Diffusion Noise Feature: Accurate and Fast Generated Image Detection

Yichi Zhang[1] and Xiaogang Xu[1,2*]

[1*]Zhejiang University, Zhejiang, China .
[2]The Chinese University of Hong Kong, Hong Kong, China.

*Corresponding author(s). E-mail(s): xiaogangxu@zju.edu.cn;
Contributing authors: yichizhang@zju.edu.cn;

Abstract

Generative models have advanced to the point where they can produce remarkably realistic images. However, this capability also introduces the risk of spreading false or misleading information. Current methods for identifying generated images face challenges such as low accuracy and limited generalization. This paper aims to address these issues by developing a representation with strong generalization capabilities to improve the detection of generated images. Our research has shown that real and generated images exhibit distinct latent Gaussian representations when processed through an inverse diffusion process within a pretrained diffusion model. By leveraging this disparity, we can enhance subtle artifacts in generated images. Based on this insight, we propose a novel image representation called **D**iffusion **N**oise **F**eature (**DNF**). DNF is derived from the estimated noise generated during the inverse diffusion process. A simple classifier, such as ResNet50, trained on DNF, achieves high accuracy, robustness, and generalization capabilities in detecting generated images, even when the corresponding generator is constructed with datasets or structures not encountered during the classifier's training. Our experiments using four training datasets and five test sets demonstrate state-of-the-art detection performance.

Keywords: Generated Image Detection, Cross-Datasets/Models Generalization, Feature Engineering for Detection, Diffusion Model, AI-Generated Content

1 Introduction

In recent years, generative models have achieved remarkable success, with Diffusion Models (Ho et al., 2020; Dhariwal and Nichol, 2021; Rombach et al., 2022) acting as catalysts for a new wave of image generation techniques (Song et al., 2020; Gu et al., 2022; Peebles and Xie, 2023; Liu et al., 2022). Due to their large-scale training datasets and numerous parameters, these models can produce highly realistic images. However, their widespread use has introduced significant risks, including the spread of false information (Castillo et al., 2011; Giglietto et al., 2019; Qi et al., 2019), fabrication of evidence (Fanelli, 2009), breaches of privacy (Murdoch, 2021), and fraudulent activities (Uyyala and Yadav, 2023). Malicious actors can exploit these models to create convincing fake images for activities such as telecommunications fraud, leading to substantial losses. Consequently, the ability to discern whether an image is real or generated has become an urgent and critical issue that demands attention.

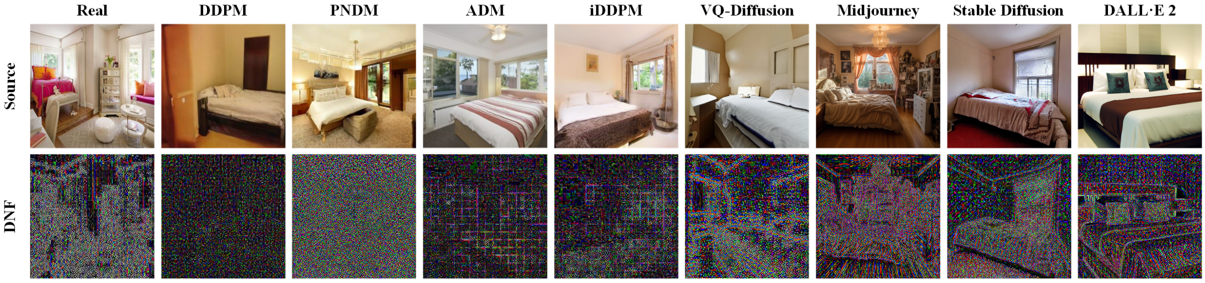

Fig. 1 **DNF of real image and generated images from eight different generators.** Nine pairs of images (Source-DNF) from LSUN-Bedroom (Yu et al., 2015) and various generators: DDPM (Ho et al., 2020), PNDM (Liu et al., 2022), ADM (Dhariwal and Nichol, 2021), iDDPM (Nichol and Dhariwal, 2021), VQ-Diffusion (Gu et al., 2022), Midjourney (Midjourney, 2022), Stable Diffusion (Rombach et al., 2022) and DALL·E 2 (Ramesh et al., 2022).

Several methods have been developed to detect generated images (Wang et al., 2020; Sinitsa and Fried, 2023; Shi et al., 2023; Tan et al., 2023; Wang et al., 2023; Zhong et al., 2023; Qian et al., 2020; Frank et al., 2020). Although these methods have proven effective for images synthesized by earlier models such as GANs (Goodfellow et al., 2020; Brock, 2018; Karras, 2017; Karras et al., 2019; Zhu et al., 2017; Choi et al., 2018), they often struggle with images produced by state-of-the-art generative models like DALL·E (Ramesh et al., 2021, 2022; Betker et al., 2023), Stable Diffusion (Rombach et al., 2022), and Midjourney (Midjourney, 2022). The realism of images generated by these advanced models closely approximates that of real images, rendering previous detection features insufficient for distinguishing between real and fake images. To address this challenge, two primary approaches should be considered. One approach involves extending current classifiers for detection. However, even with advanced models, subtle differences between real and fake data can often lead to classification failures. The alternative approach, which is the central focus of this paper, is to design a novel representation with exceptional generalization capabilities that can significantly improve the detection of generated images.

In this paper, we introduce a novel representation for detecting generated images, known as **D**iffusion **N**oise **F**eature (**DNF**). Unlike previous methods that extract features in the spatial or frequency domains (Wang et al., 2020; Frank et al., 2020), we leverage pre-trained diffusion models to construct image representations. The rationale behind this approach is that large-scale generative diffusion models are trained to learn the distribution of real images. When images from different distributions undergo the inverse diffusion process, they are unified into the same distribution, appearing as pure noise in the diffusion model. The estimated noises generated during this process contain significant information from the original image distribution, amplifying subtle differences between real and generated images, and manifesting distinct features in the estimated noise.

To implement this approach, we input the image to be detected into a pre-trained diffusion model and perform the inverse diffusion process. During this process, we collect the estimated noises generated at each step and then use a fusion strategy, determined experimentally, to obtain the DNF used for classification. As visualized in Figure 1, the estimated noise corresponding to images from different sources is significantly different. Moreover, as shown in Figure 2, the latent representation separation between real and generated image distributions in the DNF domain is more pronounced compared to other domains (Wang et al., 2020, 2023).

Extensive experiments conducted on four training datasets and five testsets have validated the state-of-the-art performance of the classifier trained on DNF in generated image detection. (i) The classifier trained on DNF demonstrated **99.8% accuracy** in both validation and testing, which significantly surpassed the average accuracy of 87.7% achieved by other methods (Wang et al., 2020; Frank et al., 2020; Chai et al., 2020; Shiohara and Yamasaki, 2022; Wang et al., 2023). (ii) The

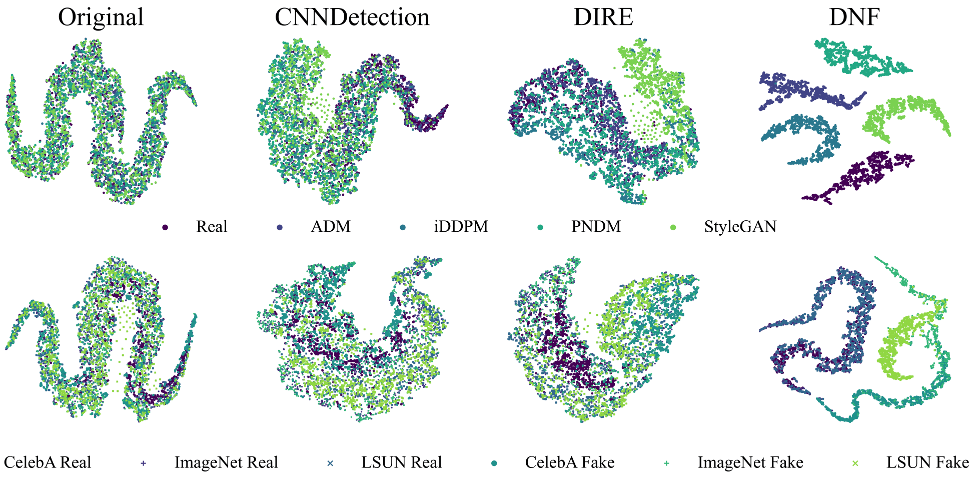

Fig. 2 Classifier's Latent Representations Visualization. We visualize the latent representations learned by different classifiers using t-SNE (Van der Maaten and Hinton, 2008). Compared to original images, CNNDetection (Wang et al., 2020), and DIRE (Wang et al., 2023), classifiers trained on DNF achieve more separable latent representations. These representations not only distinguish between real and fake images but also effectively classify images generated by different models.

