Spectrum Translation for Refinement of Image Generation (STIG) Based on Contrastive Learning and Spectral Filter Profile

Seokjun Lee[1,2], Seung-Won Jung[2], Hyunseok Seo[1]*

[1]Biomedical Research Division, Korea Institute of Science and Technology
[2]School of Electrical Engineering, Korea University
ykykyk112@kist.re.kr, swjung83@korea.ac.kr, seo@kist.kr

Abstract

Currently, image generation and synthesis have remarkably progressed with generative models. Despite photo-realistic results, intrinsic discrepancies are still observed in the frequency domain. The spectral discrepancy appeared not only in generative adversarial networks but in diffusion models. In this study, we propose a framework to effectively mitigate the disparity in frequency domain of the generated images to improve generative performance of both GAN and diffusion models. This is realized by spectrum translation for the refinement of image generation (STIG) based on contrastive learning. We adopt theoretical logic of frequency components in various generative networks. The key idea, here, is to refine the spectrum of the generated image via the concept of image-to-image translation and contrastive learning in terms of digital signal processing. We evaluate our framework across eight fake image datasets and various cutting-edge models to demonstrate the effectiveness of STIG. Our framework outperforms other cutting-edges showing significant decreases in FID and log frequency distance of spectrum. We further emphasize that STIG improves image quality by decreasing the spectral anomaly. Additionally, validation results present that the frequency-based deepfake detector confuses more in the case where fake spectrums are manipulated by STIG.

Introduction

In recent years, image generation and synthesis have noticeably progressed with various generative approaches (Radford, Metz, and Chintala 2015; Ho, Jain, and Abbeel 2020; Song et al. 2021). Further, many studies have shown ways to make the generative model produce more photo-realistic images which are not distinguishable by the human visual system (Karras, Laine, and Aila 2019; Karras et al. 2020; Choi et al. 2020; Nichol et al. 2021). In spite of their remarkable success, it is easy to find apparent discrepancies in the frequency spectrum of generated images. First, an upsampling operation in the most generative network can have aliasing on the spectrum (*i.e.*, checkerboard artifact in image domain) and insufficient high-frequency components (Zhang, Karaman, and Chang 2019; Frank et al. 2020). In
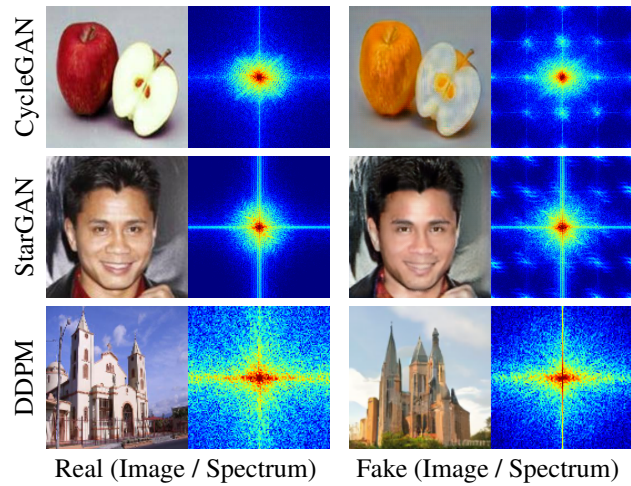
Figure 1: Spectral discrepancies between the real and generated image in various generative networks. CycleGAN (Zhu et al. 2017) and StarGAN (Choi et al. 2018), which include the transposed convolution layer (row 1 and row 2), produce the grid-based aliasing in the spectrum of the generated image. On the other hand, DDPM (row 3) (Ho, Jain, and Abbeel 2020) includes another type of discrepancy, the lack of high-frequency components.

addition, the spectral disparity has been reported in the diffusion model, which currently stands out in image generation and synthesis fields. Researchers note that generated images from the diffusion model have insufficient high-frequency components (Yang et al. 2023). Therefore, spectrum discrepancy has been an intrinsic challenge of the generative models. We visualized examples of these discrepancies in Fig. 1.

To close gaps in the frequency domain, several studies suggest approaches that modify the generative network architecture or objective function providing more guidelines for the real spectrum (Durall, Keuper, and Keuper 2020; Jung and Keuper 2021; Jiang et al. 2021). Although they can restore the distortion in the power spectrum, there is still room for improvement in terms of frequency domain. To overcome these partial success of the conventional methods, we analyze the current generative networks in the frequency domain. We first check the spectrum profiles of the

filters with respect to the kernel size to show the limitation of the conventional spatial domain approaches. We further revisit the denoising process of diffusion models using the Wiener filter-based approach so that the intrinsic problem of diffusion models (*i.e.*, insufficient high-frequencies) in the frequency domain comes out. Inspired by these analyses, we propose a frequency domain framework, **STIG**, that generally reduces the spectral distortion of the generated image. We evaluate the effectiveness of the proposed framework on various generative adversarial networks (GANs). Furthermore, we extend our experiments to another mainstream of image generation, diffusion models (DMs). The proposed framework successfully closes the gaps in the frequency domain while improving the quality of the generated image in the spatial domain as well.

The media forensic community has used spectral features to detect the synthetic image (Jeong et al. 2022; Zhang, Peng, and Long 2018; Marra et al. 2019). Thus, spectral discrepancies can be clues for promising fake image detection not only for the GANs but also DMs (Frank et al. 2020; Ricker et al. 2023). In this study, our framework is validated by FID, log frequency distance of spectrum, and frequency-based fake image detector for both GANs and DMs. In particular, the generated images manipulated by STIG make severe confusion in frequency-based fake image detection. Our contributions can be summarized as follows:

- We analyze the intrinsic limitation of generative models based on the signal processing theorem to explain the spectral discrepancy of the current generative models.

