American University of Armenia

Akian College of Science & Engineering

Bachelor of Science in Data Science

Deepfake Detection with Frequency-Enhanced Self-Blended Images

Author: Arevik Papikyan

Supervisor: Varduhi Yeghiazaryan, PhD

May, 2025

Abstract

Deepfakes are synthetically manipulated media that replace one person's appearance with another, posing a growing threat to digital integrity, personal security, and public trust. These realistic forgeries are used for misinformation, identity theft, and reputational harm. While most existing models operate on standard RGB or BGR images, this research explores alternative color spaces as potential enhancements. We employ the Frequency-Enhanced Self-Blended Images (FSBI) model due to its effectiveness and novelty in the field. Experimental results on benchmark datasets, FaceForensics++ and Celeb-DF-v2, reveal that while HSV underperforms compared to RGB, YCbCr may still offer improved detection accuracy for certain manipulation types. These findings suggest that color space transformations influence model generalization and robustness.

Contents

Chapter 1

Introduction

Deepfakes, media that have been digitally manipulated to replace a person's representation with that of another, are often used maliciously. They appear to depict situations or actions that did not happen in reality. They are created and exploited for the purpose of misinformation, identity fraud, and reputational damage. Thus, they pose a significant threat to digital security and public trust. With recent technology advancements, they have become more realistic, making manual detection almost impossible. Hence, detecting deepfake content accurately is crucial. The consequences of undetected deepfake content can be severe, from influencing public opinion with fabricated political statements to impersonating individuals for financial gain.

As the fields of computer vision and deep learning continue to advance rapidly, the capabilities of deepfake generation methods improve to produce more convincing media. Deepfake detection methods need continuous research because of the urgency for effective and adaptive detection techniques. This ongoing technological race creates the need for research to ensure that detection methods keep pace with evolving deepfake generators. This work seeks to add to this expanding field of research.

The literature review (Chapter 2) shows that papers in the deepfake detection field typically use the RGB (red, green, blue) or BGR (blue, green, red) color space. That means the input to the detection architecture is an RGB/BGR image or its channels. Therefore, we explore other color spaces to see if they help to improve the deepfake detection process. HSV (hue, saturation, value) is considered the main experimental color space. As a base model for our experiments, FSBI (Frequency-Enhanced Self-Blended Images) [16] is chosen because of its robustness, effectiveness, and novelty.

Chapter 2

Literature Review

In recent years, the rapid rise of synthetic media, especially deepfakes, has prompted research into trustworthy detection techniques. A considerable part of the existing literature has focused on detecting deepfake content through innovative analyses of visual cues, convolutional traces, and frequency information. This chapter reviews a range of studies, exploring the latest approaches, common challenges, and emerging trends in deepfake detection research.

2.1 Deepfake Detection Approaches

Early work in deepfake detection explored the residual artifacts left by generative models. For example, Guarnera et al. [6] utilized subtle convolutional inconsistencies to differentiate between authentic and manipulated images. That line of research underscored the value of capturing low-level spatial features and convolutional patterns that were frequently disrupted during deepfake synthesis. Later, Joshi and Nivethitha [9] replaced traditional convolutions with the Xception model, which embodied a fundamental shift in convolutional neural network (CNN) design by embracing the concept of depthwise separable convolutions. The Xception model turned out to be more efficient and powerful.

A recurring theme in recent literature has been exploring frequency-domain features as a complementary signal to spatial information. Works such as [7] and [10] proposed techniques incorporating frequency analysis to improve model generalization.

Beyond convolutional and frequency cues, other approaches have leveraged graph-based learning to capture unnatural visual distortions, inconsistent lighting, or irregularities in subtle features that break the natural coher-

ence of the real image. In [2], the authors proposed a method that modeled the inherent relationships in facial structures through a multi-graph attention mechanism. Similarly, Samrouth et al. [17] employed paired network architectures to detect identity inconsistencies, particularly in impersonation scenarios. The input to the architecture was a reference image of a specific individual and an image that must be verified as real or fake. By ensuring that both the reference and the suspect images were processed the same way, the network could accurately learn and detect subtle differences, enabling it to determine whether the second image was a genuine representation of the individual or a deepfake.