DNF classifier demonstrates **excellent robustness**, achieving an accuracy of over 99.2% in generated image detection even when the images undergo common perturbations such as Gaussian blur and JPEG compression during network transmission. (iii) The DNF classifier exhibits **strong cross-dataset generalization capabilities**. For example, when trained on LSUN-Bedroom (Yu et al., 2015), ImageNet (Deng et al., 2009), or CelebA (Liu et al., 2018), and tested on the other two datasets, the classifier demonstrates significantly higher accuracy compared to other detection methods. (iv) The DNF classifier possesses **remarkable cross-generator generalization capabilities**, achieving high-accuracy detection of images generated by a wide variety of generators after being trained on just a few generators. This characteristic holds true even across generators with different principles, such as GANs and DMs. Our main contributions are as follows:

- We introduce DNF, pioneering the use of estimated noise from the inverse diffusion process to construct an image representation for generated image detection.
- We conducted comprehensive experiments to prove DNF classifier achieves state-of-the-art performance in generated image detection, significantly outperforming existing methods.
- We propose a new real-world evaluation pipeline, with a particular emphasis on robustness, cross-dataset and cross-generator generalization capabilities. Our strategy shows outstanding performance in these aspects.

2 Related Work

2.1 Generative AI

Generative Adversarial Networks (GANs) GANs (Goodfellow et al., 2020) are a class of unsupervised machine learning frameworks to generate more realistic images. Unconditional GANs such as BigGAN (Brock, 2018) and StyleGANs (Karras et al., 2019, 2020, 2021) learn the latent distribution of real samples and generate high-quality images by randomly sampling from the learned latent space. Conditional GANs can use input images as conditional constraints to shape the generated images. Models like CycleGAN (Zhu et al., 2017) and StarGAN (Choi et al., 2018) are designed with this objective in mind, enabling tasks such as image translation and style transfer. Text-conditioned GANs require only a textual prompt to specify the content and style of the generated images, represented by GALIP (Tao et al., 2023) and GigaGAN (Kang et al., 2023).

Diffusion Models (DMs) DMs (Gu et al., 2022; Phung et al., 2023; Li et al., 2023; Peebles and Xie, 2023), with their remarkable image generation performance, have demonstrated the potential to become the next generation of generative models. DDPM (Ho et al., 2020) achieves high-quality image generation by injecting noise into images during the inverse diffusion process and learning how to reconstruct the original image during diffusion process. DDIM (Song et al., 2020) proposes using a method of deterministic Markov chains to reduce the number of diffusion steps, thus accelerating the image generation speed. ADM (Dhariwal and Nichol, 2021), for the first time, surpasses GANs in image generation by introducing classifier guidance to enhance the quality of generated images. Models such as Stable Diffusions (Rombach et al., 2022) and Midjourney (Midjourney, 2022), as text-to-image diffusion models, have achieved remarkable image generation capabilities through executing diffusion processes in latent space and training on large-scale datasets.

2.2 Generated Image Detection

To mitigate potential risks associated with generated images, researchers are gradually paying attention to generated image detection (Sinitsa and Fried, 2023; Tan et al., 2023; Chai et al., 2020; Corvi et al., 2023; Cao et al., 2022). CNNDetection (Wang et al., 2020) has discovered artifacts in the frequency domain of CNN-generated images, making detection of generated images feasible. It has constructed the first universal CNN-generated image detector through post-processing of images. Similarly, FrequencyDetection (Qian et al., 2020) classifies generated and real images by observing features presented after discrete cosine transformation. DisGRL (Shi et al., 2023) incorporates three proposed components to learn both forgery-sensitive and genuine compact visual patterns. DIRE (Wang et al., 2023) utilizes the Diffusion Model to reconstruct images and observes the differences between the original and reconstructed images for image detection. However, these methods struggle to cope with the increasingly evolving generative models, exhibiting significant deficiencies in cross-dataset and cross-generator detection capabilities, whereas our approach addresses this critical shortfall.

3 Method

In this section, we will first provide a brief introduction to the relevant background of the Diffusion Model, and then give the implementation details of Diffusion Noise Feature.

3.1 Preliminaries

Given an initial data distribution $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, DDPM (Ho et al., 2020) first uses a manually designed Markov chain to invert the data to a noise distribution according to Equation 1. Then the model is trained to learn a Markov chain in Equation 2 to gradually restore noisy images to their original state.

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad (1)$$

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (2)$$

β is a hyper-parameter that controls the manner noise is added in the inverse diffusion process. θ represents the parameters learned by the model during the training process. μ_θ and $\mathbf{\Sigma}_\theta$ denote the mean and covariance decided by the model during the diffusion process.

DDIM (Song et al., 2020), allowing the process to be significantly accelerated without the Markov assumption, can sample $\mathbf{x}_{t-1}$ from $\mathbf{x}_t$ via

$$\mathbf{x}_{t-1} = \sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(\mathbf{x}_t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\epsilon_\theta^t(\mathbf{x}_t) + \sigma_t\epsilon_t. \quad (3)$$

$\epsilon_t \sim \mathcal{N}(0, \mathbf{I})$ is standard Gaussian noise independent of $\mathbf{x}_t$. $\epsilon_\theta^t(\mathbf{x}_t)$ represents the estimated noise generated by the model at time step t and α is a hyper-parameter that controls the diffusion process. In fact, σ_t controls the entire diffusion process. For example, when $\sigma_t = \sqrt{(1-\alpha_{t-1})/(1-\alpha_t)}\sqrt{1-\alpha_t/\alpha_{t-1}}$ for all t, the process in Equation 3 represents the diffusion process in DDPM. Let $\sigma_t = 0$, making the forward process in DDIM determined by the given $\mathbf{x}_t$ and $\mathbf{x}_0$.

Assuming T is the total number of steps required in DDPM, when T is sufficiently large (*e.g.*, 1000), Equation 3 be interpreted as an Euler method for integrating ordinary differential equations (ODEs) akin to

$$\frac{\mathbf{x}_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{\mathbf{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t-\Delta t}}{\alpha_{t-\Delta t}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right) \epsilon_\theta^t(\mathbf{x}_t). \tag{4}$$

Defining $\gamma = \sqrt{(1-\alpha)/\alpha}$, $\bar{\mathbf{x}} = \mathbf{x}/\sqrt{\alpha}$, the corresponding ODE in Equation 4 is then reformulated as

$$\mathrm{d}\bar{\mathbf{x}} = \epsilon_\theta^t(\frac{\bar{\mathbf{x}}(t)}{\sqrt{\gamma^2+1}}, t)\mathrm{d}\gamma(t). \tag{5}$$

Now the inverse diffusion process, which progress from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$, can be rewritten as

$$\frac{\mathbf{x}_{t+1}}{\sqrt{\alpha_{t+1}}} = \frac{\mathbf{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t+1}}{\alpha_{t+1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right) \epsilon_\theta^t(\mathbf{x}_t). \tag{6}$$

To enhance the computational efficiency of this equation, the Denoising Diffusion Implicit Model (DDIM) employs a strategy of subsampling. It selects a strategic subsequence $\{\tau_0, \tau_1, \ldots, \tau_S\}$ from the comprehensive sequence $\{0, 1, \ldots, T\}$. Here, S denotes the reduced total number of steps necessary for the optimized forward process, which could be, for instance, 20 steps. By employing this approach, Equation 6 adeptly reformulated as:

$$\frac{\mathbf{x}_{\tau_{t+1}}}{\sqrt{\alpha_{\tau_{t+1}}}} = \frac{\mathbf{x}_{\tau_t}}{\sqrt{\alpha_{\tau_t}}} + \left(\sqrt{\frac{1-\alpha_{\tau_{t+1}}}{\alpha_{\tau_{t+1}}}} - \sqrt{\frac{1-\alpha_{\tau_t}}{\alpha_{\tau_t}}}\right) \epsilon_\theta^{\tau_t}(\mathbf{x}_{\tau_t}). \tag{7}$$

By selecting the subsequence $\{\tau_i\}$ appropriately, we can significantly expedite the entire diffusion process. Our designed Diffusion Noise Feature (DNF) is derived from the estimated noise sequence $\{\epsilon_\theta^{\tau_t}(\mathbf{x}_{\tau_t})\}$.

