Deepfake Video Detection

Report of the Summer Internship to be submitted in Partial Fulfillment of the Requirements for the Award of the Degree of

Master of Technology
in
Electrical Engineering
with specialization in Signal Processing and Machine Learning

by

Piyush Dangi

23EE65R08

Under the supervision of

Dr. Rajiv Ranjan Sahay

Department of Electrical Engineering
Indian Institute of Technology Kharagpur

August 2024

Abstract

DETECTION of deepfake videos is crucial due to the increasing realism and potential misuse of such synthetic content. Identifying deepfakes requires sophisticated methods that can discern subtle discrepancies between real and artificially created videos. In this context, Photo Response Non-Uniformity (PRNU) analysis emerges as a promising technique. PRNU leverages the unique sensor noise patterns inherent to individual cameras, which are typically absent in GAN-generated content.

Traditional deepfake detection methods often struggle to generalize across different GAN architectures and manipulation techniques. To address this challenge, we propose a novel approach that integrates PRNU-based feature extraction with Support Vector Machine (SVM) classification. The PRNU technique is employed to extract distinctive noise residual features from video frames, while SVM is utilized for robust classification. Additionally, we enhance the feature extraction process using Discrete Fourier Transform (DFT) methods, capturing frequency-domain characteristics that further refine detection accuracy.

This project aims to advance the state-of-the-art in deepfake detection by demonstrating the effectiveness of combining PRNU with DFT and SVM.

Keywords: *Denoising, photo response non-uniformity, support vector machine.*

Contents

1 Introduction

THE rapid evolution of deepfake technology has made it increasingly easy to create highly realistic fake videos by swapping the faces of individuals. This advancement presents a significant challenge, as deepfakes can be used maliciously to spread misinformation, manipulate public opinion, and damage reputations. Detecting such sophisticated fakes is crucial to maintaining trust and security in digital media.

Deepfake detection has traditionally relied on deep learning approaches, which, while powerful, often come with high computational costs and complexity. These methods require substantial processing power and memory, making them less suitable for real-time or large-scale applications, such as monitoring content on social media platforms. Moreover, deep learning methods can be prone to overfitting, making them less effective as deepfake techniques continue to evolve and generate new types of artifacts.

In contrast, some methods focus on detecting specific anomalies or inconsistencies in videos, such as unnatural eye movements [7] or irregular head poses [12]. However, these techniques can struggle to keep up with the rapid advancement of deepfake technologies, as new methods of creating fake content can bypass these detection strategies.

To address these challenges, The method [9] is using Photo Response Non-Uniformity (PRNU) [1] analysis combined with Discrete Fourier Transform (DFT) and Support Vector Machine (SVM) classification for deepfake detection. PRNU is a well-established technique in digital forensics that examines the unique noise patterns introduced by individual camera sensors. This noise, which is inherent to each sensor, remains consistent across images taken with the same device and can be used to identify whether an image or video has been manipulated [6].

DFT enhances this approach [9] by extracting frequency-domain features from images, which helps in capturing detailed information about the image's texture and structure. By analyzing both the PRNU patterns and frequency-domain features, we can create a more robust detection system that is less dependent on the specific artifacts produced by different deepfake technologies.

This method [9] leverages SVM for classification, which allows us to effectively differentiate between real and fake content based on the features extracted from PRNU and DFT. This combination offers a balance between accuracy and computational efficiency, making it feasible for large-scale applications.

This work represents one of the first comprehensive evaluations of PRNU-based deepfake detection on extensive and varied datasets.

2 Prior Art

The detection of deepfakes has garnered significant attention due to the increasing sophistication of generative models like Generative Adversarial Networks (GAN). Various methods have been proposed to address the challenge of identifying fake images and videos, each leveraging different techniques and technologies. Conventional Detection Methods explores the key research directions in multimedia forensics prior to the advent of deep learning technologies. Conventional methods primarily focus on identifying artifacts linked to either the camera's internal processing or subsequent editing actions.

2.1 Image power spectrum based method

The image power spectrum-based method [3] addresses the issue of CNN-based up-sampling techniques, like up-convolution and transposed convolution, which often distort the spectral distribution of natural images. By introducing a spectral regularization term during training, this method enhances spectral consistency, reduces high-frequency errors, and improves the quality of generative models.

