TRAINING-FREE AI-GENERATED IMAGE DETECTION VIA SPECTRAL ARTIFACTS

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid progress of generative models has enabled the synthesis of photorealistic images that are often indistinguishable from real photographs, raising serious concerns about misinformation and malicious use. While most existing AI-generated image (AIGI) detection methods rely on supervised training with labeled synthetic data, they struggle to generalize to unseen generators and incur substantial overhead for retraining. In this work, we propose SpAN, a simple yet effective *training-free* detection framework based on spectral analysis. Our key observation is that upsampling operations in generative models inevitably introduce spectral artifacts, which remain most pronounced at the axial Nyquist frequencies, even when images appear realistic. Building on this insight, we design two techniques to enhance detection reliability: (1) power calibration via azimuthal integration to mitigate bias from image-specific frequency distributions, and (2) autoencoder-based reconstruction to amplify residual artifacts and enable discrepancy-based scoring between original and reconstructed images. Extensive experiments across multiple datasets and generative models demonstrate that SpAN achieves robust and generalizable detection performance. For example, SpAN outperforms other training-free detection methods by a substantial margin (+0.241 AUROC) in the Synthbuster benchmark, which contains recent generative models.

1 INTRODUCTION

Recent advances in generative models, including GANs (Huang et al., 2024) and diffusion models (Wang et al., 2024; Podell et al., 2023; Zhang et al., 2023; Zheng et al., 2023), have enabled the synthesis of highly realistic images that are often indistinguishable from real photographs. These *AI-generated images (AIGIs)* are now widely used for creative content generation (OpenAI, 2024), artistic design (Adobe, 2023), and educational support (Synthesia AI, 2023). However, they also raise serious concerns, such as deepfakes (Samantha Murphy Kelly, 2025), misinformation (Daniel Dale, 2025), and potential misuse in security-sensitive domains (Elizabeth Howcroft, 2025). As a result, reliable *detection of AIGIs* has become an urgent and important research problem.

Most existing AIGI detection methods rely on training-based detectors (Corvi et al., 2023; Karageorgiou et al., 2025; Dzanic et al., 2020; Chandrasegaran et al., 2021), where the detectors trained on a labeled binary classification dataset of real and AI-generated images, *e.g.*, ImageNet (Deng et al., 2009) *vs* Stable Diffusion (Rombach et al., 2022). While these methods have shown effective, they fundamentally suffer from several limitations: (*i*) they often fail to generalize to unseen generators or cross-domain scenarios (Jia et al., 2025), (*ii*), they require to collect AIGIs from diverse generators, and (*iii*) the training-based detectors must be frequently updated to remain effective. All these limitations could be problematic given the rapid development of new generative models.

To address these limitations, researchers have recently explored *training-free* approaches that detect AIGIs without relying on specific generative models or predefined real-image distributions (Ricker et al., 2024; He et al., 2024; Tsai et al., 2024; Brokman et al., 2025). For instance, AEROBLADE (Ricker et al., 2024) leverages reconstruction errors by passing an image through a pretrained autoencoder (*e.g.*, from Stable Diffusion), while RIGID (He et al., 2024) measures robustness to image perturbations in the latent embedding space of self-supervised models such as DINOv2 (Oquab et al., 2023). These approaches demonstrate the feasibility of AIGI detection without training. However, as generative models continue to improve (*e.g.*, producing high-resolution images with accurate

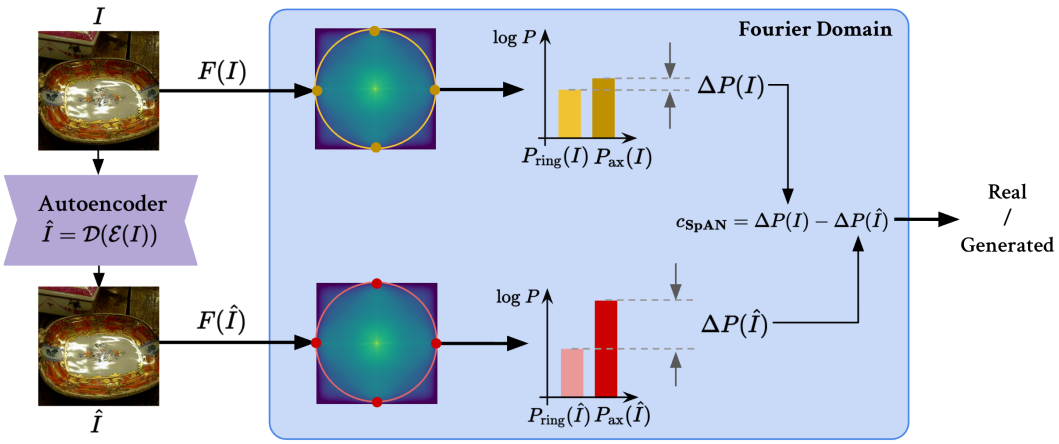


Figure 1: Overview of SpAN, the proposed training-free AIGI detection framework. SpAN first measures the power gap ΔP between P_{ax}, the power at axial Nyquist frequencies, and P_{ring}, the azimuthal integration of the high-frequency power. Then, the spectral discrepancy of ΔP between the original image I and its reconstruction $\hat{I}$ is used for our criterion $c_{\text{SpAN}} = \Delta P(I) - \Delta P(\hat{I})$.

