ECAI 2025
I. Lynce et al. (Eds.)

doi:10.3233/FAIA251048

Multi-Perspective Frequency Domain Learning for Generalizable AI-Generated Image Detection

Zili Xu[a], **Jianjie Luo**[a,*], **Fuqiang Yu**[b] **and Zhenguo Yang**[a]

[a]Guangdong University of Technology
[b]Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences)

Abstract. The prevalence of generative models in image and video generation has raised extensive concerns about potential harm and misuse. To identify the truthfulness of generated images, most of the existing methods typically apply Fast Fourier Transform (FFT) for frequency extraction. An existing problem is that the frequency-domain representations extracted by FFT are not comprehensive for AI-generated image detection. In this paper, we propose a Multi-perspective Frequency Domain Learning (MFDL) framework, which aims to learn both generalized and discriminative frequency representations via DWT and FFT. Specifically, we design a Frequency Representation Enhancement (FRE) module using the Discrete Wavelet Transform (DWT) and incorporating a multi-granularity enhancement strategy that amplifies all subbands across high frequency to improve discriminability. Additionally, we introduce a Frequency Representation Consistency (FRC) module, which employs complex convolution to capture and preserve forgery patterns in the real and imaginary components derived from FFT. By integrating complementary frequency representations from the DWT and FFT domains obtained through the FRE and FRC modules, MFDL achieves a comprehensive understanding of forgery traces in the frequency domain. This enhances the model's generalization capability for detecting generated content. Extensive experiments conducted on 32 distinct datasets, covering both GAN-generated and Diffusion-based images, demonstrate the effectiveness of our proposed MFDL framework. These experiments validate the effectiveness of multi-perspective frequency domain learning and show that MFDL outperforms existing detection methods, confirming its strong generalization ability across diverse generative models.

1 Introduction

In recent years, the rapid advancement of generative technologies [1, 45, 30, 31, 18] has increasingly blurred the boundary between reality and fiction. These technologies have reached a remarkable level, producing forgeries that closely mimic genuine pictures. Consequently, their malicious use has raised significant societal concerns, including fraud and the spread of misinformation. There is an urgent need for an effective detector to distinguish real images from generated ones.

Previous works [7, 41, 33] explore the classification of generated images, yet they face challenges in cross-domain generalization. To overcome this limitation, some recent studies [3, 14, 40] have investigated generalizable image detectors that exploit visual forgery traces, such as blurriness and unnatural artifacts. However, as generative techniques continue to evolve, the resulting images have become increasingly photorealistic, making it difficult to detect forgeries from newer models.

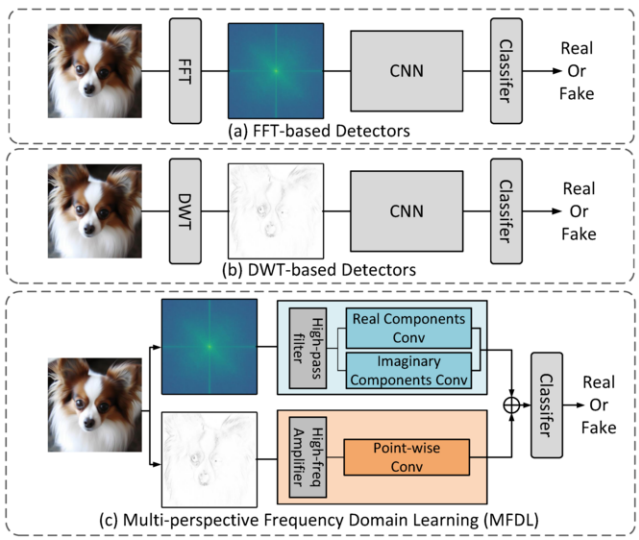

Figure 1: Previous frequency-based methods typically utilize either (a) the Fast Fourier Transform (FFT) for global frequency discrepancy traces or (b) the Discrete Wavelet Transform (DWT) to extract local edge frequency traces. (c) Our proposed MFDL integrates both FFT and DWT frequency representations, enabling more extensive frequency-based forgery recognition.

To address this challenge, researchers [6, 28, 24] have found frequency-based methods as a generalizable solution. Recent works [21, 25, 13] have explored frequency-domain forgery traces extracted using techniques such as the Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT), focusing on discrepancies in frequency components. However, relying solely on either FFT or DWT limits the detector's ability to capture comprehensive frequency traces, as shown in Figure 1.

In this paper, we propose a novel framework named Multi-perspective Frequency Domain Learning (MFDL) for the generaliz-

* Corresponding Author. Email: jianjieluo@gdut.edu.cn

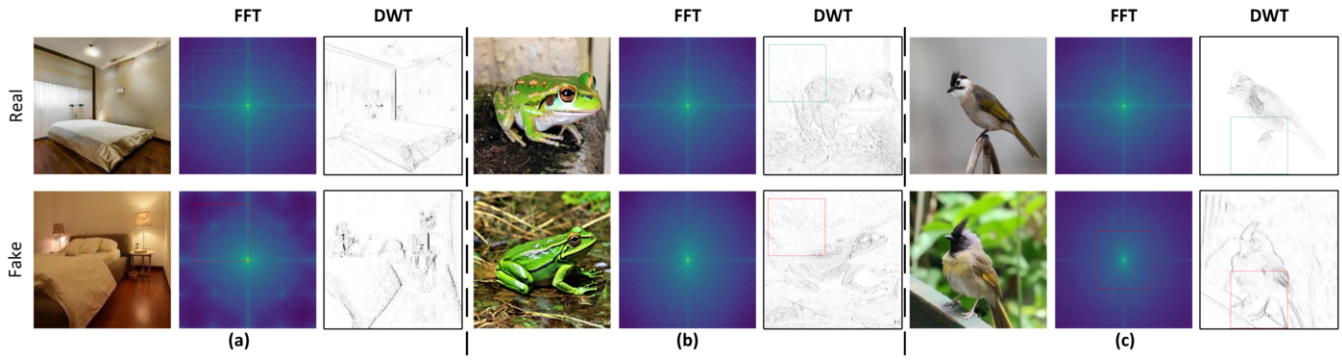

Figure 2: Frequency representations extracted by FFT and DWT. (a) The FFT-based frequency representation reveals strong global discrepancies between real and fake images. (b) The DWT-based frequency highlights unnatural and chaotic background textures. (c) The fake image exhibits subtle frequency representation in FFT and anomalous texture components in DWT.

able and robust detection of AI-generated images. To this end, MFDL adopts a dual-network architecture consisting of two complementary modules: the Frequency Representation Consistency (FRC) module and the Frequency Representation Enhancement (FRE) module.

