Enhanced Deep Learning DeepFake Detection Integrating Handcrafted Features

Alejandro Hinke-Navarro[1], Mario Nieto-Hidalgo[1][0000−0003−0623−6455], Juan M. Espín[1][0000−0001−6521−7890], and Juan E. Tapia[2][0000−0001−9159−4075]

[1] Facephi Biometrics S.A., Alicante, Spain.
{alejandrohinke, marionieto, jmespin}@facephi.com
[2] da/sec-Biometrics and Internet Security Research Group, Darmtstadt, Germany
juan.tapia-farias@h-da.de

Abstract. The rapid advancement of deepfake and face swap technologies has raised significant concerns in digital security, particularly in identity verification and onboarding processes. Conventional detection methods often struggle to generalize against sophisticated facial manipulations. This study proposes an enhanced deep-learning detection framework that combines handcrafted frequency-domain features with conventional RGB inputs. This hybrid approach exploits frequency and spatial domain artifacts introduced during image manipulation, providing richer and more discriminative information to the classifier. Several frequency handcrafted features were evaluated, including the Steganalysis Rich Model, Discrete Cosine Transform, Error Level Analysis, Singular Value Decomposition, and Discrete Fourier Transform.

Keywords: Face Manipulation · Handcrafted Features · Digital Forensics.

1 Introduction

Image manipulation has become a widely discussed topic over the years, with its detection posing an increasingly complex challenge because of the rapid advancements in generative techniques. Among the various forms of digital face manipulation, key methods include face swap, identity swap, attribute manipulation, and entire face synthesis [15].

Face swapping involves replacing a target individual's face with another person's, effectively altering the subject's appearance while retaining their original context. In contrast, full-face synthesis refers to the complete generation of facial images from scratch using advanced generative models, such as Generative Adversarial Networks (GANs) or diffusion models. These techniques enable the creation of highly realistic facial representations, often indistinguishable from authentic images.

Nowadays, most users perceive this technology as harmless entertainment; however, it is increasingly being misused for malicious purposes, such as spreading fake news, generating illicit content, and engaging in political manipulation, among others. These harmful applications have a significant impact on social media, undermining trust and contributing to a crisis of authenticity in digital content across the Internet.

Traditional methods, based on RGB pixel values and convolutional neural networks (CNNs), perform well on intra-datasets but struggle to generalize across unseen datasets [15]. This limitation arises because artifacts indicative of manipulation can be significantly diminished due to factors such as image compression or manual editing, making it challenging for these models to detect subtle inconsistencies across diverse sources.

To address this challenge, the following research questions are explored:

- How can generalization across datasets be improved?
- Which frequency-domain representations are most effective?
- How can frequency-domain features be integrated into deep learning models?

This work focuses specifically on identity swapping and full-face synthesis, two of the most prevalent manipulation techniques in digital media.

The main contributions of this work are:

- A study of frequency-domain features for face manipulation detection.
- An evaluation of several handcrafted features, identifying Discrete Cosine Transform as the most effective.
- A demonstration that minimum score-level fusion between intensity pixel values and frequency features yields improved performance over baseline models.

The remainder of this article is structured as follows: **Section** 2 reviews related work on deepfake and face manipulation detection. **Section** 3 describes the datasets and the proposed method. **Section** 4 reports experimental results. Finally, **Section** 5 concludes the paper and outlines directions for future research.

2 Related Works

Traditional methods based on intensity values (RGB) images and convolutional neural networks (CNNs) have demonstrated high performance on intra-datasets but lower generalization capabilities to perform with high rates on cross-datasets. To overcome this challenge, new approaches like frequency domain have been explored based on the changes in frequencies (high and low) that are produced when the image is manipulated. A similar effect is dedicated to compression.

Conventional deepfake detection approaches predominantly leverage spatial domain features extracted from pixel values across RGB images using deep neural networks.

Studies such as Luo et al. [6] highlight that CNN-based models often overfit to method-specific color textures, which limits their generalization capabilities when tested against unseen manipulations.

Similarly, the work by Ibsen et al. [1] emphasizes that RGB-only models struggle to differentiate real from synthetic faces when exposed to novel generative models or post-processing operations. These limitations underline the need for more robust detection techniques that incorporate additional feature representations beyond RGB data.

Recent advances in deepfake detection have highlighted the effectiveness of frequency-domain analysis in identifying manipulated content [16, 6].

