Your diffusion model is an implicit synthetic image detector

Xi Wang, Vicky Kalogeiton

Your diffusion model is an implicit synthetic image detector

Xi Wang and Vicky Kalogeiton

LIX, CNRS, Ecole Polytechnique, IP Paris
{xi.wang,vicky.kalogeiton}@polytechnique.edu

Abstract. Recent developments in diffusion models, particularly with latent diffusion and classifier-free guidance, have produced highly realistic images that can deceive humans. In the detection domain, the need for generalization across diverse generative models has led many to rely on frequency fingerprints or traces for identifying synthetic images therefore often compromising the robustness against complex image degradations. In this paper, we propose a novel approach that does not rely on frequency or direct image-based features. Instead, we leverage pre-trained diffusion models and a sampling technique to detect fake images. Our methodology is based on two key insights: (i) pre-trained diffusion models already contain rich information about the real data distribution, enabling the differentiation between real and fake images through strategic sampling; (ii) the dependency of textual conditional diffusion models on classifier-free guidance, coupled with higher guidance weights, enforces the discernibility between real and diffusion generated fake images. We evaluate our method across the GenImage dataset, with eight distinct image generators and various image degradations. Our method demonstrates its efficacy and robustness in detecting multiple types of AI-generated synthetic images, setting the new state of the art. Code is available on our project page[1]

1 Introduction

The evolution of generative AI models has enabled the production of images that can deceive humans, raising potential legal and ethical concerns. Consequently, there is a pressing need for enhanced detection techniques to match the pace of advancements in image generation. Notably, recent progress has been driven by diffusion techniques, in contrast to their predecessors, such as Generative Adversarial Networks (GANs) [12, 29]. Diffusion models aim to learn the data distribution by adding iterative noise to images and subsequently learning to denoise them. This approach effectively addresses the mode collapse issue prevalent in GAN-based methods. The introduction of classifier-guidance [9] and later, classifier-free guidance [13] has further enhanced these models by allowing for complex conditioning and image quality improvement through tuning required hyperparameter guidance weights ω. The Latent Diffusion Model (LDM) [27]

[1] https://www.lix.polytechnique.fr/vista/projects/2024_detector_wang.

represents a significant advancement, enabling the generation of high-resolution images at increased speeds by performing diffusion in the latent space of a pre-trained VAE, a technique foundational to the popular Stable Diffusion series.

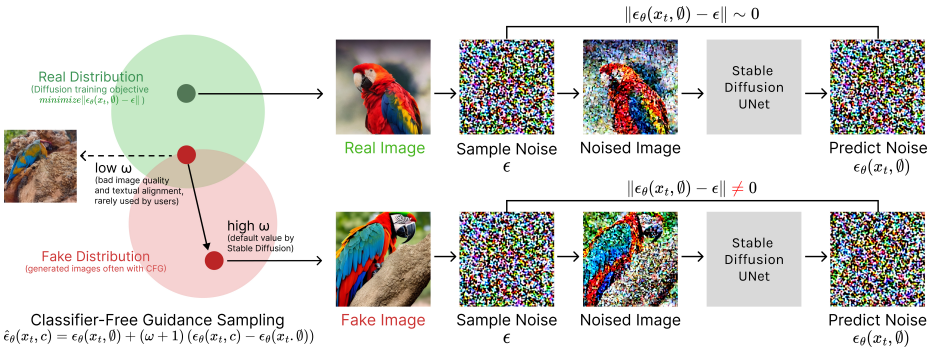


Fig. 1: Our main intuition is twofold: (i) since the diffusion model is trained solely with real images towards the objective of $\|\epsilon_\theta(x_t, c) - \epsilon\| \sim 0$, images generated by other methods are likely to elicit a different response; (ii) the prevalent use of classifier-free guidance for sampling and generation to adhere to textual prompts and improve image quality facilitates the detection of images generated by diffusion methods, which adding on the objective an extra guidance term $(\omega+1)\,(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t.\emptyset))$. By comparing the response of the diffusion network (i.e., the predicted noise $\epsilon_\theta(x_t, c)$) against sampled noise ϵ, we can detect fake images. Additionally, because our detection method does not rely on frequency or direct image-based forensic cues, it demonstrates strong robustness against image degradations such as down-sampling and blurring.

These advancements present new challenges for traditional synthetic image detection methods, which have primarily focused on GAN-generated images. To maintain generality across various generative models, many detection methods rely on local forensic clues or frequency-based analysis [28, 32]. This reliance on pixel or frequency analysis, however, often results in *diminished robustness* against real-world image degradations e.g.: compression, resampling, or blurring, due to their significant impact on the frequency characteristics of images. To address this, some works [23, 36] successfully leverage large-scale pre-trained models like CLIP [25], originally designed to learn universal text-image embeddings. Intuitively, since diffusion models are trained on large-scale datasets of real images and can generate realistic images, they shall have the potential to be used as tools for confirming whether an image belongs to real image distributions. One example is presented in [15], which proposes a sampling-based method using a pre-trained classifier-guided diffusion model for traditional image classification tasks. This work showcases the effective zero-shot labelling capabilities of diffusion-based generative models. However, this method is constrained by the need for prior knowledge of all in-domain class labels to compute relative information among labels, which limits its direct applicability for fake image

detection. In the realm of forgery detection, DIRE [33] utilizes the deterministic nature of DDIM [29] networks to invert images back to their initial Gaussian distributions. The discrepancy between inversion and generation serves as a measure of authenticity. While this approach is effective for certain diffusion-based models [37], its reliance on textual conditioning and inversion information restricts its generalizability. For example, inverting with different prompts may yield inconsistent results [37]. Additionally, since the final data representation remains image-based, it is vulnerable to frequency perturbations, a common limitation in other image-centric methods.

