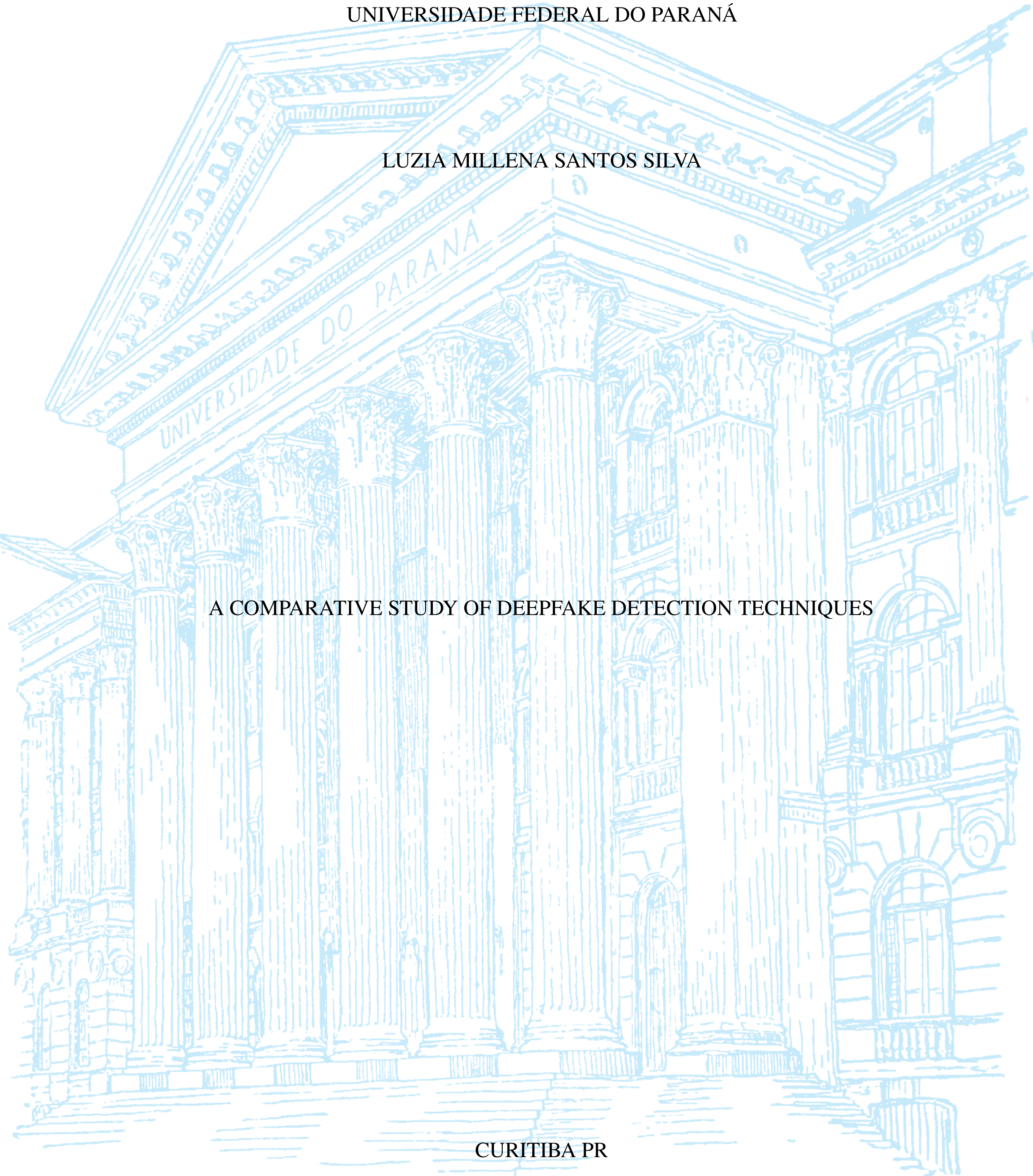
UNIVERSIDADE FEDERAL DO PARANÁ

LUZIA MILLENA SANTOS SILVA

A COMPARATIVE STUDY OF DEEPFAKE DETECTION TECHNIQUES



CURITIBA PR

2024

LUZIA MILLENA SANTOS SILVA

A COMPARATIVE STUDY OF DEEPFAKE DETECTION TECHNIQUES

Trabalho apresentado como requisito parcial à conclusão do Curso de Bacharelado em Ciência da computação, Setor de Ciências Exatas, da Universidade Federal do Paraná.

Área de concentração: *Computação*.

Orientador: David Menotti Gomes.

CURITIBA PR

2024

Ficha catalográfica

Substituir o arquivo `0-iniciais/catalografica.pdf` pela ficha catalográfica fornecida pela Biblioteca da UFPR (PDF em formato A4).

Instruções para obter a ficha catalográfica e fazer o depósito legal da tese/dissertação (contribuição de André Hochuli, abril 2019. Links atualizados Wellton Costa, Nov 2022):

1. Estas instruções se aplicam a dissertações de mestrado e teses de doutorado. Trabalhos de conclusão de curso de graduação e textos de qualificação não precisam segui-las.
2. Verificar se está usando a versão mais recente do modelo do PPGInf e atualizar, se for necessário (`https://gitlab.c3sl.ufpr.br/maziero/tese`).
3. conferir o *checklist* de formato do Sistema de Bibliotecas da UFPR, em `https://bibliotecas.ufpr.br/servicos/normalizacao/`
4. Enviar e-mail para "`referencia.bct@ufpr.br`" com o arquivo PDF da dissertação/tese, solicitando a respectiva ficha catalográfica.
5. Ao receber a ficha, inseri-la em seu documento (substituir o arquivo `0-iniciais/catalografica.pdf` do diretório do modelo).
6. Emitir a Certidão Negativa (CND) de débito junto a biblioteca, em `https://bibliotecas.ufpr.br/servicos/certidao-negativa/`
7. Avisar a secretaria do PPGInf que você está pronto para o depósito. Eles irão mudar sua titulação no SIGA, o que irá liberar uma opção no SIGA pra você fazer o depósito legal.
8. Acesse o SIGA (`http://www.prppg.ufpr.br/siga`) e preencha com cuidado os dados solicitados para o depósito da tese.
9. Aguarde a confirmação da Biblioteca.
10. Após a aprovação do pedido, informe a secretaria do PPGInf que a dissertação/tese foi depositada pela biblioteca. Será então liberado no SIGA um link para a confirmação dos dados para a emissão do diploma.

Ficha de aprovação

Substituir o arquivo 0-iniciais/aprovacao.pdf pela ficha de aprovação fornecida pela secretaria do programa, em formato PDF A4.

ACKNOWLEDGEMENTS

Primeiramente, agradeço a Deus pela energia, força e sabedoria durante toda essa jornada.

À minha mãe Marlene por me encorajar tanto a iniciar esse curso, a toda a minha família, pelo apoio incondicional e por sempre acreditarem em mim.

A Vinícius e Louise, pela companhia, apoio, por estarem presentes por todo esse tempo e por marcarem momentos tão incríveis desde o início dessa trajetória.

Ao meu namorado, Caio, obrigada por todo o carinho, paciência e por me encorajar todos os dias.

A todos os meus amigos, pelo companheirismo e pelos momentos de descontração que tornaram tudo mais leve.

Por fim, agradeço aos professores e colegas que contribuíram direta ou indiretamente para a realização deste trabalho.

Muito obrigada a todos que fizeram parte dessa trajetória!

RESUMO

Deepfake envolve a criação de imagens manipuladas para forjar a identidade de um indivíduo por diversos motivos, incluindo uso político, criminal ou de entretenimento. O uso malicioso dessa tecnologia é uma preocupação crescente, pois pode impactar significativamente as reputações de indivíduos e figuras públicas, além de moldar percepções sociais e políticas. Dessa forma, este trabalho tem como objetivo explorar estudos recentes sobre técnicas de deepfake, analisar os desafios associados a esse tema e conduzir experimentos utilizando os datasets DFDC e FaceForensics++ seguindo protocolos de treinamento de intra-dataset, cross-dataset e fusion-dataset. Para avaliar e comparar a eficácia na detecção de deepfakes, foram utilizadas as arquiteturas Xception e EfficientNet como baseline, enquanto os Vision Transformers foram escolhidos como uma abordagem de estado da arte. Os resultados obtidos nesse trabalho indicam que o treinamento do modelo utilizando Vision Transformers no protocolo intra-dataset com o DFDC demonstram performance supeior entre os experimentos realizados. Com 0,98 de acurácia, 0,27 em Equal Error Rate (EER), 0,007 em Half Total Error Rate (HTER) e 0,02 em Detection Cost Function (DCF). Além disso, a abordabem seguindo o protocolo de fusion-dataset (treinamento utilizando DFDC e Faceforensics++) demonstrou robustez razoável, reduzindo overfitting e melhorando a capacidade de detecção entre diferentes metodos de deepfake.

Palavras-chave: Detecção de Deepfakes. Vision transformers. Face-swap. Face-reenactment.

ABSTRACT

Deepfake technology involves the generation of manipulated images to forge someone's identity for various purposes, including politics, crime or entertainment. Malicious applications of deepfakes extend beyond reputation damage. This work aims to explore recent studies on deepfake techniques, examine the associated challenges, and conduct experiments using the DFDC and FaceForensics++ datasets, following intra-dataset, cross-dataset and fusion-dataset training protocols. To evaluate and compare the efficiency in deepfake detection, the experiments leveraged Xception and EfficientNet as baseline architectures, while the Efficient Vision Transformers were used as the state-of-the-art approach. The results obtained in this work indicate that training the Vision transformer model in the intra-dataset protocol using the DFDC dataset demonstrated the best performance, with 0.98 accuracy, 0.27 eer, 0.007 hter and 0.02 dcf. Additionally, the approach leveraging a fusion-dataset protocol (combining DFDC and Faceforensics++ datasets during training) showed reasonable robustness, reducing overfitting and improving detection capabilities across different deepfake generation methods.

Keywords: Deepfake detection. Vision transformers. Face-swap. Face-reenactment.

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

DINF	Departamento de Informática
PPGINF	Programa de Pós-Graduação em Informática
UFPR	Universidade Federal do Paraná

LIST OF SYMBOLS

β	Gradient decay rate
ϵ	Numerical stability constant

CONTENTS

1 INTRODUCTION

With the ever-increasing advancement of artificial intelligence (AI) generative models in recent years, numerous applications in diverse areas such as healthcare, security, and entertainment have gained popularity. In the academic field, studies have been made to enhance and generate medical images to aid disease diagnosis and treatment. In addition, assistive technologies use these models with tools such as speech-to-text (Sand et al., 2024) and image descriptions (Fernandes et al., 2022) to support people with disabilities. In art and culture, these models are used for the restoration and preservation of art pieces (Gaber et al., 2023), the revival and preservation of endangered languages by generating written and spoken content (Mgimwa and Dash, 2024), and the translation of historical texts (Liu et al., 2023).

