Loss and Likelihood Based Membership Inference of Diffusion Models

Hailong Hu[1 [0000−0001−5138−4014]] and Jun Pang[2,1 [0000−0002−4521−4112]]

[1] SnT, University of Luxembourg, Esch-sur-Alzette, Luxembourg
[2] FSTM, University of Luxembourg, Esch-sur-Alzette, Luxembourg
hailong.hu@uni.lu, jun.pang@uni.lu

Abstract. Recent years have witnessed the tremendous success of diffusion models in data synthesis. However, when diffusion models are applied to sensitive data, they also give rise to severe privacy concerns. In this paper, we present a comprehensive study about membership inference attacks against diffusion models, which aims to infer whether a sample was used to train the model. Two attack methods are proposed, namely loss-based and likelihood-based attacks. Our attack methods are evaluated on several state-of-the-art diffusion models, over different datasets in relation to privacy-sensitive data. Extensive experimental evaluations reveal the relationship between membership leakages and generative mechanisms of diffusion models. Furthermore, we exhaustively investigate various factors which can affect membership inference. Finally, we evaluate the membership risks of diffusion models trained with differential privacy.

Keywords: Membership inference attacks · Diffusion models · Human face synthesis · Medical image generation · Privacy threats

1 Introduction

Diffusion models [34] have recently made remarkable progress in image synthesis [16,19,38], even being able to generate better-quality images than generative adversarial networks (GANs) [11] in some situations [8]. They have also been applied to sensitive personal data, such as the human face [19, 37] or medical images [21,30], which might unwittingly lead to the leakage of training data. As a consequence, it is paramount to study privacy breaches in diffusion models.

Membership inference (MI) attacks aim to infer whether a given sample was used to train the model [33]. In practice, they are widely applied to analyze the privacy risks of a machine learning model [27,35]. To date, a growing number of studies concentrate on classification models [2,25,32,33,40], GANs [6,13], text-to-image generative models [39], and language models [4,5]. However, there is still a lack of work on MI attacks against diffusion models. In addition, data protection regulations, such as GDPR [29], require that it is mandatory to assess privacy

Our code is available at: https://github.com/HailongHuPri/MIDM.

threats of technologies when they are involving sensitive data. Therefore, all of these drive us to investigate the membership vulnerability of diffusion models.

In this paper, we systematically study the problem of membership inference of diffusion models. Specifically, we consider two threat models: in threat model I, adversaries are allowed to obtain the target diffusion model, and adversaries also can calculate the loss values of a sample through the model. This scenario might occur when institutions share a generative model with their collaborators to avoid directly sharing original data [24, 28]. We emphasize that obtaining losses of a model is realistic because it is widely adopted in studying MI attacks on classification models [2, 25, 33, 40]. In threat model II, adversaries can obtain the likelihood value of a sample from a diffusion model. Providing the exact likelihood value of any sample is one of the advantages of diffusion models [38]. Thus, here we aim to study whether the likelihood value of a sample can be considered as a clue to infer membership. Based on both threat models, two types of attack methods are developed respectively: loss-based attack and likelihood-based attack. They are detailed in Section 3.

We evaluate our methods on four state-of-the-art diffusion models: DDPM [16], SMLD [37], VPSDE [38] and VESDE [38]. We use two privacy-sensitive datasets: a human face dataset FFHQ [20] and a diabetic retinopathy dataset DRD [18]. Extensive experimental evaluations show that our methods can achieve excellent attack performance, and provide novel insights into membership vulnerabilities in diffusion models (see Section 5). For instance, the loss-based attack demonstrates that different diffusion steps of a diffusion model have significantly different privacy risks, and there exist high-risk regions which lead to leakage of training samples. The likelihood-based attack shows that the likelihood values of samples from a diffusion model provide a strong indication to infer training samples. We also analyze attack performance with respect to various factors in Section 6. For example, we find that the high-risk regions still exist with the increase in the number of training samples (see Figure 5). This indicates that it is urgent to redesign the current noise mechanisms used by almost all diffusion models. Finally, we evaluate our attack performance on a classical defense - differential privacy [10] (see Section 7). Specifically, we train target models using differentially-private stochastic gradient descent (DP-SGD) [1]. Extensive evaluations show that although the performance of both types of attack can be alleviated on models trained with DP-SGD, they sacrifice too much model utility, which also gives a new research direction for the future.

Our contributions in this paper are twofold. (1) We propose two types of attacks to infer the membership of diffusion models. Our attack methods reveal the relationship between the leakage of training samples and the generative mechanism of diffusion models. (2) We evaluate our attacks on one classical defense — diffusion models trained with DP-SGD, showing that it mitigates our attacks at the cost of the quality of synthetic samples.

In the end, we want to emphasize that although we study membership inference from the perspective of attackers, our proposed methods can directly be

applied to audit the privacy risks of diffusion models when model providers need to evaluate the privacy risks of their models.

2 Background: Diffusion Models

Diffusion models [34] are a class of probabilistic generative models. They aim to learn the distribution of a training set, and the resulting model can be utilized to synthesize new data samples.

In general, a diffusion model includes two processes: a forward process and a reverse process [34]. In the forward process, i.e. the diffusion process, it aims to transform a complex data distribution p_{data} into a simple prior distribution, e.g. Gaussian distribution $\mathcal{N}(0,\, \sigma^2\mathrm{I})$, by gradually adding different levels of noise $0 = \sigma_0 < \sigma_1 <, ..., < \sigma_T = \sigma_{max}$, into the data x. In the reverse process, it targets at synthesizing a new data sample $\tilde{x}_0$ through step by step denoising a data sample $\tilde{x}_T \sim \mathcal{N}(0,\, \sigma^2_{max}\mathrm{I})$. Both processes are defined as Markov chains, and the transitions from one step to another step are described by transition kernels. In the following, we briefly introduce three typical diffusion models.

DDPM. A denoising diffusion probabilistic model (DDPM) proposed by Ho et al. [16] defines the forward process: $q(x_1, ..., x_T|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1})$,where T is the number of diffusion steps. The transition kernel uses a Gaussian transition kernel: $q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t\mathrm{I})$, where the hyperparameter $\beta_t \in (0, 1)$ is a variance schedule. Based on the transition kernel, we can get a perturbed sample by: $x_t \leftarrow \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, \mathrm{I})$. The transition kernel from the initial step to any t step can be expressed as: $q(x_t|x_0) = \mathcal{N}(x_t; \sqrt{\bar{\alpha}_t}x_0, (1-\bar{\alpha}_t)\mathrm{I})$, where $\bar{\alpha}_t = \prod_{i=0}^{t} \alpha_i$ and $\alpha_t := 1 - \beta_t$. Therefore, any perturbed sample can be obtained by: $x_t \leftarrow \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\varepsilon$. In the reverse process, DDPM generates a new sample by: $\tilde{x}_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(\tilde{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\tilde{x}_t, t)) + \sigma_t\epsilon$, where $\epsilon_\theta(x_t, t)$ is a neural network predicting noise. In practice, DDPM is trained by minimizing the following loss:

$$L(\theta) = \mathbb{E}_{t\sim[1,T], x\sim p_{data}, \varepsilon\sim\mathcal{N}(0,\mathrm{I})}[||\varepsilon - \varepsilon_\theta(\sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)||^2]. \tag{1}$$