Inspired by these findings, we hypothesize that AI-generated images, especially from diffusion models, can be distinguished by analyzing their response to sampled noise, similar to the diffusion training process. This hypothesis comprises **two key considerations** (see Figure 1): (i) diffusion models contain extensive information about the *real data* distribution, which can be exploited to differentiate between real and fake images (e.g., generated by GANs or other methods) through sampling; (ii) text-to-image diffusion models (e.g. Stable Diffusion [27]) use classifier-free guidance [13] to adhere to textual prompts, which often requires a higher guidance weight ω, e.g., $\omega = 7.5$ for Stable Diffusion. Consequently, this enhances textual alignment but reduces image fidelity, thereby amplifying the differences between real and synthetic images in our sampling-based detection method and making text-to-image diffusion-generated images more detectable.

Motivated by these observations, our proposed method consists of two steps: (i) We input noised images into a diffusion model (e.g., Stable Diffusion v1.4 [27]) and collect the unconditional predicted denoising images, combined with sampled noise over multiple timesteps. This step aims to leverage a pre-trained diffusion model to gather informative data, helping discern between real and fake images. The main intuition consists in that real and synthetic images exhibit different responses to applied noise in the diffusion process, thus providing valuable information for identification. (ii) We use a deep network classifier to detect real and fake images based on the collected information. Given the stochastic sampling process and latent representation of Stable Diffusion, our method demonstrates generalized performance across multiple image generators and robustness against various image degradation techniques.

Contributions: We make the following contributions:

1. We show that pre-trained diffusion models can provide useful discriminative information for synthetic image detection.
2. We propose a learning-based framework with a novel design that uses unconditional model responses to detect fake images, avoiding reliance on frequency cues and prompts.
3. Our method exhibits generalized and robust detection performance across various AI-generated images and image degradation techniques, as evidenced by extensive experimental validation.

2 Related Works

Synthetic Image Detection. Discerning real images from forged ones has been a critical task even before the advent of deep learning-based generative methods. Traditional approaches have primarily focused on identifying compression artefacts, sampling anomalies, or incorrect physical phenomena such as reflections or perspectives to detect manipulated images [1, 2, 24, 34]. With the progress of GAN-based generative techniques, the focus shifted towards employing deep neural networks for the detection of synthesized images, leveraging the capabilities of deep network classifiers [21]. Recently, increasing attention has been placed on the generalization of classifiers, aiming to develop models capable of detecting fake images generated by various methods [6, 7, 14, 35]. Despite this shift, many strategies continue to rely on frequency analysis to identify characteristic frequency fingerprints or low-level forensic cues associated with synthetic image generation [16, 20]. To overcome image degradation such as compression, lower resolution etc. CNNSpot [32] merged the preprocessing and data augmentation before the training process to improve the robustness.

Diffusion Models. Diffusion models have recently demonstrated remarkable capability in generating photorealistic synthetic images, surpassing the performance of previous state-of-the-art GAN-based models. Unconditional generation models, such as DDPM [12], laid the groundwork for image synthesis diffusion models operating in pixel space. DDIM [29] introduced an alternative approach by relaxing the Markovian assumption in DDPM, enabling faster sampling with minimal quality degradation. Subsequent developments, including classifier-guided [9] and classifier-free guidance [13], introduced more versatile textual prompt conditioning, often utilizing textual encoders like CLIP [25]. This paved the way for numerous text-to-image synthesis models, such as GLIDE [22] and the Latent Diffusion Model [27], where the diffusion process is executed in a VAE latent space. VQDiffusion [10] leverages the VQ-VAE [31] space for its diffusion process. Noteworthy applications, including the Stable Diffusion series and Midjourney, have also gained significant attention.

Synthetic Detector for Diffusion Models. These powerful generative models pose new challenges for detection methods, particularly in terms of generalization and domain transfer capabilities. Ricker et al. [26] observed that diffusion models do not exhibit distinct frequency patterns, which are less pronounced than those of GAN-based methods. DIRE [33] suggested using the distance between DDIM-inverted and reconstructed images for fake image detection, but this method struggles with generalization due to its reliance on the conditional inversion of specific models. Ojha et al. [23] introduced UnivFD, a novel approach that utilizes the feature space of a frozen pre-trained vision-language model (CLIP-ViT) to focus on image features rather than frequency information. However, the discriminative nature of CLIP-based models limits their generalization across different generative models and degradation types. Concurrently, building on UnivFD [23], GenDet [36] proposed a teacher-student model to accentuate subtle differences in the feature space. SSP [5] demonstrated that synthetic images could be detected using a simple low-variance patch. Although the paper

does not address degradation and robustness, given the method's mechanism, it is plausible to be vulnerable to noise perturbation. It is also worth mentioning, yet outside the scope of fake image detection, Li et al. [15] proposed a novel classification method that involves sampling noise on the responses of diffusion models to input images. This approach demonstrates the versatility of generative models, particularly diffusion models, in performing classification tasks, highlighting their potential beyond image synthesis. However, their method cannot be directly applied to synthetic image detection, as it requires prior knowledge of all class labels to compute the relative distance among difference labels, which is not feasible in real and fake image detection scenarios. Moreover, classification labels require a strong dependence on the conditioning prompt, which is unsuitable for generated image detection scenarios where (i) the conditioning prompt can be agnostic, and (ii) the performance should be as general as possible across prompt variations.

3 Method

Building upon insights from prior research, we introduce our method. The main intuition of our method is to train a classifier by noising input images and collecting the predicted noise from diffusion models, mirroring the diffusion training process. By analyzing the predicted noise responses, we can effectively discern synthetically generated images from real ones. These images often exhibit deviations from the original data distribution, particularly when conditioned with techniques like Classifier-Free Guidance (CFG) with high guidance weight. Unlike many frequency-based detection strategies, our method exhibits resilience to frequency perturbations and image degradation. This robustness stems from the stochastic nature of our sampling and noising process and the latent representation inherent in the VAE framework. Before delving into the specifics of our proposed method, we first introduce the fundamental background of diffusion models and guidance methods.