Building upon the effectiveness of existing FFT-based methods that utilize high frequency components for AI-generated image detection [37, 11], the FRC module further explores the potential of the frequency domain to detect generative artifacts. The FRC module uses a high-pass filter to reduce low frequency noise, which is similar in both real and generated images, while preserving high frequency details where manipulation traces are typically introduced and detected. To enhance detection capability, the module applies complex convolutions to both the real and imaginary components of the FFT features. This allows the model to capture more detailed frequency discrepancies and learn robust, discriminative patterns in the FFT frequency space.

The FRE module complements the FRC by focusing on enhancing frequency-based features. It amplifies essential high frequency components and processes them with element-wise convolutions to refine subtle details. By introducing a multi-scale enhancement strategy, the module emphasizes informative patterns across different frequency levels. This design enables the FRE module to capture fine-grained frequency discrepancies, thereby improving the model's ability to distinguish between real and AI-generated images. The complementary enhancement in another frequency domain further boosts the robustness and generalization of the overall framework.

Finally, the representations extracted from both the FFT and DWT frequency domains are fused to provide a more comprehensive view of frequency-based artifacts. As shown in Figure 2, the FFT and DWT frequency representations are complementary. This fusion allows the model to leverage the strengths of both the FFT's global frequency analysis and the DWT's multi-scale local frequency details. Their combination enhances the model's ability to detect diverse generative artifacts and improves generalization across different types of AI-generated content, outperforming methods based on a single frequency domain or spatial features. In summary, the contributions of this paper are listed below:

- We propose a Multi-perspective Frequency Domain Learning framework named MFDL, which incorporates distinctive DWT and FFT frequency information to identify GAN and Diffusion generated images.
- We design two key modules: the Frequency Representation Consistency (FRC) module and the Frequency Representation Enhancement (FRE) module. The FRC module leverages FFT and complex convolutions to capture global frequency patterns, while the FRE module utilizes DWT to extract local frequency features and enhance high frequency forgery components.
- Extensive experiments validate the effectiveness of the proposed MFDL, demonstrating strong generalization capabilities across 32 different generation models.

2 Related works

2.1 *Pixel-based Generated Image Detection*

The rise of GAN [8, 15, 16] and Diffusion models [10, 26, 35] has spurred growing interest in detecting AI-generated images. Early studies [14, 9, 19] focused on pixel-level artifacts such as blurriness or unnatural textures, while others [4, 2] introduced data augmentation to improve robustness. However, as generative models like Stable Diffusion and DALL·E [32, 31] advance, these artifacts have become more subtle and harder to detect. Recent works [39, 5, 17] have explored vision-language models (VLMs) for improved detection, such as retraining CLIP [29] to distinguish real and fake content [27]. Another line of work, NPR [38], investigates the inherent operations of generative models to detect inconsistencies in the relationships between neighboring pixels in generated images. While these approaches aim to capture more intrinsic distinctions between real and synthetic content, they still face significant challenges in handling increasingly sophisticated generative models.

2.2 *Frequency-based Generated Image Detection*

As generative models advance, detecting images from newer models has become increasingly difficult. Prior works [7, 12, 20] explored frequency-domain discrepancies between real and fake images. For example, Wang et al. [41] used frequency representations for detection, while F3-Net [6] captured key frequency patterns of generated content. However, with the emergence of Diffusion models, such artifacts have become less visible. Recent studies [44, 22, 42] thus explore alternative frequency cues to boost detection. F2-Trans [25], for instance, applies DWT to extract high frequency components for better generalization across models. BIHPF [11] adopted the Fast Fourier Transform (FFT) to isolate high frequency signals and proposed a dual high-pass filtering network to suppress low frequency noise. Similarly, FreqNet [37] exploited high frequency features across both spatial and channel dimensions to enhance the de-