The advent of diffusion models also influenced detection strategies. For instance, Chen et al. [3] demonstrated that guided stable diffusion can be used to generate enhanced training samples, improving the ability of the detection model to generalize to unseen manipulations. Their approach was built on the idea that deepfake images contain information from both source and target identities, whereas authentic faces maintain a consistent identity. Based on that principle, they proposed a novel framework that reversed the face forgery generation process, guiding the detection model to learn more robust and identity-aware features for improved cross-domain generalization.

2.2 Common Findings & Challenges

A significant challenge across the literature has been the issue of generalization. Many studies, including the works by Liu et al. [13] and Abbasi et al. [1], highlighted that while many detection models achieve high accuracy on known datasets, their performance often degrades when exposed to novel deepfake generation techniques or adversarial attacks. That problem has motivated researchers to develop approaches that leverage multi-scale features, combine spatial and frequency signals, etc.

Several works emphasize the importance of integrating features across multiple scales. For example, the study by Gu et al. [5] shows that combining low-level (pixel and frequency) features with higher-level semantic cues can significantly boost performance. This fusion of complementary information helps detect subtle manipulations that might escape methods focusing on a single feature domain.

In addition, trustworthiness has been a central concern. Jiang [8] called attention to the need for models that were not only accurate but also robust against intentional manipulation or adversarial attacks. These studies advocated for comprehensive evaluations that include resilience metrics, ensuring that detection systems remain reliable under real-world conditions.

The evaluation of deepfake detection models has benefited from the availability of standardized datasets. FaceForensics++ [16] and Celeb-DF [12], for example, have frequently been used as benchmarks for comparing different methods. Systematic reviews, such as the one by Rana et al. [15], consolidated findings from multiple studies and provided insights into performance trends, evaluation protocols, and open research questions. These benchmarks served as critical tools for understanding the strengths and limitations of various detection approaches.

Despite significant advancements, several challenges remain open. The continual improvement of generative models demands adaptive detection strategies that can respond to new forms of manipulation. Accordingly, the literature reflects a dynamic and rapidly evolving field. While early methods focused on exploiting convolutional artifacts, more recent approaches have successfully incorporated frequency-domain features, self-blending techniques, and advanced network architectures to address generalization and robustness challenges. As deepfake generation continues to advance, integrating complementary features and comprehensive evaluation frameworks will be critical for developing truly reliable detection systems.

Chapter 3

Experiments

We conducted a preliminary experiment to evaluate the impact of different color space representations on the visual and structural characteristics of facial images within our chosen datasets. For this experiment, RGB, BGR, HSV, CIE 1976 (L*a*b*), CIE 1976 (L*u*v*), YCbCr, CIE 1931, XYZ, and YUV color spaces were explored. An example video frame from FaceForensics++ [16] is shown in Figure 3.1. A recurrent theme across images was that HSV enhanced visual contrast and exposed subtle texture anomalies around the facial region more effectively than other color spaces. Because of that difference with other color spaces, HSV was chosen as our main color space when experimenting with the FSBI model [7].

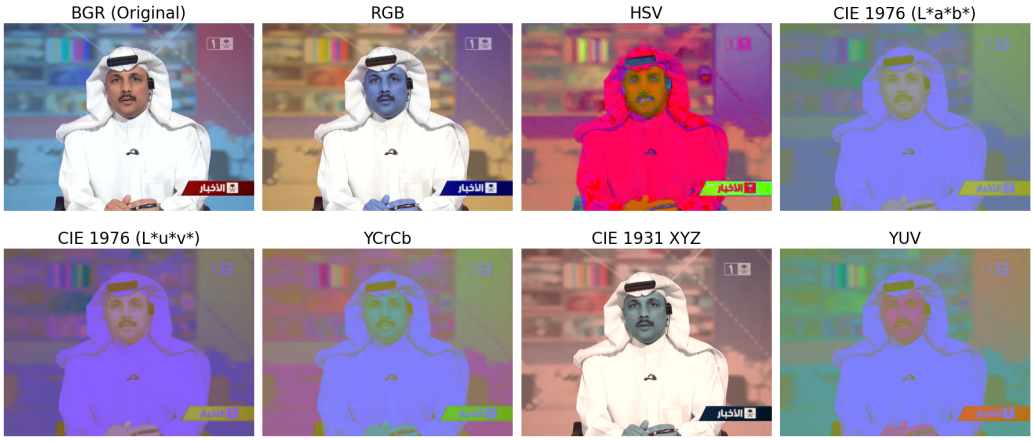

Figure 3.1: The same frame from the FaceForensics++ [16] dataset presented in different color spaces.