Impact on cpectral consistency

This method uses a 1D Fourier power spectrum to analyze the effects of up-sampling on the spectral properties of images. The Discrete Fourier Transform (DFT) is applied to the image, and the resulting 2D power spectrum is converted into a 1D representation through azimuthal integration. The analysis shows that common up-sampling techniques can significantly alter the image's frequency spectrum, impacting the visual quality of the output. By incorporating spectral regularization, these distortions are minimized, leading to more accurate and visually pleasing generative models.

2.2 Color filter array interpolation based method

Color Filter Array (CFA) artifacts are crucial in digital image forensics due to their role in identifying manipulations, including deepfakes. Most digital cameras employ a CFA with a periodic pattern, where each sensor element captures light in a specific wavelength range (red, green, or blue). The missing color information is interpolated from surrounding pixels through a process called demosaicing, introducing a subtle periodic correlation pattern across the image [4].

When an image is manipulated, such as through splicing or deepfake creation, this periodic pattern is disturbed. Since the CFA pattern and interpolation algorithms are unique to each camera model, any tampered region spliced from an image captured by a different camera will exhibit an anomalous pattern. One of the pioneering methods

to detect these anomalies was introduced by Popescu and Farid in 2005 [**?**], utilizing a linear model to capture periodic correlations.

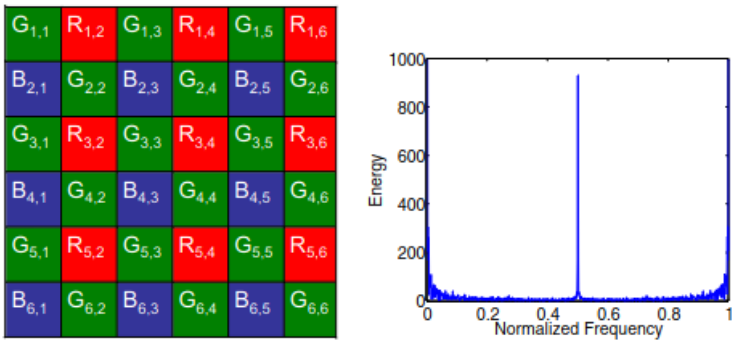


Figure 1: Digital cameras use a color filter array (CFA) on their sensors, capturing only one color per pixel. Demosaicing then interpolates these values to create a full-color image. The process introduces detectable artifacts in the image data. By analyzing these artifacts, we can identify whether an image is from a digital camera or computer-generated [4].

These periodic signals manifest as strong peaks in the Fourier domain, which can be exploited to differentiate natural images from computer-generated ones, especially after high-pass filtering the image to extract more effective features. The problem can also be addressed using a Bayesian framework, producing a probability map that allows for fine-grained localization of tampered regions. Further analysis extends this approach to consider pixel correlations across color channels, enhancing the detection of manipulations.

2.3 Photo-response non-uniformity (PRNU) based method

Deepfake detection using Photo-Response Non-Uniformity [1] is a technique that leverages the unique sensor noise pattern inherent to each camera. PRNU serves as a digital fingerprint, consistently present across all images and videos captured by the same sensor. The detection process begins with the extraction of the PRNU pattern from an image or video, focusing on the high-frequency noise components that are typically invisible to the naked eye. This extracted pattern is then compared to the expected PRNU pattern, either from a reference database or from other images known to be from the same camera. Any significant deviations between the extracted and expected PRNU patterns can indicate tampering or manipulation, such as the presence of deepfake content.