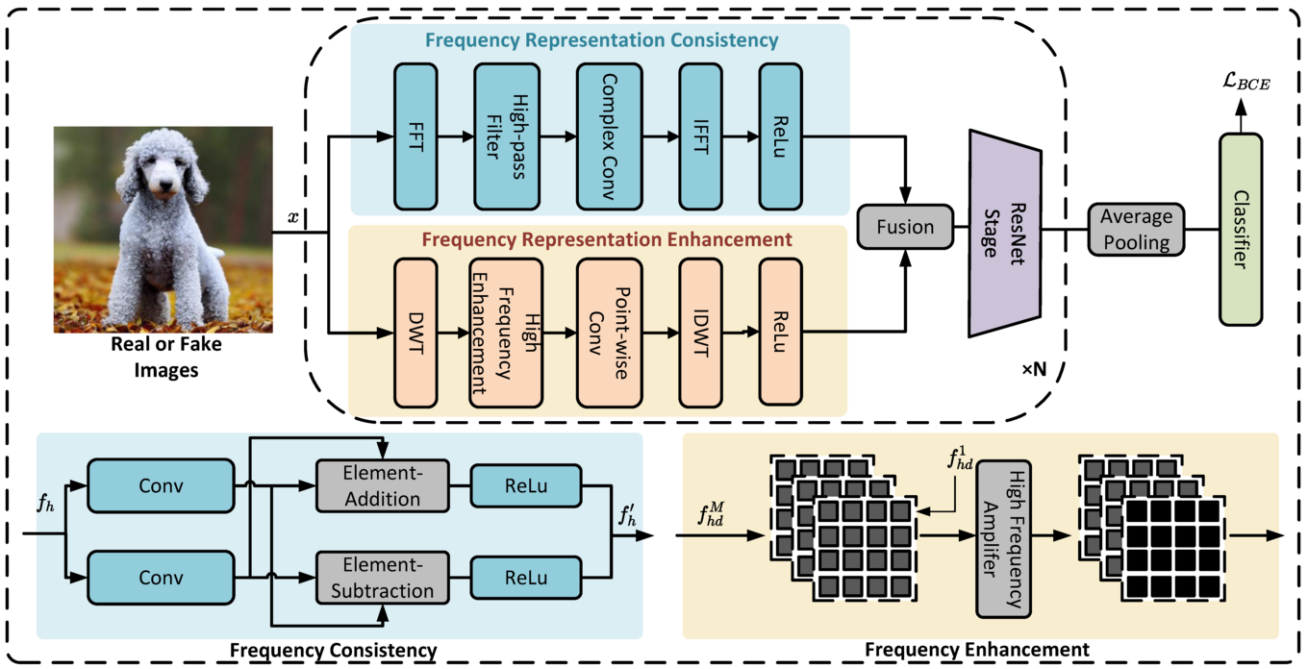

Figure 3: Overview of the proposed framework. In Frequency Representation Consistency (FRC), the input image x is first transformed into the frequency domain using FFT, followed by a high-pass filter to remove low frequency components. Meanwhile, in Frequency Representation Enhancement (FRE), the input image x is processed using DWT to extract multi-scale frequency features, where a high frequency enhancement module is applied. In the FRC branch, convolution is applied separately to the real and imaginary components of f_h, leveraging complex-valued computation to preserve detailed frequency representations. In the FRE branch, the highest-resolution frequency component f^1_{hd} is amplified across the HL, LH, and HH sub-bands to enhance local frequency features, such as background anomalies.

tector's capacity to capture frequency details. Nevertheless, these approaches typically focus on a single aspect of frequency representation, which may limit their ability to generalize across different generative traces or structures introduced by newer models. To address this limitation, we design a detector that comprehensively captures multi-perspective frequency representations, improving detection effectiveness and robustness in generative models.

3 Methodology

3.1 Overview

Intuitively, Figure 3 illustrates the Multi-perspective Frequency Domain Learning (MFDL) framework, which captures diverse aspects of frequency information through a dual-branch architecture that incorporates the Frequency Representation Enhancement (FRE) and Frequency Representation Consistency (FRC) modules. FRC employs the Fast Fourier Transform (FFT) to capture global frequency patterns and preserve frequency characteristics. Additionally, FRE leverages the Discrete Wavelet Transform (DWT) to extract local frequency features, identifying forgery traces to distinguish between real and generated images.

The task aims to identify generated images by learning a mapping $f : I \rightarrow Y$, where I denotes the set of input images and $Y = \{0, 1\}$ represents the binary labels (fake or real). Specifically, generated images may be produced by various techniques such as GANs and Diffusion models. Thus, we define $I_k \in I = \{I_1, I_2, \ldots, I_n\}$ as a collection of generated images produced by a particular generative model, where $I_k = \{(x_i, y_i)|1 \leq i \leq n\}$, with x_i representing an image sample and y_i denoting its associated class label. The task of generated image detection typically involves taking an image $x_i \in \mathbb{R}^{W \times H \times C}$ as input and predicting the image label.

3.2 Frequency Representation Consistency

To ensure generalization across different generative models and retain the frequency-domain information after convolutions, we propose the Frequency Representation Consistency (FRC) module. FRC focuses on both real and imaginary components by extracting global frequency representations using the Fast Fourier Transform (FFT).

Given an input image $x_i \in \mathbb{R}^{W \times H \times C}$, we first apply the Fast Fourier Transform (FFT) to convert it into the frequency domain. High frequency components are particularly important because they contain critical image details such as textures and edges, which are essential for distinguishing between real and generated images. To emphasize these components, we apply a high-pass filter that suppresses low frequency noise. This process retains high frequency information that is more sensitive to local artifacts and subtle manipulations found in generated images.

FRC uses the Fast Fourier Transform to extract the frequency representation, as shown in the equation:

$$f_h(u, v) = F_{\text{filter}}(u, v) \cdot \sum_{k=1}^{W} \sum_{q=1}^{H} x_i(k, q) \cdot e^{-j2\pi\left(\frac{uk}{W} + \frac{vq}{H}\right)}, \quad (1)$$

where j is the imaginary unit, and the Fourier kernel $e^{-j2\pi(\cdot)}$ focuses on transforming the spatial image $x_i(k, q)$ into its global frequency

representation f_h. The high-pass filter is defined as:

$$F_{\text{filter}}(u,v) = \begin{cases} 0, & \text{if } |u| < \frac{W}{4} \text{ and } |v| < \frac{H}{4} \\ 1, & \text{otherwise} \end{cases}, \tag{2}$$

where (u, v) are the frequency coordinates. This filter effectively suppresses the low frequency components within the central region, which correspond to the background and large image structures. To preserve the frequency forgery traces while learning features in the frequency domain, we utilize the complex convolution. Unlike traditional convolutions, complex convolutions operate on both the real and imaginary components of the frequency representation, ensuring that frequency information is retained throughout the learning process. The real and imaginary components of high frequency representation f_h are expressed as:

$$f_h = \Re(f_h) + j\Im(f_h), \tag{3}$$

where $\Re(f_h)$ and $\Im(f_h)$ are the real and imaginary parts of the frequency representation, respectively. The complex convolution kernel is defined as $K = \Re(K) + j\Im(K)$. The complex convolution operation is performed as follows:

$$\begin{aligned} \Re(f_h)' &= \Re(f_h) * \Re(K) - \Im(f_h) * \Im(K), \\ \Im(f_h)' &= \Re(f_h) * \Im(K) + \Im(f_h) * \Re(K), \end{aligned} \tag{4}$$

where $*$ denotes the convolution operation. The output frequency representation is then reconstructed from its learned real and imaginary frequency components:

$$f_h' = \Re(f_h)' + j\Im(f_h)', \tag{5}$$

where f_h' represents the enhanced frequency feature that captures forgery frequency information, crucial for recognizing subtle texture and structure details.

3.3 Frequency Representation Enhancement

To complement the FFT-based frequency extraction in FRC, we introduce the Frequency Representation Enhancement (FRE) module, which also focuses on high frequency forgery traces commonly found in generated images. This module leverages the Discrete Wavelet Transform (DWT) to capture multi-scale frequency features that span both low and high frequencies. While FFT captures global frequency information, DWT allows the detector to focus on the granular details of the image, which are essential for detecting subtle artifacts often introduced by generative models.

