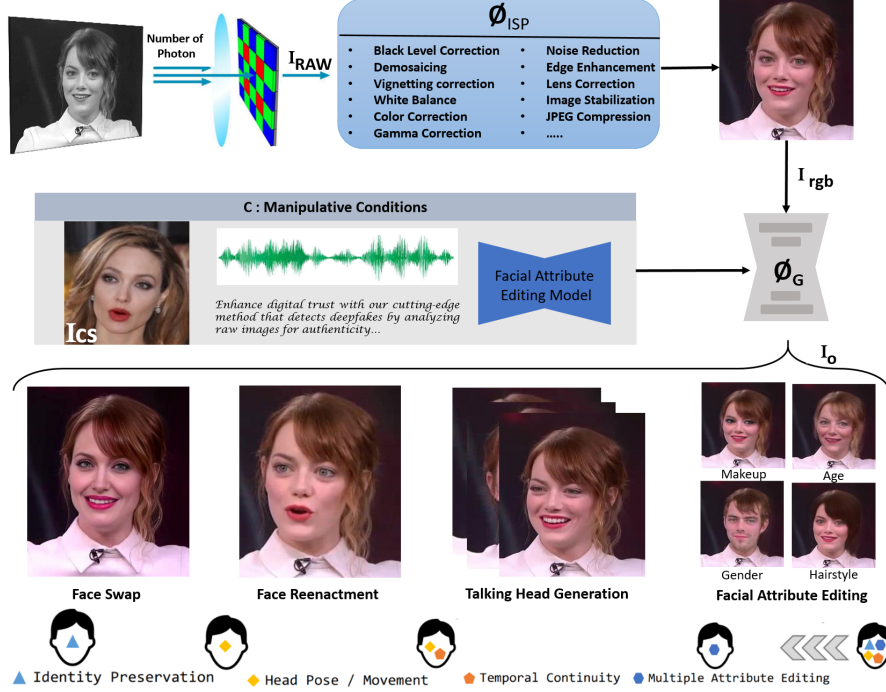


Figure 2.3: Various deepfake generation methods ϕ_G manipulate the input image I_{rgb} based on conditions like audio, video frames, or text. Typically, detection models process ISP-transformed images. The ISP process starts with light focused on the CFA sensor, producing raw pixel values I_{RAW} , which undergo stages ϕ_{ISP} such as white balance and noise removal to yield the final RGB output.

as a controlled content creation problem, where an RGB image I_{rgb} is manipulated based on specific conditional information $C = \{\text{Image, Audio, Text, ...}\}$. The generation process can be mathematically formulated as follows:

$$I_o = \phi_G(I_{rgb}, C), \quad (2.1)$$

where ϕ_G represents the specific generative network, $I_o = \{I_{rgb}^0, I_{rgb}^1, \dots, I_{rgb}^{N-1}\}$ denotes the sequence of generated contents and N is the total number of frames. This technology is broadly categorized into four main research areas:

- **Face Swapping:** Swaps the identity information of a face I_{rgb} with that of a source face $C = I_{cs}$, preserving ID-irrelevant attributes such as skin color and expressions.
- **Face-Reenactment:** Alters the facial movements of an image I_{rgb} without changing its identity or other attributes, influenced by external factors such as driving image or video.

- **Talking Face Generation:** Generates a video sequence where the character in I_{rgb} engages in conversation driven by external modalities such as text, audio, or video. This involves synchronizing lip movements and facial expressions to match the conversational content.
- **Facial Attribute Editing:** Modifies specific semantic attributes of the face I_{rgb} , such as age, gender, or expressions, based on individual preferences.

Among the various deepfake techniques, face reenactment and face swapping have gained significant popularity. These methods pose a particularly high risk to authentication services, making them a focal point of this thesis. Accordingly, our work places greater emphasis on analyzing and addressing these two techniques.

2.4.1 Overview of Face-Swapping Methods

Face-swapping algorithms can be broadly categorized into traditional graphics-based methods and more recent approaches leveraging GANs and diffusion models.

Traditional face-swapping methods primarily relied on region-based feature matching or the construction of a 3D prior model for facial parameterization [36]. The region-based approach focuses on identifying and aligning specific facial features, such as the eyes, nose, and mouth, within Regions Of Interest (ROIs) in both source and target images. Once aligned, techniques such as boundary blending and lighting adjustments are applied to create a seamless transition between the swapped regions. For example, Sunkavalli et al. [37] introduced lighting adjustments to enhance the realism of blended regions, while Bitouk et al. [38] developed an automated face replacement system that relied on a comprehensive face database to find suitable substitutes with matching poses and lighting conditions. On the other hand, the 3D prior model approach leveraged 3D Morphable Models (3DMM) to provide more robust and dynamic face-swapping capabilities. By constructing a 3D facial parameter model using a database of facial images, this method enabled the matching of the source image's parameters to the model and allowed for modifications that generated realistic face swaps. This approach outperformed in scenarios requiring adaptations to pose and lighting variations, as demonstrated by works such as those by Blanz [39] and Dale et al. [40]. While more adaptable, these methods were computationally intensive and less effective in handling extreme occlusions or lighting conditions.

With advancements in CNN models, the generation of face-swapping videos saw significant improvements [41], [42]. The advent of GAN-based methods

marked a major breakthrough in face-swapping technology. Early GAN approaches enhanced traditional algorithms by improving alignment, adapting to head pose variations, and addressing lighting inconsistencies between source and target images [43], [44]. However, these methods often required training for each identity, limiting their ability to generalize across diverse identities. To address this, researchers combined GANs with variational autoencoders (VAEs), significantly enhancing model generalizability [45], [46]. More recent GAN-based advancements have further refined face-swapping by incorporating techniques such as facial masking artifacts [47] and methods to decouple identity and attribute information [48], [49].

More recently, diffusion-based models have emerged as powerful tools for creating highly realistic face-swapping deepfakes. Zhao et al. [50] redefined the face-swapping process as a conditional inpainting task, reconstructing altered or missing facial regions based on predefined conditions. Similarly, Liu et al. [51] proposed a multi-modal face generation framework that integrates balanced identity and expression encoders within a conditional diffusion model. This approach achieves a harmonious balance between identity replacement and attribute preservation, producing exceptionally realistic outputs.

2.4.2 Overview of Face-reenactment Methods

Traditionally, facial reenactment or animation was achieved by fitting a 3D Morphable Model (3DMM) and modifying its estimated parameters. Early methods utilizing 3DMM incorporated high-detail features in animated frames either through detailed 3D scans or by learning 3DMM parameters directly from RGB images. However, these approaches involved additional steps to accurately transfer finer details, which made the process more complex and resource-intensive [52].

With advancements in deep learning, particularly in GANs, new methods emerged for generating facial animations. The literature is rich with both supervised and self-/unsupervised approaches. Supervised approaches aim to control facial animations by modeling factors of variation—such as lighting, pose, and expression—by conditioning the generated images on known ground-truth information, such as head pose, expressions, or landmarks [53], [54]. However, these methods require annotated datasets with detailed pose or expression information, which can be expensive to obtain or rely on subjective judgment (e.g., labeling expressions).

To address these challenges, self-supervised and unsupervised approaches have been developed to automatically learn factors of variation, such as optical flow and pose, without relying on labeled data. These methods often maximize mutual information or train networks to predict video frames [55].

A notable example is X2Face [56], which utilizes a two-stage training process. In the first stage, self-supervised learning is employed. Given multiple frames from a video, one frame is designated as the source frame, while the remaining frames serve as driving frames to control the source. During this stage, the model learns a bilinear sampler that encodes the pixel-wise flow $(\delta x, \delta y)$ and maps pixels from the source frame to the driving frame. In the second stage, a convolutional neural network pre-trained for face identification is incorporated to impose additional constraints based on the identity of the faces in the source and driving frames. This fine-tuning step helps the model better preserve the identity of the source face while adapting to the pose and expression of the driving frames.

Furthermore, Soumya et al. [57] proposed an interpretable and controllable face reenactment network, ICface. Similar to X2Face, their model employs a two-stage training process based on self-supervised learning. In the first stage, facial attributes such as emotion and pose are extracted from the driving image in terms of Action Units (AUs). In the second stage, these extracted attributes are transferred to the frontalized source image using a conditional generative model, specifically CycleGAN. This approach enables the synthesis of a source image with the pose and expression of the driving image while maintaining the identity of the source.

Siarohin et al. [58] proposed a self-supervised approach where sparse keypoints are extracted in an unsupervised manner. These keypoints, combined with local affine transformations, are used to generate dense motion vectors and an occlusion map. The source image, along with the dense motion vectors and occlusion map, is then utilized to render the target image, enabling realistic animation of the source image based on the motion of the driving image.

To enhance emotion transfer, micro-expression accuracy, and background quality, Hong et al. [59] introduced the Depth-Aware Generative Adversarial Network (DaGAN). This approach leverages self-supervised learning to recover dense 3D facial geometry through pixel-wise face depth maps, eliminating the need for expensive 3D annotations. DaGAN integrates depth information into the generation process using two innovative mechanisms. First, it combines geometric features from depth maps with appearance features from RGB images to predict more accurate facial keypoints, effectively capturing critical head movements. Second, it employs a 3D-aware cross-modal attention mechanism to seamlessly fuse depth and RGB information, allowing the model to capture subtle expression-related micro-movements and generate finer facial details.

Keypoint information plays a vital role in transferring motion from driving videos to still images. However, such methods often struggle when there are

significant appearance variations between the source image and the driving video. To overcome this challenge, Wang et al. [60] introduced the Latent Image Animator (LIA), a self-supervised autoencoder that transfers motion from driving videos to source images by leveraging linear transformations within the latent space.

LIA achieves animation by linearly navigating the source latent code along a learned trajectory to reach the target latent code, which encodes the high-level transformations needed for animating the source image. To enhance this process, the authors propose a Linear Motion Decomposition (LMD) approach, which represents the latent path as a linear combination of learned motion directions and their respective magnitudes. Importantly, these directions are constrained to form an orthogonal basis, where each vector corresponds to a fundamental visual transformation. By encapsulating the entire motion space within this learned basis, LIA eliminates the reliance on explicit structural representations, thereby streamlining the animation process.

2.5 Evaluation Techniques for Face Manipulation Methods

In this section, we review algorithms and metrics for assessing the quality of deepfake-generated frames. Sections 2.5.1 and 2.5.2 outline different evaluation approaches used in current methods.

2.5.1 Evaluation Techniques for Face-Swapping Methods

Identity preservation is considered a key metric in the evaluation of face-swapping methods, ensuring that the identity of the source subject is maintained in the generated face. This is typically achieved by pretrained face recognition models, which compute similarity using metrics such as cosine similarity or euclidean distance. In addition to identity preservation, expression and pose errors are critical for assessing the quality of face-swapping methods. These metrics measure how accurately the generated face replicates the target subject’s facial expression and head pose. Expression and pose accuracy are commonly evaluated using pose estimators [61] or 3D facial models [62], which extract expression and pose vectors.

Natsume et al. [45] utilized OpenFace [63] to evaluate identity preservation by calculating the squared euclidean distance between feature vectors of the

input and face-swapped images. Nirkin et al. [64] employed dlib [65] to extract identity embeddings and compare the face-swapping result of each frame to its nearest neighbor in pose from the source subject’s face views, ensuring identity consistency across frames. Pose error was computed as the euclidean distance between the Euler angles of the generated and target images, while expression error was measured as the pixel-wise euclidean distance between corresponding 2D landmarks.

Visual quality is another important aspect of face-swapping evaluation. The Structural Similarity Index (SSIM) has been used in several studies, including [64], to assess the visual similarity between images. However, SSIM evaluations can be limited by misalignment between corresponding pixels in the compared images. The Fréchet Inception Distance (FID) [66] is another widely used metric to assess the overall quality of generated images. FID measures the similarity between the distributions of real and generated images in feature space, with lower FID scores indicating higher visual fidelity and closer resemblance to real data. The aforementioned metrics are discussed in greater detail in Subsection 4.4.

2.5.2 Evaluation Techniques for Face Manipulation Methods

Evaluation techniques for face-reenactment can be classified into three categories: self-reenactment evaluation, cross-reenactment evaluation, and subjective test evaluation. The self-reenactment evaluation protocol, as depicted in Figure 2.4a, involves selecting a single frame from a video as the source image and using the remaining frames from the same video sequence to animate it. Since the source and driver identities originate from the same video sequence, the driver frames serve as a reliable ground-truth reference for comparing the generated images. This ensures a consistent and controlled evaluation of the reenactment process.

To assess the quality of the generated images in self-reenactment studies, image quality metrics such as SSIM and Peak Signal-to-Noise Ratio (PSNR) [67] are commonly employed [56], [58], [60], [68], [69]. These metrics rely on ground-truth data and provide objective measures of image similarity and fidelity. Additionally, the self-reenactment technique enables the measurement of facial keypoint error such as Average Keypoint Distance (AKD) and Missing Keypoint Rate (MKR) which offers further insights into the accuracy of the reenactment process [60], [68]. To quantitatively evaluate the quality of generated frames, Siarohin et al. [68] utilizes self-reenactment to measure the L1 error, AKD, and Average Euclidean Distance (AED) between the

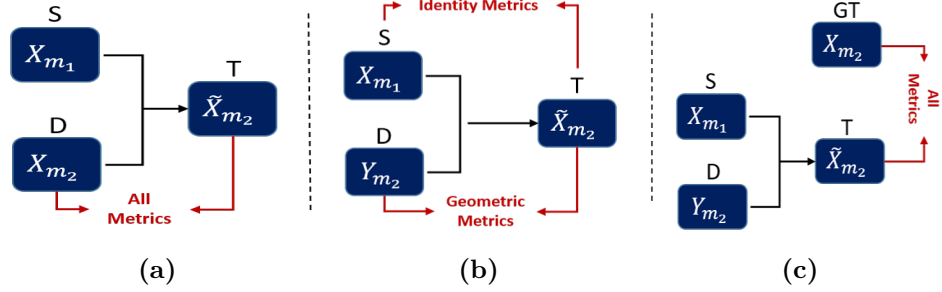


Figure 2.4: Face-reenactment evaluation protocols: self-reenactment (a), cross-reenactment (b), and our proposed evaluation protocol (c). In this illustration, X and Y represent identities, while S, D, T, and GT correspond to source, driving, target, and ground-truth, respectively. Additionally, m_1 and m_2 indicate the movement of the source and driving, respectively

generated frames and the ground-truth frames. Similarly, Gao et al. [70] reports the L1, SSIM, PSNR, FID and AKD error between the generated frames and the corresponding ground-truth frames for the self-reenactment scenario. Wang et al. [60] and Yang et al. [71] utilized the LPIPS to compute the similarity score between generated and ground-truth frames.

To quantitatively evaluate the generated frames in cross-reenactment scenarios and address the absence of ground-truth data, researchers employ a set of metrics that do not rely on explicit ground-truth comparisons. For the evaluation, researchers commonly utilize a cross-reenactment protocol, as illustrated in Figure 2.4b. In the existing cross-reenactment protocol, a prevalent method involves utilizing a pretrained network to extract identity features from the source and reenacted images [60], [68]. Alternatively, geometric features can be extracted from the driving and reenacted images [72], [73]. These extracted embeddings capture essential characteristics of the face, such as appearance and face pose. The quality of the generated frames can be assessed by computing the distance or dissimilarity between these embeddings. For instance recent face-reenactment methods [70], [74]–[76] evaluate the identity preservation by computing Cosine Similarity (CSIM) of embedding vectors between the generated frame and the source face [5]. Furthermore, Ha et al. [74] leverage pretrained networks to estimate the head pose angles and Facial Action Units (FAU) of generated image and compare these estimates with the corresponding driver’s head pose and action units, providing insights into the accuracy of the reenactment process.

Subjective test form the third category of evaluation techniques for cross-reenactment. In these evaluations, human observers play a crucial role by

providing judgments on various aspects such as the visual quality, realism, and coherence of the generated cross-reenactment frames. For instance, Siarohin et al. [68] and Wang et al. [60] conducted a user study in which participants were presented with a source image, a driving video, and the corresponding results of their method and a competitive method. Participants were asked to select the most realistic image animation. Despite the significant advancements in cross-reenactment evaluation, there is still a need for an automated protocol that can compute errors for metrics relying on explicit ground-truth data. The establishment of such a protocol would contribute to a comprehensive and robust evaluation of cross-reenactment methods, enabling a deeper understanding of their performance and fostering further advancements in the field.

In **Chapter 4**, we present the evaluation technique illustrated in Figure 2.4c, designed to assess image quality for face-reenactment methods. This chapter outlines the proposed evaluation protocol and discusses the results from the three author’s publications [P2]–[P4].

2.6 Deepfake Attacks and Countermeasures

Deepfake attacks can occur in various forms, with one of the most common being the media content attack. In this type of attack, the attacker obtains an image of the victim, often sourced from social media platforms, and uses it to create a deepfake video. These attacks typically happen offline and pose significant risks to the victim. For instance, the victim may be depicted in a compromising situation or made to say things he/she have never actually said. Such attacks can result in severe consequences, including damage to reputation and personal harm.

Another form of deepfake attack is the real-time deepfake attack, where the attacker targets an authentication system in real time. This can take the form of a presentation attack or an injection attack.

In a presentation attack scenario, the deepfake video is usually generated on a separate device, such as a laptop, and then presented to the authentication sensor. This allows the attacker to deceive the system by showcasing artificial biometric traits, effectively simulating the victim’s appearance. Injection attacks take this a step further by streaming the manipulated videos directly to authentication systems using virtual camera software (e.g., OBS), bypassing the need for physical camera sensors entirely.

2.6.1 Deepfake Detection

To address the challenges posed by deepfakes, several deepfake databases have been developed [77]–[79], offering extensive datasets that feature a diverse range of individuals with varying facial attributes and expressions.

Researchers have approached deepfake detection from multiple perspectives to enhance detection methods. Some techniques focus on the image level, aiming to identify fake images by recognizing spatial artifacts within individual frames [13], [14], while others focus on the video level, leveraging temporal information by analyzing multiple frames to detect deepfake videos [80], [81]. Furthermore, certain detection methods utilize frequency information, which proves particularly effective on highly compressed videos. The LRL [82] and FRDM [15] combine representations from both RGB and frequency domains to learn inconsistencies in the video frames.

Another direction in deepfake detection involves the use of training data synthesis, generating synthetic data that includes common deepfake artifacts. These techniques do not rely on existing fake data but generate their own. For instance, DSPFWA [83] focuses on identifying artifacts that arise during face warping, where a source face undergoes transformations such as scaling, rotation, and shearing to match the pose of target face it aims to replace. These transformations create artifacts and resolution inconsistencies between the warped face area and the surrounding context. During training, the algorithm generates synthetic data containing these affine face warping artifacts to improve detection accuracy. Similarly, Face X-ray [84] generates synthetic fake samples, called BI, by blending two images from different videos and attempts to detect deepfakes by segmenting the blending boundaries between the source and target images. SBI [85] follows a comparable approach but blends faces from the same frame to produce higher-quality images, making them more challenging for deepfake detectors to identify.

Several deepfake detectors focus on image patches rather than the entire image. PCL [16] detects deepfakes by assessing consistencies between patches of input images. The model is trained with an Inconsistency Image Generator (I2G), akin to BI [84]. CADDM [14] identifies that the stumbling block of deepfake detectors' generalization ability lies in the mistakenly learned identity representation in images. Therefore, they propose a model containing an anchor-based detector which detects deepfake artifacts in local areas. Other notable directions for deepfake detection involves focusing on specific representations such as head pose [86], eye blinking [87], mouth movements [88], and heart rate estimation [89]. While many current methods perform well in detecting known manipulations, certain studies [14], [85] have identified limitations in their ability to generalize to fake faces forged by unknown

manipulations.

To assess the good generalization of the deepfake detectors to different image processing algorithms, studies have been conducted in real-case scenarios, i.e. detecting deepfake uploaded online with video compression [42], [90], [91]. The higher the compression rate, the lower they can correctly classify. Indeed, the effect of compression can be seen in the classification Area Under the Curve (AUC) of the deepfake detectors with the low-quality videos of FaceForensics++ [92]–[94].

2.6.2 The Image Signal Processing (ISP) Pipeline and Its Impact on Deepfake Detection

Current detection models often struggle to differentiate between real and fake images because the existing definition of real images includes both raw content captured by camera sensors and content processed through various stages of image and video enhancement, including both linear and nonlinear adjustments. As a result, images are considered real even after undergoing multiple processing operations such as denoising, compression, deblurring, and white balance adjustments. These ISP processing steps are designed to produce aesthetically pleasing images for human viewers, but they pose significant challenges for deepfake detection. The issue arises from the fact that each device employs a unique ISP pipeline with distinct enhancement blocks, which obscure subtle cues that are crucial for detecting deepfakes. This variability forces detection models to adapt to unseen ISP configurations, making it more difficult to accurately identify real images.

Modern digital cameras aim to capture and render images that are both visually pleasing and accurate to human perception. However, the raw sensor data initially captured by these cameras does not resemble a finished photograph. To convert this noisy, linear intensity data into a polished image, an in-camera image ISP pipeline is employed. This pipeline transforms the sensor’s raw image into the standard RGB (sRGB) format, producing perceptually pleasant RGB images suitable for the human visual system [95]. The ISP methods can generally be categorized into two main types: model-based and learning-based approaches. Model-based methods rely on conventional blocks and learning-based methods directly acquire raw-to-RGB conversion from paired RGB and raw images via end-to-end training [96]–[98].

The common blocks of a model-based ISP pipeline include demosaicing, denoising, White Balance(WB), and Lens Shading Correction (LSC). The demosaicing is used to convert a single-channel raw image into a full-color RGB image by interpolating the raw Color Filter Array (CFA) patterned image [99].

Denoising aims to eliminate acquisition noise from images while maintaining their details to enhance quality. This is achieved through techniques such as spatial filtering, frequency domain filtering, and wavelet-based methods [100], [101]. Furthermore, white balance removes undesirable color casts caused by environmental lighting, allowing objects to be perceived as the same color regardless of the scene’s lighting. WB involves estimating the scene’s illuminant color using an algorithm and then correcting the captured image based on this estimation [102], [103]. In addition to accurate color reconstruction, ensuring uniform light distribution is essential for high-quality image processing. Lens shading correction addresses the radial decrease in light intensity towards the edges caused by sensor optics, resulting in a vignetting effect. LSC adjusts for this uneven light distribution using a pre-calibrated mask to ensure uniform light response across the sensor [104], [105].

While model-based ISP methods generally provide superior interpretability and control, they often require manual parameter adjustments and depend on camera metadata, such as color correction matrices. On the other hand, learning-based methods eliminate the necessity for such metadata but require significant amounts of data. Despite this requirement, they generally achieve superior reconstruction accuracy compared to model-based approaches [106].

Inverse ISP

A sensor raw image captured by a digital camera retains all the information captured by the sensor. The relationship between ambient light, pixel intensity, and noise distribution in the raw domain is typically much simpler compared to that in the RGB domain [107]. Hence, leveraging raw images directly for subsequent tasks can potentially yield superior performance compared to methods based on RGB images, across both low-level and high-level computer vision tasks. Recent research indicates that raw image-based approaches for tasks such as image recognition [107], [108] and denoising [109] have demonstrated higher performance levels than their RGB image-based counterparts. The use of raw images is expected to improve performance especially in difficult scenes such as extremely dark or blurry scenes that should be covered in practical application. However, the scarcity of annotated raw data has been a barrier to machine learning-based approaches. Consequently, several reversed ISP methods that convert existing large-scale RGB datasets into pseudo raw datasets have been studied [104], [106], [109], [110].

