Exposing the Fake: Effective Diffusion-Generated Images Detection

Ruipeng Ma [1] **Jinhao Duan** [2] **Fei Kong** [1] **Xiaoshuang Shi** [1] **Kaidi Xu** [2]

Abstract

Image synthesis has seen significant advancements with the advent of diffusion-based generative models like Denoising Diffusion Probabilistic Models (DDPM) and text-to-image diffusion models. Despite their efficacy, there is a dearth of research dedicated to detecting diffusion-generated images, which could pose potential security and privacy risks. This paper addresses this gap by proposing a novel detection method called *Stepwise Error for Diffusion-generated Image Detection* (SeDID). Comprising statistical-based $\text{SeDID}_{\text{Stat}}$ and neural network-based $\text{SeDID}_{\text{NNs}}$, SeDID exploits the unique attributes of diffusion models, namely deterministic reverse and deterministic denoising computation errors. Our evaluations demonstrate SeDID's superior performance over existing methods when applied to diffusion models. Thus, our work makes a pivotal contribution to distinguishing diffusion model-generated images, marking a significant step in the domain of artificial intelligence security.

1. Introduction

Generative diffusion models have made significant strides in the field of image generation, demonstrating remarkable capabilities (Song and Ermon, 2019; Song et al., 2020b; Ho et al., 2020; Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022), but have also raised privacy and abuse concerns (Zhao et al., 2023; Kong et al., 2023). Previous works (Wang et al., 2023; Ricker et al., 2022; Corvi et al., 2022) have laid the groundwork for detecting diffusion-generated images, and some have successfully leveraged the deterministic reverse and denoising processes inherent to diffusion models.

However, while detection methods such as DIRE (Wang et al., 2023) indeed leverage some deterministic aspects, they may not fully exploit the entirety of these features. In particular, DIRE concentrates on the reconstruction at the initial x_0 timestep, which may overlook the valuable information encapsulated in the intermediate steps throughout the diffusion and reverse diffusion processes. In contrast, the proposed SeDID exploits these intermediate steps, which could potentially enhance the detection efficacy. Additionally, we reveal that the distribution of real images could potentially diverge from the distribution of images generated by diffusion models, given the inherently complex and diverse characteristics of natural images. This indicates that the real-image distribution might not align perfectly with the regular patterns learned by the diffusion process.

Given these observations, we distinctly formulate our research question as follows:

Can we discriminate between real and diffusion-generated images by harnessing the inherent distributional disparities between naturally occurring and diffusion-synthesized visuals?

In our work, we address these issues by delving deeper into the deterministic reverse and denoising properties of diffusion models, proposing a novel and more encompassing detection approach. Our proposed method, the *Stepwise Error for Diffusion-generated Image Detection* (SeDID), is designed to comprehensively utilize these unique diffusion properties to improve detection performance, thereby presenting a more generalized and robust solution for detecting diffusion-generated images.

Our approach draws inspiration from SecMI (Duan et al., 2023), a Membership Inference Attack (MIA) that differentiates training data and hold-out data on the assumption that the model overfits the training data. It's intuitive that a model better fits generation data than training samples is evident. Under this perception, we believe that the MIA-style method might be suitable for generation detection. Our method, which we have dubbed the *Stepwise Error for Diffusion-generated Image Detection* (SeDID), utilizes the error between the reverse sample and the denoise sample at a specific timestep T_{SE}.

Our major contributions in this paper can be summarized as:

[1]Department of Computer Science and Engineering, University of Electronic Science and Technology of China [2]Drexel University. Correspondence to: Xiaoshuang Shi <xsshi2013@gmail.com>.

2nd AdvML Frontiers workshop at 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.
Copyright 2023 by the author(s).