The frequency representation $f \in \mathbb{R}^{W\times H\times C\times M}$ extracted via DWT (db=3) consists of M levels of multi-scale features, each revealing distinct frequency-domain discrepancies that are indicative of forged images. These discrepancies often emerge at different scales, making them important for accurate image classification. The FRE module is designed to amplify high frequency components across these multi-scale levels to enhance the model's ability to differentiate between real and generated images.

The FRE module works as follows:

$$f_{hd}^M = \text{Conv}(H(x)), \quad f_{ld} = \text{Conv}(L(x)), \tag{6}$$

where $H(\cdot)$ and $L(\cdot)$ are high-pass and low-pass filters, respectively, and the convolution operation helps extract both high and low frequency representations from the input image. The f_{hd}^M and f_{ld} denote the high and low frequency. The process for reconstructing the enhanced frequency representation is as follows:

$$f_D = \mathcal{IDWT}(f_{ld}, \{\alpha \cdot f_{hd}^1\}, f_{hd}^2, \ldots, f_{hd}^M), \tag{7}$$

where we apply the Inverse Discrete Wavelet Transform (IDWT) to the granular frequency representations to transform them back to the spatial domain, with the result denoted as f_D. The hyperparameter α is set to 2.

To capture forgery traces within the frequency representation, we apply a feature extractor that consists of a point-wise convolution layer followed by a ReLU activation function:

$$f_D' = \text{ReLU}(\text{Conv}(f_D; w, b)), \tag{8}$$

where w and b represent the convolution weights and bias parameters, and f_D' is the output feature with extracted forgery traces.

Then, we integrate the features from FRE and FRC, and learn the forgery frequency representation through a ResNet Stage Block,

$$f_{re} = \text{ResN}(f_D' + f_h'), \tag{9}$$

where f_{re} denotes the learned multi-perspective frequency feature map, and ResN represents the ResNet stage block.

Finally, MFDL leverages the Binary Cross-Entropy (BCE) loss, denoted as $\mathcal{L}_{BCE}$ to optimize the classification objective. The loss is defined as:

$$\mathcal{L}_{\text{BCE}} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \cdot \log(p_i) + (1-y_i)\cdot \log(1-p_i)\right], \tag{10}$$

where $y_i \in \{0, 1\}$ is the ground-truth label, and $p_i \in (0, 1)$ is the predicted probability.

4 Experiment

4.1 Experimental Settings

Training Dataset. To ensure a consistent basis for comparison, we follow the baseline [27, 37] and train our MFDL model exclusively on ProGAN-generated images from the ForenSynths [41] dataset. Specifically, the training set contains over 70,000 ProGAN images in four object categories: car, cat, chair, and horse.

Test Dataset. To demonstrate the effectiveness and generalization capability of our proposed model, we conduct a comprehensive evaluation on 32 generated image datasets. These datasets cover a wide range of generation sources, including open-source models such as LDM200, LDM100, PNDM, DDPM, VQDiffusion, and Guided Diffusion, as well as commercial models like Stable Diffusion v4, Glide, and DALL·E. Furthermore, we evaluate our model on eight Diffusion-generated datasets from DiffusionForensics [43], including ADM, DDPM, Stable Diffusion v1, Stable Diffusion v2, and others. In addition, we test on the large-scale GenImage [46] dataset, which consists of millions of images from advanced generators, including BigGAN, Wukong, Midjourney, and Stable Diffusion v5.

Implementation Details. During training, we use the Adam optimizer with a learning rate of 1×10^{-3}. The batch size is set to 32, and the model is trained for 65 epochs. Following the baselines [37, 38], we evaluate the models by calculating the average precision (AP) and accuracy (Acc).

4.2 Performance Evaluation

4.2.1 Performance on GAN-generated Datasets

To evaluate the effectiveness of MFDL on images generated by the GAN model, we conduct experiments on eight GAN-generated image datasets. As shown in Table 1, MFDL consistently outperforms existing baseline models across these datasets.

Table 1: Performance of the approaches on the GAN datasets. The best results are highlighted in bold and second-best results are underlined. The Mean denotes the average of Acc and AP.

Method	Test Models																	
	CycleGAN		StyleGAN		StyleGAN2		BigGAN		ProGAN		StarGAN		AttGAN		RelGAN		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
Wang [41]	80.7	96.8	73.4	98.5	68.4	97.9	45.8	95.6	91.4	99.4	80.9	95.4	57.4	90.0	63.1	94.8	70.1	96.0
Frank [7]	75.5	71.2	74.5	72.0	73.1	71.4	88.7	86.0	90.3	85.2	**99.5**	99.5	65.0	74.4	69.2	96.2	79.4	81.9
Durall [6]	69.0	64.0	54.4	52.6	66.8	62.0	60.1	56.3	81.1	74.4	98.1	98.1	39.9	38.2	80.0	88.2	68.6	66.7
SelfBland [34]	59.2	65.3	50.0	47.7	48.6	47.4	51.1	51.9	58.8	65.2	74.5	89.2	63.1	66.1	73.6	77.8	59.8	63.8
GANDetection [23]	85.2	95.5	74.4	92.9	69.9	87.9	76.3	89.9	82.7	95.1	78.8	90.2	57.4	75.1	60.9	86.2	73.2	89.1
Ojha [27]	**98.4**	**99.8**	86.1	97.5	75.2	97.5	92.7	97.8	97.8	95.5	95.5	99.4	90.5	96.7	92.8	97.5	91.1	97.7
FreqNet [37]	91.8	96.6	90.2	99.7	89.1	99.5	92.8	99.2	99.6	**100.0**	95.2	99.0	89.8	98.8	99.0	**100.0**	93.4	99.1
NPR [38]	89.6	99.4	94.7	99.7	**98.6**	99.7	81.4	99.1	99.7	**100.0**	95.6	**100.0**	73.9	84.2	99.7	**100.0**	91.6	97.8
MFDL	92.3	97.7	**94.9**	**99.9**	95.5	**99.9**	**97.4**	**99.8**	**99.8**	**100.0**	96.0	**100.0**	**97.4**	**99.5**	**99.9**	**100.0**	**96.5**	**99.6**

Table 2: Performance of the approaches on the Diffusion datasets. The best results are highlighted in bold and second-best results are underlined. The Mean denotes the average of Acc and AP.