- We propose a novel framework, **STIG**, that effectively reduces the intrinsic discrepancy of the generated images observed in the frequency domain.

- We verify the effectiveness of the proposed framework on various GANs and another mainstream of image generation, DMs.

- We present that between the real and the generated, STIG makes difficult discrimination of the frequency-based detectors.

Related Work

Frequency Discrepancy in Generative Networks

Earlier, Odena *et al*. (2016) noted the checkerboard artifacts in the generated images. They found that the transposed convolution is relevant to these artifacts. Further studies (Frank et al. 2020; Durall, Keuper, and Keuper 2020; Zhang, Karaman, and Chang 2019) showed that the artifacts originate from the high-frequency replica during the upsampling. Especially, Durall *et al*. (2020) presented that the image generated with an interpolation does not include enough high-frequency components. Wang *et al*. (2020) presented various types of spectral discrepancy in the GANs. Recently, researchers explore the spectrum inconsistency of the DMs. Rissanen *et al*. (2022) analyzed the diffusion process in the frequency domain. They stated that the DMs have an inductive bias; during the reverse process, low-frequency components are synthesized first and high-frequencies are added

to it later. Yang *et al*. (2023) has also presented an experiment that the DMs have intrinsic defects in producing high-frequencies.

To improve the quality of the generated images, many studies have made an effort to narrow gaps in the frequency domain. Durall *et al*. (2020) proposed a spectral regularization with the one-dimensional power distribution. Following studies (Jung and Keuper 2021; Chen et al. 2021) proposed a discriminator that enables the network to utilize the power spectral density (PSD). Jiang *et al*. (2021) defined a frequency-level objective function, focal frequency loss (FFL), to increase the flexibility in frequency components. Recently, the frequency domain approach to correct the spectral discrepancy has been reported (Dong, Kumar, and Liu 2022). They refined the magnitude spectrum of the generated image using CycleGAN (Zhu et al. 2017) architecture. Although current SOTA methods improve the quality of image generation and reduce spectral discrepancies, most of them still have problems of aliasing-related artifacts and insufficient high-frequency components. Besides, they basically focused on GANs. So, we propose STIG compatible to both GANs and DMs.

Frequency-Based Deepfake Detection

Due to the social impact of photo-realistic generated images in terms of security and privacy, computer vision and forensic research groups have paid attention to detecting the generated images (Cozzolino and Verdoliva 2020; Matern, Riess, and Stamminger 2019; McCloskey and Albright 2018). In addition to the earlier spatial domain approaches, many studies attempted to utilize spectral discrepancies. The periodic patterns on the spectrum and inconsistency of the spectral distribution can be a good clue for detecting fake images. Early studies boosted the detection performance by employing the magnitude of the spectrum as a detector input (Frank et al. 2020; Zhang, Karaman, and Chang 2019). In particular, Frank *et al*. (2020) pointed out the effectiveness of the frequency-based detector for performance and computational efficiency.

Method

Frequency Domain Refinement

Frequency Analysis of GANs. Basically, ideal upsampling requires an ideal low-pass filter to remove the aliasing. Upsampling with classical interpolation (*e.g*., bi-linear or bi-cubic) effectively suppresses aliasing, however, produces lack of high-frequency components. Although transposed convolution can be commonly adopted in GANs for image details, there is also a limitation that cannot be overlooked. In Fig. 2a, we demonstrate the effect of the kernel size on profile of an ideal low-pass filter from a sinc function. Although all kernels in Fig. 2a represent the low-pass filter, some kernels have severe stop-band ripples, *i.e*., the fluctuation after the cut-off frequency. This observation supports the following limitation of the existing approaches. Conventional CNN kernels commonly used in deep neural networks are difficult to remove high-frequency replicas due to small kernel size. Previous studies for reducing spectral
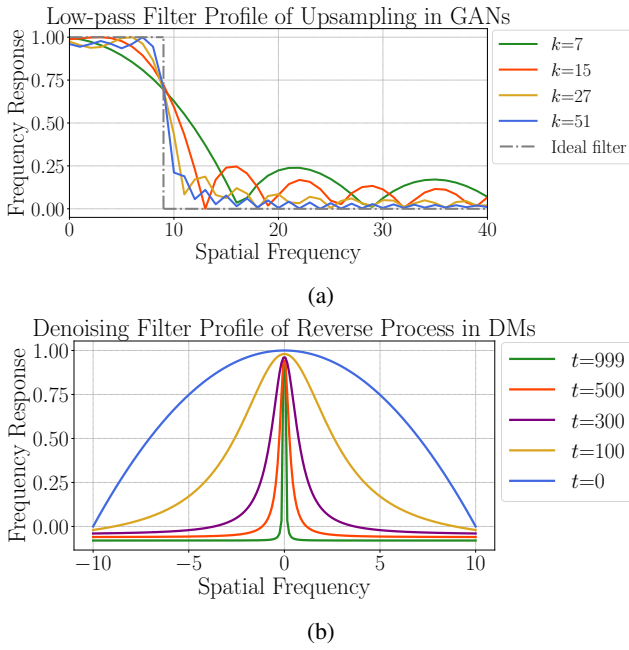
(a)

(b)

Figure 2: (a) Estimation of an ideal low-pass filter by simulation with an example sinc function in the spatial domain. The sinc function with a finite kernel size causes an obvious ripple (*i.e.*, fluctuation pattern) in the cut-off frequency band. (b) Frequency response of the denoising filter for reverse process in diffusion models. The filter still blocks the high-frequency band even if at the end of the reverse process. We notice $t \in [0, 1000]$ in this example.

discrepancies mainly focus on modifying the network architecture and adopting regularization of the training objective in the spatial domain. Their goal can be described as encouraging the network to build the appropriate filters in the frequency domain, *e.g.*, ideal rectified filter. However, kernels with a finite kernel size cannot cut off the aliasing perfectly due to the incomplete shape of the low-pass filter in the frequency domain.