Inverse ISP methods based on deep learning can be categorized into end-to-end methods and hybrid methods. Hybrid methods offer greater interpretability for intermediate representations and enable better control over the ISP layout. MBISPLD [104] employs a hybrid approach that combines

classical reversible ISP blocks with shallow CNNs to transform RGB images into raw images. In this framework, each ISP block possesses learnable parameters that are optimized using RGB and raw image pairs. On the contrary, CycleISP [109] and InvISP [98] perform RGB-to-raw (forward) and raw-to-RGB (reverse) mapping in an end-to-end fashion. CycleISP adopts cycle consistency to learn both the forward reverse directions of the ISP using two different networks. These networks are jointly fine-tuned to ensure cycle consistency. On the other hand, InvISP learns a reversible ISP using normalizing flow techniques, such as those introduced in Glow [111], to generate an invertible RGB image from the original raw image. Further more, MiAlgo [110] introduces an end-to-end encoder-decoder network. During training, it gets a full-resolution RGB image as input and leverages its corresponding paired raw image to optimize the network. The network adopts a UNet-like structure, comprising multiple sampling blocks and residual groups, to extract deep features and reconstruct the raw image data.

In **Chapters 5 and 6**, we present the results from two of the author’s publications focused on deepfake detection. The first publication, [P5], investigates the impact of image processing techniques, particularly beautification filters, on the performance of deepfake detection. The second publication, [P6], introduces a novel pipeline that utilizes raw domain data as input to enhance deepfake detection. In light of the scarcity of large-scale datasets for training on raw images, the MiAlgo framework is employed to preprocess the input images by converting them into a raw image format, a crucial step prior to performing deepfake detection. By focusing on raw data, this approach seeks to constrain the distribution of real images, making it easier for the model to learn distinctive features and generalize effectively to authentic images.

2.7 Digital Replay Attack Detection

Digital replay attacks involve injecting a genuine video stream of the victim into a facial recognition system, often using webcam simulation tools like OBS Studio on computers. On mobile devices, more sophisticated software solutions are required to achieve similar results [7], [112]. Detecting these attacks is challenging, as the biometric data used is authentic and lacks detectable anomalies. As a result, defense strategies against digital replay attacks remain underexplored, with limited research addressing this issue within the biometric community.

Some studies have proposed methods to counter digital replay attacks by leveraging external signals to verify the presence of a live user in front

of the camera. The authors of [113] and [114] suggest using a smartphone screen to emit randomly flashing colors onto the user's face. These flashes serve as a dynamic watermark within the video sequence, as the light is reflected off the user's face and captured by the camera. By analyzing the reflected light colors, these methods aim to distinguish between a live face and a replayed video or image. Similarly, [115] and [116] propose utilizing specific light patterns as an authentication mechanism for the captured content. However, the effectiveness of these approaches is limited due to their reliance on external signals, which are often too weak to detect, especially under strong ambient lighting conditions or on individuals with low skin reflectance. These constraints highlight the need for more robust and practical solutions to combat digital replay attacks.

2.8 Compression Detection for Video Forensics: Effectiveness in Replay Attacks and Deepfake Detection

A significant area of research in digital video forensics involves the detection and analysis of compression artifacts, which provide valuable insights into the editing history of videos. These artifacts are typically examined through spatial statistics within individual frames and temporal statistics embedded in the Group of Pictures (GOP) structure. The GOP defines the types and sequence of frames in a video, establishing the foundation for compression analysis.

Video manipulation often entails decompression, editing, and recompression, making the detection of double compression artifacts particularly important. These artifacts serve as crucial evidence for identifying the sequence of edits and determining whether a video has been tampered with. Techniques such as the analysis of quantization artifacts or blockiness patterns have been developed to detect traces of recompression in both images and videos. For example, the authors of [117] propose a SVM-based classifier to determine the number of compression steps applied to a video sequence. Their method relies on Benford's law, analyzing the statistics of the most significant digit in quantized transform coefficients. Similarly, Jiang et al.[118] apply Markov statistics to identify double quantization artifacts in MPEG-4 videos. Other studies, such as [119] and [120], focus on periodicity analysis and the GOP structure to detect double compression in videos.

Despite the significant body of research on detecting double or multiple compressions, most existing methods are limited to identifying double-

compressed videos for specific codecs or compression parameters. Both replay and deepfake injection attacks remain unresolved challenges. The ideal solution to this challenge is to cryptographically sign biometric data at the hardware level, enabling hardware manufacturers to verify the authenticity of the captured content. However, implementing this approach requires seamless collaboration among hardware manufacturers, operating system developers, software providers, and face anti-spoofing service providers, a level of coordination that has not yet been achieved.

In **Chapter 7**, we present the results from the author’s publication focused on digital replay attack detection. The publication, [P7], investigates whether providing uncompressed video access to face anti-spoofing service providers can enhance the detection of injected versus authentic video streams. Building upon this, we propose bypassing the compression step and directly capturing uncompressed image data from the user’s device during authentication.

Chapter 3

Enhancing Face Verification Algorithm

This chapter presents the challenges a face verification system may encounter during authentication and introduces a novel face normalization algorithm designed for preprocessing in face verification. The proposed method adaptively aligns head pose, expression, and illumination, resulting in significant performance improvements.

Section 3.1 provides an overview of the current state-of-the-art face verification methods and outlines our motivation. Section 3.2 details the proposed approach, while Section 3.3 presents the experimental evaluation and results. Finally, Section 3.4 summarizes the key contributions of this chapter.

3.1 Introduction

Face Verification (FV) has gained significant attention due to its great potential value in practical applications such as access control and video surveillance. Recent progress in face verification heavily depends on the utilization of deep convolutional neural networks, consistently showcasing notable accuracy that frequently exceeds human-level performance. In face verification models, where two images are used as input, effectiveness is indeed influenced by several factors, including scene illumination during image capture, camera parameters, image quality, alterations in facial expressions, and changes in the head pose of the subjects. Hence, it is crucial to direct the model's focus exclusively towards distinctive features crucial for individual recognition while neglecting extraneous elements. For this purpose, diverse strategies have been investigated, falling into two primary categories: incorporating image quality-related factors, such as head pose and illumination, into the loss function

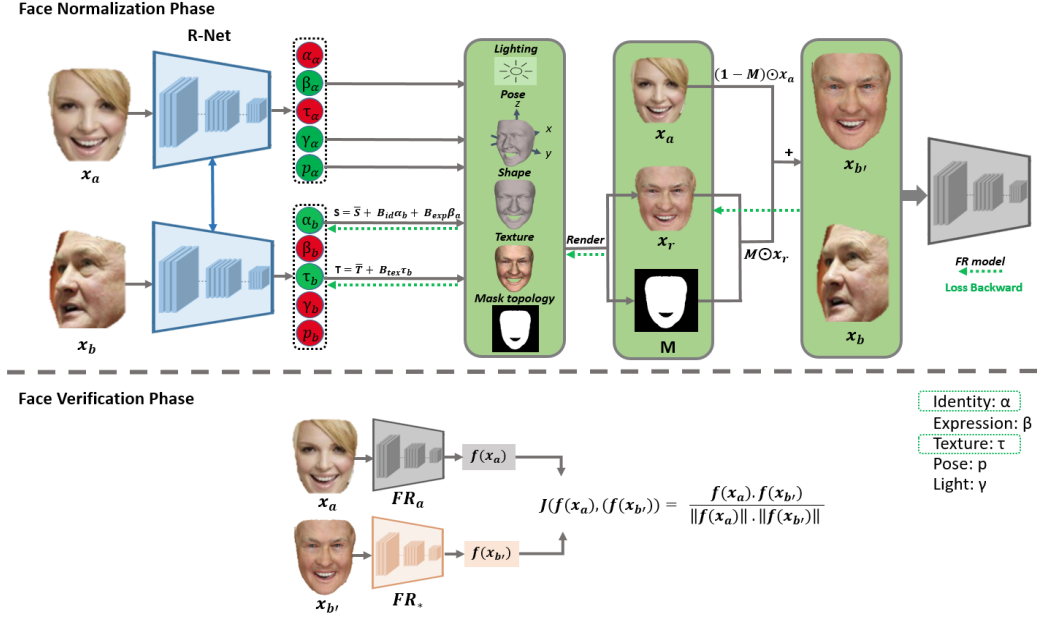


Figure 3.1: Overview of Face Normalization by AlignFace: For each image, 3DMM coefficients, including identity (α), expression (β), texture (τ), illumination (γ), and head pose (p) are extracted using the R-Net model. The image x_b is then normalized to produce $x_{b'}$, aligning its expression, head pose, and lighting conditions with those of x_a . During normalization process $x_{b'}$'s identity and texture coefficients are iteratively updated (n iterations) while keeping the parameters of the FR and R-Net models frozen. Although images generated as $x_{b'}$ closely follow the distribution of real images, discrepancies might exist between the distributions of generated $x_{b'}$ and real faces x_a . To ensure accuracy at the face verification phase, the FR model used for extracting face embeddings for $x_{b'}$ is fine-tuned, denoted as FR_* . x_a and $x_{b'}$ represent different identities.

[28]. The other approach focuses on the implementation of preprocessing techniques to normalize elements such as head pose and expression [20], and to address variations in illumination [1].

Recent advancements, focusing on enhancing performance through improved loss functions, often involve the incorporation of margin-based loss functions [121], with the primary objective of minimizing intra-class variation and maximizing inter-class distinction. A widely adopted margin-based loss function is ArcFace [5], which introduces an angular margin term into the standard softmax classification loss, significantly enhancing class separability. Nevertheless, recent investigations have pointed out that ArcFace exhibits a degree of quality-agnostic behavior, leading to instability in within-class distributions [28]. To address these challenges and improve performance,

AdaFace [121] integrates image quality information into the loss function.

In parallel, another crucial technique for improving face verification is face normalization. This involves synthesizing and transforming a face with arbitrary pose, illumination, and expression into a desired pose, balanced illumination, and neutral expression to enhance recognition. Through the normalization of images to a shared representation, the model is enabled to concentrate its discriminative capacity on the intrinsic characteristics of individuals, thereby fostering more reliable and accurate face verification outcomes.

Normalization of face pose is widely adopted in the field, typically with the desired pose specified as frontal [122]. In [21], a combination of a 3D morphable model and a generative adversarial network is employed to generate frontal face images from input profile images. Likewise, in [20], face frontalization is accomplished entirely through a generative adversarial network. The DVN [123] utilizes two layers of dual-view generators to normalize a face in dual views - one in frontal view and the other in a yaw 45° side view. MVN [122] is designed to learn the transformation from an input set to seven output sets, encompassing seven face poses from 0° to 90° in yaw with a 15° interval, utilizing seven generators. However, transferring faces to specific head poses is not always advisable due to several reasons:

- **Training Data Distribution:** The majority of the training data may not be centered within the frontal pose range and could be distributed across various angles. As illustrated in the DVN [123] framework, the face encoder exhibits greater expertise with faces within a 45° range, reflecting the predominant distribution of training data in their database. Therefore, to ensure effective of face verification in diverse scenarios, it is essential not to exclusively rely on normalization at specific poses, as optimal results may vary.
- **Photo-Realism and Texture Loss:** Generated frontalized (or at any other specific degree) face images from GANs may lack photo-realism and exhibit artifacts and texture loss, especially in occluded regions. Counterfeiting features in synthetic generated images may degrade recognition performance. For example, if a particular facial feature, such as a birthmark or mole, is obscured in the original image and remains ungenerated by the GAN model during frontalization, while being visible in the second image, the face verification model may incorrectly categorize these two images as representing different identities.

The reasons mentioned above could also be applicable to the normalization of illumination, expression, and other extraneous elements.

In this thesis, we introduce an innovative normalization algorithm designed for preprocessing input images in the context of face verification. Diverging from conventional methods, our approach places a distinctive emphasis on achieving consistency in head pose, expression, and illumination conditions between two images, avoiding an exclusive focus on the normalization of extraneous elements at specific values. Specifically, our methodology involves estimating the head pose, expression, and illumination conditions in one image, followed by the reconstruction of the second image to align with the same head pose, expression, and illumination conditions while preserving its own unique identity features. This ensures the constancy of real features in one of the images, providing a more authentic representation of the facial distribution. By adopting this approach, our algorithm allows the verification process to concentrate solely on identity evaluation, unaffected by variations in non-essential extraneous and synthesized features. This refined focus contributes to a more accurate and reliable assessment of facial identity in face verification scenarios.

3.2 Methodology

In face verification, a pair of images $\{x_a, x_b\} \subset X$ is examined using a face recognition model denoted as $f(x) : X \rightarrow \mathbb{R}^d$. This model extracts feature embeddings from the faces in the images, placing them in the \mathbb{R}^d space. The similarity between a pair of images can be commonly calculated using the cosine similarity formula:

$$J(f(x_a), f(x_b)) = \frac{f(x_a) \cdot f(x_b)}{\|f(x_a)\| \cdot \|f(x_b)\|} \quad (3.1)$$

where $\langle \cdot, \cdot \rangle$ represents the inner product of the vectors. The function J denotes the cosine similarity between the feature embeddings of x_a and x_b , with values ranging from 0 to 1. The prediction for face verification is formulated as:

$$C(x_a, x_b) = \begin{cases} 1 & \text{if } J(f(x_a), f(x_b)) \geq \delta \\ 0 & \text{otherwise} \end{cases} \quad (3.2)$$

Here, δ represents the threshold. When $C(x_a, x_b)$ equals 1, the two images are considered to depict the same identity; otherwise, they represent different identities.

3.2.1 3D Face Model Reconstruction

Given a facial image, denoted as x , R-Net model [62] is employed to regress the 3D Morphable Model (3DMM) coefficients denoted as $\alpha \in \mathbb{R}^{80}$, $\beta \in \mathbb{R}^{64}$, and $\tau \in \mathbb{R}^{80}$ corresponding to the image x . Once these coefficients are obtained, the 3D face shape (S) and texture (T) can be represented by an affine model:

$$\begin{aligned} S &= S(\alpha, \beta) = \bar{S} + B_{id}\alpha + B_{exp}\beta \\ T &= T(\delta) = \bar{T} + B_{tex}\tau \end{aligned} \quad (3.3)$$

where \bar{S} and \bar{T} denote the averages of face shape and texture, while B_{id} , B_{tex} , and B_{exp} represent the Principal Component Analysis (PCA) bases of identity, texture, and expression, respectively. The values of \bar{S} , \bar{T} , B_{id} , and B_{tex} are derived from the well-established 2009 Basel Face Model [124] and the expression bases B_{exp} are sourced from [125], which constructed using data from Face-Warehouse [126]. Furthermore, the R-Net model regresses the illumination coefficients $\gamma \in \mathbb{R}^9$, and the head pose $p \in \mathbb{R}^6$.

With access to both the facial texture and shape, we are able to represent the complete 3D mesh model of the face as $M_{sh} = (S, T)$, where $S \in \mathbb{R}^{n \times 3}$ represents the XYZ coordinates of n vertices, and $T \in \mathbb{R}^{n \times 3}$ corresponds to the RGB values of these vertices [127].

3.2.2 Proposed Method

The face verification system operates on a pair of facial images as its input. In environments without constraints, these images may exhibit variations in head pose, facial expressions, and lighting conditions, thereby significantly impacting the system's performance. To effectively tackle this challenge, we propose a dedicated pipeline designed specifically for face verification, as illustrated in Figure 3.1. Our primary objective is to normalize one of the faces within the pair, ensuring alignment in terms of head pose, expression, and illumination. This normalization process optimizes the system's workload, enabling it to concentrate exclusively on identity verification. Specifically, given an image pair, our methodology entails selecting one image, denoted as x_a , which possesses a head pose closest to the frontal pose, to serve as the reference. Subsequently, the second image, x_b , undergoes normalization to become $x_{b'}$, aligning its expression, head pose, and lighting condition with those of the original image x_a . To achieve this, we follow these steps:

- Utilize the R-Net to extract 3DMM coefficients for both provided images. As face verification models exhibit more sensitivity to pose variations than to scene illumination and facial expression [21], we specifically focus

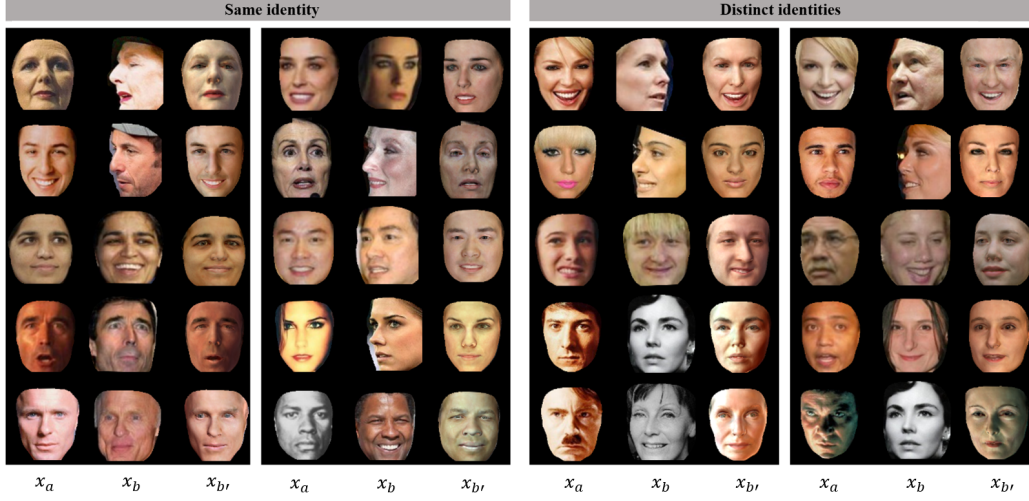


Figure 3.2: AlignFace Efficacy in Normalizing Pose, Expression, and Illumination. Displayed are the original image x_a , the comparative image x_b , and AlignFace’s reconstructed image $x_{b'}$. This demonstrates AlignFace’s ability to effectively transfer the extraneous conditions of x_a to x_b while preserving the unique identity features in x_b . Note that in the examples on the left, the identities are the same, whereas in the examples on the right, the identities are different.

on the head pose coefficient p . An image with the closest deviation from the frontal pose is denoted as x_a , while the second image x_b undergoes normalization. The coefficients for these image pairs are as follows:

$$x_a : \{\alpha_a, \beta_a, \tau_a, \gamma_a, p_a\}, x_b : \{\alpha_b, \beta_b, \tau_b, \gamma_b, p_b\}$$

- To reconstruct the 3D face model x_b with the same illumination, head pose, and expression as x_a , we initialize the 3D mesh model using the following coefficients:

$$\{\alpha_{b'}^{[0]}, \beta_{b'}^{[0]}, \tau_{b'}^{[0]}, \gamma_{b'}^{[0]}, p_{b'}^{[0]}\} \leftarrow \{\alpha_b, \beta_a, \tau_b, \gamma_a, p_a\}$$

- Since the regressed 3DMM coefficients are all differentiable, we employ $L_f = -J(f(x_{b'}), f(x_b))$ as loss function and update coefficients $\alpha_{b'}$, $\tau_{b'}$ for n iterations. The objective function can be expressed as:

$$\min_{\alpha_{b'}, \tau_{b'}} L_f(x_{b'}, x_b) \quad (3.4)$$

where,

$$x_{b'} = M \odot x_r + (1 - M) \odot x_a \quad (3.5)$$

and x_r , M are computed using the rendering function $R(\bar{S} + B_{id}\alpha_{b'} + B_{exp}\beta_a, \bar{T} + B_{tex}\tau_{b'}, \gamma_a, p_a)$. The symbol \odot denotes element-wise multiplication, and R represents the rendering function, which takes into consideration factors such as camera position and illumination. The variable M signifies the binary mask used in this process.

- After successfully reconstructing $x_{b'}$, the next step involves conducting face verification between the image pair x_a and $x_{b'}$ using a trained FR model as a feature extractor and computing the cosine distance between their feature vectors. Models such as ArcFace [5], MagFace [28], or AdaFace [121] can be employed for this purpose, considering that our algorithm serves as a face normalization tool.
- The images generated as $x_{b'}$ follow the distribution of real images; however, FR models are typically trained on real datasets, and a potential discrepancy may exist between the distributions of generated $x_{b'}$ and real faces x_a . To ensure result precision, we created a training dataset normalized by our proposed normalization tool. Subsequently, we fine-tuned the selected FR model with the generated $x_{b'}$ data, denoted as FR_* .
- After completing the fine-tuning process, the face recognition models FR_a and FR_* are employed to extract feature embeddings from x_a and $x_{b'}$, followed by computing the cosine similarity distance between these embeddings.

R-Net: In this thesis, we employ R-Net [62], a CNN-based model, to perform 3D face reconstruction from a single image. This model is trained using a hybrid-level loss function that seamlessly integrates both low-level and perception-level information, enhancing its reconstruction capabilities. The model’s strength lies in its robustness in handling challenges such as occlusion and extreme poses. It achieves this robustness by incorporating a skin color-based photometric error attention strategy, making it adaptable to scenarios with occlusions and other intricate appearance variations, such as beards and heavy makeup. The backbone of this model is the ResNet-50 network, which plays a crucial role in regressing the 3D Morphable Model (3DMM) coefficients required for accurate 3D reconstruction.

3.3 Experimental Evaluation

3.3.1 Datasets

In our experiment, the MS1M-V2 dataset [5], containing 5.8 million images and 85,000 identities, was utilized to fine-tune face recognition model. For the evaluation purposes, we selected four widely recognized unconstrained face verification benchmarks, namely Labeled Faces in the Wild (LFW) [128], Celebrities in Frontal-Profile (CFP) [129], AgeDB [130], and IARPA Janus Benchmark-B (IJB-B) [131] dataset. The LFW dataset comprises 13,233 facial images from 5,749 individuals, showcasing various poses, facial expressions, and lighting conditions. The CFP dataset, with 7,000 facial images, emphasizes extreme head poses, such as profiles, leading to significant occlusion. AgeDB, consisting of 16,516 images, focuses on age-related variations. The IJB-B dataset features 21.8K still images and 55K frames from 7,011 videos, representing 1,845 subjects with diverse qualities. All images are resized to 112×112 dimensions before the verification step.

3.3.2 Face Normalization

We employ a PyTorch implementation of R-Net [62] to acquire the 3DMM coefficients for image pairs. Within our pipeline, the FR encoder model is utilized to extract feature embeddings from both x_b and $x_{b'}$. It's important to note that the FR model is pretrained and fixed under the normalization framework. Since the entire pipeline, including the rendering procedure, is differentiable, $x_{b'}$ can be iteratively updated through backpropagation on the low-dimensional identity (α) and texture (β) coefficients of the 3DMM. We set the number of iterations to $N = 300$, the learning rate to $\alpha = 1.5$, and the decay factor to $\mu = 1$. This iterative process results in the reconstruction of x_b , aligning its expression, head pose, and lighting conditions with those of the image x_a while preserving its unique identity features.

3.3.3 Face Verification Models

In our experiments, we benchmark and utilize the encoders of two SoTA face recognition models: MagFace [28] and AdaFace [121], to serve as facial feature extractors. We employed the official implementations of MagFace and AdaFace, both utilizing ResNet100 backbones trained on the MS1M-V2 dataset. The encoder used for x_a feature extraction does not require fine-tuning. However, since a potential discrepancy may exist between the distributions of generated $x_{b'}$ and real faces, on which the original face

Table 3.1: Comparative analysis on benchmark datasets: Accuracy metrics for 1:1 verification are presented for LFW, CFP, and AgeDB datasets. For the IJB-B dataset, we report the TAR@FAR=0.01%. Red: best, blue: second-best.

Method	Dataset				
	LFW [128]	CFP [129]	AgeDB [130]	AVG	IJB-B [131]
LFW CosFace [4]	99.81	98.12	98.11	98.68	94.80
ArcFace [5]	99.83	98.27	98.28	98.79	94.25
MV-Softmax [132]	99.80	98.28	97.95	98.68	93.60
MagFace[28]	99.83	98.46	98.17	98.82	94.51
AdaFace [121]	99.82	98.49	98.05	98.79	95.67
R-Net α coefficient	92.76	84.65	87.25	86.22	87.13
R-Net α coefficient after normalization	97.46	95.32	94.11	95.63	93.46
AlignFace+MagFace	99.82	98.73	98.33	99.29	94.46
AlignFace+AdaFace	99.82	98.85	97.95	98.87	96.02

recognition models are trained, we ensure result precision by creating a training dataset. This dataset, normalized using our proposed normalization tool and derived from MS1M-V2, serves as the basis for fine-tuning the selected FR model. The fine-tuned model, denoted as FR_* in Figure 3.1, is trained with the generated x_b datatype. The fine-tuning process follow the same parameters and instructions specified in the official implementation.

