TOWARDS THE DETECTION OF AI-SYNTHESIZED HUMAN FACE IMAGES

Yuhang Lu and Touradj Ebrahimi

Multimedia Signal Processing Group (MMSPG)
École Polytechnique Fédérale de Lausanne (EPFL)

ABSTRACT

Over the past years, image generation and manipulation have achieved remarkable progress due to the rapid development of generative AI based on deep learning. Recent studies have devoted significant efforts to address the problem of face image manipulation caused by deepfake techniques. However, the problem of detecting purely synthesized face images has been explored to a lesser extent. In particular, the recent popular Diffusion Models (DMs) have shown remarkable success in image synthesis. Existing detectors struggle to generalize between synthesized images created by different generative models. In this work, a comprehensive benchmark including human face images produced by Generative Adversarial Networks (GANs) and a variety of DMs has been established to evaluate both the generalization ability and robustness of state-of-the-art detectors. Then, the forgery traces introduced by different generative models have been analyzed in the frequency domain to draw various insights. The paper further demonstrates that a detector trained with frequency representation can generalize well to other unseen generative models.

***Index Terms*—** Synthetic face image, detection, GANs, diffusion models, frequency analysis

1. INTRODUCTION

In recent years, rapid advances have been made in image manipulation and synthesis techniques, such as generative adversarial networks (GANs) [1, 2, 3, 4, 5] and variational autoencoders (VAE) [6]. In practice, these deep learning-based techniques facilitate the creation of counterfeit images or video by manipulating the face of a person, which refers to the popular term "Deepfake". The generated human face images are often too realistic to be distinguished by human observers, raising social trust concerns due to their potential exploitation for malicious purposes. Consequently, considerable efforts have been dedicated to detecting face manipulations and promising progress has been demonstrated [7, 8, 9].

Nevertheless, another source of deepfake, i.e. entire face synthesis, has not received adequate attention so far. Various GAN-based models [1, 4, 5] have been designed to create face images that do not exist in the world and produce surprisingly realistic results. More recently, the surge of Diffusion Models (DMs) has started a new paradigm in photorealistic image synthesis and an increasing number of researchers have been using them to further improve the quality of produced results. With a publicly available model in the open-source community, one can easily create tons of fake human face images with little effort. Although this type of deepfake holds potential utility for applications such as video game character modeling, it can also be abused to create fake profiles for fraud or assist in spreading misinformation. Due to the diversity of different generative models, it remains a big challenge to develop a universal detection method that can identify synthetic face images created by arbitrary models.

Fig. 1: Realistic synthetic human face images generated by ProGAN [1], StyleGAN2 [4], DDPM [10], DDIM [11], PNDM [12], and LDM [13] respectively.

Fortunately, a growing number of detection methods [14, 15, 16, 17, 18, 19, 20] have been developed for purely AI-synthesized images. Some rely on simple training of convolutional neural network (CNN) classifiers with various data preprocessing or augmentation strategies [14, 18], while others exploit specific fingerprints left by the generation techniques [15, 16, 20]. Despite these advances, several concerns still persist in current studies. First, most of the detection methods only focus on images produced by a specific type of generative model. The generalization ability of such detectors to

Support from the Swiss National Science Foundation (SNSF) 20CH21_195532 for XAIface CHIST-ERA-19-XAI-011 is acknowledged.

images created by different GAN models or recent diffusion models is not sufficiently studied. Although recent studies [21, 22] have made preliminary progress in the right direction, their focus has predominantly centered on general categories of synthetic images with rich contextual information, such as bedrooms, outdoor churches, etc. This brings a second concern, i.e. whether detectors for generic fake images can perform well for synthetic human face images. Third, the resistance of a detector against common image perturbations, particularly on DM-generated face images, remains unexplored.

