DIRE for Diffusion-Generated Image Detection

Zhendong Wang[1,*] Jianmin Bao[2,*] Wengang Zhou[1,3,†]
Weilun Wang[1] Hezhen Hu[1] Hong Chen[4] Houqiang Li[1,3,†]

[1] CAS Key Laboratory of GIPAS, EEIS Department, University of Science and Technology of China

[2] Microsoft Research Asia

[3] Institute of Artificial Intelligence, Hefei Comprehensive National Science Center

[4] Merchants Union Consumer Finance Company

{zhendongwang,wwlustc,alexhu}@mail.ustc.edu.cn

jianbao@microsoft.com, {zhwg,lihq}@ustc.edu.cn, chenhong@mucfc.com

Abstract

Diffusion models have shown remarkable success in visual synthesis, but have also raised concerns about potential abuse for malicious purposes. In this paper, we seek to build a detector for telling apart real images from diffusion-generated images. We find that existing detectors struggle to detect images generated by diffusion models, even if we include generated images from a specific diffusion model in their training data. To address this issue, we propose a novel image representation called ***DI****ffusion* ***R****econstruction* ***E****rror (****DIRE****), which measures the error between an input image and its reconstruction counterpart by a pre-trained diffusion model. We observe that diffusion-generated images can be approximately reconstructed by a diffusion model while real images cannot. It provides a hint that DIRE can serve as a bridge to distinguish generated and real images. DIRE provides an effective way to detect images generated by most diffusion models, and it is general for detecting generated images from unseen diffusion models and robust to various perturbations. Furthermore, we establish a comprehensive diffusion-generated benchmark including images generated by various diffusion models to evaluate the performance of diffusion-generated image detectors. Extensive experiments on our collected benchmark demonstrate that DIRE exhibits superiority over previous generated-image detectors. The code, models, and dataset are available at* https://github.com/ZhendongWang6/DIRE.

1. Introduction

Recently, Denoising Diffusion Probabilistic Models (DDPMs) [17, 41] have set up a new paradigm in image generation due to their strong ability to generate high-quality images. There arises plenty of studies [32, 10, 42, 23, 37] exploring the improvement of the network architecture, acceleration of sampling, and so on. As users enjoy the strong generation capability of diffusion models, there are concerns about potential privacy problems. For example, diffusion models may memorize individual images from their training data and emit them at the generation stage [3, 52]. Moreover, some attackers may develop new deepfake techniques based on diffusion models. Therefore, it is an urgent demand for a diffusion-generated image detector.

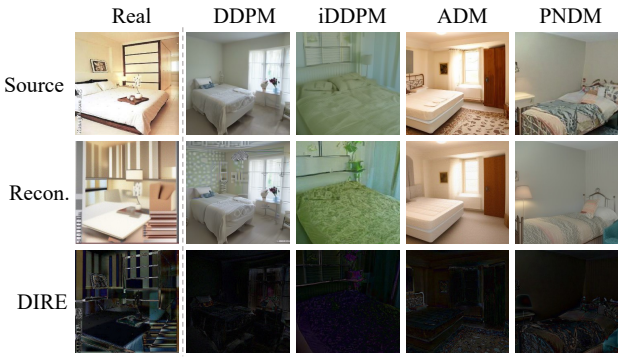


Figure 1: **The DIRE representation** of a real image and four generated images from diffusion models: DDPM [17], iDDPM [32], ADM [10], and PNDM [23], respectively. The DIREs of real images tend to have larger values compared to diffusion-generated images.

Our focus in this work is to develop a general diffusion-generated image detector. We notice that there are various detectors for detecting generated images available. Despite the fact that most diffusion models employ CNNs as the network, the generation processes between diffusion models and previous generators (*e.g.*, GAN, VAE) are entirely different, rendering previously generated image detectors ineffective. A naïve thought is to train a CNN binary classifier on diffusion-generated and real images. However, we find that such a naïve scheme suffers limited generalization to unseen diffusion-generated images.

*Equal contribution.

†Corresponding authors: Wengang Zhou and Houqiang Li.

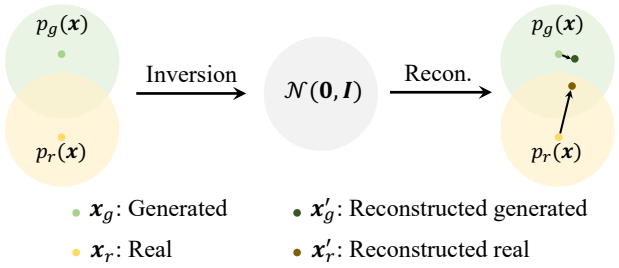


Figure 2: **Illustration of the difference between a real sample and a generated sample from the DIRE perspective.** $p_g(\mathbf{x})$ represents the distribution of generated images while $p_r(x)$ represents the distribution of real images. $\mathbf{x}_g$ and $\mathbf{x}_r$ represent a generated sample and a real sample, respectively. Using the inversion and reconstruction process of DDIM [42], $\mathbf{x}_g$ and $\mathbf{x}_r$ become $\mathbf{x}'_g$ and $\mathbf{x}'_r$, respectively. After the reconstruction, $\mathbf{x}'_r$ is actually within the $p_g(\mathbf{x})$, which leads to a noticeably different DIRE in real samples compared to generated samples.

In this paper, we propose a novel image representation, called **DI**ffusion **RE**construction **E**rror (**DIRE**), for detecting diffusion-generated images. The hypothesis behind DIRE is that images produced by diffusion processes can be reconstructed more accurately by a pre-trained diffusion model compared to real images. The diffusion reconstruction process involves two steps: (1) inverting the input image $\mathbf{x}$ and mapping it to a noise vector $\mathbf{x}_T$ in the noise space $\mathcal{N}(\mathbf{0}, \mathbf{I})$, and (2) reconstructing the image $\mathbf{x}'$ from $\mathbf{x}_T$ using a denoising process. The DIRE is calculated as the difference between $\mathbf{x}$ and $\mathbf{x}'$. As a sample $\mathbf{x}_g$ from the generated distribution $p_g(\mathbf{x})$ and its reconstruction $\mathbf{x}'_g$ belong to the same distribution, the DIRE value for $\mathbf{x}_g$ would be relatively low. Conversely, the reconstruction of a real image $\mathbf{x}_r$ is likely to differ significantly from itself, resulting in a high amplitude in DIRE. This concept is depicted in Figure 2.

DIRE offers a reliable method for differentiating between real and diffusion-generated images. By training a simple binary classifier based on DIRE, it becomes possible to detect diffusion-generated images with ease. The DIRE is general and flexible since it can generalize to images generated by unseen diffusion models during inference time. It only assumes the distinct reconstruction errors of real images and generated ones as shown in Figure 1.