Frequency Analysis of DMs. We further extend the frequency analysis to the diffusion models. We can show an intrinsic reason for spectrum discrepancy in the diffusion model by modification of the Wiener filter-based approach (Yang et al. 2023) for the denoising (reverse) process. First, if we assume the linear denoising Wiener filter $\mathbf{h}_t$, the objective function in the conventional diffusion model (Ho, Jain, and Abbeel 2020) can be rewritten as:

$$\|\boldsymbol{\epsilon} - \gamma \cdot (\mathbf{1} - \mathbf{h}_t) * \mathbf{x}_t\|^2 \quad (1)$$

where $\boldsymbol{\epsilon}$ is the Gaussian noise, $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\gamma = 1/\sqrt{1 - \bar{\alpha}_t}$. $\mathbf{X}_t$ and $\mathbf{H}_t$ represent the frequency response of $\mathbf{x}_t$ and $\mathbf{h}_t$ in diffusion models, respectively. Here, the Wiener filter denoises $\mathbf{x}_t$ by optimizing $\|\sqrt{\bar{\alpha}_t}\mathbf{x}_0 - \mathbf{h}_t * \mathbf{x}_t\|^2$, where $\sqrt{\bar{\alpha}_t}$ is a scaling factor in original DDPM study (Ho, Jain, and Abbeel 2020). From the power law, $\mathbf{X}_0 \approx 1/f^2$ (van der Schaaf and van Hateren 1996), we can define the conjugate

transposed Wiener filter $\mathbf{H}_t^*$ as follows:

$$\mathbf{H}_t^*(f) = \frac{\bar{\alpha}_t}{\bar{\alpha}_t + (1 - \bar{\alpha}_t) \cdot f^2} \quad (2)$$

More detailed expression is included in the supplementary materials. Since $|\mathbf{H}_t^*| = |\mathbf{H}_t|$, we can describe the magnitude profile of Wiener filter $|\mathbf{H}_t|$ in Fig. 2b. In DMs, the denoising filter focuses on composing low-frequency at the beginning of the sampling process and gradually adds high-frequencies as t decreases. We can easily see this in common DMs. However, there is a wideband low-pass filter even if at the end of the reverse process, $t = 0$. Therefore, generated images can suffer from insufficient high-frequency details. Based on the aforementioned frequency domain analyses, we precisely look up the spectrum in the frequency domain rather than in the spatial domain.

STIG via Contrastive Learning

Generative Adversarial Learning. To reduce the spectral discrepancy in the generated images, we first assume the generated and real images to be bilateral distributions. One could recognize that the former has spectral anomalies and the latter is faultless. Then, we can define the spectrum domains $X \subset \mathbb{C}^{H \times W \times C}$ and $Y \subset \mathbb{C}^{H \times W \times C}$ which correspond to generated spectrum and real spectrum, respectively. From the domains, unpaired samples are given, $\mathbf{x} \in X$ and $\mathbf{y} \in Y$. Our goal is to find the mapping function between both domains to refine a generated spectrum $\mathbf{x}$. Therefore, we adopt an adversarial loss (Mao et al. 2017), defined as:

$$\mathcal{L}_{adv}(G, D, X, Y) = \mathbb{E}_{\mathbf{y} \sim Y}(D(\mathbf{y}) - 1)^2 + \mathbb{E}_{\mathbf{x} \sim X}(D(G(\mathbf{x})))^2 \quad (3)$$

where G denotes the mapping function between X and Y; D aims to distinguish $G(\mathbf{x})$ from $\mathbf{y}$.

Patch-Wise Contrastive Learning. When the network decreases the discrepancy in the generated spectrum, it is important to preserve the original frequency components which are not related to the discrepancy. This is a significant issue in our study because a single frequency component affects the overall image contents. To resolve this challenge, we employ contrastive learning (Chen et al. 2020). For each spectrum of images, we can define the frequency patch by extracting the spectral feature during the feedforward process. Since frequency patches in the same location share equivalent spatial frequencies, a patch that has abnormal frequencies (*e.g.*, aliased frequency and insufficient high-frequency) can be recognized as a negative sample in terms of contrastive learning. We define this contrastive tactic (Park et al. 2020; Han et al. 2021) as follows:

$$\mathcal{L}_{pcl}(G, H, X) = \mathbb{E}_{\mathbf{x} \sim X} \sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell_{\text{CE}}(\hat{\mathbf{f}}_l^s, \mathbf{f}_l^s, \mathbf{f}_l^{S \setminus s}) \quad (4)$$

