Available online at www.qu.edu.iq/journalcm
JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS
ISSN:2521-3504(online) ISSN:2074-0204(print)

Enhanced Detection of Diffusion Model-Generated Deepfakes Using CNN Feature Maps and Forgery Traces

Banan Jamil Awrahman[a],Zhir Jamil Awrahman[b], Chya Fatah Aziz[c]

[a] *Information Technology, Halabja Technical Institute , Sulaimani Polytechnic University , Sulaimani, Iraq. Email:Banan.awrahman@spu.edu.iq*

[b]*Computer engineening, Komar University, Sulaimani, Iraq .Email: Zhirjamilf21@komar.edu.iq*

[C]*Halabja Technical College, Sulaimani Polytechnic University, Sulaimani , Iraq . Email: Chia.aziz@spu.edu.iq*

ARTICLE INFO

Article history:
Received: 29 /11/2024
Rrevised form: 4 /12/2024
Accepted : 26 /12/2024
Available online: 30 /12/2024

Keywords:

GAN,

CNN,

DDPM,

LDM,

VQGAN,

Celeb-HQ

ABSTRACT

Recent advancements in generative artificial intelligence powered by deep learning have significantly improved image generation and manipulation, resulting in highly realistic images that pose substantial risks to multimedia security. The increasing similarity between authentic and deepfake images, especially those generated by Diffusion Models, highlights the pressing need for effective detection mechanisms. This study provides a comprehensive evaluation of counterfeit facial image detection, focusing on the generalizability and robustness of various detection methods. Using a dataset comprising real images from CelebA-HQ and synthetic images generated by five state-of-the-art models (StyleGAN2, VQGAN, DDPM, PNDM, and LDM), we benchmark four leading detection algorithms: Wang2020, Grag2021, Mandelli2022, and Ojha2023.

The performance of these detectors was evaluated across different generative models and under various image perturbations, such as resizing, noise, blur, and compression. Additionally, we analyzed frequency-domain artifacts, revealing that GAN-generated images exhibit distinct frequency patterns, whereas DM-generated images closely resemble authentic ones. A novel hybrid approach combining spatial and frequency-domain features was proposed, yielding superior performance in detecting AI-generated human faces. Among the methods tested, Mandelli2022 achieved an AUC of 98.38%, while our ResNet-50+FFT model outperformed it slightly with an AUC of 98.42%. These results highlight the effectiveness of hybrid approaches in improving detection accuracy. However, detectors still face challenges in generalizing to diverse datasets, emphasizing the need for more adaptable and robust detection strategies.

MSC..

https://doi.org/ 10.29304/jqcsm.2024.16.41790

∗Corresponding author: Banan Jamil Awrahman

Email addresses: banan.awrahman@spu.edu.iq

Communicated by ‘sub etitor’

1. Introduction

Over the past years, advances in manufacturing have made edited media more and more realistic and available. Even though, creative and beneficial applications can be attained easily through image generation such as entertainment and advertising, but rather ethical concerns formed as deep learning techniques where images can masquerade as humans, highlight the potential harms of media manipulation and artificial generation. For example deep learning techniques that change the speeches of politicians [1] can be a potential threat since it can spread misleading information, threats to individuals, and harm reputations[2]. Recently, the techniques have become an integral part of fraudulent schemes, results in scams as high as $25 million [3]. Given the emerging use of social media and online media, it's not surprising that AI-generated images are becoming more intrusive in people's lives.