To evaluate the diffusion-generated image detectors, we create a comprehensive diffusion-generated dataset, the *DiffusionForensics* dataset, including three-domain images (LSUN-Bedroom [48], ImageNet [9], and CelebA-HQ [19]) generated by eleven different diffusion models. DiffusionForensics involves unconditional, conditional, and text-to-image diffusion generation models.

Extensive experiments show that the DIRE representation significantly enhances generalization ability. We show that our framework achieves a remarkably high detection accuracy and average precision on generated images from unseen diffusion models, as well as robustness to various perturbations. In comparison with existing generated image detectors, our framework largely exceeds the competitive state-of-the-art methods.

Our main contributions are three-fold as follows.

- We propose a novel image representation called DIRE for detecting diffusion-generated images.
- We set up a new dataset, DiffusionForensics, for benchmarking the diffusion-generated image detectors.
- Extensive experiments demonstrate that the proposed DIRE sets a state-of-the-art performance in diffusion-generated detection.

2. Related Work

Since our focus is to detect diffusion-generated images and the proposed DIRE representation is based on the reconstruction error by a pre-trained diffusion model, we briefly introduce recent diffusion models in image generation and generalizable generated image detection in this section.

2.1. Diffusion Models for Image Generation

Inspired by nonequilibrium thermodynamics [41], Ho *et al*. [17] propose a new generation paradigm, denoising diffusion probabilistic models (DDPMs), which achieves a competitive performance compared to PGGAN [19] on 256 $\times$ 256 LSUN [48]. Since then, more and more researchers turn their attention to diffusion models for improving the architectures [10, 37], accelerating sampling speed [32, 42, 23, 24], exploring downstream tasks [16, 31, 1, 33], and *etc*. Nichol *et al*. [32] find that learning variances of the reverse process in DDPMs can contribute to an order of magnitude fewer sampling steps. Song *et al*. [42] generalize DDPMs via a class of non-Markovian diffusion processes into denoising diffusion implicit models (DDIMs), which leads to more high-quality samples with fewer sampling steps. A later work ADM [10] finds a much more effective architecture and further achieves a state-of-the-art performance compared to other generative models with classifier guidance. From the perspective that DDPMs can be treated as solving differential equations on manifolds, Liu *et al*. [23] propose pseudo numerical methods for diffusion models (PNDMs), which further improves sampling efficiency and generation quality.

Besides unconditional image generation, there are also plenty of text-to-image generation works based on diffusion models [37, 14, 35, 39, 5, 38]. Among them, VQ-Diffusion [14] is based on a VQ-VAE [45] and models the latent space by a conditional variant of DDPMs. Another typical work is LDM [37] that conditions the diffusion models on the input by cross-attention mechanism, and proposes latent diffusion models by introducing latent space [11].

Recent popular Stable Diffusion v1 and v2 are based on LDM [37] and further improved to achieve a surprising generation performance.

2.2. Generalizable Generated Image Detection

Generated image detection has been widely explored over the past years. Earlier researchers focus on detecting generated images leveraging hand-crafted features, such as color cues [28], saturation cues [29], blending artifacts [22], co-occurrence features [30]. Marra *et al*. [26] study several classical deep CNN classifiers [18, 43, 7] to detect images generated by image-to-image translation networks. However, they do not consider the generalization capability to unseen generation models. In another work, Wang *et al*. [47] notice this challenge and claim that training a simple classifier on ProGAN-generated images can generalize to other unseen GAN-based generated images well. However, their strong generalization capability relies on their large-scale training and 20 different models each trained on a different LSUN [48] object category.

Besides detection by spatial artifacts, there are also frequency-based methods [12, 50]. Frank *et al*. [12] present that in the frequency domain, GAN-generated images are more likely to expose severe artifacts mainly caused by upsampling operations in previous GAN architectures. Zhang *et al*. [50] propose a GAN simulator, AutoGAN, to simulate the artifacts produced by standard GAN pipelines. Then they train a detector on the spectrum input on the synthesized images. It can generalize to unseen generation models to some extent. Marra *et al*. [27] and Yu *et al*. [49] suggest detecting generated images by fingerprints that are often produced during GAN generation. A recent work [25] proposes a detector based on an ensemble of EfficientNet-B4 [44] to alleviate the generalization problem.

However, with the boosting development of diffusion models, a general and robust detector for detecting images generated by diffusion models has not been explored. We note that some recent works also notice the diffusion-generated image detection problem [36, 8]. Different from them, the focus of our work is exploring a generalizable detector for wide-range diffusion models.

3. Method

In this paper, we present a novel representation named **DI**ffusion **R**econstruction **E**rror (**DIRE**) for diffusion-generated image detection. DIRE measures the error between an input image and its reconstruction by a pre-trained diffusion model. We observe that diffusion-generated images can be more approximately reconstructed by a pre-trained diffusion model compared to real images. Based on this, the DIRE provides discriminative properties for distinguishing diffusion-generated images from real images. The rest of this section is organized as follows. We begin with reviewing DDPMs, and the inversion and reconstruction process of the DDIM [42]. Then we present details of DIRE for diffusion-generated image detection. Finally, we introduce a new dataset, *i.e*., DiffusionForensics, for evaluating diffusion-generated image detectors.

3.1. Preliminaries

Denoising Diffusion Probabilistic Models (DDPMs). Diffusion models are first proposed in [41] inspired by non-equilibrium thermodynamics, and achieve strong performance in image generation [17, 32, 10, 37]. They define a Markov chain of diffusion steps that slowly add Gaussian noise to data until degenerating it into isotropic Gaussian distribution (forward process), and then learn to reverse the diffusion process to generate samples from the noise (reverse process). The Markov chain in the forward process is defined as:

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

in which $\mathbf{x}_t$ is the noisy image at the t-th step and $\alpha_1, \ldots, \alpha_T$ is a predefined schedule, with T denotes the total steps.