3.1 Frequency Enhanced Self-Blended Images

Since this work builds upon the FSBI framework [7] and explores the effect of color spaces on the model performance, this section describes the FSBI workflow.

FSBI is an extension to the SBI (self-blended images) [18] model. FSBI comprises three modules: an SBI generator, a frequency feature generator (FFG), and a CNN classifier. The framework of the FSBI model is presented in Figure 3.2. While the original SBI approach aimed to replicate common forgery artifacts (blending boundaries and statistical inconsistencies) by generating fake images from a single real image, FSBI further processes these synthesized images in the frequency domain. This frequency-based analysis enhances the model's ability to detect subtle forgeries and improves overall detection performance.

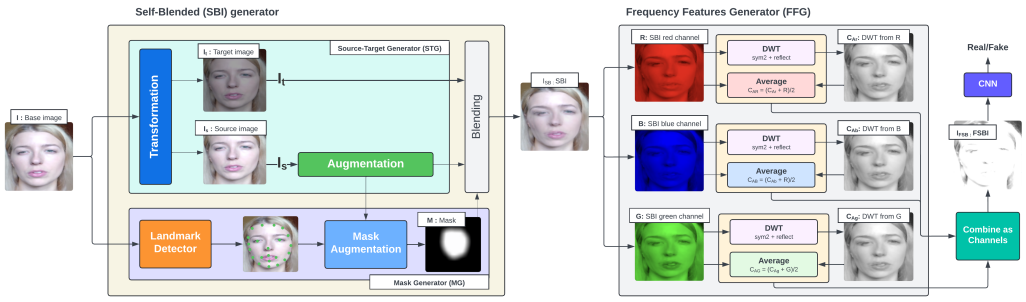

Figure 3.2: Framework of the FSBI model (reproduced from [7]). The input image is passed through the SBI generator, which is then decomposed into R, G, and B channels by the FFG. DWTs are computed for each channel individually, and the approximate coefficients are obtained. These approximate coefficients are then combined with the original channel through an averaging operation. The channels are stacked channel-wise to form the final FSBI image. The resulting FSBIs are used to train a CNN classifier to recognize real and fake images.

The process begins by passing the input image through the SBI generator, which produces pseudo source and target images from the original and blends them to simulate forgery. Next, the FFG module applies a series of Discrete Wavelet Transforms (DWT) to extract frequency-domain features from the blended image. Finally, the resulting features are passed to a CNN classifier, which decides whether images are real or fake.

3.1.1 Self-Blended Image (SBI)

The SBI generator serves as a synthetic forgery creation pipeline. It operates in three stages: source-target generation (STG), mask generation (MG), and blending.

Given an input image I, the STG creates two versions of the image: a pseudo source and target, both initially identical to I. Then it randomly applies color transformations (such as shifting the values of RGB channels, hue, saturation, value, brightness, and contrast), downsampling, sharpening, and translation to one of the images to create landmark mismatches.

MG creates a grayscale mask image to blend the source and target images. It applies a landmark detector to the input image to predict a facial region and initializes a mask by calculating the convex hull from the predicted facial landmarks. The mask is then deformed with the landmark transformation as used by Li et al. [11] and by elastic deformation as adopted by Zhao et al. [22]. It is then smoothed by two Gaussian filters. After the first smoothing, any pixel with a value less than 1 is set to 0. As a result, the mask is eroded if the first Gaussian filter has a larger kernel than the second, and dilated otherwise. Lastly, the blending ratio for the source image is modified by scaling the mask with a factor $r \in (0, 1]$, where r is uniformly selected from the multiset $\{0.25, 0.5, 0.75, 1, 1, 1\}$.

In the final step, the source image I_s and the target image I_t are blended with the generated mask M,

$$I_{SBI} = I_s \odot M + I_t \odot (1 - M). \tag{3.1}$$

3.1.2 Frequency Feature Generator (FFG)

The blended image produced by the SBI module is input into the FFG module, which uses DWT to extract more distinctive features. DWT is a technique used in signal and image processing that breaks down a signal into multiple frequency components by iteratively applying filters and dividing it into approximate and detailed coefficients at various resolutions. This enables multiresolution analysis, capturing both global and local details within the image.