3.3.4 Comparisons With State-of-The-Art Methods

To assess the efficacy of our proposed method, we conducted a comprehensive comparative analysis with SoTA methods. The results, encompassing 1:1 verification accuracy for LFW, CFP, and AgeDB, as well as TAR@FAR=0.01% for the IJB-B dataset, are showcased in Table 3.1. Notably, all models featured in this table were trained utilizing the MS1M-V2 dataset and the ResNet100 backbone. In our evaluation, we incorporated our novel normalization method as a preprocessing step for two specific models: AdaFace and MagFace. These modified models are referred to as "AlignFace+MagFace" and "AlignFace+AdaFace," respectively. The results presented in Table 3.1 for the LFW, CFP, and AgeDB datasets demonstrate that, although face verification performance is approaching saturation on these benchmarks, our proposed enhancements have yielded significant improvements. However, this increased accuracy results in higher processing times and resource consumption.

In particular, on the CFP benchmark, the incorporation of our normalization technique with MagFace (denoted as AlignFace+MagFace) led to an improvement in performance by 0.24% in accuracy compared to the previous best method. Additionally, when combined with AdaFace (AlignFace+AdaFace), there was a further increase of 0.36% in accuracy, thereby exceeding the capabilities of the previously established best-performing method. This im-

provement can be attributed to the distinct advantages of our normalization method in minimizing head pose differences between image pairs. This is particularly significant in the CFP dataset, which comprises images with both frontal and profile head poses.

The results from the IJB-B dataset indicate that the integration of AlignFace with AdaFace (AlignFace+AdaFace) yields a 0.36% increase in performance compared to using AdaFace alone. The IJB-B dataset is specifically designed to incorporate low-quality images within its validation protocol. The improvement underscores our algorithm’s robustness with varying image qualities. Additionally, the average values (AVG) presented in Table 3.1 indicate that the accuracy for the LFW, CFP, and AgeDB datasets generally improves when our normalization method is incorporated, further validating the efficacy of our approach. Figure 3.2 highlights the efficacy of the proposed method in normalizing faces in scenarios with variations in pose, expression, and illumination between input pairs.

3.3.5 Ablation Study

Assessment of Verification Accuracy via 3DMM Identity Coefficients: To assess verification accuracy using 3DMM identity coefficients, we conducted an ablation analysis in this study. We compared the identity coefficients directly extracted from the input pairs using the R-Net, labeled as ‘R-Net α coefficient’ in Table 3.1, with those coefficients post-normalization, termed ‘R-Net α coefficients after normalization’. Initial results indicated that the verification accuracy with ‘R-Net α coefficients’ was lower than that of SoTA methods. Nevertheless, upon updating these coefficients to derive x_{ν} (‘R-Net α coefficients after normalization’), a significant improvement was observed. It is crucial to note that even with the enhanced coefficients from the R-Net post-normalization, the verification accuracies did not exceed those of SoTA methods. The further improvement was observed only after processing the normalized faces, utilizing optimized identity coefficients, through the FR model for feature embedding. This advancement can be attributed to the training of FR models (such as AdaFace) and the evolution of margin-based loss functions, which have markedly increased the discriminative power of face embeddings.

3.4 Conclusions

Our proposed solution, AlignFace, introduces a novel approach to face normalization, specifically designed for preprocessing input images within the

context of face verification. AlignFace stands out by focusing on aligning head pose, expression, and illumination conditions between two images. It proficiently estimates these conditions in one image and reconstructs the other to match, while carefully preserving the unique identity features of each image. This method ensures the preservation of genuine facial features in one image, providing a more accurate representation of facial characteristics. Our experimental results underscore AlignFace’s superiority over existing state-of-the-art methods in face verification across multiple benchmark datasets.

Chapter 4

Deepfake Quality Assessment

This chapter focuses on the evaluation techniques for face-reenactment methods. In Section 4.1, we introduce various face manipulation techniques, with a particular emphasis on face-reenactment. We also discuss the challenge of deepfake quality assessment and present our motivation for this study. In section 4.2, we propose a method for assessing images generated by face-reenactment techniques. Section 4.4 details our experiments and results, while Section 4.5 and 4.6 concludes the chapter with a summary of our findings and future research directions.

4.1 Introduction

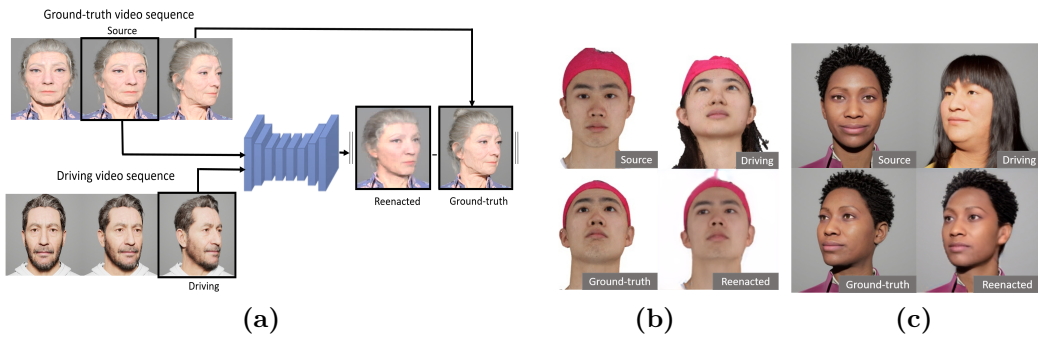


Figure 4.1: Proposed protocol (a). Examples of the source image, driving video frame, generated frame, and corresponding ground-truth provided by our proposed protocol for both the real (b) and synthesized (c) datasets.

The face serves as a highly expressive and complex nonverbal communication channel for humans. The advancements in AI-generated synthetic

faces, known as Deepfakes, have brought about significant benefits in various domains, including education, film production, and dubbing.

Among the fundamental techniques in DeepFake face manipulation are face swapping and face-reenactment. Face swapping involves transforming a face from a source image to seamlessly replace the face in a target image, achieving a result where the replacement blends naturally into the target image. Face-reenactment methods, on the other hand, aim to generate a synthesized video that animates a target face based on the movements captured from a driving video, while preserving the identity conveyed by the source image. This process involves treating the person in the source image as a controllable puppet, with the facial expressions, head pose, and movements from the driving video defining the corresponding actions in the synthesized video.

Recent face-reenactment techniques [56], [68], [74], [133]–[135] have leveraged generative models such as Encoder-Decoder (ED) networks [60], Variational Auto-Encoders (VAEs) [32], and Generative Adversarial Networks (GANs) [8] to generate images that push the boundaries of realism, making it increasingly challenging to discern between what is real and what is artificially generated. Despite the progress made in the development and application of face-reenactment methods, evaluating the realism and accuracy of the generated images, particularly in cross-reenactment scenarios where a different individual’s face is used to reenact the source face, remains a significant challenge. Directly employing image based quality metrics, such as Structural Similarity Index (SSIM) [67] or facial keypoint errors is impractical due to the absence of ground-truth data.

To address this challenge and quantitatively assess the quality of images generated through cross-reenactment, researchers have investigated the extraction of feature embeddings from both the source and generated faces. Subsequently, they calculate the errors or discrepancies between these extracted features [70], [75], [76]. Although this approach offers partial solutions for cross-reenactment evaluation, it is confined to specific metrics and lacks a comprehensive assessment.

consuming, especially when dealing with a large number of samples. Therefore, there is an urgent need to develop a new evaluation protocol that can effectively assess the fidelity of cross-reenactment images.

This work introduces a novel protocol for the quantitative evaluation of images produced by face-reenactment techniques, particularly in cross-reenact scenarios. The protocol enables assessment of cross-reenactment images using metrics that rely on explicit ground-truth such as SSIM and LPIPS. To overcome the limited availability of appropriate datasets, two video generation approaches are proposed. The first approach involves the utilization of 3D

models of real heads acquired using a multi-view system. In the second approach, realistic synthesized head models are employed, encompassing a wide range of human subjects, facial expressions, pose variations, and lighting conditions.

Our proposed protocol is applied using these datasets, along with established metrics such as SSIM [67], Cosine Similarity (CSIM) [136], Learned Perceptual Image Patch Similarity (LPIPS) [137], Average Keypoint Distance (AKD), Fréchet Inception Distance (FID) [66], and Fréchet Video Distance (FVD) [138] to assess the quality of four well known and state-of-the-art reenactment methods: FOMM [68], X2Face [56], LIA [60], and DaGAN [59].

In addition to quantitative evaluation, a series of user studies are conducted to investigate the effectiveness of our proposed protocol. These studies analyze the generated images in terms of identity preservation, head pose and facial expression replication, and overall image similarity, providing further validation of our quantitative results.

4.2 Proposed Methodology

This section presents our proposed protocol for evaluating the image quality of reenactment methods, with a focus on cross-reenactment scenarios. To fulfill this objective, we generate video sequences comprising different identities with precisely controlled and known head pose and expression for each frame. These video sequences are then employed in conjunction with our proposed protocol and a set of quantitative metrics to measure the fidelity of images generated by various reenactment methods. In the following subsections, we provide a detailed description of the proposed protocol and the process of data generation.

4.2.1 Protocol

The pipeline of our proposed protocol is depicted in Figure 4.1. The protocol involves two video sequences, denoted as A and B, representing distinct identities. For each frame, the head pose and expression are identical in both sequences. Initially, any frame can be selected from video sequence A as the source image, representing the face to be reenacted. Subsequently, the video frames of identity B are utilized to animate the source image, resulting in frames of identity A that simulate the expressions and movements of identity B. These generated frames, known as deepfake frames, are then compared with the original frames of identity A in the ground-truth video sequence to evaluate the accuracy of the cross-reenactment process. The evaluation

protocol can be summarized as follows:

1. Select a frame from video sequence A as the source image. In our experiments, we begin with frames displaying a frontal head pose and a natural expression, gradually introducing extreme variations in head pose and expression.
2. Select a driving video sequence, comprising video frames of identity B, to animate the source image. The head pose and expression in all frames of the driving video correspond to those of the source face.
3. Input the source image and driving video frames into a face-reenactment method to generate a new video sequence representing source identity A. This generated video sequence should accurately reflect the facial expressions and movements that match those of the driving video sequence.
4. Assess the accuracy of the generated frames by comparing them to the ground-truth video using metrics such as SSIM, CSIM, LPIPS, AKD, FID, and FVD.

4.2.2 Dataset Generation

Two video datasets were generated for evaluating face-reenactment techniques: one comprised real face models generated from the FaceScape dataset [139], and the other consisted of synthesized MetaHumans [140].

4.2.3 Real Face Dataset

To create a dataset comprising real human subjects, we employed the Pyrender 3D environment and utilized FaceScape [139], a well-established 3D face dataset. The FaceScape dataset consists of multi-view RGB images and intrinsic and extrinsic camera parameters, which were captured using 68 DSLR cameras. Leveraging this data, we generate 3D head point clouds with RGB values for various individuals exhibiting a neutral expression. By placing these 3D head models in desired scenes and defining specific camera parameters, we render them in the desired head poses. Figure 4.2 illustrates the rendering process.

In our study, we generated a total of 40 video sequences to investigate the impact of head rotations on face-reenactment. These sequences included five unique identities, and for each identity, we incorporated eight specific head rotations. The primary objective was to highlight different head rotation

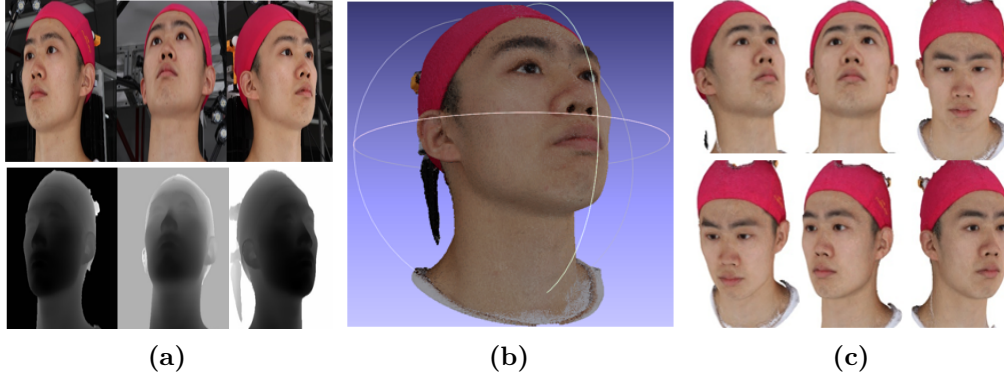


Figure 4.2: Multiview RGB images and their corresponding depth maps utilized to inverse project pixels into point clouds (a). The resulting reconstructed 3D head model (b). Rendered images of 3D models from desired angles(c).

scenarios, namely a rotation around the pitch axis, a rotation around the yaw axis, and a combined rotation involving both pitch and yaw axes. To ensure consistent evaluation, each video began with the frontal head position in the initial frame and gradually transitioned the head towards the desired rotation axis in the final frame. Throughout the duration of the clips, the facial expression of the subjects remained constant. Each video clip consisted of 100 frames with a resolution of 512×512 pixels.

4.2.4 Synthesized Dataset of MetaHumans

Evaluating the performance of face-reenactment methods solely using real data has limitations in assessing their ability to handle different facial expressions, as the individuals in the real dataset maintain a neutral expression throughout all the videos. To establish accurate ground-truth for facial expressions in the context of real datasets, image matching techniques like optical flow can be employed to reconstruct different expressions [141], [142]. However, the potential errors associated with these techniques necessitate an alternative approach. Therefore, we propose utilizing synthesized data, which provides complete control over the scene, allowing precise manipulation of geometry and appearance. This approach ensures data reliability and creates a controlled and accurate evaluation environment.

We utilized the Unreal Engine and the MetaHuman asset from the Quixel Bridge library [143] to generate a realistic synthesized face video dataset. MetaHumans are 3D human models created with advanced scanning, rigging, and animation technology, featuring high-quality photo scans of real skin textures and additional artificial textures for details like light reflection

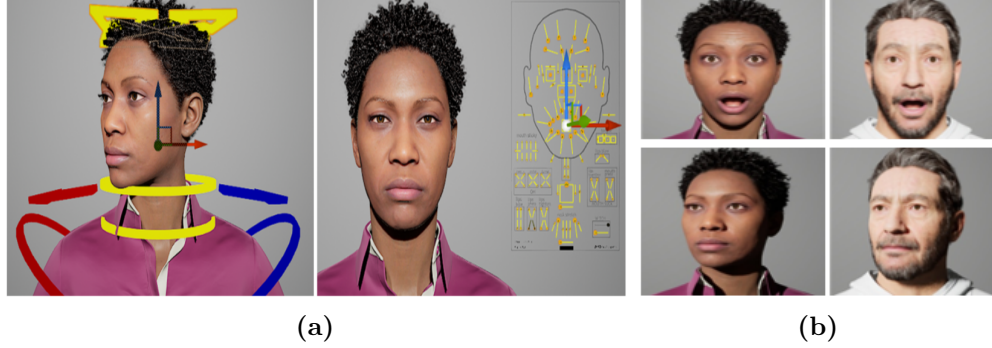


Figure 4.3: Head (left) and Face (right) Control Rig Boards enabling adjustment of pose and facial expression (a). Two MetaHumans with identical facial expressions and head poses (b).

and surface roughness. Their riggability enables precise control over facial expressions and movements. To generate the video dataset, the scene was set up in the Unreal Engine with adjusted lighting conditions and configured camera properties. MetaHuman characters were placed within the scene and animated using the Control Rig Board as shown in Figure 4.3a. The resulting animations were rendered, capturing the desired facial expressions and movements.


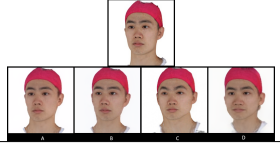
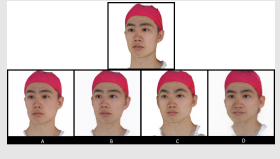
In Unreal Engine, the process of applying animations from one MetaHuman character to another is straightforward. By substituting the model references in the scene, the animations originally designed for the first character can be effortlessly transferred to the second character. This replacement ensures that both characters share the same expression setting, resulting in identical head pose and facial expression. Leveraging this capability, it becomes possible to generate multiple videos, each showcasing a different identity, while preserving consistent head pose and expression across all videos. Figure 4.3b illustrates two MetaHuman identities with the exact same head pose and expression.

The video sequences were meticulously designed to ensure a structured progression, starting with a frontal head position and neutral expression and culminating in expressive facial expressions or head rotations, or a combination of both. These sequences encompassed a diverse set of facial expressions, including amusement, anger, disgust, laughter, sadness, and surprise. The head rotations in the dataset covered 8 rotations around the yaw axis, pitch axis, and combinations of the pitch and yaw axes, including various directions such as up, down, left, right, and diagonal directions. A total of 20 distinct face movement animations were produced for each of the 10 MetaHuman identities, resulting in a dataset comprising 200 videos (10 identities \times 20 face animation). All videos were rendered at a resolution of 1920×1080 pixels,

ensuring a high level of visual quality and detail for the evaluation process.

4.3 Subjective Evaluation

Table 4.1: Summary of subjective evaluation methods.

Evaluation name	Objective	Videos/ Images	Number of scenarios	Blind comparison	Subjective Test Example
VR	Assess perceived 'Realism' of generated videos	Videos	46	Test Includes Ground-Truth	
IS	Evaluate users 'Satisfaction' with generated outputs for specific head rotations	Images	132	Explicitly Informed (On Top)	
VI VPE VS	Assess quality focusing on 'Identity' preservation (VI), head 'Pose and Expression' preservation (VPE), and overall 'Satisfaction' (VS)	Videos	20	Explicitly Informed (On Top)	

Three subjective evaluations were conducted to assess the proposed protocol and evaluate the strengths and weaknesses of each reenactment method. These evaluations utilized a set-wise ranking method, where participants were presented with a set of videos or frames and tasked with directly comparing and organizing them based on specific criteria. Table 4.1 provides a summary of the three evaluation methods along with an example of each test conducted with the participants. The evaluations involved the participation of 23 professionals specializing in computer vision, ensuring their expertise in accurately assessing the fidelity of the generated frames produced by face-reenactment methods. Prior to the evaluation, participants were provided with detailed explanations of each test and completed practice tests to ensure their comprehension of the procedures. To optimize the evaluation time per participant, the test dataset was divided randomly into two batches, allowing participants to complete half of the test. On average, each evaluation session lasted approximately one hour.

In the first evaluation, titled "Realism Assessment," participants were presented with sets of five videos that included one ground-truth video and four reenacted videos. The videos were carefully selected to cover a wide range of facial expressions and head rotations. Participants were asked to

rank the videos based on their perceived realism, using a scale from 1 to 5. To minimize bias, the order of the videos within each set was randomized, and participants were unaware of which video was the ground-truth.

The second evaluation, titled "Overall Satisfaction with Head Rotation," aimed to assess users' overall satisfaction with the generated outputs at specific head rotations. Participants were presented with sets of four frames generated by the reenactment methods, along with a ground-truth image depicting a specific head pose. Participants were explicitly informed about the identity of the ground-truth image and instructed to compare each generated image to the ground-truth. They were then asked to assign a rank to each image on a scale of 1 to 4, indicating their overall satisfaction relative to the ground-truth image.

The third evaluation aimed to assess the quality of the generated videos, focusing on three aspects: 1) identity preservation, 2) head pose and expression preservation, and 3) users' overall satisfaction. Participants were presented with sets of four videos alongside the ground-truth video and were asked to rank each video in relation to the ground-truth. The rankings were reported separately for the preservation of identity, head pose and expression, and overall satisfaction. Participants provided scores ranking from 1 to 4, with 1 indicating the highest satisfaction. The first test consisted of 46 scenarios, the second test had 132 scenarios, and the third test comprised 20 scenarios.

Statistical analysis of subjective evaluation: To assess the distance between reenactment methods through subjective evaluation, each technology is assessed by a group of observers using a set of images and videos. We utilize the outlier detection and scaling method described in the study by Perez et al.[144], which is based on Thurstone's model and its assumptions [145]. This method, given a matrix that includes the results for all participants, measures the probability of observing the data of each observer in comparison to the rest of the observers.

The method uses Maximum Likelihood Estimation (MLE) to compute an inter-quartile-normalized score for each subject. Let's suppose we aim to compare n conditions O_1, \dots, O_n (i.e., n technologies) with unknown underlying true quality scores $q = (q_1, \dots, q_n)$, where $q_i \in \mathbb{R}$ represents the quality score for condition O_i . The goal of this analysis is to estimate scores $\hat{q} = (\hat{q}_1, \dots, \hat{q}_n)$ that approximate the true quality q .

The perceived quality of a condition O_i is modeled as a random variable: $r_i \sim N(q_i, \sigma)$, where the mean of the distribution is assumed to be the true quality score q_i . When scaling the data, the focus is primarily on recovering the distance $q_i - q_j$ between underlying quality scores q_i and q_j (as scores are relative). If we know the true probability of selecting O_i as better than O_j ($P(r_i > r_j)$), the probability that O_i was selected over O_j in exactly c_{ij} trials

Table 4.2: Evaluation results for cross-identity reenactment for real dataset.

Method	Quantitative Evaluation using the Proposed Protocol						Subjective Evaluation (JOD)					Traditional CSIM↑
	SSIM↑	LPIPS↓	CSIM↑	AKD↓	FID↓	FVD↓	VR↑	IS↑	VI↑	VPE↑	VS↑	
X2Face [56]	0.749	0.260	0.695	3.892	39.4	224.0	1.065	1	1	1	1	0.52
FOMM [68]	0.788	0.222	0.867	1.983	32.2	202.4	1	1.264	1.244	1.843	2.096	0.71
DaGAN [59]	0.803	0.159	0.833	2.883	34.6	217.1	1.964	2.654	2.139	2.640	2.164	0.66
LIA [60]	0.818	0.133	0.847	2.137	30.9	210.5	3.154	3.989	4.053	4.532	4.165	0.64
Ground-truth							5.071					

out of the total number of $n_{ij} = n_{ji} = c_{ij} + c_{ji}$ trials is given by the binomial distribution.

$$\begin{aligned}
L(\hat{q}_i - \hat{q}_j \mid c_{ij}, n_{ij}) &= \\
\binom{n_{ij}}{c_{ij}} P(r_i > r_j)^{c_{ij}} (1 - P(r_i > r_j))^{n_{ij} - c_{ij}} &= \\
\binom{n_{ij}}{c_{ij}} \Phi\left(\frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}}\right)^{c_{ij}} (1 - \Phi\left(\frac{\hat{q}_i - \hat{q}_j}{\sigma_{ij}}\right))^{n_{ij} - c_{ij}}, & \quad (4.1)
\end{aligned}$$

Where, c_{ij} represents the count of cases where condition O_i was chosen as better than condition O_j , out of a total number of trials $n_{ij} = n_{ji}$. The true probability of choosing condition O_i over condition O_j can be computed using the cumulative normal distribution Φ , given two Gaussian random variables r_i and r_j .

$$P(r_i > r_j) = P(r_i - r_j > 0) = \Phi\left(\frac{q_i - q_j}{\sigma_{ij}}\right), \quad (4.2)$$

The parameter σ_{ij} represents the noise parameter in Thurstone’s model [145]. It is typically determined based on the probability p_{ij} of a 1 Just-Objectionable-Difference (JOD) unit, as described in Perez et al. [144]. The scaling of the comparisons is then performed by maximizing the products of the likelihoods.

$$\arg \max_{\hat{q}_2, \dots, \hat{q}_n} = \prod_{i,j \in \Omega} L(\hat{q}_i - \hat{q}_j \mid c_{ij}, n_{ij}) \quad (4.3)$$

where Ω denotes the number of pairs with at least one made comparison. Subjects with an inter-quartile-normalised score above a threshold of 1.5 are tagged as outliers and discarded.