Method	Test Models																	
	DALL·E		LDM200		LDM100		PNDM		DDPM		Glide		SDv4		VQDiffusion		Guided		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
Wang [41]	52.5	66.7	51.1	66.5	51.3	66.6	50.3	82.8	69.0	54.6	53.2	75.5	50.7	64.8	50.0	70.6	52.5	81.2	53.4	69.9
Frank [7]	57.0	62.5	56.4	50.9	56.6	51.3	44.0	38.2	37.0	27.6	52.0	42.4	50.5	45.5	51.7	66.7	50.6	48.5	50.6	48.1
Durall [6]	55.9	58.0	61.7	61.7	62.0	62.6	44.5	47.3	52.9	49.8	51.8	49.7	54.8	54.8	38.6	38.3	52.4	52.4	52.7	52.7
SelfBland [34]	52.4	51.6	52.6	51.9	53.0	54.0	48.2	48.2	61.9	49.6	60.8	65.2	54.8	53.4	77.2	82.7	58.0	57.6	57.6	57.1
GANDetection [23]	67.2	83.0	54.9	65.9	54.7	65.8	50.6	79.0	62.3	46.4	51.4	52.5	49.2	32.5	51.1	51.2	52.1	51.5	54.8	58.6
Ojha [27]	87.3	97.4	95.0	99.2	94.4	99.3	86.1	95.2	**74.0**	**85.1**	58.7	91.3	64.8	74.8	77.7	90.1	69.9	85.2	78.3	88.6
FreqNet [37]	97.3	**99.8**	97.5	**99.9**	97.9	99.9	85.2	99.8	69.2	32.7	87.6	95.5	70.1	92.1	**100.0**	**100.0**	71.8	88.6	86.2	89.7
NPR [38]	82.2	99.2	95.8	**99.9**	95.7	99.9	96.5	**99.9**	69.1	23.7	93.8	**98.4**	94.2	98.2	**100.0**	**100.0**	**87.4**	96.8	90.6	90.8
MFDL	**97.5**	**99.8**	**98.8**	**99.9**	**99.0**	**100.0**	**97.8**	**99.9**	69.6	35.3	**94.2**	98.3	**94.9**	**98.8**	**100.0**	**100.0**	81.9	**97.1**	**92.6**	**92.1**

Several key observations can be made: 1) The proposed MFDL achieves superior performance in detecting GAN-generated images, demonstrating the advantage of emphasizing high frequency components. 2) MFDL achieves notable AP values, indicating high confidence in classification, which benefits from the joint use of FFT and DWT for frequency domain representation. 3) Specifically, MFDL outperforms baseline methods, with an improvement of in 3.1% mean accuracy. These results demonstrate that MFDL effectively preserves frequency domain consistency, leading to enhanced robustness against various GAN architectures. Compared to existing methods, MFDL highlights the high frequency discrepancies introduced by GANs, which are crucial for reliable image generation detection.

4.2.2 *Performance on Diffusion-generated Datasets*

To further validate the generalization capability of MFDL, we conduct extensive evaluations on images generated by various Diffusion models. As shown in Table 2 and Table 3, we can observe that our method achieves comprehensive improvements in both Acc and AP compared to existing state-of-the-art detection methods.

Despite being trained exclusively on images generated by ProGAN, MFDL exhibits remarkable generalization to unseen diffusion models. As shown in Table 2, it significantly outperforms state-of-the-art approaches such as FreqNet and NPR, achieving gains of 6.4% and 2.0% in mean Acc, respectively. Furthermore, as shown in Table 3 on the challenging DIRE dataset, which features high-quality diffusion-generated images, MFDL continues to demonstrate superior performance. It surpasses both FreqNet and NPR baselines by 12.1% and 3.3% in mean accuracy, while also achieving higher and comparable average precision scores.

Additionally, in Table 4, we evaluate MFDL on a large-scale benchmark encompassing images produced by advanced diffusion models, including Midjourney and Stable Diffusion v5. MFDL achieves a mean accuracy of 88.1%, outperforming NPR and FreqNet by margins of 4.1% and 8.2%, respectively. These results highlight several key findings: 1) MFDL consistently outperforms existing baselines across diverse datasets, showcasing its ability to capture multi-perspective frequency information through the FRE and FRC modules. 2) Even when trained only on a single GAN generator, MFDL generalizes well to unseen diffusion models, demonstrating its robustness and effectiveness.

In conclusion, MFDL leverages frequency-domain representations to effectively distinguish between real and generated images. Its stable performance on challenging Diffusion-generated datasets confirms that the MFDL is an effective and generalizable AI-generated image detection framework.

4.3 *Comparison of Parameters with SOTA Models*

Table 5 reports the average performance across 32 datasets alongside the parameter counts of state-of-the-art (SOTA) methods. Compared with Ojha et al. [27], MFDL achieves better accuracy while using significantly fewer parameters, demonstrating its efficiency and the effectiveness of incorporating frequency information for detecting generated images. Moreover, MFDL surpasses FreqNet [37] and NPR [38], achieving an average accuracy gain of 7.4% and 2.5%. These results indicate that MFDL is more capable of capturing multi-perspective frequency forgery artifacts, offering superior performance with a comparable model size.

4.4 *Ablation Study*

As shown in Table 6, removing the FRC module results in a clear drop in performance, demonstrating that global frequency information plays a key role in distinguishing generated content. FRC utilizes FFT to capture broad spectral patterns, making it especially useful

Table 3: Performance of the approaches on the DiffusionForensics. The best results are highlighted in bold and second-best results are underlined. The Mean denotes the average of Acc and AP.