3.1.3 CNN Model

The final stage of the FSBI approach involves a CNN trained to identify artifacts in FSBIs and classify images as either real or fake. The pre-trained EfficientNet-B5 [19] model, originally trained on ImageNet, was fine-tuned. EfficientNet-B5, a member of the EfficientNet family, features compound

scaling, where width, depth, and resolution are uniformly scaled to balance model capacity and computational cost. This design prioritizes state-of-the-art accuracy in image classification tasks, such as those on ImageNet, while using fewer parameters compared to other models with similar performance.

3.2 Datasets

The evaluations in this work are benchmarked against FSBI using the same datasets: Celeb-DF-v2 [12] and FaceForensics++ [16]. We utilized the Celeb-DF-v2 dataset, which contains 5,639 high-quality deepfake videos of celebrities, totaling approximately two million frames. Figure 3.3 illustrates a pair of real and fake images from the Celeb-DF-v2 dataset.

(a) Real (b) Fake

Figure 3.3: Example frames from the Celeb-DF-v2 [12] dataset showing a real and a synthesized image. The faces are zoomed in and placed at the lower left part of each image to highlight distinguishing features. As we can see, the face is completely different in the fake image.

Additionally, FF++ is a dataset consisting of 1,000 original video sequences (approximately 500,000 frames) that have been manipulated with four face manipulation methods: Deepfakes (DF) [4], FaceSwap (FS) [14], Face2Face (F2F) [21], and NeuralTextures (NT) [20]. From the 1,000 real videos, FF++ has generated over 1,800,000 images derived from 4,000 fake videos. Given the large size of the FF++ dataset, we opt to use the compressed c23 version, as done in FSBI experiments. Figure 3.4 illustrates a real image from the FaceForensics++ dataset and its fakes generated by the DF, F2F, FS, and NT methods.

The download links for both datasets were provided after filling out Google Forms with information regarding the purpose of the usage and the project. Celeb-DF-v2 was provided right after submitting the form. However, there was an issue with the automated system while processing requests

for the FF++ dataset. Due to a form error, the system failed to collect email addresses and did not automatically send out the download instructions. Because of that, we managed to access the dataset only after a few weeks.

3.3 Implementation Details

The model was implemented using the PyTorch framework. For training and testing, a local computer and Google Colab Pro were used in parallel to ensure a fast pace. The local computer had an NVIDIA GeForce RTX 3090 GPU with 24 GB of memory. On Google Colab Pro, three GPU options (A100, L4, and T4) were evaluated to determine the most efficient configuration in terms of processing speed and resource utilization. Based on these evaluations, the A100 GPU was selected for training, while the L4 GPU was chosen for testing due to its favorable performance characteristics in inference tasks. For the FF++ dataset, each training session required approximately 14 hours, and testing took around 9 hours. For the Celeb-DF-v2 dataset, training and testing sessions each took approximately 8 hours. If all training and testing sessions had been performed sequentially rather than in parallel, the total runtime would have been around 214 hours. Additionally, extra time was needed for preprocessing tasks, including frame extraction from videos and the detection of facial landmarks and bounding boxes.

Only minimal modifications were made to the original FSBI method to ensure a fair and valid comparison. The proposed model was trained using an early stopping technique to avoid overfitting. The FSBI authors trained their model with 100 epochs to obtain a valid comparison with their baseline model, the SBI method. Therefore, we also used the same number of epochs (100). We trained the model on a batch size of 16. The size of the image is set to 380×380.

(a) Original

(b) Deepfakes

(c) Face2Face

(d) FaceSwap

(e) NeuralTextures

Figure 3.4: Examples from the FaceForensics++ [16] dataset showing manipulated frames using four face manipulation methods. As the images show, the difference between Face2Face and NeuralTextures manipulations is hardly visible to the naked eye.

Chapter 4

Results & Discussion

Similar to the FSBI paper [7], two settings were used to evaluate the proposed approach: within-dataset and cross-dataset. The Receiver Operating Characteristic Area Under the Curve (ROC AUC) metric was used in both settings. The AUC (%) scores for FSBI with HSV color space are shown in Table 4.1.

Table 4.1: Performance results for within- and cross-dataset evaluations across different manipulation types using the HSV color space. Since Celeb-DF-v2 does not have different manipulation types like FF++ does, the Average column represents its overall AUC score (%).