SMLD. Score matching with Langevin dynamics (SMLD) [37] first learns to estimate the *score*, then generates new samples by Langevin dynamics. The *score* refers to the gradient of the log probability density with respect to data, i.e. $\nabla_x log\ p(x)$. The transition kernel in the forward process is: $q(x_t|x_0) = \mathcal{N}(x_t; x_0, \sigma_t^2\mathrm{I})$. Thus, a perturbed sample is obtained by: $x_t \leftarrow x_0 + \sigma_t\epsilon$. In the reverse process, SMLD uses an annealed Langevin dynamics to generate a new sample by: $\tilde{x}_t \leftarrow \tilde{x}_{t-1} + \frac{\alpha_i}{2}s_\theta(\tilde{x}_{t-1}, \sigma_i) + \sqrt{\alpha_i}\epsilon$, where the hyperparameter σ_i controls the updating magnitudes and $s_\theta(x_t, \sigma_i)$ is a noise conditioned neural network predicting the *score*. Training of the SMLD is performed by minimizing the following loss:

$$L_\theta = \mathbb{E}_{t\sim[1,T], x\sim p_{data}, x_t\sim q(x_t|x)}[\lambda(\sigma_t)||s_\theta(x_t, \sigma_t) - \nabla_{x_t}\log q(x_t|x)||^2], \tag{2}$$

where $\lambda(\sigma_t)$ is a coefficient function and $\nabla_{x_t}\log q(x_t|x) = -\frac{x_t - x}{\sigma_t^2}$.

SSDE. Unlike prior works DDPM or SMLD which utilize a finite number of noise distributions, i.e. t is discrete and usually at most T, Song et al. [38] propose a score-based generative framework through the lens of stochastic differential equations (SDEs), which can add an infinite number of noise distributions to further improve the performance of generative models. The forward process which adds an infinite number of noise distributions can be described as a continuous-time stochastic process. Specifically, the forward process of the score-based SDE (SSDE) is defined as:

$$\mathrm{d}x = \mathrm{f}(x,t)dt + g(t)\mathrm{d}w, \tag{3}$$

where $\mathrm{f}(x,t)$, $g(t)$ and $d\mathrm{w}$ are the drift coefficient, the diffusion coefficient and a standard Wiener process, respectively. The reverse process corresponds to a reverse-time SDE: $dx = [\mathrm{f}(x,t) - g(t)^2\nabla_x \log\, q_t(x)]\mathrm{d}t + g(t)\mathrm{d}\bar{w}$, where $\bar{w}$ is a standard Wiener process in the reverse time. Training of the SSDE is performed by minimizing the following loss:

$$L_\theta = \mathbb{E}_{t\in\mathcal{U}(0,T),x\sim p_{data},x_t\sim q(x_t|x)}[\lambda(t)||s_\theta(x_t,t) - \nabla_{x_t}\log\, q(x_t|x)||^2]. \tag{4}$$

The SSDE is a general and unified framework. Based on different coefficients in Equation 3, the variance preserving (VP) and variance exploding (VE) are instantiated. The VPSDE is defined as: $\mathrm{d}x = -\frac{1}{2}\beta(t)x\mathrm{d}t + \sqrt{\beta(t)}\mathrm{d}w$. The VESDE is defined as: $dx = \sqrt{\frac{\mathrm{d}[\sigma^2(t)]}{\mathrm{d}t}}\mathrm{d}w$. Furthermore, the SSDE also shows the noise perturbations of DDPM and SMLD are discretizations of VP and VE, respectively. Note that, diffusion steps usually used in diffusion models also refer to time steps that are used in SDEs. In this work, we study the privacy risks of four target models: DDPM, SMLD, VPSDE, and VESDE.

3 Methodology

The objective of MI attacks is to infer if a sample was used to train a model. This provides model providers with a method to evaluate the information leakage of a machine learning model. In this section, we first introduce threat models and then present our MI methods.

3.1 Threat Models

Threat Model I. In this setting, we assume adversaries can only obtain the target model, i.e. the victim diffusion model. This setting is realistic because institutions might share generative models with their collaborators instead of directly utilizing original data, considering privacy threats or data regulations [24, 28]. We emphasize that adversaries do not gain any knowledge of the training set. Obtaining the target model indicates that adversaries can get the loss values through the model, and this is realistic because most MI attacks on classification models also assume adversaries can get loss values [2, 25, 33, 40]. Under this threat model, we propose a loss-based MI attack.

Threat Model II. In this setting, we assume adversaries can have access to the likelihood values of samples from a diffusion model. Diffusion models have advantages in providing the exact likelihood value of any sample [38]. Here we aim to study whether the likelihood values of samples can be utilized as a signal to perform membership inference. Under this threat model, we propose a likelihood-based MI attack.

3.2 Intuition

We propose MI attacks based on the following two intuitions.

Intuition I. As introduced in Section 2, a diffusion model aims to minimize the loss values over the training set in the training phase. One intuition is that member samples, i.e. the training samples, should have smaller loss values, compared to nonmember samples. This is because training/member samples involve the training process and their loss values could be minimized.

Intuition II. A diffusion model is a generative model that learns the distribution of a training set. Therefore, the likelihood values of training/member samples should be higher than these of nonmember samples. This is because training/member samples are from the distribution of the training set.

3.3 Attack Methods

Problem Formulation. Given a target diffusion model G_{tar}, the objective of MI attacks is to infer whether a sample x from a target dataset X_{tar} is used to train the G_{tar}.

Loss-based Attack. For threat model I and following intuition I, we develop a loss-based attack. As illustrated in Section 2, diffusion models can add an infinite or finite number of noise distributions, which are corresponding to continuous or discrete SDE, respectively. Therefore, we can calculate the loss value of a sample at each diffusion step t. Specifically, based on Equation 1, the loss of a sample x at t diffusion step of DDPM is calculated by:

$$L = \frac{1}{m} \sum ||\varepsilon - \varepsilon_{\theta^\star}(\sqrt{\bar{\alpha}_t}x + \sqrt{1-\bar{\alpha}_t}\varepsilon, t)||^2, \tag{5}$$

where m is the dimension of x and $\varepsilon_{\theta^\star}(.)$ is the trained network. By Equation 2, the loss of a sample x at t diffusion step of SMLD is calculated by:

$$L = \frac{1}{m} \sum \lambda(\sigma_t)||s_{\theta^\star}(x_t, \sigma_t) - \nabla_{x_t} \log q(x_t|x)||^2, \tag{6}$$

where $s_{\theta^\star}(.)$ is the trained network. Based on Equation 4, the loss of a sample x at t diffusion step of VPSDE and VESDE is:

$$L = \frac{1}{m} \sum \lambda(t)||s_{\theta^\star}(x_t, t) - \nabla_{x_t} \log\, q(x_t|x)||^2. \tag{7}$$

Then, we make a membership inference directly based on the loss value of a sample at one diffusion step. Namely, if a sample's loss value is less than certain thresholds, this sample is marked as a member sample. For one sample, we can get T or infinite losses, depending on continuous or discrete SDEs. In this work, in order to thoroughly demonstrate the performance of our attack, we compute losses of all diffusion steps T for the discrete case. We randomly select T diffusion steps for the continuous case although it has infinite steps.