This paper addresses the challenges in detecting entirely AI-synthesized human face images. The primary contribution lies in the establishment of a new benchmark for this task, achieved by systematically generating a substantial volume of synthetic human face images using seven popular generative models. Subsequently, the generalization ability and robustness of various learning-based detectors have been evaluated with the benchmark. The paper also aims to draw new insights for developing more generalizable detectors. To that end, a frequency domain analysis on the synthetic face images is carried out, examining the deviation of their spectra from that of real images. Consequently, our experimental results demonstrate that training a learning-based detector using frequency representations yields outstanding performance and generalization ability in the benchmark.

2. RELATED WORK

2.1. Generative Models for Image Synthesis

Generative adversarial networks (GANs) have long stood as the prevailing approach for numerous image synthesis tasks. In general, a GAN [23] is trained through a competing game between two models, i.e., a generator and a discriminator. The generator aims to fool the discriminator by producing images resembling those in the training data, while the latter seeks to distinguish between real and generated images. In practice, some GAN models [1, 3] take noise as input and are able to generate high-resolution images with good perceptual quality, while others [2, 24] are conditioned on additional information, such as a semantic map or another image, often employed for translation between two images. This paper focuses on unconditional face image generation and adopts three GAN models that are pre-trained on high-quality face image datasets, namely ProGAN [1], StyleGAN2 [4], and VQGAN [5].

More recently, initially inspired by non-equilibrium thermodynamics [25], diffusion models have become a new paradigm for image generation. Ho et al. [10] proposed denoising diffusion probability models (DDPM) and showed an impressive ability in image synthesis in comparison to GAN-based counterparts. Song et al. [11] explored the use of the denoising diffusion implicit model (DDIM) to improve sampling speed while maintaining good image quality. ADM [26] introduced a more effective architecture incorporating classifier guidance and demonstrated superior performance when compared to GANs. Liu et al. [12] proposed PNDM and further enhanced the sampling efficiency and generation quality. A later work LDM [13] integrated text and image inputs in latent space via a cross-attention mechanism. When these diffusion models are trained on large-scale human face datasets, they are capable of generating realistic and high-quality face images. This paper includes four diffusion models in the benchmark, namely DDPM, DDIM, PNDM, and LDM.

2.2. Detection of AI-Synthesized Images

The need for fake image detectors has existed ever since the appearance of various generative models. Some detection methods leveraged hand-crafted features, such as color cues [27], saturation cues [28], blending artifacts [29], and gradients [20], while other studies relied on CNN-based classifiers to detect fake images. Several researchers have leveraged various advanced neural network architectures as primary solutions. For example, Rössler et al. [7] retrained XceptionNet [30] with a large-scale deepfake dataset. Cozzolino et al. [31] learned a forensic embedding through an autoencoder-based architecture to distinguish between real and fake images and performed well on StyleGAN-generated images. Marra et al. [32] tested multiple CNN-based architectures for detecting GAN-generated images.

However, most of the detection methods above only show good performance when the fake images share the same distribution as the training data. This is why, more attention has been recently devoted to the detector's generalization ability. Wang et al. [14] proposed to train a basic detection network with data preprocessed by JPEG compression and Gaussian Blur, and surprisingly generalized well on other unseen GAN-generated images. Grag et al. [33] improved based on [14] by updating the network architecture. Shiohara et al. [8] fine-tuned a pre-trained EfficientNetB4 [34] to detect blending boundary artifacts and achieved promising results in cross-data evaluation on several deepfake benchmarks. Mandelli et al. [18] leveraged an ensemble of multiple EfficientNetB4 that were trained under different conditions and achieved the state-of-the-art.

An increasing number of studies have been carried out to counteract the emerging realistic fake images created by diffusion models. DIRE [35] developed an effective method to detect DM-generated images by reconstructing an input image through a pre-trained diffusion model. Lorenz et al. [36] exhibited the superiority of multi-local intrinsic dimensionality in diffusion detection. Ojha [19] proposed a universal fake image detector by leveraging a pre-trained large vision-language model and achieved excellent generalization ability across GAN and DM-based fake images.