where H denotes a two-layer MLP. When the input spectrum $\mathbf{x}$ passes through the G_{enc} which implies the encoder part of G, frequency patches $\mathbf{f}_l = H(G_{\text{enc}}^l(\mathbf{x}))$ are extracted from l-th internal layer of G_{enc}. Especially, $\mathbf{f}_l^s$ represents s-th patch of $\mathbf{f}_l$ where $S \setminus s$ stands for the set S excluding an element s and S_l denotes the pre-defined number of patches
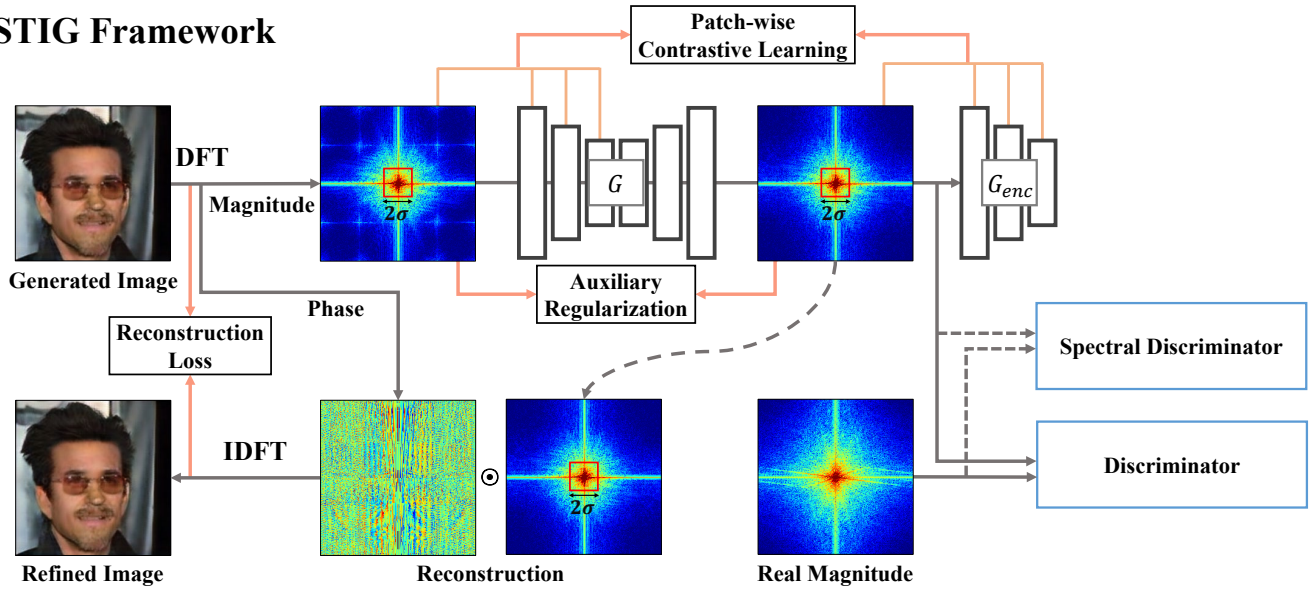
STIG Framework

Figure 3: Framework of our **STIG**. We exploit the magnitude spectrum as an input of our framework using the discrete Fourier Transform to reduce the spectral discrepancies. The spectrum of the generated image is translated into the domain of real spectrum. We, then, obtain the refined generated image by applying the inverse discrete Fourier Transform.

extracted from l-th layer. Similarly, we extract frequency patches of the translated output, $\hat{\mathbf{f}}_l = H(G_{\text{enc}}^l(G(\mathbf{x})))$. The function, $\ell_{\text{CE}}(\cdot)$ in Eq. 4 is a noise contrastive estimation (NCE) framework (Gutmann and Hyvärinen 2010) used for contrastive learning. By minimizing NCE framework, G maximizes mutual information of the input-output pair $(\mathbf{f}_l^s, \hat{\mathbf{f}}_l^s)$; the frequency patch $\hat{\mathbf{f}}_l^s$ pulls a positive sample $\mathbf{f}_l^s$ while pushes negative samples $\mathbf{f}_l^{S\setminus s}$. Therefore, the mapping function G is able to recognize the spectral discrepancies and finely keep the original frequency contents of an input spectrum while reducing the spectral anomaly.

STIG Framework. We illustrate the framework of spectrum translation, STIG, in Fig. 3. To shift the process domain, we first transform an input image into the frequency domain using the discrete Fourier Transform. STIG especially utilizes the magnitude spectrum because it includes most of the energy of spatial frequency and represents the spectral anomaly well. Then, the generator G translates the spectrum of the generated image $\mathbf{x}$ into Y, the domain of real spectrum. In spectrum translation, patch-wise contrastive loss enables G to strictly discriminate the abnormal frequencies in $\mathbf{x}$ and recover it, e.g., erasing the bundle of aliasing or filling insufficient high-frequencies. Meanwhile, we adopt the spatial domain constraint between an input image and a refined image. We implement this using the SSIM reconstruction score (Wang et al. 2004). Therefore, the objective function for spectrum translation is defined as:

$$\mathcal{L}_{trans} = \mathcal{L}_{adv} + \mathcal{L}_{pcl} + \mathcal{L}_{rec} \quad (5)$$