Likelihood-based Attack. For threat model II and following intuition II, we propose to utilize the likelihood value of a sample to infer membership. We compute the log-likelihood of a sample x based on the equation proposed by [38].

$$\log p(x) = \log p_T(x_T) - \int_0^T \nabla \cdot \tilde{\mathrm{f}}_{\theta^\star}(x_t, t)dt, \tag{8}$$

where $\nabla \cdot \tilde{\mathrm{f}}_{\theta^\star}(x, t)$ is estimated by the Skilling-Hutchinson trace estimator [12]. If the log-likelihood value of a sample is higher than certain thresholds, this sample is predicted as a member sample. As introduced in Section 2, the work SSDE [38] is a unified framework. In other words, DDPM, SMLD, VPSDE and VESDE can be described by Equation 3. Therefore, Equation 8 can be applied to these models to estimate the likelihood of one sample. In this work, we compute the likelihood values of all training samples.

4 Experiments

4.1 Datasets

We use two different datasets to evaluate our attack methods. They cover the human face and medical images, which are all considered privacy-sensitive data.

FFHQ. The Flickr-Faces-HQ dataset (FFHQ) [20] is a new dataset that contains $70,000$ high-quality human face images. In this work, we randomly choose $1,000$ images to train target models. We also explore the effect of the size of the training set in Section 6.1.

DRD. The Diabetic Retinopathy dataset (DRD) [18] contains $88,703$ retina images. In this work, we only consider images that have diabetic retinopathy, which is a total of $23,359$ images. Furthermore, we randomly choose $1,000$ images to train target models. Note that images in all datasets are resized to 64×64 just for the purpose of computation efficiency.

4.2 Metrics

Evaluation metrics for diffusion models. We use the popular Fréchet Inception Distance (FID) metric to evaluate the performance of a diffusion model [14]. A lower FID score is better, which implies that the generated samples are more realistic and diverse. Considering the efficiency of sampling, in our work the FID score is computed with all training samples and $1,000$ generated samples.

Evaluation metrics for MI attacks. We primarily use the full log-scale receiver operating characteristic (ROC) curve to evaluate the performance of our attack methods, because it can better characterize the worst-case privacy threats of a victim model [2]. We also report the true-positive rate (TPR) at the false-positive rate (FPR) as it can give a quick evaluation. We use average-case metrics — accuracy as a reference, although it cannot assess the worst-case privacy.

4.3 Experimental Setups

In terms of target models, we use open source codes [36] to train diffusion models, and their recommended hyperparameters about training and sampling are adopted. More specifically, the number of training steps for all models is fixed at $500,000$. For discrete SDEs, T is fixed as $1,000$ while T is set as 1 for continuous SDEs. In terms of our attack methods, we evaluate the attack performance using all training samples as member samples and equal numbers of nonmember samples.

5 Evaluation

5.1 Performance of Target Models

Considering their excellent performance in image generation, we choose DDPM [16], SMLD [37], VPSDE [38] and VESDE [38] as our target models. They are trained on the FFHQ dataset containing 1k samples. Target models with the best FID during the training progress are selected to be attacked. Table 1 shows the performance of the target models. Figure 9 in Appendix shows the qualitative results for these target models. Overall, all target models can synthesize high-quality and realistic images.

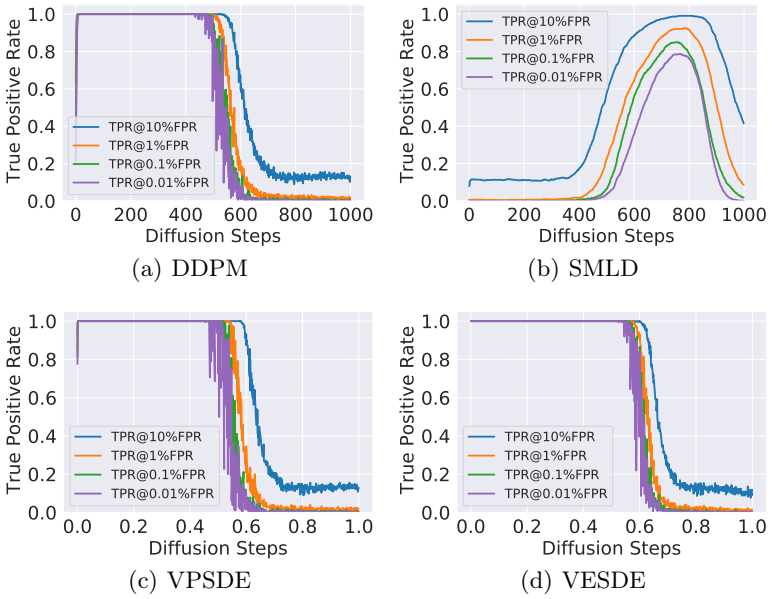
Table 1: The performance of target models on FFHQ.

Target Models	DDPM	SMLD	VPSDE	VESDE
FID	57.88	92.81	20.27	63.37

5.2 Performance of Loss-based Attack

We present our attack performance from two aspects: TPRs at fixed FPRs for all diffusion steps and log-scale ROC curves at one diffusion step. The former aims to provide the holistic performance of our attacks in diffusion models. In contrast, the latter concentrates on one diffusion step and is able to exhaustively show TPR values at a wide range of FPR values, which is key to assessing the worst-case privacy risks of a model.

TPRs at fixed FPRs for all diffusion steps. Figure 1 shows the performance of our loss-based attack on four target models trained on FFHQ. We plot TPRs at different FPRs with regard to diffusion steps for each target model. Recall DDPM and SMLD models are discrete SDEs while VPSDE and VESDE models are continuous SDEs. Thus, the number of diffusion steps for DDPM and SMLD