Wang et al. [16] introduced a Frequency Domain Filtered Residual Network, which enhances detection robustness by fusing wavelet-transformed frequency information with RGB data, particularly improving performance on compressed deepfake images.

Luo et al. [6] showed that multi-scale SRM filtering strengthens cross-dataset generalization by detecting high-frequency noise residuals.

More recently, Tan et al. [12] proposed FreqNet, a frequency-aware model that enhances deepfake detector generalization by learning high-frequency features independently of their source.

Li et al. [4] introduced FreqBlender, a method that synthesizes pseudo-fake faces by manipulating frequency information, improving the learning of generic forgery traces, and enhancing detection accuracy.

Tapia et al. [14] also demonstrate that frequency-based filters can be used to detect digital manipulation attacks, such as Morphing.

Rahaman et al. [10] introduced the concept of spectral bias, demonstrating through Fourier analysis that neural networks exhibit a learning preference for low-frequency functions. This spectral bias explains why neural networks often generalize well to natural data and highlights the robustness of low-frequency components to parameter perturbations.

Many of these findings suggest that the frequency domain contains valuable information that can be effectively leveraged to improve the detection of manipulated images.

3 Proposed Method

This work proposes a method for detecting face digital manipulation attacks, based on handcrafted frequency features and fusion with intensity values at the score level. Several datasets were employed to evaluate the generalization capabilities of the proposed approach. A diagram illustrating the method is presented in Figure 1.

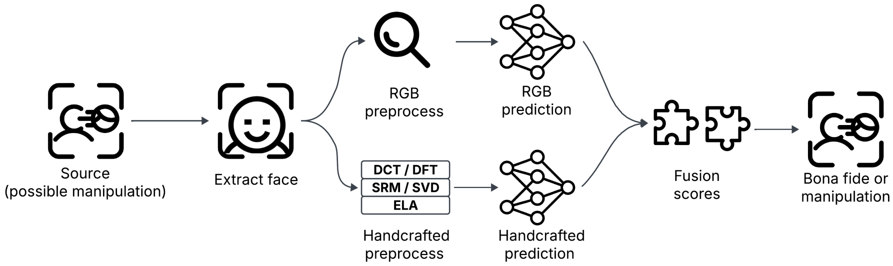

Fig. 1. Manipulation Attack Detection Framework.

3.1 Datasets

In this study, six different datasets of digitally manipulated face images were used:

- **FaceForensics++[11]:** A widely used dataset for deepfake detection, comprising 4,320 videos, including 720 original videos sourced from YouTube and 3,600 manipulated videos generated using FaceShifter, FaceSwap, Face2Face, Deepfakes, and NeuralTextures. The official dataset split was followed, with 720 videos for training, 140 for validation, and 140 for testing. Five random frames per video were used in this study.
- **Celeb-DF[5]:** A deepfake dataset specifically designed for identity-swapping manipulations, containing 5,639 deepfake videos generated from 590 original videos sourced from YouTube. Due to its real-world origin, the dataset is highly compressed, often exhibiting lower visual quality and compression artifacts, making detection more challenging. Only one frame per video was used in this study.
- **DeepfakeTIMIT[2]:** This dataset comprises videos where faces are swapped using a GAN-based approach developed from the original autoencoder-based Deepfake algorithm. It includes 620 videos with faces swapped, using the VidTIMIT database as the source. Two different qualities are provided: lower quality (LQ) with 64×64 input/output size models and higher quality (HQ) with 128×128 size models. One frame per video was used in this study.
- **DeePhy[9]:** This dataset employs sequential face swapping based on a phylogenetic approach. It contains 468 spoof videos sourced from YouTube, encoded in MPEG4 format with a resolution of 720p, using a single frame per video. One frame per video was used in this study.
- **Defacto[7]:** This dataset includes face-swapped images generated from MS-COCO images through automated forgery generation techniques, resulting in semantically meaningful and detailed manipulations. It contains 3,000 spoof images of variable sizes. One frame per video was used in this study.
- **SWAN-DF[3]:** The first high-fidelity publicly available dataset of realistic audio-visual deepfakes, where both faces and voices appear and sound like the target person. Based on the public SWAN database of real videos recorded in HD on iPhone and iPad Pro, it includes 30 pairs of manually selected individuals. Faces and voices were swapped using several autoencoder-based face-swapping models and blending techniques from DeepFaceLab, along with voice conversion methods such as YourTTS, DiffVC, HiFiVC, and FreeVC. A random selection of 10% of the dataset was used in this study.