Despite the evident benefits that generative models bring to society, some applications can be hazardous and negatively impact people's lives. For instance, voice impersonation in calls and audio messages can facilitate scams and fraud, potentially leading to unfounded accusations. Among the possible media manipulation approaches, Deepfake is a recent class of methods that can generate synthetic human images. Highly realistic counterfeit images and videos can also be used to defame, harass, and spread misinformation, damaging the reputations of both personal and professional individuals.

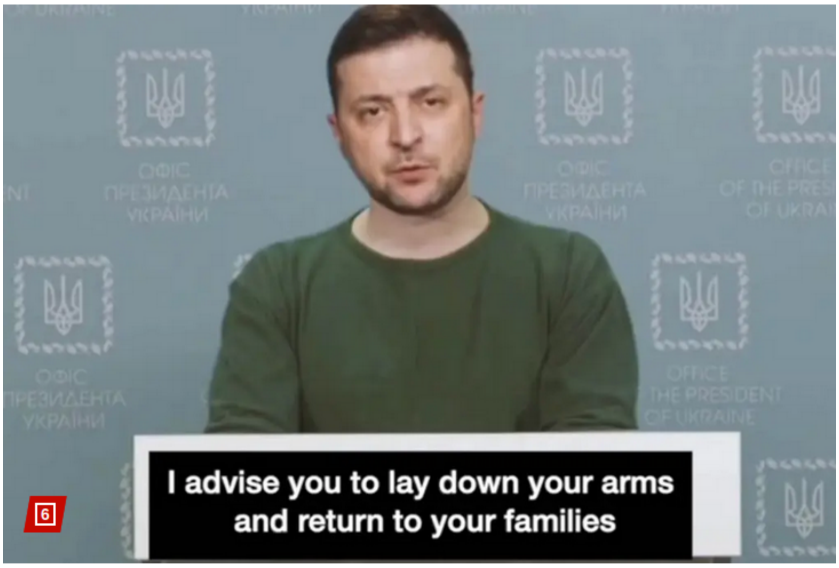

ОФІС ПРЕЗИДЕНТА УКРАЇНИ
OFFICE OF THE PRESIDENT OF UKRAINE
I advise you to lay down your arms
and return to your families

Figure 1.1: A fake video depicting Ukrainian President Volodymyr Zelensky telling his countrymen to surrender to Russia that circulated on social media and was placed on a Ukrainian news website by hackers. The video can be viewed at `https://nypost.com/2022/03/17/deepfake-video-shows-volodymyr-zelensky-telling-ukrainians-to-surrender/`.

Another significant concern is how this type of manipulation might create false beliefs in audiences through fake propaganda that affects political perceptions in electoral processes. Figure 1.1 illustrates a deep fake video depicting Ukrainian President Volodymyr Zelensky that circulated on social media and was placed on a Ukrainian news website by hackers before it was debunked and removed (Allyn, 2022). Although the video has been removed from its primary source, it is still accessible, and the impact on the masses may already have been caused by the speed at which messages are shared nowadays.

This example highlights that the malicious use of artificial intelligence to generate false data might erode trust in legitimate sources of information, as Heidari et al. (2024) observes that visual evidence no longer guarantees truth.

1.1 MOTIVATION

The duality between the benefits and dangers of deepfake models highlights the need to study methods to mitigate the risks posed by deepfakes. This is important not only because of the immediate harm these technologies can cause individuals and institutions but also because new threats appear as these technologies continue to evolve.

1.2 CHALLENGES

The primary challenges in current research are related to the rapid advancements in deepfake generation techniques, making it difficult for detection methods to keep pace with these model's evolutions. Additionally, existing detection models struggle to generalize across various types of forgery, as models trained on a specific dataset often fail to identify new or unseen deepfakes.

1.3 OBJECTIVES

Considering the harmful uses of generative models, this dissertation briefly describes strategies and methods for detecting deepfakes and explores techniques such as Convolutional Neural Networks and Vision transformers.

Thus, this work's objective is to evaluate state-of-the-art deepfake detection methods by leveraging Convolutional Neural Networks (CNNs) and Vision Transformers, performing a comparative study of their performance and practicality using the FaceForensics++ and Facebook DFDC datasets.

With the main objective outlined, the following specific objectives were defined:

- Conduct a comprehensive review of the literature to select existing methodologies, databases used, and recent advances in deep-fake detection.
- Identify limitations in current research to build the scope of the study
- Utilize the selected networks and databases to understand what models are best suited for the deepfake detection task by evaluating their performance, viability in real-world scenarios, and effectiveness in verifying the authenticity of videos.

1.4 CONTRIBUTIONS

The contributions of this work are the study of the deepfakes detection problem and the reproduction of experiments using existing models and public datasets to evaluate the effectiveness of state-of-the-art methods for detecting deepfakes. In addition, the metrics Detection Cost Function (DCF), Equal Error Rate (EER), and Half Total Error (HTER) are explored to verify which model yields the best results when evaluating prediction performance.

1.5 DOCUMENT ORGANIZATION

This work is organized into 6 chapters: Chapter 2 describes the theoretical background, and concepts such as forgery types and fingerprints, and introduces model architectures and evaluation metrics related to the deepfake detection task.

Chapter 3 presents the analysis of related works, comprising similar existing studies, and highlights the challenges faced by the authors.

Chapter 4 presents the methodology of this study, detailing the selected datasets and the specific models' architectures used.

Chapter 5, experiments, shows how the experimental setup, the dataset pre-processing, training parameters and settings and the results obtained.

Finally, in Chapter 6, the conclusions of this study and future work are discussed.

2 THEORETICAL BACKGROUND

This chapter introduces deepfake-related concepts such as forgery types used to create deepfakes and fingerprints left by manipulation techniques. Networks and metrics used throughout this study are also addressed, as they are essential to understand this topic.

2.1 DEEPFAKE FINGERPRINTS

The fundamental idea behind deepfake detection consists of the fact that a neural network during the process of generating fake content should leave a trace or anomaly embedded as a fingerprint on the manipulated data (Ciamarra et al., 2024). Most existing approaches try to expose these traces to detect potential manipulations. However, these forensic traces can be subtle and challenging to detect, particularly in cases where videos have undergone excessive compression, multiple simultaneous editing operations, or significant downsampling (Bonettini et al., 2021; Milani et al., 2012).

As seen in Figure 2.1, a fake generation algorithm introduced abnormal frequencies in the face region. By transforming images into the frequency domain, differences between real and fake faces become more apparent, helping detect deepfakes.

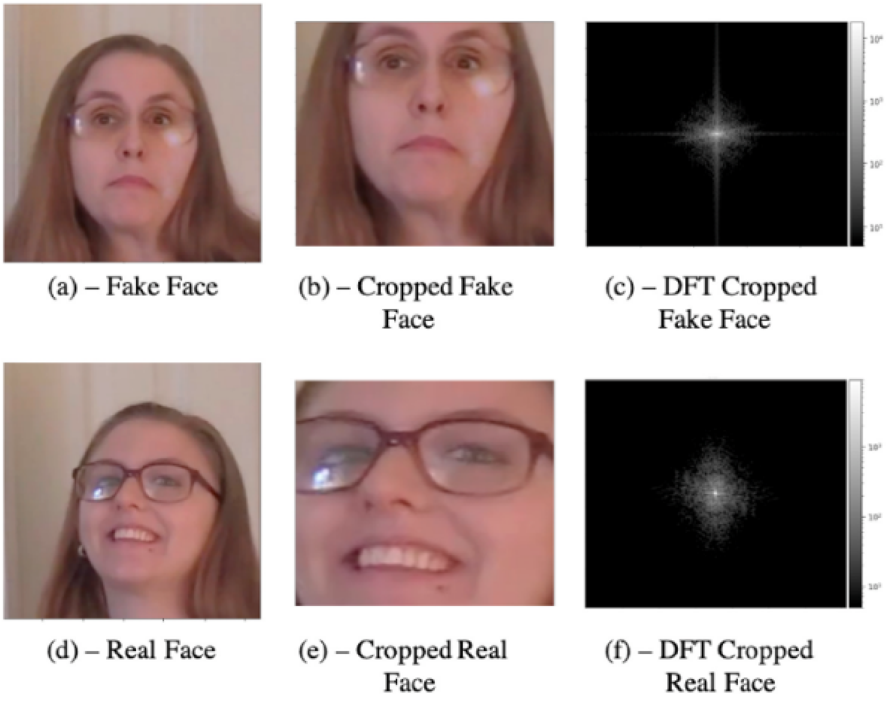


Figure 2.1: A comparison of the frequency analysis of a real and a fake sample shows the presence of abnormal frequencies (Silva et al., 2022). DFT stands for Discrete Fourier transform, a frequency domain representation of the original input sequence.

Modern techniques, such as Convolutional Neural Networks and Vision Transformers, are employed to analyze input images and extract meaningful information from them, including potential inconsistencies or anomalies, to detect deepfakes effectively.

2.2 FORGERY TYPES

The main techniques used to generate Deepfakes are generally divided into face-swapping and face-reenactment. Both techniques can be used to portray individuals behaving in a specific way, potentially misleading viewers. The DFDC dataset (Dolhansky et al., 2020) comprises videos generated using face-swapping techniques, whereas the FaceForensics++ dataset (Rossler et al., 2019) incorporates a variety of face manipulation methods, including face-swapping, face reenactment, and others.

When exploring the forgery techniques, the target and source are used to explain these methods. In general, target refers to the base video in which a face will be swapped; source refers to the source content that is used to extract the identity that will be swapped onto the target video (Dolhansky et al., 2020).

2.2.1 Face-swapping

The idea behind the face-swapping technique is to put the source face onto the target. In other words, it replaces a face in a video with someone else's face, so the person's identity in the video changes. Figure 2.2 shows an example of the application of this technique, in which the artist Kendrick Lamar uses deepfake in his video clip to impersonate other famous men.