(a) DDPM (b) SMLD

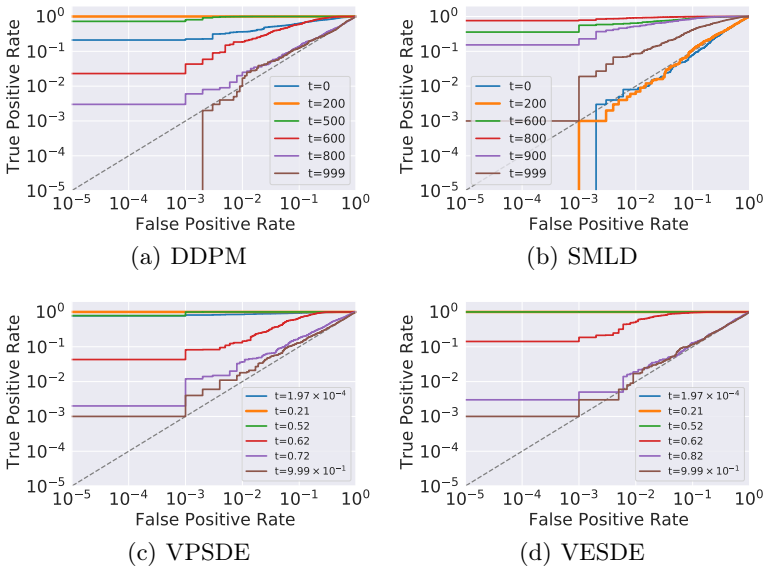
(c) VPSDE (d) VESDE

Fig. 1: Performance of the loss-based attack on all diffusion steps. Target models are trained on FFHQ.

is finite and is fixed as 1,000, while for VPSDE and VESDE models, we uniformly generate $1,000$ points within $[0, 1]$ and compute corresponding losses. Overall, all models are vulnerable to our attacks, even under the worst-case, i.e. TPR at 0.01% FPR, depicted by the purple line of Figure 1.

We observe that *our attack presents different performances in different diffusion steps*. To be more specific, there exist high privacy risk regions for diffusion models in terms of diffusion steps. In these regions (i.e. diffusion steps), our attack can achieve as high as 100% TPR at 0.01% FPR. Even for the SMLD model, close to 80% TPR at 0.01% FPR can be achieved. Recall the training mechanisms of diffusion models, different levels of noise at different diffusion steps are added during the forward process. DDPM and VPSDE and VESDE are added growing levels of noise while SMLD starts with maximum levels of noise and gradually decreases the levels of noise. Thus, we can see that these models (DDPM and VPSDE, and VESDE) are more vulnerable to leak training samples in the first half part of the diffusion steps while the SMLD model shows membership vulnerability in the second half part of the diffusion steps.

In brief, all models are prone to suffer from membership leakage in low levels of noise while they become more resistant in high levels of noise. In fact, in these diffusion steps where high levels of noise are added to training data, perturbed data is almost close to pure Gaussian noise, which can enhance the privacy of training data to some degree. We also notice that at the starting diffusion step, our attack performance suffers from a decrease. This is because there is an instability issue at this step during the training process [38]. Despite this, these peak regions still show the effectiveness of our attack.



(a) DDPM (b) SMLD

(c) VPSDE (d) VESDE

Fig. 2: Performance of the loss-based attacks at one diffusion step. Target models are trained on FFHQ.

Log-scale ROC curves at one diffusion step. Figure 2 plots full log-scale ROC curves of the loss-based attack on four target models. We choose six different diffusion steps for each target model. The rules of choosing diffusion steps for discrete SDEs (i.e. DDPM and SMLD) are: starting and ending diffusion step and the diffusion step that experiences significant changes in terms of attack performance. For continuous SDEs (i.e. VPSDE and VESDE), we first get $1,000$ points that are uniformly sampled from $[0, 1]$. Then, we choose diffusion steps from these points based on the same rule of discrete SDEs. Overall, our excellent attack performance is exhaustively shown through log-scale ROC curves.

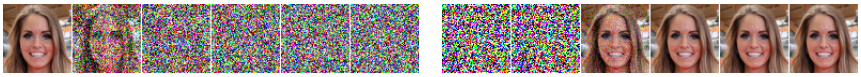
We can observe that when the levels of noise are not too large, our method can achieve a perfect attack, such as at $t = 200$ for the DDPM model, $t = 800$ for the SMLD model, and $t = 0.21$ for the VPSDE and VESDE models. Again, we can clearly see that the ROC curves on all target models are more aligned with the grey diagonal line with the increase in the magnitudes of noise. The grey diagonal line means that the attack performance is equivalent to random guesses. For example, the ROC curves are almost close to the grey diagonal line when the maximal level of noise is added, such as the DDPM model at $t = 999$, the SMLD model at $t = 0$, and the VPSDE and VESDE models at $t = 9.99 \times 10^{-1}$. It is not surprising because at that time the input samples are perturbed as Gaussian noise data in theory and indeed do not have something with original training samples.

Table 2 summarizes our attack performance on four target models with regard to diffusion steps and FPR values. We also report the average metric accuracy for reference. Here, we emphasize that only focusing on average metrics cannot

Table 2: Performance of the loss-based attack on target models trained on FFHQ.

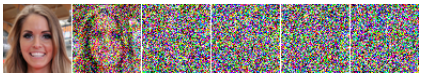
Models	T	TPR@ 10%FPR	TPR@ 1%FPR	TPR@ 0.1%FPR	TPR@ 0.01%FPR	Accuracy	Models	T	TPR@ 10%FPR	TPR@ 1%FPR	TPR@ 0.1%FPR	TPR@ 0.01%FPR	Accuracy
DDPM	0	63.50%	36.40%	22.50%	21.10%	78.25%	SMLD	0	7.90%	0.80%	0.00%	0.00%	51.20%
	200	100.00%	100.00%	100.00%	100.00%	100.00%		200	11.20%	0.70%	0.10%	0.00%	52.30%
	500	100.00%	99.50%	80.80%	72.50%	99.30%		500	88.50%	64.40%	56.10%	35.70%	89.50%
	600	59.50%	18.80%	4.30%	2.30%	81.15%		800	99.10%	91.70%	78.60%	76.10%	96.40%
	800	13.90%	2.50%	0.60%	0.30%	52.80%		900	85.80%	52.00%	22.80%	15.30%	88.80%
	999	12.60%	1.70%	0.00%	0.00%	52.45%		999	41.50%	8.60%	1.90%	0.10%	70.55%
VPSDE	1.97×10^{-4}	93.00%	85.00%	81.60%	77.60%	93.15%	VESDE	1.97×10^{-4}	100.00%	100.00%	100.00%	100.00%	100.00%
	0.21	100.00%	100.00%	100.00%	100.00%	100.00%		0.21	100.00%	100.00%	100.00%	100.00%	100.00%
	0.52	100.00%	100.00%	99.50%	78.40%	99..90%		0.52	100.00%	100.00%	100.00%	99.90%	99.95%
	0.62	66.50%	14.50%	8.20%	4.30%	85.70%		0.62	96.00%	53.60%	18.60%	14.20%	93.25%
	0.72	17.90%	3.70%	1.20%	0.20%	57.30%		0.82	13.10%	1.90%	0.50%	0.30%	52.50%
	9.99×10^{-1}	13.00%	1.8%	0.40%	0.10%	52.20%		9.99×10^{-1}	11.60%	1.70%	0.30%	0.10%	51.50%

assess the worst-case privacy risks. For instance, for the DDPM model at $t = 800$, the attack accuracy is 52.80%, which indicates the model at this diffusion step almost does not lead to the leakage of training samples, because it is close to 50% (the accuracy of random guesses). In fact, the TPR is 0.3% at the false positive rate of 0.01%, which is 30 times more powerful than random guesses. It means that adversaries can infer confidently member samples under extremely low false positive rates.