high-level semantics and realistic low-level details), visual cues become increasingly subtle and unreliable. This motivates the following research question: *rather than searching for elusive signals in the image space, can we uncover systematic traces that persist even as image fidelity improves, for example in the Fourier domain?*

Contribution. In this paper, we mainly focus on *spectral artifacts* as a robust detection signal. It is well known that upsampling operations (*e.g.*, transposed convolutions) in generative models introduce *checkerboard patterns* in the Fourier domain (Karageorgiou et al., 2025; Zhang et al., 2019). In particular, the operations induce periodic replications in the power spectrum density, as shown in Figure 2. Although subsequent convolutional layers can reduce them, residual artifacts consistently remain at specific frequency locations. Our key observation is that these artifacts are most pronounced at the axial Nyquist frequencies, *i.e.*, $(\pm 0.5, 0)$ and $(0, \pm 0.5)$, because natural images typically concentrate most of their power near the zero frequency $(0, 0)$.

Based on our observation, we propose **SpAN**, a simple yet effective AIGI detection framework that leverages **Sp**ectral **A**rtifacts at **N**yquist frequencies of AI-generated images. Our key idea is to use the power at the axial Nyquist frequencies as the base criterion for detection. Since this can be biased by image content, we introduce two complementary techniques. First, we calibrate the criterion using azimuthal integration of high-frequency power, which mitigates bias from image-specific frequency distributions. Second, we exploit the discrepancy between the criterion computed on the original image and that on its autoencoder-based reconstruction, where the reconstruction process deliberately introduces artifacts, thereby allowing the original image's spectral characteristics to be assessed relatively. By integrating these steps, our final criterion becomes more robust and reliable. To the best of our knowledge, this work is the first to directly leverage spectral-domain information in the Fourier space as a metric for training-free AIGI detection. The overall framework is illustrated in Figure 1.

Extensive experiments demonstrate that our SpAN achieves state-of-the-art performance across standard AI-generated image detection benchmarks, Synthbuster (Bammey, 2023) and GenImage (Zhu et al., 2023), as reported in Table 1 and 2, respectively. Notably, for high-resolution images (*i.e.*, Synthbuster), our SpAN outperforms the second best baseline by a large margin (**+0.241** AUROC). Furthermore, SpAN exhibits robustness over image corruptions compared to other baselines, as shown in Figure 4. These results highlights that spectral artifacts consistently exist across diverse generative models, even when AI-generated images appear photorealistic. We believe our findings shed light on fundamental properties of generation models and can inspire future advances in the field of AI-generated image detection.

2 PRELIMINARIES

2.1 PROBLEM STATEMENT: TRAINING-FREE AI-GENERATED IMAGE DETECTION

We formulate *AI-generated image (AIGI) detection* as the task of defining a classification criterion that distinguishes between images synthesized by any generative model and real-world images captured from diverse sources (*e.g.*, cameras, digital drawings). Concretely, given an image $I \in \mathcal{I}$, our goal is to design a score function $c : \mathcal{I} \to \mathbb{R}$ that assigns higher values to AI-generated images $\mathcal{D}_{\text{gen}}$ and lower values to real-world images $\mathcal{D}_{\text{real}}$. In the standard evaluation practice, $\mathcal{D}_{\text{real}}$ is sampled from a real dataset such as ImageNet (Deng et al., 2009), and $\mathcal{D}_{\text{gen}}$ is constructed by a generative model, *e.g.*, Stable Diffusion (Rombach et al., 2022).

Most prior works (Karageorgiou et al., 2025; Wang et al., 2020; Tan et al., 2024), adopt a training-based approach, using AI-generated images $\mathcal{D}_{\text{gen}}$ from a specific generative model to learn the score $c(\cdot)$. While these methods have achieved strong detection performance, but they often fails to generalize to unseen generative models. Therefore, we mainly focus on a *training-free* setting where no prior information of $\mathcal{D}_{\text{real}}$ and $\mathcal{D}_{\text{gen}}$ is available in advance, and we aim to design a model-agnostic score $c(\cdot)$ that remains effective across diverse generative models.

2.2 FREQUENCY ANALYSIS OF IMAGES

In computer vision, frequency information provides a complementary perspective to spatial-domain representations, revealing structural patterns such as edges, textures, and periodic artifacts. These characteristics are often more easily captured in the frequency domain, making spectral analysis a powerful tool for image understanding and manipulation. Given an image I of $H \times W$ pixels, its frequency representation can be obtained via the discrete Fourier transform (DFT):

$$F(u, v) = \sum_{x=0}^{W-1} \sum_{y=0}^{H-1} I(x, y) \cdot e^{-i2\pi(\frac{ux}{W} + \frac{vy}{H})},$$

where (u, v) denote frequency coordinates. For convenience, the coordinates are often normalized to the range $[-0.5, 0.5]$. This normalization places the zero frequency at the center of the spectrum, with higher frequencies distributed toward the boundaries.

From the frequency coefficients F, one can compute the *power spectrum density (PSD)* as $P(u, v) = |F(u, v)|^2$, which quantifies the amount of power contained at each frequency. The PSD provides a concise characterization of the distribution of frequency components in the image, enabling analysis of whether most power is concentrated at low frequencies (*e.g.*, smooth variations) or high frequencies (*e.g.*, fine details or noise).