Table 1 shows a summary of all the datasets used in this research.

Table 1. Summary of datasets. DF, FS, NT, FSW, and F2F represent DeepFake, FaceShifter, NeuralTransfer, FaceSwap, and Face-to-Face, respectively.

Database	**Nº of Images**	**Manipulation algorithm**
FF++	25,000 fake, 5000 real	DF, FS, NT, FSW, F2F
CelebDF	5,639 fake, 590 real	Improved DF
DeepfakeTIMIT	640 fake	GAN-based (face swap-GAN)
DeePhy	468 fake, 100 real	Phylogenetic sequential FS
Defacto	3,000 fake, 200 real	Automated semantic FS
SWAN-DF	11,940 fake	Autoencoder-based (DeepFaceLab)

3.2 Metrics

To evaluate the effectiveness of the proposed method, the ISO/IEC 30107-3 was followed[3]. Detection Equal Error Rate (D-EER) metric was employed, which represents the point at which the Attack Presentation Classification Error Rate (APCER) and the Bona fide Presentation Classification Error Rate (BPCER) are equal. The APCER indicates the proportion of attack presentations incorrectly classified as bona fide (false positives), while BPCER denotes the proportion of bona fide presentations incorrectly classified as attacks (false negatives). A lower D-EER value reflects the higher accuracy and robustness of the detection system. D-EER is widely used in biometric systems and forgery detection tasks due to its balanced assessment of both types of error rates, providing a comprehensive measure of system performance.

3.3 Feature Extraction

Several feature extraction techniques based on handcrafted features have been employed to distinguish between bona fide and digitally manipulated images [14]. In this study, five frequency handcrafted feature extraction methods were used individually and in combination to improve the detection of manipulated faces: Color (RGB), which is represented by the pixel values, Discrete Cosine Transform (DCT), Steganalysis Rich Model (SRM), Discrete Fourier Transform (DFT), Error Level Analysis (ELA), and Singular Value Decomposition (SVD). All features were extracted from grayscale versions of the images using to emphasize structural and frequency domain characteristics, except for color pixel values of RGB images.

Discrete Cosine Transform (DCT) DCT is a widely used technique in image processing that transforms spatial domain information into the frequency domain. It decomposes an image into a sum of cosine functions oscillating at different frequencies, which helps detect hidden artifacts introduced during manipulations, particularly in compressed images, as it is a core component of popular formats like JPEG that exploit frequency information for efficient compression.

In this study, DCT was applied to the entire image as well as to sub-blocks of varying sizes, specifically 8×8, 12×12, 16×16, 20×20, and 24×24 pixels, with the 20×20 configuration proving to be the most effective.

[3] https://www.iso.org/standard/79520.html

Steganalysis Rich Model (SRM) SRM is a feature extraction technique commonly used in digital forensics to detect hidden modifications in images. It focuses on capturing high-frequency noise patterns that arise from manipulation processes. In this study, an SRM filter using a kernel described in Eq. 1 was applied to the grayscale images to enhance edge detection and expose subtle alterations.

$$SRM = \begin{bmatrix} 0.0 & 1.0 & 0.0 \\ 1.0 & -4.0 & 1.0 \\ 0.0 & 1.0 & 0.0 \end{bmatrix} \tag{1}$$

This filter emphasizes discrepancies in the high-frequency domain by highlighting regions where pixel intensities exhibit irregular patterns, which are often indicative of tampering.

The Discrete Fourier Transform (DFT) DFT converts an image from the spatial domain to the frequency domain, representing it in terms of sinusoidal components. This transformation helps analyze periodic patterns and identify inconsistencies introduced by generative models or post-processing operations. It is particularly useful for detecting manipulation artifacts that manifest as unnatural frequency distributions.

Error Level Analysis (ELA) ELA is a forensic technique used to detect areas of an image that have undergone different levels of compression. Repeated compression of an image and comparison with the original reveal discrepancies in compression artifacts, which can indicate tampered regions. In this study, it was applied to grayscale images to identify potential manipulation traces based on differences in compression levels across various regions of the image. Areas with significant discrepancies often correspond to edited portions, making ELA a valuable tool for forgery detection.