An important property brought by the Markov chain is that we can obtain $\mathbf{x}_t$ from $\mathbf{x}_0$ directly via:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\alpha_t}\mathbf{x}_0, (1-\alpha_t)\mathbf{I}). \quad (2)$$

The reverse process in [17] is also defined as a Markov chain:

$$p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \boldsymbol{\mu}_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)). \quad (3)$$

Diffusion models use a network $p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)$ to fit the real distribution $q(\mathbf{x}_{t-1}|\mathbf{x}_t)$. The overall simplified optimization target is a sampling and denoising process as follows,

$$L_{\text{simple}}(\theta) = \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\alpha_t}\mathbf{x}_0 + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}, t)\right\|^2\right], \quad (4)$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Denoising Diffusion Implicit Models (DDIMs). DDIM [42] proposes a new deterministic method for accelerating the iterative process without the Markov hypothesis. The new reverse process in DDIM is as follows,

$$\begin{aligned}\mathbf{x}_{t-1} = &\sqrt{\alpha_{t-1}}\left(\frac{\mathbf{x}_t - \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)}{\sqrt{\alpha_t}}\right) \\ &+\sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) + \sigma_t\epsilon_t.\end{aligned} \quad (5)$$

If $\sigma_t = 0$, the reverse process becomes deterministic (reconstruction process), in which one noise sample determines one generated image. Furthermore when T is large enough (*e.g*., $T = 1000$), Eqn. (5) can be seen as Euler integration for solving ordinary differential equations (ODEs):

$$\frac{\mathbf{x}_{t-\Delta t}}{\sqrt{\alpha_{t-\Delta t}}} = \frac{\mathbf{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t-\Delta t}}{\alpha_{t-\Delta t}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}}\right)\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t). \quad (6)$$

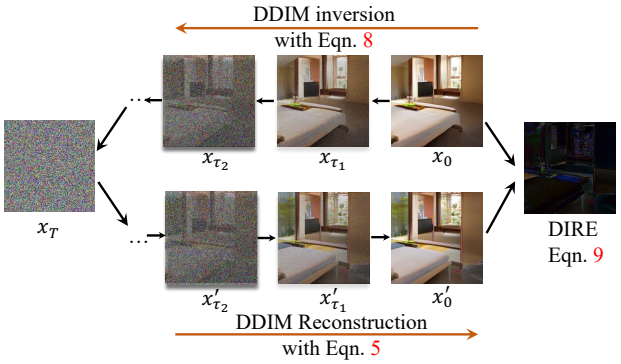


Figure 3: **Illustration of the process of computing DIRE given an input image** $\mathbf{x}_0$. The input image $\mathbf{x}_0$ is first gradually inverted into a noise image $\mathbf{x}_T$ by DDIM inversion [42], and then is denoised step by step until getting a reconstruction $\mathbf{x}'_0$. DIRE is simply defined as the residual image got from $\mathbf{x}_0$ and $\mathbf{x}'_0$.

Suppose $\sigma = \sqrt{1-\alpha}/\sqrt{\alpha}$, $\bar{\mathbf{x}} = \mathbf{x}/\sqrt{\alpha}$, the corresponding ODE becomes:

$$\mathrm{d}\bar{\mathbf{x}}(t) = \boldsymbol{\epsilon}_\theta \left(\frac{\bar{\mathbf{x}}(t)}{\sqrt{\sigma^2+1}}, t \right) \mathrm{d}\sigma(t). \tag{7}$$

Then the inversion process(from $\mathbf{x}_t$ to $\mathbf{x}_{t+1}$) can be the reversion of the reconstruction process:

$$\frac{\mathbf{x}_{t+1}}{\sqrt{\alpha_{t+1}}} = \frac{\mathbf{x}_t}{\sqrt{\alpha_t}} + \left(\sqrt{\frac{1-\alpha_{t+1}}{\alpha_{t+1}}} - \sqrt{\frac{1-\alpha_t}{\alpha_t}} \right) \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t). \tag{8}$$

This process is to obtain the corresponding noisy sample $\mathbf{x}_T$ for an input image $\mathbf{x}_0$. However, it is very slow to invert or sample step by step. To speed up the diffusion model sampling, DDIM [42] permits us to sample a subset of S steps $\tau_1, \ldots, \tau_S$, so that the neighboring $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$ become $\mathbf{x}_{\tau_t}$ and $\mathbf{x}_{\tau_{t+1}}$, respectively, in Eqn. (8) and Eqn. (5).

3.2. DIRE

Due to the intrinsic differences between diffusion models and previous generative models (*i.e.*, GANs, Flow-based models, VAEs), existing generated image detectors experience dramatic performance drops when facing images generated by diffusion models. To avoid the abuse of diffusion models, it is urgent to develop a detector for diffusion-generated image detection. A straightforward approach would be to train a binary classifier using a dataset of both real and diffusion-generated images. However, it is difficult for such a method to guarantee generalization to diffusion models that have not been previously encountered.

Our research takes note of the fact that images generated by diffusion models are essentially sampled from the distribution of the diffusion generation space ($p_g(\mathbf{x})$), while real images are sampled from another distribution ($p_r(\mathbf{x})$) although it may be near to $p_g(\mathbf{x})$ but not exactly the same.

Image Source	Denoising Condition	Generator	# of images
LSUN -Bedroom [48]		real	42,000
	unconditional	ADM [10]	42,000
		DDPM [17]	42,000
		iDDPM [32]	42,000
		PNDM [23]	42,000
	text-to-image	LDM [37]	42,000
		SD-v1 [37]	42,000
		SD-v2 [37]	42,000
		VQ-Diffusion [14]	42,000
		IF [39]	1,000
		DALLE-2 [35]	500
		Midjourney	100
ImageNet [9]		real	50,000
	conditional	ADM [10]	50,000
	text-to-image	SD-v1 [37]	50,000
CelebA-HQ [19]		real	42,000
	text-to-image	SD-v2 [37]	42,000
		IF [39]	1,000
		DALLE-2 [35]	500
		Midjourney	100

Table 1: **Composition of the DiffusionForensics dataset.** It includes real images from LSUN-Bedroom [48] ImageNet [9], CelebA-HQ [19] and generated images from pre-trained diffusion models. According to the class of diffusion models, the containing images are divided into three categories: unconditional, conditional, and text-to-image.

Our core motivation is that samples from the diffusion generation space $p_g(\mathbf{x})$ are more likely to be reconstructed by a pre-trained diffusion model while real images cannot.

So the key idea of our work is to make use of the diffusion model to detect diffusion-generated images. We find that images generated by diffusion models are more likely to be reconstructed by a pre-trained diffusion model. On the other hand, due to the complex characteristics of real images, real images can not be well constructed. As shown in Figure 1, the reconstruction error of real and diffusion-generated images shows dramatically different properties.

Given an input image $\mathbf{x}_0$, we wish to judge whether it is synthesized by diffusion models. Take a pre-trained diffusion model $\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t)$. As shown in Figure 3, we apply the DDIM [42] inversion process to gradually add Gaussian noise into $\mathbf{x}_0$ via Eqn. (8). After S steps, $\mathbf{x}_0$ becomes a point $\mathbf{x}_T$ in the isotropic Gaussian noise distribution. The inversion process is to find the corresponding point in noisy space, then the DDIM [42] generation process (Eqn. (5)) is employed to reconstruct the input image and produces a recovered version $\mathbf{x}'_0$. The differences between $\mathbf{x}_0$ and $\mathbf{x}'_0$ help to distinguish real or generated. Then the DIRE is defined as:

$$\mathrm{DIRE}(\mathbf{x}_0) = |\mathbf{x}_0 - \mathbf{R}(\mathbf{I}(\mathbf{x}_0))|, \tag{9}$$

where $|\cdot|$ denotes computing the absolute value, $\mathbf{I}(\cdot)$ is a series of the inversion process with Eqn. (8) and $\mathbf{R}(\cdot)$ is a series of the reconstruction process with Eqn. (5).

