Enhancing Frequency Forgery Clues for Diffusion-Generated Image Detection

Daichi Zhang, Tong Zhang, Shiming Ge, *Senior Member, IEEE*, and Sabine Süsstrunk, *Fellow, IEEE*

***Abstract*—Diffusion models have achieved remarkable success in image synthesis, but the generated high-quality images raise concerns about potential malicious use. Existing detectors often struggle to capture discriminative clues across different models and settings, limiting their generalization to unseen diffusion models and robustness to various perturbations. To address this issue, we observe that diffusion-generated images exhibit progressively larger differences from natural real images across low- to high-frequency bands. Based on this insight, we propose a simple yet effective representation by enhancing the Frequency Forgery Clue ($\mathbf{F}^2\mathbf{C}$) across all frequency bands. Specifically, we introduce a frequency-selective function which serves as a weighted filter to the Fourier spectrum, suppressing less discriminative bands while enhancing more informative ones. This approach, grounded in a comprehensive analysis of frequency-based differences between natural real and diffusion-generated images, enables general detection of images from unseen diffusion models and provides robust resilience to various perturbations. Extensive experiments on various diffusion-generated image datasets demonstrate that our method outperforms state-of-the-art detectors with superior generalization and robustness.**

***Index Terms*—Diffusion-generated image detection, diffusion models, frequency clue.**

I. Introduction

Diffusion models [1]–[12] have achieved remarkable success in image synthesis, producing high-quality and diverse results. However, the accessibility and realism of these generated images pose significant challenges, as they can be easily misused for malicious purposes, such as fabricating evidence or misleading the public, raising serious social, privacy, and ethical concerns [13]. Therefore, how to effectively detect diffusion-generated images has become an urgent and critical issue recently.

There are various detectors that have been proposed to address this issue [14]–[21], such as using reconstruction error [22], leveraging pre-trained vision-language models [23], or identifying artifacts introduced by the upsampling layers [24]. However, these approaches usually rely on specific image patterns or diffusion models, such as relying on specific models for reconstruction [22], requiring relevant generated images as reference sets [23], depending on architectural features like upsampling layers [24], or relying on model-specific patterns in training dataset [25], making them less effective when detecting images generated from unseen diffusion models or under various perturbations and leading to limited generalization and robustness.

With all the above concerns in mind, we raise the following question: Can we develop a general and robust diffusion-generated image detector based on their inherent difference with natural real images? As we do not rely on specific diffusion-generated images, the detector should be sufficiently general and robust. To this end, we first analyze the difference between natural real images and diffusion-generated images in the frequency domain, as the frequency domain contains more distinguishable information than the pixel domain [26]–[39]. The analysis results are shown in Fig. 1. We can observe that there exists a clear discrepancy between natural real images and diffusion-generated images in their Fourier spectrum, specifically in the mid- and high-frequency bands, which could serve as discriminative clues for detection.

To fully leverage this, we conduct a comprehensive frequency analysis on diffusion-generated images and natural real images to explore the intrinsic frequency clues, which indicates that the discriminability increases from low- to high-frequency bands. Based on our analysis, we further propose a simple yet effective representation by enhancing the **F**requency **F**orgery **C**lue ($\mathbf{F}^2\mathbf{C}$) across all frequency bands. Specifically, we design a specific frequency-selective function that serves as the weighted filter banks, which restrains the less discriminative bands (*i.e.*, low frequency) and enhances more significant discriminative frequency bands (*i.e.*, high-frequency) in the Fourier spectrum, thus leading to more discriminative representation than original images. Compared to detectors that focus on certain patterns in diffusion models, our representation leverages the intrinsic difference between diffusion-generated images and natural real images, which should be more general to discriminate unseen diffusion-generated images from real ones, *i.e.*, regardless of architectures or hyperparameters. Compared to detectors only focusing on certain bands, *i.e.*, high frequency, our representation explores the discriminative clues existing across all different bands, which should be more robust to different settings and various perturbations, such as noise or compression. Our main contributions are summarized as follows:

- We conduct a comprehensive frequency analysis on natural real images and diffusion-generated fake images and find that the diffusion-generated images exhibit increasingly significant differences with natural real images, from low- to high-frequency bands.
- To leverage this, we propose a simple yet effective representation $\mathbf{F}^2\mathbf{C}$ by designing a frequency-selective function that serves as weighted filter banks for restrain-

Daichi Zhang, Tong Zhang and Sabine Süsstrunk are with the School of Computer and Communication Sciences, EPFL, 1015 Lausanne, Switzerland (e-mail: daichi.zhang@epfl.ch; tong.zhang@epfl.ch; sabine.susstrunk@epfl.ch).

Shiming Ge is with the Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100092, China, and University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: geshiming@iie.ac.cn).

Work done when Daichi Zhang at EFPL.

Shiming Ge is the corresponding author (e-mail: geshiming@iie.ac.cn).

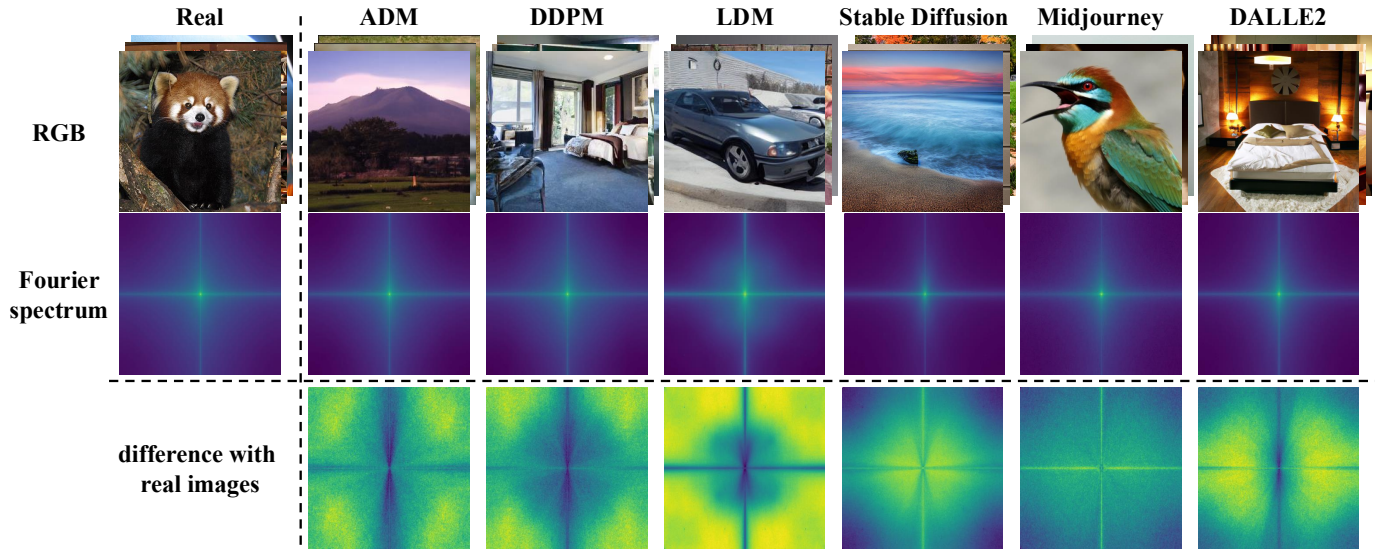

Fig. 1. **The magnitude difference between real image and different diffusion-generated models.** The fake images generated by different diffusion models (top) leave traces in their Fourier spectrum (middle). We explore their differences with natural real images for the detection of the diffusion-generated images (bottom). The Fourier spectrums are averaged on 1000 sampled images, which contain the same classes to avoid variance in content. The darker the color, the smaller the magnitude; the lighter the color, the larger the magnitude.

ing the less discriminative bands (*i.e.*, low-frequency) and enhancing the more discriminative frequency bands(*i.e.*, high-frequency) in the Fourier spectrum.