A key concept in spectral analysis is the *Nyquist frequency* f_N, defined as half of the sampling rate along each dimension. After coordinate normalization, this corresponds to the highest representable frequency at $u = \pm f_N$ and $v = \pm f_N$ where $f_N = 0.5$. Frequencies beyond this limit cannot be uniquely represented and are instead folded back into the base spectrum, a phenomenon known as aliasing. Formally, due to the periodicity of DFT, $F(u+1, v) = F(u, v)$ and $F(u, v+1) = F(u, v)$.

3 METHODOLOGY

In this section, we propose **SpAN**, a simple yet effective training-free AIGI detection framework using **Sp**ectral **A**rtifacts at **N**yquist frequencies of AI-generated images. To illustrate our framework, we first describe our observation of spectral artifacts at the axial Nyquist frequencies (Section 3.1). We then suggest a calibration technique for the artifacts to consider the amount of high-frequency information (Section 3.2). Finally, we design our detection criterion based on the spectral discrepancy between an input image and its reconstruction (Section 3.3). The overall framework is illustrated in Figure 1.

3.1 SPECTRAL ARTIFACTS AT AXIAL NYQUIST FREQUENCIES

We begin by describing our key observation on the spectral artifacts exhibited by generative models. It is widely known that a common artifact in images synthesized by generative models is the

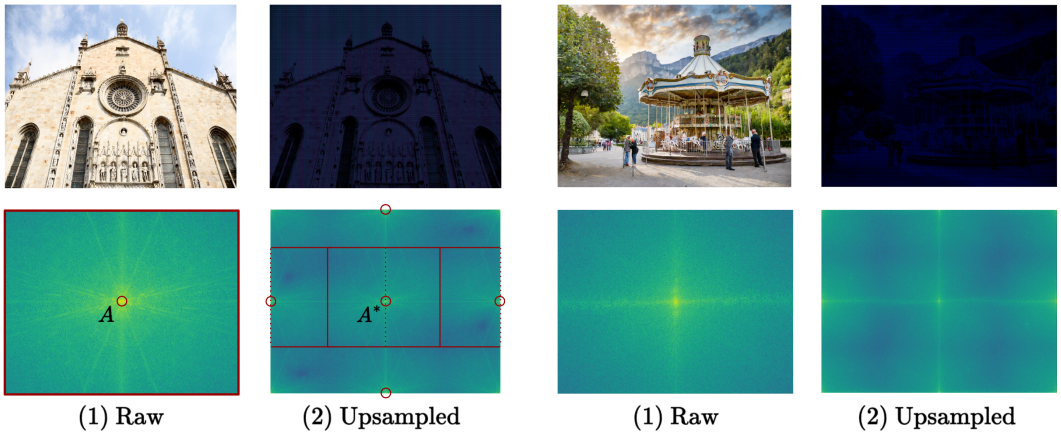


Figure 2: Visualization of the power spectrum density (PSD) of real-world and generated images, each sampled from the RAISE-1K (Dang-Nguyen et al., 2015) and Firefly (Adobe, 2023) in the Synthbuster (Bammey, 2023) benchmark, respectively. The images at the first row correspond to the raw image and its upsampled image via single transposed convolution. The second row correspond to the PSD of the image in the same column.

appearance of *checkerboard patterns* in the frequency domain (Karageorgiou et al., 2025). Prior work has shown that these artifacts arise from the use of transposed convolutions of stride 2, where zeros are inserted in a "bed-of-nails" fashion during upsampling. This operation induces a periodic replication in the power spectrum density (PSD), as formally proven by Zhang et al. (2019).

Even when convolutional layers are subsequently applied, these artifacts do not fully vanish, especially at specific frequency regions. We find that the artifacts are most clearly preserved at the *axial Nyquist frequencies*, *i.e.*, $(\pm f_N, 0)$ and $(0, \pm f_N)$. This can be attributed to the fact that natural images typically exhibit their highest power near the zero frequency $(0, 0)$, and thus such upsampling operations also leads to relatively high power concentrated at the axial Nyquist frequencies.

For example, consider a real image $I \in \mathbb{R}^{C\times H\times W}$ and its PSD shown in Figure 2a. After upsampling I to another image $I_{\text{up}} \in \mathbb{R}^{C\times 2H\times 2W}$ by a single transposed convolution as shown in Figure 2a(2), in the PSD of I_{up}, the midpoint of each edge of the spectrum acquires a significant power, the power of which is widely deviated from its adjacent region. In particular, the power of point A (*i.e.*, $P(0, 0)$) in the original image I is conveyed not only to the corresponding point A^* in the upsampled image I_{up}, but also to midpoints of each side edge (*e.g.*, $P(-f_N, 0), P(f_N, 0)$), due to the periodic replication caused by the transposed convolution. Although the generated image in Figure 2b(1) has exhibits fewer checkboard artifacts as it is generated through multiple convolutional layers, significant power still remains along the central axis of the spectrum, including the axial Nyquist frequencies.