Then for real images and diffusion-generated images, we can get their DIRE representations, we train a binary classifier to distinguish their DIREs by a simple binary cross-entropy loss, which is formulated as follows,

$$L(\mathbf{y}, \mathbf{y}') = -\sum_{i=1}^{N}(\mathbf{y}_i \log(\mathbf{y}'_i) + (1-\mathbf{y}_i)\log(1-\mathbf{y}'_i)), \quad (10)$$

where N is mini-batch size, $\mathbf{y}$ is the ground-truth label, and $\mathbf{y}'$ is the corresponding prediction by the detector. In the inference stage, we first apply a diffusion model to reconstruct the image and get the DIRE. Subsequently, we input the DIRE into the binary classifier, which will then classify the source image as either real or generated.

3.3. DiffusionForensics: A Dataset for Evaluating Diffusion-Generated Image Detectors

To better evaluate the performance of diffusion-generated detectors, we establish a dataset, DiffusionForensics, which is comprised of images generated by various diffusion models for comprehensive experiments. Its composition is shown in Table 1. The images can be roughly divided into three classes by their source: LSUN-Bedroom [48], ImageNet [9], and CelebA-HQ [19].

LSUN-Bedroom. We collect bedroom images generated by 11 diffusion models, in which four subset images (ADM [10], DDPM [17], iDDPM [32], PNDM [23]) are generated by unconditional diffusion models and the other seven (LDM [37], SD-v1 [37], SD-v2 [37], VQ-Diffusion [14], IF [39][1], DALLE-2 [35], and Midjourney[2]) are generated by text-to-image diffusion models. The text prompt for all the text-to-image generation is "A photo of bedroom". The numbers of real and generated images in each image domain are listed in Table 1. Subsets containing 42,000 images are divided into 40,000 for training, 1,000 for validation, and 1,000 for testing. The remaining subsets (IF, DALLE-2, Midjourney) are only used for testing.

ImageNet. We further collect images from ImageNet for evaluating detectors when facing more universal image generation and cross-dataset evaluation. To be specific, we collect images from a conditional diffusion model (ADM [10]) with class condition. Applying the pre-trained ADM model [10], we generate 50,000 images in total (50 images for each class in ImageNet), *i.e.*, 40,000 for training, 5,000 for validation, and 5,000 for testing. And for text-to-image diffusion generation, we employ SD-v1 [37] in which the text prompt for generation is "A photo of {class}"(1,000 classes from ImageNet [9]). The number and split of images is the same as conditional ADM.

CelebA-HQ. Besides the bedroom and universal ImageNet scenarios, one may be curious about face domain. We further collect 42,000 real images from CelebA-HQ [19]. And sampling 42,000 face images using the pre-trained SD-v2 model with the prompt "A professional photograph of face". The 40,000/1,000/1,000 images in this SD-v2 subset are used as the face-domain training/validation/testing dataset. Further, we collect 1,000 IF images, 500 DALLE-2 images, and 100 Midjourney images only for face-domain evaluation.

The split of real images for training/validation/testing is 40,000/1,000/1,000 when the number of real images is 42,000, and 40,000/5,000/5,000 when the number of real images is 50,000. Besides, all the data in the DiffusionForensics dataset are triplet, *i.e.*, source image, reconstructed image, and corresponding DIRE image. In general, the proposed DiffusionForensics dataset contains unconditional, conditional, and text-to-image generated images, which is fertile for evaluation from various aspects.

4. Experiment

In this section, we first introduce the experimental setups and then provide extensive experimental results to demonstrate the superiority of our approach.

4.1. Experimental Setup

Data pre-processing and augmentation. All the experiments are conducted on our DiffusionForensics dataset. To calculate DIRE for each image, we use the ADM [10] network pre-trained on LSUN-Bedroom as the reconstruction model, and the DDIM [42] inversion and reconstruction process in which the number of diffusion steps $S = 20$ by default. We employ ResNet-50 [15] as our forensics classifier. The size of most images (ADM [10], DDPM [17], iDDPM [32], PNDM [23], VQ-Diffusion [14], LDM [37]) in the dataset is 256×256. For Stable Diffusion [37] v1 and v2, IF, DALLE-2, and Midjourney, the generated images are resized into 256×256 with bicubic interpolation. During training, the images fed into the network are randomly cropped with the size of 224×224 and horizontally flipped with a probability of 0.5. During testing, the images are center-cropped with the size of 224×224.

Evaluation metrics. Following previous generated-image detection methods [47, 46, 51], we report accuracy (ACC) and average precision (AP) in our experiments to evaluate the detectors. The threshold for computing accuracy is set to 0.5 following [47].

Baselines. 1) CNNDetection [47] proposes a CNN-generated image detection model that can be trained on one CNN dataset and then generalized to other CNN-synthesized images. 2) GANDetection [25] applies an ensemble of EfficientNet-B4 [44] to increase the detection performance. 3) SBI [40] trains a general synthetic-image detector on images generated by blending pseudo source and target images from single pristine images. 4) Patchforensics [4] employs a patch-wise classifier which is claimed to be better than

[1]Reproduced version of Imagen by DeepFloyd Lab at StabilityAI: https://github.com/deep-floyd/IF