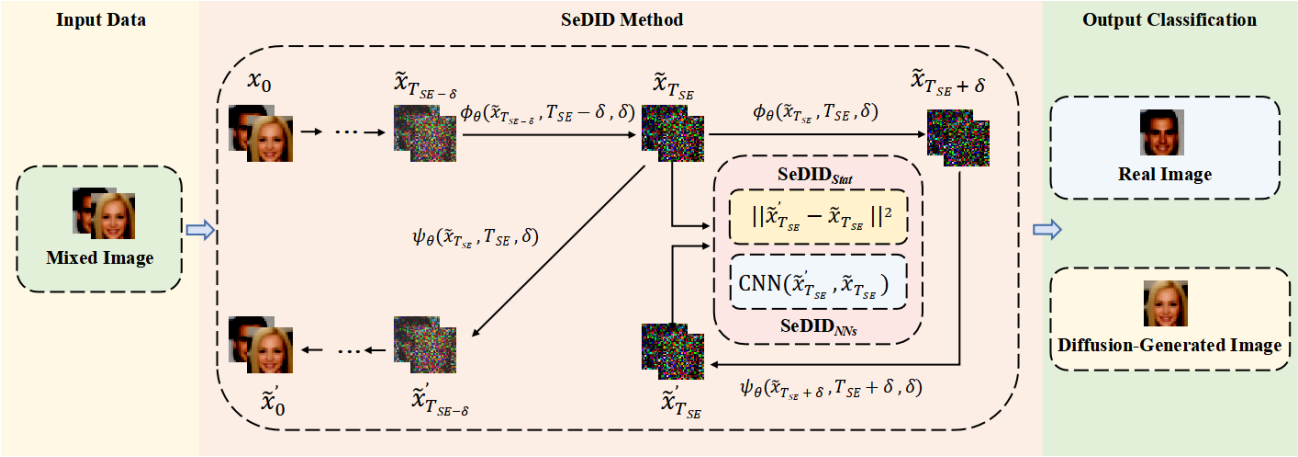

Figure 1. **The pipeline of our SeDID method.** Given mixed image data, it is processed through the SeDID method to compute the noise profile, a characterization of the noise patterns inherent to diffusion-generated images. Then, SeDID provides two branches: the Statistical-Based Synthetic Image Detection $\text{SeDID}_{\text{Stat}}$ and the Neural Network-Based Synthetic Image Detection $\text{SeDID}_{\text{NNs}}$. The $\text{SeDID}_{\text{Stat}}$ branch involves statistical analysis, error calculation, and model evaluation. The $\text{SeDID}_{\text{NNs}}$ branch employs a ResNet-18 model, which computes prediction errors, and updates weights via backpropagation. Both branches calculate the Area Under the Receiver Operating Characteristic Curve (AUC), the Accuracy (ACC), and the True Positive Rate at a given False Positive Rate (TPR@FPR), and classify images with *real* or *generated* output.

- We propose SeDID, a novel detection scheme for diffusion-generated images. SeDID uniquely exploits the distinct properties of diffusion models, particularly focusing on the errors between reverse and denoise samples at specific timesteps during the generation process.
- We adapt insights from membership inference attacks to emphasize the distributional disparities between real and generated data. This perspective enhances our understanding of diffusion models' security and privacy implications and underpins the design of SeDID.
- We present an extensive empirical evaluation of SeDID on three distinct datasets. The results demonstrate SeDID's superior performance in detecting diffusion-generated images, surpassing existing methodologies.

The remainder of this paper is organized as follows: Section 2 discusses the related work; Section 3 elaborates on our proposed methodology, SeDID; Section 4 presents the comprehensive evaluation of SeDID and discusses the results; finally, Section 5 concludes the paper and provides directions for future research.

2. Related Works

2.1. Generative Diffusion Models

Diffusion models, introduced by Sohl-Dickstein et al. (Sohl-Dickstein et al., 2015), offer an approach distinct from Generative Adversarial Networks (GANs) (Goodfellow et al., 2020; Yuan and Moghaddam, 2020; Yuan et al., 2023). These models gradually convert real data into noise and then learn to reverse this transformation. Ho et al. (Ho et al., 2020) enhanced this process, leading to a better approximation of the real data distribution. Such improvement has significantly influenced our work, particularly our emphasis on the reverse and denoising steps. This field's versatility is demonstrated by Kong et al. (Kong et al., 2020), who employed diffusion models in audio synthesis, inspiring our method's adaptability.

Diffusion models have been broadly employed for accelerating inference (Song et al., 2020a; Salimans and Ho, 2022; Dockhorn et al., 2021) and conditional generation (Dhariwal and Nichol, 2021; Ho and Salimans, 2022). Several recent studies have addressed challenges such as improving inference speed and developing innovative methods (Xiao et al., 2021; Watson et al., 2022; Rombach et al., 2022; Meng et al., 2021).

2.2. Diffusion-generated Image Detection

Research on image detection originated with a focus on black-box (Shokri et al., 2017; Salem et al., 2018; Yeom et al., 2018) and white-box attacks (Nasr et al., 2019; Rezaei and Liu, 2021), both primarily developed for classification models (Sablayrolles et al., 2019; Song and Mittal, 2021; Choquette-Choo et al., 2021; Hui et al., 2021; Truex et al., 2019; Salem et al., 2018; Pyrgelis et al., 2017). Black-box attacks assume limited knowledge about the model's internals, whereas white-box attacks presume complete model visibility. The research arena then widened to the detection of synthetic images, particularly those generated by diffusion models (Corvi et al., 2023). This evolution incorporated the examination of forensic traces in diffusion-generated

synthetic images and the performance evaluation of GAN-dedicated detectors when applied to these images, even in challenging contexts involving image compression and resizing in social networks.

DIRE (Wang et al., 2023) represents a preliminary exploration in detecting diffusion-generated data, utilizing the reconstruction error of images using Denoising Diffusion Implicit Models (DDIM) (Song et al., 2020a) for inversion and reconstruction. While this investigation is unfolding, other strides have been made within the broader field of diffusion models. Architectural advancements have been achieved with ADM (Dhariwal and Nichol, 2021), while PNDMs (Liu et al., 2022) have focused on accelerating the sampling speed. Furthermore, Stable Diffusion (Rombach et al., 2022) v1 and v2 have delved into exploring downstream tasks.

(Duan et al., 2023) proposed the Step-wise Error Comparing Membership Inference (SecMI) approach for membership inference attack (MIA), leveraging the error comparison of the posterior estimation from the forward process. This concurrent work in the field inspired us in developing our current work, SeDID, which aims to detect diffusion-generated images effectively.

In summary, our SeDID method is a refined version of the concurrent work, SecMI, refocused specifically on detecting generated images, we adapt its technique to compute errors. This approach improves both the Area Under the Curve (AUC) and Accuracy (ACC) metrics in spotting diffusion-generated images, compared to methods solely relying on DDIM for image inversion and reconstruction.