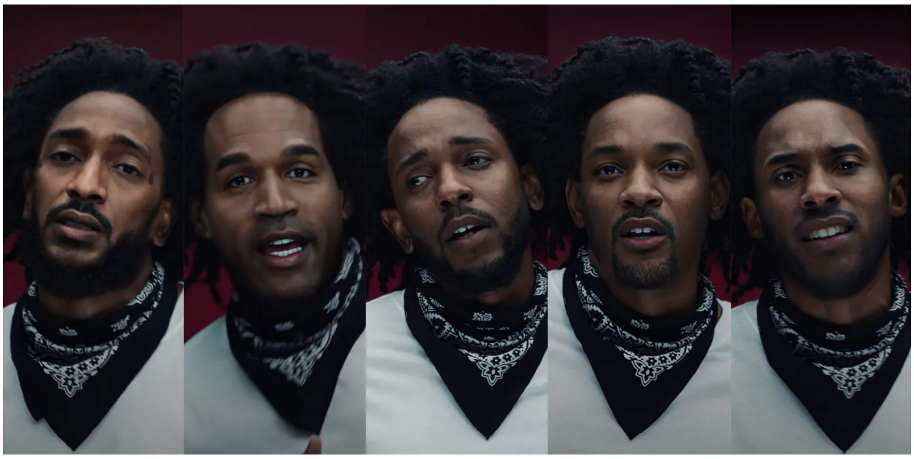


Figure 2.2: The center image shows Kendrick Lamar's real face, while the surrounding images are the faces of Nipsey Hussle, O.J. Simpson (to his left), Will Smith, and Kobe Bryant (to his right) swapped onto Kendrick Lamar's original image (Pitchfork, 2022).

2.2.2 Face-reenactment

Face-reenactment on the other hand, involves transferring the posture and expressions from the source scene to manipulate the target video while keeping the essence of the target identity unchanged (Wang et al., 2023). An example is shown in Figure 2.3

2.2.3 VISION TRANSFORMER

As explained by Chen et al. (2021a) with the scheme of Figure 2.4 of the Vision Transformer (ViT) architecture, an image is converted into a sequence of patch tokens by dividing it into patches and linearly projecting each patch into a token. A classification token (CLS) is added to the sequence, similar to BERT (Kenton and Toutanova, 2019). To incorporate positional information, which is important for vision tasks, position embeddings are added to all tokens,

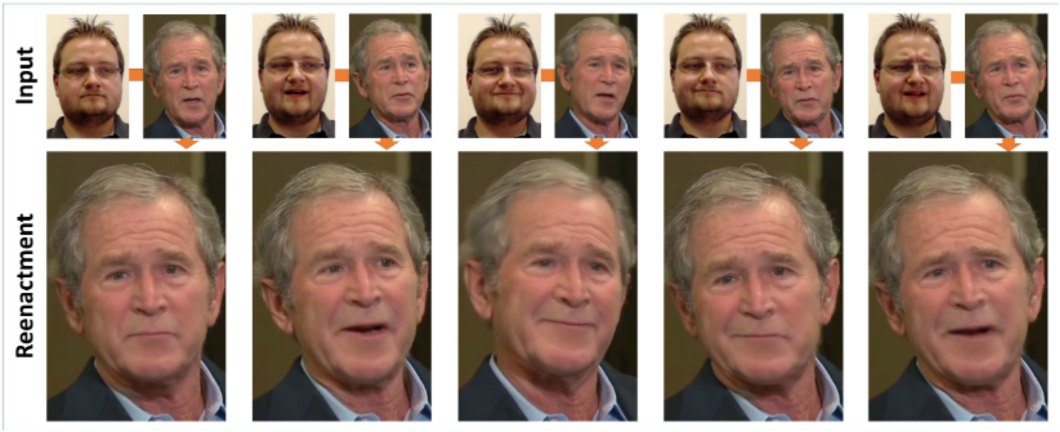


Figure 2.3: Real-time Face Capture and Reenactment of RGB Videos (Thies et al., 2016)

including the CLS token. The sequence is then processed through stacked transformer encoders, with the CLS token used for classification.

Each transformer encoder consists of blocks with multiheaded self-attention and a feed-forward network. The feed-forward network includes a two-layer multilayer perceptron with an expansion ratio, applying GELU activation after the first linear layer. Layer normalization (LN) is applied before each block and residual connections are used throughout (Chen et al., 2021a).

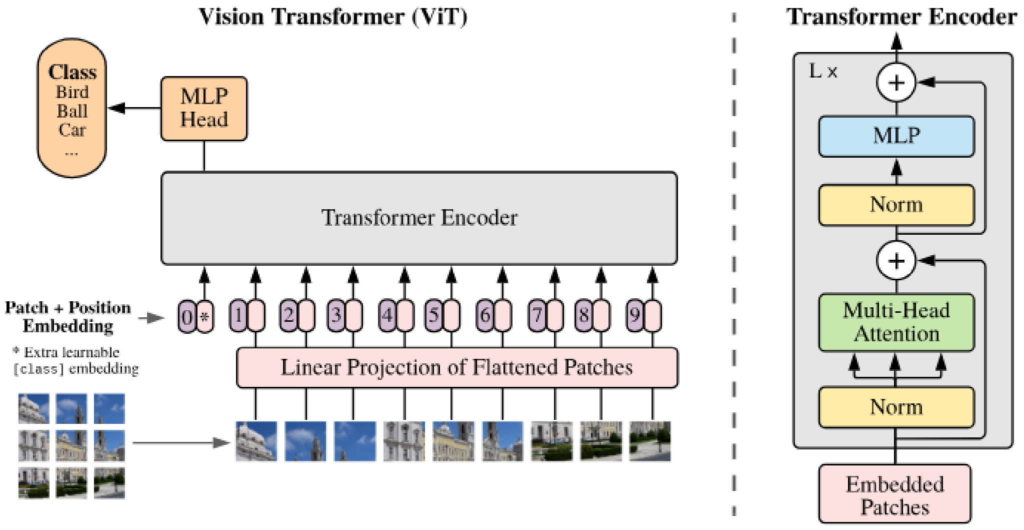


Figure 2.4: Vision transformer model overview

2.3 METRICS

This section explains the metrics selected to evaluate the models: EER, HTER, and DCF.

2.3.1 Equal Error Rate (EER)

EER is calculated by threshold when the false acceptance or positive rate (FPR) and the false rejection or negative rate (FNR) are equal in the validation set. This value indicates that a proportion of false rejections is equal to a proportion of false acceptances (Wang et al., 2023). When the EER value is lower, the accuracy of the classification algorithm is higher. The EER is defined as follows:

$$EER = FPR_{val} = FNR_{val} \tag{1}$$

In Eq. 1, FPR is the false alarm rate, FNR is the missed detection rate, and val represents the result on the verification/validation set.

The FPR is calculated as follows:

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

And the FNR is:

$$FNR = \frac{FN}{TP + FN} \tag{3}$$

Here, FP stands for false positive, FN for false negative, TN for true negative, and TP for true positive.

2.3.2 Half Total Error Rate (HTER)

This metric averages the False Acceptance Rate (FPR) and the False Rejection Rate (FNR). Mathematically:

$$HTER = \frac{FPR_{test} + FNR_{test}}{2}. \tag{4}$$

HTER is computed using the Equal Error Rate (EER) threshold, ensuring a fair comparison between false acceptance and false rejection cases. HTER provides a balanced overview of the system's performance, considering both the cases when a genuine user is incorrectly rejected and when an attacker is wrongly accepted. A lower HTER means a more robust and reliable biometric system (Kuznetsov et al., 2024).

2.3.3 Detection Cost Function (DCF)

DCF serves as a unified measure for evaluating the performance of detection models and provides insights into new advanced methods. It is defined as a weighted sum of two types of errors: miss detection P_{miss} and false alarm (acceptance) P_{fa} (Kukanov et al., 2020). The detection cost function (DCF) is given by:

$$\text{DCF}(t) = C_{\text{miss}} \cdot P_{\text{tar}} \cdot P_{\text{miss}}(t) + C_{\text{fa}} \cdot (1 - P_{\text{tar}}) \cdot P_{\text{fa}}(t), \tag{5}$$

where it depends on the decision threshold t, applied to the scores. The parameters C_{miss} (cost of a miss detection) and C_{fa} (cost of a false alarm) are usually set to one. P_{tar} is the prior probability of the target class, which takes values from $\{0.1, 0.05, 0.01\}$ (Kukanov et al., 2020).

2.4 CONCLUDING REMARKS

This chapter provided an overview of key deepfake-related concepts necessary to understand the detection techniques used in this study. It covered forensic fingerprints left by deepfake methods, forgery techniques like face-swapping and face-reenactment, evaluation metrics like Equal Error Rate (EER) and Half Total Error Rate (HTER) and Detection Cost Function (DCF).

It also introduced the Vision Transformers architecture, a state-of-the-art network used in the following chapters.

Regarding the DCF metric, this work attempts to minimize its value in validation experiments to identify the best model. By adjusting DCF parameters to penalize specific misclassifications (such as assigning higher costs to false negatives) the metric can be aligned with the priorities of a given application. In deepfake detection, false negatives are particularly concerning, as undetected deepfake content can spread widely. Therefore, monitoring DCF values and adjusting penalties allows tailoring the evaluation metric to real-world consequences.

This theoretical background establishes a basis for the related work in the field, the experiments, and the analysis in the following chapters, highlighting the challenges and progress in deepfake detection.

3 RELATED WORKS

The purpose of this chapter is to present techniques, models, and methods related to the deepfake detection task, the primary focus of this study. The subsequent sections review the work and results from related works published in the past five years. The content is organized into three sections: works based on convolutional neural networks, the ones exploring the frequency domain, and those involving fingerprint and watermark techniques. Finally, we present concluding remarks.

3.1 CONVOLUTIONAL NEURAL NETWORKS BASED WORKS

SurFake is a method proposed by Ciamarra et al. (2024) that examines how deepfake algorithms create inconsistencies in an image's original features. By analyzing surface characteristics, they generate a descriptor to train a Convolutional Neural Network (CNN) for deepfake detection using the Global Surface Descriptor (GSD). The GSD captures geometric features like cheek curvature, jawline contours, and nose structure. When a forgery algorithm alters an image to create a deepfake, it might change the GSD information, resulting in inconsistent patterns. These inconsistencies can be detected and used to determine whether an image is authentic or a deepfake. Experimental results performed on the FF++ (FaceForensics++) (Rossler et al., 2019) dataset show that using only the GSD feature to train a CNN model gives an accuracy of 75% and when it is combined with RGB frames it results in 97.75% accuracy. Despite the good results achieved by combining RGB frames with the GSD features the improvement was limited compared to the detection using RGB frames only (97.57%). Thus, evaluating the effectiveness of the GSD feature shows that it performs poorly compared to average results using RGB frames alone. This reliance on RGB images to enhance the GSD feature's performance makes the approach less effective because of the computational cost. Essentially, while the combination slightly improves detection accuracy, the added complexity and resource requirements make the technique less practical.