Evaluation	Train	Test	Manipulations				Average
			DF	FS	F2F	NT	
Within-dataset	FF++	FF++	95.28	95.29	92.20	79.48	90.56
	Celeb-DF	Celeb-DF					73.36
Cross-dataset	Celeb-DF	FF++	85.64	83.63	85.01	68.79	80.77
	FF++	Celeb-DF					52.29

As Table 4.1 shows, the AUC scores for HSV are lower than the ones for RGB reported in the original FSBI paper [7]. While experiments using FaceForensics++ as the test set yielded results comparable to the RGB baseline, the model exhibited significantly weaker performance on the Celeb-DF-v2 dataset. In the cross-dataset setting with Celeb-DF-v2 as the test set, the model's performance was close to random classification. The difference in model performance for RGB and HSV is illustrated in Figure 4.1. Additionally, the figure clearly illustrates that experiments using FaceForensics++ for both training and testing achieved higher AUC scores compared

to those where the model was trained on FaceForensics++ and tested on Celeb-DF-v2. That result aligns with expectations, given the domain shift and increased realism of Celeb-DF-v2 deepfakes. This trend was also observed in the original FSBI experiments using RGB, further highlighting the challenge of generalization across datasets.

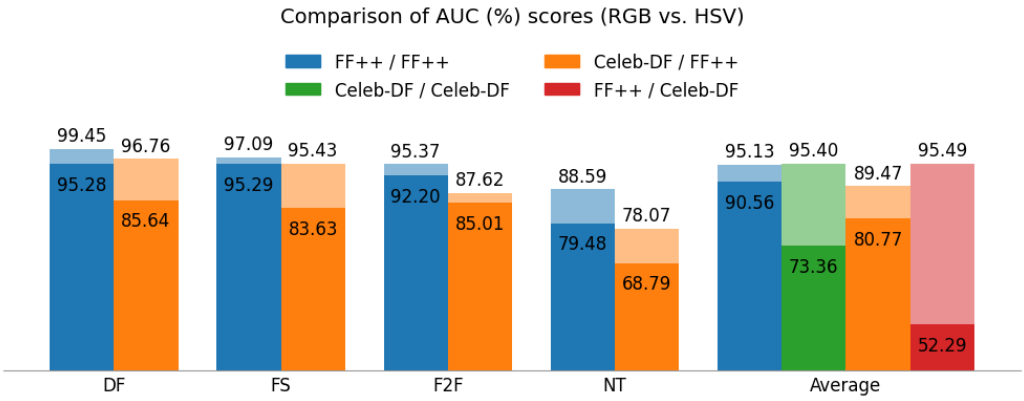

Figure 4.1: AUC score (%) comparison between RGB and HSV color spaces across manipulation types in the FaceForensics++ and Celeb-DF-v2 datasets. Light-colored bars represent RGB, while dark-colored bars indicate HSV.

Several factors may account for the poor performance on Celeb-DF-v2. The deepfakes in this dataset are widely considered more realistic and contain fewer visible artifacts than those in FF++, making them more difficult to detect. Nonetheless, this alone does not fully explain the drastic drop in AUC scores, particularly given that the original RGB-based experiments demonstrated better generalization to Celeb-DF-v2. If the HSV color space were solely responsible for the performance degradation, one would expect consistently low scores across all test datasets, which was not the case. This highlights the need for additional experiments and research to find out the contributing factors and assess the specific role of color space changes in limiting generalization.

We also conducted some experiments using the YCbCr color space. When training on the FaceForensics++ dataset, the model achieved an AUC of 97.97% for the Deepfakes manipulation and 96.82% for the Face2Face manipulation. Notably, the 96.82% AUC for Face2Face exceeds the corresponding result from the RGB-based experiment. Furthermore, when training on Celeb-DF-v2, the model achieved an AUC of 84.28% on Face2Face. These preliminary results are promising, and further experiments with the YCbCr color space still need to be explored.

Chapter 5

Conclusion & Future Work

This thesis investigated the impact of alternative color spaces on the performance of the FSBI model. While prior work in the field has predominantly relied on RGB and BGR color representations, this study aimed to determine whether using other color spaces can improve the ability of the deepfake detection model to identify facial manipulations.

Experimental results demonstrated that while HSV color space exhibited promising visual characteristics, its performance was lower compared to the RGB baseline. Notably, the FSBI model with HSV inputs performed considerably worse on the Celeb-DF-v2 dataset, which contains highly realistic deepfakes. In contrast, one of the preliminary experiments with the YCbCr color space surpassed the AUC score of the same experiment done using RGB. These outcomes indicate that while HSV may not be ideal in the FSBI context, YCbCr still needs further exploration as a potential enhancement. The findings suggest the importance of continuing research into color space transformations as a dimension for improving deepfake detection models.