- Extensive experiments on various public diffusion-generated image datasets demonstrate the superiority of our proposed method against other state-of-the-art competitors with impressive generalization and robustness.

II. Related Work

A. Diffusion Models

Diffusion Models have achieved remarkable success in image synthesis task [1]–[12]. The main idea of diffusion models is inspired by the non-equilibrium thermodynamics proposed in [40]. Typically, diffusion models define two Markov chains of diffusion steps that first slowly add Gaussian noise to clean images, until disturbing them into isotropic Gaussian noise (termed diffusion or forward process); then they learn to reverse the diffusion process to generate clean samples from the noise (termed denoising or reverse process). Due to substantial efforts focusing on improving model architectures [5], [41], sampling methods [42]–[44], and optimizing processes [9], [45], [46], recent diffusion models are capable of generating high-quality images beyond human imagination at an extremely low cost, which can be a double-edged sword. Thus, developing general and robust diffusion-generated image detectors has recently become a critical issue.

B. Generated Image Detection

Recent generative models, such as GANs [47]–[50] and Diffusion models [1]–[12], have achieved remarkable success in image generation task. Various detectors have been proposed to prevent the malicious use of generated images [14]–[21]. To develop a general GAN-generated image detector, [25] introduce carefully designed pre- and post-processing with data augmentation. To detect generated and manipulated images, [14] and [51] propose to use patch-level artifacts. Recently, [22] found that diffusion-generated images are easier to reconstruct by diffusion models than real images. They propose a representation DIRE based on reconstruction error. [23] propose to use the pre-trained vision-language models to learn discriminative clues for detection. NPR [24] explore the artifacts introduced by the up-sampling layer in diffusion model architectures. These methods still, however, highly rely on specific patterns in training diffusion-generated images, which could lead to performance drops when detecting unseen models. Instead, we leverage the intrinsic statistic difference with natural real images in the frequency domain to discriminate from diffusion-generated images.

C. Frequency Artifacts in Generated Images

Some prior works have demonstrated that generated images exist artifacts in the frequency domain [27]–[39]. To detect GAN-generated images, [52] analyzes artifacts by discrete cosine transform (DCT). [27] leverage the Fourier spectrum discrepancy in GAN- and VAE-generated images. [15], [53] propose exploring the frequency clues to detect generated and manipulated images. There are already some works that find specific patterns in diffusion-generated images in the frequency domain which could serve as clues for detection, such as [28] analyze the frequency fingerprints of different diffusion models, [29], [54] analyze the artifacts in both spatial and frequency domains, [55] train a mask on Fourier spectrum and use the cosine similarity with a reference set for detection. These methods, however, focus mainly on the frequency distribution of specific diffusion-generated images or certain frequency bands, which may lead to limited generalization and robustness. Instead, we focus on the more general frequency difference between natural real images and diffusion-generated

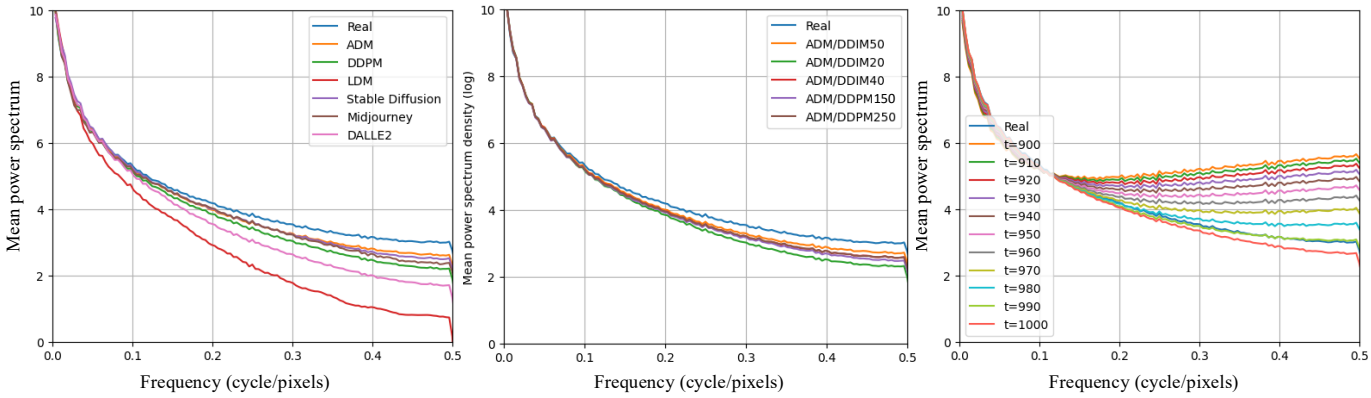

Fig. 2. **Mean power spectrum** of natural real and diffusion-generated images from different diffusion models (a) and different time steps (b). We further explore the spectrum during the denoising process in (c).

images and propose a more robust representation across all different bands by restraining the less discriminative bands and enhancing the more discriminative ones.

III. Methodology

A. Frequency Forgery Clue

We first analyze the difference between natural real images and diffusion-generated images in the frequency domain, as shown in Fig 3 (a). Specifically, to transform an image $\mathbf{x}$ from the spatial domain to the frequency domain, we first transform it into a grayscale image since the color information contributes less to the frequency distribution of an image. Then we compute the Discrete Fourier Transform and the mean power spectrum $F(\mathbf{x})$ that can be formulated as follows:

$$F(\mathbf{x}) = \log|\mathrm{DFT}(\mathbf{x})|^2, \tag{1}$$

where $\mathbf{x}$ is the input image, the $\mathrm{DFT}(\cdot)$ is the Discrete Fourier Transform and $|\cdot|$ computes the magnitude on each pixel. In practice, we use the FFT algorithm to compute the DFT. We compute and visualize the mean power spectrum on images generated from different diffusion models and different time steps, as shown in Fig. 2 (a) and (b), respectively.