3.1 Background of Diffusion and CFG

Diffusion Models for Image Synthesis. Building upon the foundational work of DDPM [12], the core objective of the diffusion model is to train a neural network ϵ_θ, to effectively denoise data that has been previously noised over a series of timesteps. As a generative model, the training of DDPM seeks to approximate generated data towards the entire data distribution, denoted as p_{data}.

The model aims to reconstruct the original data instance x_0, drawn from p_{data}, from its noised counterpart x_t. This noised data x_t is represented as a combination of the original data x_0 and Gaussian noise ϵ, scaled by the noise level $\gamma(t)$: $x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon$, where $\gamma(t) \in [0,1]$ is a monotonically decreasing function of the timestep t, and $\epsilon \sim \mathcal{N}(0,1)$ denotes standard Gaussian noise. DDPM [12] also finds that predicting the added noise ϵ rather than

directly reconstructing x_0, and omitting the variational lower bound (VLB) scaling factors, enhances the model's performance. Consequently, a simpler training objective for ϵ_θ is defined by a loss function that minimizes the distance between the predicted noise and the sampled noise, as follows:

$$L_{\text{simple}} = \mathbb{E}_{x_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0,1), t \sim \mathcal{U}[0,1]} \left[\|\epsilon_\theta(x_t) - \epsilon\| \right] \quad . \tag{1}$$

Upon completion of the training, the model can generate new data instances from the distribution of the original dataset, p_{data}. This is achieved by initiating with a sample from a standard Gaussian distribution, $x_T = \epsilon \sim \mathcal{N}(0, 1)$, and iteratively denoising it to produce a synthetic data instance $\hat{x_0}$ that approximates to the original data distribution. This denoising process can be implemented using various sampling strategies, e.g., DDPM [12] and DDIM [29].

Classifier-Free Guidance. To enable conditioning the output of diffusion models such as $p(x_t|c)$, Dhariwal et al. [9] propose Classifier-Guidance (CG) that utilizes a pre-trained classifier $p(c|x_t)$, thereby forming the conditional input as $\nabla_{x_t} \log p(x_t|c) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(c|x_t)$, in accordance with Bayes' rule. More importantly, they propose to add a scaling ω on the classifier gradient to make the conditioning process manipulable and amplifiable. However, CG requires training a noise-dependent classifier, which can be cumbersome, particularly for novel classes, and poses challenges for more complex conditioning formats, such as textual prompts. To address this, Ho et al. [13] suggest an alternative approach by employing an implicit classifier, expressed as $\nabla_{x_t} \log p(c|x_t) = \nabla_{x_t} \log p(x_t, c) - \nabla_{x_t} \log p(x_t)$. This method involves training the diffusion network on the joint distribution of data and condition, modifying the loss function L_{simple} to include the condition c by replacing $\epsilon_\theta(x_t)$ with $\epsilon_\theta(x_t, c)$. To represent the condition of unconditional generation, during the training, the condition c is occasionally dropped with a constant probability p_{cond}, resulting in an unconditional response $\epsilon_\theta(x_t, \emptyset)$.

$$\begin{aligned} L_{\text{simple cfg}} = \mathbb{E}_{x_0 \sim p_{\text{data}}, \epsilon \sim \mathcal{N}(0,1), t \sim \mathcal{U}[0,1]} \left[\|\epsilon_\theta(x_t, c) - \epsilon\| \right] \\ c = \emptyset \text{ with } p_{\text{cond}}. \end{aligned} \tag{2}$$

With a single network learns both conditional predicted noise $\epsilon_\theta(x_t, c)$ and unconditional predicted noise $\epsilon_\theta(x_t.\emptyset)$, we can formulate the predicted noise for generation process of *classifier-free guidance (CFG)*, also controlled by a guidance weight ω:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + \omega \left(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t.\emptyset) \right) \quad . \tag{3}$$

Or an alternative form:

$$\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, \emptyset) + (\omega + 1) \left(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t.\emptyset) \right) \quad . \tag{4}$$

In practice, the guidance weight ω is often set to a higher value, such as $\omega = 7.5$ for Stable Diffusion 1.5, to ensure adherence to conditional inputs and empirically improve image quality.

Latent Diffusion Model Instead of operating directly in the pixel space as in DDPM [12], subsequent developments such as the Latent Diffusion Model [27] suggest conducting the diffusion process within a pre-trained VAE-like latent space. This approach not only accelerates training and generation times but also shifts the model's focus towards capturing perceptual content rather than merely replicating low-level pixel details. This design is used in various architectures of text-to-image applications e.g. Stable Diffusion, which leverages a pre-trained VAE space and CLIP [25] embeddings with classifier-free guidance to generate images from textual prompts.

3.2 Your diffusion model is an implicit synthetic image detector

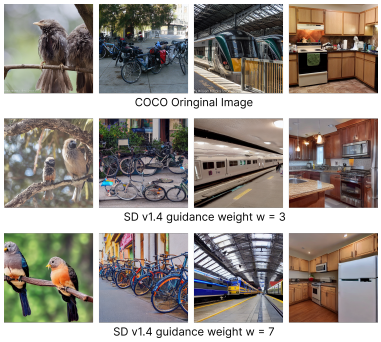
Our main intuition of this paper is to leverage the diffusion model's sampled responses for determining whether an image belongs to the distribution of real images or originates from the generated models.