3.2 Diffusion Noise Feature

Shortcomings of previous methods Given an initial real image distribution $p_r(\mathbf{x})$ and a generated image distribution $p_g(\mathbf{x})$, the advanced generative capability of existing models allows $p_g(\mathbf{x})$ to closely approximate with $p_r(\mathbf{x})$, thereby complicating the task of discerning samples from these two distributions. Prior approaches (Frank et al., 2020; Wang et al., 2020) have primarily focused on identifying subtle distinctions between $p_g(\mathbf{x})$ and $p_r(\mathbf{x})$ using methods like frequency domain analysis which encounter limitations when confronted with the realistic images generated by DM-based generators (Dhariwal and Nichol, 2021; Rombach et al., 2022).

Goal Our goal is to design a innovative image representation that recasts the original image distribution, $p_g(\mathbf{x})$ and $p_r(\mathbf{x})$, into distinct new distributions, $p'_g(\mathbf{x})$ and $p'_r(\mathbf{x})$. These new distributions should be readily distinguishable, facilitating the classification of samples by various detectors. In essence, this novel representation is intended to the subtle differences between generated images and real images, thereby assisting the detectors to accurately categorize samples as either generated or real.

Motivation In our research, we have discerned a pronounced divergence in the characteristics of images originating from disparate distributions when undergoing the identical inverse diffusion process according to Equation 6. This contrast is specifically evident in the manifestation of the estimated noise sequence $\{\epsilon_\theta^{\tau_t}\}$. As shown in Figure 1, the patterns of $\{\epsilon_\theta^{\tau_t}\}$ corresponding to generated and real images are entirely distinct, with even images from different generators showing different characteristics in $\{\epsilon_\theta^{\tau_t}\}$.

The rationale behind this lies in the diffusion models' inherent tendency to consolidate samples from diverse distributions into a unified distribution,thereby exaggerating the nuanced discrepancies in the fine details among various samples. Our subsequent experiments have demonstrated that this phenomenon is not influenced by factors such as image content, image resolution, or the number of iterations within the inverse diffusion process. The manifestation of estimated noise sequence $\{\epsilon_\theta^{\tau_t}\}$ corresponding to images is strongly correlated with the source of image generation. Therefore, this observation can be harnessed to devise an image representation that is conducive to the detection of generated images.

Implementation Suppose we have an image $\mathbf{x}_0$ that needs to be detected, a pre-trained diffusion model $\mathcal{F}_\theta$ with parameter θ and a time step sequence $\{\tau_0, \tau_1, \ldots, \tau_S\}$ sampled from $\{0, 1, \ldots, T\}$. We can obtain the estimated noise sequence $\{\epsilon_\theta^{\tau_t}\}$ by inputting $\mathbf{x}_0$ into the diffusion model $\mathcal{F}_\theta$ and executing the inverse diffusion process in Equation 7. In order to obtain a usable DNF from the estimated noise sequence $\{\epsilon_\theta^{\tau_t}\}$ for detection, we need to design a fusion strategy $\mathcal{G}$. Since the estimated noise $\epsilon_\theta^{\tau_i}$ at different time step τ_i during the diffusion process may contain

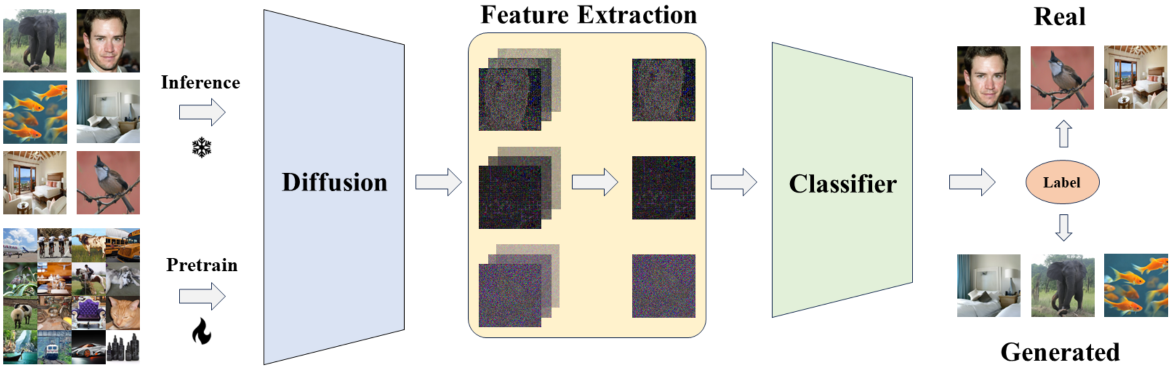

Fig. 3 **Illustration of Generated Image Detection based on DNF.** We leverage a Diffusion Model $\mathcal{F}_\theta$ pre-trained on a large-scale dataset to perform the inverse diffusion process on the image under detection. By employing a fusion strategy $\mathcal{G}$, we convert the collected estimated noises $\{\epsilon_\theta^{\tau_t}\}$ into a classifiable DNF, which is then utilized to train the classifier.

different information, different fusion strategies may impact detection accuracy. A comprehensive discussion on various fusion strategies will be provided later. In this section, we define the default fusion strategy $\mathcal{G}$ to extract the first sample $\epsilon_\theta^{\tau_0}$ from the estimated noise sequence $\{\epsilon_\theta^{\tau_t}\}$ as the DNF for detection.

$$\mathbf{DNF}(\mathbf{x}_0) = \mathcal{G}(\{\mathcal{F}_\theta(\mathbf{x}_i), t_i\}), \; i \in \{0, \ldots, S\}. \quad (8)$$

Upon completion of the DNF computation, the resulting DNF can be harnessed to train a detector for generated images. This implementation is shown in Equation 8 and explained in detail in Figure 3.

4 Experiments

Our experimental framework encompasses four primary components: comparative experiments with existing methods, generalization capability evaluation, perturbation robustness evaluation, and ablation studies. Check the supplementary file for more experimental results and details.

4.1 Experimental Setup

Datasets To ensure a fair comparison with these methods, we conducted our experiments on three widely used datasets, DiffusionForensics (Wang et al., 2023), CNNSpot (Wang et al., 2020) and GenImage (Zhu et al., 2024), which contain authentic images from sources such as ImageNet (Deng et al., 2009), LSUN-Bedroom (Yu et al., 2015), and CelebA (Liu et al., 2018), as well as images generated by various types of generative models, *e.g.*, GANs, DMs.

Baselines We select CNNDetection, SBI, PatchForensics, F3Net, and DIRE as baselines, all widely recognized generated image detection methods, covering approaches like image post-processing, frequency domain detection, and image reconstruction.

CNNDetection (Wang et al., 2020) introduced a detection model trained to distinguish images generated by CNN-base models. However, this model exhibits strong generalization ability, meaning it can effectively detect images generated by various CNN models, not just the one it was trained on.

SBI (Shiohara and Yamasaki, 2022) is a method applied for DeepFake detection. It trains a universal generated image detector by blending fake source images with target images derived from a single original image. This approach enables the detector to learn and recognize synthetic images, regardless of the specific source or target images used in the blending process.

Patchforensics (Chai et al., 2020) utilizes a patch-wise classifier, which has been reported to outperform simple classifiers in detecting fake images. Instead of analyzing entire images, Patchforensics focuses on examining smaller patches within an image to identify inconsistencies or anomalies that indicate image manipulation or forgery.

F3Net (Qian et al., 2020) emphasizes the significance of frequency information in detecting generated images. By analyzing the frequency components of an image, F3Net can identify discrepancies or irregularities that are indicative of image tampering or generation.

DIRE (Wang et al., 2023) utilizes diffusion models to reconstruct images and uses the difference between the original image and the reconstructed image as the feature for classification. By comparing the differences between the features of real images and generated images, excellent performance in generated image detection can be achieved.

Training Details Before training, we preprocess the images, which consists of two parts: calculating DNF and data augmentation. We rescale all images to a uniform resolution of 256×256 to ensure that the Diffusion Model pretrained on LSUN-Bedroom can correctly compute the DNF for each image. Before training, we also perform horizontal flipping with a probability of 50%. During training, we adopt a training setup similar to CNNDetection and DIRE, using ResNet50 (He et al., 2016) as the classifier. We use the Adam (Kingma and Ba, 2015) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, a batch size of 64, and an initial learning rate of 10^{-4}. The learning rate is reduced by a factor of 10 if after 5 epochs the validation accuracy does not increase by 0.1%, and we terminate training when the learning rate reaches 10^{-6}.