3. Methodology

In this section, we detail our novel synthetic image detection method in diffusion models, namely SeDID, which builds upon the work of Duan (Duan et al., 2023) and the Diffusion Denoising Probabilistic Models (DDPM) (Ho et al., 2020). We start by defining key notations and outlining the fundamental principles of DDPM.

3.1. Notations

We use standard notations as defined by Ho et al. (2020) (Ho et al., 2020). We denote the real data distribution as $q(x_0)$ and the latent variable model approximating $q(x_0)$ as $p_\theta(x_0)$. The "noise-prediction" model, ϵ_θ, is parameterized by weights θ.

The diffusion model comprises a T-step diffusion process $q(x_t|x_{t-1})$ and a denoising process $p_\theta(x_{t-1}|x_t)$ for $1 \leq t \leq T$:

$$\begin{aligned} q(x_t|x_{t-1}) &= \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I), \\ p_\theta(x_{t-1}|x_t) &= \mathcal{N}(x_{t-1}; \mu_\theta(x_t,t), \Sigma_\theta(x_t,t)), \end{aligned} \tag{1}$$

where x_t refers to the diffusion result at timestep t, β_t is the noise factor at timestep t, I is the identity matrix, μ_θ and Σ_θ are the mean and variance matrix of the denoising distribution respectively. The forward sampling at time step t is:

$$q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)I), \tag{2}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$.

3.2. Definitions

In the context of our work, we primarily focus on the deterministic denoising function ψ_θ, and the deterministic reverse function ϕ_θ.We also introduce the notion of (t,δ-*error*), quantifying the posterior estimation error at timestep t under stepsize δ, and the Stepwise Error Calculation Time Step, T_{SE}.

Definition 3.1 (Deterministic denoising function ψ_θ)**.** The deterministic denoising function $\psi_\theta(x,t)$, following the denoising process from DDIM (Song et al., 2020a), recovers the original data from the noised input x at timestep t. It is defined as:

$$\psi_\theta(x,t,\delta) = \sqrt{\bar{\alpha}_{t-\delta}}f_\theta(x,t) + \sqrt{1-\bar{\alpha}_{t-\delta}}\epsilon_\theta(x,t), \tag{3}$$

where $\epsilon_\theta(x,t)$ represents the stochastic noise, $\bar{\alpha}_t$ is the noise scale at timestep t and $\psi_\theta(x,t,\delta) = x_{t-\delta}$. The definition of f_θ is Eq. (4). We recover x by applying the formula recurrently. Generally, we can use $\delta > 1$ to accelerate the denoising process.

$$f_\theta(x,t) = \frac{x - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x,t)}{\sqrt{\bar{\alpha}_t}} \tag{4}$$

Definition 3.2 (Deterministic reverse function ϕ_θ)**.** The deterministic reverse function $\phi_\theta(x,t)$, following the reverse process from DDIM (Song et al., 2020a), is the reversed process of denoising process. Given a sample x_t, we can leverage $\phi_\theta(x,t)$ to obtain $x_{t+\delta}$:

$$\phi_\theta(x,t,\delta) = \sqrt{\bar{\alpha}_{t+\delta}}f_\theta(x,t) + \sqrt{1-\bar{\alpha}_{t+\delta}}\epsilon_\theta(x,t), \tag{5}$$

where $\phi_\theta(x,t,\delta) = x_{t+\delta}$.

The operations of ϕ_θ, ψ_θ, and f_θ are applied during the diffusion process at specific timesteps determined by the Stepwise Error Calculation Time Step, T_{SE}.

Definition 3.3 (t,δ-*error*)**.** For a sample $x_0 \sim D$ and its deterministic reverse result $\tilde{x}_t = \phi_\theta(\phi_\theta(\ldots\phi_\theta(x_0,0,\delta),t-2\delta,\delta),t-\delta,\delta)$, we define ($t$,$\delta$-*error*) as:

$$E_{t,\delta} = ||\psi_\theta(\phi_\theta(\tilde{x}_t,t,\delta),t,\delta) - \tilde{x}_t||^2, \tag{6}$$

where ψ_θ is the deterministic denoising function and ϕ_θ is the deterministic reverse function.

In other word, we first use ϕ_θ to obtain the $\tilde{x}_{t+\delta}$, and then, we use ψ_θ to get the reconstruction $\tilde{x}'_t$. $E_{t,\delta}$ is the difference

beteen them. It is intuitive that the model is more familiar to the sample generated, so the sample with smaller $E_{t,\delta}$ will be more likely to be the sample generated.

Definition 3.4 (Stepwise Error Calculation Time Step, T_{SE} and stepsize, δ)**.** T_{SE} is denoted to be the target timestep t to compute (t,δ-*error*), and is called **Stepwise Error Calculation Timestep**. We abuse δ to represent the selected **stepsize**, since it is not ambiguous.

3.3. Detailed Design of the SeDID Method

The SeDID method applies the concept of (t,δ-*error*) in combination with noise information during each diffusion process step, under the hypothesis that the process for real images differs from generated images. Experiments are conducted at various time steps T_{SE}, representing different stages in the diffusion process.