Auxiliary Regularizations. Another key point to close the gap in the frequency domain is matching the distribution of the power spectrum. It is well known that the

power spectral density of natural images follows a power law $1/f^\alpha$, where $\alpha = 2$, approximately (van der Schaaf and van Hateren 1996). To adjust the power spectral density of the generated spectrum, we introduce the chessboard integration which follows Chebyshev (i.e., chessboard) distance (Cantrell 2000) and the spectral discriminator:

$$A_k(F) = \sum_{u=-k}^{k} \sum_{v=-k}^{k} |F(u,v)|, \ k = 0, \cdots, M/2 - 1.$$
$$(6)$$
$$CI_k(F) = A_k(F) - A_{k-1}(F), \ k = 1, \cdots, M/2 - 1.$$
$$(7)$$

where $|F(u,v)|$ denotes the magnitude of the discrete Fourier Transform whose size is $M \times N$. Especially, $CI_0(F)$ is defined as $A_0(F)$ which implies the magnitude of DC frequency. More details are included in the supplementary materials. Since the aliased signal during the upsampling is lying on a Cartesian grid, our chessboard integration easily grabs the aliasing pattern in addition to the power disparity. We use the same adversarial objective in Eq. 3 for training the spectral discriminator D_s, which is formulated as:

$$\mathcal{L}_{spec}(G, D_s, X, Y) = \mathbb{E}_{\mathbf{y}\sim Y}(D_s(CI(\mathbf{y})) - 1)^2 \quad (8)$$
$$+ \mathbb{E}_{\mathbf{x}\sim X}(D_s(CI(G(\mathbf{x}))))^2$$

Here, $CI_k(\cdot) \in CI(\cdot)$.

Basically, the low-frequency part is significant for spectrum translation and reconstruction into the image. Therefore, we exploit the regularization term that constrains the low-frequency distance to maintain the energy level of an input image as follows:

$$\mathcal{L}_{lf} = \sum_{u=-\sigma}^{\sigma} \sum_{v=-\sigma}^{\sigma} (|F(u,v)| - |F_r(u,v)|)^2 \quad (9)$$

Method	Domain	CycleGAN		StarGAN		StarGAN2		StyleGAN	
		FID↓	LFD↓	FID↓	LFD↓	FID↓	LFD↓	FID↓	LFD↓
Original Generated	-	58.75	15.31	192.48	15.43	15.03	16.10	3.59	13.65
Durall *et al.*	S	64.53	16.11	74.00	14.71	21.03	16.00	3.46	13.47
Jung and Keuper	S	33.83	16.60	120.87	15.28	19.01	15.99	14.12	14.56
FFL	S	50.29	15.62	184.84	15.31	17.74	16.08	3.98	13.76
SpectralGAN	F	14.22	14.50	19.96	15.02	7.31	14.44	3.91	12.38
STIG (ours)	F	**6.97**	**12.77**	**6.30**	**12.87**	**3.10**	**12.58**	**2.37**	**12.26**

Table 1: FID and log frequency distance (LFD) of the magnitude spectrum for GANs. The upper three methods stand for the spatial domain method (S), *i.e.*, Durall *et al.* , Jung and Keuper, and FFL. On the other hand, SpectralGAN and STIG stand for the frequency domain method (F). The arrow implies the direction of the best score in FID and LFD, respectively.

Method	DDPM-Face		DDPM-Church		DDIM-Face		DDIM-Church	
	FID↓	LFD↓	FID↓	LFD↓	FID↓	LFD↓	FID↓	LFD↓
Original Generated	43.54	15.28	59.66	15.49	51.49	15.23	55.25	15.48
STIG (ours)	**21.91**	**14.52**	**22.53**	**15.38**	**8.56**	**12.70**	**16.47**	**14.72**

Table 2: FID and log frequency distance (LFD) of the magnitude spectrum for DMs.

where $|F(u, v)|$ and $|F_r(u, v)|$ are an input magnitude and a refined magnitude spectrum, respectively; σ denotes the bandwidth taken to estimate distance. To prevent including artifact components, the bandwidth σ can be configured carefully with the number of upsampling layers in the source network. We typically select $\sigma = 8$ for all experiments in this paper. Finally, the training objective for STIG is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_{trans} + \mathcal{L}_{spec} + \mathcal{L}_{lf} \qquad (10)$$

Experiments

Experimental Setup

Training Details. We evaluate STIG on datasets generated from diverse generative networks, not only GANs but DMs as well. For GANs, we select models of CycleGAN, StarGAN, StarGAN2, and StyleGAN (Zhu et al. 2017; Choi et al. 2018, 2020; Karras, Laine, and Aila 2019) following a setting in (Dong, Kumar, and Liu 2022). We further generate fake images from two popular diffusion models, DDPM and DDIM (Ho, Jain, and Abbeel 2020; Song, Meng, and Ermon 2020). For training the DMs, we especially employ two benchmarks (*i.e.*, FFHQ (Karras, Laine, and Aila 2019) and LSUN Church (Yu et al. 2015)) to investigate the generalization ability of the proposed framework on diffusion models. We set the resolution to 128×128 on DMs and 256×256 on GANs. STIG is composed of the generator G and two kinds of discriminators, D and D_s for spectrum translation. We apply a Nested U-Net (Zhou et al. 2019) architecture to the generator and PatchGAN to the discriminator (Isola et al. 2017). For the spectral discriminator, we employ a simple fully connected layer with the sigmoid activation. More details about implementation are included in the supplementary materials.

Comparison Methods. We compare our STIG with SOTA methods which aim to correct the spectral discrep-

ancy for generative adversarial networks. We first verify our method with spectral regularization methods (Durall, Keuper, and Keuper 2020; Jung and Keuper 2021). Then, the distance-based approach, focal frequency loss (FFL) (Jiang et al. 2021) is applied for another validation. Besides, a frequency domain approach, SpectralGAN (Dong, Kumar, and Liu 2022) is compared as well. For the diffusion models, we verify superiority of our STIG in comparison to the original generated images.