Evaluation metrics. We primarily evaluate our model using two metrics: *Accuracy* and *Average Precision*. These are two commonly used and effective metrics in generated image detection (Wang et al., 2020, 2023).

4.2 Baseline Comparison

We evaluated the performance of the DNF classifier and other baselines in generated image detection on DiffusionForensics (Wang et al., 2023). When evaluating the baselines, we made every effort to use their officially released code and model parameters for testing. Additionally, we retrained CNNDetection (Wang et al., 2020), Patchforensics (Chai et al., 2020), and F3Net (Frank et al., 2020) with the same settings as the DNF classifier to demonstrate DNF classifier's superior performance compared to these methods when using the same dataset. We conducted comprehensive evaluations on the LSUN-Bedroom Split of DiffusionForensics and present the results in Table 1.

Our experiments indicates that traditional generated image detection methods represented by CNNDetection, Patchforensics, and SBI (Shiohara and Yamasaki, 2022) cannot effectively detect images generated by diffusion models. While they perform well in detecting images generated by GANs, they exhibit a significant performance decline when faced with images from these previously unseen types of generators. Their average accuracy stands at 55.8%, with an average precision of 71.7%.

After retraining CNNDetection, F3Net, and Patchforensics on the LSUN-Bedroom Split of DiffusionForensics, we found that these methods indeed exhibit good detection performance for images generated by diffusion models, achieving average accuracy of 86.0% and average precision of 94.1%. However, this detection performance seems to not generalize to generator categories unseen during training. While achieving average accuracy of 94.8% and average precision of 99.2% on seen generators, they only achieve average accuracy of 82.2% and average precision of 91.9% on unseen generators.

The most outstanding performance among the baselines is achieved by DIRE, reaching an accuracy of 92.6% and a precision of 99.4%, and it can generalize detection capability to the vast majority of generators. Our method, the detector trained on DNF achieves remarkably impressive performance, surpassing all previous methods on this dataset, achieving 99.8% accuracy and 99.9% average precision.

4.3 Generalization Capability

In generalization capability evaluation, we selected the best-performing CNNDetection (Wang et al., 2020) after retraining and the overall best-performing DIRE (Wang et al., 2023) to conduct a generalization evaluation experiment with our DNF. In this experiment, each method will be retrained on three training splits of DiffusionForensics (Wang et al., 2023) and CNNSpot (Wang et al., 2020) and tested on

Table 1 **Comparison to Existing Methods.** We evaluate the performance of the baseline and our proposed method for generated image detection on the LSUN-Bedroom split of the DiffusionForensics dataset. "*" indicates that the method was retrained on the training set, and "†" denotes that images generated by the model were included in the training data. We report Acc (%) / AP (%) as the metrics.

Method	Testing Generators										Total
	ADM†	DDPM	iDDPM†	LDM	PNDM†	SD-v2	VQ-D	DALL-E 2	IF	Midjourney	Avg.
CNNDet	50.1/63.5	50.2/79.4	50.2/78.0	50.1/61.4	50.1/60.3	50.8/80.7	50.1/70.8	52.8/87.4	51.3/79.9	50.9/58.5	50.6/71.9
Patchfor	50.2/67.4	53.2/74.2	51.2/63.4	56.7/89.1	56.5/72.4	54.2/72.7	87.2/95.4	50.1/68.9	50.0/56.3	56.1/57.2	56.5/71.7
SBI	53.4/60.8	56.9/50.8	58.4/56.2	83.4/90.2	73.1/95.6	59.2/70.9	56.2/74.2	51.2/56.4	61.3/72.3	52.3/87.9	60.5/71.5
DIRE	94.7/99.7	92.6/99.6	94.6/99.7	94.6/99.5	94.3/99.1	94.6/99.7	94.6/99.8	89.5/99.5	94.6/99.7	82.1/98.0	92.6/99.4
F3Net*	91.2/97.8	90.7/98.5	89.9/99.2	98.1/**100**	92.3/97.2	81.1/90.4	92.4/97.3	78.1/86.2	73.6/82.2	75.9/81.1	86.3/92.9
Patchfor*	94.1/99.8	72.9/98.2	95.2/99.4	97.2/**100**	94.2/**100**	74.5/90.2	95.4/100	85.2/98.2	65.4/82.3	53.2/88.6	83.7/95.7
CNNDet*	98.8/99.9	98.5/99.9	99.1/99.9	97.9/99.8	99.1/99.9	80.4/93.5	78.8/94.6	94.5/98.5	80.3/94.0	53.4/58.1	88.1/93.8
DNF (Ours)	**100/100**	99.7/**100**	**100/100**	**100/100**	**100/100**	**100/100**	99.8/**100**	**100/100**	99.9/**100**	98.9/99.9	**99.8/99.9**

Table 2 **Generalization Capability Evaluation - I.** We evaluate the generalization capability of the baseline and our proposed method on the ImageNet and CelebA split of the DiffusionForensics dataset and GenImage. "†" denotes that images generated by the model were included in the training data. We report Acc (%) / AP (%) as the metrics.

Method	Training	DF ImageNet		GenImage				DF CelebA				Total
	Dataset	ADM†	SD-v1	SD-v1.4	SD-v1.5	Glide	wukong	SD-v2†	Mid.	DALL-E 2	IF	Avg.
CNNDet	LSUN-B.	63.6/80.6	53.3/63.8	52.8/55.0	53.0/56.0	78.3/88.1	50.8/51.8	12.9/9.8	11.8/7.7	49.0/49.4	12.8/9.6	43.8/47.2
	ImageNet	71.6/79.8	51.0/51.2	41.3/40.9	40.6/40.5	60.5/63.4	45.9/48.9	37.0/41.6	48.4/49.1	54.2/52.2	36.5/41.2	48.7/50.9
	CelebA	51.0/58.8	52.6/68.0	51.1/50.3	52.9/57.5	50.5/50.0	53.1/57.1	78.4/69.9	73.6/67.7	54.2/52.2	53.6/53.9	57.1/58.5
	CNNSpot	51.2/82.0	50.5/69.5	50.4/59.4	50.6/60.1	52.4/68.6	50.6/59.0	52.8/87.4	54.9/90.1	53.8/87.9	50.3/61.3	51.8/72.5
DIRE	LSUN-B.	99.8/99.8	**99.1**/99.9	91.2/98.6	91.6/98.8	92.4/99.5	90.1/98.3	49.9/49.9	50.4/50.2	50.4/50.2	50.3/50.2	76.5/79.5
	ImageNet	99.8/99.9	98.2/99.9	95.4/99.7	96.3/99.9	67.2/73.1	52.8/63.8	50.0/50.0	50.0/50.0	50.0/50.0	50.0/50.0	71.0/73.6
	CelebA	99.8/99.9	58.2/66.2	53.4/62.1	55.8/67.8	63.1/71.5	66.8/78.8	96.7/**100**	95.0/**100**	93.4/**100**	96.8/**100**	77.9/84.6
	CNNSpot	72.8/83.4	50.1/50.1	51.2/53.6	49.8/50.1	73.4/76.8	58.6/61.2	50.1/50.2	58.2/62.9	67.2/75.3	52.1/53.3	58.4/61.7
DNF	LSUN-B.	98.0/**100**	96.3/**100**	98.6/99.9	98.6/99.9	99.9/**100**	99.7/**100**	75.5/99.8	97.5/99.9	**100/100**	**100/100**	96.4/**99.9**
	ImageNet	**100/100**	98.9/99.9	**100/100**	**100/100**	**100/100**	**100/100**	98.7/**100**	99.0/**100**	**100/100**	**100/100**	**99.9/99.9**
	CelebA	**100/100**	98.9/**100**	99.7/99.9	99.8/**100**	99.7/99.9	99.8/**100**	**100/100**	**100/100**	**100/100**	**100/100**	99.7/**99.9**
	CNNSpot	86.9/**100**	77.7/**100**	77.5/99.1	77.8/99.1	79.2/96.6	80.3/98.7	60.6/99.1	86.1/99.7	85.0/99.7	75.8/99.6	78.7/99.5

five test set from DiffusionForensics, CNNSpot and GenImage (Zhu et al., 2024) to assess the methods' cross-dataset and cross-generator generalization capabilities. The evaluation results across multiple datasets can be found in the Table 2, Table 3 and Table 4.