Method	Test Models																	
	ADM		DDPM		IDDPM		LDM		PNDM		VQDiffusion		SDv1		SDv2		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
Wang [41]	53.9	71.8	62.7	76.6	50.2	82.7	50.4	78.7	50.8	90.3	50.0	71.0	38.0	76.7	52.0	90.3	51.0	79.8
Frank [7]	58.9	65.9	37.0	27.6	51.4	65.0	51.7	48.5	44.0	38.2	51.7	66.7	32.8	52.3	40.8	37.5	46.0	50.2
Durall [6]	39.8	42.1	52.9	49.8	55.3	56.7	43.1	39.9	44.5	47.3	38.6	38.3	39.5	56.3	62.1	55.8	47.0	48.3
SelfBland [34]	57.0	59.0	61.9	49.6	63.2	66.9	83.3	92.2	48.2	48.2	77.2	82.7	46.2	68.0	71.2	73.9	63.5	67.6
GANDetection [23]	51.1	53.1	62.3	46.4	50.2	63.0	51.6	48.1	50.6	79.0	51.1	51.2	39.8	65.6	50.1	36.9	50.8	55.4
Ojha [27]	78.4	92.1	72.9	78.8	75.0	92.8	82.2	97.1	75.3	92.5	83.5	97.7	56.4	90.4	71.5	92.4	74.4	91.7
FreqNet [37]	66.7	85.2	90.3	99.1	60.1	92.9	97.5	**100.0**	84.9	99.3	99.9	**100.0**	93.8	99.6	70.7	96.5	83.0	96.6
NPR [38]	83.7	98.0	**97.4**	**100.0**	85.8	99.1	**100.0**	**100.0**	96.0	**100.0**	**100.0**	**100.0**	91.1	**99.9**	80.5	**99.8**	91.8	99.5
MFDL	**86.3**	**98.4**	95.4	**100.0**	**88.5**	**99.2**	**100.0**	**100.0**	**98.9**	**100.0**	**100.0**	**100.0**	**94.3**	**99.9**	**98.0**	**99.8**	**95.1**	**99.6**

Table 4: Performance of the approaches on the GenImage. The best results are highlighted in bold and second-best results are underlined. The Mean denotes the average of Acc and AP.

Method	Test Models															
	BigGAN		Wukong		VQDM		Glide		Midjourney		ADM		SDv5		Mean	
	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP	Acc	AP
Wang [41]	71.1	85.9	60.3	75.6	51.0	57.3	56.4	68.8	58.0	72.2	51.3	66.2	50.5	61.5	56.9	69.6
Frank [7]	81.9	93.6	63.4	61.7	40.3	39.5	77.8	85.1	54.1	52.9	45.8	46.0	39.2	37.7	57.5	59.5
Fusing [14]	77.4	90.7	59.0	94.1	51.7	64.6	55.1	75.6	57.2	77.5	52.2	70.0	51.4	65.7	57.7	76.8
LGrad [36]	85.6	92.9	67.1	72.9	59.5	62.4	72.9	77.4	66.1	80.4	65.3	71.8	63.6	62.8	68.5	74.3
DIRE [43]	70.1	75.2	54.4	55.3	53.6	54.5	71.7	78.3	58.0	61.8	75.7	85.8	49.8	49.5	61.9	65.7
Ojha [27]	**95.0**	**99.2**	70.9	91.0	85.3	96.5	62.4	83.8	56.1	74.0	66.8	86.8	63.4	85.8	71.4	88.1
FreqNet [37]	89.8	94.6	85.5	90.5	78.0	83.1	78.3	84.0	63.1	68.2	78.0	84.7	86.5	91.1	79.9	85.1
NPR [38]	87.9	95.5	85.4	94.5	79.4	94.6	92.1	97.5	73.7	87.8	79.0	94.9	90.7	96.6	84.0	94.4
MFDL	90.8	96.9	**88.9**	**95.1**	**86.1**	**96.9**	**92.9**	**97.7**	**83.8**	**90.3**	**81.9**	**95.2**	**92.4**	**96.9**	**88.1**	**95.5**

Table 5: Comparison of Model Parameters and Accuracy.

Method	Parameters (↓)	Time (↓)	Accuracy (↑) on 32 models
Ojha [27]	427M	201.5ms	79.0%
FreqNet [37]	1.9M	50.7ms	85.8%
NPR [38]	**1.4M**	**32.5ms**	89.7%
MFDL	2.0M	59.8ms	**93.2%**

Table 6: Ablation Study of our proposed modules.

FRE	FRC	Accuracy (↑) on 32 models
×	×	63.4%
✓	×	87.6%
×	✓	89.5%
✓	✓	**93.2%**

when dealing with varying generative styles. On the other hand, FRE focuses on high frequency representation at different scales, which are typically associated with subtle texture inconsistencies or local artifacts. These fine-grained cues complement the global features extracted by FRC. These two modules form a complementary structure that captures both global and local frequency irregularities, contributing to robust detection across diverse generative sources.

Complex convolutions process the real and imaginary components independently, making them a natural fit for the FRC module, where FFT produces complex-valued features. In contrast, as DWT in FRE outputs only real-valued subbands, applying complex convolutions there distorts the representations and harms performance. Table 7 shows that using complex convolution solely in FRC yields the best results, while applying it in FRE alone or in both modules degrades performance. In the former case, both modules fail to capture useful cues, and in the latter, FRE interferes with FRC. Removing complex convolution entirely also reduces accuracy. These findings indicate that complex convolution should be selectively applied in FRC, and that both FRC and FRE are essential for learning complementary frequency representations.

Table 7: Ablation Study on Complex Convolution. The w/o and w/ in denote the without and with.

Complex Conv	Accuracy (↑) on 32 models
w/o	90.4%
w/ in FRE&FRC	88.7%
w/ in FRE	86.6%
w/ in FRC	**93.2%**

4.5 *Visualization on Forgery Traces*

The Grad-CAM visualizations highlight the regions where the model focuses when detecting generated images, effectively revealing potential forgery traces. As shown in Figure 4, the attention maps indicate that MFDL is capable of locating manipulated regions in images produced by both GAN and Diffusion models. Compared to GAN-generated images, those generated by Diffusion models exhibit subtler artifacts, making them more challenging to detect. Nevertheless, MFDL successfully identifies forgery-related regions across a wide range of generative models, demonstrating its strong generalization ability. Notably, despite being trained only on images generated by ProGAN, MFDL can still capture discriminative frequency-domain cues from unseen generative models. This highlights the importance of multi-perspective frequency representation in enhancing the model's effectiveness in generated image detection.