Singular Value Decomposition (SVD) SVD is a matrix factorization technique that decomposes an image into three matrices, representing its intrinsic structure in terms of singular values and orthogonal components. It is effective in identifying structural changes caused by manipulations, as alterations typically disrupt an image's natural rank and singular value distribution. This study applied SVD to grayscale images to capture global structural inconsistencies with a component of 50. Figure 2, shows an example of the frequency features extracted.

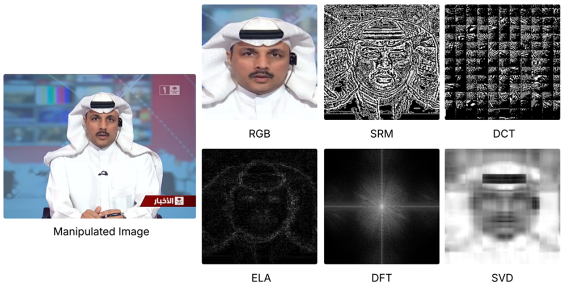

Fig. 2. Feature extraction example for a manipulated image.

3.4 Models

Preprocesing. The preprocessing pipeline begins by cropping faces and adding a 50% padding around each crop. This wider margin exposes background context that face-swap and similar attacks typically leave unaltered, allowing the network to contrast manipulated pixels with their unmodified surroundings. Subsequently, the images are resized to a fixed resolution of 384 × 384 pixels.

The resized images serve as inputs to either EfficientNetV2 B0 [13] or MobileViT-S [8] models, both of which are initialized using ImageNet pre-trained weights. EfficientNetV2-B0 was selected due to its well-balanced trade-off between computational efficiency and performance, making it suitable for deployment in resource-constrained environments. MobileViT-S, chosen for its compact size and rapid inference capability, leverages transformer-like attention mechanisms to capture detailed feature interactions through self-attention maps.

For handcrafted models, images are converted to grayscale. Several data augmentation techniques were employed during training to enhance model robustness, including horizontal flipping, random contrast adjustment, random brightness variation, random hue shifts, random saturation changes, and random JPEG compression, which were applied exclusively to manipulated images. This choice is motivated since GAN generated images often lack of compression artifacts, random JPEG compression was applied to manipulated samples to prevent the model from relying on this pattern and instead focus on manipulation-related traces.

Model weights were optimized using the AdaGrad algorithm with a minibatch size of 32 and an initial learning rate of $1e-4$. Training was conducted for up to 225 steps, equivalent to approximately 65 epochs. For the MobileViT-S architecture, a patch size of 2 was explicitly adopted. All training was performed using an NVIDIA A100 GPU.

Fusion at Score Level The fusion of scores involves combining the outputs of different models (RGB and Frequency) based on specific aggregation rules. The fusion strategies considered in this experiment include weighted fusion, where models contribute based on assigned importance; minimum fusion, which selects the lowest score among the models; mean fusion, which computes the average score; and maximum fusion, which takes the highest score from each model.

4 Experiments

Three experiments were proposed to show and compare the results with different frequency filters.

4.1 Experiment 1: Handcrafted features benchmark

All filters were trained and evaluated using the datasets mentioned in Table 1 to measure the impact of each one.

Table 2. D-EER % for the different handcrafted features.

		Intra	**Cross**		
		FF++	Celeb-DF	Dephy	Defacto
Effv2b0	RGB	6.20	35.87	11.72	30.00
	SRM	5.75	51.53	14.11	61.10
	DCT	3.18	46.96	8.19	48.37
	ELA	8.58	45.96	11.61	38.92
	DFT	35.23	46.54	36.30	48.00
	SVD	31.69	44.77	17.63	33.07
MobileViT-S	RGB	1.37	36.76	7.05	31.97
	SRM	13.93	53.98	25.83	53.50
	DCT	5.03	49.18	20.02	50.50
	ELA	13.59	46.11	9.44	33.50
	DFT	40.27	48.67	45.85	48.03
	SVD	35.43	41.56	20.02	32.50

Observing the cross-dataset performance in Table 2, the results are generally suboptimal. These findings suggest that mismatched bona fide distributions due to dataset-specific conditions such as varying image acquisition settings, compression methods, and quality negatively affect model generalization. Experiment 2 explores a potential solution by using a single, consistent source of bona fide images from FaceForensics++(FF++).

4.2 Experiment 2: Handcrafted features benchmark using FF++ bona fides

Due to significant discrepancies in the distribution of bona fide images across the public datasets used in this study (see Fig. 3), only bona fide images from the FF++ dataset were employed for final model evaluation. This decision is supported by several considerations.