Cross-dataset Generalization We found that when the training and testing datasets are from the same source, all three methods demonstrate good detection performance. Taking DIRE as an example, the DIRE detector trained on the LSUN-Bedroom Split or CelebA Split achieves accuracies of 92.6% and 95.4% respectively on the corresponding test sets. However, when tested on a test set from a different source than the training set, the accuracy of DIRE in cross-validation between LSUN-Bedroom Split and CelebA Split drops to 50.2% and 67.5% respectively. In comparison, DNF achieves average accuracy of 96.2% and average precision of 99.8% in all cross-dataset generalization tests, demonstrating superior cross-dataset generalization performance compared to other methods. Especially, the DNF detector trained on the ImageNet Split achieves an average precision and accuracy of 99.9% across the five test sets. The DNF detector trained on CNNSpot, which exhibits poorer generalization performance. Considering that this training set only contains images generated by ProGAN and real images from ImageNet, it is indeed quite challenging for the model trained on CNNSpot to generalize to other datasets.

Cross-generator Generalization Existing detection methods often successfully identify the generator of the training set images. CNNDetection trained on CNNSpot can detect images generated by ProGAN (Karras, 2017) with 100% accuracy, but when detecting other generators such as StyleGAN (Karras et al., 2019), the accuracy drops to 66.3%. DIRE trained on the ImageNet Split achieves a 99.8% accuracy in detecting images generated by ADM (Dhariwal and Nichol, 2021), and 96.6% accuracy in Stable Diffusions, but when detecting images from other generators, *e.g.*, Glide, wukong, it only achieves an 60.0% accuracy. Meanwhile, DNF

Table 3 Generalization Capability Evaluation - II. We evaluate the generalization capability of the baseline and our proposed method on the LSUN split of the DiffusionForensics dataset. ”†” denotes that images generated by the model were included in the training data. We report Acc (%) / AP (%) as the metrics.

Method	Training				LSUN-Bedroom Split						Total
	Dataset	ADM†	DDPM	iDDPM†	LDM	PNDM†	SD-v2	VQ-D	DALL-E 2	IF	Mid.	Avg.
CNNDet	LSUN-B.	98.8/99.9	98.5/99.9	99.1/99.9	97.9/99.8	99.1/99.9	80.4/93.5	78.8/94.6	94.5/98.5	80.3/94.0	53.4/58.1	88.1/93.8
	ImageNet	72.4/74.1	71.2/65.7	76.8/80.8	64.0/60.1	76.7/85.8	67.4/61.3	78.4/93.1	77.2/80.4	72.1/69.1	70.1/73.8	72.6/81.8
	CelebA	55.1/63.3	49.1/48.3	51.9/69.0	56.6/64.8	45.9/34.0	83.7/92.9	52.1/60.9	50.0/51.3	55.1/69.0	50.9/60.3	55.0/61.4
	CNNSpot	50.1/63.5	50.2/79.4	50.2/78.0	50.1/61.4	50.1/60.3	50.8/80.7	50.1/70.8	52.8/87.4	51.3/79.9	50.9/58.5	50.6/71.9
DIRE	LSUN-B.	94.7/99.7	92.6/99.6	94.6/99.7	94.6/99.5	94.3/99.1	94.6/99.7	94.6/99.8	89.5/99.5	94.6/99.7	82.1/98.0	92.6/99.4
	ImageNet	60.2/91.3	54.9/86.8	60.3/91.7	57.9/89.1	57.6/79.6	58.9/90.5	57.5/94.0	40.6/64.2	47.6/66.1	28.2/61.3	52.3/81.5
	CelebA	67.8/82.7	62.6/62.9	62.4/67.1	75.3/97.9	57.4/68.0	74.3/93.0	75.2/95.0	67.1/93.7	78.3/97.0	54.8/39.9	67.5/79.7
	CNNSpot	74.8/86.9	72.3/86.3	65.4/81.3	66.1/75.2	52.1/56.8	50.1/52.1	55.4/58.9	72.9/78.3	53.6/65.2	61.3/64.9	62.4/70.6
DNF	LSUN-B.	**100**/**100**	99.7/**100**	**100**/**100**	**100**/**100**	**100**/**100**	**100**/**100**	99.8/**100**	**100**/**100**	99.9/**100**	98.9/99.9	99.8/**99.9**
	ImageNet	**100**/**100**	**100**/**100**	**100**/**100**	99.9/**100**	**100**/**100**	**100**/**100**	**100**/**100**	**100**/**100**	98.8/99.2	**100**/**100**	**99.9**/**99.9**
	CelebA	**100**/**100**	99.2/**100**	**100**/**100**	99.2/99.9	**100**/**100**	**100**/**100**	**100**/**100**	98.6/99.9	**100**/**100**	98.1/99.3	99.5/**99.9**
	CNNSpot	99.7/99.7	97.7/99.7	99.9/99.7	99.9/99.7	97.7/99.7	82.6/99.5	99.9/99.7	90.9/99.6	97.5/99.7	99.7/99.7	96.5/99.7

Table 4 Generalization Capability Evaluation - III. We evaluate the generalization capability of the baseline and our proposed method on the CNNDetection dataset. ”†” denotes that images generated by the model were included in the training data. We report Acc (%) / AP (%) as the metrics.

Method	Training					CNNSpot						Total
	Dataset	ProGAN†	StyleGAN	StyleGAN2	StarGAN3	BigGAN	CycleGAN	GuaGAN	StarGAN	ProjGAN	Diff-ProjGAN	Avg.
CNNDet	LSUN-B.	68.6/89.7	71.1/88.5	65.8/83.2	97.9/99.8	58.3/89.7	54.9/60.3	64.8/74.4	75.5/86.1	74.1/91.4	68.3/88.6	69.9/85.2
	ImageNet	84.4/92.8	82.8/88.8	84.3/89.2	50.1/61.4	80.3/86.4	57.4/53.9	75.3/84.1	94.5/98.8	63.3/62.2	59.2/56.8	73.2/77.4
	CelebA	50.3/51.6	54.3/62.2	53.7/70.4	97.9/99.8	51.1/53.2	50.0/48.4	52.3/59.3	50.0/44.0	51.8/52.8	52.4/55.3	56.4/59.7
	CNNSpot	**100**/**100**	73.4/98.5	68.4/97.9	50.1/61.4	59.0/88.2	80.7/96.8	79.2/98.1	80.9/95.4	52.8/90.0	52.0/88.3	69.7/91.5
DIRE	LSUN-B.	52.8/58.8	51.1/56.7	51.7/58.0	84.6/99.6	49.7/46.9	49.6/50.1	51.3/47.4	47.8/40.7	84.6/99.6	84.6/99.5	60.8/65.7
	ImageNet	51.6/56.2	52.3/58.9	50.1/50.3	67.5/78.9	66.9/73.2	53.3/60.1	51.2/65.8	88.2/95.7	56.2/62.1	54.9/60.2	59.2/66.1
	CelebA	62.1/75.2	66.3/69.3	50.1/56.2	53.2/62.1	52.1/53.2	56.8/52.1	51.3/56.3	52.1/56.3	63.2/71.2	66.6/73.2	57.4/62.5
	CNNSpot	95.2/99.3	82.5/93.2	74.8/88.9	82.1/91.2	72.1/78.9	72.9/80.1	65.8/73.5	96.7/99.6	67.2/76.9	67.8/76.8	77.7/85.8
DNF	LSUN-B.	99.9/**100**	99.3/**100**	97.8/**100**	99.3/**100**	**100**/**100**	**100**/**100**	**100**/**100**	**100**/**100**	99.9/**100**	99.9/**100**	99.6/**100**
	ImageNet	**100**/**100**	98.6/99.7	99.9/**100**	99.9/**100**	**100**/**100**	**100**/**100**	99.9/**100**	**100**/**100**	**100**/**100**	**100**/**100**	**99.8**/99.9
	CelebA	**100**/**100**	**100**/**100**	99.8/**100**	**100**/**100**	99.9/**100**	**100**/**100**	98.3/**100**	**100**/**100**	**100**/**100**	**100**/**100**	**99.8**/**100**
	CNNSpot	**100**/**100**	99.6/**100**	97.2/**100**	97.7/99.7	90.5/**100**	78.8/**100**	85.5/**100**	**100**/**100**	99.8/99.7	99.8/99.7	94.9/99.9

achieves detection accuracies for unseen generators of 96.2%, 99.8%, 99.6%, and 78.7% across these three test sets. Crucially, according to the comprehensive result in supplementary material, DNF can even generalize between generators with different principles.