From this observation, one can expect that the power at the axial Nyquist frequencies is high for AI-generated images due to the use of transposed convolutions, while real images have a low power at the frequencies. Motivated by this, we propose to use the power as a simple criterion for AIGI detection, formally defined as:

$$P_{\text{ax}}(I) = \frac{1}{4} \sum_{(u,v)\in\{(\pm f_N,0),(0\pm f_N)\}} P(u, v),$$

where $P(u, v)$ denotes the PSD at frequency (u, v) of the image I. In practice, we simply compute the average of nearest points of the axial Nyquist frequencies in the discrete PSD.

3.2 Power Calibration via Azimuthal Integration

The distribution of frequency components may vary across images depending on their content, resulting in different amounts of high-frequency information and dominant frequency directions. Con-

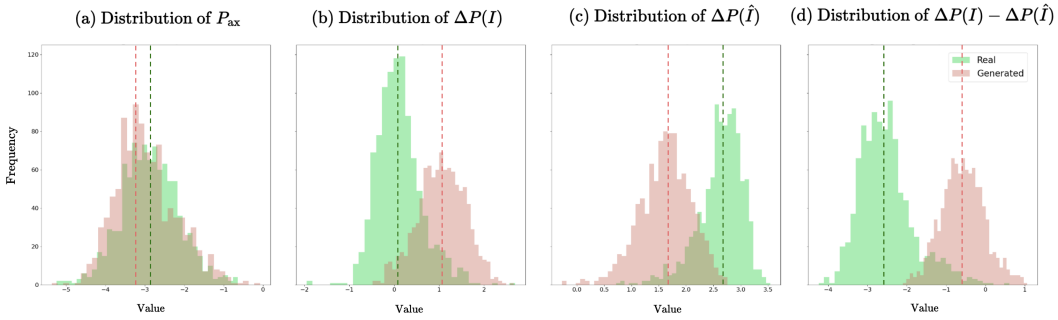


Figure 3: The distribution of $P_{\text{ax}}(I)$, $\Delta P(I)$, $\Delta P(\hat{I})$ and $\Delta P(I) - \Delta P(\hat{I})$. For visualization, 1000 images are sampled from Midjourney and ImageNet in the GenImage benchmark, respectively. The vertical dashed-line denotes the mean value of each distribution.

sequently, the power at the axial Nyquist frequencies $P_{\text{ax}}(I)$ may be biased by the amount of high-frequency information. To calibrate this, we normalize $P_{\text{ax}}(I)$ using the azimuthal integration of power at the same frequency magnitude as follows:

$$\Delta P(I) = \log P_{\text{ax}}(I) - \log P_{\text{ring}}(I) \quad \text{where} \quad P_{\text{ring}}(I) = \frac{1}{2\pi} \int_0^{2\pi} P(f_N \cos\phi, f_N \sin\phi) d\phi.$$

To examine the efficacy of this calibration technique, we calcuate $P_{\text{ax}}(I)$ and $\Delta P(I)$ for 1000 ImageNet (Deng et al., 2009) images and 1000 Midjournery-generated (Midjourney Inc., 2023) images, and visualize their distributions in Figure 3a and 3b, respectively. Although $P_{\text{ax}}(I)$, the power at the Nyquist frequencies, is not a sufficient detection metric (see Figure 3a), its calibrated version ΔP provides much stronger discriminative power. These results highlight the importance of assessing how strongly certain artifacts appear relative to the overall frequency distribution, rather than relying solely on absolute power values.

In practice, the azimuthal integration is approximated by averaging the power over all frequency points that fall within a ring of width δ around the target magnitude as follows:

$$P_{\text{ring}}(I) = \frac{1}{|\mathcal{R}(f_N, \delta)|} \sum_{(u,v) \in \mathcal{R}(f_N, \delta)} P(u, v),$$

where $\mathcal{R}(r, \delta) = \{(u, v) : r - \delta \leq \sqrt{u^2 + v^2} < r\}$ denotes the set of frequency points that fall within the ring of radius r and width δ.

3.3 Spectral Artifact Detection with Reconstruction

The calibrated power ΔP introduced in Section 3.2 may not be sufficient as a criterion when artifacts are relatively weak, such as in low-resolution generated images. To further enhance detection capability, we exploit the difference between an original image and its autoencoder-based reconstruction. The key idea is that the reconstruction can be regarded as an AI-generated image, since the autoencoder inevitably performs upsampling operations (*e.g.*, transposed convolutions) that generate grid-aligned spectral artifacts. Therefore, if the spectral discrepancy between the original and reconstructed images is large, the original is likely a real-world image because real images often have less artifacts; otherwise, it is likely to have been generated by a generative model.

Based on this intuition, we propose to use the discrepancy in the calibrated power ΔP between an original image I and its reconstruction $\hat{I} = \mathcal{D}(\mathcal{E}(I))$, where $\mathcal{E}$ and $\mathcal{D}$ are the encoder and decoder of an autoencoder, respectively. Formally, our final detection criterion $c_{\text{SpAN}}(\cdot)$ is defined as follows:

$$\begin{aligned} c_{\text{SpAN}}(I) &= \Delta P(I) - \Delta P(\hat{I}) \\ &= \Big(\log P_{\text{ax}}(I) - \log P_{\text{ring}}(I)\Big) - \Big(\log P_{\text{ax}}(\hat{I}) - \log P_{\text{ring}}(\hat{I})\Big). \end{aligned}$$

This criterion is particularly effective for high-resolution images, which tend to exhibit stronger spectral artifacts due to multiple upsampling operations in a generation process. For low-resolution

images, we upsample the image I while preserving its aspect ratio before feeding it into the autoencoder. This step ensures that the reconstruction process induces sufficient spectral artifacts, thereby making our discrepancy-based score a more reliable detection signal.