The work of Patel et al. (2023) presents an enhanced deep-CNN (D-CNN) used to detect deepfakes with reasonable accuracy and high generalization. The authors state that a CNN trained on one dataset may not perform well on a different dataset. To address this type of inconsistency, they propose using a D-CNN model that can interpret data from various domains while maintaining the robustness and generalizability of the deepfake detection method. This approach aims to achieve high accuracy through an effective ensemble of the proposed CNN models. The model is trained on synthetic and real images from different sources, aiming to improve the generalizability and cross-learning accuracy. The images are resized and fed into the D-CNN model, using binary cross entropy and the Adam optimizer to enhance the learning rate. The proposed architecture reaches an accuracy of 97.2% in the test dataset, considering 5 datasets for deepfake images and 2 datasets for real ones. More specifically: AttGAN (Facial Attribute Editing by Only Changing What You Want) (He et al., 2019), Group-wise deep whitening-and-coloring transformation (GDWCT) (Cho et al., 2019), StyleGAN (Karras et al., 2019), StyleGAN2 Karras et al. (2020), and StarGAN (Choi et al., 2018), used for fake images. CelebA (Large-scale Celeb Faces Attributes) (Liu et al., 2015) and Flickr Faces High Quality(FFHQ) (Karras et al., 2019) were used for real images. Thus, the model performs well over low-resolution images but slightly drops over high-resolution ones. Despite a great difference between these images' resolutions, the proposed architecture provides a well-balanced performance in all the data sources.

Zhang et al. (2020) developed a feature extraction technique based on deep learning and Error Level Analysis (ELA). The ELA method can obtain the compression distortion during lossy image compression. The local minimum in the image difference represents the original regions, and the local maximum represents tampered regions. Thus, a CNN can use this information as a feature to detect whether an image is a deepfake. The main goal is to increase the efficiency of distinguishing deepfake-generated images from real faces. Experiments show that the ELA method can improve the training efficiency of the CNN model and effectively distinguish fake facial images generated by deep learning. The accuracy obtained was 97% on the Milborrow University of Cape Town Database (MUCT) (Milborrow et al., 2010). Although the study presented good results, the proposed method only works well with compressed images in the JPEG format using lossy techniques, so detecting falsification under low-quality compression without data loss is not ideal. A downside is that it limits the method's real-world applicability, reducing its effectiveness in detecting deepfakes across various image types and compression methods.

Guarnera et al. (2020) proposed a method to detect counterfeit images generated by GANs using an Expectation-Maximization algorithm that identifies and extracts a unique "fingerprint" of the traces left by the convolutions on the generated images. When detecting these characteristics, it is possible to determine whether an image is authentic or fake. The experiments showed an accuracy of more than 98% in deep fake images generated by 10 different GAN architectures: CYCLEGAN (Park et al., 2019), STARGAN (Choi et al., 2018), ATTGAN (He et al., 2019), GDWCT (Cho et al., 2019), STYLEGAN (Karras et al., 2019), STYLEGAN2 (Karras et al., 2020), PROGAN (Karras et al., 2018), FACEFORENSICS++ (Rossler et al., 2019), IMLE (Li et al., 2019) and SPADE (Park et al., 2019). The work demonstrates that the efficiency of training models like CNNs can be improved with the ELA method and does not depend on image semantics (recognizing a person, identifying a specific object, or understanding the context of a scene). Besides that, tests on Deepfakes generated by the app FACEAPP (FaceApp, 2024) reached 93% accuracy showing the technique's efficiency in real scenarios. However, the technique faces challenges with current methods such as GANprintR (GAN-fingerprint Removal approach) (Neves et al., 2020) that can remove the "fingerprints" left by GANs.

For the movement pattern detection task in videos, Caldelli et al. (2021) proposed a technique using frame sequences along the time and Optical flow (OF) fields to train CNNs. This method distinguishes fake videos from real ones by identifying manipulations based on movement dissimilarities. The optical flow is employed to capture the movement patterns between the frames. These patterns analysis with a CNN allows the detection of structural alterations typical of different deepfake methods, providing a more robust detection. The results obtained with the FaceForensics++ dataset show that this technique is effective for distinguishing between fake videos and real ones, especially in scenarios of cross-forgery. The OF method alone is less accurate overall when compared to RGB frames, with 70.76% in the C40 (low visual quality) dataset and 88.92% in the C23 (high visual quality) dataset, while the RGB yields 97.72% accuracy in the C23 dataset and 95.38% in the C40 dataset. When combining RGB and Optical flow the performance is comparable to state-of-the-art methods that resort to separate frames of video, improving the effectiveness of individual methods, with 98.41% on C23 (high-quality version of the dataset) and 95.70% on the C40 dataset. These results show that even with the high accuracies on cross-forgery scenarios, the technique still relies on the RGB images for better results. Also, the approach depends on the consistency of optical flow fields, which may fail to detect anomalies if deepfake algorithms do not significantly alter motion patterns.

Saikia et al. (2022) also used feature extraction based on optical flow to obtain temporal data from videos. In that work, these features are fed into a hybrid model for classification. This hybrid model is based on a combination of CNN and recurrent neural networks. From this study, it has been noticed that the fake videos also have distorted movement vectors when compared to the real ones. The proposed method shows an accuracy of 66.26%, 91.21%, and 79.49% on DFDC, FF++, and Celeb-DF, respectively, with a very reduced number of samples (≤ 100 frames). This promises a preemptive detection of fake data compared to the existing models. The model shows varying performance across different datasets. For instance, it performs best on the FaceForensics++ dataset and least on the DFDC dataset. This inconsistency suggests that the model might not generalize well across different types of deepfake videos. Also, the study aims to reduce computational complexity by limiting the number of frames and the sample size. While this approach is beneficial for faster processing, it might compromise the model's ability to capture detailed temporal dynamics essential for accurate deepfake detection.

To detect and localize manipulations in a facial image, Liang et al. (2023) proposed a network composed of three parts: LSTM network, FGPM, and decoder (classifier). The LSTM network was used with characteristics of resampling to learn the correlation between different patches. In contrast, the FGPM architecture was used to learn the facial characteristics to localize the manipulation. The experiments were conducted on CelebA and FF++ datasets for training and FF++, DeeperForensics Dataset (Jiang et al., 2020), Celeb-DFv1 (Li et al., 1909), Celeb-DFv2 (Li et al., 2020), and Google Deepfake Detection(DFD) (AI, 2020) for testing. The best test result was obtained in the FF++ dataset. The proposed approach achieved an F-score of 97%, surpassing CNNDetect (85%), Xception (91%), DSP-FWA (87%), and Face X-ray (88%). Although the proposed method yielded good results, its best performance was on the same dataset it was trained on (FF++), indicating limited generalization capability. Additionally, its high computational complexity may limit its applicability in resource-constrained environments, which is particularly important for real-time applications where processing speed is critical.

3.2 WORKS BASED ON FREQUENCY DOMAIN

Wolter et al. (2022) introduces a method for detecting synthetic images using wavelet-packet representation, which captures both spatial and frequency data, unlike Fourier transforms that lose spatial information. The study reproduces the experimental setup using the spatial approach from Yu et al. (2019) and the frequency-only Discrete cosine transform (DCT) representation from Frank et al. (2020). They integrate existing Fourier-based methods with fusion networks, improving performance compared to previous Fourier or pixel-based methods on the CelebA, FFHQ, FF++, and LSUN-data (Large-scale Scene UNderstanding) (Yu et al., 2015) datasets. The best-performing network achieved 99.45% accuracy on the CelebA dataset and, despite being lightweight with only 109k parameters, it performs comparably to much larger models used by (Yu et al., 2019) (9 million parameters) and (Frank et al., 2020) (170k parameters). However, fusing Fourier and wavelet packet features does not enhance performance when including StyleGAN2 (Karras et al., 2020) and StyleGAN3 (Karras et al., 2021) generated images. The classifiers rely heavily on high-frequency information, which makes them vulnerable to methods that remove these details, such as JPEG compression. This reliance limits their effectiveness in certain practical scenarios where image quality is compromised.

Jeong et al. (2022) state that when training GAN models to detect deepfakes via frequency level information, the network is prone to overfitting. To address this, the authors designed a framework to generalize the deepfake detector by creating frequency-level perturbation maps to make the created images indistinguishable from the real ones. This process enhances the deepfake detector's ability to generalize across different GAN models by shifting focus from specific artifacts to overall image irregularities. The study performed four experiments:

manipulated face images, resized face images, unseen categories, and unseen models using datasets such as FFHQ, LSUN, CelebA, Imagenet (Russakovsky et al., 2015), COCO (Lin et al., 2014), and Deepfake dataset (Rossler et al., 2019). The model achieved 97.16% accuracy on manipulated face images and 97.8% accuracy on resized face images. Using various categories, for the unknown GAN models experiment, the accuracy was 74.93% for one training category, 75.11% for two training categories, and 79.40% for four training categories. Overall, FrePGAN demonstrated superior performance compared to other models. Not focusing on frequency-level artifacts helps the model become more versatile in detecting a wider range of deepfakes. Still, it may reduce its accuracy for the specific types of deepfakes it was originally trained to identify. Also, the proposed model involves complex transformations and alternating updates between the perturbation generator and the classifier, which could result in high computational costs and longer training times.

3.3 WORKS BASED ON FINGERPRINT AND WATERMARK

Yu et al. (2020) work allows deepfake developers to fingerprint their models for accurate detection and attribution of generated samples, enabling the regulation of generative models.

The novel technique involves an ad-hoc generation of a large population of models with distinct fingerprints. The experiments were conducted on the CelebA, LSUN Bedroom, and Cat datasets. The proposed model outperforms previous state-of-the-art models, particularly in robustness and immunizability against common image perturbations such as cropping, resizing, blurring, JPEG compression, and additive Gaussian noise. The method maintains high detection accuracy ($\geq$ 99%) for fingerprint verification, showcasing its robust and scalable approach to deepfake detection and attribution. The work relies on responsible disclosure, where developers create and share mechanisms like fingerprints to detect and attribute deep-fakes, enhancing AI security. However, its success depends on widespread adoption since without broad implementation by developers and organizations, the effectiveness of this strategy might be limited, and it remains uncertain whether this level of adoption will be achieved.