From the results, we observe that the diffusion-generated images and natural real images exhibit significant differences in the frequency domain: their discrepancy becomes increasingly discriminative from low- to high-frequency bands. This phenomenon can be reflected in pixel space as the diffusion-generated images are usually smoother and lack the high-fidelity details that indicate the high-frequency parts. Besides, the images generated by different diffusion models and different timesteps exhibit different frequency patterns, and more time steps lead to high similarity to real images. And all of the generated images follow the above observation that their discrepancies with real images become increasingly discriminative from the low to high frequencies. Thus, we can draw following conclusion from above analysis:

Remark 1: Diffusion-generated images exhibit significant discrepancies with natural real images. These discrepancies are increasingly discriminative from low- to high-frequency bands.

Furthermore, we investigate what causes this phenomenon during the diffusion forward/backward processes. Specifically, we analyze the mean power spectrum of the intermediate results of the last 100 time steps during the DDPM [1] 1000-step denoising process by using the ADM [2] since the image details are synthesized around the end step [56], as illustrated in Fig. 2 (c). We observe that the mid-high-frequency parts of generated images are generated towards mainly the end steps of the denoising process. And these end steps determine the underestimation of mid-high-frequency bands. Note that during training of diffusion models, a neural network ϵ_θ is optimized to predict the added noise, given the noisy image $\mathbf{x}_t$ and corresponding time step t, the optimization target is a sampling and denoising process which can be defined as follows:

$$L_\theta(\mathbf{x}_0, t) = \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2, \tag{2}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, $\mathbf{x}_0$ is clean image, and α_t is the predefined noise schedule. By analyzing the optimization process above, we argue that the underestimation at the end steps is caused by the optimization objective described in Eq. 2 because the generated images towards the end steps are closer to denoised clean images. This makes it more difficult to predict the added noise when compared to pure Gaussian distributions, but the Eq. 2 treats equally the denoising tasks at different noise levels. There are also some existing analyses [9], [28] that agree that this objective cannot lead to good likelihood values, and some objectives are designed to address this issue [57], [58]. But still, these methods struggle to synthesize the high-frequency details. Thus, we can draw the following conclusion for the cause of the frequency discrepancy:

Remark 2: The spectrum discrepancy is highly related to the challenging optimization objective, when towards denoising step $t = 0$.

Moreover, some prior studies [26], [59], [60] show that the mean power spectrum of natural real images has the following rule:

$$S(f) \propto f^{-\alpha}, \alpha \approx 2, \tag{3}$$

where $S(\cdot)$ is the mean power spectrum and f is the frequency.

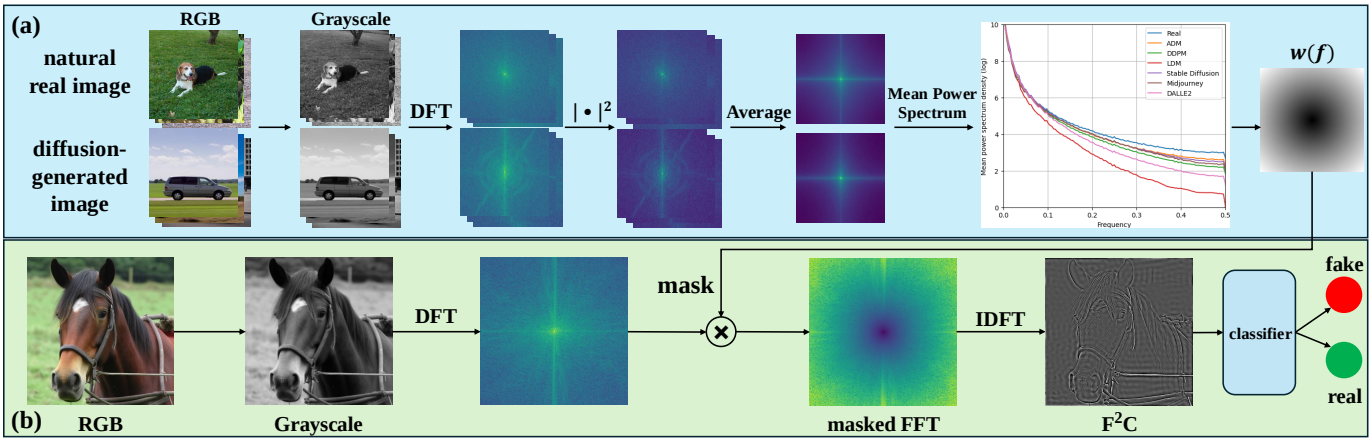

Fig. 3. **Overview of our proposed method.** We first analyze the discrepancy of mean power spectrum between natural real and diffusion-generated images, as shown at the top (a). Based on the analysis, we design a specific frequency-selective function $w(f)$ that serves as the filter banks on the Fourier spectrum to restrain the less discriminative frequency bands and to enhance the more discriminative ones, thus leading to more discriminative representation, as shown at the bottom (b).

B. Frequency-Selective Function

With the above observation that diffusion-generated images exist frequency discrepancy with natural real images, it comes to our mind that if we could design a representation that exploits the most discriminative clues and removes the similar patterns in the frequency domain, we could obtain a more effective representation for distinguishing the diffusion-generated images from real ones. To this end, we propose a simple yet effective representation by enhancing **F**requency **F**orgery **C**lue (**F**^{2}**C**) for diffusion-generated image detection. Specifically, we design a weighted filter which restrains the less discriminative (low frequency) bands and enhances those more discriminative (mid-high frequency), as shown in Fig 3 (b). As the representation is designed by the general observation and the principle on natural real and diffusion-generated fake images, our representation should be general and robust for detection.

To achieve this, we first compute the frequency spectrum deviation between natural real and diffusion-generated images, based on Fig. 2 (a)&(b). Specifically, we compute the subtraction of each spectrum with the one on natural real images, and visualize them with a scaling factor, as shown in Fig. 4 (a)&(b). We aim to design a frequency-selective function $w(\cdot)$ based on the distribution above to serve as the weighted filter banks applying on the Fourier spectrum to restrain the less discriminative bands and to enhance the more discriminative bands. Then, we inverse the enhanced Fourier spectrum to RGB space, thus leading to a more discriminative representation than the original RGB images. This process can be formulated as follows:

$$\mathbf{F}^2\mathbf{C}(\mathbf{x}) = \mathrm{IDFT}(\mathrm{DFT}(\mathbf{x}) \cdot w(f)), \tag{4}$$

where the $\mathbf{x}$ is the input image, $\mathrm{DFT}(\cdot)$ is the Discrete Fourier Transform, and the $\mathrm{IDFT}(\cdot)$ is the Inverse Discrete Fourier Transform. In practice, we use the FFT algorithm to compute them.

