Generated Distributions Are All You Need for Membership Inference Attacks Against Generative Models

Anonymous authors
Paper under double-blind review

Abstract

Generative models have shown their promising performance on various real-world tasks, which, at the same time, introduce the threat of leaking private information of their training data. Several membership inference attacks against generative models have been proposed in recent years and exhibit their effectiveness in different settings. However, these attacks all suffer from their own limitations and cannot be generalized to all generative models under all scenarios. In this paper, we propose the first generalized membership inference attack for generative models, which can be utilized to quantitatively evaluate the privacy leakage of various existing generative models. Compared with previous works, our attack has three main advantages, i.e., (i) only requires black-box access to the target model, (ii) is computationally efficient, and (iii) can be generalized to numerous generative models. Extensive experiments show that various existing generative models in a variety of applications are vulnerable to our attack. For example, our attack could achieve the AUC of 0.997 (0.997) and 0.998 (0.999) against the generative model of DDPM (DDIM) on the CelebA and CIFAR-10 datasets. These results demonstrate that private information can be effectively exploited by attackers in an efficient way, which calls on the community to be aware of privacy threats when designing generative models.

1 Introduction

Recent arms race in the visual generation has reached a new peak. Dall-E-2 (Ramesh et al., 2022), Imagen (Saharia et al., 2022), and Parti (Yu et al., 2022), driven by big data and empowered by deep generative models, have emerged one after the other in a short period of time to showcase their improved power. The realism and creativity of generated media have set up a new state of the art that were never envisioned one year ago. In the backend, Generative Adversarial Networks (GANs), Variational AutoEncoders (VAEs), and diffusion probabilistic models (for short, we use "diffusion models" to refer to this family of models) are three representative milestones. While enjoying the benefits of these technologies, the threat of privacy leaks Chen et al. (2020); Azadmanesh et al. (2021); Hu & Pang (2021); Zhou et al. (2022) comes with them. To this end, we aim to reveal the hidden privacy issues posed by generative models through the lens of membership inference attacks.

Membership Inference Attacks (MIAs), which aim to infer whether a query sample is in the training set of the target model or not, could be leveraged by malicious adversaries to cause severe consequences. For example, an adversary might infer that an individual is in the training set of a disease diagnosis classifier trained on data from high-income people. The membership status, in this case, directly reveals the income of the target individual, which is a typical privacy breach. On the other hand, MIAs can also be used for benign purposes such as auditing, i.e. quantitatively assessing the privacy breach of machine learning models. Moreover, MIAs are commonly used as the building block for more sophisticated attacks, which makes it an important problem and has attracted lots of attention in recent years.

Most existing MIAs aim at discriminative models Shokri et al. (2017); Salem et al. (2019); Choo et al. (2021); Zhang et al. (2021); Li & Zhang (2021), which cannot directly extend to generative models due to the heterogeneous architecture of target models. A few existing MIAs explore privacy risks in generative models Hayes et al. (2019); Chen et al. (2020); Azadmanesh et al. (2021), however, they are limited to certain attack assumptions, e.g. requiring white-box access. These limitations hinder their application in state-of-the-art generative models like diffusion models, which motivates researchers to propose more practical and generalized MIAs.

In this paper, we propose the first generalized membership inference attack against generative models with relaxed assumptions. Compared with existing attacks, our method has three main advantages: (i) Relaxed assumptions. Our work assumes that the adversary only has black-box access to the target model. That is, the attacker could only query the target model in an API manner, without the knowledge of the training data and the architecture of the target model. (ii) Computationally efficient. Contrary to existing attacks, our method does not require training shadow models, which makes the method computationally efficient. (iii) Good generalizability. Our attack can be applied to extensive generative models including the state-of-the-art diffusion models. Whereas previous works are limited by their generalizability.

We extensively evaluate our work on the state-of-the-art generative models. Concretely, we conduct experiments on four benchmark datasets (i.e., CIFAR-10 CIF, CelebA Liu et al. (2015) ,LSUN-Bedroom and LSUN-Church Yu et al. (2015) with three commonly-used diffusion models structures (i.e., DDPM Ho et al. (2020), DDIM Song et al. (2021) and FastDPM Kong & Ping (2021)). Experimental results show the effectiveness of our work. For instance, when the target models are established on the CelebA and CIFAR-10 datasets with the backbone of DDPM (DDIM), our attack achieves the AUC of 0.997 (0.997) and 0.998 (0.999). We further validate the effectiveness of our work on other generative tasks, including the class-conditional generation (Esser et al., 2021), semantic-conditional generation (Liu et al., 2019) and text-conditional generation (Rombach et al., 2022), the stylization (Park et al., 2020), the superresolution (Chen et al., 2021), the image inpainting (Li et al., 2022), the denoising (Zamir et al., 2021), the colorization (Zhang et al., 2016) and the artifact reduction (Zamir et al., 2021). Our attack is proved to be effective on all of these tasks.