(a) DDPM

(b) SMLD

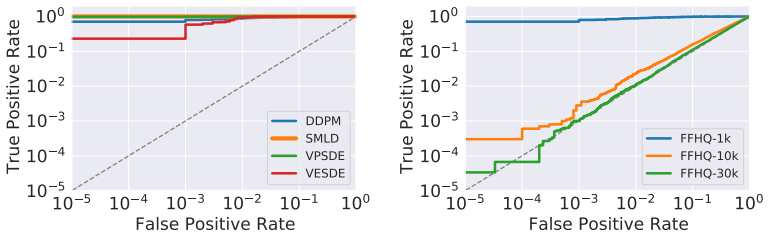


(c) VPSDE

(d) VESDE

Fig. 3: Perturbed data of four target models under different diffusion steps. The diffusion steps correspond to these in Figure 2. Specifically, from left to right for each model: DDPM (0, 200, 500, 600, 800, 999); SMLD (0, 200, 600, 800, 900, 999); VPSDE (1.97×10^{-4}, 0.21, 0.52, 0.62, 0.72, 9.99×10^{-1}); VESDE (1.97×10^{-4}, 0.21, 0.52, 0.62, 0.82, 9.99×10^{-1}).

Figure 3 shows perturbed data of four target models under different diffusion steps. The diffusion steps in Figure 3 are corresponding to these in Figure 2. We observe that even when some perturbed data that is almost not recognized by human beings is used to train the model, it seems not to prevent model memorization. For example, for the DDPM model at $t = 600$, the perturbed image is meaningless for humans. However, the attack accuracy is as high as 81.15%. At the same time, the TPR at 0.01% FPR is 2.30%, which is 230 times more powerful times than random guesses. It indicates that models trained on perturbed data, except for Gaussian noise data, can still leak training samples. *The noise mechanism of diffusion models does not provide privacy protection.*



(a) Likelihood-based attack on different target models.

(b) Likelihood-based attacks on models trained on different sizes of datasets.

Fig. 4: Performance of the likelihood-based attack.

5.3 Performance of Likelihood-based Attack

Figure 4(a) demonstrates our likelihood-based attack performance on four target models. Overall, our attacks still perform well on all target models. For example, our attack on the SMLD and VPSDE models almost remains 100% true positive rates on all false positive rate regimes. For the VESDE model, attack results are slightly inferior to the SMLD model, yet still higher than the 10% true positive rate at an extremely low 0.001% false positive rate.

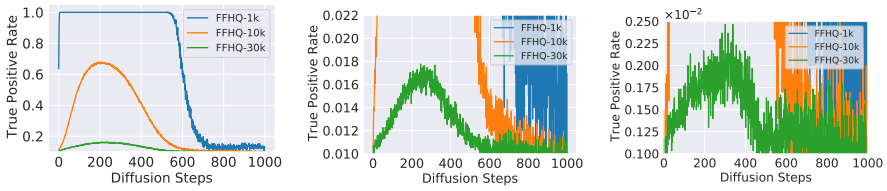
Table 3 shows our attack results at different FPR values for all target models. Once again, we can clearly see that even at the 0.01% FPR, the lowest TPR among all models is as high as 23.10%, which is 2,310 times than random guesses. In addition, we also observe that the attack accuracy is above 98% for all target models. Our attack results also remind model providers that they should be careful when using likelihood values.

Table 3: Likelihood-based attack. Target models are trained on FFHQ.

Models	TPR@ 10%FPR	TPR@ 1%FPR	TPR@ 0.1%FPR	TPR@ 0.01%FPR	Accuracy
DDPM	98.00%	89.00%	79.70%	71.00%	95.75%
SMLD	100.00%	100.00%	100.00%	100.00%	100.00%
VPSDE	100.00%	99.60%	98.90%	98.20%	99.45%
VESDE	100.00%	93.80%	58.40%	23.10%	98.50%

5.4 Takeaways

Our loss-based attack utilizes loss values to make a membership inference. Although the loss-based attack requires adversaries to choose a suitable diffusion step to mount the attack, our extensive experiments identify the high privacy risk region. More importantly, our loss-based attack reveals the relationship between membership risks and the generative mechanism of diffusion models. This provides a new angle to mitigate membership risks by designing novel noise mechanisms of diffusion models. Our likelihood-based attack does not need to choose a diffusion step and infers membership directly based on likelihood values. Both loss and likelihood information can lead to the leakage of training samples.



(a) TPR at 10% FPR (b) TPR at 1% FPR (c) TPR at 0.1% FPR

Fig. 5: Performance of loss-based attack with different sizes of datasets. The target model is DDPM trained on FFHQ. Each subfigure shows attack performance with different sizes of datasets on fixed FPRs.

6 Analysis

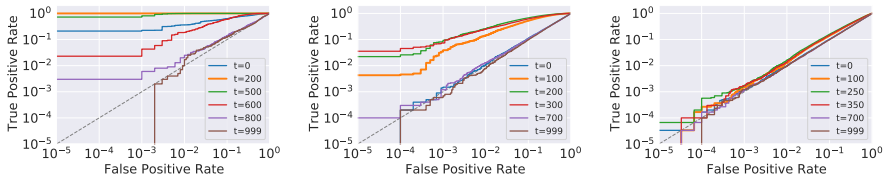
6.1 Effects of Size of a Training Dataset

We study attack performance with regard to different sizes of the training set of a target model. Here, we choose the DDPM models trained on FFHQ as target models. We use FFHQ-1k, FFHQ-10k, and FFHQ-30k to represent different sizes of a dataset, which refer to $1,000$, $10,000$, and $30,000$ training samples in each dataset respectively. The FID of the target model DDPM trained on FFHQ-1k, FFHQ-10k, and FFHQ-30k are 57.88, 34.34, and 24.06, respectively. In the following, we present the performance of our both attacks.

Performance of loss-based attack. Figure 5 depicts the performance of loss-based attacks on all diffusion steps under different sizes of a training set. Overall, we can observe that attack performance gradually becomes weak when the size of training sets increases. For example, at diffusion step $t = 200$, the TPR at 10% FPR decreases from 100% to about 15% when the training samples increase from 1k to 30k. Here, note that the starting points of the y-axis in Figure 5 are not 0 and we set them as the probability of random guesses. Thus, as long as the lines can be shown in the figure, it indicates this is an effective attack.

However, *the peak regions still exist even if the number of training samples increases to 30k and the FPR value is as low as 0.1%.* For instance, as shown in Figure 5(c), it shows our attack performance of 0.1% FPR on all models. Diffusion steps in the range of 0 to 400 are still vulnerable to our attack, compared to other steps. It indicates that these diffusion steps indeed lead a model to more easily leak training data. We further show the attack performance based on each dataset in Figure 11 in Appendix.