Neekhara et al. (2022) introduce FaceSigns: a deep learning-based semi-fragile watermarking system to verify the authenticity of digital images and detect facial manipulations. Unlike previous digital watermarking and steganography approaches, the method's purpose is to ensure the watermark is robust against benign image transformations (e.g., compression, color adjustments) but fragile to malicious manipulations like Deepfake transformations. Thus, the non-tampered image will contain an intact watermark, whereas a manipulated image will have a corrupted watermark. To evaluate whether an image has been manipulated, checking for the integrity of a watermark should suffice. The experiments conducted on the CelebA dataset showed that the method can detect manipulated content with an AUC score of 99.6%. The authors propose that embedding a secret verifiable message into images at the time of acquisition can establish the provenance of real images and videos, thereby addressing the limitations of Deepfake detection. However, this approach faces challenges, including the difficulty of ensuring consistent adoption and implementation across a wide range of devices, manufacturers, and platforms. Additionally, verifying provenance requires preventing attackers from detecting, removing, or altering the embedded watermark.

Huang et al. (2022) introduce the CMUA-Watermark technique to combat deepfakes using adversarial watermarks. These watermarks protect facial images from various deepfake models by disrupting their ability to generate convincing deepfakes. The study uses the CelebA and LFW (Huang et al., 2008) datasets for training and testing, and 100 randomly selected images from Films100 to evaluate real-world effectiveness. The study compares the CMUA-Watermark with state-of-the-art attack methods on CelebA. The proposed technique shows SRmask scores of 1.0000 for StarGAN (Choi et al., 2018), 0.8708 for AttGAN, and 0.9987 for HiSD(Li et al., 2021), outperforming other methods, except for AGGAN, where Momentum iterative method (MIM) (Dong et al., 2018) achieves a slightly higher SRmask of 0.9994. When applied to real social media platforms like Tantan and Jimu, images protected by the CMUA watermark fail liveness detection modules, while some StarGAN-generated images pass. These results highlight the watermark's robust effectiveness across various datasets and deepfake models. The Fréchet Inception Distance (FID) scores indicate high-quality generation with effective watermarking, especially for the proposed method with the AttGAN and HiSD models, scoring 1.8133 and 1.9672, respectively. However, the best scores for AGGAN and StarGAN are achieved by the MIM (Dong et al., 2018), with 1.8435 and 2.5281, respectively. Despite proposing a more comprehensive evaluation method, the work acknowledges the limitations of current evaluation techniques. The new metrics require further validation and comparison with other established evaluation frameworks to ensure their robustness.

Wang et al. (2022) also developed an anti-forgery method to protect shared facial images from being manipulated by deepfake models. They studied proactive defense techniques by adding adversarial noises into the source data to disrupt the deepfake manipulations, exposing artifacts that could be easily spotted even with simple deepfake detectors. These perturbations are robust enough to resist common image transformations, such as compression, Gaussian blur, and the evasion technique MagDR via image reconstruction. For the latter, Chen et al. (2021b) illustrated that a simple input reconstruction could destroy the added adversarial noises. CelebA dataset was used to create deepfakes using three deepfake models: StarGAN, AttGAN, and Fader Network (Lample et al., 2017). The best results for Attack Success Rate (ASR) and Structural Similarity Index Measure (SSIM) metrics were 100.0 and 0.251, where the proposed method outperformed the others for StarGan, AttGAN, and Fader Networks(Lample et al., 2017). For the PSNR, the model outperforms the others in AttGAN, Identity swap, and Face reenactment, with 15.975 as the best result. For the L2 norm[1], it only surpassed other methods in AttGAN with a score of 0.103. The results show that the novel technique outperforms existing ones in terms of robustness. However, the presented method lacks generalization to various deepfake techniques, raising a need for broader validation to ensure the technique's effectiveness against other forms of deepfake manipulations and emerging forgeries.

3.4 WORKS BASED ON VISION TRANSFORMERS

Wodajo and Atnafu (2021) proposed a novel DeepFake detection method by enhancing a vision transformer (ViT) model with CNN features and patch-embedding techniques, supported by a distillation method to improve accuracy. The model captures spatial and temporal video characteristics, overcoming the limitations of traditional CNN-based models. The study demonstrated that this combined architecture could achieve competitive results with an accuracy of 91. 5% and an area under the curve (AUC) of 0.91 in the DeepFake Detection Challenge Dataset (DFDC). However, its computational complexity is 8 to 10 times higher than standard ViT, which poses challenges for real-time applications and resource-limited devices. The authors emphasize the need to address these challenges and advance DeepFake detection methods.

[1] The L2 norm, also called the Euclidean norm, is defined as $\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$.

Table 3.1: Datasets used throughout the studies reviewed. Each row shows the amount of images or videos, if there are real or fake contents, their dimensions, source and popularity.

Name	Amount of images/videos	Content	Real or Fake	Dimensions	Source	Popularity
FaceForensics++ (FF++)	1000	Videos	Real & Fake	256x256	Link	2.3k stars
Celeb-df	6229	Videos	Real & Fake	256x256	Link	1506 citations, 269 stars
DFDC	128,154	Videos	Real & Fake	256x256	Link	870 citations
Vox-DeepFake	1M	Videos	Real & Fake		-	-
Deeper-Forensics (Deeper)	60000 videos, 17.6 million frames	Videos	Real & Fake		Link	22 papers, 526 stars
LFW	13,233	Images	Real	250x250	Link	-
CelebA	202,599	Images	Real	178×218	Link	3,097 papers
FFHQ	70000	Images	Real	1024×1024	Link	1,238 papers
MUCT	3,523	Images	Real	480x640	Link	233 stars

3.5 CONCLUDING REMARKS

CNN is the most used technique reviewed in this study, followed by fingerprint, watermark, and frequency domain analysis. Among the methods analyzed, the Optical Flow-based CNN achieved the highest accuracy for detecting unseen deepfake manipulations in videos, with 98.41% accuracy. The top result for image-based deepfake detection was 99.45% accuracy, using a combination of spatial and frequency features via wavelet packets.

The DFDC dataset stands out for its size and popularity, featuring 128,154 videos and cited in over 870 papers, whereas FaceForensics++ (FF++) is the most cited dataset in the reviewed articles, with 2560 citations and 1,000 videos containing both real and fake content.

This study focuses on deepfake detection in videos, as this type of media is more frequently used to spread fake content, presents greater analytical challenges due to motion, and has a stronger psychological impact on viewers compared to images. Furthermore, deepfake videos present significant threats, making the accurate detection of deepfakes in this format crucial.

Taking these conclusions into account, this work aims to implement CNN and Vision Transformer architectures trained on the FaceForensics++ and DFDC video datasets. It evaluates their performance and analyzes their effectiveness in verifying video authenticity in the following chapters.

Article	Method	Year of Publication	Dataset	Results
Ciamarra et al. (2024)	Use surface geometry features to train CNN	2024	FaceForensics++	RGB frames + GSD features: 97.75% accuracy
Patel et al. (2023)	D-CNN-based architecture	2023	CelebA, FFHQ, GDWCT, AttGAN, STARGAN, StyleGAN, StyleGAN2	97.2% accuracy
Zhang et al. (2020)	Error Level Analysis and CNN	2020	MUCT (Milborrow University of Cape Town)	97% accuracy
Guarnera et al. (2020)	Extract convolutional traces to train a CNN	2020	FaceForensics++	93% accuracy
Caldelli et al. (2021)	Optical flow fields and CNN	2021	FaceForensics++	RGB + Optical flow: 98.41% accuracy
Saikia et al. (2022)	CNN and LSTM	2022	DFDC, FF++ and Celeb-DF	66.26%, 91.21%, and 79.49% accuracy for each dataset respectively
Liang et al. (2023)	CNN and LSTM	2023	CelebA, FaceForensics++	-
Wolter et al. (2022)	Wavelet-packet analysis (space and frequency)	2022	FFHQ, LSUN-data, CelebA, FF++	99.45% accuracy
Jeong et al. (2022)	FrePGAN: GAN and CNN	2022	FFHQ, LSUN, CelebA, Imagenet, COCO and Deepfake dataset	97.16% accuracy
Yu et al. (2020)	Generate models with distinct fingerprints	2022	CelebA, LSUN Bedroom and Cat datasets	99% accuracy
Neekhara et al. (2022)	Watermark and CNN	2022	CelebA dataset	AUC score of 99.6%
Huang et al. (2022)	Adversarial watermark	2022	CelebA, LFW, Films100 (real); StarGAN, AttGAN, HiSD (fake)	SRmask of 1.0000 and FID of 1.9672
Wang et al. (2022)	Adversarial Perceptual-aware Perturbations	2022	CelebA	-
Wodajo and Atnafu (2021)	Convolutional Vision Transformer	2021	DFDC	91.5% accuracy

Table 3.2: Summary of various DeepFake detection and disruption methods, their publication years, datasets used, and results achieved.

4 METHODOLOGY

This chapter describes the methodology employed to perform the comparative study of deepfake detection methods. First, it addresses the datasets and networks used and then explains the experiments conducted in intra-dataset and cross-dataset protocols.

4.1 DATASETS

The experiments were carried out on the FF++ and DFDC datasets. 30,000 images were obtained from each dataset, totalling 60,000, where 75% was used to train the models, 15% to validate them, and 10% to test their performances. Figure 4.1 shows samples of frames of each dataset.

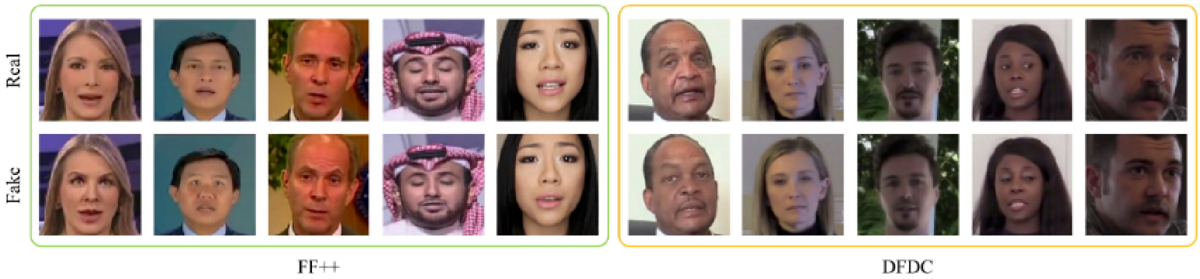


Figure 4.1: Sample faces extracted from FF++ and DFDC datasets.

4.1.1 FaceForensics++

This is a large-scale facial manipulation dataset generated using state-of-the-art automated video editing methods (Heo et al., 2023). It comprises 1000 original and fake videos generated through different deepfake generation techniques. For this work, the sub-dataset Face2Face was used.

Face2Face is a facial reenactment system that transfers expressions from a source video to a target video while preserving the target person's identity (Rossler et al., 2019).