[2]https://www.midjourney.com

Method	Training dataset	Generation model	Recon. model	Diffusion-generated bedroom images											Total Avg.
				ADM	DDPM	iDDPM	PNDM	SD-v1	SD-v2	LDM	VQD	IF	DALLE-2	Mid.	
CNNDet [47]	LSUN	ProGAN	–	50.1/63.4	56.7/74.6	50.1/77.6	50.3/82.9	50.2/70.9	50.8/80.4	50.1/60.2	50.1/70.6	51.3/79.7	68.4/78.9	90.8/11.1	56.3/68.2
GANDet [25]	LSUN	ProGAN	–	54.2/43.6	52.2/47.3	45.7/57.3	42.1/77.6	68.1/78.5	61.5/52.7	79.2/57.1	64.8/52.3	90.6/16.1	95.7/11.8	92.3/24.1	67.9/47.1
Patchfor [4]	FF++	Multiple	–	50.4/74.8	56.8/67.4	50.3/69.5	55.1/78.5	49.9/84.7	50.0/52.8	54.0/92.0	92.8/99.7	55.3/88.1	66.9/65.1	90.9/81.5	61.1/77.6
SBI [40]	FF++	Multiple	–	53.6/57.7	55.8/47.4	54.0/58.2	46.7/44.8	65.6/75.9	55.0/59.8	81.0/88.3	59.6/66.6	70.8/78.1	67.7/52.5	76.5/9.6	62.4/58.1
CNNDet* [47]	LSUN-B.	ADM	–	**100/100**	83.7/99.5	**100/100**	71.2/98.6	77.4/85.8	85.9/98.4	98.9/**100**	72.9/97.2	99.8/**100**	90.9/95.1	92.4/52.0	88.5/93.3
Patchfor* [4]	LSUN-B.	ADM	–	**100/100**	72.9/**100**	**100/100**	96.6/**100**	63.2/71.3	97.2/**100**	97.3/**100**	**100/100**	99.8/**100**	**100/100**	99.4/**100**	93.3/97.4
F3Net* [34]	LSUN-B.	ADM	–	96.0/99.7	95.5/99.6	96.4/99.9	96.0/99.7	86.1/95.3	81.1/91.5	93.8/98.4	90.1/96.7	89.4/96.6	92.9/95.8	86.9/23.1	91.3/90.6
DIRE (ours)	LSUN-B.	ADM	ADM	**100/100**	**100/100**	**100/100**	99.7/**100**	**99.7/100**	**100/100**	**100/100**	**100/100**	**100/100**	**100/100**	**100/100**	**99.9/100**
	LSUN-B.	PNDM	ADM	**100/100**	**100/100**	**100/100**	**100/100**	89.4/99.9	**100/100**	**100/100**	**100/100**	**100/100**	**100/100**	**100/100**	99.0/**100**
	LSUN-B.	iDDPM	ADM	99.6/**100**	**100/100**	**100/100**	89.7/99.8	99.7/**100**	**100/100**	99.9/**100**	99.9/**100**	99.9/**100**	99.6/**100**	**100/100**	98.9/**100**
	LSUN-B.	StyleGAN	ADM	98.8/**100**	99.8/**100**	99.9/**100**	89.6/**100**	95.2/**100**	**100/100**	**100/100**	**100/100**	**100/100**	99.9/**100**	**100/100**	98.5/**100**

Table 2: **Comprehensive comparisons of our DIRE and other generated image detectors on the LSUN-Bedroom split of DiffusionForensics.** The previous detectors including CNNDet [47], GANDet [25], Patchfor [4], and SBI [40] are evaluated with their provided weights. * denotes our reproduced training with the official codes. All the used diffusion-generation models [10, 23, 32] for preparing training data are unconditional models pre-trained on LSUN-Bedroom (LSUN-B.) [48]. Generated images from StyleGAN [20] trained on LSUN-Bedroom are downloaded from the official repository. All the testing images produced by text-to-image generators (SD-v1 [37], SD-v2 [37], LDM [37], VQDiffusion [14], IF, DALLE-2, Midjourney) are prompted by "A photo of bedroom". We report ACC (%) and AP (%) (ACC/AP in the Table).

Method	Generated face images				
	SD-v2	IF	DALLE-2	Midjourney	StarGAN
CNNDet* [47]	95.7/99.8	71.1/82.7	64.8/33.7	90.4/69.3	30.7/45.3
F3Net* [34]	89.9/99.1	75.2/84.9	75.2/69.8	82.5/87.9	27.0/45.2
DIRE (ours)	**96.7/100**	**96.8/99.9**	**95.6/99.9**	**99.1/100**	**97.9/99.8**

Table 3: **Face domain evaluation.** All detectors are trained on CelebA-HQ [19] and diffusion images generated by SD-v2 [37]. * denotes our reproduced training with the official codes. When generating images using SD-v2 and IF, the prompts used is "A professional photograph of face". ACC (%) and AP (%) are reported (ACC/AP in the Table).

simple classifiers for fake image detection. 5) F3Net [34] proposes that the frequency information of images is essential for fake image detection.

4.2. Comparison to Existing Detectors

Diffusion models [17, 10] are claimed to exhibit better generation ability than previous generation models (*e.g.*, GAN [13], VAE [21]). We notice that previous detectors achieve surprising performance on images generated by CNNs [20, 6, 2], but the generalization ability to recent diffusion-generated images has not been well explored. Here, we evaluate CNNDetection [47], GANDetection [25], Patchforensics [4], and SBI [40] on the proposed DiffusionForensics dataset using the pre-trained weights downloaded from their official repositories.

First, we conduct experiments on the LSUN-Bedroom split of DiffusionForensics. The quantitative results can be found in Table 2. We find that existing detectors have a significant performance drop when dealing with diffusion-generated images, with ACC results lower than 70%. We also include diffusion-generated images (ADM [10]) as training data and re-train CNNDetection [47], Patchforensics [4], and F3Net [34], whose training codes are publicly available. The resulting models get a significant improvement on images generated by the same diffusion models as used in training, but still perform unsatisfactorily facing unseen diffusion models. In contrast, our DIRE shines with excellent generalization performance. Concretely, DIRE with the generation model used to prepare training data and the reconstruction model used to compute DIRE set to ADM achieves an average of 99.9% ACC and 100% AP when detecting bedroom images generated by various diffusion models.

Besides the comprehensive comparisons on bedroom images, we further conduct a comparison of DIRE and previous detectors on the CelebA-HQ split of DiffusionForensics. The result is reported in Table 3. Our DIRE, CNNDet [47], and F3Net [34] are trained with images generated by SD-v2, and evaluated with images generated by SD-v2, IF, DALLE-2, Midjouney, and StarGAN. The results demonstrate our DIRE encompasses a much stronger capability when detecting generated face images.

4.3. Generalization Capability Evaluation

Effect of choice of generation and reconstruction models. We evaluate the impact of different choices of the generation and reconstruction models on the generalization capability. We employ the ADM [10] model as the reconstruction model and apply different models for generating images. After generation, the ADM reconstruction model converts these images into their DIREs for training a binary classifier. In this evaluation, we select three different generation models for preparing training data: PNDM [23] and iDDPM [32] (diffusion models) and StyleGAN [20] (GAN model). The results are reported in Table 2. Despite the inconsistent use of generation and reconstruction models when training, DIRE still keeps a strong generalization capability. Specifically, when

Method	Generation model	Generated IN images	
		ADM	SD-v1
CNNDet* [47]	ADM	66.2/82.3	47.0/80.4
F3Net* [34]	ADM	69.5/87.3	44.8/82.6
DIRE (ours)	ADM	**98.4/99.9**	**97.2/99.6**
	iDDPM	93.4/99.4	92.5/98.8
	StyleGAN	85.6/98.4	85.4/98.1

Table 4: **Cross-dataset evaluation** on the ImageNet (IN) split using the detectors trained on the LSUN-Bedroom split of DiffusionForensics. * denotes our reproduced training with the official codes. Each testing set is generated by corresponding generation model pre-trained on the corresponding dataset. Images generated by SD-v1 are prompted by " A photo of {class}" in which the classes are from [9]. ACC (%) and AP (%) are reported (ACC/AP in the Table).