Table 1: Generative models used in this work, including three GAN models and four diffusion models. The quality of a face dataset produced by each model is reported with FID scores. 10k images are randomly sampled for each model to calculate FID score. A lower FID refers to higher quality.

Model Family	Method	Publication	FID
GANs	ProGAN	Karras et al. (2018) [1]	12.39
	StyleGAN2	Karras et al. (2019) [4]	15.17
	VQGAN	Esser et al. (2021) [5]	12.99
DMs	DDPM	Ho et al. (2020) [10]	16.64
	DDIM	Song et al. (2020) [11]	14.36
	PNDM	Liu et al. (2022) [12]	13.97
	LDM	Rombach et al.(2022) [13]	7.28

3. DETECTION BENCHMARK FOR SYNTHETIC HUMAN FACE IMAGES

This paper contributes a comprehensive benchmark for synthetic face image detection. This section first introduces the collected dataset along with the detectors incorporated into the benchmark. Then, two major objectives of the benchmark, i.e., evaluating the generalizability and robustness of a detector, and how to achieve them are elaborated.

3.1. A Dataset of Synthetic Face Images

This work first collects a dataset that comprises real images from the CelebA-HQ [1] dataset and synthetic human face images created by seven cutting-edge generative models. As shown in Table 1, the fake face images are synthesized by GANs, including ProGAN [1], StyleGAN2 [4], and VQGAN [5], and DMs, including DDPM [10], DDIM [11], PNDM [12], and LDM [13]. For LDM, the unconditional mode is employed because bimodal inputs (e.g. incorporating texts) will result in unnatural face images. More specifically, StyleGAN2 is pre-trained on the FFHQ [3] dataset and all other models are pre-trained on CelebA-HQ [1] dataset. The default resolution for the entire dataset is set to 256×256 because it is the most common output size among the selected generative models. Higher-resolution images generated by certain models are downscaled to 256×256 using bilinear interpolation. Under these settings, all models are capable of generating realistic human face images, see examples in Figure 1. The Fréchet inception distances (FID) reported in Table 1 further show that all models produce images of comparable quality.

For each generation technique, 40k images are collected in total and they are by default split into 38k, 1k, and 1k for training, validation, and testing purposes.

3.2. Detectors

Several learning-based methods for synthetic image detection are selected for experiments in the benchmark. All methods report satisfactory performance in prior studies on general fake image detection tasks. However, their performance specifically concerning synthetic human face images, their adaptability to DM-generated images, and their robustness against common perturbations have not been investigated until this paper. The selected detectors are outlined as follows.

Wang2020 [14] employed a ResNet-50 architecture and trained with JPEG compression and Gaussian blurring as augmentation, obtaining fair generalization ability among GAN-synthesized images. Grag2021 [33] built upon this approach by further exploring different variations of ResNet-50 to enhance performance in real-world scenarios. Mandelli2022 [18] used an ensemble of five orthogonal EfficientNetB4 [34] networks to detect fake images. Each model was trained on different datasets created by various GAN models and augmented using different techniques. This strategy significantly improved the overall performance and generalization ability. Ojha2023 [19] leveraged a large pre-trained vision-language model and exhibited exceptional generalization ability in detecting fake images across a variety of generative models.

3.3. Generalizability

One of the main objectives of the proposed benchmark is to assess the generalization ability of a detector in the presence of synthetic human face images. This can be interpreted in one of the following two ways: (i) whether a detector trained with other categories of fake images can effectively generalize to synthetic human face images; (ii) whether a detector trained on images created by a specific generative model can still obtain satisfactory performance on unseen GANs and DMs. To address the first way, this work employs open-source detection methods pre-trained on various categories of synthetic images, such as bridge, church, etc., sourced from the LSUN dataset [37]. They are directly evaluated by the synthetic face images from the benchmark. For the latter, training data is selected from one GAN model and one diffusion model in the benchmark and used to train detectors from scratch. Then, they are tested using the out-of-distribution fake face images synthesized by other GANs and DMs.