Evaluation Metrics. In this paper, we assess STIG with respect to the spectral discrepancy, image quality, and performance of deepfake detectors. For evaluation of STIG in the frequency domain, we first employ FID (Heusel et al. 2017) to calculate the similarity between the real spectrum and each fake spectrum. Then, we report log frequency distance (LFD) (Jiang et al. 2021) by assessing the logarithm difference in averaged spectra. We convert the magnitude spectrum to be log-scaled and normalize it to [-1, 1]. In the spatial domain, we also verify the refined image from STIG using FID to show the improvement of the proposed method. Finally, we evaluate it with the deepfake detectors, which can show the effect of STIG as well.

Quantitative Analysis

Frequency Domain Results. We evaluate STIG in comparison with other methods that aim to reduce the spectral discrepancy in generated images. We report the FID between the real and fake magnitude spectrum and the LFD of the averaged spectrum in Table 1. It shows that the originally generated spectrum has a severe statistical inconsistency except for StyleGAN. As can be seen, the spatial domain methods mitigate spectral disparity, however, they rather deepen the discrepancy in some cases, *e.g.*, StarGAN2 and StyleGAN. On the other hand, the frequency domain methods successfully reduce the discrepancy in the frequency domain. Although SpectralGAN works well for most benchmarks, it

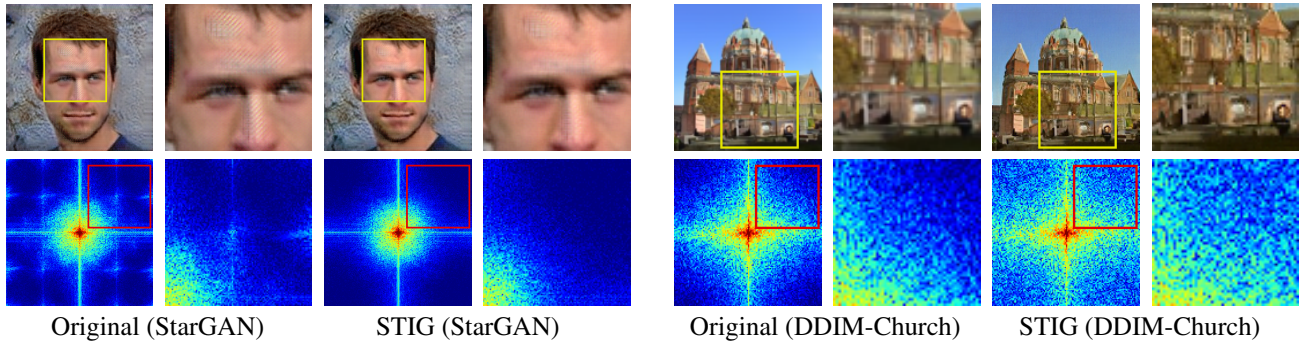| Original (StarGAN) | STIG (StarGAN) | Original (DDIM-Church) | STIG (DDIM-Church) |

Figure 4: STIG examples on StarGAN and DDIM-Church. We magnified the image and corresponding magnitude spectrum relevant to the spectral discrepancies (yellow and red boxed). The lefts indicate the original generated image and the corresponding spectrum. On the other hand, the rights indicate STIG-refined ones.

Type	Benchmark	FID of Image↓	
		w/o STIG	w/ STIG
GANs	CycleGAN	61.69	**59.85**
	StarGAN	26.67	**25.38**
	StarGAN2	13.52	**13.36**
	StyleGAN	20.17	**19.92**
DMs	DDPM-Face	40.19	**26.57**
	DDPM-Church	31.59	**29.35**
	DDIM-Face	36.70	**22.19**
	DDIM-Church	27.62	**22.53**

Table 3: FID of the generated and STIG-refined image for various generative networks.

Loss function	FID of Image↓			
$\mathcal{L}_{lf}/\mathcal{L}_{spec}$	O/O	O/X	X/O	X/X
CycleGAN	**59.85**	60.62	60.70	60.62
StarGAN2	**13.36**	13.41	13.58	13.59
DDPM-Church	**29.35**	29.42	31.15	31.25
DDIM-Face	**22.19**	22.23	25.02	24.90

Table 4: Ablation study for the auxiliary regularization. We compare the performance of STIG under different combinations of auxiliary regularization terms on various benchmarks.

makes worse in StyleGAN, where the disparity is relatively gentle. On the other hand, we show that our method effectively closes the gaps in the frequency domain, outperforming other methods. STIG presents well-generalized performance regardless of the architecture of the generative network. We also provide the frequency domain effect of STIG on diffusion models. Table 2 shows that our method reduces the spectral discrepancy on diffusion models as well even though the source of the disparity is different from the GAN-based models. STIG decreases 67% of FID and 1.04 of LFD on average for diffusion models.

Improvement on Image Quality. From the duality of the discrete Fourier Transform, spectrum translation improves the quality of the generated image as well. The discrepancy in the frequency domain can be recognized in specific ways in the spatial domain, *e.g.*, periodic checkerboard artifact patterns or blurry texture. To evaluate the effect of STIG in the image domain, we calculate the FID from the originally generated images and STIG-refined images with real ones. Table 3 presents the improvements of the generated images, processed with STIG in the frequency domain. STIG enhances the generated images for all benchmarks with diffusion models as well as generative adversarial networks. We emphasize that the proposed framework is especially effective for diffusion models. The diffusion-generated images

are improved by 26.1% in terms of FID on average.