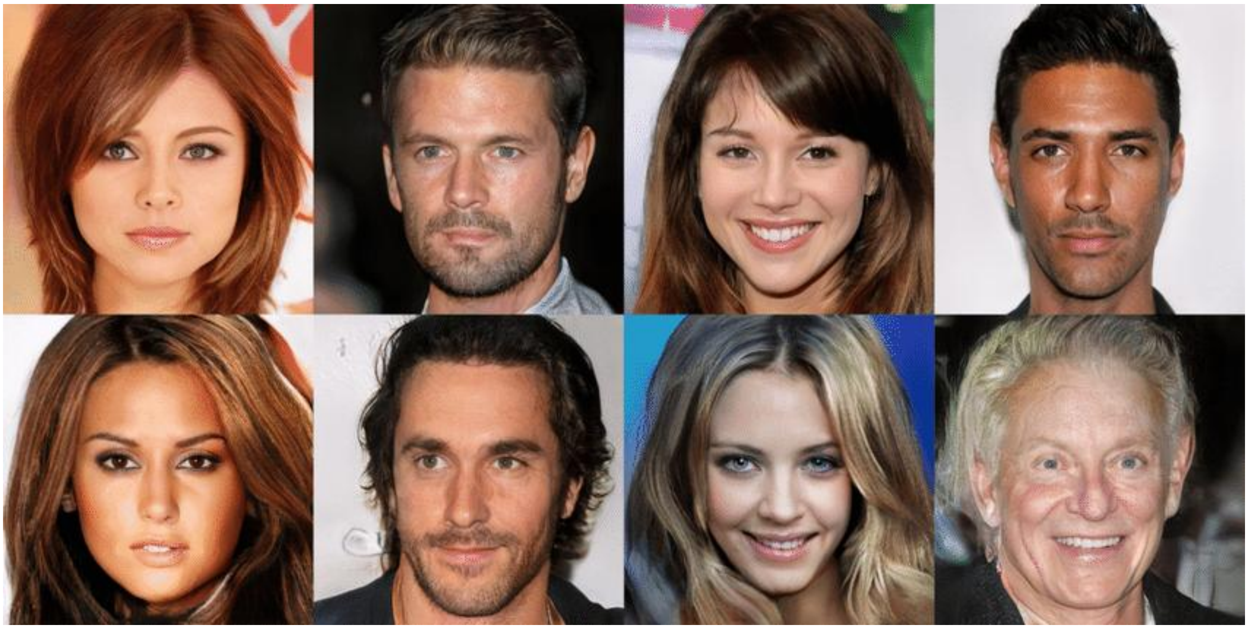

Fig. 1: Images Of Human Faces, Though Entirely Synthetic, Created By A GAN Trained Using The CelebA-Hq Dataset [29].

Recently image generation evolved rapidly by introducing generative adversarial networks (GANs) [4,5,6,7,8] and variational autoencoders (VAE)[9]. In practical applications, these techniques grounded in deep learning promote the dissemination of misleading or deceptive information through the manipulation of images, commonly referred to by the contemporary term "Deepfake". The human face images produced exhibit a remarkable level of detail, often rendering them nearly indistinguishable from authentic images to the human eye. As a result, numerous methodologies aimed at the identification of counterfeit content have been articulated. [10, 11, 12].

Furthermore, several platforms, both commercial and open-source, have contributed to bringing this technology closer to the people, such as Stable Diffusion [17], Imagen [15], Dreambooth [16], Dall-e 2 [14], and Midjourney [13]. However, diffusion models (DMs) have facilitated image generation with high-quality, while the presence of manipulated content is increasing leading to hidden privacy issues [18, 19].

Nevertheless, Various GAN-based models [4, 7, 8] to generate non worldly face images produce surprisingly realistic results, but their output often contains patterns beyond what CNN can learn to detect effortlessly. Recently, Diffusion Models (DMs) have ushered in a new era in creating photorealistic images and many researchers have been carrying out to implement them to further enhance the production quality. Due to open-source communities making their model publicly accessible, individuals can easily create countless images of fake human faces without any effort. While this type of deepfake is potentially useful for applications such as entertainment and advertising, it can also be abused to create fraudulent schemes or assist in spreading misinformation. It is impossible to develop a single detection method that can classify generated face images generated by random models due to the diversity of generative models.

Fortunately, detection methods [20, 21, 22, 23, 24, 25, 26] have been developed con- currently with generation methods. Some depends on easy training of convolutional neural network (CNN) classifiers with several dataset augmentation or preprocessing strategies [20, 24], while others utilizes special patterns left by the generation techniques [21, 22, 26]. Despite these advancements, numerous issues continue to persist in contemporary experimental frameworks. Primarily, a significant number of detection methodologies predominantly depend on extensive datasets for the purpose of training. Consequently, these methodologies primarily focus on images generated by a particular category of generative model. The generalization capabilities of such detectors have not been adequately explored. Although recent investigations [27, 28] have achieved some progress in the appropriate direction, their emphasis has been predominantly on broad categories characterized by abundant contextual information, such as outdoor churches, bedrooms, and similar environments. This introduces a new challenge. Whether these detectors can achieve success for synthetic images of the human face. Also, a detector's resistance to normal image degradation, especially on DM-generated face images.

This paper focuses on detecting AI-synthesized human face images, particularly those generated by Diffusion Models (DMs). To achieve this, a novel feature fusion pipeline is introduced, comprising two distinct branches: one processes spatial features of an image using a ResNet-50, while the other extracts frequency-domain features. These feature maps are then concatenated and passed through fully connected layers to determine the authenticity of the input image.

The proposed architecture demonstrates exceptional performance and generalization capabilities, validated through benchmarks on synthetic human faces generated by seven popular generative models, including StyleGAN2, VQGAN, DDPM, PNDM, and LDM. The feature fusion pipeline effectively combines spatial and frequency-domain features, enabling robust classification of synthetic images. Benchmark results show the model surpasses state-of-the-art methods such as Mandelli2022 and exhibits strong resilience against common image perturbations, including resizing, noise, and compression.

Furthermore, this study emphasizes detecting DM-generated face images, addressing a critical gap in the current detection landscape. By integrating innovative methodologies and tackling previously underexplored challenges, this research advances the field of synthetic image detection and sets a new standard for future studies.

2. Related Works

In this section, explore a brief overview of existing systems for generating deep- fake images such as "Generative adversarial networks (GANs) and Diffusion models (DMs)", and examine different detection and analysis techniques.