4.4 Perturbation Robustness

When images are shared on social networks, they often undergo perturbations such as Gaussian blur or JPEG compression, resulting in the loss of image details. This loss significantly impacts the performance of generated image detection. To evaluate the robustness of various methods under such perturbations, we designed robustness experiments using images from the LSUN-Bedroom split of the DiffusionForensics.

We applied varying degrees of perturbation to the images, including Gaussian blur with $\sigma \in \{0, 1, 2, 3\}$ and JPEG compression with $\mathit{Quality} \in \{100, 65, 30\}$, to explore the performance fluctuations of different methods as perturbation intensity increases. In Figure 4, as the perturbation intensifies, the detailed information in the images becomes less distinct, making detection more challenging. Methods like PatchForensics (Chai et al., 2020) and F3Net (Qian et al., 2020) demonstrate robustness to specific types of perturbations due to their design principles. In contrast, DNF, with its inherent ability to enhance image details, ensures that even weakened details remain effective for image generation detection. As a result, DNF showed minimal performance degradation, consistently achieving detection accuracy above 99.2%.

Additionally, we assessed the robustness of the detector against disturbances such as resizing, cropping, and rotation to evaluate their impact on DNF's performance. We present the results in Table 5. Since these transformations are common

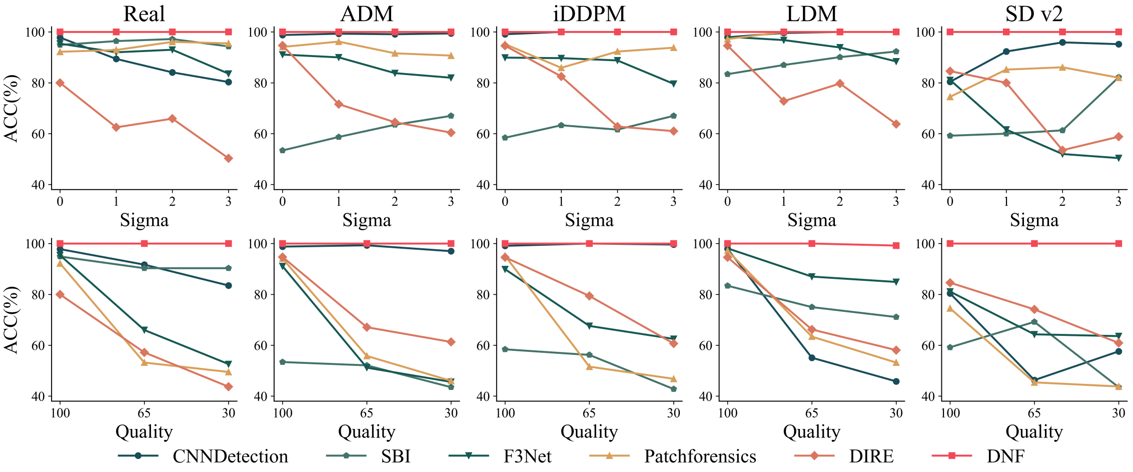

Fig. 4 Perturbation Robustness Evaluation - I. The perturbations added to the images include Gaussian blur with $\sigma \in \{0, 1, 2, 3\}$ and JPEG compression with $Quality \in \{100, 65, 30\}$. DNF demonstrates excellent resistance to perturbations, consistently maintaining an accuracy over 99.2%. For more types of perturbations, please refer to the supplementary materials.

Table 5 Perturbation Robustness Evaluation - II. Before resizing the images to a uniform size for calculating DNF, we perform the perturbation on the test images. When encountering significant perturbation such as cropping to 64×64 pixels or resizing to 128×128 pixels, the detection accuracy will significantly decrease. The former is due to severe semantic information loss, while the latter is due to pixel detail loss during the resizing process. Minor disturbances do not cause significant damage to the detector's performance. We report Acc (%) / AP (%) as the metrics.

Perturbation	ADM	iDDPM	LDM	SD v2
None	**100**/**100**	**100**/**100**	**100**/**100**	**100**/**100**
Crop 64	92.1/98.7	91.7/98.8	93.4/99.8	87.2/88.2
Crop 224	99.9/**100**	**100**/**100**	**100**/**100**	**100**/**100**
Resize 128	98.2/99.6	99.9/**100**	95.2/99.9	**100**/**100**
Resize 1024	99.8/**100**	99.9/**100**	**100**/**100**	99.9/**100**
Rotation $\pi/2$	**100**/**100**	99.9/**100**	99.9/**100**	**100**/**100**
Rotation π	99.9/**100**	99.7/**100**	**100**/**100**	99.8/**100**

both in image transmission and data augmentation, they did not significantly degrade most methods' performance. However, excessive cropping of image content caused a noticeable performance drop in DNF.

4.5 Ablation Studies

Effectiveness of Diffusion Model $\mathcal{F}_\theta$. In Equation 8, we require a pre-trained diffusion model $\mathcal{F}_\theta$ to compute estimate noise at specific timesteps. Considering that diffusion models with different architectures pre-trained on various datasets may produce differing noise estimates, we investigate their impact on detection performance. Specifically, we conduct experiments using DDIM (Song et al., 2020) or ADM (Dhariwal and Nichol, 2021) pre-trained on LSUN-Bedroom (Liu et al., 2018) or ImageNet (Deng et al., 2009) datasets to construct DNF. ADM, differing in structure from DDIM, incorporates a classifier guidance mechanism to enhance the quality of generated images. The experimental results are presented in Table 6.

We observed that diffusion models $\mathcal{F}_\theta$ with different architectures, pre-trained on various datasets, did not significantly affect detection performance. However, it is worth noting that, based on the original intention behind designing DNF and issues identified during experiments, we emphasize the importance of ensuring that the diffusion model is sufficiently pre-trained.

Table 6 **Ablation Studies on Effectiveness of Diffusion Model $\mathcal{F}_\theta$.** We evaluate the impact of diffusion model $\mathcal{F}$ and pre-train parameters θ on the performance of our proposed method on the LSUN-Bedroom split of the DiffusionForensics dataset. "†" denotes that images generated by the model were included in the training data. We report Acc (%) / AP (%) as the metrics.

Diffusion	Pretrain	DF LSUN-Bedroom										Total
Model $\mathcal{F}$	Dataset θ	ADM†	DDPM	iDDPM†	LDM	PNDM†	SD-v2	VQ-D	DALL-E 2	IF	Midjourney	Avg.
DDIM	LSUN-B.	**100/100**	99.7/**100**	**100/100**	**100/100**	**100/100**	**100/100**	99.8/**100**	**100/100**	99.9/**100**	98.9/99.9	**99.8/99.9**
	ImageNet	**100/100**	99.2/**100**	**100/100**	**100/100**	**100/100**	99.6/99.9	**100/100**	**100/100**	99.6/**100**	**100/100**	**99.8/99.9**
ADM	LSUN-B.	**100/100**	**100/100**	**100/100**	99.8/**100**	**100/100**	99.1/**100**	**100/100**	98.9/99.4	**100/100**	99.2/99.9	99.7/**99.9**
	ImageNet	**100/100**	**100/100**	**100/100**	97.2/98.9	**100/100**	99.2/99.9	**100/100**	99.8/**100**	99.7/**100**	**100/100**	99.5/99.8

Table 7 **Ablation Studies on Effectiveness of Fusion Strategy $\mathcal{G}$.** We evaluate the impact of fusion strategy $\mathcal{G}$ on the performance of our proposed method. We report ACC (%) / AP (%) as the metrics.

Method	Testing Generators					Avg.
	ADM	iDDPM	LDM	StyleGAN	SD-v1	
$\mathcal{G}_{first}$	**100/100**	**100/100**	**100/100**	99.3/**100**	96.3/**100**	**99.1/100**
$\mathcal{G}_{avg}$	98.2/99.9	99.1/**100**	97.5/**100**	98.2/99.9	**98.2/100**	98.2/99.9
$\mathcal{G}_{last}$	**100/100**	**100/100**	99.2/99.9	96.2/99.4	93.1/98.8	97.7/99.6

A well-converged diffusion model should unify images from diverse data distributions into a single noise distribution. This is crucial as a fully converged diffusion model not only conveys information across different data distributions but also captures high-frequency image details more effectively through the denoising task. The estimated noises generated by such diffusion models can then reflect the unique characteristics of images from various data distributions, meeting the necessary requirements for reliable generated image detection.