Hypothesis. For a given image (or the image latent from VAE) x, the noised image x_t at a specific timestep t can be achieved using a noise schedule and by sampling standard Gaussian noise $\epsilon \sim \mathcal{N}(0, 1)$: $x_t = \sqrt{\gamma(t)}x_0 + \sqrt{1-\gamma(t)}\epsilon$. When the noised image x_t is input into a diffusion model, such as Stable Diffusion, with a known condition c or unconditional label $\emptyset$, we can compute both the conditional and unconditional predicted noise, denoted as $\epsilon_\theta(\hat{x}_{t,\epsilon}, c)$ and $\epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset)$ (recall that $\emptyset$ is computed by dropping the condition, therefore can be treated as a special label), respectively. We refer to these two $\epsilon_\theta(\hat{x}_{t,\epsilon}, \cdot)$ as the model's response at a given timestep.

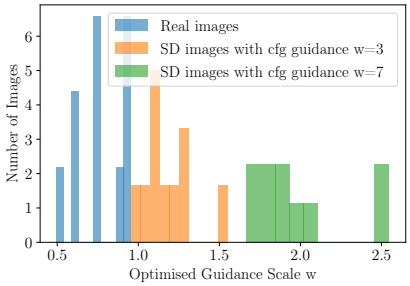
Our main hypothesis is that a diffusion model can be effectively transformed into a detector when collecting sampled responses. Given that the diffusion model is trained on the *real* dataset, the sampled noise ϵ should be close to the predicted response $\epsilon \approx \epsilon_\theta(\hat{x}_{t,\epsilon}, c)$ or $\epsilon \approx \epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset)$, indicating the model's proficiency in handling real data of the training loss, i.e. *minimizing*: $\|\epsilon_\theta(x_t, c) - \epsilon\|$ (see Eq. 2).

However, for fake images, two distinct cases arise: (1) Images from non-diffusion-based methods (e.g., GANs or direct editing) often show less adherence to the objective loss, since they are never trained in this manner, leading to a distinct model response from that of real images; (2) Images from diffusion models are expected to behave as $\epsilon \approx \epsilon_\theta(\hat{x}_{t,\epsilon}, c)$, since using the same objective function. However, to enhance generation quality and textual alignment, the majority of generation methods employ guidance strategies (e.g., CFG [13]) with high guidance weights ω: $\hat{\epsilon}_\theta(x_t, c) = \epsilon_\theta(x_t, c) + \omega\,(\epsilon_\theta(x_t, c) - \epsilon_\theta(x_t, \emptyset))$. The extra guidance term skews the generation away from the real dataset distribution, causing the synthetic images to diverge from the sampled noise ϵ, and leading to a different behaviour that $\epsilon \neq \epsilon_\theta(\hat{x}_{t,\epsilon}, c)$.

Validation. To validate our hypothesis, especially the diffusion with CFG case, we use the COCO dataset [17] along with its associated prompts to generate a series of images using Stable Diffusion v1.4 [27], each with varying guidance weights. Figure 2a showcases the VAE-encoded and then re-decoded images produced under different guidance settings with the original images. The corresponding prompts for these images, arranged from left to right, are as follows:



(a) Comparison of the real and generated image with different ω from same prompt ($\omega = 7$ is recommended by Stable Diffusion).



(b) Optimised ω^* from real images and Stable Diffusion generated images with different classifier-free guidance scales.

Fig. 2: Figure of comparison of real images from the COCO dataset with Stable Diffusion v1.4 generated ones under different Classifier-Free Guidance weight: $\omega = 3$ and $\omega = 7$ (recommended value to ensure textual alignment). and **(b) the histogram of the optimized ω^* from 10 real and generated images** by sampling method, reveals that images generated with higher guidance (i.e. $\omega = 7$) can be successfully retrieved as higher optimised ω than the real image's results. This pattern underscores the feasibility of utilizing the model's response and the associated sampled noise as reliable indicators to distinguish between real and synthesized images.

"Two birds sitting on a narrow branch next to each other, looking in opposite directions", *"A bunch of bicycles parked on the street, surrounded by various items"*, *"Several trains parked next to a platform, beneath an overhead ceiling"*, and *"A kitchen area featuring a white refrigerator, a stove, other appliances, and brown cabinets."*. We see clearly when $\omega = 7$, the results are visually and textual better than when $\omega = 3$, justifying the default recommended higher guidance.

We can then optimise and retrieve the ω by re-injecting the generated or real images back to the diffusion model and collect their conditional and unconditional response by defining an optimisation problem (see also algorithm 1):

$$w_x^* = \arg\min_{\omega \in R} E_{\epsilon \sim \mathcal{N}(0,1), t \sim \mathcal{U}[0,1]} \left[\| \epsilon - (\epsilon_\theta(\hat{x}_{t,\epsilon}, c) + \omega(\epsilon_\theta(\hat{x}_{t,\epsilon}, c) - \epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset))) \|^2 \right] \tag{5}$$

Ideally, an optimized guidance weight $\omega^* \approx 0$ indicates that the image closely resembles a real image, as the diffusion model is trained according to the equation given in Eq. 2. Figure 2b presents the distribution of optimized ω^* values for 10 images, where it is evident that synthetically generated images (represented in green and orange) exhibit higher ω^* values. This observation strongly suggests that the conditional or unconditional responses from the diffusion model do not match the sampled noise ϵ, thereby differentiating generated images from real ones. In other words, for real images we could hypothesise $\epsilon \approx \epsilon_\theta(\hat{x}_{t,\epsilon}, c)$ or $\epsilon \approx \epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset)$, which is not the case for fake images using higher CFG guidance

weight, where $\epsilon \neq \epsilon_\theta(\hat{x}_{t,\epsilon}, c)$. More interestingly, this behaviour does not take any *shortcut* from frequency analysis or recognizing special frequency fingerprints. Instead, it only exploits the pre-trained diffusion model to stochastically judge if an image lies on the real data distribution; hence, theoretically, it should be robust against perturbations and degradation.