2.1. Generative Models

Generative adversarial networks (GANs) are one method of image synthesis tasks. A GAN [31] is trained through competition between two models, i.e., a generator and discriminator. The discriminator seeks to classify whether an image is real or fake, while the generator aims to trick the discriminator by creating images like those in the dataset. Practically speaking, some GAN models [4, 6] take noise as input and can extract high-resolution images, while others [5, 32] are conditioned on additional information, such as another image. In this paper we will only focus on un- conditional face image generation and adopt three pre-trained GAN models, namely- ProGAN [4], StyleGAN2 [7], and VQGAN [8].

Lately, the Diffusion Models have emerged as a preeminent methodology for image generation, with their conceptual foundations rooted in non-equilibrium thermodynamics [33]. Ho et al. [34] introduced denoising diffusion probability models (DDPM), which have exhibited remarkable efficacy in image generation when juxtaposed with the GANs paradigm [6]. Song et al. [35] subsequently presented denoising diffusion implicit models (DDIMs) to accelerate the image generation process while upholding the integrity of image quality. Further research, namely the ablated diffusion model (ADM) [36] proposed a more efficient architectural framework enhanced by classifier support. In subsequent investigations, Liuet al. [37] introduced pseudo numerical methods for diffusion models (PNDMs), resulting in augmented sampling efficiency and enhanced generative quality. Ultimately, the latent diffusion model (LDM) [38], which has exhibited superior performance relative to its diffusion counterparts, employs a cross-attention mechanism predicated on transformers [39] to effectively integrate image and text within the latent space. The generation performance has been significantly improved by stable diffusion v2, which is founded upon the principles of LDM and minimizes computational demands. These diffusion models demonstrate the capability to generate high-fidelity facial images. This paper encompasses four diffusion models within the benchmark, specifically DDPM, DDIM, PNDM, and LDM.

2.2. Detection Models

The necessity for synthetic image generators has been present since the inception of image synthesis technology. Some detection methodologies have employed various features, including chromatic indicators [39], saturation indicators [40], blending anomalies [41], and gradient pattern [27], whereas alternative investigations rely on convolutional neural network (CNN)-based classifiers to identify counterfeit images. Scholars have adopted sophisticated neural network architectures as principal methodologies. For example, Rossler et al. [10] implemented the training of Xcep- tionNet [30] utilizing a comprehensive deepfake dataset. Marra et al. [43] examined several CNN-based architectures for the identification of deepfake imagery.

Notwithstanding, the efficacy of most of the models declines in empirical assessments due to insufficient generalizability. Consequently, researchers have redirected their focus toward enhancing the generalization capabilities of detectors. Wang et al. [20] introduced a model exhibiting robust generalization on previously unencountered GAN-generated images by training a fundamental detection network using data subjected to JPEG compression and Gaussian Blur preprocessing. Grag et al. [45] augmented the work of [20] by refining the configuration of the network architecture. Shiohara et al. [11] fine-tuned a pre-trained Effi- cientNetB4 [46] and realized favorable outcomes in cross-data evaluations across multiple deepfake benchmarks. Mandelli et al. [24] trained several instances of Efficient NetB4 under varied conditions, thereby achieving state of the art performance.

The investigations conducted by Dong et al. [47] and Ricker et al. [27 have elucidated that detectors designed for recognizing GAN-generated images have become obsolete, as they are predicated on the extraction of trainable noise patterns, which no longer hold value in the context of DM-generated images. Wang et al. [44] ascertained that images produced by diffusion models possess characteristics that facilitate reconstruction more readily than their natural counterparts. The Diffusion Reconstruction Error (DIRE) metric is employed to discern such characteristics. Lorenz et al. [48] substantiated the successful attainment of multi-local intrinsic dimensionality within the domain of diffusion detection. Amoroso et al. [2] illustrated the segregation of semantic and stylistic features in images, further demonstrating that a heightened level of separability for style domains exists within synthetic images. Nevertheless, the implementation of semantic-style decoupling presents a considerable challenge, as it necessitates specifically tailored training datasets.

This study presents a novel feature fusion pipeline that integrates spatial features extracted from ResNet-50 with frequency-domain features, enabling superior detection capabilities across a diverse range of synthetic images, including those generated by both GANs and diffusion models (DMs). This hybrid framework demonstrates enhanced generalization and robustness, consistently outperforming state-of-the-art methods such as Mandelli2022 [24], even under challenging image perturbations like resizing, noise, and compression. A significant contribution of this work lies in addressing the challenges posed by DM-generated face images, a critical gap in existing research. The proposed approach undergoes comprehensive evaluation across seven generative models, including advanced diffusion models, underscoring its versatility and resilience in detecting synthetic images.

By combining spatial and frequency-domain features, this unified framework effectively addresses the limitations of prior methods, offering robust generalization across diverse generative models. The proposed method adapts to the rapidly evolving landscape of synthetic image generation, providing a scalable and reliable solution for detecting AI-generated content.