Effectiveness of Fusion Strategy $\mathcal{G}$. During the image generation process by the diffusion model, the estimated noise generated at each time step contains different information. The fusion strategy $\mathcal{G}$ determines how to convey this information into the DNF used for detection. We have devised three straightforward fusion tactics: $\mathcal{G}_{first}$ selects the first element from an estimated noise sequence; $\mathcal{G}_{avg}$ takes the average of the entire sequence; and $\mathcal{G}_{last}$ extracts the last element of the sequence. In previous experiments, we commonly used $\mathcal{G}_{first}$ as the default fusion strategy. To investigate the impact of different fusion strategies on DNF detection, corresponding experiments were designed to verify the effects of using different fusion strategies on the same dataset, and the experimental results are presented in Table 7. $\mathcal{G}_{first}$ exhibits the best detection performance, while $\mathcal{G}_{last}$'s detection performance is relatively poorer.

We visualized the estimated noise sequences for a image generated by Stable Diffusion in Figure 5. It can be observed that as the inverse process progresses, the estimated noise gradually approaches pure noise. In our understanding, the estimated noise contains a significant amount of high-frequency information and pixel distribution information of the current image. Therefore, the $\epsilon_\theta^{\tau_0}$ contains more information beneficial for image detection than the subsequent noise estimations. This well explains why, in comparison to $\mathcal{G}_{avg}$ and $\mathcal{G}_{last}$, $\mathcal{G}_{first}$ provides better detection performance.

Of course, this does not mean that subsequent estimated noises are meaningless; it is simply because we are currently using a simple classifier like ResNet50 (He et al., 2016), and the representation in the form of a single image better illustrates the simplicity of DNF. The entire estimated noise sequence $\{\epsilon_\theta^{\tau_i}\}$ contains information from the original image distribution to the pure noise distribution. If we could design a detection structure tailored for the sequence, it should achieve better detection results, which is also the direction of our future efforts.

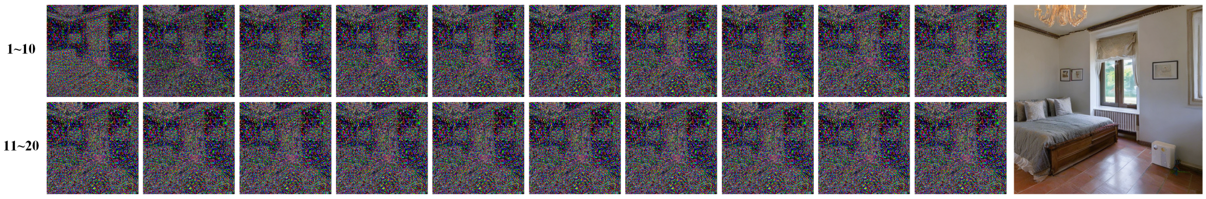

Fig. 5 Estimated Noise Sequences Visualization. We visualized the estimated noise sequence of an image generated during the inverse diffusion process performed by DDIM. As the timesteps increase, the estimated noise progressively approaches pure noise, resulting in performance differences when using different fusion strategies $\mathcal{G}$.

Table 8 Image Format Bias. Tests are conducted using training sets and test sets in different formats. The vertical columns representing the training set formats and the horizontal rows representing the test set formats.

	Original	PNG	JPEG
Original	**98.6/99.9**	93.1/98.2	95.7/99.7
PNG	96.4/**99.9**	**98.9/99.9**	94.3/99.7
JPEG	94.3/99.6	86.9/92.5	**96.1/99.9**

5 Discussion

5.1 Image Format Bias

Considering that our dataset consists of images stored in different formats, classifiers may learn biases introduced by the dataset composition (Grommelt et al., 2024). To demonstrate the impact of this factor on classifier performance, we conducted experiments on the DiffusionForensics dataset. We saved the training set exclusively in either PNG or JPEG format and evaluated performance on test sets saved in both PNG and JPEG formats. The experimental results are presented in Table 8. We observed that the image format bias introduced by the dataset composition does indeed affect the classifier. However, the performance differences were acceptable and did not impact the overall experimental outcomes.

5.2 Frequency Domain Analysis

For the frequency domain analysis, we randomly selected 1,000 images generated by each generator and performed Fourier transforms on them. We then calculated the average of the Fourier transform results. Our findings revealed that for the same image content "Bedroom", there were notable differences for DNF in the frequency domain among images generated from different sources. The separation of real and generated images in the frequency domain helps explain the superiority of using DNF to achieve high-performance detection. The results of the frequency domain analysis have been visualized in Figure 6, with all data processed using logarithmic transformation. It can be observed that the frequency domain spectrum of generated images exhibits periodic grid-like and dot-like features, while the frequency distribution of real images is relatively uniform.

5.3 Speed Advantage

One important point is that $\mathcal{G}_{first}$ can significantly reduce the running time of the pipeline compared to other strategies. For instance, when using a 20-step DDIM, DIRE requires the execution of 40 steps of estimated noise inference, whereas computing DNF$_{first}$ only requires a single step. Ideally, this could achieve a 40-fold acceleration during the dataset processing phase. It is worth noting that compared to training a ResNet50 as a classifier, preprocessing the dataset in diffusion process often takes more time, hence this acceleration is crucial. Furthermore, it is noteworthy that during the detection inference of a small number of images, since only one step of Diffusion Inversion is needed, the Diffusion Backbone, such as U-Net, can be directly connected to our classifier Backbone for integrated deployment, reducing the time spent on inference. This is why DNF is considered a fast method for generating image detection.

5.4 Excellent Performance of DNF

We have demonstrated through extensive experiments the superiority of DNF in the task of generated image detection compared to other methods, achieving higher detection accuracy, stronger robustness, and faster detection speed. Based on the analysis of these experimental results, we

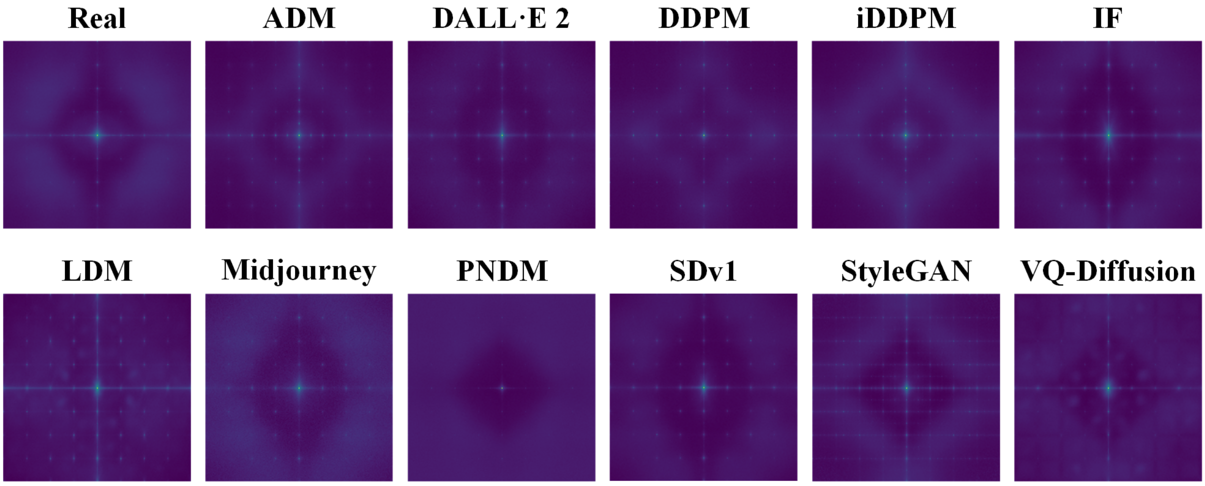

Fig. 6 Frequency Domain Visualization. We visualized the DNF of images with similar content and observed that the frequency domain spectrum of generated images exhibits periodic grid-like and dot-like features, while the frequency distribution of real images is relatively uniform.

believe the high performance primarily stems from the following reasons.

The training process of diffusion models inherently focuses on learning image details. During training, diffusion models repeatedly perform denoising tasks to generate high-quality image details, acquiring prior knowledge of image details, denoted as $\mathcal{P}_m$. In designing DNF, we leverage this prior knowledge $\mathcal{P}_m$ by calculating the estimated noise during the inverse diffusion process. This effectively utilizes the diffusion model's understanding of image details, which is critical for detecting generated images.