In addition to the quantitative results, we provide the visual effect of STIG by displaying the qualitative outcomes. First, we present a STIG example on the generative adversarial network, StarGAN, on the left side of Fig. 4. There are noticeable checkerboard artifacts (*i.e.*, checkerboard patterns on the human face) in the original generated image which correspond to grid and dot patterns in the spectrum. By adopting STIG, artifacts in the generated image are successfully reduced. The spectral anomalies (*e.g.*, grid and dots) in the spectrum are completely removed as well. Figure 4 also shows the visual example of the diffusion model on the right side. The original generated image from the DDIM-Church benchmark has insufficient high-frequency content. We can see that STIG obviously produces appropriate high-frequency components while maintaining original spatial frequencies. As a result, the reconstructed image from STIG has sophisticated image features (*e.g.*, edges and texture) like a real image. We include more visual examples in the supplementary materials.

Ablation Study

We conducted ablation experiments to analyze the effect of auxiliary loss terms of STIG. Because auxiliary terms are adopted to encourage the proposed framework to preserve fundamental frequencies during spectral translation, we evaluate them in the spatial domain using FID. In Table 4, we present the performance of STIG under different

Method	CycleGAN	StarGAN	StarGAN2	StyleGAN	DDPM-F*	DDPM-C*	DDIM-F*	DDIM-C*	Avg
Original Fake	98.85%	100.00%	99.78%	97.24%	99.73%	99.27%	99.75%	99.39%	99.25%
STIG	52.98%	50.00%	50.14%	53.78%	86.22%	62.03%	49.87%	53.52%	57.32%

Table 5: Detection accuracy of the CNN-based fake image detector in the frequency domain. F* and C* after the name of the diffusion models indicate Face and Church, respectively.

Method	CycleGAN	StarGAN	StarGAN2	StyleGAN	DDPM-F*	DDPM-C*	DDIM-F*	DDIM-C*	Avg
Original Fake	100.00%	100.00%	99.98%	99.89%	99.97%	99.92%	99.97%	99.95%	99.96%
STIG	50.25%	50.00%	50.16%	53.01%	70.55%	57.63%	49.96%	50.46%	54.00%

Table 6: Detection accuracy of the ViT-based fake image detector in the frequency domain.
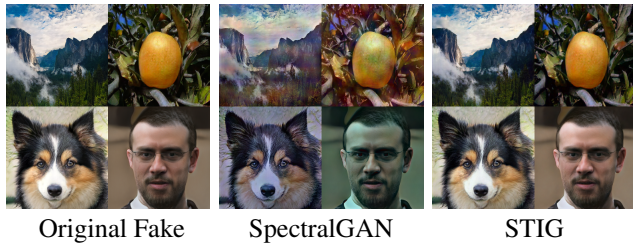

Original Fake   SpectralGAN   STIG

Figure 5: Comparison with a frequency domain method, SpectralGAN (Dong, Kumar, and Liu 2022), for color tone. Examples are sampled from CycleGAN, StarGAN2, and StyleGAN benchmarks.

combinations of auxiliary regularizations. For most benchmarks, the low-frequency loss $\mathcal{L}_{lf}$ improves the visual quality of the refined image. Although the spectral discriminator does not affect the quality by itself, it performs well combined with $\mathcal{L}_{lf}$. Especially, auxiliary terms are effective for diffusion models, *i.e.*, 8.2% decrease of FID on average.

Additionally, we compare STIG with another frequency domain approach, SpectralGAN (Dong, Kumar, and Liu 2022) in Fig. 5. SpectralGAN breaks the color tone of original images. On the other hand, STIG preserves the energy level of color distribution by adopting auxiliary terms. $\mathcal{L}_{lf}$ aims to preserve the low-frequency energy of the original generated image while $\mathcal{L}_{spec}$ is matching the power distribution of the spectrum, *e.g.*, producing the insufficient high-frequencies. Thus, the two auxiliary terms enable more efficient learning.

STIG on Frequency-Based Detector

In this section, we investigate the effect of STIG on the frequency-based fake image detectors. First, we examine the conventional CNN-based frequency domain detector which is composed of shallow convolutional layers (Frank et al. 2020). And then, we also consider the cutting-edge detector which has more network capacity and performs well in various computer vision tasks. We employ the ViT-B16 (Dosovitskiy et al. 2020) for fake image detection as well. We trained and evaluated classifiers with the dataset consisting of real and fake (generated) image pairs. We provide more details for training detectors in the supplementary materials.

Shallow CNN-Based Detector. We provide the detection accuracy of the CNN-based detector in Table 5. The CNN-based detector well discriminates the fake spectrum from real spectrum in the frequency domain even though the detector has a very shallow architecture. The network can make the decision boundary by detecting spectral clues which are only in the spectrum of the fake image. However, the detector can't recognize the refined spectrum from real spectrum except DDPM-Face and DDPM-Church benchmark. The performance drop in average accuracy (*i.e.*, 41.93%) implies that the CNN-based detector can be easily confused by STIG.

ViT-Based Detector. Table 6 shows the detection performance of the ViT-based frequency domain detector. We can see that the ViT-based detector perfectly detects the fake spectrum with an average accuracy of 99.96%. After STIG applies spectrum translation to fake images, the average detection accuracy drops to 54%. STIG makes a fool of the detector even though each detector is trained with generated images that are from diverse generative networks.

Neutralized Frequency-Based Detector. By investigating the effect of STIG on frequency-based detectors, we can show the promising detectors easily make confusion. First, CNN-based and ViT-based detector report perfect detection ability, around 99%, for all benchmarks with the original generated. However, detection performance severely drops when improvement of the spectrum is applied.

Conclusion

In this paper, we propose the frequency domain approach that reduces the spectral discrepancy in the generated image. We generally analyze the intrinsic limitation of generative models in the frequency domain. Our method considerably improves spectral realism and image quality by directly manipulating the frequency components of the generated image. Experimental results on eight fake image benchmarks show that STIG significantly mitigates spectral anomalies. Our method also reports improved image qualities not only in generative adversarial networks but also in diffusion models. From the results with STIG, current frequency-based detectors are not enough to replace image-based detectors.