3.4. Robustness against Image Perturbation

Synthetic face images often undergo various processing operations before dissemination, such as compression, resizing, etc. Therefore, in addition to the generalization ability, the proposed benchmark further measures the robustness of detectors against common image perturbations. Specifically, the impact of the following perturbations is analyzed:

- **JPEG compression** is performed and the impact of different quality factors are measured individually, i.e., $\{10, 20, \ldots, 90\}$.
- **Blurry effect** is applied via a Gaussian Blur kernel. The kernel size is selected from $\{3, 5, \ldots, 15\}$.

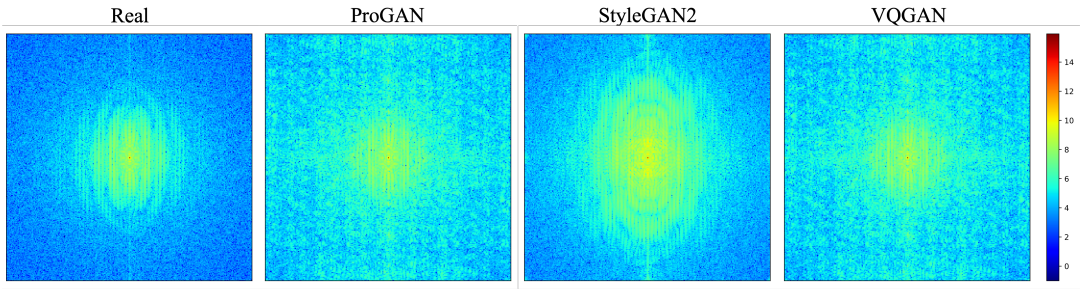


Fig. 2: Mean frequency spectra of real images from CelebA-HQ [1] and synthetic human face images created by three GAN models, namely ProGAN [1], StyleGAN2 [4], and VQGAN [5].

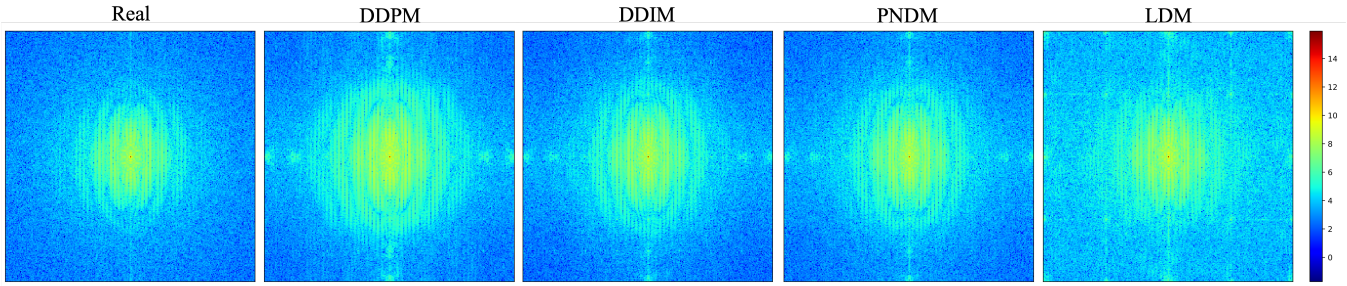


Fig. 3: Mean frequency spectra of real images from CelebA-HQ [1] and synthetic human face images created by four diffusion models, namely DDPM [10], DDIM [11], PNDM [12], and LDM [13].

- **Gaussian noise** with zero mean is added and the standard deviation is selected from $\{5, 10, \ldots, 30\}$.
- **Resizing operation** is employed by first downsampling the image to lower resolutions by a scale of $\{2, 4, \ldots, 12\}$, with bicubic interpolation and then upscaling to 256×256.

Notably, only one type of perturbation of a fixed intensity is applied to the entire test set in each evaluation to avoid randomness.

4. FREQUENCY ARTIFACTS ANALYSIS

As the generation tools become more advanced, their results become indistinguishable from real images when observed by human subjects in the spatial domain and even some CNN-based detectors. Studies [38, 14, 21, 22] have identified characteristic fingerprints present in GAN-generated images via frequency analysis and observed grid-like artifacts in general categories of synthetic images.