4.4 Experiment and Results

Dataset: Two video datasets were compiled to assess face-reenactment techniques. The first dataset included 40 videos of real face models, featuring

Table 4.3: Evaluation Results for Cross-identity Reenactment for synthesized MetaHuman Dataset.

Method	Quantitative Evaluation using the Proposed Protocol						Subjective Evaluation (JOD)					Traditional
	SSIM \uparrow	LPIPS \downarrow	CSIM \uparrow	AKD \downarrow	FID \downarrow	FVD \downarrow	VR \uparrow	IS \uparrow	VI \uparrow	VPE \uparrow	VS \uparrow	CSIM \uparrow
X2Face [56]	0.656	0.190	0.652	4.821	50.6	293.5	1	1	1	1	1	0.61
FOMM [68]	0.687	0.182	0.838	3.971	41.6	257.7	2.159	2.918	1.805	2.187	2.293	0.67
DaGAN [59]	0.821	0.147	0.865	1.902	45.4	271.5	3.075	4.034	2.557	2.789	3.320	0.64
LIA [60]	0.836	0.142	0.874	2.159	43.6	255.2	4.004	5.438	2.996	3.300	3.490	0.68
Ground-truth							5.269					

five identities with 8 head rotations each. The second dataset comprised 200 synthesized videos of MetaHumans, exhibiting 10 identities with 20 variations of head movement and facial expressions. A systematic approach was employed for both datasets, selecting first frame of one video as the source for each identity and utilizing the remaining videos from the same face animation type but different identities as driving videos. This methodology yielded a total of 1960 scenarios, with 160 scenarios derived from the real dataset and 1800 scenarios from the synthesized dataset. Table 4.1 provides an overview of the scenario distribution in the three subjective tests, ensuring an equal representation of synthesized and real scenarios in each test. These datasets offer a comprehensive and diverse range of scenarios, providing valuable insights into the performance of face-reenactment methods.

Methods and Metrics: In our evaluation, we compare the performance of four face-reenactment methods: FOMM [68], X2Face [56], LIA [60], and DaGAN [59]. The effectiveness of these methods is evaluated using six widely recognized metrics: SSIM [67], CSIM [136], LPIPS [137], FID [66] and FVD [138]. The CSIM metric utilizes facial embeddings extracted through the ArcFace [5] face recognition model to measure content similarity between generated and ground-truth images. The AKD metric quantifies keypoint discrepancies by extracting 468 facial landmarks using the MediaPipe library [146]. To interpret subjective evaluation results, we employ Thurstone’s model assumptions to scale the ranking scores, as detailed in Section 4.3. The scores are represented on the Just-Objectable-Difference (JOD) scale, where a difference of 1 JOD signifies that 75% of observers favored one condition over another.

Evaluation and Analysis of Protocol Performance: Table 4.2 presents the performance evaluation results of cross-reenactment methods on real datasets, while Table 4.3 showcases the results on synthesized Metahuman datasets. The evaluation is conducted using various quantitative metrics, including SSIM, AKD, and LPIPS, which are computed based on 1960 scenarios derived from 240 videos. These metrics are employed to measure the disparities between the reenacted images and the corresponding ground-truth

images provided by our protocol design. Additionally, our evaluation protocol incorporates the utilization of CSIM, FID, and FVD, which are commonly employed in existing face-reenactment evaluation.

FID assesses the photo-realism of the generated samples by comparing them to the ground-truth images at a deep feature level, while FVD, a modified version of FID, accounts for temporal coherence by considering spatial-temporal features. Notably, these metrics operate at the data distribution level, rather than focusing on individual frames. The calculation of FID and FVD metrics remains consistent with existing approaches since the ground-truth comprises data distributions of the Metahumans and real head videos.

In our analysis, we also incorporate the calculation of CSIM using the existing protocol depicted in Figure 2.4b, referred to as $\text{CSIM}_{\text{trad}}$. This metric evaluates the cosine similarity between the source and reenacted faces. However, the presence of distinct head poses between the source and reenacted faces poses a challenge, resulting in lower CSIM scores in traditional evaluation compared to the measurements obtained through our protocol.

Furthermore, the subjective test results are reported in both Table 4.2 and Table 4.3. The subjective evaluation serves multiple objectives in our study: firstly, it allows for the identification of strengths and weaknesses of each face-reenactment method, providing qualitative insights into their performance. Secondly, it enables the assessment of the effectiveness of our proposed protocol compared to existing evaluation approaches. Lastly, the subjective results aid in determining the most informative quantitative metrics within our protocol that best describe the quality of reenacted images, thereby facilitating the identification of suitable metrics for future evaluations.

During the subjective tests, the reenactment methods are evaluated based on their performance in generating realistic content (VR_{JOD}), preserving identity (VI_{JOD}), transferring pose and expression (VPE_{JOD}), and overall satisfaction (VS_{JOD}). Statistical analysis reveals that the LIA method consistently achieves the highest scores in all subjective tests, slightly surpassing DaGAN. Both LIA and DaGAN consistently outperform X2Face and FOMM. A significant finding emerges from the blind comparison between the ground-truth and reenacted videos. The VR_{JOD} scores, calculated based on blind comparisons where the ground-truth is questioned alongside the reenacted videos, indicate that all four reenactment methods fail to generate sufficiently realistic content. Human subjects were able to distinguish reenacted content from the ground-truth images. FID and FVD are commonly used metrics to assess image and video realism. It is noteworthy that although FOMM demonstrates a good FID score, it does not align with the qualitative results (VR_{JOD}).

Furthermore, FOMM exhibits good scores in CSIM and AKD, which are considered identity preservation metrics in the literature. For example, its CSIM and CSIM_{trad} scores in real dataset evaluation outperform other methods. It should be noted that FOMM employs relative keypoint locations to address the identity preservation problem, which seemingly increases CSIM, CSIM_{trad} , and AKD scores. However, its subjective score VI_{JOD} is lower than both LIA and DaGAN. To determine which quantitative metrics better describe the quality of reenacted images, the Pearson correlation coefficient is presented in Figure 4.5. The results demonstrate that the frame-based metrics within our protocol, where the ground-truths are provided, exhibit the strongest correlation with subjective evaluations.

Pose Transferability Evaluation Using Our Proposed Protocol: Supplementing the results in Tables 4.2 and 4.3, we conducted a comprehensive analysis encompassing subjective and quantitative results using both the real head dataset and the synthesized dataset. A dedicated subjective test was conducted to assess overall satisfaction with image-based reenactment, with a specific focus on head rotation at different degrees. The driving sequences were incrementally rotated up to 50 degrees while maintaining natural facial expressions. The resulting overall scores, denoted as IS_{JOD} scores, were calculated for various head rotation scenarios, including rotations around the pitch axis, yaw axis, and combinations of pitch and yaw axes. The obtained scores are presented in Table 4.2 for the real head dataset and in Table 4.3 for the synthesized MetaHuman dataset.

To further analyze the quality of generated images under specific rotation conditions, we provide results for yaw rotation (right) and yaw-pitch rotation (up and left) in Figure 4.4. In addition to the subjective evaluations, quantitative scores such as SSIM, CSIM, and AKD were computed using ground-truth data as per our proposed protocol. Based on the findings presented in Tables 4.2 and 4.3, both the LIA and DaGAN methods demonstrate comparable performance in generating animated faces. However, based on Figure 4.4 they exhibit distinguishable sensitivities to head rotation. Through the subjective tests and SSIM evaluation, it is evident that LIA performs better in scenarios with more significant head movement in the driving video. Conversely, DaGAN exhibits superior performance in scenarios involving minimal head rotation, particularly those closer to the frontal head pose. Notably, DaGAN’s quality deteriorates gradually, and beyond a certain threshold (approx. 30°), it becomes comparable to or even worse than FOMM. In contrast, the FOMM method showcases resilience to head rotation, as the quality of reenacted images remains relatively unaffected and comparable to scenarios with a frontal head pose. When evaluating the CSIM and AKD metrics, FOMM achieves scores on par with those of LIA and DaGAN. However, its SSIM

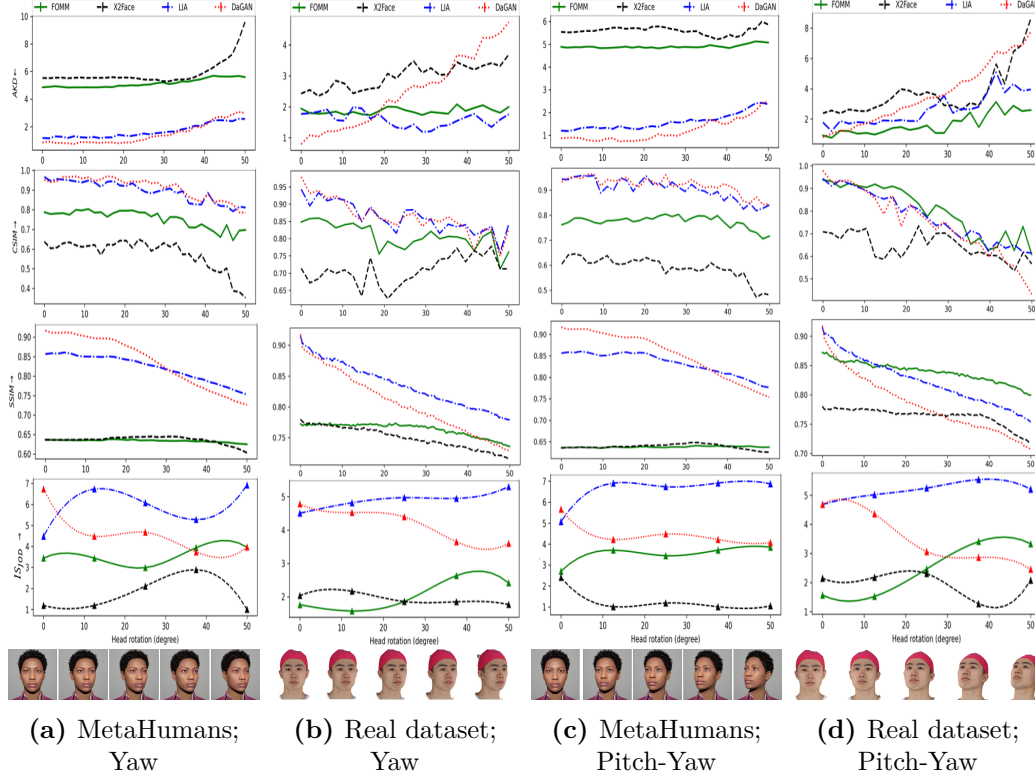


Figure 4.4: Pose transferability evaluation using our proposed protocol. The figure presents the results of the image-based overall satisfaction subjective test scores (IS_{JOD}) for different head degrees, along with the corresponding quantitative scores such as SSIM, CSIM, and AKD, computed using ground-truth data following our proposed protocol.

score is notably lower.

4.5 Future Work

The application of our proposed protocol to face swapping methods shows great promise for future research. To implement our protocol for face swapping, we recommend utilizing our MetaHumans dataset and creating a comprehensive ground-truth by integrating elements generated from diverse sources. Specifically, the backgrounds, body and hairstyles can be preserved and rendered similarly to the driving videos, while the face identities should be derived from the source images.

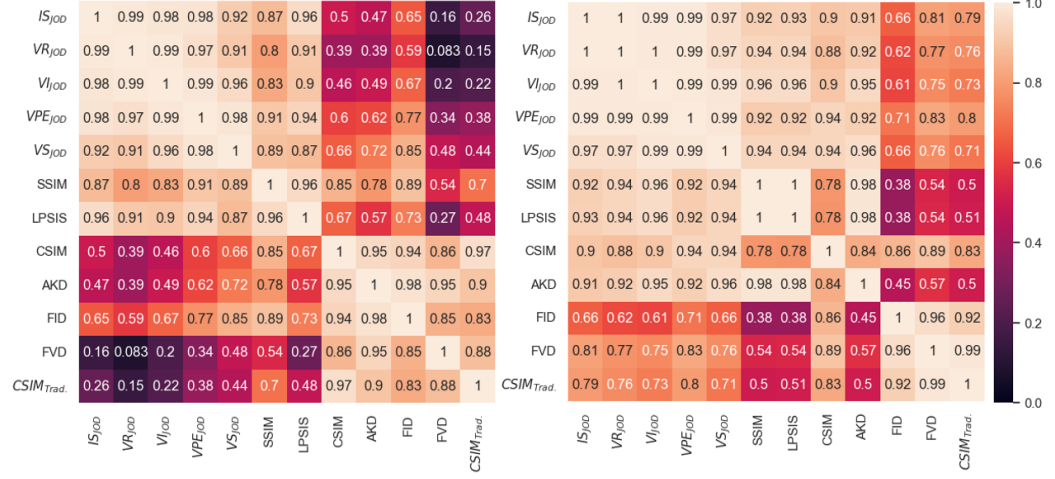


Figure 4.5: Confusion matrix depicting the correlation of metrics within Real (left) and synthesized (right) datasets

4.6 Conclusion

This work presents a novel protocol for evaluating the realism and accuracy of face-reenactment generators in cross-reenactment scenarios. Comparative analysis with existing evaluation approaches demonstrates the effectiveness of our protocol, supported by user studies validating its efficacy in analyzing identity preservation, head pose, and facial expression replication. The results reveal a strong correlation between subjective evaluations and frame based metrics (e.g., SSIM and LPIPS) within our protocol.

Chapter 5

Effect of Beautification Filters on Deepfake Detectors

5.1 Introduction

Artificial Intelligence Generated Content has drawn significant attention in both academic and industrial realms in recent years, particularly with the notable advancements in deepfake technology within the generative domain [147]. This technology performs remarkably well at creating highly realistic facial media content, transitioning from traditional graphics-based methods to sophisticated deep learning approaches by initially employing advanced techniques such as Variational Autoencoders (VAEs) [35] and Generative Adversarial Networks (GANs) [8].

Recent advancements in diffusion models [9] have significantly enhanced the capability to generate high-quality images and videos, advancing deepfake technology into various practical applications such as entertainment, art, and education. However, these technological advancements also introduce substantial risks [148], [149]. Deepfakes create opportunities for criminal misuse, such as impersonating individuals to commit fraud or deceive others into divulging sensitive information. For example, deepfake audio or video calls can convincingly mimic trusted contacts like family members or professionals, exploiting established trust. Additionally, deepfakes pose a threat to security systems by potentially circumventing facial recognition and fooling biometric authentication software, thereby granting unauthorized access to restricted areas or sensitive data. This vulnerability is particularly concerning for mobile devices used for secure unlocking, financial transactions, or access to medical records, posing significant security risks for both users and applications.

To mitigate the risks posed by deepfakes, detection methods have evolved

from early handcrafted feature-based techniques [10], [11] to modern deep learning approaches [12]–[16] and more recent hybrid models [17], [18]. Detection tasks are typically framed as classification problems, applied either at the image level or the video level, depending on the practical application. These models can be represented as:

$$S_o = \phi_D(I_o), \quad (5.1)$$

where ϕ_D abstracts the specific detection network, and S_o represents the fake score of the generated content I_o .

Existing methodologies in deepfake detection typically rely on a supervised approach. This involves developing a real vs fake image classifier by assembling a large dataset of generated images from multiple generative models and training a binary classifier. However, in practical scenarios, the specific techniques used for facial manipulation are unknown beforehand, and access to the attacker’s model is typically unavailable. Despite achieving high detection accuracies, approaching 98%, these classifiers are prone to overfitting. This limitation restricts their effectiveness to the manipulation techniques on which they were originally trained, resulting in significant performance degradation when confronted with forgeries generated by new, previously unseen methods.

Recent studies have recognized this challenge and aimed to enhance the robustness of detection algorithms by focusing on intrinsic indicators of forgery that go beyond relying solely on known manipulation characteristics. For example, methodologies such as Face X-ray [84] and SBIs [85] target blending artifacts directly rather than general forgery traces, thereby significantly improving generalization capabilities. Nevertheless, these approaches remain vulnerable to common perturbations, as blending artifacts can be influenced by various image and video processing operations such as compression.

In practical settings, content captured by cameras often undergoes various digital image and video processing operations, including post-processing such as stylization filters [150] before dissemination. Recent efforts have systematically quantified the impact of such operations on detection accuracy [22]–[25]. These investigations consider factors such as noise, resizing, compression, denoising, contrast and brightness adjustments, and changes in resolution. Consequently, strategies like stochastic degradation-based augmentation, driven by typical image and video processing operations, have been proposed [151] to enhance the generalization of deepfake detection tasks. Among these operations, compression has garnered particular attention, as applying deepfake detectors to compressed videos often results in decreased detection performance [152], [153].

Recently, beautification filters have gained widespread popularity, enabling users to enhance their facial appearance through integrated social media tools.

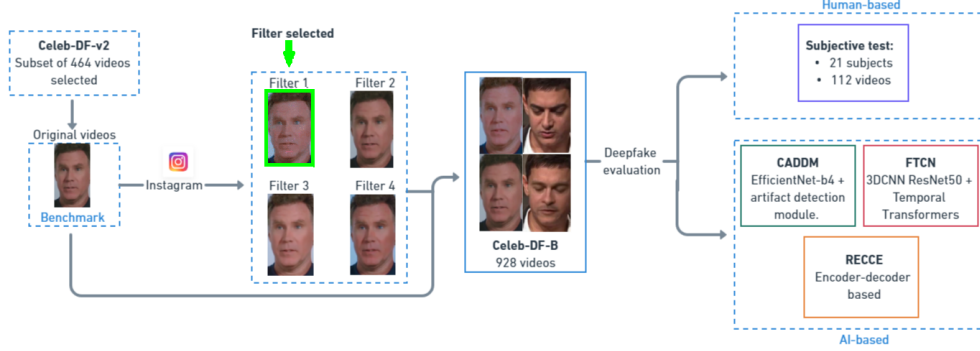


Figure 5.1: Pipeline of the proposed method. A subset of 464 videos (50% Real and 50% Fake) are selected. Each video is uploaded to the social network Instagram, where one of the four different filters is randomly selected and applied to it. The four filters uniformly appear in the Celeb-DF-B database. The final database has a size of 928 videos and it is used to perform a human-based deepfake detection and to evaluate the robustness of three SotA AI-based detectors.

This raises the question of whether such filters amplify manipulation artifacts in deepfake videos or suppress them, potentially affecting the performance of deepfake detection algorithms. In this section, we describe our approach and experimental analysis.

Social media platforms offer a diverse range of tools referred to as "filters" designed to automatically enhance user's image, demanding minimal or no user proficiency [154]. Certain types of filters are designed to tweak different facial features such as skin, lips, eyes, and nose to enhance the beauty of the user. We will refer to those filters as *beautification filters*. Some common modifications are makeup addition, narrow noses, skin tanning and smoothening. *Beautification filters* have been demonstrated as a disturbance factor for AI facial processing tasks such as face recognition and gender classification [154]. Despite deepfake detection technology being challenged against several video processing operations [155], its robustness against social media beautification effects has not yet been tested.

In this work, we study the behavior of 3 SotA passive deepfake detectors trained on the FaceForensics++ (FF++) dataset [77]. Our objective is to test the robustness of deepfake detectors against beautified videos and measure the impact of the *beautification filters* on the classification score. Moreover, we compare the performance of those detectors with the ability of an average user to classify real and fake videos when they are beautified. The pipeline is presented in Figure 5.1. The key contributions of this study include:

- We introduce a new benchmark dataset, the Celeb-DB-B database based

on a subset of videos from the Celeb-DF dataset and composed of 928 videos balanced in terms of four categories Real, Real-Beautified, Fake, Fake-Beautified;

- We study the impact of those filters on three deepfake detectors finding a drop in performance for video-level AUC and revealing how social media beautification can be used to make fake videos look more authentic;
- Finally a subjective evaluation is conducted to investigate whether the utilization of beautification filters presents challenges for human observers when distinguishing between the authenticity of deepfake and real videos.

The rest of this work is organized as follows. Section 5.2 presents the creation of the Celeb-DF-B dataset. In Sections 5.3 and 5.4, we present the performance of deepfake detectors and users on the Celeb-DF-B dataset. Finally, conclusions are summarized in Section 5.5

5.2 Dataset

In this section, we introduce the protocol employed in the creation of the Celeb-DF-B database, its composition, and the specific social media filters chosen for face beautification.

The Celeb-DF [156] dataset consists of 590 real videos and 5639 DeepFake videos. The average duration of all videos is approximately 13 seconds, with a standard frame rate of 30 frames per second. The real videos are sourced from publicly accessible YouTube content corresponding to interviews featuring 59 celebrities. Among these, for the creation of the Celeb-DF-B database, we chose a subset consisting of 232 real and 232 fake videos. The selection of videos followed three criteria: 1) an equal sampling from each identity in the real videos; 2) pairing each real video with a fake counterpart created through FaceSwap; and 3) maintaining a balance in the number of time one identity is used as source and target.

Once the data was sampled, *beautification filters* were applied to the videos as depicted in Figure 5.1. Instagram was selected as the filter provider due to its large selection of available *beautification filters* which users regularly apply to enhance their multimedia content. In Table 5.1 we present the selected *beautification filters* along with the facial traits modified by them. Each of the 464 non-beautified videos is beautified with one of the four selected filters resulting in the creation of 928 videos that constitute the Celeb-DF-B dataset. Example frames of the videos belonging to the Celeb-DF-B database are displayed in Figure 5.2.

Table 5.1: Characteristics of the selected Instagram filters. Traits modifications were assessed by visual inspection of pixel differences between original and filtered images.

Filter	Color	Skin	Makeup	Eyes	Nose	Lips
BROWN	x	x	x		x	
California dreamin	x	x			x	x
Relax! You Pretty!	x	x			x	x
Hawaii Grain	x	x	x	x	x	x

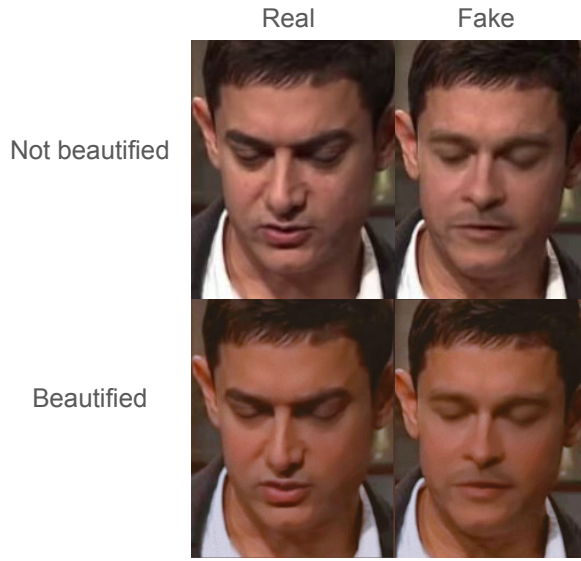


Figure 5.2: Frames extracted from four distinct videos within the Celeb-DF-B database are depicted here.

5.3 Experimental Setup and Results

5.3.1 Deepfake Detectors

We selected 3 different passive deepfake detectors for this experiment: CADDM [14], RECCE [93], and FTCN [157].

CADDM detects traces of forgery on the frame level. First, the image is passed through an EfficientNet-b4 [158] backbone to extract useful features for the classification task. Then, it detects forgery locations on different scales through an artifact detection module trained with a custom Multi-scale Face Swap algorithm to generate forgery location ground truth. The average of the scores between the individual frames becomes the classification score of the video. With this architecture, the model focuses more on local forgeries instead of learning face distribution to perform better while detecting fakes

of unseen faces.