We further examine the effectiveness of this reconstruction-based technique by visualizing the distributions of $\Delta P(I)$, $\Delta P(\hat{I})$, and $\Delta P(I) - \Delta P(\hat{I})$ in Figure 3b-d, respectively. For generated images, comparing the calibrated power of the original image $\Delta P(I)$ with that of the reconstructed image $\Delta P(\hat{I})$ shows little difference (*e.g.*, $\Delta P(I) \approx 1 \rightarrow \Delta P(\hat{I}) \approx 1.6$ in average), since artifacts are already present in the original. In contrast, for real images, new artifacts are introduced during reconstruction, leading to a significant increase (*e.g.*, $\Delta P(I) \approx 0 \rightarrow \Delta P(\hat{I}) \approx 2.5$ in average). Consequently, when examining the distribution of our final score $c_{\text{SpAN}} = \Delta P(I) - \Delta P(\hat{I})$, we observe substantially improved discriminative power.

4 Experiment

We design our experiments to validate the followings:

- Does our metric achieve strong performance in diverse AIGI detection tasks? (§4.1)
- How much does each component contribute to overall performance ? (§4.2)
- Is our method robust over corruptions on the raw images ? (§4.3)

Evaluation Benchmark. We conducted evaluations on two widely used benchmarks in the field of AI-generated image detection: Synthbuster (Bammey, 2023) and GenImage (Zhu et al., 2023). The Synthbuster benchmark is composed of high-resolution images generated from 9 recent diffusion models, including commercial models, such as Firefly (Adobe, 2023), Midjourney (Midjourney Inc., 2023), DALL-E 2 (Ramesh et al., 2022), DALL-E 3 (Ramesh et al., 2023), and Stable Diffusion (Rombach et al., 2022), and real images are come from the subset of the Raise-1k dataset (Dang-Nguyen et al., 2015), which contains up to 4K resolution (that is, 3840×2160) images. The GenImage benchmark contains relatively low-resolution images from 8 different generators. It includes images generated from GAN (Brock et al., 2018) and diffusion models where resolution ranges from 128×128 to 1024×1024. The detection performance is measured by the area under ROC curve (AUROC).

Implementation Details. For implementing our method, the ring width is set to $\delta = 0.01$, and SDv1.4 is used for the autoencoder. For amplifying artifacts, we increased the image resolution by doubling its size until the smaller side of the image becomes at least 1024 pixels, preserving the original aspect ratio. For a complete evaluation on Synthbuster and GenImage benchmarks, we used 2 NVIDIA RTX 4090 GPUs, each taking 3.5 and 40 hours, respectively.

Baselines. We compare our performance with recent training-free AIGI detection methods, RIGID He et al. (2024), MINDER Tsai et al. (2024), AEROBLADE Ricker et al. (2024), and Manifold Bias Brokman et al. (2025). To reproduce the result of AEROBLADE we used SDv1.4 as the autoencoder, identical to our selection of the autoencoder. For Manifold Bias, we follow the official code provided by the authors that applies SDv2 as the latent diffusion model. Finally, we follow the original setting of the authors, where the VIT-L14 version of the DINOv2 is applied as the feature extractor. In case of RIGID and MINDER, we used the thresholds and pretrained model, DINOv2 Oquab et al. (2023) as indicated in the paper.

4.1 Main Result

Tables 1 and 2 are the evaluation results of our method and beselines on the Synthbuster and GenImage benchmarks. At the Synthbuster benchmark, our method exhibits the best performance among 4 other baselines achieving **+0.241** AUROC at than the second best performing method, AEROBLADE. While AEROBLADE performs competitive where the generative model matches the inspected autoencoder, it fails to generalize on the proprietary models. Especially, we observe a notable gap between SpAN and the baselines on detecting recent AI-generated images, supporting the generalizability of SpAN.

Table 1: AI-generated image detection performance (AUROC) in the Synthbuster benchmark (Bammey, 2023). We denote **bold**, and underline as the best method and second best method.

Method	Firefly	GLIDE	SDXL	SDv2	SDv1.3	SDv1.4	DALL-E 3	DALL-E 2	Midjourney	Mean
RIGID	0.519	0.868	0.757	0.615	0.448	0.446	0.442	0.596	0.593	0.587
MINDER	0.440	0.568	0.472	0.721	0.656	0.668	0.346	0.445	0.345	0.518
AEROBLADE	0.592	**0.954**	0.668	0.567	0.950	0.950	0.486	0.392	0.769	0.703
Manifold Bias	0.493	0.779	0.562	0.749	0.544	0.549	0.379	0.607	0.424	0.565
SpAN (ours)	**0.945**	0.893	**0.988**	**0.948**	**0.994**	**0.994**	**0.948**	**0.795**	**0.989**	**0.944**

Table 2: AI-generated image detection performance (AUROC) of proposed method and baselines in the GenImage Zhu et al. (2023) benchmark. We denote **bold** and underline as the best method and second-best method.