This section analyzes forgery traces in the frequency domain, particularly for synthetic human face images created by various generative models. As suggested by prior work [14], each image is first converted to gray-scale by averaging over color channels and then high-pass filtered by subtracting a median-filtered version of itself. Subsequently, the Fast Fourier Transform (FFT) is applied to the processed image to extract the frequency spectrum, with magnitude values log-scaled for better visualization. Figure 2 and 3 depict the average frequency spectrum of 1,000 images randomly sampled from the real CelebA-HQ dataset and seven fake face image datasets created generative models listed in Table 1.

As shown in Figure 2, the common grid-like artifacts found in previous studies [14, 38] are notably absent in our GAN-generated face datasets. However, datasets created by ProGAN and VQGAN exhibit numerous high-frequency noises. The more advanced StyleGAN2 contains relatively fewer such artifacts but remains distinguishable from real image spectra. On the other hand, Figure 3 shows that the FFT spectra of DM-created face images closely resemble the real spectrum, except for LDM which contains both high-frequency noise and grid-form artifacts. While images produced by DDPM, DDIM, and PNDM exhibit fewer visible artifacts in the frequency domain, they tend to have higher spectra density and contain low-frequency artifacts along the vertical and horizontal impulse sequence, deviating from that of real image spectra.

Observing notable discrepancies between real and synthetic face images in their frequency representations, this paper further explores the potential utility of these differences in training a more generic detector that can identify fake face images generated by various GANs and DMs. Specifically, the detection task is framed as a binary classification process and three basic classification networks are selected, namely ResNet-50 [39], XceptionNet [30], and EfficientNetB4 [34]. These networks are trained with only frequency representa-

Table 2: Detection performance of four pre-trained detectors. The weights released by the original authors are utilized.

AUC/AP (%)	GANs			DMs				Average
	ProGAN	StyleGAN2	VQGAN	DDIM	DDPM	PNDM	LDM	
Wang2020 [14]	78.31/78.09	88.39/88.34	79.80/79.84	74.51/70.94	65.09/60.58	76.38/73.40	77.10/76.51	77.08/75.39
Grag2021 [33]	99.96/99.96	99.27/99.29	99.32/99.35	68.85/62.45	59.22/52.93	66.82/61.84	99.67/99.69	84.73/82.22
Mandelli2022 [18]	100.00/100.00	100.00/100.00	99.43/99.36	99.99/99.99	97.87/97.70	98.16/98.02	98.76/99.00	99.17/99.15
Ojha2023 [19]	96.38/96.52	73.24/69.57	96.20/96.27	97.78/97.97	93.31/93.57	96.37/96.55	98.18/98.19	93.42/92.66

Table 3: Generalization analysis of various detection techniques. All methods are trained on face images generated by ProGAN and DDIM, and tested on images created by all seven generative models. To distinguish from prior experiments, $*$ here refers to the retrained version of Wang2020 and Ojha2023 using our training set. The best result is denoted by underscore.

AUC/AP (%)	GANs			DMs				Average
	ProGAN	StyleGAN2	VQGAN	DDIM	DDPM	PNDM	LDM	
ResNet-50	100.00/100.00	52.95/50.52	100.00/100.00	61.60/61.07	73.32/71.25	60.02/60.93	37.97/43.86	69.41/69.66
XceptionNet	100.00/100.00	57.65/57.21	100.00/100.00	53.59/59.16	52.26/51.30	57.51/59.16	39.94/44.65	65.85/67.35
EfficientNetB4	100.00/100.00	52.26/50.70	100.00/99.99	77.69/78.22	71.20/70.91	81.46/81.83	61.19/62.42	77.69/77.72
Wang2020* [14]	100.00/100.00	52.41/52.98	100.00/100.00	90.28/90.34	85.17/84.42	86.97/87.27	64.77/65.52	82.80/82.93
Ojha2023* [19]	99.97/99.97	94.80/94.58	99.97/99.97	99.72/99.74	98.60/98.68	99.56/99.58	99.94/99.94	98.94/98.92
ResNet-50+FreqSpec	99.79/99.78	93.60/92.96	99.84/99.83	98.38/98.37	98.77/98.59	99.30/99.43	99.55/99.49	98.46/98.35
XceptionNet+FreqSpec	99.45/99.53	95.26/95.03	99.64/99.66	99.24/99.16	99.03/98.89	99.66/99.71	99.64/99.66	98.85/98.81
EfficientNetB4+FreqSpec	99.87/99.89	98.72/98.68	99.95/99.95	99.53/99.53	99.54/99.53	99.97/99.97	99.94/99.94	99.65/99.64