RECCE is an encoder-decoder-based model. The encoder is based on Xception [159]. The reconstruction network has been trained in an unsupervised manner to learn the representation of real faces. The face frames are passed through the encoder-decoder architecture. Then, encoder and decoder features are agglomerated together with the residual images (i.e. the difference between the reconstructed and the original frame) to classify each frame as fake or genuine. The video’s classification score is computed as the average score between each frame.

FTCN is a model trained to detect temporal inconsistencies in videos. Because deepfakes are generated frame by frame, they are likely to present temporal incoherences. FTCN network has a Resnet50 3DCNN [160] backbone to extract temporal features and a Temporal Transformers [161] as a classifier. Therefore, FTCN does not look for manipulations on each image independently but on a sequence of frames.

5.3.2 Experimental Setup

Metrics: To evaluate the three selected deepfake detectors, we compute the video-level AUC of the Receiver Operating Characteristic (ROC) curve and the False Negative Rate (FNR), i.e., the proportion of fake videos recognized as genuine, which, in a real-case scenario, is desirable to minimize. Additionally, we analyze the histogram of the classification scores before and after beautification to gain a better understanding of the behavior of deepfake detectors on beautified videos.

Evaluation protocol: We follow the evaluation process defined in [14]. We extract 32 frames at equal intervals to obtain 32 classification scores. Each evaluation score represents a real number between 0 and 1 for real and fake videos, respectively. The video score is then computed as the average of all the individual scores. FTCN, on the other hand, extracts a sequence of N consecutive frames from the video. To maintain consistency with the evaluation of CADDM and RECCE, we set $N = 32$. In our study, we define the positive class as ‘fake videos’ and the negative class as ‘genuine videos’.

Implementation details: Our implementations of CADDM¹ RECCE² and FTCN³ are based on publicly available GitHub projects. All three deepfake detector models are trained on FaceForensics++ [77] and use a backbone trained on ImageNet to extract features from images. FF++

¹<https://github.com/megvii-research/CADDM>

²<https://github.com/VISION-SJTU/RECCE>

³<https://github.com/yinglinzheng/FTCN>

Table 5.2: Type of data from FF++ seen by each deepfake detectors

Model	Compression	Seen Face Manipulation	Training strategy
CADDM [14]	Raw	DF, F2F, FSh, FS, NT	supervised
RECCE [93]	c23	DF, F2F, FSh, FS, NT	semi-supervised
FTCN [157]	c23	DF, F2F, FS, NT	supervised

contains respectively 5000 and 1000 fake and real videos divided into three subsets: train, val, and test. Five manipulation techniques were used to generate the fake videos. They are either face reenactment (Face2Face: F2F, NeuralTexture: NT) or FaceSwap (Deepfake: DF, FaceSwap: FS, FaceShifter: FSh) based methods. All the 6000 videos exist in 3 versions: *raw*, *High-Quality* (*c23*), and *Low Quality* (*c40*). Table 5.2 gives a summary of the specific data seen by each model during their training on FF++. For more information about the training of the three models, please refer to their corresponding publications.

5.3.3 Experimental Results

In Table 5.3 and Figure 5.3, we present various results of our experiments. From Table 5.3, we can observe that all detectors suffer a drop of approximately 15% in AUC when tested with beautification filters. In Figure 5.3 (a), we see the impact of the beautification process on the FNR. Specifically, beautified videos reduce the FNR for CADDM and FTCN. However, the False Negative Rate for RECCE is higher for beautified videos. This presents a significant issue, as fake videos may appear authentic due to the simple application of a beautification filter. In contrast to CADDM and FTCN, RECCE did not encounter any fake videos during the training of its reconstruction network as mentioned in Section 5.3.1. Thus, RECCE did not learn any specific features associated with face manipulation. Even if beautification introduces minor artifacts, it removes some of the manipulation introduced by deepfakes. However, supervised trained models such as CADDM and FTCN can detect these minor artifacts.

To better understand the behavior of deepfake detectors on beautified videos, we analyzed the histogram of the classification scores before and after beautification. In Figure 5.4, we illustrate the difference in the distribution of the classification scores of the deepfake detectors on Celeb-DF-B for beautified and non-beautified videos. A score of 0 is the lowest probability that a video is fake according to a deepfake detector while a score of 1 represents the highest probability. For CADDM and FTCN, we can observe higher confidence scores for the beautified videos, indicating they are more likely to be detected as fake. On average, all the confidence scores of the videos are shifted by +0.1

Table 5.3: AUC score of each detector w/o and w/ beautification on Celeb-DF-B

Model	AUC (\uparrow)	
	w/o beautification	w/ beautification
CADDM [14]	0.91	0.76 (\downarrow 0.15)
RECCE [93]	0.81	0.66 (\downarrow 0.15)
FTCN [157]	0.80	0.64 (\downarrow 0.16)

and +0.3, respectively, for CADDM and FTCN after beautification. This behavior was expected since beautification may present manipulation clues. However, the behavior is slightly different for RECCE. On average, beautified videos appear more authentic than the original ones, with an average score shift of -0.07. In summary, for RECCE, beautified videos tend to appear more real than their non-beautified counterparts.

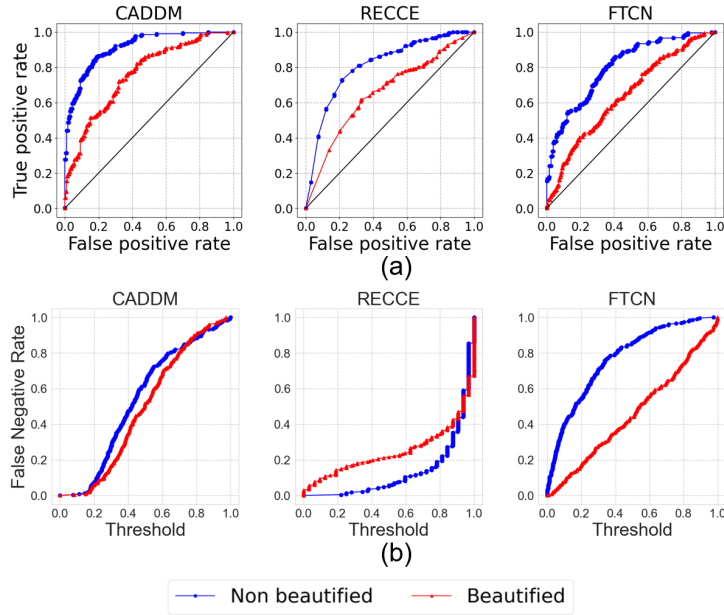


Figure 5.3: Result of the evaluation on Celeb-DF-B with the the 3 detectors. a) The video-level AUC of the ROC curve and b) The False Negative Rate for different classification score thresholds

5.4 Subjective Evaluation

In this section, we present the subjective evaluation conducted to assess the impact of applying a beautification filter to both deepfake and real videos.

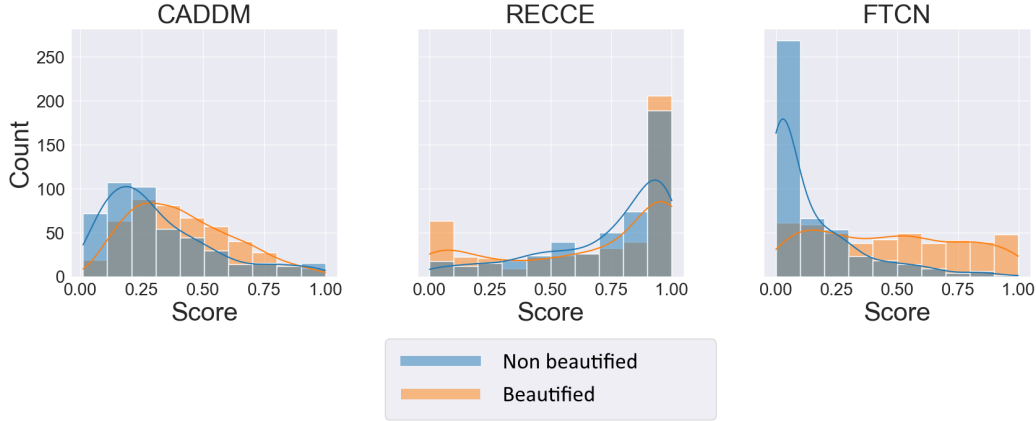


Figure 5.4: The histograms of the classification score for each deepfake detector on real and deepfake videos. In blue (resp. orange) the non-beautified (resp. beautified) subset of Celeb-DF-B. CADDM and FTCN tend to see beautification as additional face manipulation (histogram shifted to the right) whereas RECCE finds fake videos more realistic after the beautification process (histogram shifted to the left)

We performed a subjective evaluation of deepfake videos, using a web-based framework for crowdsourcing experiments. The primary objective of this subjective test was to investigate whether the utilization of such filters presents challenges for human observers when distinguishing between the authenticity of deepfake videos and real videos. To achieve this goal, we selected a total of 112 videos (56 real and 56 deepfakes) from the Celeb-DF-B database. The selection process involved the following steps:

Fake Video Selection: We randomly picked 7 videos for each type of beautification filter from the fake video category, resulting in 28 videos. These same videos were included without the filter in the subjective test dataset.

Real Video Selection: Subsequently, we chose 7 videos for each filter type from the real video dataset, and once again, we incorporated these same videos without filters into the subjective test dataset.

Test protocol: To establish a consistent benchmark for comparison with typical deepfake detection algorithms, we presented human subjects with cropped face regions. Furthermore, we extended the boundary by an additional 100 pixels into the background to assess the algorithms' ability to handle background information.

Before the evaluation, participants received comprehensive explanations of the test procedures and completed practice tests to ensure their understanding. To optimize efficiency and prevent fatigue during the evaluation, we divided the test dataset randomly into three batches. This approach allowed participants

to complete each test in separate sessions, with breaks in between. On average, each test batch lasted approximately 15 minutes, consistent with the standard recommendations [162]. The evaluation involved 21 participants with diverse backgrounds. Each video was shown to the participants three times consecutively. After viewing each video, following a procedure similar to that of Korshunov et al. [163], participants were asked, "Is the person's face in the video real or fake?". Then, they were then asked to identify the specific features or characteristics that influenced their judgment regarding the video's authenticity. The available feature options included: 1. Face contour, 2. Shadow inconsistency, 3. Inconsistency between eyes, 4. Eye blinking, 5. Mouth, 6. Teeth, 7. Lip motion, 8. Head motion, 9. Face/body mismatch, 10. Contextual mismatch, 11. Skin texture, and 12. Video quality.

5.4.1 Subjective Evaluation Results

Table 5.4 displays the results of the subjective assessment outlined in Section 5.4. The data within the table offers valuable insights into human performance in discerning deepfake videos from authentic ones, explaining the influence of beautification filters on human accuracy. The results suggest that human accuracy for non-beautified videos is higher (69%) than for beautified videos (66%), implying that human judgments are more effective at distinguishing between real and fake videos when no beautification is applied. Furthermore,

Table 5.4: Subjective evaluation results on Celeb-DF-B dataset for Beautified and Non-beautified videos

Metric	Non-beautified	Beautified
Accuracy	0.69	0.66
Recall	0.70	0.76

our analysis uncovers a significant contrast in recall rates between beautified (76%) and non-beautified (70%) videos. Recall, also known as sensitivity or true positive rate, measures the ability of a classifier to correctly identify positive instances among all actual positive instances. In the context of deepfake detection, a higher recall implies that the deepfake detection model or human evaluators are better at spotting deepfakes when the videos are beautified.

The increased recall rate in our study implies that evaluators are slightly better at identifying deepfakes when beautification filters are present. However, the observed accuracy rates suggest that while human subjects improve in detecting deepfakes with applied filters, they also tend to misclassify more genuine videos as fake in this scenario. This underscores the impact of

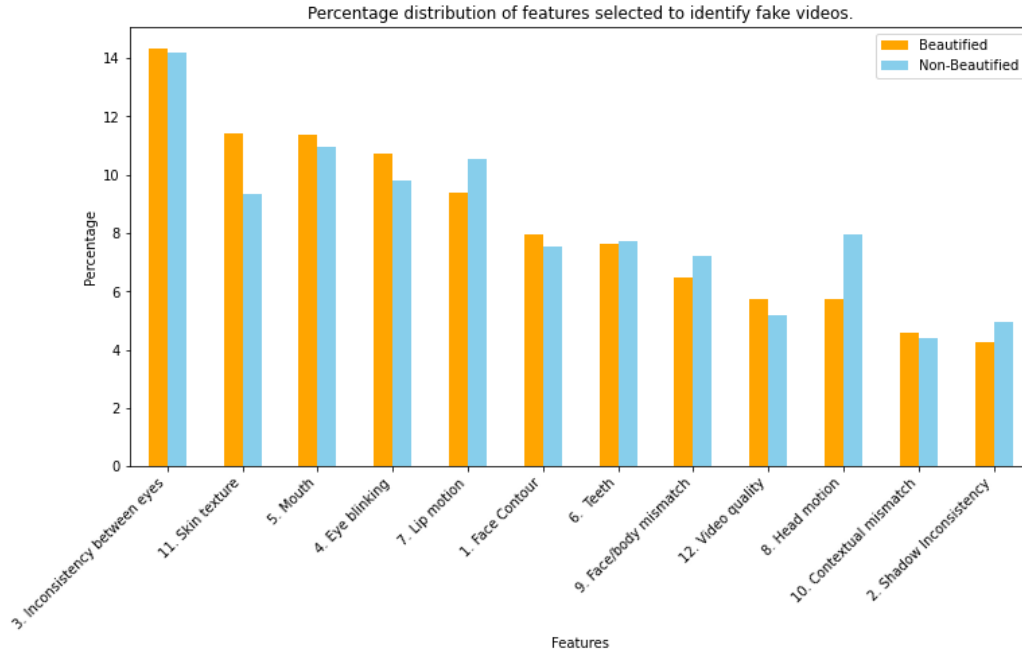


Figure 5.5: Each user’s accuracy before and after applying filters.

beautification filters on human detection capabilities: they not only aid in recognizing deepfakes but may also lead to a higher rate of false positives, where non-deepfake videos are mistakenly identified as deepfakes.

In the subjective evaluation, participants were also tasked with identifying the specific features or characteristics that played a role in shaping their judgment regarding the video’s authenticity. Among the provided feature options, the inconsistency between eyes stood out as the most frequently noted feature in both beautified and non-beautified videos. An interesting finding is that many participants highlighted modifications in skin texture as a factor influencing their categorization of videos as fake, with a notably higher percentage observed in beautified videos, as depicted in Figure 5.5.

5.5 Conclusion and Discussion

In this work, we investigate the ongoing trend of digital face beautification through social media filters and its implications for deepfake detection. The application of filters to facial multimedia is a user-friendly practice, as it does not demand any prior expertise, unlike other image editing techniques. This accessibility makes filters highly approachable for the average social media user. This study extends beyond AI-based detection, assessing three state-of-

the-art deepfake detectors and the impact that the use of filters on deepfake videos has on human detection via a subjective evaluation. Experiments are conducted in our proposed Celeb-DF-B database showing how the application of filters significantly shifts the scores of the assessed deepfake detectors and changes the perceived information for human subjects.

Our findings reveal that, depending on the classifier used, even easy-to-use social media filters can significantly increase the likelihood of a deepfake video being wrongly classified as authentic. This not only challenges the robustness of current deepfake detection methods but also raises important questions about the reliability of these systems in real-world scenarios, where such filters are commonly used. We highlighted that deepfake detection is not just a matter of identifying sophisticated manipulations but also understanding how common alterations can impact these systems. Future challenges include mitigating the effects of beautification filters. In scenarios requiring access to secure locations or sensitive information, such as government facilities, financial institutions, or military installations, it becomes imperative to minimize the risk of an impersonation attack. Retraining deep learning-based detectors with beautified data might not guarantee a solution, as filters are being created daily, making generalization difficult. Given the substantial impact of beautification filters, the use of a dedicated filter detection method is strongly advisable.

Chapter 6

RAW Data: A Key Component for Effective Deepfake Detection

6.1 Introduction

Image processing operations, such as compression or beautification filters, can obscure forgery indicators, leading to inaccurate deepfake detection. Detection models often struggle to differentiate between real and fake images because the current definition of real images encompasses both raw content captured by camera sensors and content processed through various stages of image and video processing, including linear and nonlinear adjustments. Consequently, images are considered real even after undergoing multiple processing operations such as denoising, compression, deblurring, and white balance adjustments. This work aims to redefine the boundary between real and fake images by narrowing the definition of authentic samples to a stage closer to the radiance of the scene as captured by the sensor, prior to any transformations by an Image Signal Processor (ISP).

The ISP is designed to convert raw sensor data from Bayer Color Filter Arrays (CFA) into visually appealing RGB output images. This transformation process, begins as the camera lens focuses light onto the CFA sensor, producing a digital representation of the scene in raw pixel values. These raw images undergo several ISP stages, including white balance, noise removal, deblurring, and tone mapping, to produce the final RGB output. The primary goal of each stage in an ISP is to produce images that are aesthetically pleasing to the human eye, this often involves nonlinear modifications that enhance certain visual aspects at the expense of the original data's fidelity. These enhancements are typically tuned based on subjective human ratings, aiming to maximize visual appeal rather than preserving the true characteristics

of the imaging scene. While this approach is beneficial for applications like photography, it introduces significant challenges in the context of deepfake detection, as outlined below.

- **Advantages of raw data for deepfake detection:** Raw data, as a linear representation of scene radiance, captures the original light distribution without alteration. This unaltered preservation makes raw data an ideal starting point for deepfake detection, as it retains essential details and the distribution of facial features crucial for accurately distinguishing authentic images from forgeries.
- **Effects of different ISP pipelines on detection accuracy:** Transformations in the ISP pipeline introduce nonlinear alterations to raw input images, enhancing visual appeal but modifying the geometric and color distribution of facial features. Each device employs a unique ISP pipeline configuration with distinct enhancement blocks, further complicating detection by obscuring subtle cues needed for accurate identification of deepfakes. Moreover, the proprietary nature and hardware integration of these pipelines vary significantly across different manufacturers. This variability poses a challenge for detection models, as they must adapt to novel and previously unseen ISP configurations, which can obscure the essential cues for effectively identifying and classifying real images.

Our research makes significant contributions to the field of deepfake detection through the following key innovations:

1. **Utilization of raw data for enhanced model performance:** We propose a novel pipeline that utilizes raw data as input for deepfake detection. This approach allows for easier learning and better generalization on authentic images. Incorporating raw data as input necessitates a modification in the existing detection formula (2), now represented as:

$$S_o = \phi_D(I_{\text{RAW}}), \quad (6.1)$$

where ϕ_D denotes the deepfake detection function, and I_{RAW} signifies the raw image data.

2. **Proposal for training on raw data using an RGB-to-raw auxiliary model in response to dataset limitation:** Given the scarcity of large-scale datasets for training on raw images, we propose a methodological approach, detailed in Section 6.2, to train our model on raw

images for deepfake detection. This approach addresses the limitations posed by the absence of appropriate resources and establishes a foundation for future research. Our primary objective is to focus on detection in the raw domain, assuming that attacks occur in this worst-case scenario. However, if the model encounters processed RGB images, and it is determined that the examination should be conducted in the RGB domain, our proposed methodology includes an auxiliary model to convert these images to raw format for examination.

6.2 Proposed Method

We optimize a binary classifier using cross-entropy loss, L , to perform deepfake detection defined as:

$$L = -\frac{1}{N} \sum_{i=0}^{N-1} \{t_i \log F(x_i) + (1 - t_i) \log(1 - F(x_i))\}, \quad (6.2)$$

where $F(x)$ denotes the probability of x being classified as "fake", and t_i represents the binary label associated with the input image, indicating whether it is fake (1) or real (0). To enhance the generalization and robustness of our detection algorithms, we utilize synthetic raw samples that embody common forgery traces. These include blending boundaries, source feature inconsistencies, and statistical anomalies in the frequency domain. All of these present significant challenges for detection. The samples undergo a conversion process from raw to RGB using ϕ_{ISP} , are manipulated within the RGB domain, and subsequently reconverted to raw using $\phi_{\text{ISP_inv}}$. This novel synthetic training data, called Raw Self-Blended Images ($I_{\text{RAW_SB}}$), is designed to improve model robustness.

Our key observation is that as deepfake generation techniques evolve, synthesized images will increasingly resemble pristine target images in terms of facial landmarks and pixel statistics. Based on this insight, we have utilized a synthetic data generation pipeline similar to [85] that creates fake images by blending pseudo source and target images derived from a single image (I_{base}). This presents models with a more complex and generalized task for face forgery detection. To produce ($I_{\text{RAW_SB}}$), we develop a Source-Target Generator (STG) and a Mask Generator (MG). The STG starts by generating pairs of pseudo source and target images from single pristine images using straightforward image processing techniques, and the MG creates various blending masks based on the facial landmarks of the input images.

As illustrated in Figure 6.1, the generation of raw synthesized training data begins with converting a raw image (I_{RAW}) to RGB format (I_{base}) using

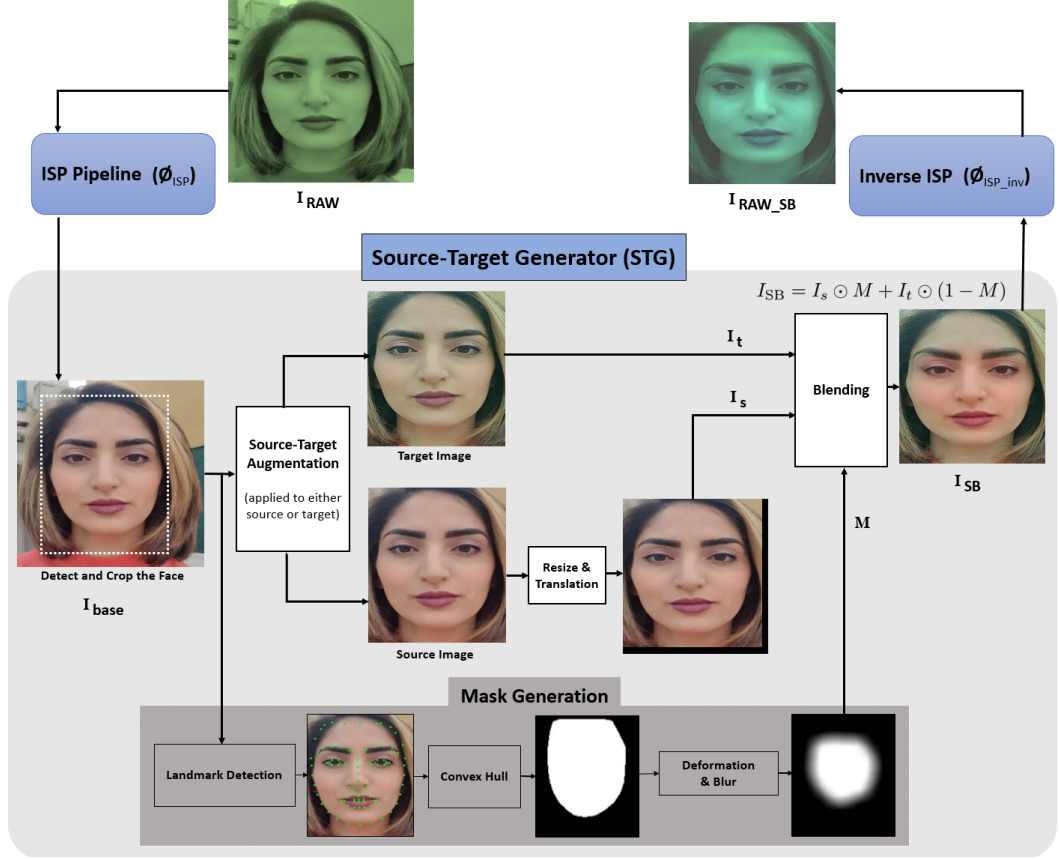


Figure 6.1: Overview of generating a RAW Self-Blended Image (I_{RAW_SB}). A base image I_{base} is fed into the Source-Target Generator (STG) and the Mask Generator (MG). The STG produces pseudo source and target images from the base image using various image augmentations, while the MG creates a blurring mask from facial landmarks and deforms it to enhance mask diversity. The source and target images are then blended with the mask and input into the Inverse ISP pipeline to reconstruct the raw format of the RGB input image.