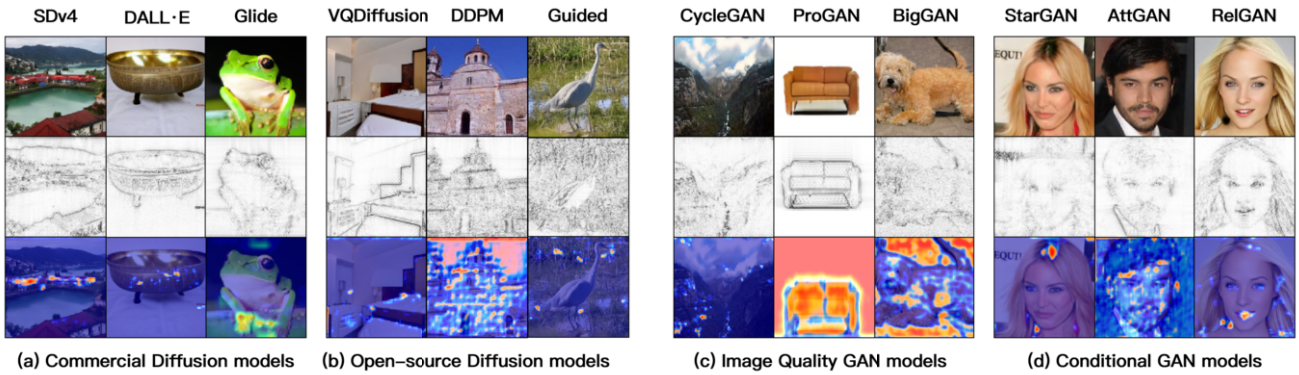

Figure 4: Visualization of the Gradient-weighted Class Activation Mapping (Grad-CAM) extracted from MFDL for both Diffusion-generated and GAN-generated images.

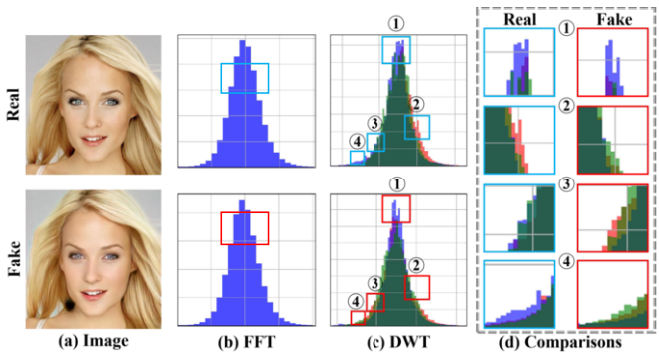

Figure 5: The discrepancies in frequency energy distributions. (a) shows sample images from the dataset. (b) presents the frequency distributions extracted using FFT. (c) illustrates the frequency distributions derived from DWT. (d) highlights the differences in DWT-based frequency distributions between real and fake images.

4.6 *Discrepancies in Frequency Domain*

To quantify the frequency-domain differences between real and generated images, we analyze the energy distributions derived from the Fast Fourier Transform (FFT) and Discrete Wavelet Transform (DWT). As visualized in Figure 5, generated images exhibit subtle energy discrepancies in the FFT spectrum, whereas the DWT analysis reveals more pronounced differences in high frequency subbands (e.g., LH, HL, and HH). These observations confirm that fake images possess abnormal frequency characteristics, providing a strong motivation for our framework to incorporate both global FFT and local DWT representations.

4.7 *Visualization on Image Distributions*

The t-SNE visualization in Figure 6 shows that MFDL effectively learns distinct frequency patterns to separate real and generated images. The clear clustering of real and fake samples demonstrates high separability for features extracted from both the FFT and DWT domains. Furthermore, the model exhibits strong generalization across diverse generative models, including both GANs and Diffusion, confirming that its multi-perspective design captures robust forgery traces. The complementary frequency representations from FRC and FRE jointly enhance the model's ability to detect a wide range of images from unseen generative models.

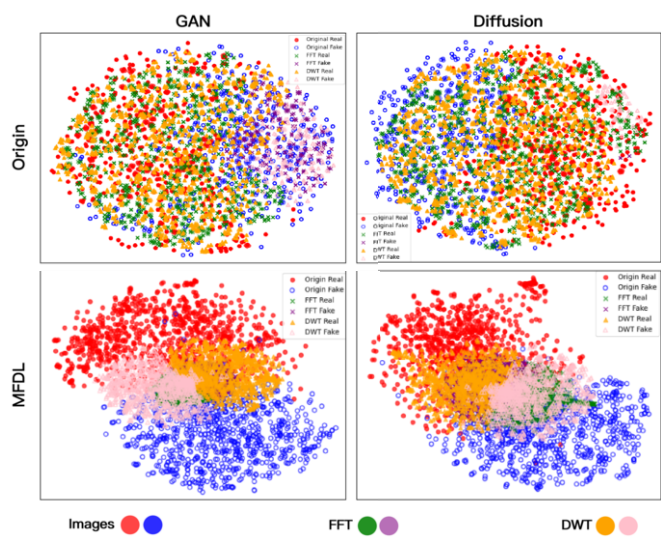

Figure 6: The image distributions via t-SNE visualization. The red, green, and yellow colors denote real image distributions, while the blue, purple, and pink colors denote generated image distributions.

5 Conclusion

We propose Multi-perspective Frequency Domain Learning (MFDL) for detecting images generated by both GANs and Diffusion models. MFDL consists of two branches: Frequency Representation Enhancement (FRE), which uses DWT with multi-scale enhancement to capture local frequency artifacts, and Frequency Representation Consistency (FRC), which applies FFT and complex convolution to model global frequency patterns. By integrating these two complementary techniques, MFDL learns a multi-perspective frequency representation, allowing it to better understand both the local and global characteristics of generative models. Experiments on 32 datasets demonstrate that MFDL achieves superior generalization and outperforms existing methods across diverse generative sources.

Acknowledgements

This work is supported by the Guangdong Basic and Applied Basic Research Foundation (No.2024A1515010237), and the Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Qilu University of Technology (Shandong Academy of Sciences) (No.2023ZD035).

References

[1] A. Brock, J. Donahue, and K. Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[2] J. Cao, C. Ma, T. Yao, S. Chen, S. Ding, and X. Yang. End-to-end reconstruction-classification learning for face forgery detection. In *Proceedings of CVPR*, pages 4113–4122, 2022.

[3] L. Chai, D. Bau, S.-N. Lim, and P. Isola. What makes fake images detectable? understanding properties that generalize. In *Proceedings of ECCV*, pages 103–120, 2020.

[4] L. Chen, Y. Zhang, Y. Song, L. Liu, and J. Wang. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of CVPR*, pages 18710–18719, 2022.

[5] D. Cozzolino, G. Poggi, R. Corvi, M. Nießner, and L. Verdoliva. Raising the bar of ai-generated image detection with clip. In *Proceedings of CVPR*, pages 4356–4366, 2024.