- **Heterogeneous capture conditions.** The datasets differ markedly in their image sources: some contain "in-the-wild" pictures taken under uncontrolled settings, whereas others include images acquired in controlled or studio environments. This mismatch produces considerable variation in visual characteristics and overall image quality.

- **Compression and format inconsistencies.** Differences in compression schemes, file formats, and orientations introduce additional divergence, making cross-dataset comparisons more difficult.
- **Shortage of bona fide samples.** Several datasets provide only a small number of bona fide images, or none at all, which limits representativeness and reduces statistical reliability during evaluation.

Figure 3 shows the different images from bona fide subsets.

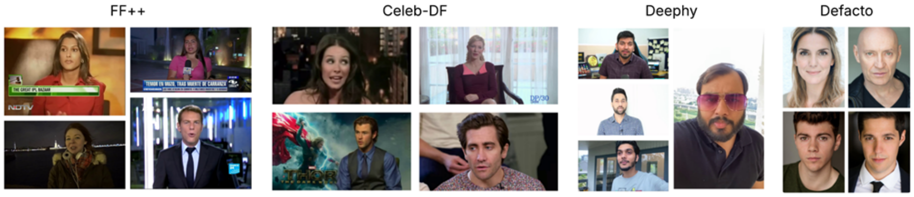

Fig. 3. Bona fide samples distribution across datasets.

4.3 Experiment 3: Fusions at score level

Building upon the findings from Experiment 2 in Table 3, RGB and DCT demonstrated superior performance. This experiment focuses on these two feature set. The DCT-based model exhibits strong detection capabilities for identity face swapping but performs poorly in full-face synthesis detection, whereas the RGB-based model shows the opposite trend. The objective is to leverage the strengths of both spatial and frequency domains to enhance detection performance by exploring various fusion strategies.

It is essential to emphasise that model calibration prior to score fusion significantly impacts overall performance. Since FF++ served as the common source of bona fide images for each dataset, calibration was conducted by targeting a BPCER value. This approach ensures a consistent thresholding strategy across datasets.

In Table 4, the default protocol refers to the configuration obtained directly from training without applying any threshold calibration. For the RGB EfficientNet v2 b0 and DCT EfficientNet v2 b0 models, the "Default" model configuration yielded a BPCER of 19.27% and 7.71%, respectively. In contrast, the RGB MobileViT-S and DCT MobileViT-S models achieved BPCERs of 4.02% and 11.07%, respectively, under the default conditions. "Protocol I" refers to each model being calibrated to BPCER 2.00% before the fusion. "Protocol II" means that each model has been calibrated to BPCER 5.00% before the fusion.

Table 3. D-EER % for the different handcrafted features using FF++ bona fides for every dataset. The highlighted numbers in bold indicate the best performance observed across the dataset.

		Intra	**Cross**					**Avg**
		FF++	Celeb-DF	Df-Timit	Dephy	Defacto	Swan DF	
Effv2b0	RGB	6.20	17.45	**6.39**	9.43	29.36	15.58	14.07
	SRM	5.75	17.12	36.25	22.95	18.43	24.17	20.78
	DCT	3.18	**3.86**	50.49	12.38	**9.93**	29.35	18.20
	ELA	8.58	8.22	13.75	14.57	44.13	33.21	20.41
	DFT	35.23	24.00	61.41	53.14	62.44	69.29	50.92
	SVD	31.69	35.61	10.92	25.53	42.89	36.04	30.78
Mobile ViT-S	RGB	**1.37**	**4.03**	27.18	**6.58**	11.40	**8.72**	9.21
	SRM	13.93	19.13	36.73	34.57	19.11	27.01	25.91
	DCT	**5.03**	7.40	49.35	25.14	**9.53**	34.27	21.79
	ELA	13.59	10.40	17.47	17.81	48.18	46.31	25.63
	DFT	40.27	36.06	47.82	52.66	64.10	66.48	51.57
	SVD	35.43	34.87	18.77	25.72	38.93	41.95	32.78

Table 4. D-EER % for the different fusions by a minimum score between RGB and DCT with the designed protocol. The highlighted numbers in bold indicate the best performance observed across the dataset.