Let $D(\cdot,\cdot)$ represent the difference in prior knowledge of image details between two models. We observe that, despite variations in model structures and datasets leading to differences in learned priors $\mathcal{P}$, when the dataset is sufficiently diverse and the model's learning capability is strong, we generally have:

$$D(\mathcal{P}_{real}, \mathcal{P}^i_{gen}) > D(\mathcal{P}^j_{gen}, \mathcal{P}^i_{gen}), \tag{9}$$

where $\mathcal{P}_{real}$ represents the prior knowledge learned from the training dataset, and $\mathcal{P}_{gen}$ denotes the priors learned by different generative models. Our method effectively replaces $\mathcal{P}^j_{gen}$ with the pre-trained diffusion model's prior $\mathcal{P}_m$, leading to:

$$D(\mathcal{P}_{real}, \mathcal{P}^i_{gen}) > D(\mathcal{P}_m, \mathcal{P}^i_{gen}). \tag{10}$$

In practice, this manifests in the form of estimated noise. Real images tend to produce estimated noise with grid-like artifacts, while generated images, although differing across models, generally exhibit similar estimated noise that is distinct from that of real images. This amplifies the differences between real and generated images, enabling high-performance detection of generated images.

6 Future Directions

The current DNF method, while effective, still has limitations. One notable shortcoming is its poor generalization across different styles of images. For example, a model trained on ImageNet struggles to determine whether electronic art paintings are generated. This is mainly due to the limitation of training resources and data. In the future, we will collaborate with more organizations and acquire more data and computation resources to scale the capability of DNF in various scenarios.

Moreover, the utilization of the estimated noise sequence in the current framework can be further improved. There is significant potential for further exploration into the role and application of these subsequent diffusion noises in the context of generated image detection. Understanding and leveraging these noise sequences in more SOTA

AIGC frameworks could uncover new avenues to enhance detection accuracy and robustness.

In addition, with the rise of video generation frameworks and advancements in real-time DeepFake technologies, there is an urgent need for scalable detection frameworks capable of addressing these rapidly evolving challenges in AIGC (AI-Generated Content). Future research should focus on designing adaptable and efficient methods for generated video detection, ensuring that detection capabilities keep pace with the rapid innovations in generative AI.

7 Conclusion

In this paper, we investigate the characteristics of estimated noise generated during the inverse diffusion process for images from different data distributions and, for the first time, utilize estimated noise to design a novel classification feature for detecting generated images. Classifiers trained with this feature demonstrate superior detection performance, stronger generalization, and enhanced robustness against perturbations compared to baseline methods. We hope this approach can inspire new ideas for future AIGC detection methods, mitigating the potential harm caused by the growing prevalence of misinformation.

8 Data Availability Statement

The data that support the results and analysis of this study is publicly available in a repository. The DiffusionForensics dataset is available at https://github.com/ZhendongWang6/DIRE. The CNNSpot dataset is available at https://github.com/PeterWang512/CNNDetection. The GenImage dataset is avaiable at https://genimage-dataset.github.io/

References

Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., *et al.*: Improving image generation with better captions. Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf **2**(3), 8 (2023)

Brock, A.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)

Chai, L., Bau, D., Lim, S.-N., Isola, P.: What makes fake images detectable? understanding properties that generalize. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVI 16, pp. 103–120 (2020). Springer

Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., Choo, J.: Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8789–8797 (2018)

Corvi, R., Cozzolino, D., Zingarini, G., Poggi, G., Nagano, K., Verdoliva, L.: 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 (2023). IEEE

Castillo, C., Mendoza, M., Poblete, B.: Information credibility on twitter. In: Proceedings of the 20th International Conference on World Wide Web, pp. 675–684 (2011)

Cao, J., Ma, C., Yao, T., Chen, S., Ding, S., Yang, X.: End-to-end reconstruction-classification learning for face forgery detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4113–4122 (2022)

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). Ieee

Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)

Fanelli, D.: How many scientists fabricate and falsify research? a systematic review and meta-analysis of survey data. PloS one **4**(5), 5738 (2009)

Frank, J., Eisenhofer, T., Schönherr, L., Fischer, A., Kolossa, D., Holz, T.: Leveraging frequency analysis for deep fake image recognition. In: International Conference on Machine Learning, pp. 3247–3258 (2020). PMLR

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: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10696–10706 (2022)

Giglietto, F., Iannelli, L., Valeriani, A., Rossi, L.: 'fake news' is the invention of a liar: How false information circulates within the hybrid news system. Current sociology **67**(4), 625–642 (2019)

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)

Grommelt, P., Weiss, L., Pfreundt, F.-J., Keuper, J.: Fake or jpeg? revealing common biases in generated image detection datasets. arXiv preprint arXiv:2403.17608 (2024)

Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)

Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. Advances in neural information processing systems **34**, 852–863 (2021)

Karras, T.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017)

Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (2015)

Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410 (2019)

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119 (2020)

Kang, M., Zhu, J.-Y., Zhang, R., Park, J., Shechtman, E., Paris, S., Park, T.: Scaling up gans for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10124–10134 (2023)

Li, X., Liu, Y., Lian, L., Yang, H., Dong, Z., Kang, D., Zhang, S., Keutzer, K.: Q-diffusion: Quantizing diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17535–17545 (2023)

Liu, Z., Luo, P., Wang, X., Tang, X.: Large-scale celebfaces attributes (celeba) dataset. Retrieved August **15**(2018), 11 (2018)

Liu, L., Ren, Y., Lin, Z., Zhao, Z.: Pseudo numerical methods for diffusion models on manifolds. arXiv preprint arXiv:2202.09778 (2022)

Midjourney: Midjourney. https://www.midjourney.com (2022)

Murdoch, B.: Privacy and artificial intelligence: challenges for protecting health information in a new era. BMC Medical Ethics **22**, 1–5 (2021)

Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: International Conference on Machine Learning, pp. 8162–8171 (2021). PMLR

Phung, H., Dao, Q., Tran, A.: Wavelet diffusion models are fast and scalable image generators. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10199–10208 (2023)

Peebles, W., Xie, S.: Scalable diffusion models with transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195–4205 (2023)

Qi, P., Cao, J., Yang, T., Guo, J., Li, J.: Exploiting multi-domain visual information for fake news detection. In: 2019 IEEE International Conference on Data Mining (ICDM), pp. 518–527 (2019). IEEE

Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision, pp. 86–103 (2020). Springer

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: 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)

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

Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: International Conference on Machine Learning, pp. 8821–8831 (2021). Pmlr

Shi, Z., Chen, H., Chen, L., Zhang, D.: Discrepancy-guided reconstruction learning for image forgery detection. arXiv preprint arXiv:2304.13349 (2023)

Sinitsa, S., Fried, O.: Deep image fingerprint: Accurate and low budget synthetic image detector. arXiv preprint arXiv:2303.10762 (2023)

Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

Shiohara, K., Yamasaki, T.: Detecting deepfakes with self-blended images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 18720–18729 (2022)

Tao, M., Bao, B.-K., Tang, H., Xu, C.: Galip: Generative adversarial clips for text-to-image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14214–14223 (2023)

Tan, C., Zhao, Y., Wei, S., Gu, G., Wei, Y.: Learning on gradients: Generalized artifacts representation for gan-generated images detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12105–12114 (2023)

Uyyala, P., Yadav, D.C.: The advanced proprietary ai/ml solution as antifraudtensorlink4cheque (aftl4c) for cheque fraud detection. The International journal of analytical and experimental modal analysis **15**(4), 1914–1921 (2023)

Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)

Wang, Z., Bao, J., Zhou, W., Wang, W., Hu, H., Chen, H., Li, H.: Dire for diffusion-generated image detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 22445–22455 (2023)

Wang, S.-Y., Wang, O., Zhang, R., Owens, A., Efros, A.A.: 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)

Yu, F., Seff, A., Zhang, Y., Song, S., Funkhouser, T., Xiao, J.: Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. arXiv preprint arXiv:1506.03365 (2015)

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. Advances in Neural Information Processing Systems **36** (2024)

Zhu, J.-Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2223–2232 (2017)

Zhong, N., Xu, Y., Qian, Z., Zhang, X.: Rich and poor texture contrast: A simple yet effective approach for ai-generated image detection. arXiv preprint arXiv:2311.12397 (2023)