Method	ADM	BigGAN	GLIDE	Midjourney	SDv1.4	SDv1.5	VQDM	Wukong	Mean
RIGID	**0.874**	0.974	0.952	0.778	0.682	0.682	**0.915**	0.699	0.820
MINDER	0.768	0.681	0.582	0.450	0.607	0.596	0.882	0.676	0.655
AEROBLADE	0.856	**0.981**	**0.989**	0.918	**0.982**	**0.984**	0.732	**0.983**	0.928
Manifold Bias	0.727	0.925	0.852	0.510	0.675	0.673	0.874	0.653	0.736
SpAN (ours)	0.791	0.957	0.935	**0.975**	0.975	0.977	0.857	0.973	**0.930**

4.2 Ablation Study

Component Ablation Study. We observed the contribution of each component of our score c_{SpAN}, by sequentially adding each component suggested from §3.1 to §3.3. As shown in Table 3, purely using the averaged power at axial Nyquist points yield less discriminative result, because the absolute value may vary within both generated images and real-world images. However, by calibrating P_{ax} by P_{ring}, we could achieve significant increase in performance, even without using any reconstruction process. As visualized in Figure 3b,d and indicated in the third row of the Table 3, subtracting $\Delta P(\hat{I})$ widens the gap, initially observed at the distribution of ΔP (see Figure 3d). This may be attributed to the fact that the reconstruction process 'cancels out' the artifacts of the original image, shifting the overall distribution of real-world images to the negative direction of the axis. Finally, by applying upsampling to original images before reconstruction, the value of $\Delta P(\hat{I}$ is intensified in case of real-world images, resulting in the best performance.

Choice of parameter. We also performed ablation on the parameter or architecture design to demonstrate that our method is not overly dependent on specific conditions. The width of the ring δ at P_{ring} designates the broadness of a region that is used for calibration. We tracked the difference in the evaluation metric while increasing δ in the power of 2 from 0.01. As reported in Table 4, AUROC is preserved within the gap of 0.001 until $\delta = 0.08$, indicating the consistency of our method to size of the adjacent calibration region. Note that decreasing δ less than 0.01 makes it unavailable to define P_{ring}, as the width of the ring becomes too small for frequency points to fall within the region.

Table 4: Ablation study on ring width δ at the Synthbuster benchmark. The best result is denoted in **bold**.

$\delta = 0.16$	0.08	0.04	0.02	0.01	0.005
0.934	0.943	0.943	0.943	**0.944**	N/A

Choice of Autoencder. When selecting autoencoder, we trailed on 3 different autoencoders including SDv1.4, SDv2 (Rombach et al., 2022), and Kandinsky v2.1 (Arseniy Shakhmatov, 2023). Although SD v1.4 gives the best performance, replacing with other autoencoders still outperformed other baseline models, which supports invariance of our method to the choice of a specific model. This suggests that our method can benefit from common autoencoders which exhibit artifacts during upsampling in the generation process.

Table 5: Ablation on autoencoder variants at the Synthbuster benchmark. The best result is in **bold**.

KD v2.1	SD v2	SD v1.4
0.857	0.860	**0.944**

4.3 Robustness to Corruptions

For practical deployment in real-world scenarios, AIGI detectors must remain robust when applied to web-collected images that may undergo various perturbations such as JPEG compres-

Table 3: Component ablation study of our method on the GenImage benchmark. For component analysis, 1,000 images are sampled per each generative model with its corresponding real images. 'Ups.' denotes the upsampling process before reconstruction. '✓' and '✗' denotes that the component is used, and not used respectively.

Component	P_{ax}	ΔP	$\Delta P - \Delta\hat{P}$	AUROC
Nyquist Frequency (§3.1)	✓	✗	✗	0.573
+ Power Calibration (§3.2)	✓	✓	✗	0.849
+ Autoencoder-based reconstruction (§3.3)	✓	✓	✓	**0.930**

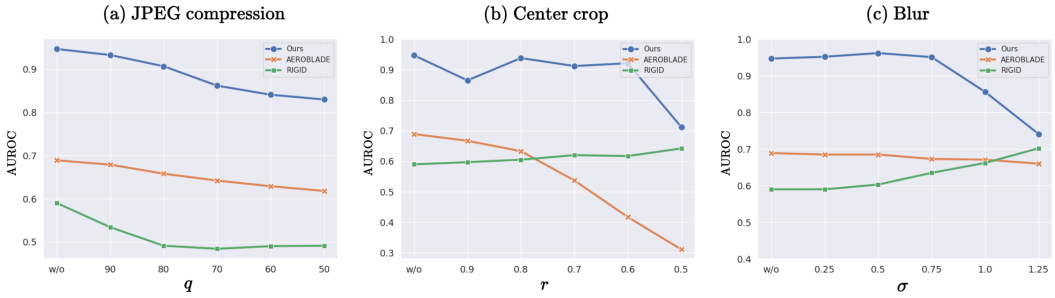


Figure 4: Robustness to image corruptions of our method and baselines. For each corruption from (a) to (c), images are compressed by quality q, cropped by ratio r and resized back to its original size, or blurred by standard deviation σ. Our method shows consistent superior result over other 2 baselines, validating its robustness to image perturbations.

sion. To evaluate this, we further assess the performance of SpAN on both real and AI-generated images under such perturbations. Specifically, we sample 500 real images from the Raise-1k dataset (Dang-Nguyen et al., 2015) and 500 generated images from each model in the Synthbuster benchmark (Bammey, 2023). We then test three types of perturbations: JPEG compression, cropping and resizing, and Gaussian blurring, following (Ricker et al., 2024; Frank et al., 2020). The results of SpAN and the baselines are presented in Figure 4. As shown, SpAN maintains strong robustness and consistently outperforms the baselines even under the most severe perturbation conditions.