For the frequency-selective function $w(\cdot)$, we introduce two following principles based on the analysis above to process the less discriminative band (*i.e.*, low frequency), and more discriminative band (*i.e.*, mid-high frequency), respectively, described as follows:

Low-frequency band. In Fig. 4 (a)&(b), the low-frequency part of real and diffusion-generated images exhibits high similarity, which indicates that there is no significant discrepancy in this band. Hence, we remove the low-frequency information to restrain the less discriminative band by simply setting the weight to zero, formulated as follows:

$$w(f) = 0, f \leq \tau, \tag{5}$$

where τ is the threshold for the low frequency, and we empirically set $\tau = 0.1$.

Mid-high frequency band. The mid-high frequency parts are increasingly discriminative, as indicated in Fig. 4 (a)&(b). Following the principle that higher weights should be assigned to more discriminative bands, we compute the weights, based on their discrepancy. To this end, we introduce another kernel function $k(\cdot)$ to fit the power spectrum discrepancy distribution in Fig. 4 (a)&(b), which can be formulated as follows:

$$k(f) = |log|G_1(f)|^2 - log|G_0(f)|^2|, \tag{6}$$

where $\{G_1(f), G_0(f)\}$ are the Discrete Fourier Transform distribution of diffusion-generated and natural real images, respectively. As we only care about the discrepancy between them, we can simplify the above equation as follows:

$$|G_1(f)| = e^{\frac{k(f)}{2}} \cdot |G_0(f)|. \tag{7}$$

Note that, for natural real images, their frequency distribution follows the principle described in Eq. 3. Therefore, we choose $\alpha = 2$ which should be an appropriate parameter to approximate the statics of real images, formulated as:

$$S_0(f) = |G_0(f)|^2 = \frac{1}{f^2}. \tag{8}$$

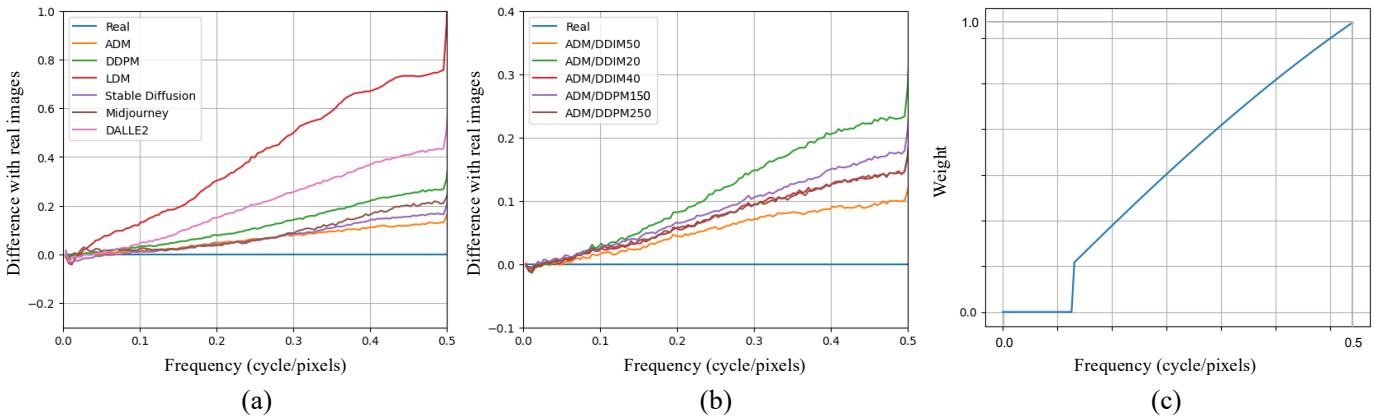

Fig. 4. **Spectrum discrepancy** between natural real and diffusion-generated images in (a) and (b). We further design the *frequency-selective function* based on the discrepancy as shown in (c).

Thus, we can further compute the discrepancy of each frequency band, based on Eq. 7 & 8 to obtain the desired frequency-selective function as follows:

$$w(f) = ||G_1(f)| - |G_0(f)|| = (e^{\frac{k(f)}{2}} - 1) \cdot \frac{1}{f} \quad (9)$$

Considering both the low and mid-high frequency described above, our final designed frequency-selective function is as follows:

$$w(f) = \begin{cases} 0, & f \leq \tau \\ (e^{\frac{k(f)}{2}} - 1) \cdot \frac{1}{f}, & f > \tau \end{cases} \quad (10)$$

Furthermore, based on our observations, we choose the two-degree linear function (quadratic function) as the kernel function with the coefficients $k(f) = -0.2f^2 + 0.8f - 0.05$ since it has the minimum error after fitting to the distributions. The corresponding designed function $w(f)$ is shown in Fig. 4 (c), which restrains the less discriminative band, (low frequency), and enhances the more discriminative band (mid-high frequency), thus leading to a more discriminative representation.

C. Diffusion-Generated Image Detection

After the representation learning stage, we can obtain the $\mathbf{F^2C}$ representations for both natural real and diffusion-generated images. We further use the representations as input to train a naive binary classifier to distinguish the real and generated images by a simple binary cross-entropy loss, which is formulated as follows:

$$L(y, \hat{y}) = -\Sigma_{i=1}^{n} \left(y_i \log(\hat{y}_i) + (1 - y_i) \log(1 - \hat{y}_i) \right), \quad (11)$$

where n is the mini-batch size, $y \in \{0, 1\}$ is the ground-truth label for real and fake, and $\hat{y}$ is the output prediction of the classifier. We choose ResNet-50 [61] with a fully-connected layer as our classifier. And during the inference stage, we input the $\mathbf{F^2C}$ representation to the trained classifier that could be classified as real or diffusion-generated. After we choose the frequency-selective function, the representation is easy to obtain and the inference stage can be simply conducted with trained classifier.

IV. Experiment

A. Experimental Setup

Dataset. Following recent state-of-the-art diffusion-generated image detectors [22]–[24], [62], we evaluate our proposed method on three public diffusion-generated image datasets, including (1) GenImage [62], (2) UniformerDiffusion [23], and (3) DiffusionForensics [22]. We present the results on challenging GenImage below, which includes ADM [2], Glide [3], Midjourney [4], Stable-Diffusion-v1.4 [5], Stable-Diffusion-v1.5 [5], VQDM [6], Wukong [7]. More results on UniformerDiffusion and DiffusionForensics are presented in our supplementary material. For training set, we use the fake images generated from ADM trained on ImageNet and real images from ImageNet, which contain 40,000 fake and real images, respectively. We choose the simple and early proposed ADM as the training set since it should be more challenging and difficult to generalize compared to using other more recent diffusion models, such as Stable-Diffusion 1.4.

Evaluation metric. Following prior state-of-the-art methods [22], [23], [25], we report the average precision (AP) and accuracy (ACC) with a fixed 0.5 threshold. We test three times and compare the averaged results.