Algorithm 1 Optimisation of guidance scale from generated images

Input: test image $\hat{x}$, conditioning inputs c (e.g., text embeddings), number of loop N per input
for $i = 1, \ldots, N$ **do**
  Sample $t \sim \text{Uniform}(0, 1)$, $\epsilon \sim \mathcal{N}(0, I)$
  $\hat{x}_t = \sqrt{\alpha_t}\hat{x} + \sqrt{1-\alpha_t}\epsilon$
  $\epsilon_\theta(\hat{x}_t, c) = UNet(\hat{x}_t, t, c)$
  $\epsilon_\theta(\hat{x}_t, \emptyset) = UNet(\hat{x}_t, t, \emptyset)$
  $Errors[N].\text{append}(\|\epsilon - (\epsilon_\theta(\hat{x}_{t,\epsilon}, c) + \omega(\epsilon_\theta(\hat{x}_{t,\epsilon}, c) - \epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset)))\|^2)$
end for
return $\arg\min_\omega \text{mean}(Errors)$

Method. Optimizing ω for each image individually presents a significant computational challenge, requiring the sampling of numerous ϵ values for accurate estimation. Furthermore, this approach fails to leverage the rich information in the model's response (with the size of $4 \times 64 \times 64$ per sample) but reduces it instead to scalar data (estimated ω^*). Consequently, we propose a method that directly distinguishes between real and fake images by analyzing the diffusion model's informative response from input images using a DNN classifier.

The framework of our method consists of two main components: a Model Response Sampler and a Classifier, depicted in Figure 3 in the left and right panels respectively.

The Model Response Sampler operates by leveraging a pre-trained diffusion model, such as Stable Diffusion, to sample and collect the corresponding unconditional response $\epsilon_\theta(x_t, \emptyset)$. Although one might consider using both conditional and unconditional responses, our empirical findings suggest that relying on the conditional response $\epsilon_\theta(x_t, c)$ can lead to overfitting to the specific generative model used for training (see Ablation study in section 4.3). In addition, sampling the conditional response requires a semantic condition or prior label information which are often not available in real-world cases.

The sampler is shown in Figure 3 left panel, we sample these responses from the diffusion model across a predefined number of timesteps N (e.g., 10) to across the entire generation process. For each image, this sampling procedure is repeated K times (e.g., 20) and averaged to mitigate stochastic variability from the diffusion model and sampling. The sampled model responses are then aggregated into a triplet of $(\epsilon, \epsilon_\theta(x_t, \emptyset), |\epsilon - \epsilon_\theta(x_t, \emptyset)|)$ as the input for the image detector. Consequently, the input dimension for the classifier becomes (B, N, C, W, H), where B is the batch size, N is the number of timesteps, C is

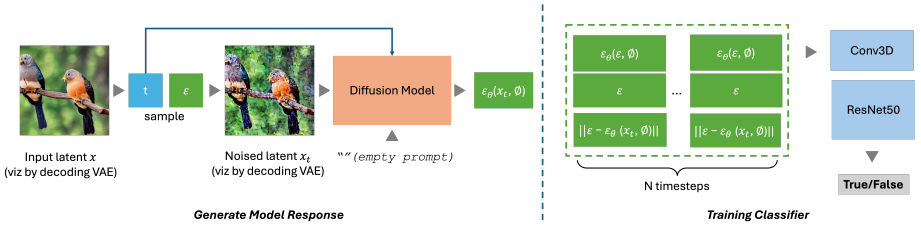


Fig. 3: The pipeline of our method contains two parts: a sampler (left) and a classifier (right), the job of the sampler is to sample and collect model response from a pre-trained diffusion model. The collected data are inputted into a classifier with ResNet50 backbone to predict if an image is real or fake.

the stacked channel of the triplet, and W and H are the width and height of the encoded images (64 in the case of VAE-encoder). Once we have the sampled data, we employ our classifier backboned by a ConvNet and a ResNet-50 [11] (see Figure 3 right panel). This network is trained with a Binary Cross-Entropy loss and outputs binary predictions, discerning between real and fake images.

4 Experiments

4.1 Dataset

For our experiments, we utilized the GenImage [37] dataset, which comprises real images sourced from ImageNet [8] and their corresponding labels, alongside synthetically generated images produced by *eight distinct generative methods*, include: Stable Diffusion V1.4 [27], Stable Diffusion V1.5 [27], GLIDE [22], VQDM [10], Wukong, BigGAN [4], ADM [9], and Midjourney (V5). The textual prompts used for generating images adhere to the format *"Photo of {label}"*, with labels drawn from the same set of 1,000 categories used in ImageNet. See Figure 4 for some snippets of the dataset on various methods.

Each generative method holds both the real and fake subsets of the training and validation sets. In total, the GenImage dataset encompasses 1,331,167 real images and 1,350,000 fake images, distributing approximately 160,000 images for each generative model.