4.1.2 DFDC

The DFDC dataset is a large and publicly available face-swap video dataset, with more than 120,000 total clips sourced from 3,426 paid actors, produced with several Deepfake, GAN-based methods (Dolhansky et al., 2020). In this study, the FaceSwap method from this dataset was used.

4.2 NETWORKS

The neural networks employed in this study were chosen based on their performance, availability of implementations, and widespread adoption within the research community. Xception and EfficientNet were selected as baseline models, while Vision Transformers (ViTs) were utilized to represent the state-of-the-art approach.

4.2.1 XCEPTION NET

Xception is a convolutional neural network built entirely with separable convolution layers in depth. It is based on the idea that cross-channel correlations and spatial correlations in feature maps can be handled separately. This concept extends the approach used in the Inception architecture, leading to the name Xception, which stands for "Extreme Inception" (Chollet, 2017).

The network consists of 36 convolutional layers organized into 14 modules. All modules, except the first and last, use linear residual connections. For image classification tasks, the convolutional base is followed by a logistic regression layer. Optionally, fully connected layers can be added before this layer, as explored in the experimental results (Chollet, 2017).

A complete description of the specifications of the network is given in figure 4.2

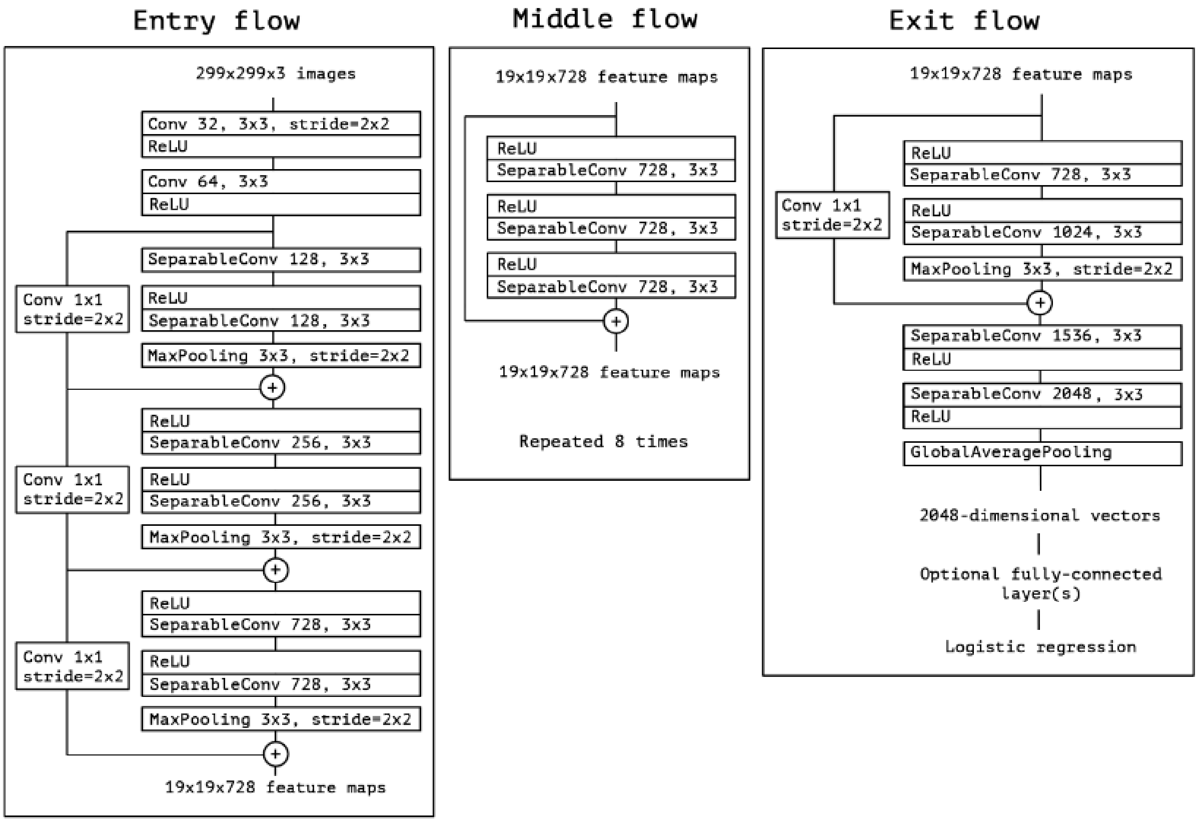


Figure 4.2: The Xception architecture: the data first goes through the entry flow, then through the middle flow which is repeated eight times, and finally through the exit flow (Chollet, 2017).

The implementation used here follows the pipeline described by di Milano Image and Lab (2020) [1]

4.2.2 EFFICIENTNET AND ATTENTION MECHANISM

The EfficientNetB4 architecture [2] , depicted in the blue block of Figure 4.3, processes a color image I (the face extracted from a video frame) as input. The network outputs a 1792-element feature vector, $f(I)$. The final score associated with the face is obtained through a classification layer. This variant of the EfficientNetB4 architecture incorporates attention mechanisms, as proposed by (Bonettini et al., 2021), enabling the neural network to focus on the most relevant

[1] The pre-trained weights are available at https://paperswithcode.com/model/xception?variant=xception-1

[2] The code implementation is available at https://github.com/polimi-ispl/icpr2020dfdc

parts of the input for deepfake detection (Bonettini et al., 2021). The implementation steps for the attention mechanism are as follows:

- Select the feature maps extracted by the Efficient-NetB4 up to a certain layer, such that these features provide sufficient information on the input frame without being too detailed or unrefined. To this purpose, the output features at the third MBConv block were selected, which have a size of 28×28×56 (Bonettini et al., 2021);
- Process the feature maps with a single convolutional layer with kernel size 1 followed by a Sigmoid activation function to obtain a single attention map (Bonettini et al., 2021);
- Multiply the attention map for each feature map at the selected layer (Bonettini et al., 2021).

The attention-based module is depicted in the red block of Figure 4.3.

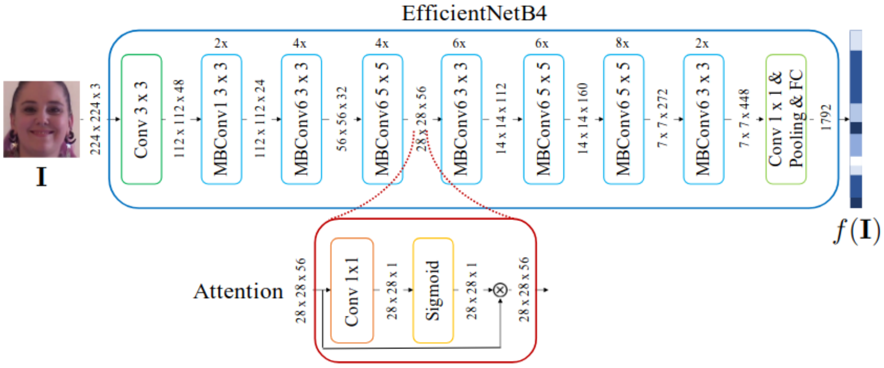


Figure 4.3: EfficientNetB4 and attention mechanism.

This mechanism enables the network to focus only on the most relevant portions of the feature maps; moreover, it provides a deeper insight into which parts of the network's input are assumed to be the most informative (Bonettini et al., 2021).

4.2.3 CROSS EFFICIENT VISION TRANSFORMERS

The Convolutional Cross ViT architecture combines features of the Efficient ViT (Coccomini et al., 2022) and multi-scale Transformer architectures (Chen et al., 2021a). As shown in Figure 4.4, it has two branches: the S-branch for smaller patches and the L-branch for larger patches to capture a wider view. Visual tokens from the Transformer Encoders in both branches interact via cross-attention. The CLS tokens from each branch generate separate logits, which are summed, and a sigmoid function outputs the final probabilities. The architecture uses two CNN backbones. The first, EfficientNet B0, processes 7 × 7 patches in the S-branch and 54 × 54 in the L-branch. The second handles 7 × 7 patches in the S-branch and 64 × 64 in the L-branch. [3]

[3] The code implementation is available at https://github.com/davide-coccomini/Combining-EfficientNet-and-Vision-Transformers-for-Video-Deepfake-Detection

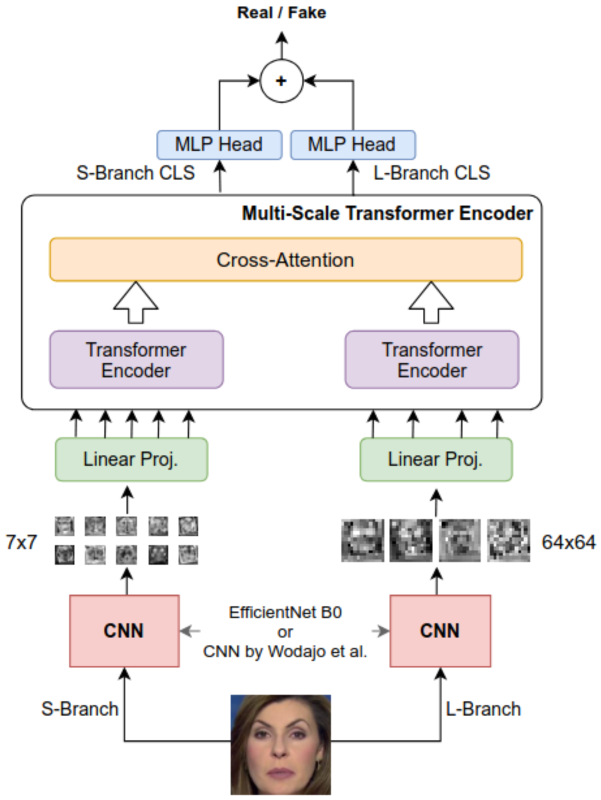


Figure 4.4: Cross efficient Vit.

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

The experimental setup uses an NVIDIA GeForce RTX 3060 GPU with 12 GiB of memory and an AMD Ryzen 9 5950X 16-core processor. The system runs on Ubuntu 20.04.6 LTS and has 125 GiB of RAM, providing plenty of capacity for large datasets and models. Python 3.7.10 is used as the programming language, along with the libraries defined by the authors in the code sources (Coccomini et al., 2022) (di Milano Image and Lab, 2020)

5.2 METHODOLOGY

The Xception, EfficientNetB4, and CrossEfficientVit models were trained using the DFDC and FF++ datasets, and the experiments were carried out as follows:

- Pre-process the train, validation, and test datasets;
- Train and evaluate the Xception model;
- Train and evaluate the EfficientNet model;
- Train and evaluate the Cross Efficient Vit model.