At each experimental time step T_{SE}, we calculate the (t,δ-*error*), $E_{T_{SE},\delta}$ under stepsize δ:

$$E_{T_{SE},\delta} = ||\psi_\theta(\phi_\theta(\tilde{x}_{T_{SE}}, T_{SE}), T_{SE}, \delta) - \tilde{x}_{T_{SE}}||^2, \quad (7)$$

where $x_{T_{SE}}$ is the intermediate result from the diffusion process at the time step T_{SE}, and $\tilde{x}_{T_{SE}}$ is the corresponding image from the denoising process at the same time step.

3.4. The Mechanics of the SeDID Method

SeDID has two variants: Statistical-Based Synthetic Image Detection ($\text{SeDID}_{\text{Stat}}$) and the Neural Networks(NNs)-Based Synthetic Image Detection ($\text{SeDID}_{\text{NNs}}$).

Statistical-Based Synthetic Image Detection – $\text{SeDID}_{\text{Stat}}$ We apply (t,δ-error) as the metrics to determine whether a sample is synthetic. If the error is smaller than a threshold h, we classify it to the synthetic sample. This can be described by:

$$g(x) = \mathbb{1}[E_{T_{SE},\delta} < h] \quad (8)$$

The AUC, best ACC, and TPR@FPR are computed based on these noise profiles.

Neural Network-Based Synthetic Image Detection – $\text{SeDID}_{\text{NNs}}$ $\text{SeDID}_{\text{NNs}}$ extends the capabilities of $\text{SeDID}_{\text{Stat}}$ by incorporating a neural network with a ResNet-18 architecture. The network is fed with the intermediate outcomes of both the diffusion and reverse diffusion processes at timestep T_{SE}. This allows the model to classify images as real or synthetic by using their respective noise profiles.

Our SeDID method leverages the strengths of statistical and machine learning approaches, accurately distinguishing real and synthetic images based on their noise profiles through a dual-phase strategy.

4. Experiment

In this section, we evaluate the performance of SeDID across various datasets and configurations.

4.1. Datasets

This study employed three publicly available datasets - CIFAR10 (Krizhevsky et al., 2009), TinyImageNet (Tavanaei, 2020), and CelebA (Liu et al., 2015) - each presenting distinct complexities due to their unique characteristics.

CIFAR10: This dataset contains 60,000 color images, evenly distributed across 10 classes. The set is divided into 50,000 training images and 10,000 validation images.

TinyImageNet: This subset of the ILSVRC-2012 classification dataset features 200 categories, each with 500 training images, 50 validation images, and 50 synthetic images.

CelebA: This large-scale face attributes dataset comprises over 200,000 celebrity images, each annotated with 40 attribute labels. The varied size and complexity of the images in this dataset make it a challenging platform for generative models.

All images across all datasets in this work are preprocessed to a uniform size of 32×32 pixels.

4.2. Experimental Settings

Dataset Preparation All datasets underwent standard preprocessing operations, including normalization with a mean of 0.5 and standard deviation of 0.5 for each color channel, as well as potential random horizontal flipping for data augmentation. The CelebA dataset was further preprocessed with center cropping to 140 pixels as an additional augmentation step.

Baseline We employ the existing Method (Matsumoto et al., 2023) as our baseline. This approach uses the DDPM's training loss for membership inference by comparing generated Gaussian noise to predicted noise.

Metrics Performance assessment of our SeDID hinged on three pivotal metrics: Accuracy (ACC), Area Under the ROC Curve (AUC), and True Positive Rate at a fixed False Positive Rate (TPR@FPR). ACC accounts for the fraction of correct predictions, AUC delineates model's class discriminative ability with higher scores signifying better performance, and TPR@FPR is an indicator of model's sensitivity at a given false-positive rate.

Experimental Setup and Implementation Details In this work, we used diffusion models with the settings from (Ho et al., 2020) and performed a diffusion process of $T = 1000$ steps. The models were trained on the entire datasets with

1,200,000 steps for Tiny ImageNet and CelebA, and 800,000 steps for CIFAR10. We synthesized a balanced dataset of 50,000 diffusion-generated and 50,000 real images, selected from the training split of each dataset.

The $\text{SeDID}_{\text{NNs}}$ was trained on 10% of this data, using a ResNet18 as the backbone. The training phase extended for 20 epochs at a learning rate of 0.001. To manage the learning dynamics, we employed Stochastic Gradient Descent (SGD) with a momentum of 0.9 and weight decay of 5e-4. The batch size was set to 128.

The computations were performed on an Intel(R) Xeon(R) Gold 6248R CPU @ 3.00GHz and an NVIDIA GeForce RTX 3090 GPU, ensuring the reproducibility of our experiments.

4.3. Selection of Optimal T_{SE} and stepsize δ

In diffusion processes, the selection of optimal Stepwise Error Calculation Time Step, T_{SE}, and stepsize δ is critical, as they directly affect the (t,δ-error) - the difference between the expected and actual states of the synthetic image at time step t. These parameters are pivotal for optimizing the quality of synthetic image generation.

To determine the optimal T_{SE} and δ, we experimented extensively on CelebA, Tiny-ImageNet, and CIFAR10 datasets. Our aim, by exploring various combinations, was to achieve the best possible AUC and ACC scores.

Figure 2 depicts the performance of varying δ values across the datasets with best T_{SE}. The optimal performance is achieved at $\delta = 165$ for all datasets, thereby validating our choice.