4.2 Benchmark and results

We conduct two experiments: cross-domain performance and robustness assessment. For cross-domain, in line with GenImage [37], we train *exclusively* on images generated by Stable Diffusion V1.4 (SD V1.4) and test on different subsets from other generative methods. This aims to assess the detection capability and generalizability of the classifiers. For robustness, following [37], both the training and test phases are conducted using images generated by SD V1.4 and

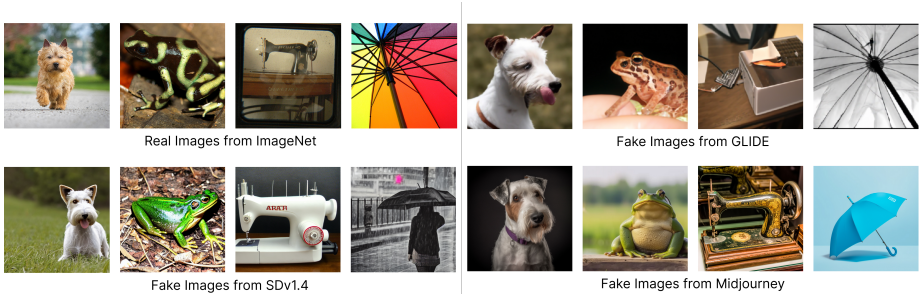


Fig. 4: Image samples from GenImage [37] dataset of real images from ImageNet and synthetic images generated by different methods.

real subset. To evaluate the robustness of detection methods, we include various degradations in the test images: Lower-resolution (LR): The resolution of input images is reduced to either 112 or 64 pixels; JPEG compression: Images are compressed using JPEG with quality settings of 65 or 30; Gaussian blur: Images are blurred using a Gaussian filter with σ=3 or σ=5. We report the binary classification accuracy for all degraded images.

Method	Midjourney	SDV1.4	SDV1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg Acc.	Rank
ResNet-50 [11]	54.9	99.9	99.7	53.5	61.9	98.2	56.6	52.0	72.1	4
DeiT-S [30]	55.6	99.9	99.8	49.8	58.1	98.9	56.9	53.5	71.6	5
Swin-T [18]	62.1	99.9	99.8	49.8	67.6	99.1	62.3	57.6	74.8	2
CNNSpot [32]	52.8	96.3	95.9	50.1	39.8	78.6	53.4	46.8	64.2	10
Spec [35]	52.0	99.4	99.2	49.7	49.8	94.8	55.6	49.8	68.8	8
F3Net [28]	50.1	99.9	99.9	49.9	50.0	99.9	49.9	49.9	68.7	9
GramNet [19]	54.2	99.2	99.1	50.3	54.6	98.9	50.8	51.7	69.9	7
DIRE [33]	60.2	99.9	99.8	50.9	55.0	99.2	50.1	50.2	70.7	6
UnivFD [3]	73.2	84.2	84.0	55.2	76.9	75.6	56.9	80.3	73.3	3
Ours	57.8	91.9	90.4	52.7	83.7	89.4	61.1	78.9	**75.7**	1

Table 1: Results of different methods trained on SDV1.4 and cross-testing with other models generated test set of GenImage Dataset.

In both experiments, we benchmark our method against models from the GenImage dataset and other state-of-the-art diffusion detection methods, including ResNet-50 [11], DeiT-S [30], and Swin-T [18], as well as specialized synthetic image detection methods including some frequency or low-level-based CNNSpot [32], Spec [35], F3Net [28]; more global feature-based GramNet [19], diffusion-inversion-based DIRE [33], and CLIP pre-trained model UnivFD [3].

Cross-Domain Results Table 1 reports the results of our cross-domain experiment. When trained on SD V1.4, most frequency-based methods and naive backbones suffer overfitting and fail to generalize to other methods. Our method achieves an accuracy of 91.9% on the test set of SD v1.4, indicating robust performance within the same domain. However, it does not reach 99.9%, as other

methods do, mainly due to the stochasticity of the sampling process. Comparable results are observed on SD v1.5 and Wukong, since these models share a similar framework and use classifier-free guidance to generate images, attributed to the consistent performance. For the in-domain evaluation, our method outperforms another latent embedding-based approach with no usage of direct image information, UnivFD [3], which reports an accuracy of 84.0% for SDV1.4.

For the image generated from different frameworks, our approach also leads the accuracy on GLIDE (83.7%) than UnivFD [3] (76.9%) and significantly surpassing other methods. We believe this is due to the guidance setting during the generation of GLIDE. A similar trend is observed with the BigGAN generator, where our method ranks second in accuracy, slightly behind UnivFD [3]. We report a leading average accuracy of 75.7% and showed good cross-domain performance, especially GLIDE and BigGAN. Many frequency-based methods [28, 32] including generative model-based detection approaches DIRE [33], despite achieving high in-domain accuracy (SDV1.4), often fail to generalize to other AI-generated content.

Compared to UnivFD [3], we attribute the reasons for lower performance on Midjourney (57.8% vs. 73.2%) as follows: (i) the potentially more discriminative embedding space provided by the CLIP model to general images; (ii) Methods like Midjourney may present less guidance phenomenon and yield better fidelity performance to the original image, makes them hard to be noticed by our sampling-based model. Notice also that we did not optimize the architecture or the hyper-parameters of the classifier which is a simple ResNet-50. Better results are to be expected by exhaustively searching for better classifiers, without deviating from the core idea proposed in this paper.

Method	LR (112)	LR (64)	JPEG (q=65)	JPEG (q=30)	Blur (σ=3)	Blur (σ=5)	Avg Acc.	Rank
ResNet-50 [11]	96.2	57.4	51.9	51.2	97.9	69.4	70.6	5
DeiT-S [30]	97.1	54	55.6	50.5	94.4	67.2	69.8	6
Swin-T [18]	97.4	54.6	52.5	50.9	94.5	52.5	67.0	8
CNNSpot [32]	50.0	50.0	97.3	97.3	97.4	77.9	78.3	3
Spec [35]	50.0	49.9	50.8	50.4	49.9	49.9	50.1	10
F3Net [28]	50.0	50.0	89	74.4	57.9	51.7	62.1	9
GramNet [19]	98.8	94.9	68.8	53.4	95.9	81.6	82.2	2
DIRE [33]	64.1	53.5	85.4	65.0	88.8	56.5	68.9	7
UnivFD [3]	88.2	78.5	85.8	83.0	69.7	65.7	78.5	4
Ours	89.2	78.5	82.6	74.4	90.6	89.5	**84.1**	1

Table 2: Results of different methods trained on SD v1.4 and tested on degraded images of SD v1.4 testset from GenImage Dataset.