5.3 DATA PRE-PROCESSING

To prepare the datasets for model training, the videos from both datasets were downloaded, and the faces were obtained using a dedicated face detection and extraction script. The extracted face images from each dataset were then partitioned into training, validation, and test sets. For each dataset, 30,000 images were generated. 75% of the samples were used for training, 15% for validation, and 10% for testing.

5.4 TRAINING AND TESTING PROTOCOLS

The three models were trained and tested using intra-dataset, cross-dataset, and fusion-dataset protocols in both DFDC and FF++.

Intra-dataset means that a model is trained and tested in the same dataset, and a cross-dataset setting, a model is trained in one dataset and tested in another one. The fusion-dataset protocol involves creating a combined dataset from the original ones, ensuring a more diverse training set. This approach aims to enhance the model's generalization capabilities by exposing it to a broader variety of data patterns. The protocols are illustrated in Figure 5.1 using the Cross Efficient Vit architecture as an example.

The performance of each model on the validation set was evaluated by monitoring the validation accuracy, the equal error rate (EER), the half-total error rate (HTER), and the detection cost function (DCF) in the training phase. Once reasonable values for these metrics were observed over several epochs, the models were evaluated on the test sets according to the same metrics.

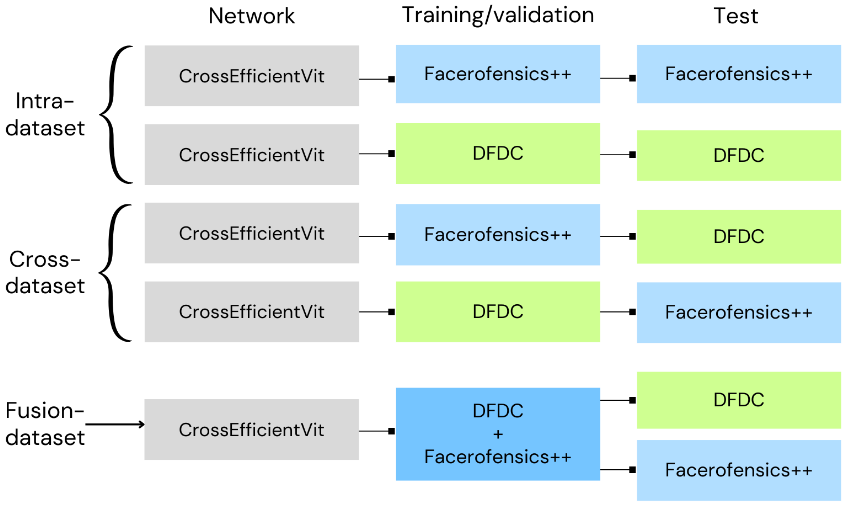


Figure 5.1: Intra-dataset, Cross-dataset, and Fusion-dataset protocols for evaluating the models' performances.

5.5 TRAINING PARAMETERS AND SETTINGS

5.5.1 Xception

The XceptionNet is configured using the same approach as the author's. It is trained on ImageNet using separable convolutions with residual connections and adapted for this task by replacing the final fully connected layer with two outputs. The remaining layers are initialized with ImageNet weights. All previous layers are frozen to configure the new fully connected layer, and the network is pre-trained for three epochs. It is then trained for an additional 15 epochs, with the best model selected based on the previously mentioned evaluated metrics. The binady cross-entropy loss was used as the objective during training. The network was optimized end-to-end using the Adam optimizer with a learning rate of 10^{-4}

5.5.2 EfficientNet

During training and validation, data augmentation is applied to enhance model robustness using the Albumentations library. The model parameters are the same as those used by the authors, being trained using the Adam optimizer, an initial learning rate of 10^{-5} and a binary cross-entropy loss with logits.

Other parameters were experimented with, but the model's performance was improved only by increasing the number of iterations from 20,000 to 30,000. The model processes batches of 32 faces evenly split between real and fake or training and is stopped earlier if validation loss stabilizes.

5.5.3 CrossEfficient VIT

The architecture exploits internally the EfficientNet-Pytorch[1] and ViT-Pytorch [2] repositories (Coccomini et al., 2022). The standard binary cross-entropy loss was used as the objective during training. The network was optimized end-to-end, using an SGD optimizer with a learning rate of

[1] https://github.com/lukemelas/EfficientNet-PyTorch

[2] https://github.com/lucidrains/vit-pytorch/tree/main/vit_pytorch

0.01. During training, data augmentation was performed using the Albumentations library, and common transformations such as the introduction of blur, Gaussian noise, transposition, rotation, and various isotropic resizes were applied.

5.6 RESULTS

Tables 4.1 and 4.2 report the results of the experiments performed in this study.

Protocol	**Model**	**Accuracy**	**EER**	**HTER**	**DCF**
DFDC → DFDC	Xception	0.96	0.65	0.04	0.14
	EfficientNet	0.97	1.10	0.03	0.12
	EfficientNet + VIT	**0.98**	**0.27**	**0.007**	**0.02**
FF++ → FF++	Xception	0.86	0.68	0.13	0.50
	EfficientNet	0.87	0.12	0.13	0.47
	EfficientNet + VIT	0.90	0.57	11.0	0.04

Table 5.1: Performance metrics for trained models under intra-dataset protocols.

Protocol	**Model**	**Accuracy**	**EER**	**HTER**	**DCF**
FF++ → DFDC	Xception	0.54	0.42	0.45	2.80
	EfficientNet	0.51	0.57	**0.35**	3.50
	EfficientNet + VIT	**0.49**	**0.03**	0.58	**2.22**
DFDC → FF++	Xception	0.50	0.50	0.49	3.50
	EfficientNet	0.49	0.49	0.49	3.42
	EfficientNet + VIT	0.50	0.93	11.0	2.41

Table 5.2: Performance metrics for models trained under cross-dataset protocols.

In general, the metrics for the cross-dataset protocols are close to 0.5, which is the performance of a random classifier. This indicates that the models struggle to generalize across datasets. This makes sense once they were trained in a specific dataset and tested in a different one. Hence, it is noticeable how difficult it is for the models to classify unseen data correctly, considering that they learned one type of method, for example, face-reenactment, and need to analyze deepfakes generated with face-swap.

This raises the need to investigate whether combining both deep-fake methods may create models that can better generalize to unseen images and methods. Therefore, in order to achieve a better generalization, the Faceforensics++ and DFDC datasets were combined to explore the fusion-dataset protocol. After training the models with the new dataset, the tests were performed on the same test partitions as the previous experiments. The results are presented in tables 5.3 and 5.4.

Dataset	**Model**	**Accuracy**	**EER**	**HTER**	**DCF**
FF++ and DFDC	Xception	0.95	0.67	0.05	0.17
	EfficientNet	0.95	0.80	0.04	0.11
	EfficientNet + VIT	**0.98**	**0.38**	**0.02**	**0.08**

Table 5.3: Performance metrics for model trained under fusion-dataset protocol and tested on DFDC dataset

Dataset	Model	Accuracy	EER	HTER	DCF
FF++ and DFDC	Xception	0.90	0.51	0.10	0.45
	EfficientNet	0.90	0.65	**0.10**	0.39
	EfficientNet + VIT	**0.98**	**0.47**	0.12	**0.08**

Table 5.4: Performance metrics for model trained under fusion-dataset protocol tested on Faceforencics++ dataset

5.6.1 CONCLUDING REMARKS

The EfficientNet + ViT model generally performs the best across intra-dataset protocol, achieving the highest accuracy of 0.98 and the lowest EER (0.27), HTER(0.007), and DCF (0.02) in the DFDC protocol. The results for the cross-dataset protocol show that the models are not capable of generalization, showing a performance comparable to random classifiers, with 0.49 acuracy, 0.03 EER, 0.58 HTER and 2.22 DCF being the best results for the EfficientNet + VIT trained on the FF++ dataset and tested on the DFDC. Conversely, the results for the fusion dataset strategy show good test results in both DFDC and FF++ datasets, showing good signs of generalization. This is probably due to the fact that the models learned both face reenactment and face swap deep-fake generation methods.

6 CONCLUSIONS AND DISCUSSIONS

Deepfake detection in videos is more commonly used to spread fake content, posing greater analytical challenges due to motion, and has a stronger psychological impact than images. Hence, accurate detection is crucial due to the significant threats the malicious use of this technology poses. This study compared deepfake detection models, including Xception, EfficientNet, and Vision Transformers trained on DFDC and Facefoensics++ datasets. Key metrics were used to perform this evaluation: Accuracy, Equal Error Rate (EER), Half Total Error Rate (HTER), and Detection Cost Function (DCF). Regarding the use of the DCF metric, this work explored the development of models that accurately classify deepfakes, particularly in scenarios where impostor attempts pose more significant risks than genuine ones. This is crucial for preventing fraud and misinformation. Therefore, during the experiments, the false acceptance rate was minimized by choosing suitable parameters to calculate the DCF metric.

The results highlight significant variations in model performance depending on the training protocol used. Models trained in a cross-dataset protocol exhibited generalization challenges, with accuracy around 0.5, EER of 0.03, HTER of 0.58, and DCF of 2.22. It is noticeable how they struggled to achieve high detection performance when tested on unseen data. This limitation underscores the difficulty of adapting deepfake detection models to diverse real-world scenarios. In contrast, models trained using intra-dataset protocols demonstrated superior performance, benefiting from consistency in training and testing data, with the best result of EfficientNet + VIT: 0.98 accuracy, 0.27 EER, 0.007 HTER and 0.02 DCF. Additionally, the fusion dataset approach (combining DFDC and Faceforensics++ datasets during training) showed reasonable robustness, avoiding overfitting and improving detection capabilities across the different deepfake methods used.

These findings highlight the need for effective training strategies to enhance the reliability and adaptability of deepfake detection models applied to systems such as media forensics, law enforcement, and financial security. Future work may focus on enhancing cross-dataset and fusion dataset generalization through advanced data augmentation techniques and using larger, more diverse training datasets, including imbalanced ones.

REFERENCES

AI, G. (2020). Contributing data to deepfake detection research. Accessed: Jan 29, 2025.

Allyn, B. (2022). Deepfake video of zelenskyy could be 'tip of the iceberg' in info war, experts warn. *NPR*. Accessed: 2024-12-26.

Bonettini, N., Cannas, E. D., Mandelli, S., Bondi, L., Bestagini, P., and Tubaro, S. (2021). Video face manipulation detection through ensemble of cnns. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 5012–5019.

Caldelli, R., Galteri, L., Amerini, I., and Del Bimbo, A. (2021). Optical flow based cnn for detection of unlearnt deepfake manipulations. *Pattern Recognition Letters*, 146:31–37.