		Intra	**Cross**				
Model	**Protocol**	FF++	Celeb-DF	Df.TIMIT	Dephy	Defacto	SwanDF
RGB Effv2b0 + DCT Effv2b0	Default	2.01	5.54	**7.69**	5.52	13.28	15.40
	Protocol I	2.03	4.73	8.74	5.52	11.76	15.97
	Protocol II	1.99	4.49	9.22	**5.33**	11.44	16.44
RGB MobileViT-S + DCT MobileViT-S	Default	1.34	2.98	38.43	8.57	9.80	11.74
	Protocol I	1.34	2.52	34.55	7.05	8.23	9.73
	Protocol II	1.34	2.98	38.43	8.57	9.80	11.74
RGB MobileViT-S + DCT Effv2b0	Protocol I	**1.17**	**2.56**	36.57	5.99	**7.73**	**9.06**

5 Conclusions

Minimum score fusion between spatial and frequency-domain features achieved the best performance. These findings suggest that integrating handcrafted frequency features with deep learning models enhances manipulation detection, demonstrating the effectiveness of this hybrid approach in improving robustness against various manipulation techniques.

Acknowledgements

This work was supported by Facephi, R&D department and the German Federal Ministry of Education and Research and the Hessian Ministry of Higher Education, Research, Science and the Arts within their joint support of the National Research Center for Applied Cybersecurity ATHENE.

References

[1] Shichao Dong et al. “Implicit identity leakage: The stumbling block to improving deepfake detection generalization”. In: *Proceedings of the IEEE/CVF Conf. on Computer Vision and Pattern Recognition.* 2023, pp. 3994–4004.

[2] Pavel Korshunov and Sébastien Marcel. *DeepFakes: a New Threat to Face Recognition? Assessment and Detection.* 2018.

[3] Pavel Korshunov et al. “Vulnerability of Automatic Identity Recognition to Audio-Visual Deepfakes”. In: *IEEE Intl. Joint Conf. on Biometrics.* Sept. 2023.

[4] Hanzhe LI et al. “FreqBlender: Enhancing DeepFake Detection by Blending Frequency Knowledge”. In: *The Thirty-eighth Annual Conference on Neural Information Processing Systems.* 2024.

[5] Yuezun Li et al. “Celeb-DF: A large-scale challenging dataset for deepfake forensics”. In: *Proceedings of the IEEE/CVF Conf. on computer vision and pattern recognition.* 2020, pp. 3207–3216.

[6] Yuchen Luo et al. “Generalizing face forgery detection with high-frequency features”. In: *Proceedings of the IEEE/CVF Conf. on computer vision and pattern recognition.* 2021, pp. 16317–16326.

[7] Gaël Mahfoudi et al. “Defacto: Image and face manipulation dataset”. In: *27Th European Signal Processing Conference (EUSIPCO).* IEEE. 2019, pp. 1–5.

[8] Sachin Mehta and Mohammad Rastegari. “MobileViT: Light-weight, General-purpose, and Mobile-friendly Vision Transformer”. In: *Intl. Conf. on Learning Representations.* 2022.

[9] Kartik Narayan et al. “Deephy: On deepfake phylogeny”. In: *IEEE Intl. Joint Conf. on Biometrics (IJCB).* IEEE. 2022, pp. 1–10.

[10] Nasim Rahaman et al. “On the spectral bias of neural networks”. In: *Intl. conference on Machine Learning.* PMLR. 2019, pp. 5301–5310.

[11] Andreas Rossler et al. “Faceforensics++: Learning to detect manipulated facial images”. In: *Proceedings of the IEEE/CVF Intl. conf. on computer vision.* 2019, pp. 1–11.

[12] Chuangchuang Tan et al. “Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning”. In: *Proceedings of the AAAI Conference on Artificial Intelligence.* Vol. 38. 5. 2024.

[13] Mingxing Tan and Quoc Le. “Efficientnetv2: Smaller models and faster training”. In: *Intl. Conf. on Machine Learning.* PMLR. 2021, pp. 10096–10106.

[14] Juan E Tapia and Christoph Busch. “Face feature visualisation of single morphing attack detection”. In: *11th Intl. Workshop on Biometrics and Forensics (IWBF).* IEEE. 2023, pp. 1–6.

[15] Ruben Tolosana et al. “Deepfakes and beyond: A survey of face manipulation and fake detection”. In: *Information Fusion* 64 (2020).

[16] Bo Wang et al. “Frequency domain filtered residual network for deepfake detection”. In: *Mathematics* 11.4 (2023), p. 816.