Robustness Results In Table 2, we detail the findings from our robustness experiment. This experiment involved training our model with images generated by SDV1.4 and subsequently testing it on degraded images within the same domain to assess the robustness. The results, as shown in Table 2, highlight our method's consistently high robustness across various image degradation techniques. While our method's average robustness also leads all the compared methods, it also

exhibits lower variability and more consistent performance across all forms of image degradation. The second best method, GramNet [19] (a global texture method) robustness is significantly compromised by JPEG compression, with its accuracy dropping to approximately 50% when the compression quality is reduced to 30. Note that UnivFD [3] exhibits commendable robustness due to its use of clip-based information. This corroborates our argument that relying directly on image information makes it vulnerable to degradation.

4.3 Ablation

As discussed in Section 3, using conditional responses may compromise the generalization of the model, e.g., erroneous or imprecise prompts could deteriorate the detection. Additionally, extra efforts are required to analyze the semantics of the image to create appropriate conditions. Here, we ablate different conditional and unconditional inputs. To obtain the conditional prompts, we follow the approach used in the GenImage Dataset [37] and generate captions for image generation using the textual prompt format *"Photo of {label}"*, where the labels are sourced from the same 1,000 categories in ImageNet.

Method	Midjourney	SD V1.4	SD V1.5	ADM	GLIDE	Wukong	VQDM	BigGAN	Avg Acc. (%)
C&N	54.2	93.9	93.9	47.3	53.1	93.0	51.9	45.4	68.4
C&U&N	55.7	94.1	94.2	49.4	49.9	91.3	56.0	44.3	66.9
C&U&N-Sub	53.2	91.6	91.6	48.6	50.7	91.7	55.9	43.8	65.9
U&N (Ours)	57.8	91.9	90.4	52.7	83.7	89.4	61.1	78.9	**75.7**

Table 3: Generalization ablation of different input types: C&N (Conditional and Noise), C&U&N (Conditional, unconditional and Noise) and ours chosen U&N (Unconditional and Noise). The model is trained on SDv1.4 and tested on different generators. We observe that the conditional response often leads to overfitting to Stable Diffusion and Wukong model, whereas using only unconditional response (U&N) helps to improve the generalization to unseen models.

We evaluate four distinct combinations of input triplets including conditional and unconditional input to assess their impact and justify the design of using unconditional response for our method:

1. **C&N**: $(\epsilon_\theta(\hat{x}_{t,\epsilon}, c), \epsilon, \|\epsilon_\theta(\hat{x}_{t,\epsilon}, c) - \epsilon\|)$, a triplet incorporates the conditional response $\epsilon_\theta(\hat{x}_{t,\epsilon}, c$, sampled random noise ϵ and their difference.
2. **U&N**: $(\epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset), \epsilon, \|\epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset) - \epsilon\|)$, a triplet utilizing the unconditional response $\epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset)$ and sampled random noise ϵ and their difference.
3. **U&C&N**: $(\epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset), \epsilon_\theta(\hat{x}_{t,\epsilon}, c), \epsilon)$, a triplet includes both conditional and unconditional responses alongside the sampled random noise.
4. **U&C&N-sub**: $(\|\epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset) - \epsilon\|, \|\epsilon_\theta(\hat{x}_{t,\epsilon}, c) - \epsilon\|, \|\epsilon_\theta(\hat{x}_{t,\epsilon}, c) - \epsilon_\theta(\hat{x}_{t,\epsilon}, \emptyset)\|)$, a triplet also integrating both conditional and unconditional responses with the sampled random noise, focusing on the differences between these responses.

In line with the experimental setup described earlier, we evaluate the performance of different input configurations in terms of cross-domain generalization and robustness, with results detailed in Table 3 and Table 4. For cross-domain generalization, it is evident that configurations incorporating conditional responses (**C&N**, **U&C&N**, and **U&C&N-sub**) tend to overfit, relying predominantly on the model's conditional response. This is highlighted by the notable performance on the Stable Diffusion series (including Wukong) across these configurations, while performance significantly drops for other generator methods.

Method	LR (112)	LR (64)	JPEG (q=65)	JPEG (q=30)	Blur (σ=3)	Blur (σ=5)	Avg Acc(%)
C&N	50.4	48.2	84.4	72.1	48.7	50.0	59.0
C&U&N	51.2	48.3	83.9	76.2	49.2	46.9	59.3
C&U&N-Sub	48.7	46.6	83.5	69.9	47.0	45.9	56.9
U&N (Ours)	89.2	78.5	82.6	74.4	90.6	89.5	84.1

Table 4: Robustness ablation of different input types: C&N (Conditional and Noise), C&U&N (Conditional, unconditional and Noise) and ours chosen U&N (Unconditional and Noise). The model is trained on SDv1.4 and tested on SDv1.4 with different types of image degradations.

Table 4 further shows the impact of input design on robustness against various types of image degradation. Except for JPEG compression, all conditional response-inclusive configurations exhibit diminished robustness. This aligns with our hypothesis that the sampling method and latent representation inherently mitigate shortcuts based on image information, thus showing less susceptibility to JPEG compression artefacts. In contrast, our chosen **U&N** configuration, which leverages only the unconditional response and sampled noise, consistently maintains both high generalization and robustness across different scenarios.