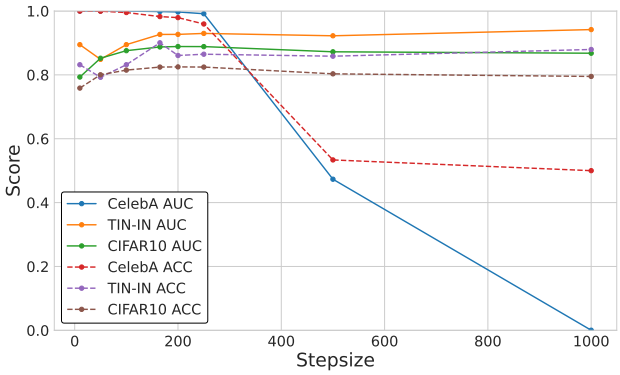

Figure 2. **Effect of different Stepsize on CelebA, TINY-IMAGENET, and CIFAR10 datasets.** The optimal Stepsize δ is highlighted with a triangle in each series.

4.4. Experimental Results

This subsection provides a comprehensive discussion of our experimental results conducted on three diverse datasets: CIFAR10, TinyImageNet, and CelebA. A visual representation of SeDID's performance across various Stepwise Error Calculation Time Steps, T_{SE}, for the STAT strategy on these datasets is depicted in Figure 3.

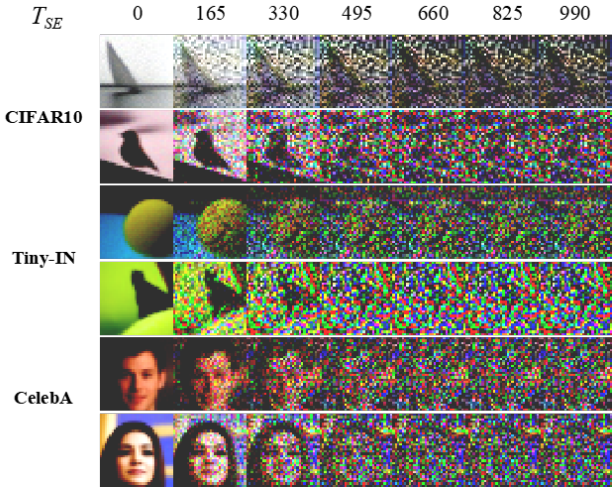

Figure 3. **Visual representation of the intermediate results of the SeDID method using the STAT strategy on the CIFAR10, Tiny-ImageNet, and CelebA datasets across various Stepwise Error Calculation Time Steps, T_{SE}, within the range of [0, 1000] with the stepsize δ = 165.** This figure illustrates the diffusion progression under the SeDID method across different Stepwise Error Calculation Time Steps, T_{SE}, at a fixed stepsize δ.

Table 1. **Performance of GID-Diff across different timesteps.** The table shows the increasing AUC and ACC values with the growth of diffusion time T.

	CIFAR10		TINY-IMAGENET		CELEBA	
T	AUC	ACC	AUC	ACC	AUC	ACC
165	0.4954	0.5089	0.1278	0.5000	**0.9985**	**0.9843**
330	0.5415	0.5279	0.0606	0.5000	0.4213	0.5103
495	0.5695	0.5579	0.5125	0.5240	0.2240	0.5001
660	0.5667	0.5845	0.4971	0.6059	0.1866	0.5489
825	0.8650	0.7992	0.9827	0.9615	0.0001	0.5000
990	**0.8875**	**0.8244**	**0.9998**	**0.9966**	0.0000	0.5000

As detailed in Table 1, SeDID consistently outperforms the existing method in detecting diffusion-generated images. This is demonstrated by its superior AUC and ACC values across all tested datasets. The results show the T_{SE} have significant impact to the performance. This observation underscores the necessity of selecting an optimal diffusion timestep to maximize detection performance and also highlights the potency of SeDID in effectively distinguishing diffusion-generated images.

Additionally, we illustrate the diverse performance capabilities of our proposed SeDID method with respect to a range of different timesteps T_{SE} in Figure 4. The results show as the value of T_{SE} progressively escalates, a distinct improvement in key performance metrics - TPR@FPR(1%) and TPR@FPR(0.1%) - is observed for the CIFAR10 and Tiny-

Table 2. **SeDID variants and the existing method (Matsumoto et al., 2023) comparison on DDPM.** The table highlights the superior AUC and ACC of SeDID variants.

	CIFAR10		TINY-IMAGENET		CELEBA		AVERAGE	
METHOD	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑	AUC↑	ACC↑
EXISTING METHOD	0.5293	0.5225	0.5303	0.5254	1.0000	1.0000	0.6865	0.6826
$\text{SeDID}_{\text{STAT}}$	0.8874	**0.8244**	0.9266	0.9004	0.9983	0.9825	0.9374	0.9024
$\text{SeDID}_{\text{NNs}}$	**0.8903**	0.8218	**0.9999**	**0.9980**	**1.0000**	**1.0000**	**0.9634**	**0.9399**

ImageNet datasets. In stark contrast, these performance metrics exhibit a decreasing trend on the CelebA dataset with an increasing T_{SE}. This intricate behavior not only underscores the inherent adaptability of the SeDID method, capable of fine-tuning T_{SE} to match the unique attributes of various datasets, but also hints at the potential dataset-dependence of the optimal T_{SE}. This interpretation suggests that a meticulous, deliberate selection of T_{SE} might be a powerful strategy in significantly enhancing the performance of synthetic image detection tasks.