Deepfake generation techniques often fail to replicate these subtle noise patterns, leading to inconsistencies that can be detected through PRNU analysis. By identifying these anomalies, PRNU-based methods provide a robust means of distinguishing between authentic and manipulated media. This approach is particularly valuable in

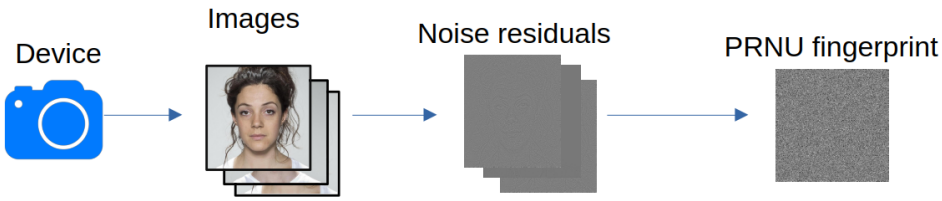


Figure 2: The device PRNU Pattern is estimated by a large number of noise residuals.

digital forensics, where maintaining the integrity and authenticity of visual content is crucial.

3 Aim and Objectives

The *aim* of this project is to develop a generalized method for detecting deepfakes, leveraging advanced image processing techniques to ensure robust and reliable detection across various types of deepfake content.

The following objectives have been set to achieve this :

1. Improving the deepfake detection framework to effectively identify various types of manipulated content through a generalized approach.

2. Integrating advanced image processing techniques, including PRNU (Photo Response Non Uniformity), to enhance the accuracy and robustness of the detection process.

3. Evaluating the performance of the detection method across diverse deepfake scenarios and datasets to ensure broad applicability.

4. Optimizing the detection algorithm for computational efficiency to facilitate real-time applications.

4 Work Progress and Achievements

4.1 Photo-response non-uniformity (PRNU) based deepfake detection

PRNU-based deepfake video detection involves analyzing the unique noise patterns inherent to each camera sensor. By extracting and comparing these noise fingerprints from images or videos, the system can detect inconsistencies that suggest manipulation. Techniques like Discrete Fourier Transform (DFT) are used to examine frequency-domain anomalies, while machine learning classifiers, such as SVM, help distinguish

between genuine and altered content. The method's effectiveness relies on its ability to handle various types of manipulation and artifacts introduced during deepfake creation.

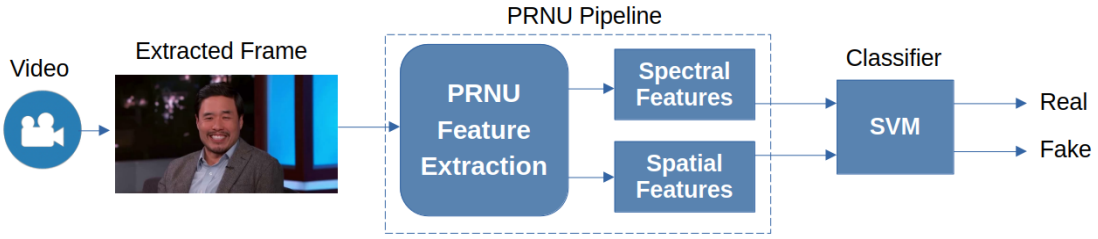


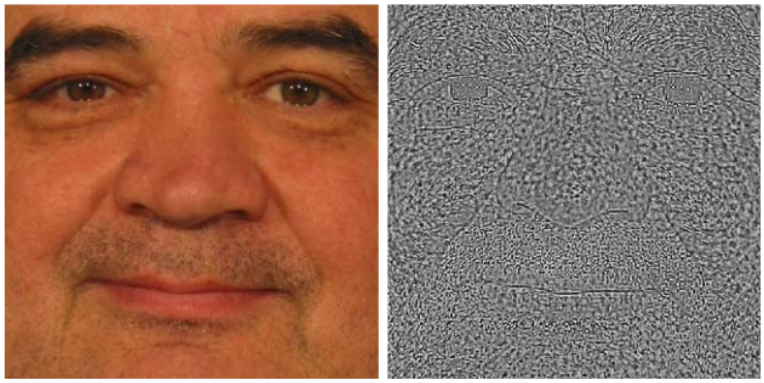
Figure 3: Overview of the pipeline structure of the PRNU based deepfake detection

4.1.1 PRNU extraction

The PRNU is a noise-like pattern created by slight variations in individual pixels during the photon-to-electron conversion in digital image sensors. This inherent pattern is embedded in every image captured by these sensors. It acts as an unintentional, stochastic spread-spectrum watermark that endures through processes like lossy compression or filtering. To extract the PRNU noise residual from an image, Fridrich's method [1] is used. For each image $\mathbf{I}$, the noise residual $\mathbf{W}_I$ is estimated according to Eq. (3).

$$\mathbf{W}_I = \mathbf{I} - F(\mathbf{I}) \tag{1}$$

where F is a denoising function designed to filter out the sensor pattern noise. In this work we use wavelet based denoising filter proposed by Mihcak et al.[10].