5 Conclusion

In conclusion, we proposed a novel method for detecting AI-generated images by harnessing the capabilities of pre-trained diffusion models and a sampling technique focused on the unconditional response of the model. Our approach transforms diffusion models into implicit detectors of synthetic images. The sampling process and the latent representation encoded within the models enable our method to achieve broad generalizability across various image generators. More importantly, we have demonstrated that our method exhibits uniformly superior robustness against various image degradations, a frequent challenge in real-world cases. Future efforts will concentrate on reducing the number of required timesteps and sampling iterations, aiming to improve the efficiency and speed of fake image detection, which remains a current limitation of our method.

Acknowledgement

This work was supported by ANR APATE ANR-22-CE39-0016, Hi!Paris grant and fellowship, and was granted access to the High-Performance Computing (HPC) resources of IDRIS under the allocations 2024-AD011014300R1 made by GENCI. We would like to thank Nicolas Dufour, David Picard, and the anonymous reviewers for their insightful comments and suggestions.

References

1. Agarwal, S., Farid, H.: Photo forensics from jpeg dimples. In: IEEE workshop on information forensics and security (WIFS) (2017)
2. Agarwal, S., Farid, H.: Photo forensics from rounding artifacts. In: Proceedings of the 2020 ACM Workshop on Information Hiding and Multimedia Security (2020)
3. Bansal, A., Chu, H.M., Schwarzschild, A., Sengupta, S., Goldblum, M., Geiping, J., Goldstein, T.: Universal guidance for diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Chen, J., Yao, J., Niu, L.: A single simple patch is all you need for ai-generated image detection. arXiv preprint arXiv:2402.01123 (2024)
6. Chen, L., Zhang, Y., Song, Y., Liu, L., Wang, J.: Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
7. Cozzolino, D., Thies, J., Rössler, A., Riess, C., Nießner, M., Verdoliva, L.: Forensictransfer: Weakly-supervised domain adaptation for forgery detection. arXiv preprint arXiv:1812.02510 (2018)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conf. Comput. Vis. Pattern Recog. (2009)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Adv. Neural Inform. Process. Syst. (2021)
10. Gu, S., Chen, D., Bao, J., Wen, F., Zhang, B., Chen, D., Yuan, L., Guo, B.: Vector quantized diffusion model for text-to-image synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conf. Comput. Vis. Pattern Recog. (2016)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Adv. Neural Inform. Process. Syst. (2020)
13. Ho, J., Salimans, T.: Classifier-free diffusion guidance. In: Adv. Neural Inform. Process. Syst. Worksh. (2021)
14. Lee, S., Tariq, S., Kim, J., Woo, S.S.: Tar: Generalized forensic framework to detect deepfakes using weakly supervised learning. In: IFIP International Conference on ICT Systems Security and Privacy Protection. Springer (2021)
15. Li, A.C., Prabhudesai, M., Duggal, S., Brown, E., Pathak, D.: Your diffusion model is secretly a zero-shot classifier. Int. Conf. Comput. Vis. (2023)
16. Li, J., Xie, H., Li, J., Wang, Z., Zhang, Y.: Frequency-aware discriminative feature learning supervised by single-center loss for face forgery detection. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)

17. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Eur. Conf. Comput. Vis. (2014)
18. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10012–10022 (2021)
19. Liu, Z., Qi, X., Torr, P.H.: Global texture enhancement for fake face detection in the wild. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
20. Luo, Y., Zhang, Y., Yan, J., Liu, W.: Generalizing face forgery detection with high-frequency features. In: IEEE Conf. Comput. Vis. Pattern Recog. (2021)
21. Marra, F., Gragnaniello, D., Cozzolino, D., Verdoliva, L.: Detection of gan-generated fake images over social networks. In: IEEE conference on multimedia information processing and retrieval (MIPR) (2018)
22. Nichol, A., Dhariwal, P.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
23. Ojha, U., Li, Y., Lee, Y.J.: Towards universal fake image detectors that generalize across generative models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2023)
24. Popescu, A.C., Farid, H.: Exposing digital forgeries by detecting traces of resampling. IEEE Trans. Image Process. (2005)
25. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. (2021)
26. Ricker, J., Damm, S., Holz, T., Fischer, A.: Towards the detection of diffusion model deepfakes. arXiv preprint arXiv:2210.14571 (2022)
27. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. (2022)
28. Shi, H., Cao, G., Zhang, Y., Ge, Z., Liu, Y., Yang, D.: F 3 net: Fast fourier filter network for hyperspectral image classification. IEEE Transactions on Instrumentation and Measurement (2023)
29. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)
30. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jégou, H.: Training data-efficient image transformers & distillation through attention. In: Int. Conf. Mach. Learn. (2021)
31. Van Den Oord, A., Vinyals, O., et al.: Neural discrete representation learning. Adv. Neural Inform. Process. Syst. (2017)
32. Wang, S.Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: Cnn-generated images are surprisingly easy to spot... for now. In: IEEE Conf. Comput. Vis. Pattern Recog. (2020)
33. Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Int. Conf. Comput. Vis. (2023)
34. Yao, H., Wang, S., Zhao, Y., Zhang, X.: Detecting image forgery using perspective constraints. IEEE Signal Processing Letters (2011)
35. Zhang, X., Karaman, S., Chang, S.F.: Detecting and simulating artifacts in gan fake images. In: 2019 IEEE international workshop on information forensics and security (WIFS) (2019)
36. Zhu, M., Chen, H., Huang, M., Li, W., Hu, H., Hu, J., Wang, Y.: Gendet: Towards good generalizations for ai-generated image detection. arXiv preprint arXiv:2312.08880 (2023)

37. Zhu, M., Chen, H., Yan, Q., Huang, X., Lin, G., Li, W., Tu, Z., Hu, H., Hu, J., Wang, Y.: Genimage: A million-scale benchmark for detecting ai-generated image. Adv. Neural Inform. Process. Syst. (2024)