4.5. Comparison to Baselines

In the context of limited research specifically dedicated to the detection of images generated by diffusion models, we employ the existing method (Matsumoto et al., 2023) as our benchmark. After evaluating our SeDID approach on the three datasets, the resulting AUC and ACC are presented in Table 2. It is evident that both $\text{SeDID}_{\text{Stat}}$ and $\text{SeDID}_{\text{NNs}}$ demonstrate superior performance over the existing method, particularly in terms of ACC with an average increase of over 30% observed for both SeDID variants. This underscores the effectiveness of SeDID, particularly when employing neural networks in the inference strategy, as indicated by the notably high performance of $\text{SeDID}_{\text{NNs}}$.Furthermore, $\text{SeDID}_{\text{NNs}}$ displays an even more pronounced performance improvement, outperforming $\text{SeDID}_{\text{Stat}}$ by over 7%. This demonstrates that training a neural network can significantly improve the performance in term of AUC and ACC.

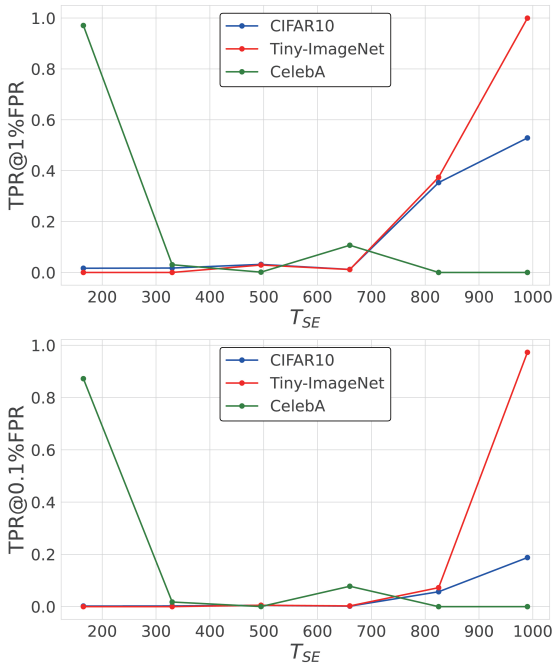

Figure 4. **Performance of the SeDID method at various timesteps.** This graph shows how SeDID's effectiveness varies with different time steps $T_{SE} \in [0, 1000]$. Key observations include a smaller improvement rate in TPR@FPR(0.1%) for CIFAR10 and Tiny-ImageNet as T_{SE} increases, and the highest performance for CelebA at small T_{SE}.

5. Conclusion and Future Work

In this paper, we have presented SeDID, a novel method for detecting diffusion-generated images. SeDID leverages unique attributes of diffusion models, specifically deterministic reverse and deterministic denoising errors, providing a powerful tool for image detection. Extensive experiments on three different datasets demonstrate SeDID's superior performance compared to existing approaches. This work contributes to the field of diffusion-generated image detection, laying the groundwork for future research. Looking forward, we intend to:

- Extend our approach to encompass other types of diffusion-based generation models, broadening its applicability.
- Investigate an automated mechanism for optimal timestep and T_{SE} selection, aiming to enhance the precision and effectiveness of detection.
- Explore the potential of a multi-step error computation approach in enhancing detection accuracy, leveraging our defined t-error concept.
- We plan to further investigate the influences on SeDID's performance across various datasets, specifically examining unusual trends in datasets like CelebA.

Through these explorations, we aspire to develop more secure and reliable systems capable of effectively identifying and neutralizing potential threats from the misuse of generative diffusion models.

References

C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot. Label-only membership inference attacks. In *International Conference on Machine Learning*, pages 1964–1974, 2021.

R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. On the detection of synthetic images generated by diffusion models. *arXiv preprint arXiv:2211.00680*, 2022.

R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva. On the detection of synthetic images generated by diffusion models. In *International Conference on Acoustics, Speech and Signal Processing*, pages 1–5, 2023.

P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794, 2021.

T. Dockhorn, A. Vahdat, and K. Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.

J. Duan, F. Kong, S. Wang, X. Shi, and K. Xu. Are diffusion models vulnerable to membership inference attacks? *arXiv preprint arXiv:2302.01316*, 2023.

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

J. Ho and T. Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851, 2020.

B. Hui, Y. Yang, H. Yuan, P. Burlina, N. Z. Gong, and Y. Cao. Practical blind membership inference attack via differential comparisons. *arXiv preprint arXiv:2101.01341*, 2021.

F. Kong, J. Duan, R. Ma, H. Shen, X. Zhu, X. Shi, and K. Xu. An efficient membership inference attack for the diffusion model by proximal initialization. *arXiv preprint arXiv:2305.18355*, 2023.

Z. Kong, W. Ping, J. Huang, K. Zhao, and B. Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

A. Krizhevsky, G. Hinton, et al. Learning multiple layers of features from tiny images, 2009.

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

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *International Conference on Computer Vision*, December 2015.

T. Matsumoto, T. Miura, and N. Yanai. Membership inference attacks against diffusion models. *arXiv preprint arXiv:2302.03262*, 2023.

C. Meng, Y. He, Y. Song, J. Song, J. Wu, J.-Y. Zhu, and S. Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2021.

M. Nasr, R. Shokri, and A. Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *IEEE Symposium on Security and Privacy*, pages 739–753, 2019.