Given the preliminary results with YCbCr, future studies should conduct full-scale evaluations across multiple datasets and manipulation types to assess its robustness and generalization capabilities.

In addition, testing the trained models against adversarial examples or deepfakes generated by unseen and state-of-the-art models (e.g., diffusion-based synthesis) would provide further insights into the models' practicality and resilience.

By building on these findings, future work can contribute to the development of deepfake detection systems that are both more accurate and more robust across real-world scenarios.

Bibliography

[1] Maryam Abbasi, Paulo Váz, José Silva, and Pedro Martins. Comprehensive evaluation of deepfake detection models: Accuracy, generalization, and resilience to adversarial attacks. *Applied Sciences*, 15(3):1225, 2025.

[2] Guorong Chen, Chongling Du, Yuan Yu, Hong Hu, Hongjun Duan, and Huazheng Zhu. A deepfake image detection method based on a multi-graph attention network. *Electronics*, 14(3):482, 2025.

[3] Shen Chen, Taiping Yao, Hong Liu, Xiaoshuai Sun, Shouhong Ding, Rongrong Ji, et al. DiffusionFake: Enhancing generalization in deepfake detection via guided stable diffusion. *Advances in Neural Information Processing Systems*, 37:101474–101497, 2024.

[4] deepfakes. faceswap: Deepfakes Software For All. https://github.com/deepfakes/faceswap. Accessed: 2025-04-23.

[5] Siqi Gu, Zihan Qin, Lizhe Xie, Zheng Wang, and Yining Hu. Multiscale features integrated model for generalizable deepfake detection. *International Journal of Intelligent Systems*, 2025(1):7084582, 2025.

[6] Luca Guarnera, Oliver Giudice, and Sebastiano Battiato. Deepfake detection by analyzing convolutional traces. In *IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 666–667, 2020.

[7] Ahmed Abul Hasanaath, Hamzah Luqman, Raed Katib, and Saeed Anwar. FSBI: Deepfake detection with frequency enhanced self-blended images. *Image and Vision Computing*, page 105418, 2025.

[8] Justin Jiang. Addressing vulnerabilities in AI-image detection: Challenges and proposed solutions. In *2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC)*, pages 1–9. IEEE, 2025.

[9] Paritosh Joshi and V Nivethitha. Deep fake image detection using Xception architecture. In *2024 5th International Conference on Recent Trends*

in Computer Science and Technology (ICRTCST), pages 533–537. IEEE, 2024.

[10] Hanzhe Li, Jiaran Zhou, Yuezun Li, Baoyuan Wu, Bin Li, and Junyu Dong. FreqBlender: Enhancing deepfake detection by blending frequency knowledge. *arXiv preprint arXiv:2404.13872*, 2024.

[11] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 5001–5010, 2020.

[12] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. Celeb-DF: A large-scale challenging dataset for deepfake forensics. In *IEEE/CVF conference on computer vision and pattern recognition*, 2020.

[13] Ping Liu, Qiqi Tao, and Joey Zhou. Robust deepfake detection by addressing generalization and trustworthiness challenges: A short survey. In *1st ACM Multimedia Workshop on Multi-modal Misinformation Governance in the Era of Foundation Models*, pages 3–11, 2024.

[14] MarekKowalski. FaceSwap: 3d face swapping implemented in python. https://github.com/MarekKowalski/FaceSwap/. Accessed: 2025-04-23.

[15] Md Shohel Rana, Mohammad Nur Nobi, Beddhu Murali, and Andrew H Sung. Deepfake detection: A systematic literature review. *IEEE Access*, 10:25494–25513, 2022.

[16] Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. FaceForensics++: Learning to detect manipulated facial images. In *ICCV/CVF International Conference on Computer Vision*, 2019.

[17] Khouloud Samrouth, Pia El Housseini, and Olivier Deforges. Siamese network-based detection of deepfake impersonation attacks with a person of interest approach. *ACM Transactions on Multimedia Computing, Communications and Applications*, 21(3):1–23, 2025.

[18] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 18720–18729, 2022.

[19] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[20] Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[21] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of RGB videos. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 2387–2395, 2016.

[22] Tianchen Zhao, Xiang Xu, Mingze Xu, Hui Ding, Yuanjun Xiong, and Wei Xia. Learning self-consistency for deepfake detection. In *IEEE/CVF conference on computer vision and pattern recognition*, pages 15023–15033, 2021.