3. Methodology:

This manuscript undertakes a thorough evaluation of counterfeit facial image identification. In this section, both the dataset and the benchmarking process are delineated. Thereafter, two fundamental aims of the benchmark, specifically the generalizability and robustness of a detection mechanism, and identifying methodologies to achieve these objectives.

3.1. Dataset

The dataset used for this experiment is composed of real images from CelebA-HQ [4] Synthetic human face datasets and images generated by the five latest synthetic models, namely StyleGAN2 [7], VQGAN, DDPM [34], PNDM [49], and LDM [17]. For this only unconditional mode is employed to avoid unnatural face images. Moreover, each of the models undergoes training utilizing the CelebA-HQ [4] dataset. A uniform resolution of 256x256 is employed throughout the entire dataset because it has a ubiquitous output size in the selected models. Lower or higher resolution created by certain model rescaled to 256x256 by using bilinear interpolation. Under these conditions, all models can effectively generate realistic human face images. For each generation model, a total of 10kimages are gathered and they are split into: 8k, 1k, and 1k for training, testing, validation.

3.2. Data Preprocessing

Before training and testing, the dataset underwent preprocessing to ensure uniformity. First, all images were rescaled to 256x256 pixels, matching the standard output size of the generative models. Subsequently, the dataset was normalized using the calculated mean and standard deviation to standardize input data distribution. Once normalization ensured uniform data distribution, the dataset was prepared for the training phase.

A total of 4 deepfake detection algorithms has been implemented for experiments in our benchmark. All methods have achieved sufficient performance on general fake image tasks. However, their performance is dropped concerning the adaptability with DM generated images. The following detectors have been selected for the experiment. Wang2020[14]

The selected detectors are outlined as follows.

- Wang2020 [20] performed data preprocessing with Gaussian blurring and JPEG compression then feeding it into a ResNet-50, attaining sufficient generalization ability for GAN generated images.

- Grag2021 [45] explored further variations of ResNet-50 to improve performance in practical scenarios.
- Mandelli2022 [24] used an ensemble of five EfficientNetB4[46] networks to detect synthesized images. Different datasets created by various GAN models and different augmentation techniques were used to train each model separately.This strategy resulted in a significant improvement in overall performance.
- A large pre- trained vision-language model was utilized by Ojha2023 [25] and manifested rare generalization ability in detecting synthesized images.

These detectors were rigorously evaluated to analyze their performance, adaptability, and limitations, particularly when applied to DM-generated images

3.3. Generalizability And Robustness

In this subsection, we evaluate the capacity for generalization of a detector when confronted with artificially generated human facial images. This inquiry can be approached through two distinct methodologies. The first involves training the detector using a multitude of categories of synthetic images to enhance its ability to generalize across various representations of human faces. To implement this initial strategy, this research employed the LSUN dataset [53], which comprises a wide array of synthetic image categories, including but not limited to office and store environments. Conversely, the second methodology investigates whether a detector trained on datasets derived from a specific generative model can maintain satisfactory performance when applied to previously unseen Generative Adversarial Networks (GANs) and diffusion models (DMs). To address this methodological concern, the training data is obtained from a consortium of GAN and diffusion models, followed by subsequent testing against previously unencountered data.

Besides generalization ability, the benchmark additional steps are the robustness of detectors against common image perturbation. Particularly effectiveness of the succeeding perturbation is analyzed.

- The operation of image resizing is conducted by initially down sampling the image to reduced resolutions utilizing a scale of {2, 4, ..., 10}, employing bicubic interpolation, followed by an upscale to dimensions of 256×256.
- Gaussian noise characterized by a mean of zero is introduced, with the standard deviation being selected from the set {5, 10, . . ., 25}.
- The application of a blurry effect is achieved through the implementation of a Gaussian Blur kernel. The dimensions of the kernel are chosen from the range {3, 5, . . ., 13}.
- JPEG compression is executed, and the effects of various quality factors are assessed individually, specifically from the set {10, 20, ..., 100}.

3.4. Frequency Analysis

As the image generation tools are continuously improving, their outputs become undetectable by humans in the spatial domain and even simple CNN-based detectors. GAN-generated images can be identified via its distinguishing features through frequency analysis observed artifacts in general categories of synthetic images [51,20,27,28]. Synthetic human face images generated by various models can be detected by analyzing forgery traces in the frequency domain. Prior studies [20] have discovered that applying fast Fourier (FFT) to an image can extract the frequency spectrum. Figure 2 average frequency spectrum of 1000 images sample from Celeb-HQ dataset.