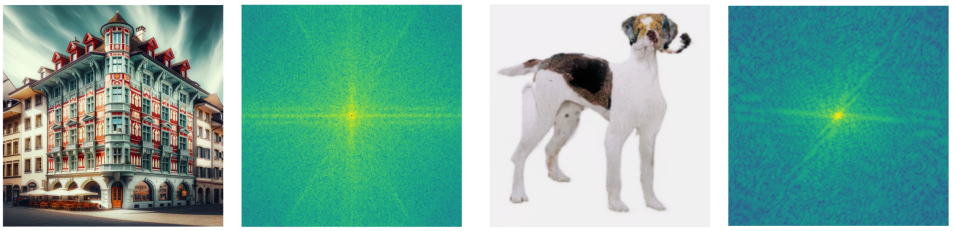
4.4 Case Study

In this subsection, we show specific cases of how our method can behave according to the characteristics of the generated images in the Fourier domain. Figure 5 shows a generated image from DALL-E3 (Ramesh et al., 2023) and ADM (Dhariwal & Nichol, 2021), and its counterpart converted by the discrete Fourier transform, respectively. For 5a, which is comprised of high-frequency details, our method can behave better by capturing the artifact from the raw state of the generative image, resulting in relatively high c_{SpAN}. This is mainly since $\Delta P(I)$ is big enough to cancel the effect of subtracting $\Delta P(\hat{I})$. In contrast, an over-blurry image, such as in Figure 5b, can unintentionally mimic the distribution of a real-world image in the Fourier domain, which paradoxically becomes relatively difficult to detect. However, considering that recent generative models are becoming closer to real-world images imitating high-frequency details, this property can become an advantage in the near future. Also, compared to the previous reconstruction-based model, which has the assumption that generated images are harder to reconstruct, our method has the strength to handle high complexity images by observing the artifact of the upsampling process.

5 Related Works

5.1 Training-free AI-generated image detection

To address the rapid proliferation of generative models, training-free detection methods which do not require AIGIs for training have recently emerged. Most existing approahces leverage the pre-



(a) DALL-E 3 (b) ADM

Figure 5: A set of generated image and its power spectrum density map each sampled from DALL-E3 in the Synthbuster benchmark, (Ramesh et al., 2023) and ADM (Dhariwal & Nichol, 2021) in the GenImage benchmark.

trained representations of large foundation models (*e.g.*, DINOv2 (Oquab et al., 2023)) for detection. For instance, Ricker et al. (2024) measures the perceptual distance between an original image and its reconstruction by the Latent Diffusion Model (LDM) autoencoder, based on the observation that images generated by LDMs exhibit lower reconstruction error when evaluated by the corresponding LDM. On the other hand, He et al. (2024) and Tsai et al. (2024) exploit the robustness of self-supervised vision foundation models to perturbations like Gaussian noise or blurring, under the hypothesis that real images are inherently more robust to such distortions. Brokman et al. (2025) assumes that real data are more likely to reside on the latent-space manifold of the LDM. In contrast to these approaches, which primarily depend on predictions from pre-trained models, we demonstrate that image-specific frequency information remains highly effective for detecting AIGIs in a training-free regime.

5.2 AI-GENERATED IMAGE DETECTION VIA FREQUENCY ANALYSIS

Several training-based AIGI methods have leveraged frequency information as the key representations (Li et al., 2024; Dzanic et al., 2020; Chandrasegaran et al., 2021) Durall et al. (2020) pointed out the spectral distortion in the images generated from the CNN-based model and utilized the gap to detect deep-fake images. Frank et al. (2020) investigates the artifacts in GAN-generated images in the frequency domain by applying the discrete cosine transform (DCT), and indicates the artifact as a result of upsampling techniques. Another approach is to learn a deepfake detector with a perturbation generator as in Jeong et al. (2022). Karageorgiou et al. (2025) employs masked spectral learning to learn the spectral distribution of real images, considering generated images as out-of-distribution samples. Although analysis based on the Fourier domain has been used as a distinctive factor for discriminating generated images, this difference has not been utilized in a training-free setting. We object to modeling this difference by observing the artifacts that generated images reveal when transformed into the Fourier domain.

6 CONCLUSION

In this work, we propose SpAN, a simple yet effective training-free AIGI detection method inspired by the spectral artifacts of generated images observed in the Fourier domain. By comparing the energy gap near the axial Nyquist frequency before and after image reconstruction, we could robustly discriminate AI-generated images from real-world images. Extensive experiments demonstrate the effectiveness of our framework across diverse benchmarks and types of generative models, as well as its robustness to image perturbations. We hope that our research will be expanded to exploit other artifacts residing in generated images in the training-free setting of AIGI detection.

REFERENCES

Adobe. Adobe firefly. https://www.adobe.com/sensei/generative-ai/firefly.html, 2023. Generative AI model, accessed 2025-09-20.