Baselines. For fair and comprehensive comparisons, we choose and categorize four different types of state-of-the-art detectors: traditional image classification backbones (including (1) ResNet-50 [61] and (2) Swin-T [63]), deepfake detectors (including (3) Patchfor [14] and (4) F3Net [15]), diffusion-generated image detectors ((5) DIRE [22]), and universal detectors (including (6) CNNDet [25], (7) uniFD [23], (8) NPR [24], (9) SAFE [19], (10) C2P-CLIP [20]), and (11) VIB-Net [21]. For all aforementioned baselines, we use the same training and testing settings. Please refer to the appendix for more details.

B. Comparison to the State-of-the-Art

Generalization to unknown models. We first evaluate the generalization of our proposed on unknown diffusion models, which is a major challenge in this task. Specifically, we train all detectors with the same training dataset generated

TABLE I
GENERALIZATION RESULTS ON GENIMAGE DATASET. WE REPORT THE DETECTION ACCURACY AND AVERAGE PRECISION (ACC/AP) AVERAGED OVER REAL AND FAKE IMAGES ON UNKNOWN DIFFUSION MODELS.

Detection method	Different Diffusion Models in GenImage							Total
	ADM	Glide	Midjourney	SD-v1.4	SD-v1.5	VQDM	Wukong	Avg.
ResNet-50 [61]	81.20/97.42	80.01/93.05	60.85/66.55	55.45/60.71	54.10/60.42	76.40/87.73	51.01/53.59	65.57/74.21
Swin-T [63]	74.84/88.45	75.69/93.59	62.48/70.85	70.69/78.65	71.19/78.02	73.49/75.61	70.03/77.46	71.20/80.38
Patchfor [14]	99.65/99.43	99.81/99.63	57.19/85.88	50.59/61.62	50.64/61.51	99.83/99.95	50.52/61.89	72.60/81.42
F3Net [15]	99.64/99.99	99.84/99.99	51.48/74.93	50.02/59.41	50.33/61.98	99.94/99.99	50.13/52.61	71.63/78.41
DIRE [22]	61.35/97.91	61.65/99.17	61.65/94.83	59.55/92.09	59.30/92.94	61.05/96.88	58.70/88.31	60.46/94.59
CNNDet [25]	64.55/83.73	62.45/70.72	51.15/50.69	56.30/54.72	54.30/54.08	62.70/69.97	56.85/57.76	58.33/63.10
UniFD [23]	72.45/91.45	62.30/63.65	53.50/50.83	67.00/78.58	67.10/74.38	72.25/95.35	70.45/85.94	66.44/77.17
NPR [24]	77.90/96.91	77.95/93.69	73.30/86.76	75.40/83.14	73.50/83.40	80.15/91.35	74.00/87.88	76.03/89.02
SAFE [19]	92.18/96.78	95.83/98.90	95.21/98.76	95.30/97.91	96.02/97.13	96.33/99.61	98.28/98.99	95.59/98.30
C2P-CLIP [20]	96.87/97.72	98.65/99.05	87.21/89.33	94.89/95.41	95.56/97.89	96.33/98.58	99.08/99.90	95.51/96.84
VIB-Net [21]	93.58/95.64	84.87/97.18	89.51/97.85	95.56/98.19	95.21/98.72	89.83/97.33	98.39/99.63	92.42/97.79
F^2C	99.95/100.0	99.95/100.0	99.95/100.0	99.90/99.99	99.95/100.0	99.90/100.0	99.80/100.0	99.91/100.0

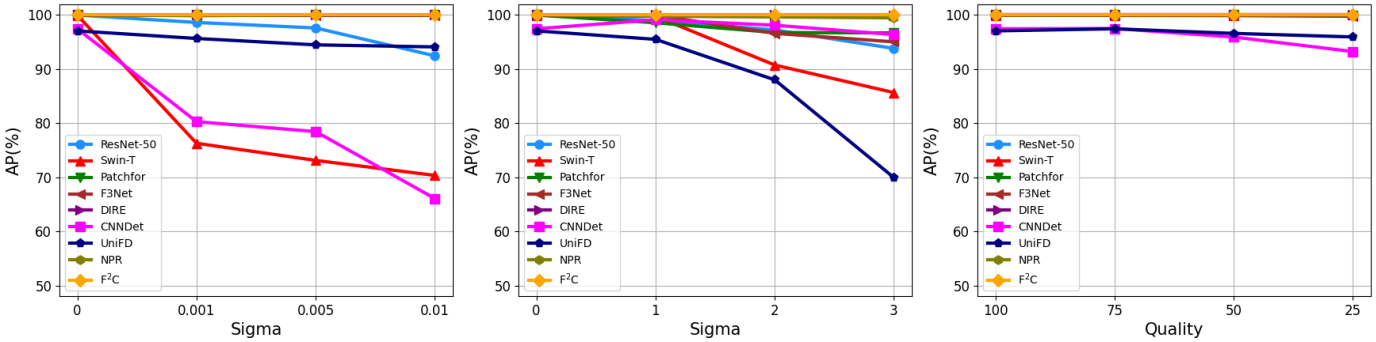

Fig. 5. **Robustness results to unseen perturbations.** Average precision (AP) of different methods, when detecting real/fake images from ProGAN under three different types of perturbations with three different severity levels: Gaussian Noise ($\sigma = 0.001, 0.005, 0.01$), Gaussian Blur ($\sigma = 1, 2, 3$), and JPEG Compression ($quality = 75, 50, 25$) (from left to right).

TABLE II
ABLATION STUDY ON DIFFERENT IMAGE REPRESENTATION. WE REPORT THE ACC/AP RESULTS ON THE GENIMAGE DATASET THAT INDICATES OUR DESIGNED REPRESENTATION CAN ACHIEVE IMPROVED PERFORMANCE.

Representation	Different Diffusion Models in GenImage							Total
	ADM	Glide	Midjourney	SD-v1.4	SD-v1.5	VQDM	Wukong	Avg.
RGB	78.40/96.28	76.95/89.49	60.10/65.20	55.60/60.94	53.75/59.68	71.65/80.70	50.60/54.01	63.86/72.33
Grayscale	97.90/99.75	98.00/99.81	75.85/89.44	65.65/81.87	65.65/82.68	92.90/98.61	67.00/83.52	80.42/90.81
F^2C	99.95/100.0	99.95/100.0	99.95/100.0	99.90/99.99	99.95/100.0	99.90/100.0	99.80/100.0	99.91/100.0

from ADM on ImageNet, then we evaluate them on the three aforementioned public datasets. We first evaluate on the challenging GenImage, which is a recent and diversified dataset with multi classes trained on ImageNet. The ACC/AP results are presented in Tab. I. From the results, we observe that all baseline detectors have a slight performance drop when encountering more diversified generated images, which is a challenging setting for existing detectors. Among these detectors, our method still achieves impressive generalization with 99.91% average ACC, with 5.41% and 10.98% AP improvements compared to the recent DIRE and NPR.