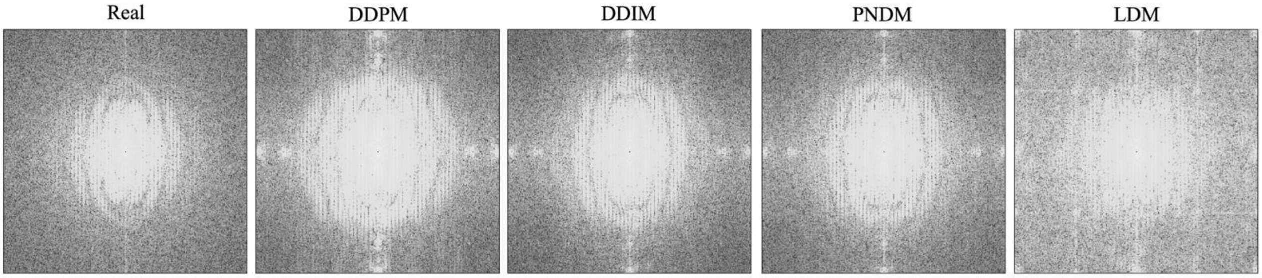

Fig. 2: Frequency Spectra Of Real Images Sourced From Celebq Dataset And Generated Images From(Progan,Stylegan2,Vqgan)

The frequency spectrum depicted by figure 3 manifests high contrast from prior studies [20, 51]. Nevertheless, high frequency noises are manifested by the dataset created by VQGAN. Also, StyleGAN2 exhibits quite fewer noises but still discernible from the real image spectra. On the other hand, the spectra of DM-created images are closely like the real images spectrum as shown in figure 2, excluding LDM which includes high frequency noise. Whilst images generated by DDPM and PNDM contain fewer visible artifacts in the frequency domain, they are more prone to high spectra density which deviated from real images spectra.

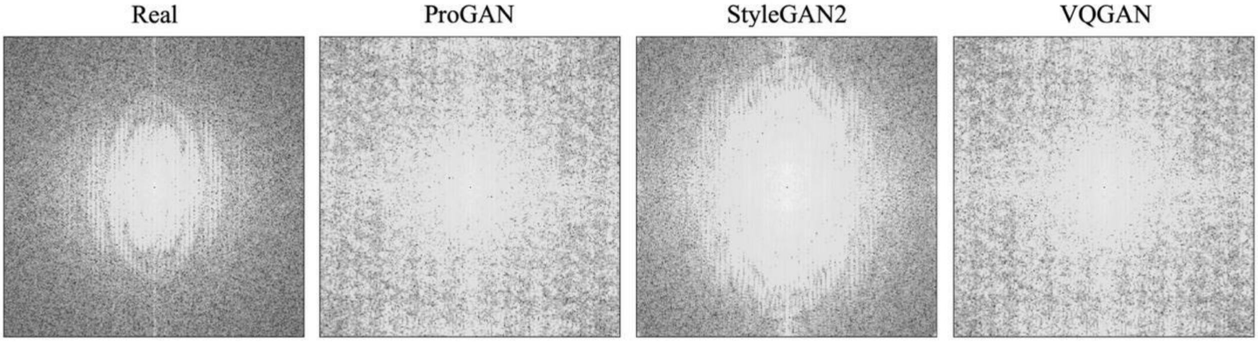

Fig. 3: (Progan,Stylegan2,Vqgan)Noises.

Significant discrepancies can be discerned between authentic images and artificially generated images in their frequency representations. The implications of these distinctions are further examined in this manuscript by employing a hybrid model, as depicted in figure 2, which can identify generated images produced by various Generative Adversarial Networks (GANs) and Diffusion Models (DMs). In particular, the detection output layer is designed as a categorical classification mechanism, with a fundamental pre-trained classification network, such as ResNet-50 [52] forming one of the branches and functioning as a feature extractor for both authentic and synthetic images. Conversely, the second branch applies Fast Fourier Transform (FFT) to the image, followed by a convolutional block that subsequently extracts features; these features, derived from both branches, are concatenated and subsequently input into a fully connected network. Furthermore, a categorical classification approach is employed as the final layer to assign labels to the images.

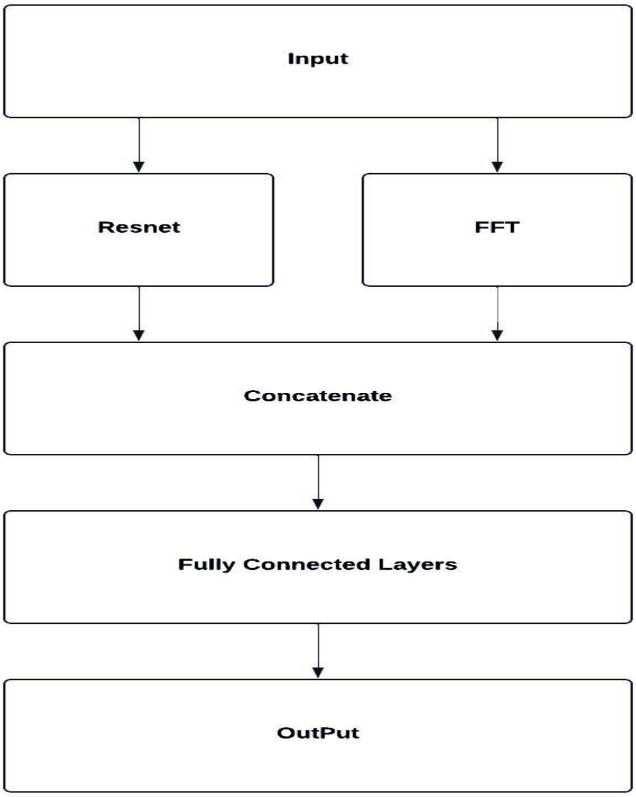

Fig. 4: Hybrid Detection Model Architecture

Figure 4, shows the diagram of the hybrid detection model architecture, combining ResNet-50 for spatial domain analysis and FFT for frequency domain analysis, with feature concatenation and fully connected layers for final classification. This model aims to enhance detection capabilities for GAN and DM-generated images.