A. Pyrgelis, C. Troncoso, and E. De Cristofaro. Knock knock, who's there? membership inference on aggregate location data. *arXiv preprint arXiv:1708.06145*, 2017.

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

S. Rezaei and X. Liu. On the difficulty of membership inference attacks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7892–7900, 2021.

J. Ricker, S. Damm, T. Holz, and A. Fischer. Towards the detection of diffusion model deepfakes. *arXiv preprint arXiv:2210.14571*, 2022.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou. White-box vs black-box: Bayes optimal strategies for membership inference. In *International Conference on Machine Learning*, pages 5558–5567, 2019.

C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, volume 35, pages 36479–36494, 2022.

A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes. Ml-leaks: Model and data independent

membership inference attacks and defenses on machine learning models. *arXiv preprint arXiv:1806.01246*, 2018.

T. Salimans and J. Ho. Progressive distillation for fast sampling of diffusion models. *arXiv preprint arXiv:2202.00512*, 2022.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *IEEE Symposium on Security and Privacy*, pages 3–18, 2017.

J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265, 2015.

J. Song, C. Meng, and S. Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020a.

L. Song and P. Mittal. Systematic evaluation of privacy risks of machine learning models. In *USENIX Security Symposium*, volume 1, page 4, 2021.

Y. Song and S. Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020b.

A. Tavanaei. Embedded encoder-decoder in convolutional networks towards explainable ai. *arXiv preprint arXiv:2007.06712*, 2020.

S. Truex, L. Liu, M. E. Gursoy, L. Yu, and W. Wei. Demystifying membership inference attacks in machine learning as a service. *IEEE Transactions on Services Computing*, 14(6):2073–2089, 2019.

Z. Wang, J. Bao, W. Zhou, W. Wang, H. Hu, H. Chen, and H. Li. Dire for diffusion-generated image detection. *arXiv preprint arXiv:2303.09295*, 2023.

D. Watson, W. Chan, J. Ho, and M. Norouzi. Learning fast samplers for diffusion models by differentiating through sample quality. In *International Conference on Learning Representations*, 2022.

Z. Xiao, K. Kreis, and A. Vahdat. Tackling the generative learning trilemma with denoising diffusion gans. *arXiv preprint arXiv:2112.07804*, 2021.

S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. Privacy risk in machine learning: Analyzing the connection to overfitting. In *IEEE Computer Security Foundations Symposium*, pages 268–282, 2018.

C. Yuan and M. Moghaddam. Attribute-aware generative design with generative adversarial networks. *Ieee Access*, 8:190710–190721, 2020.

C. Yuan, T. Marion, and M. Moghaddam. Dde-gan: Integrating a data-driven design evaluator into generative adversarial networks for desirable and diverse concept generation. *Journal of Mechanical Design*, 145(4):041407, 2023.

Z. Zhao, J. Duan, X. Hu, K. Xu, C. Wang, R. Zhang, Z. Du, Q. Guo, and Y. Chen. Unlearnable examples for diffusion models: Protect data from unauthorized exploitation. *arXiv preprint arXiv:2306.01902*, 2023.

Appendix

A. Generalization to Latent Diffusion Model (LDM)

For the Latent Diffusion Model (LDM), the calculation of *t-error* is similar to DDPM, except the intermediate latent variables v_t are in the latent space and the reverse process is conditioned by text embeddings. Specifically, we denote by V the Variational Autoencoders (VAEs) utilized to encode the original images into the latent space, i.e., $v_0 = V(x_0), x_0 \sim D$, and denote by C the text condition. The diffusion process and the denoising process can be derived as:

$$\begin{aligned} q(v_t|v_{t-1}) &= \mathcal{N}(v_t; \sqrt{1-\beta_t} v_{t-1}, \beta_t \mathbf{I}) \\ p_\theta(v_{t-1}|v_t) &= \mathcal{N}(v_{t-1}; \mu_\theta(v_t, t, C), \Sigma_\theta(v_t, t)). \end{aligned} \tag{9}$$

Then, the *t-error* can be rewritten as:

$$\tilde{\ell}_{t,v_0} = ||\psi_\theta(\phi_\theta(\tilde{v}_t, t, C), t, C) - \tilde{v}_t||^2, \tag{10}$$

where we reuse the symbols ϕ_θ and ψ_θ as the deterministic reverse and sampling regarding v_t:

$$\begin{aligned} v_{t+1} &= \phi_\theta(v_t, t, C) \\ &= \sqrt{\bar{\alpha}_{t+1}} f_\theta(v_t, t, C) + \sqrt{1-\bar{\alpha}_{t+1}} \epsilon_\theta(v_t, t, C), \end{aligned} \tag{11}$$

$$\begin{aligned} v_{t-1} &= \psi_\theta(v_t, t, C) \\ &= \sqrt{\bar{\alpha}_{t-1}} f_\theta(v_t, t, C) + \sqrt{1-\bar{\alpha}_{t-1}} \epsilon_\theta(v_t, t, C), \end{aligned} \tag{12}$$

where

$$f_\theta(v_t, t, C) = \frac{v_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(v_t, t, C)}{\sqrt{\bar{\alpha}_t}}. \tag{13}$$

B. Adopted Diffusion Model and Datasets

In our study, we employ CIFAE10, Tiny-ImageNet and CelebA on DDPM. The detailed settings are demonstrated in Table 3.