Acknowledgments

References

Cantrell, C. D. 2000. *Modern Mathematical Methods for Physicists and Engineers*. USA: Cambridge University Press. ISBN 0521598273.

Chen, T.; Kornblith, S.; Norouzi, M.; and Hinton, G. E. 2020. A Simple Framework for Contrastive Learning of Visual Representations.

Chen, Y.; Li, G.; Jin, C.; Liu, S.; and Li, T. 2021. SSD-GAN: Measuring the Realness in the Spatial and Spectral Domains. In *AAAI*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8789–8797.

Choi, Y.; Uh, Y.; Yoo, J.; and Ha, J.-W. 2020. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Cozzolino, D.; and Verdoliva, L. 2020. Noiseprint: A CNN-Based Camera Model Fingerprint. *IEEE Transactions on Information Forensics and Security*, 15: 144–159.

Dong, C.; Kumar, A.; and Liu, E. 2022. Think Twice Before Detecting GAN-generated Fake Images from their Spectral Domain Imprints. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7855–7864.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Durall, R.; Keuper, M.; and Keuper, J. 2020. Watch Your Up-Convolution: CNN Based Generative Deep Neural Networks Are Failing to Reproduce Spectral Distributions. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7887–7896.

Frank, J.; Eisenhofer, T.; Schönherr, L.; Fischer, A.; Kolossa, D.; and Holz, T. 2020. Leveraging frequency analysis for deep fake image recognition. In *International conference on machine learning*, 3247–3258. PMLR.

Gutmann, M.; and Hyvärinen, A. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, 297–304. JMLR Workshop and Conference Proceedings.

Han, J.; Shoeiby, M.; Petersson, L.; and Armin, M. A. 2021. Dual contrastive learning for unsupervised image-to-image translation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 746–755.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems*, volume 33, 6840–6851. Curran Associates, Inc.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. *CVPR*.

Jeong, Y.; Kim, D.; Ro, Y.; and Choi, J. 2022. FrePGAN: Robust Deepfake Detection Using Frequency-Level Perturbations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(1): 1060–1068.

Jiang, L.; Dai, B.; Wu, W.; and Loy, C. C. 2021. Focal Frequency Loss for Image Reconstruction and Synthesis. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 13899–13909.

Jung, S.; and Keuper, M. 2021. Spectral Distribution Aware Image Generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(2): 1734–1742.

Karras, T.; Laine, S.; and Aila, T. 2019. A Style-Based Generator Architecture for Generative Adversarial Networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4396–4405.

Karras, T.; Laine, S.; Aittala, M.; Hellsten, J.; Lehtinen, J.; and Aila, T. 2020. Analyzing and Improving the Image Quality of StyleGAN. In *Proc. CVPR*.

Mao, X.; Li, Q.; Xie, H.; Lau, R. Y.; Wang, Z.; and Smolley, S. P. 2017. Least Squares Generative Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2813–2821.

Marra, F.; Gragnaniello, D.; Verdoliva, L.; and Poggi, G. 2019. Do gans leave artificial fingerprints? In *2019 IEEE conference on multimedia information processing and retrieval (MIPR)*, 506–511. IEEE.

Matern, F.; Riess, C.; and Stamminger, M. 2019. Exploiting Visual Artifacts to Expose Deepfakes and Face Manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, 83–92.

McCloskey, S.; and Albright, M. 2018. Detecting GAN-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*.

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

Odena, A.; Dumoulin, V.; and Olah, C. 2016. Deconvolution and Checkerboard Artifacts. *Distill*.

Park, T.; Efros, A. A.; Zhang, R.; and Zhu, J.-Y. 2020. Contrastive Learning for Unpaired Image-to-Image Translation. In *European Conference on Computer Vision*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Ricker, J.; Damm, S.; Holz, T.; and Fischer, A. 2023. Towards the Detection of Diffusion Model Deepfakes. arXiv:2210.14571.

Rissanen, S.; Heinonen, M.; and Solin, A. 2022. Generative modelling with inverse heat dissipation. *arXiv preprint arXiv:2206.13397*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. *arXiv:2010.02502*.

Song, Y.; Sohl-dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2021. Score-based generative modeling through stochastic differential equations.

van der Schaaf, A.; and van Hateren, J. 1996. Modelling the Power Spectra of Natural Images: Statistics and Information. *Vision Research*, 36(17): 2759–2770.

Wang, S.-Y.; Wang, O.; Zhang, R.; Owens, A.; and Efros, A. A. 2020. CNN-Generated Images Are Surprisingly Easy to Spot. . . for Now. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8692–8701.

Wang, Z.; Bovik, A.; Sheikh, H.; and Simoncelli, E. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612.

Yang, X.; Zhou, D.; Feng, J.; and Wang, X. 2023. Diffusion Probabilistic Model Made Slim. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 22552–22562.

Yu, F.; Zhang, Y.; Song, S.; Seff, A.; and Xiao, J. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365*.

Zhang, L.-B.; Peng, F.; and Long, M. 2018. Face Morphing Detection Using Fourier Spectrum of Sensor Pattern Noise. In *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Zhang, X.; Karaman, S.; and Chang, S.-F. 2019. Detecting and Simulating Artifacts in GAN Fake Images. In *WIFS*.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2019. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE Transactions on Medical Imaging*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2242–2251.