The impressive performance across the various diffusion models and settings further demonstrates the superiority of our proposed **F^2C** representation, as it restrains the less discriminative clues and enhances those more discriminative in the frequency domain for detection.

Robustness to unseen perturbations. The robustness is also a major concern for existing detectors, as there are various but common post-preprocessing perturbations in real-scenario applications, such as compression. To address this issue, we evaluate all detectors' robustness against three common but widely used perturbations on images generated from

TABLE III
ABLATION STUDY ON DIFFERENT KERNEL FUNCTIONS FOR MID-HIGH FREQUENCIES. WE REPORT THE ACC/AP ON THE GENIMAGE DATASET, FROM WHICH WE OBSERVE THAT ONLY A SUITABLE FUNCTION CAN ACHIEVE IMPRESSIVE PERFORMANCE.

Kernel function	Different Diffusion Models in GenImage							Total
	ADM	Glide	Midjourney	SD-v1.4	SD-v1.5	VQDM	Wukong	Avg.
$k(f) = f$	99.15/99.93	99.50/99.91	99.30/99.97	98.85/99.80	99.20/99.97	99.10/99.97	98.35/99.87	99.06/99.92
$k(f) = a \cdot e^{bf} + c$	50.00/54.00	50.00/54.00	50.00/54.00	50.00/54.00	50.00/54.00	50.00/54.00	50.00/52.92	50.00/53.85
$k(f) = a \cdot log(bf) + c$	99.75/99.99	99.95/100.0	99.95/100.0	99.95/100.0	99.90/99.97	99.90/99.98	99.90/99.98	99.90/99.99
$k(f) = af^2 + bf + c$	99.95/100.0	99.95/100.0	99.95/100.0	99.90/99.99	99.95/100.0	99.90/100.0	99.80/100.0	99.91/100.0

TABLE IV
ABLATION STUDY ON DIFFERENT LINEAR DEGREES. WE REPORT THE ACC/AP RESULTS ON THE GENIMAGE DATASET, FROM WHICH WE OBSERVE THAT BOTH TOO SIMPLE AND COMPLEX LINEAR FUNCTIONS CAN LEAD TO A SLIGHT PERFORMANCE DROP.

Kernel function	Different Diffusion Models in GenImage							Total
	ADM	Glide	Midjourney	SD-v1.4	SD-v1.5	VQDM	Wukong	Avg.
$k(f) = af^3 + bf^2 + cf + d$	100.0/100.0	99.90/100.0	99.95/100.0	99.90/99.99	99.80/99.99	99.85/100.0	99.40/99.99	99.83/100.0
$k(f) = af + b$	99.85/100.0	99.90/100.0	99.80/99.99	99.50/99.96	99.70/99.99	99.70/100.0	99.10/99.98	99.65/99.99
$k(f) = af^2 + bf + c$	99.95/100.0	99.95/100.0	99.95/100.0	99.90/99.99	99.95/100.0	99.90/100.0	99.80/100.0	99.91/100.0

ADM (the same as the training set), including Gaussian Noise, Gaussian Blur, and JPEG Compression, following [22], [25]. For each perturbation, we employ three different severity levels to disrupt images: $\sigma = 0.001, 0.005, 0.01$ for Gaussian Noise, $\sigma = 1, 2, 3$ for Gaussian Blur, and $quality = 75, 50, 25$ for JPEG Compression. The results are shown in Fig. 5. From the results, we first observe that existing detectors would suffer from common perturbations, especially for Gaussian Noise. This indicates that some of the representations these detectors rely on might not be sufficiently robust to real scenario disruptions. Our proposed representation suffers significantly less from the above three perturbations, with only slight or even no performance drops. This indicates that, by exploring the discriminative clues with natural real images across all frequency bands, our proposed representation has impressive robustness against common perturbations.

Evaluation on GAN-generated images. The generalization to GAN-generated images is also a critical issue. To validate this, we conduct further cross-domain experiments on GAN-generated images from DiffusionForensics [64](including ProjGAN [65], StyleGAN [49], Diff-ProjGAN [66], and Diff-StyleGAN [67]) as shown in Tab. V. There is no GAN-generated image in training data in this setting. We observe that our method can still achieve impressive performance on GAN-generated images, which indicates the phenomenon of frequency distribution we find on Diffusion-generated images may also benefit detecting GAN-generated images.

TABLE V
GENERALIZATION (ACC/AP) ON GAN-GENERATED IMAGES.

Method	Different GAN Models in DiffusionForensics			
	StyleGAN	ProjGAN	Diff-StyleGAN	Diff-ProjGAN
F3Net	88.1/95.5	74.4/86.0	85.5/94.4	70.2/83.0
CNNDet	94.3/99.8	62.2/93.2	68.1/91.4	60.0/92.6
$\mathbf{F^2C}$	99.8/99.9	99.7/99.1	99.5/99.7	99.0/99.8

C. Ablation Study

Comparison with different image representations. To examine whether our proposed representation is better than other image representations for detecting diffusion-generated images, we first conduct further ablation studies on various inputs for detection, including RGB and grayscale images. The results on GenImage dataset are presented in Tab. II, which indicates that RGB and grayscale images cannot achieve the desired generalization on unknown diffusion models. One explanation could be that pixel space does not share common distributions among different diffusion models. Their comparisons with our proposed $\mathbf{F^2C}$ demonstrate that our representation serves as a general and robust image representation, thus contributing to a generalizable detector than simply using RGB images. This also provides more evidence for the superiority of our method by exploring the discriminative clues with natural real images in the frequency domain.

Effect of different kernel functions on mid-high frequency. We use the two-degree linear function (quadratic function) as the kernel function $k(f)$ to achieve the enhanced $\mathbf{F^2C}$ representation during the evaluation above. To examine whether this is an optimal function, we conduct further ablation studies by using different kernel functions to fit the frequency distributions. We choose the following functions: simple linear function $k(f) = f$, exponential function, and logarithm function. The parameters of exponential and logarithm functions are set by fit to distributions above ($k(f) = 650 \cdot e^{0.3f} - 650$ for exponential and $k(f) = 0.18 \cdot log(0.25f) + 0.48$). The results are presented in Tab. III. We observe that different kernel functions lead to different performance, which indicates that a suitable frequency-selective function is necessary for the enhanced

TABLE VI
ABLATION STUDY ON LOW-FREQUENCY BANDS. WE REPORT THE ACC/AP RESULTS ON GENIMAGE DATASET, WHICH INDICATES INTRODUCING TOO MUCH LOW-FREQUENCY OR IGNORING TOO MUCH MID-HIGH-FREQUENCY INFORMATION CAN UNDERMINE THE PERFORMANCE.