Method	GAN-generated bedroom images			
	StyleGAN	ProjGAN	Diff-StyleGAN	Diff-ProjGAN
CNNDet* [47]	94.3/99.8	62.2/93.2	68.1/91.4	60.0/92.6
F3Net* [34]	88.1/95.5	74.4/86.0	85.5/94.4	70.2/83.0
DIRE (ours)	99.8/**100**	**100/100**	**100/100**	**100/100**

Table 5: **GAN evaluation** on the LSUN-Bedroom split. * denotes our reproduced training on the LSUN-Bedroom-ADM subset of DiffusionForensics with the official codes. Each testing set is generated by corresponding GAN model pre-trained on the corresponding dataset. ACC (%) and AP (%) are reported (ACC/AP in the Table).

pairing iDDPM [32] as the generation model and ADM [10] as the reconstruction model, DIRE achieves 98.9% ACC and 100% AP on average, highlighting its adaptation with images generated by different diffusion models. It's worth noting that when the generation model is StyleGAN, DIRE still exhibits excellent performance. This might be attributed to DIRE's capability of incorporating the generation properties of other generation models besides diffusion models.

Cross-dataset evaluation. We further design a more challenging scenario, *i.e.*, training the detector with images generated by models pre-trained on LSUN-Bedroom [48] and then testing it on images produced by models pre-trained on ImageNet [9]. We choose three different generators for generating training images: ADM [10], iDDPM [32], and StyleGAN [20]. The evaluation results on ADM (IN) are shown in Table 4. We find that CNNDet [47] and F3Net [34] get a dramatically performance. But DIRE still maintains a satisfactory generalization capability even though facing unseen datasets, *i.e.*, ACC/AP: 98.4%/99.9% and 93.4%/99.4% when training on images generated by ADM and iDDPM, respectively. This evaluation further validates that the proposed DIRE is a general image representation for detecting diffusion-generated images.

Unseen text-to-image generation evaluation. Furthermore, we seek to verify whether DIRE can detect images generated by unseen text-to-image models. We adopt SD-v1 as the generation model and generate images based on the class label of ImageNet [9]. The results are shown in Table 4. Our detector DIRE trained with images generated by ADM pre-trained on LSUN-Bedroom achieves a 97.2% ACC and 99.6% AP, demonstrating the strong generalization capability of DIRE to text-to-image generation models.

Unseen GAN evaluation. Besides generalization between diffusion models, we further evaluate the performance of DIRE for images generated by GANs. We evaluate the performance of our DIRE, CNNDet [47], and F3Net [34] trained on the ADM subset of the LSUN-Bedroom split. The results are reported in Table 5. In this setting, all the trained detectors are not trained with any GAN image. Our reproduced CNNDet and F3Net experience significant performance drop, which suggests that previous generated image detectors fail across diffusion and GAN models. In contrast, DIRE achieves surprising performance when detecting GAN-generated images. This indicates that DIRE is not only an effective image representation for diffusion-generated image detection but also may beneficial to detect GAN-generated images even though DIRE is built upon a mathematical formulation of the diffusion forward and reverse processes.

4.4. Robustness to Unseen Perturbations

Besides the generalization to unseen generation models, the robustness to unseen perturbations is also a common concern since in real-world applications images are usually perturbed by various degradations. Here, we evaluate the robustness of detectors in two-class degradations, *i.e.*, Gaussian blur and JPEG compression, following [47]. The perturbations are added under three levels for Gaussian blur ($\sigma = 1, 2, 3$) and two levels for JPEG compression ($quality = 65, 30$). We explore the robustness of our baselines CNNDetection [47], GANDetection [25], SBI [40], F3Net [34], Patchforensics [4], and our DIRE. The results are shown in Figure 4. We observe that at each level of blur and JPEG compression, our DIRE gets a perfect performance without performance drop. It is worth noting that our reproduction of CNNDetection* [47] and Patchforensics* [4] trained on LSUN-Bedroom-ADM subset of DiffusionForensics also get satisfactory performance while they experience a dramatic performance drop facing JPEG compression, which further reveals training on RGB images may be not robust.

4.5. More Analysis of the Proposed DIRE

In this subsection, we conduct experiments on the LSUN-Bedroom split of DiffusionForensics to help better understanding of DIRE.

How do the inversion steps in DDIM affect the detection performance? Recent diffusion models [42, 10] find that more steps contribute to more high-quality images and