ϕ_{ISP} . The bounding box and facial landmarks of the face in the input image are then detected, leading to the cropping of the face area. This cropped image serves as both the source and target image for further processing. To introduce inconsistencies between the source and target images, a series of augmentations, such as brightness and contrast adjustments, are applied to either source or target image. The source image is also resized and translated to replicate blurring boundaries and landmark mismatches. Additionally, the Mask Generator (MG) produces a grayscale mask specifying the manipulated region.

The initial mask is defined as the convex hull of the facial landmarks in the input image. Given that face manipulation techniques may target various areas of the face, resulting in diverse shapes of manipulated regions (e.g., affecting only the mouth, eye, or entire face region), the mask is altered to accommodate these variations using landmark augmentation [84]. The mask is then smoothed using a Gaussian filter. Finally, the mask is applied to blend the source and target images according to Equation 6.3, resulting in the creation of the self-blended image (I_{SB}).

$$I_{SB} = I_s \odot M + I_t \odot (1 - M) \quad (6.3)$$

Where, I_s , I_t , and M represent the source image, target image, and the generated mask, respectively. After preparing the self-blended image, we use an inverse ISP model, ϕ_{ISP_inv} to transform the RGB image (I_{SB}) back into the raw image domain (I_{RAW_SB}). Finally, this synthesized raw data, along with the original raw image (I_{RAW}), is input to the detection algorithm for training.

6.3 Experiment

Our proposed methodology employs raw domain data as the primary input for deepfake detection. Recognizing the scarcity of large-scale datasets specifically designed for training on raw domain images, we have developed a methodological approach appropriate to train deep neural networks directly on raw images. To implement this, our method reverses the data processing flow during the training phase. Instead of converting raw data into RGB images as depicted in Figure 6.1, we initiate with RGB images from a publicly available dataset, labeled as I_{base} . These images are then transformed into raw format, labeled as I_{RAW} . The generated I_{RAW} and synthetically created fake images in raw format (I_{RAW_SB}) are utilized for training. It is noteworthy that during inference, our model processes raw data directly. In cases where RGB images are received, they are converted to raw as a preprocessing step before being input into our model. This reversal strategy provides multiple strategic benefits:

1. It allows us to leverage the existing wealth of large-scale RGB data while still capitalizing on the advantages of training in the raw domain.
2. Additionally, by training our model directly on raw data, we are enabled to benchmark our methodologies against existing RGB-based state-of-the-art models. This setup not only offers a direct comparison of

the effectiveness of our raw-based approach but also underscores its adaptability and robustness, particularly in scenarios where RGB images are received.

6.3.1 Experimental Setup

Training Dataset: We trained our model on the widely used benchmark dataset, Face-Forensics (FF++) [10], which contains 1,000 real videos and 4,000 fake videos forged by four manipulation methods, i.e., Deepfakes (DF) [164], Face2Face (F2F) [165], FaceSwap (FS) [166], and NeuralTextures (NT) [167]. For the purpose of training and validation, we exclusively utilized the real videos from the FF++ dataset, comprising 720 videos for training and 140 videos for validation. Both the training and validation sets included the raw real videos as well as their corresponding synthesized self-blended images.

Evaluation Dataset:

To evaluate the performance of our approach, we utilized the test set from the FaceForensics++ (FF++) dataset, which includes both authentic and manipulated videos. For cross-dataset evaluation, we employed three recent deepfake datasets to assess the generalizability and robustness of our model across various sources and manipulation techniques:

- Celeb-DF v2 (CDF) [78] applies more advanced deepfake techniques to celebrity videos sourced from YouTube, providing a realistic benchmark that reflects higher quality deepfake generation.
- Deep-Fake Detection (DFD) [168] offers thousands of deepfake videos created with the consent of the actors, aiming to represent a diverse set of facial expressions and lighting conditions.
- DeepFake Detection Challenge public test set (DFDC) [79] include videos with various disturbances such as compression, downsampling, and noise, presenting challenges typical of real-world scenarios.

Each dataset offers unique challenges and helps in validating the effectiveness and adaptability of our detection system across different domains and attack vectors.

Data Preprocess: We utilize Dlib and RetinaFace [169] for extracting facial landmarks and bounding boxes from individual video frames, respectively. Within Dlib, we employ an 81-point facial landmarks shape predictor [170]. The dimensions of the face, computed from the bounding box, determines the cropping of the facial region. Facial landmarks are required during the training phase for generating self-blended images and they are dispensable

during inference.

Data Augmentation: During the generation of synthesized fake raw images, we employ various frequency and color transformations using the Albumenations toolbox [171] to introduce variations between the source and target images. In the Source-Target Generator (STG) phase, augmentations such as image compression, RGB shift, Hue Saturation Value (HSV) adjustments, and random brightness and contrast transformations are applied either to the source or the target image. These modifications enhance the robustness of our model by simulating a range of potential distortions encountered in practical scenarios.

Raw Data Generation: In the absence of authentic raw image data, and to facilitate the generation of synthetic raw domain images via our proposed pipeline, we employ the state-of-the-art inverse ISP model, known as MiAlgo [110]. This method is trained to recover raw data from the RGB Huawei P20 model; however, it can effectively generalize to noisy and unseen similar sensors. Moreover, the method doesn't require any metadata or specific camera parameters (e.g., correction matrices, digital gains), which are typically inaccessible. MiAlgo utilizes an end-to-end encoder-decoder UNet-like structure, incorporating key components such as the residual group [109] and the enhanced block [104]. The algorithm takes a full-resolution RGB image and converts the input image to a raw RGGB pattern. Consequently, we perform demosaicing on the RGGB image and then feed this raw data into a binary classifier for deepfake detection.

Evaluation Metrics: We utilize the video-level area under the receiver operating characteristic curve (AUC) for comparison with previous research. Normally, predictions at the frame level are averaged across video frames.

Training details: We utilize the state-of-the-art convolutional network architecture EfficientNet-b4 (EFNB4) [174], pretrained on ImageNet, as our classifier and train it for 100 epochs using the SAM optimizer [175]. We set the batch size to 32 and the learning rate to 0.001. During training, we sample only eight frames per video. Each batch includes both real images and their self-blended counterparts.

6.3.2 Experimental Results and Analysis

The use of raw data as an input for deepfake detection represents a novel approach within this field, where no directly comparable methods currently exist. Consequently, we assess the performance of our proposed method by comparing it with established methods that operate within the RGB domain. This comparison not only facilitates a direct evaluation of the

Table 6.1: Cross-Dataset Evaluation on CDF, DFD, and DFDC Datasets Using raw data for deepfake detection is a novel approach with no direct comparisons in the field. We compare our method’s performance against established RGB domain techniques. Our methodology includes an inverse ISP model to convert RGB images to raw format for analysis. The model is trained exclusively on the high-quality FF++ dataset using only real data. Results from previous methods are cited from their original papers. Bold values indicate the best performance, while underlined values denote the second-best performance.

Method	Input Type	Training set		Test Set AUC (%)		
		Real	Fake	CDF	DFD	DFDC
DSP-FWA [83]	Frame	✓	✓	69.30	-	-
Face X-ray + BI [84]	Frame	✓		-	93.47	-
Face X-ray + BI [84]	Frame	✓	✓	-	95.40	-
LRL [82]	Frame	✓	✓	78.26	89.24	-
FRDM [15]	Frame	✓	✓	79.4	91.9	-
PCL + I2G [16]	Frame	✓		90.03	99.07	67.52
EFNB4 + SBIs [85]	Frame	✓		<u>93.18</u>	97.56	72.42
Two-branch [172]	Video	✓	✓	76.65	-	-
DAM [173]	Video	✓	✓	75.3	-	-
LipForensics [88]	Video	✓	✓	82.4	-	-
FTCN [157]	Video	✓	✓	86.9	94.40	<u>71.00</u>
EFNB4 + I_{RAW_SB} (ours)	Frame	✓		94.23	<u>98.46</u>	69.42

effectiveness of our raw-based approach but also highlights its versatility and robustness, especially in scenarios involving the receipt of RGB images. For this comparative analysis, our methodology is benchmarked using the public dataset described in Section 6.3.1, against other RGB-domain deepfake detection techniques [15], [16], [82]–[85], [88], [157], [172], [173].

Quantitative Results:

Table 6.1 presents the Area Under the Curve (AUC) evaluation metrics of our method compared with existing methods to benchmark against prior works. Our method, EFNB4 + I_{RAW_SB} , demonstrated state-of-the-art performance on the CDF dataset and showed competitive results across other deepfake detection datasets, as detailed below:

On the CDF dataset, our approach achieved an impressive AUC of 94.23%, surpassing EFNB4 + SBIs, which obtained 93.18%. For the DFD dataset, PCL + I2G attained the highest AUC of 99.07%, while our method followed closely with an AUC of 98.46%. In the DFDC dataset, EFNB4 + SBIs led with an AUC of 72.42%, and FTCN, which utilized both fake and real sets during training, achieved 71.00%. Our method secured an AUC of 69.42%,

Table 6.2: Cross-manipulation evaluation on FF++.

Method	Test Set AUC (%)				
	DF	F2F	FS	NT	FF++
Face X-ray + BI [84]	99.17	98.57	98.21	98.13	98.52
PCL + I2G [16]	100	98.97	99.86	97.63	99.11
EFNB4 + SBIs [85]	99.99	99.88	99.91	99.79	99.64
EFNB4 + $I_{\text{RAW_SB}}$	99.92	99.16	99.46	99.72	99.56

ranking as the third-best performer on this dataset. Table 6.2 presents our cross-manipulation evaluation results on FF++. We can see that even if our method is not trained on specific artifacts within the FF++ dataset, still our method performs well in recognizing different deepfake artifacts generated by different methods or nearly matches existing methods across four manipulations (99.92% on DF, 99.16% on F2F, 99.46% on FS, and 99.72% on NT) and achieves 99.56% the overall performance on the entire FF++ dataset.

Qualitative Analysis of Inverse ISP:

Figure 6.2 presents examples of I_{base} , I_{RAW} , I_{SB} , $I_{\text{RAW_SB}}$ and blending mask. The I_{base} samples are sourced from FF++ dataset [10]. I_{RAW} and $I_{\text{RAW_SB}}$ are transformed from I_{base} and I_{SB} respectively, using inverse ISP model (see Figure 6.1). Although the inverse ISP model was not specifically trained for face images, it successfully reconstructs the raw domain data with high fidelity. Furthermore, we observe that the artifacts in synthetic images (I_{SB}), generated from I_{base} , are distinctly visible and can also be recognized in $I_{\text{RAW_SB}}$, indicating effective transformation and consistency across image representations.

6.3.3 Limitations and Future Work

While our approach shows promise, particularly in cross-dataset evaluations, several limitations merit attention and set the direction for future research. Firstly, our methodology primarily utilizes raw domain data for deepfake detection to mitigate the impact of nonlinear transformations caused by ISP processes. However, the scarcity of large-scale facial datasets specifically tailored for training on raw domain images presents a significant challenge. Currently, we rely on an inverse ISP model to generate raw data during the training phase. Future work could explore the development of large-scale datasets directly from sensor outputs, which would likely enhance the

effectiveness and authenticity of the training data. Additionally, the inverse ISP currently employed is trained predominantly on non-facial images and from a limited set of sensor samples. Expanding the diversity of sensor types in the training set and tailoring the model to better handle facial images could substantially improve its utility and accuracy. One potential approach involves utilizing a camera source identification algorithm [176] to detect the sensor model and applying the appropriate inverse ISP method to reconstruct the raw data.



Figure 6.2: Example images of I_{base} , I_{RAW} , I_{SB} , and $I_{\text{RAW_SB}}$. The I_{base} samples are sourced from FF++ dataset. I_{RAW} and $I_{\text{RAW_SB}}$ are transformed from I_{base} and I_{SB} respectively, using inverse ISP model.

6.4 Conclusion

This study introduces a novel pipeline that leverages raw domain data as input to enhance deepfake detection. Our approach addresses the limitations of current models, which often struggle with genuine images modified by ISP pipelines. By focusing on raw data, we simplify the detection process and improve generalization, effectively bounding the distribution of real images. This method allows for easier learning and better generalization on authentic images. Given the scarcity of large-scale datasets designed for training on raw images, we propose a methodological approach to train our model on

raw image data. Our primary objective is to focus on detection in the raw domain, however, if the model encounters processed RGB images, our proposed methodology includes an auxiliary model to convert these images to raw format for examination. Our method demonstrated state-of-the-art performance on the CDF dataset and showed competitive results across other RGB domain deepfake detection datasets. Future work could explore the development of large-scale datasets directly from sensor outputs, which would likely enhance the effectiveness and authenticity of the training data. This approach holds promise for advancing deepfake detection technologies, contributing to more reliable defenses against the malicious use of digital imagery in various domains. Specially, it enhances the security of devices with authentication systems by allowing manufacturers to use raw data for identification, making impersonation attacks more challenging.

Chapter 7

Towards Secure Authentication: Detecting Replay Attacks via Compression Artifacts

7.1 Introduction

Advancements in biometric techniques and the increasing demand for seamless identity verification have led to the widespread adoption of remote face verification systems. As these systems become more popular, they also become increasingly attractive targets for malicious actors attempting to spoof the system and impersonate legitimate users. These attacks can occur on both the client side, where users interact with the system, and the server side, where data is processed and validated. In this work, we assume that both the application running on the device and the server handling the data are securely protected. Our focus is on addressing client-side vulnerabilities, particularly injection attacks, where attackers bypass the sensor and directly inject malicious digital content into the data stream [7]. Figure 7.1 provides an illustration of injection attacks, which can be broadly categorized into two types: deepfake attacks and digital replay attacks.

1. **Deepfake Attacks:** In this type of attack, the adversary begins by obtaining images of the victim. Using a deepfake algorithm, the attacker generates a manipulated video in real time, replicating the facial expressions and head movements required by an active liveness detection system [25]. This deepfake video is then streamed to the authentication system via virtual camera software (e.g., OBS), which acts as an intermediary to bypass the physical camera sensor.

2. **Digital Replay Attacks:** In these attacks, adversaries leverage authentic video footage of the victim, often sourced from publicly available platforms such as social media, and inject it into the system using virtual camera software. Since the video is genuine and lacks manipulation artifacts, it poses significant challenges for the system to distinguish between a live user and a replayed video.

To mitigate deepfake attacks researchers have proposed various detection methods, ranging from early handcrafted feature-based techniques to modern deep learning-based approaches [10], [12], [14]. These detection methods aim to identify subtle inconsistencies or artifacts introduced during video synthesis and they often rely on supervised learning, where models are trained to recognize known deepfake artifacts.

To address digital replay attacks, remote face authentication systems commonly implement active liveness detection, which requires users to perform specific actions such as nodding, blinking, or opening their mouth to verify their physical presence and confirm they are alive. While this approach is effective in countering replayed videos, it is not always user-friendly and may

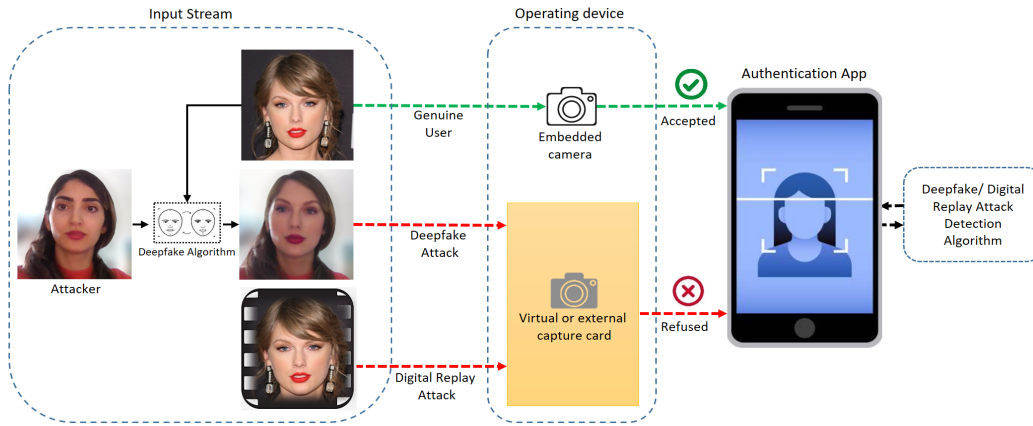


Figure 7.1: Illustration of various input streams to a remote face authentication service. The input can originate from different scenarios. In the first scenario, the face of a genuine user is provided to the service, granting access to the application upon successful authentication. In the second scenario, a deepfake injection attack is performed. Here, the attacker generates a real-time video mimicking the victim's expressions and head movements using a single image. This video is streamed via virtual camera software to imitate a legitimate webcam feed, deceiving the authentication system. In the third scenario, the attacker uses either a single image or a pre-recorded video of the victim. The virtual camera streams a genuine video of the victim that lacks visible artifacts. Our main goal is to exploit compression artifacts for detecting digital replay attack.

be perceived as inconvenient, potentially impacting the overall usability of the system. Another method for detecting replay injection attacks involves analyzing the metadata of the user’s device and camera, flagging suspicious camera names as potential indicators of an injection. However, this method is highly vulnerable, as attackers can easily manipulate the camera metadata to bypass detection algorithms. This type of attack is the focus of our investigation in this work. To the best of our knowledge, no machine learning-based approaches currently exist to mitigate digital replay attacks, as the injected video is authentic and lacks detectable artifacts.

Content captured by cameras in practical scenarios often undergoes various digital image and video processing operations, including post-processing techniques such as stylization filters and beautification [154], before being disseminated. Recent studies have systematically evaluated the adverse effects of these operations on the performance of biometric algorithms [22], [25]. Hence in this work we investigate whether providing uncompressed video access to face anti-spoofing service providers can improve the detection of injected versus authentic video streams. Building on this, we propose bypassing the compression step and directly capturing uncompressed image data from the user’s device during authentication, rather than relying on video content that has passed through the complete post-processing stages of the Image Signal Processing (ISP) pipeline [102]. This strategy enables the detection system to better differentiate between authentic and injected video streams, as injected videos are often sourced from the internet and are typically compressed using widely-used algorithms. By ensuring that the user’s camera captures uncompressed images during authentication, the detection model can focus on identifying compression artifacts. The presence of such artifacts strongly indicates that the image frame has been injected, thereby significantly improving the system’s ability to detect and prevent injection attacks.

For this purpose, we utilize raw video datasets in their original, uncompressed form to simulate real-world scenarios. To generate corresponding compressed versions, widely-used video compression algorithms are applied. A classifier is then designed and trained on both compressed and uncompressed frames. Through this process, we aim to evaluate the classifier’s ability to detect compression artifacts and assess its effectiveness in accurately distinguishing between compressed and uncompressed frames. This analysis provides valuable insights into the role of compression artifacts as distinguishing features in video authentication tasks.

7.2 Proposed Method

This work investigates whether providing uncompressed video access to face anti-spoofing service providers can improve the detection of injected versus authentic video streams. We hypothesize that uncompressed video frames from a user’s device would lack compression artifacts, while injected videos, such as deepfakes or replays, would show detectable artifacts due to compression during their creation or transmission. The aim of this study is to analyze these compression artifacts for effective differentiation.

To achieve this, we propose a machine learning-based model trained on both compressed and uncompressed video frames. Videos are compressed using four widely used algorithms—H.264, H.265, VP8, and VP9—chosen for their popularity. Both compressed and uncompressed video versions are converted into individual frames.

For training, random patches of size 224×224 are extracted from the frames, ensuring balanced representation from both compressed and uncompressed frames. By focusing on image patches rather than full frames, the model captures localized compression artifacts, which are key for accurate detection. These patches are then used to train a binary classifier. The model is optimized using cross-entropy loss, L , defined as:

$$L = -\frac{1}{N} \sum_{i=0}^{N-1} \{t_i \log F(x_i) + (1 - t_i) \log(1 - F(x_i))\}$$

where $F(x)$ represents the probability of classifying a patch x as compressed, and t_i is the binary label associated with the input patch, where $t_i = 1$ for compressed and $t_i = 0$ for uncompressed.

7.3 Experiment

7.3.1 Experimental Setup

To train and test a classifier for compression detection, we utilize six well-established video datasets that are commonly employed in video quality assessment and coding algorithm evaluation.

- **Xiph.org Video Test Media Dataset** [177]: contains a diverse collection of video clips with varying resolutions (240 to 2160), frame rates (25–60 fps). A subset of 47 videos, featuring resolutions of CIF (352×288), HD (1280×720), and Full HD (1920×1080), is selected from this dataset.

Table 7.1: Performance metrics of the model on trained codecs (H.264, H.265, VP8, VP9) and an unseen codec (MPEG-4), demonstrating its detection accuracy and generalization capability.

Codec	AUC	Precisio	Recall	F1
H.264	0.979	0.868	0.970	0.916
H.265	0.986	0.870	0.986	0.924
VP8	0.993	0.871	0.998	0.930
VP9	0.988	0.869	0.982	0.922
Mpeg4	0.965	0.864	0.939	0.900

- **SJTU-4K Video Sequence Dataset** [178]: contains 15 4K (3840×2160) sequences captured with a Sony F65 camera at 30 fps. For our experiments, we utilize the 8-bit YUV 4:2:0 format videos.
- **SJTU-HDR Video Sequence Dataset** [179]: contains 16 High Dynamic Range video sequences, captured at 60 fps using Sony F65 and F55 cameras. The sequences, originally provided in 16-bit OpenEXR format. We convert videos to 8-bit YUV 4:2:0 for our experiments.
- **UVG-Dataset** [180]: comprises 16 4K video sequences captured at 50 or 120 fps in raw 8-bit and 10-bit YUV 4:2:0 formats. We utilize the 4:2:0 YUV format for this study.
- **USTC-TD Video Dataset** [181]: This dataset contains 10 video sequences, captured at 30 fps using Nikon D3200 and Nikon Z-fc cameras. The videos are provided in Full HD and were converted to the YUV 4:2:0 format using FFmpeg library [182].
- **MCL-JCV** [183]: Comprises 30 HD/Full HD uncompressed video sequences. Additionally, it includes encoded videos produced using the H.264/AVC codec, with their quality determined by the quantization parameter (QP), which varies from 1 to 51. For this study we utilize HD videos.

We use the videos from the first five datasets for training. The data is split into 70% for training, 20% for evaluation, and 10% for testing. To assess the model’s ability to generalize to datasets beyond the training set, we use the **MCL-JCV** [183] dataset as a cross-test set.

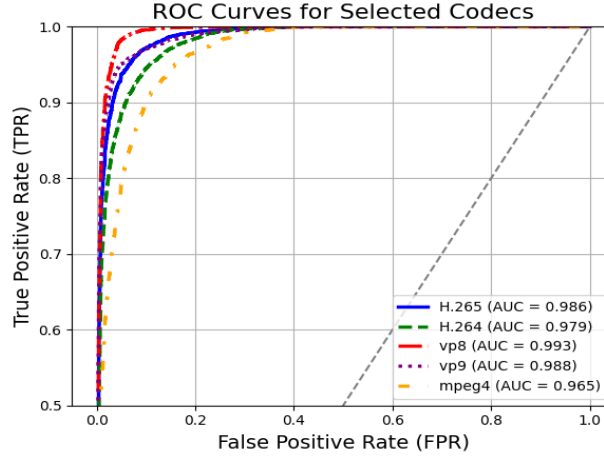


Figure 7.2: ROC curves for different codecs

We leverage both compressed and uncompressed video frames to train our classifier. Uncompressed frames are generated by converting raw videos into .png format without any compression. Compressed frames are obtained by applying compression algorithms (H.264, H.265, VP8, and VP9) to the raw videos, followed by saving the resulting frames in .png format. The default compression parameters of FFmpeg for each algorithm are used to simulate typical compression scenarios. We ensure an equal number of uncompressed and compressed frames for training. To introduce greater variability during training, we randomly crop image patches of size 224x224 and apply both horizontal and vertical flipping as augmentations. The ResNet-50 model architecture, pretrained on ImageNet, is applied and trained for 20 epochs using the Adam optimizer. A batch size of 128 is used, with an initial learning rate of 0.001.