Minimum threshold τ	Different Diffusion Models in GenImage							Total
	ADM	Glide	Midjourney	SD-v1.4	SD-v1.5	VQDM	Wukong	Avg.
0.00	99.35/99.99	99.80/99.99	99.75/99.95	99.30/99.96	99.20/99.98	99.65/99.96	98.75/99.96	99.40/99.97
0.20	99.75/100.0	99.85/100.0	99.95/100.0	99.90/99.99	99.95/100.0	99.90/100.0	99.78/100.0	99.87/100.0
0.10	99.95/100.0	99.95/100.0	99.95/100.0	99.90/99.99	99.95/100.0	99.90/100.0	99.80/100.0	99.91/100.0

representation, *e.g.*, the exponential function is not a suitable function. We argue that a suitable and desired function should fit the distributions properly without overfitting or underfitting. The logarithm functions achieve impressive performance, and quadratic functions further improve the results, which indicates that our specifically designed frequency-selective function is a suitable function for restraining the less discriminative bands and enhancing those more discriminative ones.

Effect of different linear degrees for kernel function. We conduct ablation experiments by employing linear function with different degrees, *i.e.*, linear/quadratic/cubic functions. The results are shown in Tab. IV (the coefficients for three- and one-degree linear functions are: $k(f) = -4f^3 + 3.2f^2 - 0.16f + 0.02$ and $k(f) = f - 0.04$), from which we observe that both low- and high-degree linear functions lead to slight performance drops. One explanation could be that both the too simple and the complicated linear functions cannot fit the distributions properly, *i.e.*, with underfitting for the low degree functions and overfitting for the high-degree ones. The comparisons also demonstrate that, to obtain the desired representation, a two-degree linear function is a simple yet suitable choice for restraining and enhancing different frequency bands. Moreover, we also believe that if there exist other functions that fit the distribution properly, they can also be employed to obtain the enhanced representation.

Effect of the minimum threshold on low frequency We conduct further ablation studies on the threshold for restraining low-frequency bands by employing a different minimum threshold τ or not, as presented in Tab. VI. We observe that the performance is improved when employing a suitable minimum threshold to restrain the low-frequency band. This demonstrates that low-frequency band cannot provide discriminative information for diffusion-generated image detection and that eliminating them could boost the performance. Additionally, the performance will drop when the threshold is too low or too high, which indicates that both introducing too much low-frequency information or ignoring too much mid-high frequency information could undermine the performance. Thus, the threshold for low-frequency band should be chosen carefully and $\tau = 0.1$ is a suitable choice.

D. Visualization

To analyze our designed representation more directly, we visualize the Fourier spectrum and the corresponding enhanced representation on real and ADM-generated images respectively, as shown in Fig. 6. We observe that our designed representations remove the low-frequency information, which is less discriminative, and that they enhance the mid- and high-frequency parts, such as edges and details, which are more discriminative. The representations on real images preserve more mid- and high-frequency information of original images compared to diffusion-generated images that are more distinguishable serving as clues for the detection task.

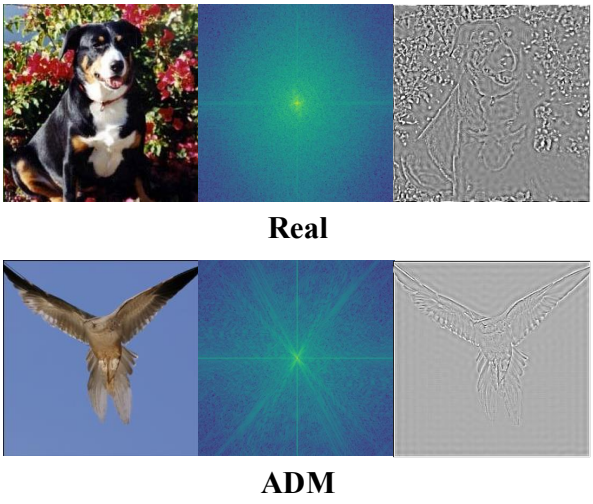

Fig. 6. **The visualization of Fourier spectrum and our designed representation** on real and different diffusion-generated images. We observe that our representation enhances the mid-high-frequency clues and removes low-frequency information, which makes it more discriminative.

V. CONCLUSION

In this paper, we first conduct a comprehensive analysis that shows the diffusion-generated images exhibit increasing differences with natural real images from low- to high-frequency bands. Upon this observation, we propose a simple yet effective representation $\mathbf{F}^2\mathbf{C}$ by designing a specific frequency-selective function that serves as the weighted filter banks on the Fourier spectrum to restrain the less-discriminative frequency bands, *low-frequency* and to enhance the more discriminative ones, *high-frequency*. Extensive experiments on various public diffusion-generated image datasets demonstrate the superiority of our proposed method. We hope our method could provide insights for detecting generated data from the perspective of natural real data. In the future, we aim to extend our idea and method to other AI-generated content (AIGC) detection tasks to facilitate the development of AIGC safety.

REFERENCES

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in Neural Information Processing Systems*, vol. 33, pp. 6840–6851, 2020.

[2] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in Neural Information Processing Systems*, vol. 34, pp. 8780–8794, 2021.

[3] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[4] Midjourney, 2023. [Online]. Available: https://www.midjourney.com/home/

[5] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 684–10 695.

[6] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo, "Vector quantized diffusion model for text-to-image synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 696–10 706.

[7] Wukong, 2022. [Online]. Available: https://xihe.mindspore.cn/modelzoo/wukong

[8] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, 2021, pp. 8821–8831.

[9] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International Conference on Machine Learning*, 2021, pp. 8162–8171.

[10] L. Liu, Y. Ren, Z. Lin, and Z. Zhao, "Pseudo numerical methods for diffusion models on manifolds," *arXiv preprint arXiv:2202.09778*, 2022.

[11] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[12] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[13] K. Devlin and J. Cheetham, "Fake trump arrest photos: How to spot an ai-generated image," *BBC News*, vol. 24, 2023.

[14] L. Chai, D. Bau, S.-N. Lim, and P. Isola, "What makes fake images detectable? understanding properties that generalize," in *Proceedings of the European Conference on Computer Vision*, 2020, pp. 103–120.

[15] 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 the European Conference on Computer Vision*, 2020, pp. 86–103.

[16] D. Cozzolino, G. Poggi, M. Nießner, and L. Verdoliva, "Zero-shot detection of ai-generated images," in *Proceedings of the European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., vol. 15076, 2024, pp. 54–72.

[17] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE transactions on image processing*, vol. 23, no. 2, pp. 710–724, 2013.

[18] N. Zhong, H. Chen, Y. Xu, Z. Qian, and X. Zhang, "Beyond generation: A diffusion-based low-level feature extractor for detecting ai-generated images," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 8258–8268.

[19] O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and F. Feng, "Improving synthetic image detection towards generalization: An image transformation perspective," in *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 2025, pp. 2405–2414.

[20] C. Tan, R. Tao, H. Liu, G. Gu, B. Wu, Y. Zhao, and Y. Wei, "C2p-clip: Injecting category common prompt in clip to enhance generalization in deepfake detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, no. 7, 2025, pp. 7184–7192.