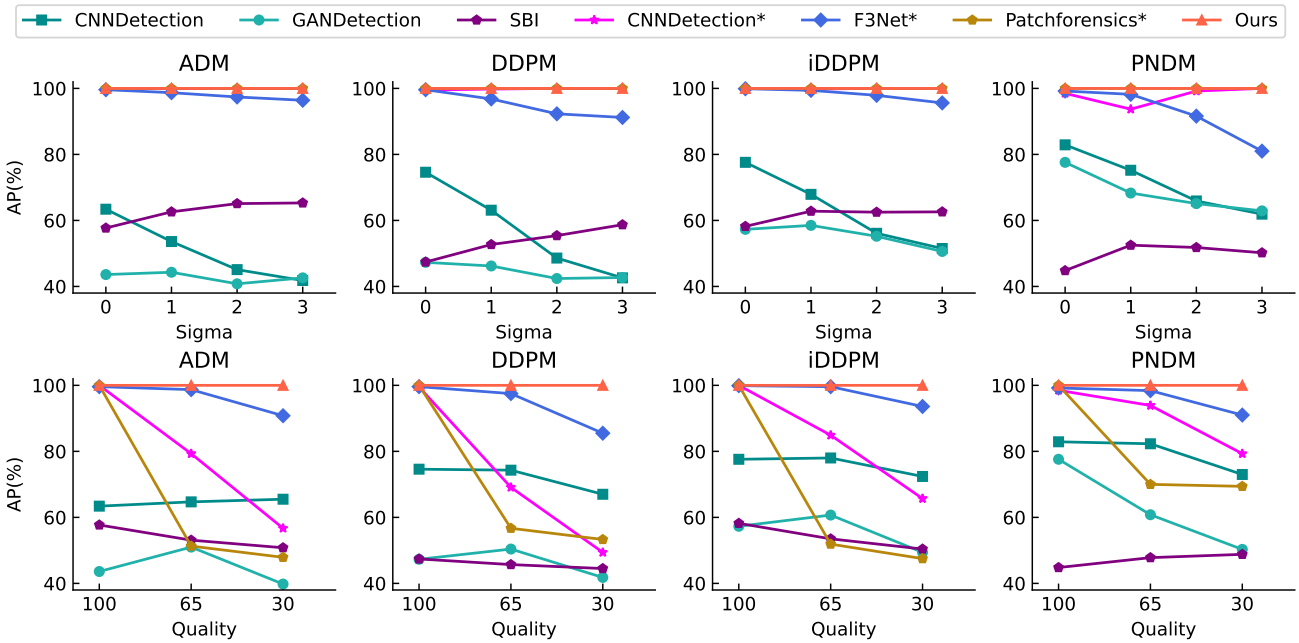


Figure 4: **Robustness to unseen perturbations.** The top rows show the robustness to Gaussian blur, and the bottom rows show the robustness to JPEG compression. * denotes our reproduced training on the LSUN-Bedroom-ADM subset of DiffusionForensics with AP (%) reported for robustness comparison.

S	ADM	DDPM	iDDPM	PNDM	SD-v1
5	**100/100**	**100/100**	**100/100**	97.5/**100**	87.5/99.8
10	**100/100**	**100/100**	**100/100**	99.4/**100**	98.2/**100**
20	**100/100**	**100/100**	**100/100**	99.7/**100**	99.7/**100**
50	**100/100**	**100/100**	**100/100**	**100/100**	**99.9/100**

Table 6: **Influence of different inversion steps.** All the models in this experiment are trained on the ADM subset and tested on other subsets of LSUN-Bedroom. ACC (%) and AP (%) are reported (ACC/AP in the Table).

Input	ADM	DDPM	iDDPM	PNDM	SD-v1
REC	**100/100**	57.1/57.7	49.7/92.6	87.1/98.7	46.9/57.0
RGB	**100/100**	87.3/99.6	**100/100**	77.8/99.1	77.4/85.8
RGB&DIRE	**100/100**	99.8/**100**	99.9/**100**	99.2/**100**	62.4/92.4
DIRE	**100/100**	**100/100**	**100/100**	**99.7/100**	**99.7/100**

Table 7: **Influence of different input information.** All the models in this experiment are trained on the ADM subset and tested on other subsets of LSUN-Bedroom. ACC (%) and AP (%) are reported (ACC/AP in the Table).

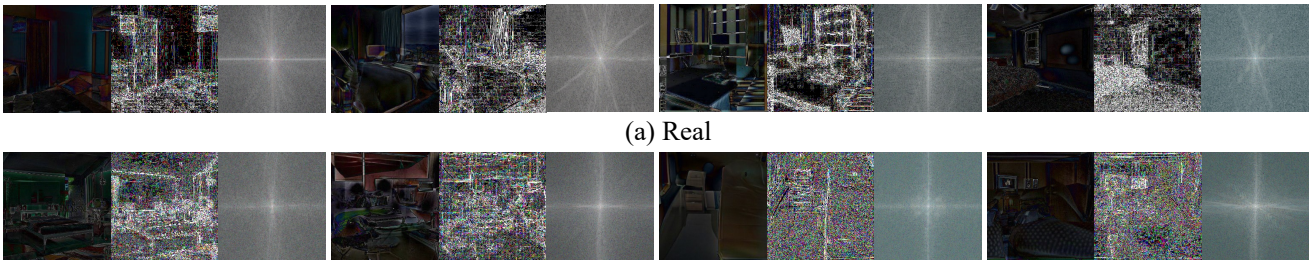
DDIM [42] sampling can improve the generation performance compared to original DDPM [17] sampling. Here, we explore the influence of different inversion steps in diffusion-generated image detection. Note that the steps in reconstruction are the same as in the inversion by default. The results are reported in Table 6. We observe that more steps in DDIM benefit the detection performance of DIRE. Considering the computational cost, we choose 20 steps by default.

Is DIRE really better than the original RGB for detecting diffusion-generated images? We conduct an experiment on various forms of input for detection, including RGB images, reconstructed images (REC), DIRE, and the combination of RGB and DIRE (RGB&DIRE). The results displayed in Table 7 reveal that REC performed much worse than RGB, suggesting that reconstructed images are not suitable as input information for detection. One possible explanation is the loss of essential information during reconstruction by a pre-trained diffusion model. The comparison between RGB and DIRE also demonstrates that DIRE serves as a stronger image representation, contributing to a more generalizable detector than simply training on RGB images. Furthermore, we find that combining RGB with DIRE together hurts the generalization compared to pure DIRE. Therefore, we use DIRE as the default input for detection by default.

	ADM	DDPM	iDDPM	PNDM	SD-v1
w/o ABS	**100/100**	99.4/**100**	**100/100**	98.2/**100**	87.0/93.0
w/ ABS	**100/100**	**100/100**	**100/100**	**99.7/100**	**99.7/100**

Table 8: **Effect of computing the absolute value (ABS) when obtaining DIRE.** All the models in this experiment are trained on the ADM subset and tested on other subsets of LSUN-Bedroom. ACC (%) and AP (%) are reported (ACC/AP in the Table).

Effect of different calculation of DIRE. After computing the residual result of the reconstructed image and source image, whether to compute the absolute value should be considered. As reported in Table 8, we find that the absolute operation is critical for achieving a strong diffusion-generated image detector, particularly on SD-v1 [37] where it improves



(a) Real (b) ADM (c) DDPM (d) iDDPM (e) PNDM

Figure 5: **Noise pattern and frequency analysis of DIRE of real and generated images.** Noise pattern is regular to portray the shape of objects in DIRE of real images, while it is messy in DIRE of diffusion-generated images. For frequency analysis, the frequency bands in DIRE of real images are more abundant than that of diffusion-generated images, *i.e.*, the white regions in the frequency domain are larger.

ACC/AP from 87.0%/93.0% $\rightarrow$ 99.7%/100%. So by default, the absolute operation is applied in all our models.

Qualitative Analysis of DIRE. The above quantitative experiments have indicated the effectiveness of the proposed DIRE. As analyzed before, the key motivation behind DIRE is that generated images can be approximately reconstructed by a pre-trained diffusion model while real images cannot. DIRE makes use of the residual characteristic of an input image and its reconstruction for discrimination. To gain a better understanding of its intrinsic properties, we conduct a further qualitative analysis of DIRE, utilizing noise pattern and frequency analysis for visualization.