(a) (b)

Figure 4: Example of PRNU extraction (a) Original (b) PRNU

4.1.2 PRNU splitting

The system can work with the PRNU extracted from the entire image or by dividing the PRNU into multiple equisized cells. In this study, we are dividing it into $N = 8 \times 8$ cells.

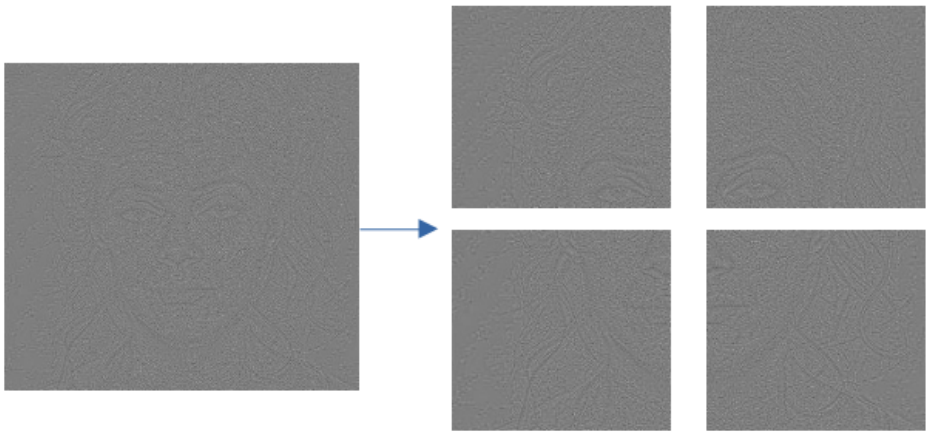
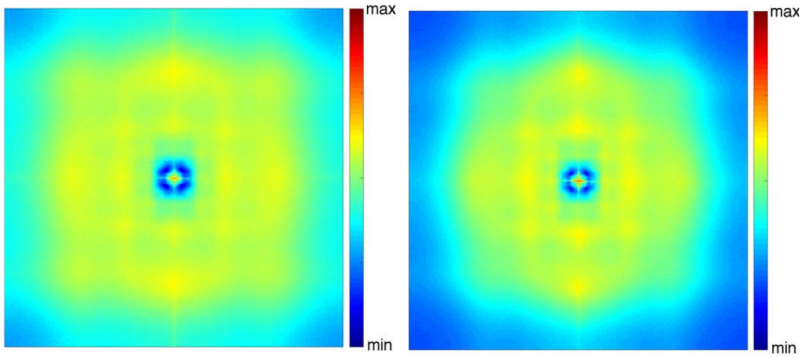


Figure 5: Example for splitting the PRNU into $N = 4$ Cells (2×2) of equal size.



(a) (b)

Figure 6: DFT magnitude spectra of the PRNUs extracted from (a) real (b) fake images

Increasing the number of cells is expected to highlight the non-linear transformations of the PRNU. Ultimately, this results in N distinct cells $C_1, \ldots, C_N$. Figure 4 illustrates an example where the PRNU is divided into $N = 2 \times 2$ equisized cells.

4.1.3 Cell-wise feature extraction

Feature extraction is conducted individually for each cell. The first step involves obtaining the frequency spectrum of the PRNU within each cell using the discrete Fourier transform (DFT). The resulting magnitude spectrum reveals how the PRNU signal has been altered by the manipulation process. To quantify these changes, we compute the histogram of the DFT magnitudes, which represents the magnitude distribution within the spectrum. Figure 6 presents the DFT magnitude spectra for both a real and a fake sample image, along with their corresponding histograms, where a shift in the magnitude distribution is noticeable. All DFT magnitude histograms are divided into 100 bins.