[6] R. Durall, M. Keuper, and J. Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of CVPR*, pages 7890–7899, 2020.

[7] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz. Leveraging frequency analysis for deep fake image recognition. In *Proceedings of ICML*, pages 3247–3258, 2020.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proceedings of NeurIPS*, 2014.

[9] Z. Gu, Y. Chen, T. Yao, S. Ding, J. Li, and L. Ma. Delving into the local: Dynamic inconsistency learning for deepfake video detection. In *Proceedings of AAAI*, volume 36, pages 744–752, 2022.

[10] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Proceedings of NeurIPS*, 2020.

[11] Y. Jeong, D. Kim, S. Min, S. Joe, Y. Gwon, and J. Choi. Bihpf: Bilateral high-pass filters for robust deepfake detection. In *Proceedings of WACV*, pages 48–57, 2022.

[12] Y. Jeong, D. Kim, Y. Ro, and J. Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of AAAI*, volume 36, pages 1060–1068, 2022.

[13] G. Jia, M. Zheng, C. Hu, X. Ma, Y. Xu, L. Liu, Y. Deng, and R. He. Inconsistency-aware wavelet dual-branch network for face forgery detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 3(3):308–319, 2021.

[14] Y. Ju, S. Jia, L. Ke, H. Xue, K. Nagano, and S. Lyu. Fusing global and local features for generalized ai-synthesized image detection. In *Proceedings of ICIP*, pages 3465–3469, 2022.

[15] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[16] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of CVPR*, pages 4401–4410, 2019.

[17] M. Keita, W. Hamidouche, H. Bougueffa, A. Hadid, and A. Taleb-Ahmed. Harnessing the power of large vision language models for synthetic image detection. *arXiv preprint arXiv:2404.02726*, 2024.

[18] D. Kingma, T. Salimans, B. Poole, and J. Ho. Variational diffusion models. In *Proceedings of NeurIPS*, 2021.

[19] C. Li, Z. Huang, D. P. Paudel, Y. Wang, M. Shahbazi, X. Hong, and L. Van Gool. A continual deepfake detection benchmark: Dataset, methods, and essentials. In *Proceedings of WACV*, pages 1339–1349, 2023.

[20] J. Li, H. Xie, J. Li, Z. Wang, and Y. Zhang. Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In *Proceedings of CVPR*, pages 6458–6467, 2021.

[21] H. Liu, X. Li, W. Zhou, Y. Chen, Y. He, H. Xue, W. Zhang, and N. Yu. Spatial-phase shallow learning: rethinking face forgery detection in frequency domain. In *Proceedings of CVPR*, pages 772–781, 2021.

[22] H. Liu, Z. Tan, Q. Chen, Y. Wei, Y. Zhao, and J. Wang. Unified frequency-assisted transformer framework for detecting and grounding multi-modal manipulation. *International Journal of Computer Vision*, pages 1–18, 2024.

[23] S. Mandelli, N. Bonettini, P. Bestagini, and S. Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *Proceedings of ICIP*, pages 3091–3095, 2022.

[24] F. Marra, D. Gragnaniello, L. Verdoliva, and G. Poggi. Do gans leave artificial fingerprints? In *Proceedings of MIPR*, pages 506–511, 2019.

[25] C. Miao, Z. Tan, Q. Chu, H. Liu, H. Hu, and N. Yu. F 2 trans: High-frequency fine-grained transformer for face forgery detection. *IEEE Transactions on Information Forensics and Security*, 18:1039–1051, 2023.

[26] A. Q. Nichol and P. Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of ICML*, pages 8162–8171, 2021.

[27] U. Ojha, Y. Li, and Y. J. Lee. Towards universal fake image detectors that generalize across generative models. In *Proceedings of CVPR*, pages 24480–24489, 2023.

[28] Y. Qian, G. Yin, L. Sheng, Z. Chen, and J. Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *Proceedings of ECCV*, pages 86–103, 2020.

[29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of ICML*, pages 8748–8763, 2021.

[30] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever. Zero-shot text-to-image generation. In *Proceedings of ICML*, pages 8821–8831, 2021.

[31] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of CVPR*, pages 10684–10695, 2022.

[33] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of ICCV*, pages 1–11, 2019.

[34] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of CVPR*, pages 18720–18729, 2022.

[35] J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[36] C. Tan, Y. Zhao, S. Wei, G. Gu, and Y. Wei. Learning on gradients: Generalized artifacts representation for gan-generated images detection. In *Proceedings of CVPR*, pages 12105–12114, 2023.

[37] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. Frequency-aware deepfake detection: Improving generalizability through frequency space domain learning. In *Proceedings of AAAI*, volume 38, pages 5052–5060, 2024.

[38] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of CVPR*, pages 28130–28139, 2024.

[39] A. Uchendu, T. Le, and D. Lee. Topformer: Topology-aware authorship attribution of deepfake texts with diverse writing styles. In *Proceedings of ECAI*, pages 1446–1454. 2024.

[40] C. Wang and W. Deng. Representative forgery mining for fake face detection. In *Proceedings of CVPR*, pages 14923–14932, 2021.

[41] S.-Y. Wang, O. Wang, R. Zhang, A. Owens, and A. A. Efros. Cnn-generated images are surprisingly easy to spot... for now. In *Proceedings of CVPR*, pages 8695–8704, 2020.

[42] Y. Wang, K. Yu, C. Chen, X. Hu, and S. Peng. Dynamic graph learning with content-guided spatial-frequency relation reasoning for deepfake detection. In *Proceedings of CVPR*, pages 7278–7287, 2023.

[43] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. Dire for diffusion-generated image detection. In *Proceedings of ICCV*, pages 22445–22455, 2023.

[44] S. Woo et al. Add: Frequency attention and multi-view based knowledge distillation to detect low-quality compressed deepfake images. In *Proceedings of AAAI*, volume 36, pages 122–130, 2022.

[45] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of ICCV*, pages 2223–2232, 2017.

[46] M. Zhu, H. Chen, Q. Yan, X. Huang, G. Lin, W. Li, Z. Tu, H. Hu, J. Hu, and Y. Wang. Genimage: A million-scale benchmark for detecting ai-generated image. In *Proceedings of NeurIPS*, 2023.