4. Experimental Results

The experimental procedure is systematically delineated into two distinct phases. In the first phase, a variety of prominent detection methodologies are scrutinized, specifically Ojha2023 [25], Wang2020 [20], Mandelli2022 [24], Grag2021 [45] utilizing the weights disseminated by the respective authors. The detection algorithms were primarily trained on the LSUN dataset [53] which encompasses a wide array of general categories of synthetic imagery, whereas the latter was subjected to training on a synthetic face dataset [54] for comparative analysis. All detection algorithms were thoroughly evaluated using the entirety of the five test sets provided within this benchmark.

In the second phase, the robustness of each detector against image perturbations is systematically evaluated. This evaluation encompasses all detectors from the first phase, which were tested on facial images generated by ProGAN and DDIM, inclusive of various perturbations.

Following previous work about synthetic image detection and benchmarking Area Under Receiver Operating Characteristic Curve (AUC) scores are used to evaluate the detectors. Firstly, table 1 reports the performance of the detectors. Wang2020 performance is relatively low on images generated by both GAN and DM, however reasonable generalization among extensive categories of fake images has been documented by previous studies. High generalizability on GAN generated images was attained by Grag2021 but fails to achieve good performance on DM generated images. Ojha2023 Exhibits impressive transferability across images synthesized by both GANs and DMs, excluding Style- GAN2. Relative to, the Mandelli2022 manifests exceptional performance in our benchmark. The combination of ResNet-50 and frequency representation even outshines the cutting-edge performance on certain GAN models and DMs. In conclusion, detectors trained on extensive categories struggle to adapt synthetic face images.

Table 1:performance of the detectors, Whites belong to the authors

AUC (%)	GANs		DMs			Average
	StyleGAN2	VQGAN	DDPM	PNDM	LDM	
Wang2020	87.4	81.2	68.1	79.4	80.1	**79.24**
Grag2021	99.1	99.4	60.2	68.9	99.2	**85.36**
Mandelli2022	94.6	99.9	98.4	99.3	99.7	**98.38**
Ojha2023	73.7	96.3	93.5	96.7	98.2	**91.48**
ResNet-50+FFT	**94.6**	**99.8**	**98.7**	**99.4**	**99.6**	**98.42**