7.3.2 Experimental Results

In our experiment, the MCL-JCV dataset is employed as a cross-test dataset to evaluate the model's performance on previously unseen data. Table 7.1 highlights the model's performance on the codecs it was trained on, using the default parameters of the FFmpeg library. Furthermore, the table evaluates the model's generalization capability by testing its performance on an unseen codec, specifically the MPEG-4 compression method.

The results in Table 7.1 and AUC curves in Figure 7.2 reveal outstanding performance for the H.265, VP8, and VP9 codecs, with AUC values near or equal to 0.99. Among these, the VP8 codec achieves the best results across all metrics, suggesting that its compression artifacts are the most

distinguishable by the detector. In contrast, the H.264 codec demonstrates slightly lower performance across all metrics, indicating that its compression artifacts are less prominent and harder for the model to detect. For the unseen MPEG-4 codec, the results show a decline in performance compared to the seen codecs. Nevertheless, the model maintains a reasonably high level of accuracy, showcasing its adaptability to compression methods it was not explicitly trained on.

Table 7.2 and Figure 7.3 showcase the model’s capacity to generalize to compressed frames across a range of quantization parameters, despite being trained with FFmpeg’s default quantization values. The H.264 codec is tested with Quantization Parameters (QP) ranging from 1 to 50. A QP value of 1 represents the highest image quality, while 50 corresponds to the lowest. The results in Table II reveal that at QP=1, the model struggles to distinguish between compressed and uncompressed frames, as the quality is nearly indistinguishable from uncompressed images. From QP=20 onward, the model’s performance improves significantly, with metrics approaching near-perfect values. This trend indicates that higher compression levels introduce more noticeable artifacts, making them easier for the model to detect.

To gain qualitative insights, Figure 7.4 visualizes the results of guided backpropagation [184], which highlights all contributing features that influence the prediction. Additionally, we use Grad-CAM [185] to visualize the regions where the model’s attention is concentrated, specifically for the compressed category. As seen in the visualizations, the model’s attention is more sparse when analyzing uncompressed frames, whereas on compressed frames (e.i. H.264, H.265, VP8, VP9 and Mpeg4) the attention is more concentrated and coarser, focusing on areas where compression artifacts are most prominent.

Based on the quantitative and qualitative results, we can conclude that all compression methods introduce artifacts that are distinguishable from uncompressed frames. Notably, even under moderate compression with a quantization parameter (QP) of 20—where image quality remains nearly flawless—the model reliably detects compression artifacts.

If the authentication service can access uncompressed frames directly from the device, it can focus exclusively on identifying compression artifacts. The presence of such artifacts would indicate that the frame is injected, simplifying the detection process. This approach eliminates the need for the algorithm to recognize specific artifacts left behind by various deepfake generators. Furthermore, in digital replay attacks, where virtual cameras often apply video compression, injected videos should be distinguishable from genuine videos due to the compression artifacts they inevitably contain. This distinction significantly facilitates the task of detecting injected content.

Table 7.2: Performance metrics of the model for H.264 compression across varying quantization parameters (QP).

Codec-QP	AUC	Precision	Recall	F1
H.264-Q01	0.498	0.492	0.143	0.221
H.264-Q20	0.948	0.859	0.903	0.881
H.264-Q30	0.984	0.869	0.981	0.922
H.264-Q45	0.994	0.871	0.997	0.930
H.264-Q50	0.995	0.871	0.999	0.931

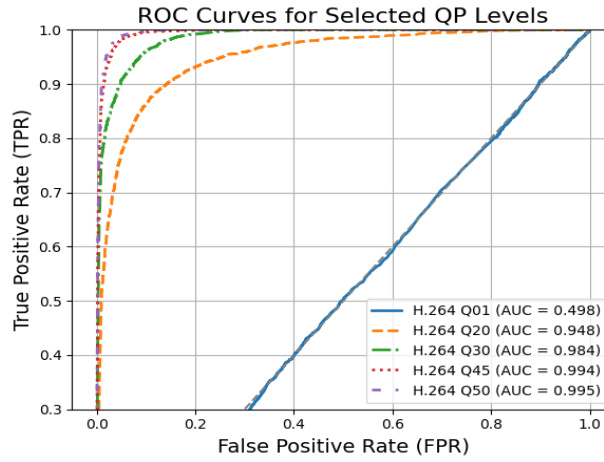


Figure 7.3: ROC curves for H.264 codec with different QP levels

7.4 Conclusion

This study introduces a novel approach for mitigating replay attacks by utilizing compression artifacts to differentiate between compressed and uncompressed video frames. These artifacts act as reliable indicators of injected content, thereby facilitating the detection process. In the context of digital replay attacks, where virtual cameras typically apply video compression, injected videos can be distinguished from genuine ones based on the inherent compression artifacts. By using raw video datasets and applying common video compression algorithms, a classifier was trained to differentiate between compressed and uncompressed frames. Experimental results demonstrate that the model effectively performs this distinction, highlighting the significance of compression artifacts in video authentication.

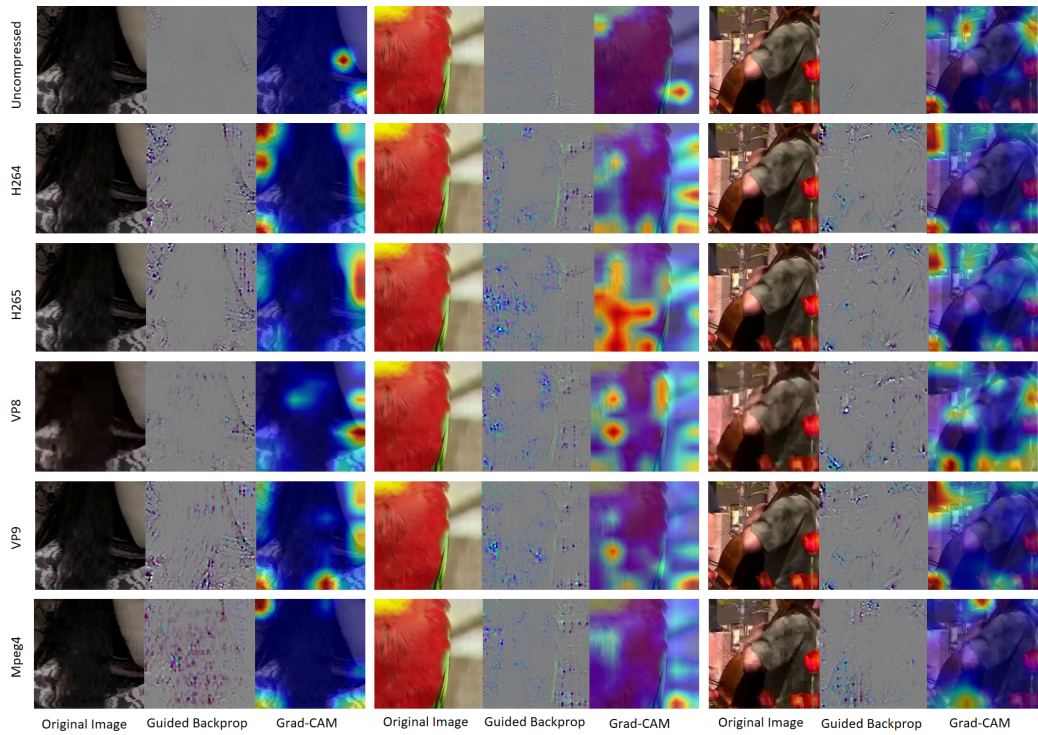


Figure 7.4: Guided Backpropagation and Grad-CAM visualizations for uncompressed and compressed video frames, highlighting the areas of support for the compressed category.

Chapter 8

Conclusion and Future Directions

8.1 Conclusion

Biometric face authentication leverages the unique biological characteristics of an individual’s face, eliminating the need to remember passwords or carry physical tokens. This method offers significant advantages, including enhanced security due to the difficulty of replicating biometric traits, improved user convenience, and a seamless authentication experience. As a result, face authentication is increasingly integrated into remote authentication services and portable devices, such as laptops and smartphones, to provide a secure and user-friendly solution for access control.

Traditional biometric threat models primarily considered different attack points within biometric verification systems. However, with advancements in AI-generated algorithms, a new type of attack—deepfake—has emerged. Deepfake technology enables real-time manipulation of a victim’s facial images, posing a significant challenge to biometric security by spoofing authentication systems. To better understand the vulnerabilities of face verification systems, the threats posed by deepfake attacks, and to enhance both the accuracy and security of face verification systems, the following contributions are made in this thesis:

Enhancing Face Verification Algorithm

Given that face verification systems are highly sensitive to variations in head pose, facial expressions, and illumination conditions, factors that often differ between ID card images and real-world user data, this thesis introduces a novel face alignment algorithm designed to enhance preprocessing for face verification. The proposed approach estimates the head pose, expression, and

illumination conditions from the first image pair and reconstructs the second image pair to align with these attributes while preserving its unique identity features.

Deepfake Quality Assessment

To better understand the threat posed by deepfake attacks, this thesis analyzes the quality of deepfake video frames generated by face reenactment techniques, with a focus on variations in facial expressions and head movements. This analysis helps to identify the weaknesses of different deepfake generators. Additionally, a gap in the literature is identified regarding the assessment of face reenactment quality. To address this, a novel protocol is introduced for the quantitative evaluation of images produced by face reenactment techniques, particularly in cross-reenactment scenarios. The protocol enables the assessment of cross-reenacted images using metrics that rely on explicit ground truth, such as SSIM and LPIPS. Given the limited availability of suitable datasets, two video generation approaches are proposed. The first approach utilizes 3D models of real human heads captured through a multi-view imaging system. The second approach employs realistic, synthetically generated head models that encompass a diverse range of human subjects, facial expressions, pose variations, and lighting conditions.

Effect of Beautification Filters on Deepfake Detectors

To assess the vulnerabilities and strengths of state-of-the-art deepfake detection methods against image processing techniques, the robustness of deepfake detectors was analyzed in this thesis when applied to beautified videos. Specifically, the impact of beautification filters on detection performance was measured by evaluating classification scores before and after applying these filters. Furthermore, the effectiveness of automated detectors was compared with the ability of an average human user to distinguish between real and fake videos, providing insights into both machine and human susceptibility to such alterations.

Finding Key Components for Effective Deepfake Detection

The performance of current deepfake detectors significantly degrades when common image processing techniques such as compression, resizing, and beautification filters are applied. To address this challenge, this thesis introduces a novel pipeline that leverages raw domain data as input to enhance deepfake detection. By focusing on raw data, the aim is to constrain the distribution of real images, making it easier for the model to learn distinctive features and generalize effectively to authentic images. This approach improves robustness against image alterations, leading to more reliable deepfake detection.

Detecting Replay Attacks via Compression Artifacts

In digital replay attacks, adversaries leverage authentic video footage of the victim, often sourced from publicly available platforms such as social media,

and inject it into the system using virtual camera software. Since the video is genuine and lacks manipulation artifacts, it poses significant challenges for the system to distinguish between a live user and a replayed video. This thesis investigates whether providing uncompressed video access to face anti-spoofing service providers can improve the detection of injected versus authentic video streams. This strategy enables the detection system to better differentiate between authentic and injected video streams, as injected videos are often sourced from the internet and are typically compressed using widely-used algorithms.

Participation in Deepfake Detection Challenges and Projects

In addition to the contributions mentioned above, I, along with other group members, participated in and won the Défi Hermès Deepfake Challenge [186]. In this competition, we developed an ensemble model that made decisions based on multiple specialized detectors, each expert in identifying different types of manipulations, including color modifications, splicing, face alterations, and entirely synthesized images. The final classification was determined by agreement among these detectors on an unseen dataset.

Furthermore, we contributed to a project defense in SPRIN-D [187] on deepfake detection, which successfully secured a grant. This project focuses on developing a deepfake detection system based on the JPEG AI compression algorithm, leveraging its ability to reconstruct real images with higher fidelity while degrading fake images, thereby enhancing detection accuracy.

8.2 Directions for Future Research

The field of deepfake generation is evolving rapidly, presenting an ongoing challenge for detection systems. Despite significant advancements in deepfake detection, the problem remains far from fully resolved. Based on the findings and contributions of this thesis, several promising directions for future research emerge.

One key avenue for improvement in authentication services lies in leveraging device sensors more effectively. If device manufacturers allow authentication providers to access raw sensor data, detection capabilities could be significantly enhanced. This access would enable more precise differentiation between genuine and manipulated content, strengthening security in biometric authentication systems.

This thesis highlights the potential of raw sensor data in deepfake detection. However, a major limitation is the absence of large-scale facial datasets specifically designed for training deepfake detection models on raw-domain images. To address this, an inverse Image Signal Processing (ISP) model was

introduced in this work to generate raw data during training. Future research could focus on developing extensive datasets captured directly from sensor outputs. Such datasets would enhance both the authenticity and robustness of deepfake detection models, improving their generalization across real-world scenarios.

Improving deepfake detection technologies is crucial for strengthening digital security, particularly in biometric authentication systems. As deepfake attacks become increasingly sophisticated, they pose significant threats to identity verification and access control mechanisms. By developing more robust detection methods, we can enhance the reliability of face authentication systems, preventing unauthorized access and mitigating security risks in applications such as online banking, identity verification, and secure communications.

Author's Publications

Conferences

- [P1] S. Hussein and J.-L. Dugelay, “Alignface: Enhancing face verification models through adaptive alignment of pose, expression, and illumination,” in *2024 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2024, pp. 3243–3249.
- [P2] S. Hussein, J.-L. Dugelay, F. Aili, and E. Nars, “A 3d-assisted framework to evaluate the quality of head motion replication by reenactment deepfake generators,” in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, pp. 1–5.
- [P3] S. Hussein and J.-L. Dugelay, “Metahumans help to evaluate deepfake generators,” in *2023 IEEE 25th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2023, pp. 1–6.
- [P4] S. Hussein and J.-L. Dugelay, “A comprehensive framework for evaluating deepfake generators: Dataset, metrics performance, and comparative analysis,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 372–381.
- [P5] A. Libourel, S. Hussein, N. Mirabet-Herranz, and J.-L. Dugelay, “A case study on how beautification filters can fool deepfake detectors,” in *IWBF 2024, 12th IEEE International Workshop on Biometrics and Forensics*, 2024.
- [P6] S. Hussein and J.-L. Dugelay, “Raw data: A key component for effective deepfake detection,” in *ICASSP 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2025, pp. 1–5.
- [P7] S. Hussein, “Towards secure authentication: Detecting replay attacks via compression artifacts,” *Submitted to IWBF*, 2025.

Challenges

- [C1] J.-L. Dugelay, S. Hussein, M. E. A. Bellebna, A. Libourel, A. Sitcharn, and A. E. Mennaoui, *Défi Hermès - Images Falsifiées*, Accessed: 2025-03-15, 2025.

Project Defenses

- [PD1] J.-L. Dugelay, S. Hussein, and A. Libourel, *Funke Deepfake Challenge*, Accessed: 2025-03-15, 2024.

References

- [1] M. Hassan, M. Suhail Shaikh, and M. A. Jatoi, “Image quality measurement-based comparative analysis of illumination compensation methods for face image normalization,” *Multimedia Systems*, pp. 1–10, 2022.
- [2] S. Dalal and V. P. Vishwakarma, “A novel approach of face recognition using optimized adaptive illumination–normalization and kelm,” *Arabian Journal for Science and Engineering*, vol. 45, pp. 9977–9996, 2020.
- [3] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, “Sphereface: Deep hypersphere embedding for face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 212–220.
- [4] H. Wang, Y. Wang, Z. Zhou, *et al.*, “Cosface: Large margin cosine loss for deep face recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5265–5274.
- [5] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, “Arcface: Additive angular margin loss for deep face recognition,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.
- [6] S. Bhattacharjee, A. Mohammadi, A. Anjos, and S. Marcel, “Recent advances in face presentation attack detection,” *Handbook of Biometric Anti-Spoofing: Presentation Attack Detection*, pp. 207–228, 2019.
- [7] K. Carta, C. Barral, N. El Mrabet, and S. Mouille, “Video injection attacks on remote digital identity verification solution using face recognition,” in *13th International Multi-Conference on Complexity, Informatics and Cybernetics*, 2022.

- [8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, “Generative adversarial networks,” *Communications of the ACM*, vol. 63, no. 11, pp. 139–144, 2020.
- [9] L. Yang, Z. Zhang, Y. Song, *et al.*, “Diffusion models: A comprehensive survey of methods and applications,” *ACM Computing Surveys*, vol. 56, no. 4, pp. 1–39, 2023.
- [10] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “Faceforensics++: Learning to detect manipulated facial images,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 1–11.
- [11] D. Siegel, C. Kraetzer, S. Seidlitz, and J. Dittmann, “Media forensics considerations on deepfake detection with hand-crafted features,” *Journal of Imaging*, vol. 7, no. 7, p. 108, 2021.
- [12] A. Heidari, N. Jafari Navimipour, H. Dag, and M. Unal, “Deepfake detection using deep learning methods: A systematic and comprehensive review,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 14, no. 2, e1520, 2024.
- [13] J. Choi, T. Kim, Y. Jeong, S. Baek, and J. Choi, “Exploiting style latent flows for generalizing deepfake detection video detection,” *arXiv preprint arXiv:2403.06592*, 2024.
- [14] S. Dong, J. Wang, R. Ji, J. Liang, H. Fan, and Z. Ge, “Implicit identity leakage: The stumbling block to improving deepfake detection generalization,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 3994–4004.
- [15] Y. Luo, Y. Zhang, J. Yan, and W. Liu, “Generalizing face forgery detection with high-frequency features,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 16 317–16 326.
- [16] T. Zhao, X. Xu, M. Xu, H. Ding, Y. Xiong, and W. Xia, “Learning self-consistency for deepfake detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 023–15 033.
- [17] A. Megahed, Q. Han, and S. Fadl, “Exposing deepfake using fusion of deep-learned and hand-crafted features,” *Multimedia Tools and Applications*, vol. 83, no. 9, pp. 26 797–26 817, 2024.
- [18] A. Al-Adwan, H. Alazzam, N. Al-Anbaki, and E. Alduweib, “Detection of deepfake media using a hybrid cnn–rnn model and particle swarm optimization (psa) algorithm,” *Computers*, vol. 13, no. 4, p. 99, 2024.

- [19] European Union Agency for Cybersecurity (ENISA), *Remote id proofing - good practices*, Accessed: 2025-01-30.
- [20] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9851–9858.
- [21] H. Wu, J. Gu, X. Fan, H. Li, L. Xie, and J. Zhao, "3d-guided frontal face generation for pose-invariant recognition," *ACM Transactions on Intelligent Systems and Technology*, vol. 14, no. 2, pp. 1–21, 2023.
- [22] Y. Lu and T. Ebrahimi, "Assessment framework for deepfake detection in real-world situations," *EURASIP Journal on Image and Video Processing*, vol. 2024, no. 1, p. 6, 2024.
- [23] F. Cocchi, L. Baraldi, S. Poppi, M. Cornia, L. Baraldi, and R. Cucchiara, "Unveiling the impact of image transformations on deepfake detection: An experimental analysis," in *International Conference on Image Analysis and Processing*, Springer, 2023, pp. 345–356.
- [24] X. Zhang, S. Karaman, and S.-F. Chang, "Detecting and simulating artifacts in gan fake images," in *2019 IEEE international workshop on information forensics and security (WIFS)*, IEEE, 2019, pp. 1–6.
- [25] S. Gal and B. Bulgurcu, "Exploring factors influencing internet users' susceptibility to deepfake phishing," 2024.
- [26] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, IEEE, vol. 1, 2005, pp. 539–546.
- [27] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.
- [28] Q. Meng, S. Zhao, Z. Huang, and F. Zhou, "Magface: A universal representation for face recognition and quality assessment," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 14 225–14 234.
- [29] N. K. Ratha, J. H. Connell, and R. M. Bolle, "Enhancing security and privacy in biometrics-based authentication systems," *IBM systems Journal*, vol. 40, no. 3, pp. 614–634, 2001.

- [30] X. Xu, T. Zhao, Z. Zhang, *et al.*, “Principles of designing robust remote face anti-spoofing systems,” *arXiv preprint arXiv:2406.03684*, 2024.
- [31] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, *et al.*, *Generative adversarial networks*, 2014. arXiv: 1406.2661 [stat.ML].
- [32] D. P. Kingma, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013.
- [33] J. Ho, A. Jain, and P. Abbeel, *Denoising diffusion probabilistic models*, 2020. arXiv: 2006.11239 [cs.LG].
- [34] J. Ho, C. Saharia, W. Chan, D. J. Fleet, M. Norouzi, and T. Salimans, “Cascaded diffusion models for high fidelity image generation,” *Journal of Machine Learning Research*, vol. 23, no. 47, pp. 1–33, 2022.
- [35] Y. Pu, Z. Gan, R. Henao, *et al.*, “Variational autoencoder for deep learning of images, labels and captions,” *Advances in neural information processing systems*, vol. 29, 2016.
- [36] G. Pei, J. Zhang, M. Hu, *et al.*, “Deepfake generation and detection: A benchmark and survey,” *arXiv preprint arXiv:2403.17881*, 2024.
- [37] K. Sunkavalli, M. K. Johnson, W. Matusik, and H. Pfister, “Multi-scale image harmonization,” *ACM Transactions on Graphics (TOG)*, vol. 29, no. 4, pp. 1–10, 2010.
- [38] D. Bitouk, N. Kumar, S. Dhillon, P. Belhumeur, and S. K. Nayar, “Face swapping: Automatically replacing faces in photographs,” in *ACM SIGGRAPH 2008 papers*, 2008, pp. 1–8.
- [39] V. Blanz, K. Scherbaum, T. Vetter, and H.-P. Seidel, “Exchanging faces in images,” in *Computer Graphics Forum*, Wiley Online Library, vol. 23, 2004, pp. 669–676.
- [40] K. Dale, K. Sunkavalli, M. K. Johnson, D. Vlastic, W. Matusik, and H. Pfister, “Video face replacement,” in *Proceedings of the 2011 SIGGRAPH Asia conference*, 2011, pp. 1–10.
- [41] I. Korshunova, W. Shi, J. Dambre, and L. Theis, “Fast face-swap using convolutional neural networks,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 3677–3685.
- [42] P. Yu, Z. Xia, J. Fei, and Y. Lu, “A survey on deepfake video detection,” *Iet Biometrics*, vol. 10, no. 6, pp. 607–624, 2021.
- [43] J. R. A. Moniz, C. Beckham, S. Rajotte, S. Honari, and C. Pal, “Unsupervised depth estimation, 3d face rotation and replacement,” *Advances in neural information processing systems*, vol. 31, 2018.