Aleksandr Nikolich Arseniy Shakhmatov, Anton Razzhigaev. kandinsky 2.1, 2023.

Quentin Bammey. Synthbuster: Towards detection of diffusion model generated images. *IEEE Open Journal of Signal Processing*, 5:1–9, 2023.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

Jonathan Brokman, Amit Giloni, Omer Hofman, Roman Vainshtein, Hisashi Kojima, and Guy Gilboa. Manifold induced biases for zero-shot and few-shot detection of generated images. *arXiv preprint arXiv:2504.15470*, 2025.

Keshigeyan Chandrasegaran, Ngoc-Trung Tran, and Ngai-Man 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*, pp. 7200–7209, 2021.

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

Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. Raise: A raw images dataset for digital image forensics. In *Proceedings of the 6th ACM multimedia systems conference*, pp. 219–224, 2015.

Daniel Dale. Fact check: The fake photos, false claims and wild conspiracy theories swirling around the murder of charlie kirk. https://edition.cnn.com/2025/09/20/politics/fact-check-charlie-kirk-murder, 2025. News Article, accessed 2025-09-20.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Ricard Durall, Margret Keuper, and Janis Keuper. Watch your up-convolution: Cnn based generative deep neural networks are failing to reproduce spectral distributions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7890–7899, 2020.

Tarik Dzanic, Karan Shah, and Freddie Witherden. Fourier spectrum discrepancies in deep network generated images. *Advances in neural information processing systems*, 33:3022–3032, 2020.

Elizabeth Howcroft. Ai-generated content raises risks of more bank runs, uk study shows. https://www.reuters.com/technology/artificial-intelligence/, 2025. acccessed: 2025-04-13.

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

Zhiyuan He, Pin-Yu Chen, and Tsung-Yi Ho. Rigid: A training-free and model-agnostic framework for robust ai-generated image detection. *arXiv preprint arXiv:2405.20112*, 2024.

Nick Huang, Aaron Gokaslan, Volodymyr Kuleshov, and James Tompkin. The gan is dead; long live the gan! a modern gan baseline. *Advances in Neural Information Processing Systems*, 37: 44177–44215, 2024.

Yonghyun Jeong, Doyeon Kim, Youngmin Ro, and Jongwon Choi. Frepgan: robust deepfake detection using frequency-level perturbations. In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 1060–1068, 2022.

Zexi Jia, Chuanwei Huang, Yeshuang Zhu, Hongyan Fei, Xiaoyue Duan, Zhiqiang Yuan, Ying Deng, Jiapei Zhang, Jinchao Zhang, and Jie Zhou. Secret lies in color: Enhancing ai-generated images detection with color distribution analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 13445–13454, 2025.

Dimitrios Karageorgiou, Symeon Papadopoulos, Ioannis Kompatsiaris, and Efstratios Gavves. Any-resolution ai-generated image detection by spectral learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18706–18717, 2025.

Yanhao Li, Quentin Bammey, Marina Gardella, Tina Nikoukhah, Jean-Michel Morel, Miguel Colom, and Rafael Grompone 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*, pp. 3855–3865, 2024.

Midjourney Inc. Midjourney. https://www.midjourney.com/, 2023. Generative AI model, accessed 2025-09-20.

OpenAI. Sora. https://openai.com/index/sora/, 2024. Generative AI Model, accessed 2025-09-20.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. URL https://arxiv.org/abs/2204.06125.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Pamela Mishkin, Brooks Chan, Christopher Hesse, Alec Radford, and Ilya Sutskever. Dall·e 3, 2023. URL https://cdn.openai.com/papers/dall-e-3.pdf. OpenAI Technical Report.

Jonas Ricker, Denis Lukovnikov, and Asja Fischer. Aeroblade: Training-free detection of latent diffusion images using autoencoder reconstruction error. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9130–9140, 2024.

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

Samantha Murphy Kelly. Hollywood celebrities are being deepfaked into porn. https://edition.cnn.com/2025/03/08/tech/hollywood-celebrity-deepfakes-congress-law, 2025. News Article, accessed 2025-04-13.

Synthesia AI. Synthesia. https://www.synthesia.io/, 2023. Large Language Model, accessed 2025-09-20.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Chung-Ting Tsai, Ching-Yun Ko, I Chung, Yu-Chiang Frank Wang, Pin-Yu Chen, et al. Understanding and improving training-free ai-generated image detections with vision foundation models. *arXiv preprint arXiv:2411.19117*, 2024.

Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei 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*, pp. 8695–8704, 2020.

Yufei Wang, Wenhan Yang, Xinyuan Chen, Yaohui Wang, Lanqing Guo, Lap-Pui Chau, Ziwei Liu, Yu Qiao, Alex C Kot, and Bihan Wen. Sinsr: diffusion-based image super-resolution in a single step. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 25796–25805, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3836–3847, 2023.

Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *2019 IEEE international workshop on information forensics and security (WIFS)*, pp. 1–6. IEEE, 2019.

Kaiwen Zheng, Cheng Lu, Jianfei Chen, and Jun Zhu. Dpm-solver-v3: Improved diffusion ode solver with empirical model statistics. *Advances in Neural Information Processing Systems*, 36: 55502–55542, 2023.

Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image. *Advances in Neural Information Processing Systems*, 36:77771–77782, 2023.