Figure 6 shows ROC curves of our attack against target models trained on different sizes of training sets. Based on the same rules described in Section 5.2, we select several different diffusion steps and plot their ROC curves. On the one hand, we can see that indeed models become less vulnerable as the number of training samples increases. For instance, Figure 6(c) shows the DDPM trained on FFHQ-30K is more resistant to MI attacks on the full log-scale TPR-FPR curve. On the other hand, when diffusion step t equals 250, our attack shows higher attack performance than random guesses at the low false positive rate, such as 10^{-4}. This is also corresponding to the peak steps in Figure 5.



(a) FFHQ-1k (b) FFHQ-10k (c) FFHQ-30k

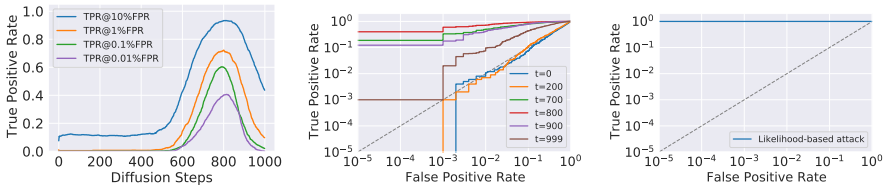
Fig. 6: Performance of loss-based attack with different sizes of datasets. The target model is DDPM. TPR-FPR Curves under different time steps.

We also observe from Figure 6 that TPR values in diffusion steps of high privacy risks do not further go down with the increase in FPR values, especially in extremely low FPR regimes. Take the DDPM trained on FFHQ-30K as an example (see Figure 6(c)), the TPR value at diffusion step $t = 250$ are still about 10^{-4} at the FPR value of 10^{-5}, while at $t = 999$, the TPR value at 10^{-5} FPR value is 0. This indicates that at $t = 250$, there are some training samples whose loss values are always smaller than the minimal loss value of the nonmember sample. Otherwise, the green line ($t = 250$) will go down to zero, similar to the brown line ($t = 999$). In other words, there are partial training samples that can be inferred with 100% confidence at this diffusion step. Note that in reality, even if only one sample can be inferred as a member confidently, it still constitutes a severe privacy violation [2, 17, 22].

Performance of likelihood-based attack. Figure 4(b) shows the performance of likelihood-based attacks in terms of different sizes of training sets. Similar to the loss-based attack, the performance of the likelihood-based attack decrease with an increase in the sizes of training sets. Specifically, the likelihood-based attack shows perfect performance on the target model trained on FFHQ-1k. When the size of a training set increases to 10K, there is a significant drop but still better than random guesses on the full log-scale ROC curve. In particular, in the extremely low false positive rate regime, such as 10^{-4}, the true positive rate is about 6×10^{-4}, which is 6 times more powerful than random guesses. In the model trained on FFHQ-30K, the ROC curve is almost close to the diagonal line, which indicates that adversaries are difficult to infer member samples through likelihood values.

6.2 Effects of Different Datasets

In this subsection, we show our attack performance on a medical image dataset about diabetic retinopathy. We choose the medical image dataset because the number of images that have diabetic retinopathy is usually insufficient in practice [21]. These types of images could be used for training a diffusion model and later the trained model is utilized to generate more novel images. We have described this dataset DRD in Section 4.1. We choose the SMLD as the target model and the number of training samples is $1,000$. Overall, the SMLD model can achieve excellent performance in image synthesis, with an FID of 33.20. Figure 10 in Appendix visualizes synthetic samples, which all show good quality.



(a) Loss-based attack (b) Loss-based attack (c) Likelihood-based attack

Fig. 7: Attack performance on the DRD dataset.

Performance of loss-based attack. Figure 7 shows the performance of loss-based attacks for the target model SMLD trained on DRD. Here, note that the levels of the noise of the SMLD model gradually become small with an increase in diffusion steps. Figure 7(a) shows the performance of our loss-based attack on all diffusion steps. Figure 7(b) depicts ROC curves for different diffusion steps on target model SMLD trained on DRD. We can again observe our attacks can still perform perfectly on DRD at diffusion steps of low levels of noise.

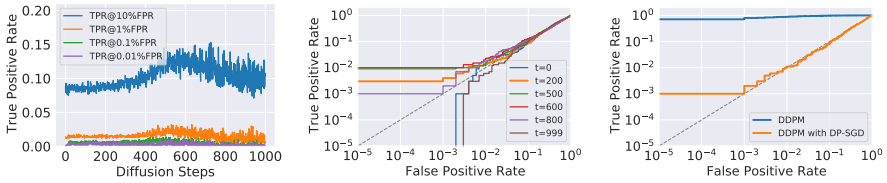
Performance of likelihood-based attack. Figure 7(c) reports the performance of our likelihood-based attack on the SMLD model trained on DRD. As expected, our attack still shows excellent performance. We can clearly find that the attack achieves 100% TPR on all FPR values, which means that all member samples are inferred correctly. Table 4 in Appendix reports the quantitative results of both attacks.

7 Defenses

Differential privacy (DP) [1, 10] is considered as a common defense measure for preventing the leakage of training samples of a machine learning model. In this section, we present our attack results on diffusion models using the DP defense.

We adopt Differentially-Private Stochastic Gradient Descent (DP-SGD) [1] to train diffusion models. DP-SGD is widely used for privately training a machine learning model. Generally, DP-SGD achieves differential privacy by adding noise into per-sample gradients. In our work, we implement DP diffusion models through the Opacus library [41] which allows us to set privacy budgets through hyperparameters. Here, we set the clip bound C and the failure probability δ as 1 and 5×10^{-4}. The batch size and the number of epochs are 64 and $1,800$. Thus, the final privacy budget ϵ is 19.62. Generally, a smaller privacy budget means a higher privacy setting and more severe model utility loss. The common choice of privacy budget is $\epsilon \leq 10$ [1, 41], and in this work we choose a higher privacy budget because we consider the utility of a diffusion model. We choose the DDPM model as the target model. It is trained on FFHQ containing $1,000$ training samples, and the FID is 393.94.

Performance of loss-based attack. Figure 8 shows the performance of both types of attacks on DDPM trained with DP-SGD on FFHQ. In Figure 8(a), we present the performance of the loss-based attack on all diffusion steps. Clearly, we can see that although differentially training DDPM, i.e. DDPM with DP-SGD,



(a) Loss-based attack (b) Loss-based attack (c) likelihood-based attack

Fig. 8: Attack performance on DDPM with DP-SGD.

indeed can significantly decrease the membership leakages, the peak regions can be still identified between 400 and 800 diffusion step. Figure 8(b) further shows ROC curves of our loss-based attack on different diffusion steps. Again, we can observe that in the low FPR regimes, some training samples are still inferred with a higher probability, such as 10^{-2} TPR at 10^{-4} FPR at t=500. This is higher than 100 times than random guesses (TPR is 10^{-4} at 10^{-4} FPR).