Chen, C.-F. R., Fan, Q., and Panda, R. (2021a). Crossvit: Cross-attention multi-scale vision transformer for image classification. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 357–366.

Chen, Z., Xie, L., Pang, S., He, Y., and Zhang, B. (2021b). Magdr: Mask-guided detection and reconstruction for defending deepfakes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9014–9023.

Cho, W., Choi, S., Park, D. K., Shin, I., and Choo, J. (2019). Image-to-image translation via group-wise deep whitening-and-coloring transformation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10639–10647.

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797.

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ciamarra, A., Caldelli, R., Becattini, F., Seidenari, L., and Del Bimbo, A. (2024). Deepfake detection by exploiting surface anomalies: the surfake approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1024–1033.

Coccomini, D. A., Messina, N., Gennaro, C., and Falchi, F. (2022). Combining efficientnet and vision transformers for video deepfake detection. In Sclaroff, S., Distante, C., Leo, M., Farinella, G. M., and Tombari, F., editors, *Image Analysis and Processing – ICIAP 2022*, pages 219–229, Cham. Springer International Publishing.

di Milano Image, P. and Lab, S. P. (2020). Icpr 2020 dfdc solution. `https://github.com/polimi-ispl/icpr2020dfdc`. Accessed: 2025-01-19.

Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., and Ferrer, C. C. (2020). The deepfake detection challenge (dfdc) dataset. *arXiv preprint arXiv:2006.07397*.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. (2018). Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193.

FaceApp (2024). Faceapp. `https://www.faceapp.com/`. Accessed: 2024-07-02.

Fernandes, D. L., Ribeiro, M. H. F., Cerqueira, F. R., and Silva, M. M. (2022). Describing image focused in cognitive and visual details for visually impaired people: An approach to generating inclusive paragraphs. *arXiv preprint arXiv:2202.05331*.

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., and Holz, T. (2020). Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, pages 3247–3258. PMLR.

Gaber, J. A., Youssef, S. M., and Fathalla, K. M. (2023). The role of artificial intelligence and machine learning in preserving cultural heritage and art works via virtual restoration. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, X-1/W1-2023:185–190.

Guarnera, L., Giudice, O., and Battiato, S. (2020). Fighting deepfake by exposing the convolutional traces on images. *IEEE Access*, 8:165085–165098.

He, Z., Zuo, W., Kan, M., Shan, S., and Chen, X. (2019). Attgan: Facial attribute editing by only changing what you want. *IEEE transactions on image processing*, 28(11):5464–5478.

Heidari, A., Jafari Navimipour, N., Dag, H., and Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 14(2):e1520.

Heo, Y.-J., Yeo, W.-H., and Kim, B.-G. (2023). Deepfake detection algorithm based on improved vision transformer. *Applied Intelligence*, 53(7):7512–7527.

Huang, G. B., Mattar, M., Berg, T., and Learned-Miller, E. (2008). Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*.

Huang, H., Wang, Y., Chen, Z., Zhang, Y., Li, Y., Tang, Z., Chu, W., Chen, J., Lin, W., and Ma, K.-K. (2022). Cmua-watermark: A cross-model universal adversarial watermark for combating deepfakes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 989–997.

Jeong, Y., Kim, D., Ro, Y., and Choi, J. (2022). Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pages 1060–1068.

Jiang, L., Li, R., Wu, W., Qian, C., and Loy, C. C. (2020). Deeperforensics-1.0: A large-scale dataset for real-world face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karras, T., Aila, T., Laine, S., and Lehtinen, J. (2018). Progressive growing of gans for improved quality, stability, and variation. arxiv 2017. *arXiv preprint arXiv:1710.10196*, pages 1–26.

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., and Aila, T. (2021). Alias-free generative adversarial networks. *Advances in neural information processing systems*, 34:852–863.

Karras, T., Laine, S., and Aila, T. (2019). A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., and Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119.

Kenton, J. D. M.-W. C. and Toutanova, L. K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of naacL-HLT*, volume 1, page 2. Minneapolis, Minnesota.

Kukanov, I., Karttunen, J., Sillanpää, H., and Hautamäki, V. (2020). Cost sensitive optimization of deepfake detector. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1300–1303. IEEE.

Kuznetsov, O., Zakharov, D., Frontoni, E., Maranesi, A., and Bohucharskyi, S. (2024). Cross-database liveness detection: Insights from comparative biometric analysis. *arXiv preprint arXiv:2401.16232*.

Lample, G., Zeghidour, N., Usunier, N., Bordes, A., Denoyer, L., and Ranzato, M. (2017). Fader networks: Manipulating images by sliding attributes. *Advances in neural information processing systems*, 30.

Li, K., Zhang, T., and Malik, J. (2019). Diverse image synthesis from semantic layouts via conditional imle. 2019 ieee. In *CVF International Conference on Computer Vision (ICCV)*, pages 4219–4228.

Li, X., Zhang, S., Hu, J., Cao, L., Hong, X., Mao, X., Huang, F., Wu, Y., and Ji, R. (2021). Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8639–8648.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (1909). A large-scale challenging dataset for deepfake forensics (2019). *URL http://arxiv. org/abs/1909.12962*, 35:36.

Li, Y., Yang, X., Sun, P., Qi, H., and Lyu, S. (2020). Celeb-df: A large-scale challenging dataset for deepfake forensics. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3207–3216.

Liang, P., Liu, G., Xiong, Z., Fan, H., Zhu, H., and Zhang, X. (2023). A facial geometry-based detection model for face manipulation using cnn-lstm architecture. *Information Sciences*, 633:370–383.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Liu, C., Wang, D., Zhao, Z., Hu, D., Wu, M., Zhang, H., Lin, L., Liu, J., Shen, S., Li, B., et al. (2023). Sikugpt: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. *ACM Journal on Computing and Cultural Heritage*.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Mgimwa, P. A. and Dash, S. R. (2024). *Reviving Endangered Languages: Exploring AI Technologies for the Preservation of Tanzania's Hehe Language*, pages 23–33. Springer Nature Singapore, Singapore.

Milani, S., Fontani, M., Bestagini, P., Barni, M., Piva, A., Tagliasacchi, M., and Tubaro, S. (2012). An overview on video forensics. *APSIPA Transactions on Signal and Information Processing*, 1:e2.

Milborrow, S., Morkel, J., and Nicolls, F. (2010). The MUCT Landmarked Face Database. *Pattern Recognition Association of South Africa*. `http://www.milbo.org/muct`.

Neekhara, P., Hussain, S., Zhang, X., Huang, K., McAuley, J., and Koushanfar, F. (2022). Facesigns: semi-fragile neural watermarks for media authentication and countering deepfakes. *arXiv preprint arXiv:2204.01960*.

Neves, J. C., Tolosana, R., Vera-Rodriguez, R., Lopes, V., Proença, H., and Fierrez, J. (2020). Ganprintr: Improved fakes and evaluation of the state of the art in face manipulation detection. *IEEE Journal of Selected Topics in Signal Processing*, 14(5):1038–1048.

Park, T., Liu, M.-Y., Wang, T.-C., and Zhu, J.-Y. (2019). Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346.

Patel, Y., Tanwar, S., Bhattacharya, P., Gupta, R., Alsuwian, T., Davidson, I. E., and Mazibuko, T. F. (2023). An improved dense cnn architecture for deepfake image detection. *IEEE Access*, 11:22081–22095.

Pitchfork (2022). How kendrick lamar's "the heart part 5" video subverts deepfake technology. Accessed: 2025-01-29.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., and Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252.

Saikia, P., Dholaria, D., Yadav, P., Patel, V., and Roy, M. (2022). A hybrid cnn-lstm model for video deepfake detection by leveraging optical flow features. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

Sand, C., Svensson, I., Nilsson, S., Selenius, H., and Fälth, L. (2024). Speech-to-text intervention to support text production for students with intellectual disabilities. *Disability and Rehabilitation: Assistive Technology*, pages 1–8.

Silva, S. H., Bethany, M., Votto, A. M., Scarff, I. H., Beebe, N., and Najafirad, P. (2022). Deepfake forensics analysis: An explainable hierarchical ensemble of weakly supervised models. *Forensic Science International: Synergy*, 4:100217.

Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., and Nießner, M. (2016). Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395.

Wang, C., Sharifzadeh, H., Varastehpour, S., and Ardekani, I. (2023). Analysis and comparison of deepfakes detection methods for cross-library generalisation. In *2023 20th Annual International Conference on Privacy, Security and Trust (PST)*, pages 1–7.

Wang, R., Huang, Z., Chen, Z., Liu, L., Chen, J., and Wang, L. (2022). Anti-forgery: Towards a stealthy and robust deepfake disruption attack via adversarial perceptual-aware perturbations. *arXiv preprint arXiv:2206.00477*.

Wodajo, D. and Atnafu, S. (2021). Deepfake video detection using convolutional vision transformer. *arXiv preprint arXiv:2102.11126*.

Wolter, M., Blanke, F., Heese, R., and Garcke, J. (2022). Wavelet-packets for deepfake image analysis and detection. *Machine Learning*, 111(11):4295–4327.

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., and Xiao, J. (2015). Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*.

Yu, N., Davis, L. S., and Fritz, M. (2019). Attributing fake images to gans: Learning and analyzing gan fingerprints. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7556–7566.

Yu, N., Skripniuk, V., Chen, D., Davis, L., and Fritz, M. (2020). Responsible disclosure of generative models using scalable fingerprinting. *arXiv preprint arXiv:2012.08726*.

Zhang, W., Zhao, C., and Li, Y. (2020). A novel counterfeit feature extraction technique for exposing face-swap images based on deep learning and error level analysis. *Entropy*, 22(2):249.

Glossary

ASR Attack Success Rate. 24

CNN Convolutional Neural Networks. 20–22, 24

cross-forgery The term cross-forgery indicates when a model trained on a specific forgery is required to work against another unknown one. In general, state-of-the-art Deepfakes video detection methods are based on static frame features that though well-performing when trained on a specific kind of attack (same-forgery scenario), show bad performances in a cross-forgery scenario.. 21

FGPM Facial Geometry Prior Module. 22

FID Fréchet Inception Distance . 24

GAN Generative Adversarial Networks. 22

GSD Global Surface Descriptor. 20

LSTM Long Short-Term Memory. 22

MIM Momentum iterative method. 24

Optical flow Optical flow is the pattern of apparent motion of image objects between two consecutive frames caused by the movement of an object or camera.. 21

PSNR Peak Signal to Noise Ratio. 24

SRmask Success Rate of Masked Images. 24

SSIM Structural Similarity Index Measure. 24