Table 3. **Selected diffusion model and datasets.** This table outlines the model and datasets utilized in our study, detailing the image resolution and the count of real and generated images for each pair.

Model	Dataset	Resolution	# Real Images	# Generated Images
DDPM	CIFAR-10	32	50,000	50,000
	Tiny-ImageNet	32	50,000	50,000
	CelebA	32	50,000	50,000

C. Assessing the Effectiveness of Defensive Training in Diffusion Models

We explore defensive training strategies for diffusion models and assess their effectiveness with the SeCID method. Specifically, we investigate aggressive regularization and data augmentation. The Figures 5, 6, and 7 showcase samples from the resulting models.

Despite the defensive training efforts, the generated images are disappointingly vague and unrealistic, a consequence of the aggressive regularization and data augmentation. Moreover, the quality of reconstructed images from both the member and hold-out sets is significantly compromised, posing substantial challenges to the effectiveness of Membership Inference Attacks (MIA). The results underscore a fundamental trade-off between model performance and privacy protection in the context of diffusion models.

Figure 5. **Samples from DDPM on CIFAR-10 at step 800,000.**

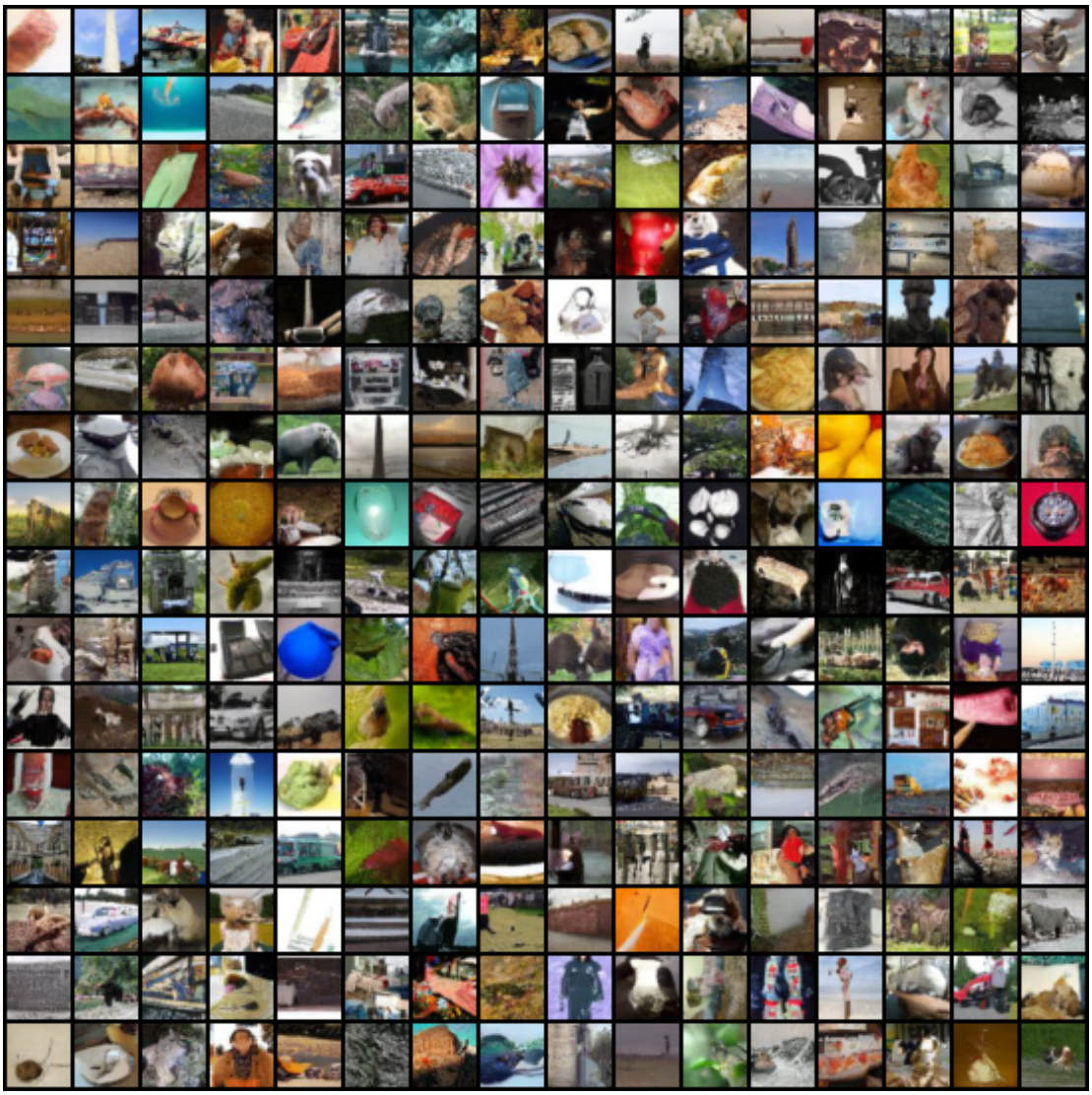

Figure 6. **Samples from DDPM on Tiny-ImageNet at step 1,200,000.**

Figure 7. **Samples from DDPM on CelebA at step 1,200,000.**