tions of both real and synthetic face images. Further details about experimental setups and results are presented in the next section.

5. DETECTION PERFORMANCE

5.1. Experimental Setup

Experiments are structured into three phases. In phase one, four popular detection methods are first evaluated on our benchmark, namely Wang2020 [14], Grag2021 [33], Ojha2023 [19], Mandelli2022 [18]. The pre-trained weights directly released by the authors are used. The former three detectors were originally trained on general categories of synthetic images sourced from the LSUN dataset [37], while the latter was trained on a synthetic face dataset [40] for comparative analysis. The four detectors are tested with all seven test sets provided by the benchmark.

In the second phase, Wang2020 and Ojha2023 are selected to evaluate their generalization ability across various generative models. Additionally, three CNN classifiers (ResNet-50, XceptionNet, EfficientNetB4) and their counterparts trained with frequency representations are also evaluated under the same setting. In detail, all the detectors are trained on real images sourced from CelebA-HQ, and fake images created by one GAN model (ProGAN) and one diffusion model (DDIM). The number of real images is upsampled accordingly to ensure a balanced training set. Then, tests are performed under the same configuration as in phase one.

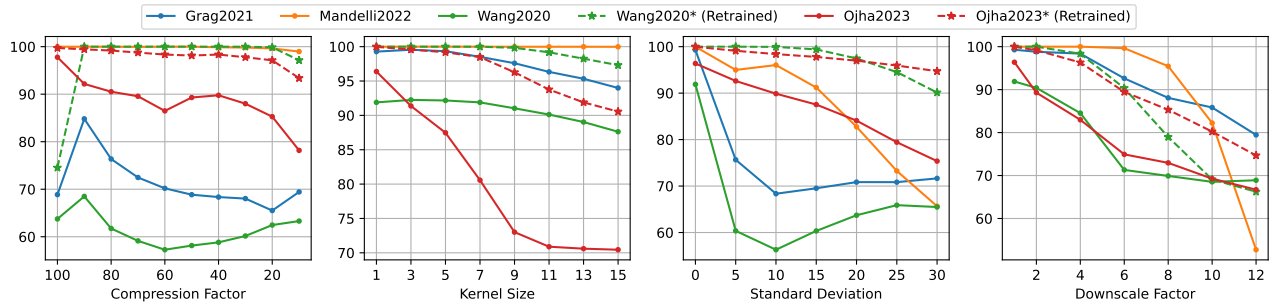
In phase three, the robustness of detectors against four common image perturbations is assessed. The evaluation involves four pre-trained detectors from phase one, tested on distorted face images generated by ProGAN and DDIM. Furthermore, Wang2020 and Ojha2023 retrained on the corresponding training set are also incorporated and tested under the same conditions.