For every cell we compute the features listed [9] in Table 1,

Table 1: PRNU features

Feature type	Feature	Description
Spatial	P_{en}	Energy of PRNU values
	P_{var}	Variance of PRNU values
	P_{skew}	Skewness of PRNU values
	P_{kurt}	Kurtosis of PRNU values
	P_{varH}	Variance of values in PRNU histogram
	P_{maxH}	Position of maximum value in PRNU histogram
Spectral	D_{en}	Energy of DFT values
	D_{var}	Variance of DFT values
	D_{skew}	Skewness of DFT values
	D_{kurt}	Kurtosis of DFT values
	D_{varH}	Variance of values in DFT histogram
	D_{maxH}	Position of maximum value in DFT histogram

4.1.4 Features calculation

Energy of PRNU values:

$$P_{en} = \sum_{i=1}^{N} |x_i|^2, \tag{2}$$

Skewness of PRNU values:

$$P_{skew} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^3}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{\frac{3}{2}}}, \tag{3}$$

where,
N is the number of PRNU values,
x_i represents each PRNU value,
μ is the mean of the PRNU values.

Kurtosis of PRNU values:

$$P_{kurt} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^4}{\left(\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2\right)^{2}} - 3, \tag{4}$$

Variance of values in PRNU histogram:

$$P_{varH} = \frac{1}{B}\sum_{n=1}^{B}\left(H_p(n) - \bar{H}_p\right)^2, \tag{5}$$

Position of maximum value in PRNU histogram:

$$P_{maxH} = \underset{n=1,\ldots,B}{\operatorname{argmin}} H_p(n), \tag{6}$$

where,

N is the number of pixels,

x_i represents each pixel value,

μ is the mean of the pixel values,

B is number of bins in PRNU cell's histogram,

$H_p(n)$ PRNU cell's histogram,

$\bar{H}_p$ represents the mean frequency of the histogram bins.

Similarly, spectral features can be calculated from the DFT magnitude spectra of the PRNU cell.

4.1.5 Feature aggregation

For each feature, we calculate the mean and variance across all cells of an image, yielding a total of 24 features per image. This approach ensures that the features are independent of the image size, making them comparable across images of different sizes. Example:- Mean of feature P_{en}:

$$\bar{P}_{en} = \frac{1}{N} \sum_{n=1}^{N} P_{en}(n) \tag{7}$$

Variance of feature P_{en}:

$$P_{enVar} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} \left(P_{en}(n) - \bar{P}_{en}\right)^2} \tag{8}$$

where N is the total number of PRNU cells, $\bar{P}_{en}$ is calculated by simply averaging the P_{en} of the individual cells, while P_{enVar} represents the variance of all PRNU cells within an image.

4.1.6 Support vector machine (SVM) classification

After extracting the features from each image in the dataset, we use these features to train a support vector machine (SVM) for binary classification. The SVM categorizes the images into two classes: $fake$ representing images manipulated by a deepfake algorithm, and $real$ representing untampered, authentic images.

4.2 Datasets

For learning-based approaches, high-quality training data is crucial. To evaluate new methods effectively, it's important to compare results across multiple datasets with diverse characteristics. While the research community has released numerous datasets with image and video manipulations, not all are suitable for developing learning-based methods. Splitting a single dataset into training, validation, and test sets can lead to polarization or overfitting if not done carefully. The Table 2 reviews the most commonly used datasets.

Table 2: List of datasets including video manipulations

Dataset	Reference	Pristine / Forged	Frame size	Year
FaceForensics++	[11]	1,000 / 4,000	480p, 720p, 1080p	2019
Celeb-DF	[8]	590 / 5,639	various	2020
DeeperForensics-1.0	[5]	50,000 / 10,000	1080p	2020
DFDC	[2]	19,154 / 100,000	240p - 2160p	2019