When images are acquired, various factors from hardware facilities, such as lens and sensors, and software algorithms, such as compression and demosaic, can impact image quality at the low level. One typical low-level analysis of images is noise pattern analysis[3], which is usually regular and corresponds to the shape of objects in real scenarios. In addition to low-level analysis, frequency analysis can provide frequency information about images. To compute the frequency information of DIRE, we used FFT algorithms.

We visualize the results of the aforementioned two analysis tools in Figure 5. The visual comparison of noise patterns highlights significant differences of the DIRE of real and diffusion-generated images from the low-level perspective, with real images tending to be regular and corresponding to the shape of objects while diffusion-generated images tend to be messy. By comparing the FFT spectrum of DIRE from real and diffusion-generated images, we observe that the FFT spectrum of real images is usually more abundant than that of diffusion-generated images, which confirms that real images are more difficult to be reconstructed by a pre-trained diffusion model.

5. Conclusion

In this paper, we focus on building a generalizable detector for discriminating diffusion-generated images. We find that previous generated-image detectors show limited performance when detecting images generated by diffusion models. To address the issue, we present an image representation called DIRE based on reconstruction errors of images inverted and reconstructed by DDIM. Furthermore, we create a new dataset, DiffusionForensics, which includes images generated by unconditional, conditional, and text-to-image diffusion models to facilitate the evaluation of diffusion-generated images. Extensive experiments indicate that the proposed image representation DIRE contributes to a strong diffusion-generated image detector, which is very effective for this task. We hope that our work can serve as a solid baseline for diffusion-generated image detection.

Acknowledgement. This work was supported by NSFC under Contract 61836011 and 62021001. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution, USTC, and the Supercomputing Center of the USTC.

[3] https://29a.ch/photo-forensics/#noise-analysis

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022. 2

[2] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 6

[3] Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. *arXiv preprint arXiv:2301.13188*, 2023. 1

[4] Lucy Chai, David Bau, Ser-Nam Lim, and Phillip Isola. What makes fake images detectable? understanding properties that generalize. In *ECCV*, 2020. 5, 6, 7

[5] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *arXiv preprint arXiv:2209.14491*, 2022. 2

[6] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *CVPR*, 2018. 6

[7] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017. 3

[8] Riccardo Corvi, Davide Cozzolino, Giada Zingarini, Giovanni Poggi, Koki Nagano, and Luisa Verdoliva. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680*, 2022. 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 2, 4, 5, 7

[10] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, 2021. 1, 2, 3, 4, 5, 6, 7

[11] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *CVPR*, 2021. 2

[12] Joel Frank, Thorsten Eisenhofer, Lea Schönherr, Asja Fischer, Dorothea Kolossa, and Thorsten Holz. Leveraging frequency analysis for deep fake image recognition. In *ICML*, 2020. 3

[13] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NeurIPS*, 2014. 6

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *CVPR*, 2022. 2, 4, 5, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022. 2

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arxiv:2006.11239*, 2020. 1, 2, 3, 4, 5, 6, 8

[18] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 3

[19] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 2, 4, 5, 6

[20] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 6, 7

[21] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

[22] Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *CVPR*, 2020. 3

[23] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 1, 2, 4, 5, 6

[24] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*, 2022. 2

[25] Sara Mandelli, Nicolò Bonettini, Paolo Bestagini, and Stefano Tubaro. Detecting gan-generated images by orthogonal training of multiple cnns. In *ICIP*, 2022. 3, 5, 6, 7

[26] Francesco Marra, Diego Gragnaniello, Davide Cozzolino, and Luisa Verdoliva. Detection of gan-generated fake images over social networks. In *MIPR*, 2018. 3

[27] Francesco Marra, Diego Gragnaniello, Luisa Verdoliva, and Giovanni Poggi. Do gans leave artificial fingerprints? In *MIPR*, 2019. 3

[28] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using color cues. *arXiv preprint arXiv:1812.08247*, 2018. 3

[29] Scott McCloskey and Michael Albright. Detecting gan-generated imagery using saturation cues. In *ICIP*, 2019. 3

[30] Lakshmanan Nataraj, Tajuddin Manhar Mohammed, Shivkumar Chandrasekaran, Arjuna Flenner, Jawadul H Bappy, Amit K Roy-Chowdhury, and BS Manjunath. Detecting gan generated fake images using co-occurrence matrices. *arXiv preprint arXiv:1903.06836*, 2019. 3

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[32] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *ICML*, 2021. 1, 2, 3, 4, 5, 6, 7

[33] Gaurav Parmar, Krishna Kumar Singh, Richard Zhang, Yijun Li, Jingwan Lu, and Jun-Yan Zhu. Zero-shot image-to-image translation. *arXiv preprint arXiv:2302.03027*, 2023. 2

[34] Yuyang Qian, Guojun Yin, Lu Sheng, Zixuan Chen, and Jing Shao. Thinking in frequency: Face forgery detection by mining frequency-aware clues. In *ECCV*, 2020. 6, 7

[35] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 2, 4, 5

[36] Jonas Ricker, Simon Damm, Thorsten Holz, and Asja Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022. 3

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1, 2, 3, 4, 5, 6, 8

[38] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. *arXiv preprint arXiv:2208.12242*, 2022. 2

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*, 2022. 2, 4, 5

[40] Kaede Shiohara and Toshihiko Yamasaki. Detecting deepfakes with self-blended images. In *CVPR*, 2022. 5, 6, 7

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015. 1, 2, 3

[42] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 1, 2, 3, 4, 5, 7, 8

[43] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[44] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *ICML*, 2019. 3, 5

[45] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. In *NeurIPS*, 2017. 2

[46] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *ICCV*, 2019. 5

[47] Sheng-Yu Wang, Oliver Wang, Richard Zhang, Andrew Owens, and Alexei A Efros. Cnn-generated images are surprisingly easy to spot...for now. In *CVPR*, 2020. 3, 5, 6, 7

[48] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 2, 3, 4, 5, 6, 7

[49] Ning Yu, Larry S Davis, and Mario Fritz. Attributing fake images to gans: Learning and analyzing gan fingerprints. In *ICCV*, 2019. 3

[50] Xu Zhang, Svebor Karaman, and Shih-Fu Chang. Detecting and simulating artifacts in gan fake images. In *WIFS*, 2019. 3

[51] Peng Zhou, Xintong Han, Vlad I Morariu, and Larry S Davis. Learning rich features for image manipulation detection. In *CVPR*, 2018. 5

[52] Derui Zhu, Dingfan Chen, Jens Grossklags, and Mario Fritz. Data forensics in diffusion models: A systematic analysis of membership privacy. *arXiv preprint arXiv:2302.07801*, 2023. 1