5.2. Evaluation Metrics

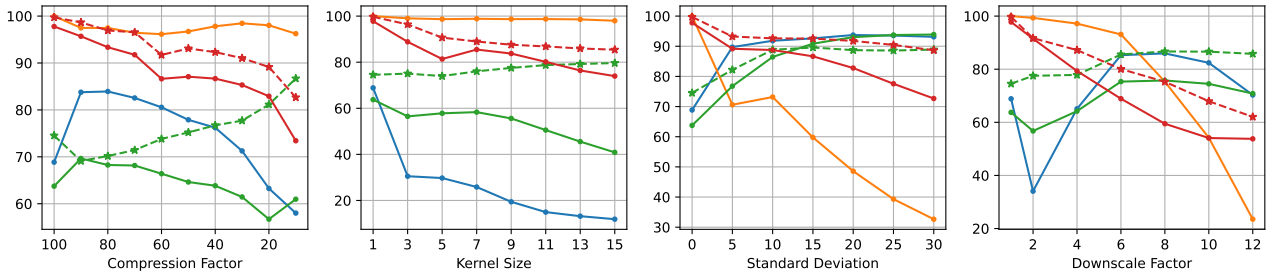
Following previous work about synthetic image detection and benchmarking, the average precision (AP) and Area Under Receiver Operating Characteristic Curve (AUC) scores are used to evaluate the detectors.

5.3. Experimental Results

This section presents the results of the three-phase evaluation. First of all, the performance of four pre-trained detectors is reported in Table 2. Wang2020 shows poor performance on both GAN and DM-generated face images, although in previous study it reported fair adaptability among general categories of fake images. Grag2021 is able to generalize among face images created by different GAN models but fails to achieve good performance on three diffusion models, i.e., DDPM, DDIM, and PNDM. Ojha2023 demonstrates good transferability across face images synthesized by most GANs and DMs, except for StyleGAN2. In comparison, the Mandelli2022 detector that has been trained on a GAN-created face dataset shows exceptional performance in our benchmark. To sum up, detectors only trained on general fake images struggle to adapt to synthetic face images.



(a) Evaluation performance on test set created by ProGAN.



(b) Evaluation performance on test set created by DDIM.

Fig. 4: Performance of various detectors under the perturbation of JPEG compression, Gaussian blur, Gaussian noise, and resizing operation (from left to right). The evaluation is conducted on two test sets created by ProGAN and DDIM respectively.

Secondly, Table 3 summarizes detectors' performance after being trained with fake face images from the benchmark. As a result, the retrained version of Wang2020 shows the potential to generalize to images created by VQGAN, DDPM, and PNDM, yet struggles to adapt to StyleGAN2 and LDM. Conversely, Ojha2023 achieves nearly flawless detection across all the GANs and DMs. Notably, after training with frequency representations of these face images, the three CNN detectors achieve much better generalization ability when compared to their counterparts that are directly trained with RGB images. The combination of EfficientNetB4 and frequency representation even surpasses the state-of-the-art performance on certain GAN models and most DMs.

Thirdly, Figure 4 illustrates the robustness of various detectors under four common image perturbations. The four solid lines represent the performance of the four pre-trained detectors from phase one. Although both Wang2020 and Grag2021 incorporate data augmentation techniques, they are notably affected by compression artifacts and noise. Similarly, the performance of Ojha2023 deteriorates as the perturbation intensity increases. Mandelli2022 remains the most robust among the four, particularly in handling data subjected to JPEG compression and Gaussian blur effect. However, its performance inevitably declines in the presence of heavy noise or low-resolution effects. The two dashed lines additionally depict the performance of the retrained version of Wang2020 and Ojha2023 using the training set generated by ProGAN and DDIM. While the overall evaluation results improve, Figure 4b reveals that they are not resilient enough to compression artifacts and low-resolution effects.

6. CONCLUSION

This paper addressed detection of entirely AI-synthesized human face images. A comprehensive benchmark was devised to assess fake image detectors in terms of adaptability and robustness. Results show that detectors only trained on general categories of fake images have difficulty generalizing to synthetic face images. The generalization across various GANs and DMs and robustness against perturbations also remain two important challenges in most detection methods. Furthermore, the paper examined forgery traces of synthetic face images in the frequency domain and demonstrated that training a detector with frequency representation can significantly enhance its performance and generalization ability.

7. REFERENCES

[1] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," *arXiv preprint arXiv:1710.10196*, 2017.