- [44] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “Cvae-gan: Fine-grained image generation through asymmetric training,” in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2745–2754.
- [45] R. Natsume, T. Yatagawa, and S. Morishima, “Rsgan: Face swapping and editing using face and hair representation in latent spaces,” *arXiv preprint arXiv:1804.03447*, 2018.
- [46] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua, “Towards open-set identity preserving face synthesis,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6713–6722.
- [47] F. Liu, L. Yu, H. Xie, *et al.*, “High fidelity face swapping via semantics disentanglement and structure enhancement,” in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 6907–6917.
- [48] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, “Information bottleneck disentanglement for identity swapping,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 3404–3413.
- [49] Z. Liu, M. Li, Y. Zhang, *et al.*, “Fine-grained face swapping via regional gan inversion,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 8578–8587.
- [50] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, “Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8568–8577.
- [51] R. Liu, B. Ma, W. Zhang, *et al.*, “Towards a simultaneous and granular identity-expression control in personalized face generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 2114–2123.
- [52] M. Zollhöfer, J. Thies, P. Garrido, *et al.*, “State of the art on monocular 3d face reconstruction, tracking, and applications,” in *Computer graphics forum*, Wiley Online Library, vol. 37, 2018, pp. 523–550.
- [53] J. Cao, Y. Hu, B. Yu, R. He, and Z. Sun, “Load balanced gans for multi-view face image synthesis,” *arXiv preprint arXiv:1802.07447*, 2018.

- [54] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow, “Interpretable transformations with encoder-decoder networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5726–5735.
- [55] E. L. Denton *et al.*, “Unsupervised learning of disentangled representations from video,” *Advances in neural information processing systems*, vol. 30, 2017.
- [56] O. Wiles, A. Koepke, and A. Zisserman, “X2face: A network for controlling face generation using images, audio, and pose codes,” in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 670–686.
- [57] S. Tripathy, J. Kannala, and E. Rahtu, “Icface: Interpretable and controllable face reenactment using gans,” in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2020, pp. 3385–3394.
- [58] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “Animating arbitrary objects via deep motion transfer,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 2377–2386.
- [59] F.-T. Hong, L. Zhang, L. Shen, and D. Xu, “Depth-aware generative adversarial network for talking head video generation,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3397–3406.
- [60] Y. Wang, D. Yang, F. Bremond, and A. Dantcheva, “Latent image animator: Learning to animate images via latent space navigation,” *arXiv preprint arXiv:2203.09043*, 2022.
- [61] N. Ruiz, E. Chong, and J. M. Rehg, “Fine-grained head pose estimation without keypoints,” in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 2074–2083.
- [62] Y. Deng, J. Yang, S. Xu, D. Chen, Y. Jia, and X. Tong, “Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019.
- [63] B. Amos, B. Ludwiczuk, M. Satyanarayanan, *et al.*, “Openface: A general-purpose face recognition library with mobile applications,” *CMU School of Computer Science*, vol. 6, no. 2, p. 20, 2016.

- [64] Y. Nirkin, Y. Keller, and T. Hassner, “Fsgan: Subject agnostic face swapping and reenactment,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7184–7193.
- [65] D. E. King, “Dlib-ml: A machine learning toolkit,” *The Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.
- [66] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, “Gans trained by a two time-scale update rule converge to a local nash equilibrium,” *Advances in neural information processing systems*, vol. 30, 2017.
- [67] A. Hore and D. Ziou, “Image quality metrics: Psnr vs. ssim,” in *2010 20th international conference on pattern recognition*, IEEE, 2010, pp. 2366–2369.
- [68] A. Siarohin, S. Lathuilière, S. Tulyakov, E. Ricci, and N. Sebe, “First order motion model for image animation,” *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [69] F. Yin, Y. Zhang, X. Cun, *et al.*, “Styleheat: One-shot high-resolution editable talking face generation via pre-trained stylegan,” in *European conference on computer vision*, Springer, 2022, pp. 85–101.
- [70] Y. Gao, Y. Zhou, J. Wang, X. Li, X. Ming, and Y. Lu, “High-fidelity and freely controllable talking head video generation,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5609–5619.
- [71] K. Yang, K. Chen, D. Guo, S.-H. Zhang, Y.-C. Guo, and W. Zhang, “Face2face ρ : Real-time high-resolution one-shot face reenactment,” in *European conference on computer vision*, Springer, 2022, pp. 55–71.
- [72] R. Abdal, Y. Qin, and P. Wonka, “Image2stylegan: How to embed images into the stylegan latent space?” In *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4432–4441.
- [73] M. Cornia, M. Tomei, L. Baraldi, and R. Cucchiara, “Matching faces and attributes between the artistic and the real domain: The personart approach,” *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 18, no. 3, pp. 1–23, 2022.
- [74] S. Ha, M. Kersner, B. Kim, S. Seo, and D. Kim, “Marionette: Few-shot face reenactment preserving identity of unseen targets,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, 2020, pp. 10 893–10 900.

- [75] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky, “Fast bi-layer neural synthesis of one-shot realistic head avatars,” in *European Conference on Computer Vision*, Springer, 2020, pp. 524–540.
- [76] L. Chen, G. Cui, Z. Kou, H. Zheng, and C. Xu, “What comprises a good talking-head video generation?” In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020.
- [77] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, “FaceForensics++: Learning to detect manipulated facial images,” in *International Conference on Computer Vision (ICCV)*, 2019.
- [78] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df: A large-scale challenging dataset for deepfake forensics,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 3207–3216.
- [79] B. Dolhansky, J. Bitton, B. Pflaum, *et al.*, “The deepfake detection challenge (dfdc) dataset,” *arXiv preprint arXiv:2006.07397*, 2020.
- [80] H. Liu, X. Li, W. Zhou, *et al.*, “Spatial-phase shallow learning: Rethinking face forgery detection in frequency domain,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 772–781.
- [81] Q. Gu, S. Chen, T. Yao, Y. Chen, S. Ding, and R. Yi, “Exploiting fine-grained face forgery clues via progressive enhancement learning,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 735–743.
- [82] S. Chen, T. Yao, Y. Chen, S. Ding, J. Li, and R. Ji, “Local relation learning for face forgery detection,” in *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, 2021, pp. 1081–1088.
- [83] Y. Li and S. Lyu, “Exposing deepfake videos by detecting face warping artifacts,” *arXiv preprint arXiv:1811.00656*, 2018.
- [84] L. Li, J. Bao, T. Zhang, *et al.*, “Face x-ray for more general face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.
- [85] K. Shiohara and T. Yamasaki, “Detecting deepfakes with self-blended images,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 720–18 729.

- [86] X. Yang, Y. Li, and S. Lyu, “Exposing deep fakes using inconsistent head poses,” in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2019, pp. 8261–8265.
- [87] Y. Li, M.-C. Chang, and S. Lyu, “In ictu oculi: Exposing ai created fake videos by detecting eye blinking,” in *2018 IEEE International workshop on information forensics and security (WIFS)*, IEEE, 2018, pp. 1–7.
- [88] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic, “Lips don’t lie: A generalisable and robust approach to face forgery detection,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5039–5049.
- [89] J. Hernandez-Ortega, R. Tolosana, J. Fierrez, and A. Morales, “Deepfakeson-phys: Deepfakes detection based on heart rate estimation,” *arXiv preprint arXiv:2010.00400*, 2020.
- [90] M. Kombrink and Z. Geradts, “The influence of compression on the detection of deepfake videos,” *Artificial Intelligence (AI) in Forensic Sciences*, p. 174, 2023.
- [91] A. Mitra, S. P. Mohanty, P. Corcoran, and E. Kougianos, “A machine learning based approach for deepfake detection in social media through key video frame extraction,” *SN Computer Science*, vol. 2, pp. 1–18, 2021.
- [92] Z. Yan, P. Sun, Y. Lang, *et al.*, *Multimodal graph learning for deepfake detection*, 2023. arXiv: 2209.05419 [cs.CV].
- [93] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang, “End-to-end reconstruction-classification learning for face forgery detection,” in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 4103–4112.
- [94] S. Dong, J. Wang, J. Liang, H. Fan, and R. Ji, “Explaining deepfake detection by analysing image matching,” in *European Conference on Computer Vision*, Springer, 2022, pp. 18–35.
- [95] T. Brooks, B. Mildenhall, T. Xue, J. Chen, D. Sharlet, and J. T. Barron, “Unprocessing images for learned raw denoising,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 036–11 045.

- [96] K. Q. Dinh and K. P. Choi, “End-to-end single-frame image signal processing for high dynamic range scenes,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2023, pp. 2449–2458.
- [97] S. Zini, C. Rota, M. Buzzelli, S. Bianco, and R. Schettini, “Back to the future: A night photography rendering isp without deep learning,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1465–1473.
- [98] Y. Xing, Z. Qian, and Q. Chen, “Invertible image signal processing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 6287–6296.
- [99] Q. Jin, G. Facciolo, and J.-M. Morel, “A review of an old dilemma: Demosaicking first, or denoising first?” In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2020, pp. 514–515.
- [100] B. Kawar, G. Vaksman, and M. Elad, “Stochastic image denoising by sampling from the posterior distribution,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1866–1875.
- [101] Y. Hou, J. Xu, M. Liu, *et al.*, “Nlh: A blind pixel-level non-local method for real-world image denoising,” *IEEE Transactions on Image Processing*, vol. 29, pp. 5121–5135, 2020.
- [102] S. Hussein, “Color constancy with small dataset via pruning of cnn filters,” M.S. thesis, 2021.
- [103] M. Afifi and M. S. Brown, “Deep white-balance editing,” in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2020, pp. 1397–1406.
- [104] M. V. Conde, S. McDonagh, M. Maggioni, A. Leonardis, and E. Pérez-Pellitero, “Model-based image signal processors via learnable dictionaries,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, 2022, pp. 481–489.
- [105] S. Chen, H. Feng, D. Pan, Z. Xu, Q. Li, and Y. Chen, “Optical aberrations correction in postprocessing using imaging simulation,” *ACM Transactions on Graphics (TOG)*, vol. 40, no. 5, pp. 1–15, 2021.
- [106] J. Otsuka, M. Yoshimura, and T. Ohashi, “Self-supervised reversed image signal processing via reference-guided dynamic parameter selection,” *arXiv preprint arXiv:2303.13916*, 2023.

- [107] M. Yoshimura, J. Otsuka, A. Irie, and T. Ohashi, “Rawgment: Noise-accounted raw augmentation enables recognition in a wide variety of environments,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 007–14 017.
- [108] Y. Hong, K. Wei, L. Chen, and Y. Fu, “Crafting object detection in very low light,” in *BMVC*, vol. 1, 2021, p. 3.
- [109] S. W. Zamir, A. Arora, S. Khan, *et al.*, “Cycleisp: Real image restoration via improved data synthesis,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2696–2705.
- [110] M. V. Conde, R. Timofte, Y. Huang, *et al.*, “Reversed image signal processing and raw reconstruction. aim 2022 challenge report,” in *European Conference on Computer Vision*, Springer, 2022, pp. 3–26.
- [111] D. P. Kingma and P. Dhariwal, “Glow: Generative flow with invertible 1x1 convolutions,” *Advances in neural information processing systems*, vol. 31, 2018.
- [112] K. Carta, C. Barral, N. El Mrabet, and S. Mouille, “On the pitfalls of videoconferences for challenge-based face liveness detection,” in *proceedings of World Multi-Conference on Systemics, Cybernetics and Informatics*, vol. 2021, 2021.
- [113] D. F. Smith, A. Wiliem, and B. C. Lovell, “Face recognition on consumer devices: Reflections on replay attacks,” *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 4, pp. 736–745, 2015.
- [114] H. Farrukh, R. M. Aburas, S. Cao, and H. Wang, “Facerevelio: A face liveness detection system for smartphones with a single front camera,” in *Proceedings of the 26th Annual International Conference on Mobile Computing and Networking*, 2020, pp. 1–13.
- [115] D. Tang, Z. Zhou, Y. Zhang, and K. Zhang, “Face flashing: A secure liveness detection protocol based on light reflections,” *arXiv preprint arXiv:1801.01949*, 2018.
- [116] H. Liu, Z. Li, Y. Xie, *et al.*, “Livescreen: Video chat liveness detection leveraging skin reflection,” in *IEEE INFOCOM 2020-IEEE Conference on Computer Communications*, IEEE, 2020, pp. 1083–1092.

- [117] S. Milani, P. Bestagini, M. Tagliasacchi, and S. Tubaro, "Multiple compression detection for video sequences," in *2012 IEEE 14th International Workshop on Multimedia Signal Processing (MMSP)*, IEEE, 2012, pp. 112–117.
- [118] X. Jiang, W. Wang, T. Sun, Y. Q. Shi, and S. Wang, "Detection of double compression in mpeg-4 videos based on markov statistics," *IEEE Signal processing letters*, vol. 20, no. 5, pp. 447–450, 2013.
- [119] Y. Li, M. Gardella, Q. Bammey, *et al.*, "A contrario detection of h. 264 video double compression," in *2023 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2023, pp. 1765–1769.
- [120] D. Vazquez-Padin, M. Fontani, T. Bianchi, P. Comesaña, A. Piva, and M. Barni, "Detection of video double encoding with gop size estimation," in *2012 IEEE International Workshop on Information Forensics and Security (WIFS)*, IEEE, 2012, pp. 151–156.
- [121] M. Kim, A. K. Jain, and X. Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18 750–18 759.
- [122] C.-H. Tang, Y.-M. Chou, and G.-S. J. Hsu, "Multi-view normalization for face recognition," in *2021 IEEE International Conference on Image Processing (ICIP)*, IEEE, 2021, pp. 2343–2347.
- [123] G.-S. Hsu and C.-H. Tang, "Dual-view normalization for face recognition," *IEEE Access*, vol. 8, pp. 147 765–147 775, 2020.
- [124] P. Paysan, R. Knothe, B. Amberg, S. Romdhani, and T. Vetter, "A 3d face model for pose and illumination invariant face recognition," in *2009 sixth IEEE international conference on advanced video and signal based surveillance*, Ieee, 2009, pp. 296–301.
- [125] Y. Guo, J. Cai, B. Jiang, J. Zheng, *et al.*, "Cnn-based real-time dense face reconstruction with inverse-rendered photo-realistic face images," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 6, pp. 1294–1307, 2018.
- [126] C. Cao, Y. Weng, S. Zhou, Y. Tong, and K. Zhou, "Facewarehouse: A 3d facial expression database for visual computing," *IEEE Transactions on Visualization and Computer Graphics*, vol. 20, no. 3, pp. 413–425, 2013.
- [127] X. Yang, C. Liu, L. Xu, *et al.*, "Towards effective adversarial textured 3d meshes on physical face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4119–4128.

- [128] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” in *Workshop on faces in ‘Real-Life’ Images: detection, alignment, and recognition*, 2008.
- [129] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, “Frontal to profile face verification in the wild,” in *2016 IEEE winter conference on applications of computer vision (WACV)*, IEEE, 2016, pp. 1–9.
- [130] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, “Agedb: The first manually collected, in-the-wild age database,” in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.
- [131] C. Whitelam, E. Taborsky, A. Blanton, *et al.*, “Iarpa janus benchmark-b face dataset,” in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 90–98.
- [132] X. Wang, S. Zhang, S. Wang, T. Fu, H. Shi, and T. Mei, “Mis-classified vector guided softmax loss for face recognition,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 2020, pp. 12 241–12 248.
- [133] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, “Few-shot video-to-video synthesis,” in *Conference on Neural Information Processing Systems (NeurIPS)*, 2019.
- [134] J.-W. Seow, M.-K. Lim, R. C.-W. Phan, and J. K. Liu, “A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities,” *Neurocomputing*, 2022.
- [135] Y. Wang, P. Bilinski, F. Bremond, and A. Dantcheva, “G3an: This video does not exist. disentangling motion and appearance for video generation,” *arXiv preprint arXiv:1912.05523*, 2019.
- [136] M. C. Doukas, S. Zafeiriou, and V. Sharmanska, “Headgan: One-shot neural head synthesis and editing,” in *Proceedings of the IEEE/CVF International conference on Computer Vision*, 2021, pp. 14 398–14 407.
- [137] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, “The unreasonable effectiveness of deep features as a perceptual metric,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 586–595.
- [138] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly, “Towards accurate generative models of video: A new metric & challenges,” *arXiv preprint arXiv:1812.01717*, 2018.

- [139] H. Yang, H. Zhu, Y. Wang, *et al.*, “Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2020.
- [140] Z. Fang, L. Cai, and G. Wang, “Metahuman creator the starting point of the metaverse,” in *2021 International Symposium on Computer Technology and Information Science (ISCTIS)*, 2021, pp. 154–157.
- [141] P. Babahajiani, “Geometric computer vision: Omnidirectional visual and remotely sensed data analysis,” 2021.
- [142] S. Husseini, “A survey of optical flow techniques for object tracking,” B.S. thesis, 2017.
- [143] Unreal Engine, *Metahumans in quixel bridge*, <https://docs.metahuman.unrealengine.com/en-US/metahumans-in-quixel-bridge/>, Accessed on May 12, 2023, 2021.
- [144] M. Perez-Ortiz, A. Mikhailiuk, E. Zerman, V. Hulusic, G. Valenzise, and R. K. Mantiuk, “From pairwise comparisons and rating to a unified quality scale,” *IEEE Transactions on Image Processing*, vol. 29, pp. 1139–1151, 2019.
- [145] L. L. Thurstone, “A law of comparative judgment.,” *Psychological review*, vol. 34, no. 4, p. 273, 1927.
- [146] C. Lugaresi, J. Tang, H. Nash, *et al.*, “Mediapipe: A framework for perceiving and processing reality,” in *Third Workshop on Computer Vision for AR/VR at IEEE Computer Vision and Pattern Recognition (CVPR)*, vol. 2019, 2019.
- [147] J. Gui, T. Chen, J. Zhang, *et al.*, “A survey on self-supervised learning: Algorithms, applications, and future trends,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
- [148] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, and Y. Liu, “Countering malicious deepfakes: Survey, battleground, and horizon,” *International journal of computer vision*, vol. 130, no. 7, pp. 1678–1734, 2022.
- [149] S. M. Abdullah, A. Cheruvu, S. Kanchi, *et al.*, “An analysis of recent advances in deepfake image detection in an evolving threat landscape,” *arXiv preprint arXiv:2404.16212*, 2024.
- [150] X. Hu, Q. Huang, Z. Shi, *et al.*, “Style transformer for image inversion and editing,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 11 337–11 346.

- [151] Y. Lu, R. Luo, and T. Ebrahimi, “Improving deepfake detectors against real-world perturbations with amplitude-phase switch augmentation,” in *Applications of Digital Image Processing XLVI*, SPIE, vol. 12674, 2023, p. 1 267 402.
- [152] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper, “Unmasking deepfakes with simple features,” *arXiv preprint arXiv:1911.00686*, 2019.
- [153] Z. Chen, X. Liao, X. Wu, and Y. Chen, “Compressed deepfake video detection based on 3d spatiotemporal trajectories,” *arXiv preprint arXiv:2404.18149*, 2024.
- [154] N. Mirabet-Herranz, C. Galdi, and J.-L. Dugelay, “Impact of digital face beautification in biometrics,” in *2022 10th European Workshop on Visual Information Processing (EUVIP)*, IEEE, 2022, pp. 1–6.
- [155] Y. Lu and T. Ebrahimi, “Impact of video processing operations in deepfake detection,” *arXiv preprint arXiv:2303.17247*, 2023.
- [156] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, “Celeb-df (v2): A new dataset for deepfake forensics [j],” *arXiv preprint arXiv*, 2019.
- [157] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen, “Exploring temporal coherence for more general video face forgery detection,” in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 044–15 054.
- [158] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG].
- [159] F. Chollet, *Xception: Deep learning with depthwise separable convolutions*, 2017. arXiv: 1610.02357 [cs.CV].
- [160] K. Hara, H. Kataoka, and Y. Satoh, *Learning spatio-temporal features with 3d residual networks for action recognition*, 2017. arXiv: 1708.07632 [cs.CV].
- [161] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, *An image is worth 16x16 words: Transformers for image recognition at scale*, 2021. arXiv: 2010.11929 [cs.CV].
- [162] C. Keimel, J. Habigt, C. Horch, and K. Diepold, “Qualitycrowd—a framework for crowd-based quality evaluation,” in *2012 Picture coding symposium*, IEEE, 2012, pp. 245–248.

- [163] P. Korshunov and S. Marcel, “Subjective and objective evaluation of deepfake videos,” in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2021, pp. 2510–2514.
- [164] Deepfakes, *Deepfakes*, [Online; accessed 2024-04-30].
- [165] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Nießner, “Face2face: Real-time face capture and reenactment of rgb videos,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2387–2395.
- [166] MarekKowalski, *Faceswap*, [Online; accessed 2024-06-12].
- [167] J. Thies, M. Zollhöfer, and M. Nießner, “Deferred neural rendering: Image synthesis using neural textures,” *Acm Transactions on Graphics (TOG)*, vol. 38, no. 4, pp. 1–12, 2019.
- [168] *Contributing data to deepfake detection research*. [Online; accessed 2024-04-30].
- [169] J. Deng, J. Guo, Y. Zhou, J. Yu, I. Kotsia, and S. Zafeiriou, “Retinaface: Single-stage dense face localisation in the wild,”
- [170] Codeniko. “Codeniko/shape_predictor_81_face_landmarks: Custom shape predictor model trained to find 81 facial feature landmarks given any image.” [Online; accessed 2024-04-30]. (2021), [Online]. Available: https://github.com/codeniko/shape_predictor_81_face_landmarks.
- [171] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, p. 125, 2020.
- [172] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed, “Two-branch recurrent network for isolating deepfakes in videos,” in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, Springer, 2020, pp. 667–684.
- [173] T. Zhou, W. Wang, Z. Liang, and J. Shen, “Face forensics in the wild,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5778–5788.
- [174] M. Tan and Q. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *International conference on machine learning*, PMLR, 2019, pp. 6105–6114.

- [175] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, “Sharpness-aware minimization for efficiently improving generalization,” in *International Conference on Learning Representations*, 2020.
- [176] C. Galdi, M. Nappi, and J.-L. Dugelay, “Multimodal authentication on smartphones: Combining iris and sensor recognition for a double check of user identity,” *Pattern Recognition Letters*, vol. 82, pp. 144–153, 2016.
- [177] Xiph.org. “Video test media.” [Online]. Available: <https://media.xiph.org/video/derf/>. Accessed: 2024-11-13., Xiph.org Foundation. (2024), [Online]. Available: <https://media.xiph.org/video/derf/>.
- [178] L. Song, X. Tang, W. Zhang, X. Yang, and P. Xia, “The sjtu 4k video sequence dataset,” in *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, IEEE, 2013, pp. 34–35.
- [179] L. Song, Y. Liu, X. Yang, G. Zhai, R. Xie, and W. Zhang, “The sjtu hdr video sequence dataset,” in *Proceedings of International Conference on Quality of Multimedia Experience (QoMEX 2016)*, 2016, p. 100.
- [180] A. Mercat, M. Viitanen, and J. Vanne, “Uvg dataset: 50/120fps 4k sequences for video codec analysis and development,” in *Proceedings of the 11th ACM Multimedia Systems Conference*, 2020, pp. 297–302.
- [181] Z. Li, J. Liao, C. Tang, *et al.*, “Ustc-td: A test dataset and benchmark for image and video coding in 2020s,” *arXiv preprint arXiv:2409.08481*, 2024.
- [182] S. Tomar, “Converting video formats with ffmpeg,” *Linux journal*, vol. 2006, no. 146, p. 10, 2006.
- [183] H. Wang, W. Gan, S. Hu, *et al.*, “Mcl-jcv: A jnd-based h. 264/avc video quality assessment dataset,” in *2016 IEEE international conference on image processing (ICIP)*, IEEE, 2016, pp. 1509–1513.
- [184] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” *arXiv preprint arXiv:1412.6806*, 2014.
- [185] 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 *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 618–626.

- [186] J.-L. Dugelay, S. Hussein, M. E. A. Bellebna, A. Libourel, A. Sitcharn, and A. E. Mennaoui, *Défi Hermès - Images Falsifiées*, Accessed: 2025-03-15, 2025.
- [187] J.-L. Dugelay, S. Hussein, and A. Libourel, *Funke Deepfake Challenge*, Accessed: 2025-03-15, 2024.