[21] H. Zhang, Q. He, X. Bi, W. Li, B. Liu, and B. Xiao, "Towards universal ai-generated image detection by variational information bottleneck network," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 23 828–23 837.

[22] Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li, "Dire for diffusion-generated image detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 445–22 455.

[23] U. Ojha, Y. Li, and Y. J. 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. 24 480–24 489.

[24] C. Tan, Y. Zhao, S. Wei, G. Gu, P. Liu, and Y. Wei, "Rethinking the upsampling operations in cnn-based generative network for generalizable deepfake detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 28 130–28 139.

[25] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8695–8704.

[26] v. A. Van der Schaaf and J. v. van Hateren, "Modelling the power spectra of natural images: statistics and information," *Vision research*, vol. 36, no. 17, pp. 2759–2770, 1996.

[27] T. Dzanic, K. Shah, and F. Witherden, "Fourier spectrum discrepancies in deep network generated images," *Advances in Neural Information Processing Systems*, vol. 33, pp. 3022–3032, 2020.

[28] J. Ricker, S. Damm, T. Holz, and A. Fischer, "Towards the detection of diffusion model deepfakes," *arXiv preprint arXiv:2210.14571*, 2022.

[29] R. Corvi, D. Cozzolino, G. Poggi, K. Nagano, and L. Verdoliva, "Intriguing properties of synthetic images: from generative adversarial networks to diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 973–982.

[30] N. Yu, L. S. Davis, and M. 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.

[31] K. Chandrasegaran, N.-T. Tran, and N.-M. Cheung, "A closer look at fourier spectrum discrepancies for cnn-generated images detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 7200–7209.

[32] Q. Bammey, "Synthbuster: Towards detection of diffusion model generated images," *IEEE Open Journal of Signal Processing*, 2023.

[33] J. Zhang, Y. Wang, H. R. Tohidypour, and P. Nasiopoulos, "Detecting stable diffusion generated images using frequency artifacts: A case study on disney-style art," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1845–1849.

[34] 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 the AAAI Conference on Artificial Intelligence*, vol. 38, no. 5, 2024, pp. 5052–5060.

[35] Y. Jeong, D. Kim, Y. Ro, and J. Choi, "Frepgan: robust deepfake detection using frequency-level perturbations," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 1, 2022, pp. 1060–1068.

[36] 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 the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7278–7287.

[37] S. Yan, O. Li, J. Cai, Y. Hao, X. Jiang, Y. Hu, and W. Xie, "A sanity check for ai-generated image detection," *arXiv preprint arXiv:2406.19435*, 2024.

[38] R. Xia, D. Zhou, D. Liu, J. Li, L. Yuan, N. Wang, and X. Gao, "Inspector for face forgery detection: Defending against adversarial attacks from coarse to fine," *IEEE Transactions on Image Processing*, 2024.

[39] E. Prashnani, M. Goebel, and B. Manjunath, "Generalizable deepfake detection with phase-based motion analysis," *IEEE Transactions on Image Processing*, 2024.

[40] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International Conference on Machine Learning*, 2015, pp. 2256–2265.

[41] S. Welker, H. N. Chapman, and T. Gerkmann, "Driftrec: Adapting diffusion models to blind jpeg restoration," *IEEE transactions on image processing*, vol. 33, pp. 2795–2807, 2024.

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

[43] C. Lu, Y. Zhou, F. Bao, J. Chen, C. Li, and J. Zhu, "Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps," *Advances in Neural Information Processing Systems*, vol. 35, pp. 5775–5787, 2022.

[44] C. Xu, J. Yan, M. Yang, and C. Deng, "Rethinking noise sampling in class-imbalanced diffusion models," *IEEE Transactions on Image Processing*, 2024.

[45] J. Ho and T. Salimans, "Classifier-free diffusion guidance," *arXiv preprint arXiv:2207.12598*, 2022.

[46] Z. Huang, Y. Fan, C. Liu, W. Zhang, Y. Zhang, M. Salzmann, S. Süsstrunk, and J. Wang, "Fast adversarial training with adaptive step size," *IEEE Transactions on Image Processing*, vol. 32, pp. 6102–6114, 2023.

[47] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[48] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of gans for improved quality, stability, and variation," in *International Conference on Learning Representations*, 2018.

[49] T. Karras, S. Laine, and T. 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.

[50] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," in *International Conference on Learning Representations*, 2018.

[51] Y. Hua, R. Shi, P. Wang, and S. Ge, "Learning patch-channel correspondence for interpretable face forgery detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 1668–1680, 2023.

[52] J. Frank, T. Eisenhofer, L. Schönherr, A. Fischer, D. Kolossa, and T. Holz, "Leveraging frequency analysis for deep fake image recognition," in *International Conference on Machine Learning*, 2020, pp. 3247–3258.

[53] C. M. et al., "F^2trans: High-frequency fine-grained transformer for face forgery detection," *IEEE Transactions on Information Forensics and Security*, vol. 18, pp. 1039–1051, 2023.

[54] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, "On the detection of synthetic images generated by diffusion models," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2023, pp. 1–5.

[55] Y. Li, Q. Bammey, M. Gardella, T. Nikoukhah, J.-M. Morel, M. Colom, and R. G. Von Gioi, "Masksim: Detection of synthetic images by masked spectrum similarity analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 3855–3865.

[56] E. Hoogeboom, J. Heek, and T. Salimans, "simple diffusion: End-to-end diffusion for high resolution images," in *International Conference on Machine Learning*. PMLR, 2023, pp. 13 213–13 232.

[57] D. Samuel, R. Ben-Ari, N. Darshan, H. Maron, and G. Chechik, "Norm-guided latent space exploration for text-to-image generation," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[58] Y. Song, C. Durkan, I. Murray, and S. Ermon, "Maximum likelihood training of score-based diffusion models," *Advances in neural information processing systems*, vol. 34, pp. 1415–1428, 2021.

[59] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *Josa a*, vol. 4, no. 12, pp. 2379–2394, 1987.

[60] G. J. Burton and I. R. Moorhead, "Color and spatial structure in natural scenes," *Applied optics*, vol. 26, no. 1, pp. 157–170, 1987.

[61] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[62] 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," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[63] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[64] Z. W. et al., "Dire for diffusion-generated image detection," in *ICCV*, 2023, pp. 22 445–22 455.

[65] A. Sauer, K. Chitta, J. Müller, and A. Geiger, "Projected gans converge faster," *Advances in Neural Information Processing Systems*, vol. 34, pp. 17 480–17 492, 2021.

[66] Z. Wang, H. Zheng, P. He, W. Chen, and M. Zhou, "Diffusion-gan: Training gans with diffusion," *arXiv preprint arXiv:2206.02262*, 2022.

[67] K. Song, L. Han, B. Liu, D. Metaxas, and A. Elgammal, "Stylegan-fusion: Diffusion guided domain adaptation of image generators," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 5453–5463.