[2] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[3] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

[4] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. CVPR*, 2020.

[5] Patrick Esser, Robin Rombach, and Bjorn Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12873–12883.

[6] Diederik P Kingma and Max Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[7] 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*, 2019, pp. 1–11.

[8] Kaede Shiohara and Toshihiko Yamasaki, "Detecting deepfakes with self-blended images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18720–18729.

[9] Yuhang Lu and Touradj Ebrahimi, "Assessment framework for deepfake detection in real-world situations," *arXiv preprint arXiv:2304.06125*, 2023.

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[11] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.

[12] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao, "Pseudo numerical methods for diffusion models on manifolds," *arXiv preprint arXiv:2202.09778*, 2022.

[13] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10684–10695.

[14] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros, "Cnn-generated images are surprisingly easy to spot...for now," in *CVPR*, 2020.

[15] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi, "Do gans leave artificial fingerprints?," in *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2019, pp. 506–511.

[16] Ning Yu, Larry S Davis, and Mario Fritz, "Attributing fake images to gans: Learning and analyzing gan fingerprints," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 7556–7566.

[17] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro, "Training CNNs in presence of JPEG compression: Multimedia forensics vs computer vision," in *IEEE International Workshop on Information Forensics and Security (WIFS)*, 2020.

[18] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro, "Detecting gan-generated images by orthogonal training of multiple cnns," in *2022 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2022, pp. 3091–3095.

[19] Utkarsh Ojha, Yuheng Li, and Yong Jae Lee, "Towards universal fake image detectors that generalize across generative models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 24480–24489.

[20] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yunchao Wei, "Learning on gradients: Generalized artifacts representation for gan-generated images detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12105–12114.

[21] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer, "Towards the detection of diffusion model deepfakes," *arXiv preprint arXiv:2210.14571*, 2022.

[22] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva, "On the detection of synthetic images generated by diffusion models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[23] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio, "Generative adversarial nets," *Advances in neural information processing systems*, vol. 27, 2014.

[24] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[25] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[26] Prafulla Dhariwal and Alexander Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[27] Scott McCloskey and Michael Albright, "Detecting gan-generated imagery using color cues," *arXiv preprint arXiv:1812.08247*, 2018.

[28] Scott McCloskey and Michael Albright, "Detecting gan-generated imagery using saturation cues," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 4584–4588.

[29] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5001–5010.

[30] François Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1251–1258.

[31] Davide Cozzolino, Justus Thies, Andreas Rössler, Christian Riess, Matthias Nießner, and Luisa Verdoliva, "Forensictransfer: Weakly-supervised domain adaptation for forgery detection," *arXiv preprint arXiv:1812.02510*, 2018.

[32] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva, "Detection of gan-generated fake images over social networks," in *2018 IEEE conference on multimedia information processing and retrieval (MIPR)*. IEEE, 2018, pp. 384–389.

[33] Diego Gragnaniello, Davide Cozzolino, Francesco Marra, Giovanni Poggi, and Luisa Verdoliva, "Are gan generated images easy to detect? a critical analysis of the state-of-the-art," in *2021 IEEE international conference on multimedia and expo (ICME)*. IEEE, 2021, pp. 1–6.

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

[35] Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, Hezhen Hu, Hong Chen, and Houqiang Li, "Dire for diffusion-generated image detection," *arXiv preprint arXiv:2303.09295*, 2023.

[36] Peter Lorenz, Ricard L Durall, and Janis Keuper, "Detecting images generated by deep diffusion models using their local intrinsic dimensionality," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 448–459.

[37] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao, "Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop," *arXiv preprint arXiv:1506.03365*, 2015.

[38] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz, "Leveraging frequency analysis for deep fake image recognition," in *International conference on machine learning*. PMLR, 2020, pp. 3247–3258.

[39] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[40] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila, "Alias-free generative adversarial networks," *Advances in Neural Information Processing Systems*, vol. 34, pp. 852–863, 2021.