4.3 Results:

Accuracy $\approx 61.3\%$, Error rate $\approx 38.5\%$

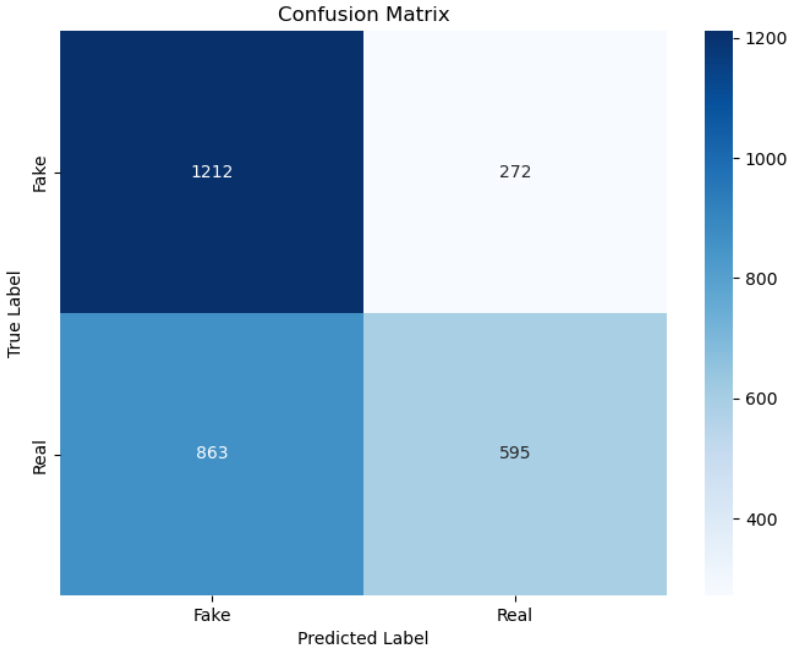


Figure 7: Confusion Matrix

5 Summary and Future Work

The rapid advancement of deepfake technology presents a growing challenge to digital media authenticity. Traditional detection methods, while powerful, often face limitations due to high computational demands and susceptibility to evolving manipulation techniques. In contrast, this study, which integrates PRNU analysis with Discrete Fourier Transform (DFT) and Support Vector Machine (SVM) classification, offers a compelling solution by focusing on the intrinsic noise-like patterns unique to camera sensors and leveraging frequency-domain features.

In my future work, I am addressing the challenges related to insufficient training data by developing a fully unsupervised deepfake detection framework. This approach will eliminate the need for any true label information during both the training and testing phases.

References

[1] Mo Chen, Jessica Fridrich, Miroslav Goljan, and Jan Lukás. Determining image origin and integrity using sensor noise. *IEEE Transactions on Information Forensics and Security*, 3(1):74–90, 2008.

[2] Brian Dolhansky, Joanna Bitton, Ben Pflaum, Jikuo Lu, Russ Howes, Menglin Wang, and Cristian Canton Ferrer. The deepfake detection challenge (dfdc) dataset. *ArXiv Preprint ArXiv:2006.07397*, 2020.

[3] Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7890–7899, 2020.

[4] Andrew C Gallagher and Tsuhan Chen. Image authentication by detecting traces of demosaicing. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8. IEEE, 2008.

[5] Liming Jiang, Ren Li, Wayne Wu, Chen Qian, and Chen Change Loy. 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*, pages 2889–2898, 2020.

[6] Marissa Koopman, Andrea Macarulla Rodriguez, and Zeno Geradts. Detection of deepfake video manipulation. In *The 20th Irish Machine Vision and Image Processing Conference (IMVIP)*, pages 133–136, 2018.

[7] Yuezun Li, Ming-Ching Chang, and Siwei Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. Ieee, 2018.

[8] Yuezun Li, Xin Yang, Pu Sun, Honggang Qi, and Siwei Lyu. 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, 2020.

[9] Florian Lugstein, Simon Baier, Gregor Bachinger, and Andreas Uhl. Prnu-based deepfake detection. In *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, pages 7–12, 2021.

[10] M Kivanc Mihcak, Igor Kozintsev, and Kannan Ramchandran. Spatially adaptive statistical modeling of wavelet image coefficients and its application to denoising. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No. 99CH36258)*, volume 6, pages 3253–3256. IEEE, 1999.

[11] Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1–11, 2019.

[12] Xin Yang, Yuezun Li, and Siwei Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.