Performance of likelihood-based attack. Figure 8(c) shows the performance of the likelihood-based attack on DDPM training with DP-SGD on FFHQ. Again, we can see that differentially private training of a diffusion model indeed can mitigate our attack. At the same time, we also see at the low false positive rate regime, our attack still remains at 0.1% true positive rate, which illustrates the effectiveness of our attack even in the worst-case. Here, we also note that the FID of the target model is 393.94, which means that the utility of the target model suffers from a severe performance drop. We leave developing more usable techniques to train a diffusion model with DP-SGD as future work. Table 5 in Appendix summarizes the quantitative results of both attacks.

8 Related Work

Diffusion models. Diffusion models have attracted increasing attention in the past years. Sohl-Dickstein et al. [34] first introduce nonequilibrium thermodynamics to build generative models. The key idea is to slowly add noise into data in the forward process and learn to generate data from noise through a reverse process. Ho et al. [16] further propose to use parameterization techniques in diffusion models, which enable diffusion models to generate high-quality images. Song et al. [37] present to train a generative model by estimating gradients of data distribution, i.e. score. Furthermore, Song et al. [38] propose a unified framework to describe these diffusion models through the lens of stochastic differential equations. However, in this work, we study diffusion models from the perspective of privacy.

Membership inference attacks. There are extensive works on membership inference (MI) attacks on classification models. Various attack methods under different threat models are proposed, such as using fewer shadow models [32], using loss values [2,25,33,40] and using labels of victim models [7,23]. In addition, there are several MI attacks on generative models [6, 13, 15]. Nevertheless, these attacks are more specific to GANs and heavily rely on the unique characteristics

of GANs, such as discriminators or generators. They cannot be extended to diffusion models, because diffusion models have different training and sampling mechanisms. Therefore, our work on MI of diffusion models aims to fill this gap.

Membership inference attacks in diffusion models. In this paragraph, we discuss our work and its relation to several similar/concurrent works studying MI attacks in diffusion models. Wu et al. [39] study MI attacks against text-to-image generative models. One diffusion-based text-to-image generative model, LDM [31], is attacked by their methods based on query data pair, i.e. text and corresponding output image. Unlike text-to-image generative models, we focus on unconditional diffusion models. Furthermore, our MI attack methods, such as the loss-based attack, are totally different from their methods [39]. Subsequently, there are several concurrent works that investigate MI attacks against diffusion models based on the loss information [3, 9, 26, 42]. However, they only consider discrete diffusion models where the number of noise distributions is finite. Our work systematically studies both discrete and continuous diffusion models. Although Carlini et al. [3] design more sophisticated and effective methods, they require extraordinarily huge computation resources, such as training hundreds of shadow diffusion models or millions of queries from diffusion models. In contrast, our method only utilizes loss values, which is much more computationally efficient. In addition, we also propose the likelihood-based method which is not considered in these works [3, 9, 26, 42].

9 Conclusion

In this paper, we have developed two types of membership inference attack methods: loss-based attack and likelihood-based attack. Our methods have demonstrated the connection between membership inference risks and the generative mechanism of diffusion models. To be more specific, our loss-based attack reveals that in terms of diffusion steps, there exist high-risk regions where training samples can be inferred with high precision. Although membership inference becomes more challenging with the increase in the number of training samples, the high-risk regions still exist. Our experimental results on classic privacy protection mechanisms, i.e. diffusion models trained with DP-SGD, further show that DP-SGD alleviates our attacks at the expense of severe model utility.

Designing an effective differential privacy strategy to produce high-quality images for diffusion models is promising and challenging, which is part of our future work. In addition, it is an interesting direction to study MI attacks of diffusion models in stricter scenarios, such as only obtaining synthetic data.

Acknowledgments

This research was funded in whole by the Luxembourg National Research Fund (FNR), grant reference 13550291.

Appendix

In this section, we show additional results and introduce each result in its caption.

(a) DDPM (b) SMLD (c) VPSDE (d) VESDE

Fig. 9: Generated images from different target models trained on FFHQ. It is corresponding to Section 5.1.

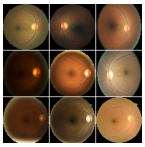


Fig. 10: Generated images from the target model SMLD trained on the DRD dataset. It is corresponding to Section 6.2.

Table 4: Quantitative results of our attacks on SMLD trained on DRD. It is corresponding to Section 6.2.

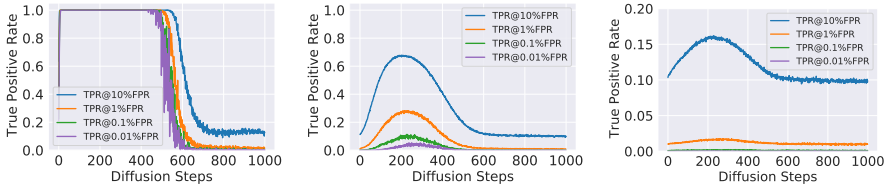
Attack	T	TPR@ 10%FPR	TPR@ 1%FPR	TPR@ 0.1%FPR	TPR@ 0.01%FPR	Accuracy
Loss-based	0	7.50%	1.10%	0.00%	0.00%	50.25%
	200	11.20%	0.70%	0.10%	0.00%	52.25%
	700	80.60%	50.50%	33.34%	18.80%	85.45%
	800	93.30%	72.20%	60.00%	40.10%	92.25%
	900	79.80%	42.40%	17.70%	12.30%	86.35%
	999	43.60%	9.70%	2.00%	0.10%	70.95%
Likelihood-based	-	100.00%	100.00%	100.00%	99.90%	99.95%

References

1. Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K., Zhang, L.: Deep learning with differential privacy. In: ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 308–318. ACM (2016)

Table 5: Quantitative results of our attacks on DDPM trained with DP-SGD. It is corresponding to Section 7.

Attacks	T	TPR@ 10%FPR	TPR@ 1%FPR	TPR@ 0.1%FPR	TPR@ 0.01%FPR	Accuracy
Loss-based	0	8.80%	1.40%	0.00%	0.00%	52.25%
	200	8.60%	1.40%	0.40%	0.30%	53.20%
	500	**10.70%**	**1.60%**	**0.90%**	**0.90%**	**51.85%**
	600	**13.00%**	**2.30%**	**1.00%**	**1.00%**	**51.85%**
	800	11.60%	2.10%	0.30%	0.30%	51.75%
	999	10.40%	0.60%	0.00%	0.00%	53.90%
Likelihood-based	-	8.40%	1.10%	0.20%	0.10%	51.75%



(a) FFHQ-1k. (b) FFHQ-10k. (c) FFHQ-30k.

Fig. 11: Performance of loss-based attacks with different sizes of datasets. The target model is DDPM trained on FFHQ. Each subfigure shows attack performance with different FPRs on fixed dataset sizes. It is corresponding to Section 6.1.