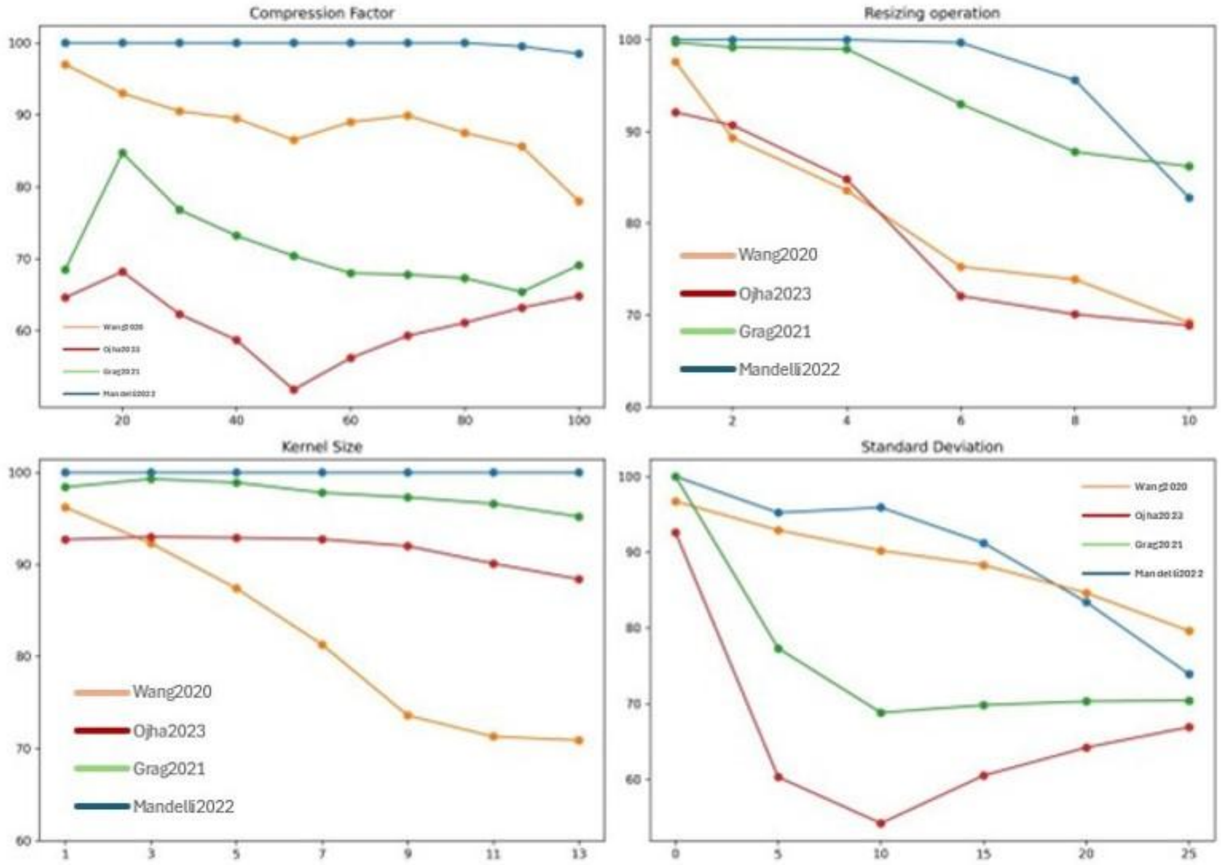

Fig. 5: Average Frequency Spectrum Of 1000 Images Sample From Celeb-Hq Dataset.

Secondly, robustness of various detection algorithms in response to prevalent image perturbations is depicted in Figure 5. Even though both Wang2020 and their performance is evidently compromised by noise and compression artifacts. Similarly, as the intensity of perturbations escalates, the efficacy of Ojha2023 markedly declines. Notably, Mandelli2022 emerges as the most resilient algorithm, particularly when exposed to Gaussian blur and JPEG compression. However, the introduction of low resolution or significant noise adversely affects its operational performance.

5. Conclusion, And Feature Work

While significant progress has been made in generative AI detection, the rapid evolution of sophisticated models, particularly Diffusion Models (DMs), presents ongoing challenges. Existing detection techniques often struggle to generalize effectively across diverse generative models and rely heavily on extensive datasets, making them sensitive to variations in the data. To address these limitations, our hybrid model leverages spatial and frequency-domain features, offering enhanced detection accuracy and robustness against common image perturbations. Our benchmarking results validate the effectiveness of this approach, with the ResNet-50+FFT model achieving the highest AUC (98.42%), surpassing the state-of-the-art Mandelli2022 model (98.38%) across a wide range of GAN and DM-generated images. This success underscores the practical potential of our method in detecting AI-synthesized human faces and safeguarding multimedia content. However, despite its strengths, the proposed method has certain limitations that warrant further investigation. The integration of Fast Fourier Transform (FFT) and deep learning introduces computational complexity, particularly when applied to large, high-resolution images. Additionally, the model's ability to generalize across diverse domains remains a challenge. For instance, a model trained on social media imagery may not perform as effectively on medical or satellite images without significant fine-tuning.

Future research should focus on designing scalable and adaptable frameworks to address these challenges. Potential directions include optimizing computational efficiency for high-resolution data, improving cross-domain generalization, and expanding the model's capabilities to incorporate tasks like image quality assessment. For example, alongside detecting authenticity (real or fake), the model could identify compression artifacts or low-quality distortions commonly associated with image manipulation. As generative AI continues to advance, the development of robust, scalable, and versatile detection techniques will be critical for protecting multimedia content and addressing the broader implications of synthetic media in various domains.

Conflicts of interest

We have no conflicts of interest to disclose.

Acknowledgements

We would like to thank Sulaimanyah Polytechnic University (www.spu.edu.iq) sualaimanyah, Iraq.

References

[1] Suwajanakorn, S., Seitz, S.M., and Kemelmacher-Shlizerman, I.Synthesizing obama: learning lip sync from audio. ACM Transactions on Graphics (ToG), 36(4):1–13, 2017.

[2] Zhou, X. and Zafarani, R. A survey of fake news: Fundamental theories, detection methods, and op- portunities, September 2020. ISSN 1557-7341. URL http://dx.doi.org/10.1145/3395046

[3] Chen, H. and Magramo, K. Finance worker pays out $25 million after video call with deepfake'chief financial officer', 2024.

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

[5] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros, "Un- paired image-to-image transla- tion using cycle-consistent adversarial networks," in *Proceedings of the IEEE international confer ence on computer vision*, 2017, pp. 2223–2232.

[6] Tero Karras, Samuli Laine, and Timo Aila, "A style-based generator architecture for generative ad versarial networks," in *Proceedings of* the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 4401–4410.

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

[8] 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.

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

[10] 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.

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

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

[13] David Holz. Midjoureny. https://docs. midjourney.com/docs/model-versions, 2022. [Online; ac-cessed 26-June-2023].

[14] David Holz. Dall-e 2. https://labs.openai.com, 2022. [Online; accessed 27-June-2023].

[15] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Sali- mans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural In- formationProcessing Systems*, 35:36479–36494, 2022.

[16] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceed- ings of the* IEEE/CVF *Conference on Computer Vision and Pattern Recognition*, pages 22500-22510, 2023.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-reso- lution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[18] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Bor- ja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023.

[19] Derui Zhu, Dingfan Chen, Jens Grossklags, and Mario Fritz. Data forensics in diffusion models: A system- atic analysis of membership privacy. *arXiv preprint arXiv*:2302.07801, 2023.

[20] 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.

[21] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi, "Do gans leave artifi- cial fingerprints?," in *2019 IEEE con- ference on multimedia information processing and retrieval (MIPR)*. IEEE, 2019, pp. 506–511.

[22] 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.

[23] SaraMandelli,Nicolo`Bonettini,PaoloBestagini,andStefanoTubaro, “Training CNNs in presence of JPEG compression: Multimedia foren- sics vs computer vision,” in *IEEE International Workshop onInforma- tion Forensics and Security (WIFS)*, 2020.

[24] SaraMandelli,Nicolo`Bonettini,PaoloBestagini,andStefanoTubaro, “Detecting gan-generated images by orthogonal training of multiple cnns,” in *2022 IEEE International Conference on Image Process ing (ICIP)*. IEEE, 2022, pp. 3091–3095.

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

[26] Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, and Yun- chao Wei, “Learning on gradi- ents: 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.

[27] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer, “Towards the detection of diffusion mo del deepfakes,” *arXiv preprint arXiv:2210.14571*, 2022.

[28] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Ver- doliva, “On the detection of synthetic images generated by diffusion models,” in *ICASSP 2023-2023IEEE*

[29] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality,Stability, and Variation. In International Conference on Learning Representa-tions. https://openreview.net/forum?id=Hk99zCeAb

[30] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embed-ding. *science*, 290(5500):2323–2326, 2000.

[31] 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 informationprocess- ing systems*, vol. 27, 2014.

[32] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, “Image-to-image translation with conditional adversarial networks,” in *Pro- ceedings of the IEEE conference on computer vision andpattern recog- nition*, 2017, pp. 1125–1134.

[33] JaschaSohl-Dickstein,EricWeiss,NiruMaheswaranathan, and Surya Ganguli. Deep unsu- pervised learning using nonequilibrium thermodynamics. In International Conference on Machine Learning, pages 2256–2265. PMLR, 2015.

[34] JonathanHo,AjayJain,andPieterAbbeel,“Denoising diffusion probabilistic models,” *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

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

[36] PrafullaDhariwalandAlexanderNichol.Diffusionmodels beat gans on image synthesis. Advances in Neural Informa- tion Processing Systems, 34:8780–8794, 2021.

[37] LupingLiu,YiRen,ZhijieLin,andZhouZhao.Pseudonu- merical methods for diffusion models on man- ifolds. arXiv preprint arXiv:2202.09778, 2022.

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-reso- lution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[39] Scott McCloskey and Michael Albright, “Detecting gan-generated im- agery using color cues,” *arXiv preprint arXiv:1812.08247*, 2018.

[40] 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.

[41] 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 com puter* vision and pattern recognition, 2020, pp. 5001–5010.

[42] 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.

[43] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva, “Detection of gan- generated fake images over social net- works,” in 2018 IEEE conference on multimedia informationprocessing and retrieval (MIPR). IEEE, 2018, pp. 384–389.

[44] 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.

[45] 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 IEEEinternational conference on multimedia and expo (ICME)*. IEEE, 2021, pp. 1–6.

[46] Mingxing Tan and Quoc Le, “Efficientnet: Rethinking model scaling for convolutional neural net- works,” in International conference on machine learning. PMLR, 2019, pp. 6105–6114.

[47] Chengdong Dong, Ajay Kumar, and Eryun Liu. Think twice before detecting gan-generated fake images from their spectral domain imprints. In Proceedings of the IEEE/CVF Conference on Com-puter Vision and Pattern Recognition, pages 7865–7874, 2022.

[48] Peter Lorenz, Ricard L Durall, and Janis Keuper, “Detecting images generated by deep diffusion models using their local intrinsic dimen- sionality,” in Proceedings of the IEEE/CVF InternationalConference on Computer Vision, 2023, pp. 448–459.

[49] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao, “Pseudo numer- ical methods for diffusion models on manifolds,” *arXiv preprint arXiv:2202.09778*, 2022.

[50] SaraMandelli,Nicolo`Bonettini,PaoloBestagini,andStefanoTubaro, “Detecting gan-generated images by orthogonal training of multiple cnns,” in *2022 IEEE International Conference on Image Process- ing (ICIP)*. IEEE, 2022, pp. 3091–3095.

[51] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz, “Leveraging frequency analysis for deep fake image recognition,” in International conferenceon machine learning. PMLR, 2020, pp. 3247–3258.

[52] 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.

[53] 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.

[54] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hell- sten, Jaakko Lehtinen, and Timo Aila, “Alias-free generative adver- sarial networks,” *Advances in Neural Information Process-ing Systems*, vol. 34, pp.8.