2. Carlini, N., Chien, S., Nasr, M., Song, S., Terzis, A., Tramer, F.: Membership inference attacks from first principles. In: IEEE Symposium on Security and Privacy (S&P). pp. 1519–1519. IEEE (2022)
3. Carlini, N., Hayes, J., Nasr, M., Jagielski, M., Sehwag, V., Tramèr, F., Balle, B., Ippolito, D., Wallace, E.: Extracting training data from diffusion models. arXiv preprint arXiv:2301.13188 (2023)
4. Carlini, N., Liu, C., Erlingsson, Ú., Kos, J., Song, D.: The secret sharer: Evaluating and testing unintended memorization in neural networks. In: USENIX Security Symposium (USENIX Security). pp. 267–284. USENIX Association (2019)
5. Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al.: Extracting training data from large language models. In: USENIX Security Symposium (USENIX Security). pp. 2633–2650. USENIX Association (2021)
6. Chen, D., Yu, N., Zhang, Y., Fritz, M.: Gan-leaks: A taxonomy of membership inference attacks against generative models. In: ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 343–362. ACM (2020)
7. Choquette-Choo, C.A., Tramer, F., Carlini, N., Papernot, N.: Label-only membership inference attacks. In: International Conference on Machine Learning (ICML). pp. 1964–1974. PMLR (2021)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 34, pp. 8780–8794. Curran Associates, Inc. (2021)
9. Duan, J., Kong, F., Wang, S., Shi, X., Xu, K.: Are diffusion models vulnerable to membership inference attacks? arXiv preprint arXiv:2302.01316 (2023)

10. Dwork, C.: Differential privacy: A survey of results. In: International conference on theory and applications of models of computation. pp. 1–19. Springer (2008)
11. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 2672–2680. Curran Associates, Inc. (2014)
12. Grathwohl, W., Chen, R.T., Bettencourt, J., Sutskever, I., Duvenaud, D.: Ffjord: Free-form continuous dynamics for scalable reversible generative models. In: International Conference on Learning Representations (ICLR) (2018)
13. Hayes, J., Melis, L., Danezis, G., De Cristofaro, E.: LOGAN: Membership inference attacks against generative models. In: Proceedings on Privacy Enhancing Technologies. pp. 133–152. Sciendo (2019)
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 6626–6637. Curran Associates, Inc. (2017)
15. Hilprecht, B., Härterich, M., Bernau, D.: Monte carlo and reconstruction membership inference attacks against generative models. In: Proceedings on Privacy Enhancing Technologies. pp. 232–249. Sciendo (2019)
16. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 33, pp. 6840–6851. Curran Associates, Inc. (2020)
17. Hu, H., Pang, J.: Membership inference attacks against GANs by leveraging over-representation regions. In: ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 2387—-2389. ACM (2021)
18. Kaggle.com: Diabetic retinopathy detection. `https://www.kaggle.com/c/diabetic-retinopathy-detection#references.` (2015)
19. Karras, T., Aittala, M., Aila, T., Laine, S.: Elucidating the design space of diffusion-based generative models. In: Advances in Neural Information Processing Systems (NeurIPS). Curran Associates, Inc. (2022)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4401–4410. IEEE (2019)
21. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models for medical image analysis: A comprehensive survey. arXiv preprint arXiv:2211.07804 (2022)
22. Leino, K., Fredrikson, M.: Stolen memories: Leveraging model memorization for calibrated white-box membership inference. In: Proceedings of USENIX Security Symposium (USENIX Security). pp. 1605–1622. USENIX Association (2020)
23. Li, Z., Zhang, Y.: Membership leakage in label-only exposures. In: ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 880–895. ACM (2021)
24. Lin, Z., Jain, A., Wang, C., Fanti, G., Sekar, V.: Using gans for sharing networked time series data: Challenges, initial promise, and open questions. In: Proceedings of the ACM Internet Measurement Conference (IMC). pp. 464–483. ACM (2020)
25. Liu, Y., Zhao, Z., Backes, M., Zhang, Y.: Membership inference attacks by exploiting loss trajectory. In: ACM SIGSAC Conference on Computer and Communications Security (CCS). pp. 2085–2098 (2022)
26. Matsumoto, T., Miura, T., Yanai, N.: Membership inference attacks against diffusion models. arXiv preprint arXiv:2302.03262 (2023)

27. Murakonda, S.K., Shokri, R.: Ml privacy meter: Aiding regulatory compliance by quantifying the privacy risks of machine learning. arXiv preprint arXiv:2007.09339 (2020)
28. Park, N., Mohammadi, M., Gorde, K., Jajodia, S., Park, H., Kim, Y.: Data synthesis based on generative adversarial networks. Proceedings of the VLDB Endowment **11**(10), 1071–1083 (2018)
29. Parliament, E., of the European Union, C.: Art. 35 gdpr: Data protection impact assessment. `https://gdpr-info.eu/art-35-gdpr/` (2016)
30. Pinaya, W.H., Tudosiu, P.D., Dafflon, J., Da Costa, P.F., Fernandez, V., Nachev, P., Ourselin, S., Cardoso, M.J.: Brain imaging generation with latent diffusion models. In: MICCAI Workshop on Deep Generative Models. pp. 117–126. Springer (2022)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695. IEEE (2022)
32. Salem, A., Zhang, Y., Humbert, M., Berrang, P., Fritz, M., Backes, M.: Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models. In: Network and Distributed Systems Security Symposium (NDSS). Internet Society (2019)
33. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership inference attacks against machine learning models. In: IEEE Symposium on Security and Privacy (S&P). pp. 3–18. IEEE (2017)
34. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning (ICML). pp. 2256–2265. PMLR (2015)
35. Song, S., Marn, D.: Introducing a new privacy testing library in tensorflow. `https://blog.tensorflow.org/2020/06/introducing-new-privacy-testing-library.html` (2020)
36. Song, Y.: Score-based generative modeling through stochastic differential equations. `https://github.com/yang-song/score_sde_pytorch` (2021)
37. Song, Y., Ermon, S.: Generative modeling by estimating gradients of the data distribution. In: Advances in Neural Information Processing Systems (NeurIPS). vol. 32. Curran Associates, Inc. (2019)
38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: International Conference on Learning Representations (ICLR) (2021)
39. Wu, Y., Yu, N., Li, Z., Backes, M., Zhang, Y.: Membership inference attacks against text-to-image generation models. arXiv preprint arXiv:2210.00968 (2022)
40. Ye, J., Maddi, A., Murakonda, S.K., Shokri, R.: Enhanced membership inference attacks against machine learning models. In: ACM SIGSAC Conference on Computer and Communications Security (CCS). p. 3093–3106 (2022)
41. Yousefpour, A., Shilov, I., Sablayrolles, A., Testuggine, D., Prasad, K., Malek, M., Nguyen, J., Gosh, S., Bharadwaj, A., Zhao, J., Cormode, G., Mironov, I.: Opacus: User-friendly differential privacy library in pytorch. arXiv preprint arXiv:2109.12298 (2021)
42. Zhu, D., Chen, D., Grossklags, J., Fritz, M.: Data forensics in diffusion models: A systematic analysis of membership privacy. arXiv preprint arXiv:2302.07801 (2023)