In summary, our contributions are as follows:

- We propose the first generalized membership inference attack against generative models. Compared with previous works, our attack requires fewer assumptions and is computationally efficient. Moreover, our attack can be generalized to extensive generative models, including state-of-the-art diffusion models.
- We empirically validate the effectiveness of our attack on four benchmark datasets with three diffusion model architectures. Experimental results show that our attack works well on diffusion models (AUC close to 1).
- We further explore the privacy threats on other generative tasks, and the results demonstrate that almost all tasks are subject to privacy threats. Our work fills the blank of understanding privacy risk for generative models, which motivates researchers to take privacy threats into concern when designing generative models.

2 Related Works

Generative models: Generative models have achieved prosperity development in recent years, and several generative model architectures have been proposed. Based on their design philosophy, current generative models can be categorized into three main classes, i.e., generative adversarial networks (GANs) Arjovsky et al. (2017); Wang et al. (2018); Zhang et al. (2018); Brock et al. (2019); Patashnik et al. (2021), variational autoencoders (VAEs) Kingma & Welling (2014); Higgins et al. (2017); Sohn et al. (2015); van den Oord et al. (2017); Tolstikhin et al. (2018); Chen et al. (2018), and diffusion probabilistic models (diffusion models for short) Ho et al. (2020); Song et al. (2021); Kong & Ping (2021); Rombach et al. (2022); Ramesh et al. (2022); Saharia et al. (2022). A GAN-based generative model basically consists of a generator and a discriminator, where the generator manages to create images which will not be regarded as fake ones by the discriminator. A VAE-based generative model is to estimate a normal distribution for each input image, then to sample a latent representation form this distribution, finally to restore the input image from this sampled representation. Regarding diffusion generative models, an input image is added noises step by step, then a U-Net model predicts the added noised for each step, finally, the noised image is denoised using the predicted noises step by step until being a clean image. In this paper, we investigate the privacy risk introduced by diffusion models, as diffusion models have attained state-of-the-art performance in quantities of generative tasks.

Membership inference attacks: Membership inference attacks aim to infer whether a query sample is involved in the model training process. Existing attacks mainly focus on the discriminative models Shokri et al. (2017); Salem et al. (2019); Choo et al. (2021); Zhang et al. (2021); Li & Zhang (2021), which are difficult to apply to the generative models. In this work, we mainly focus on membership inference attacks against generative models. To our best knowledge, current membership inference attacks against generative models mainly focus on GAN-based generative models, which limits the applicable scope of membership inference. Hayes et al. (2019) present the membership inference attack on several GANs and a single VAE, which starts with the white-box access to the target model and then involves auxiliary knowledge of dataset samples for the black-box attack. Hilprecht et al. (2019) propose two kinds of membership inference against GANs and VAEs. One works on both GANs and VAEs, and only aims at samples of the model which are close to training data records. And the other is designed specifically for VAEs. Chen et al. (2020) study the first taxonomy of membership inference applicable to various settings, which mainly focuses on GANs. And a calibration technique is discussed to boost the attack performance. Azadmanesh et al. (2021) conduct a white-box membership inference against GANs, which requires the access to generator structure of the victim GAN, then an autoencoder is built that shares the generator structure, finally the membership status is inferred according to the cost values of the autoencoder. Unfortunately, these works cannot generalize well to the state-of-the-art generative technique, i.e., diffusion models, due to the diversity property of diffusion models. For instance, regarding the black-box attack in Chen et al. (2020), it only achieves a chance-like performance on diffusion models since a single generated image is hard to represent the training data (members) of the target model. To this end, we make use of generative distributions to learn the attack model, i.e., involving a set of generated images instead of a single one. Our approach is architecture-agnostic, which is applicable to various generative models. Moreover, our attack does not build a shadow model, which is computationally efficient.

3 Our Attack

3.1 Problem Statement

In this paper, we study the problem of membership inference attack against generative models, whose goal is to infer whether a query image x_{query} belongs to the training set (members) of the target generative model $\mathcal{M}_{\text{target}}$. The attack $\mathcal{A}$ can be formulated as follow:

$$\mathcal{A} : x_{\text{query}}, \mathcal{M}_{\text{target}}, \mathcal{K} \rightarrow \{\text{member}, \text{non-member}\} \tag{1}$$

Here, $\mathcal{K}$ denotes extra information known to the attacker. Previous works always rely on strong assumptions that the attacker has information about the model, i.e., white-box access to the model. However, in our work, we only require the attacker has black-box access to the model, which means the attacker can only interact with the model through an API manner. Following convention, we also assume the attacker has access to an auxiliary dataset $\mathcal{D}_{\text{aux}}$, which comes from a similar distribution as the training dataset $\mathcal{D}_{\text{train}}$.

Our attack is architecture-agnostic and can be generalized to extensive generative models. For ease of illustration, in the following, we use the diffusion model as an example, as it outperforms other structures in numerous tasks, and no existing attacks are applicable to it. Specifically, we consider two main classes of tasks as attack scenarios, i.e., the unconditional generation task and the conditional generation task. The main difference is the input to the generative models. For the unconditional generation task, the generative model takes noises as input, and output generated images. While for conditional tasks, the generative model takes extra information to guide the generation process, e.g., data sampled from an auxiliary dataset. In our attack, we regard the samples from the auxiliary dataset as non-members and label the samples generated by generative models as members.

3.2 Methodology

Figure 1 depicts the general framework of our attack. The attack process consists of three steps: image generation, attack model establishment, and membership inference, where image generation is the core step of our attack.

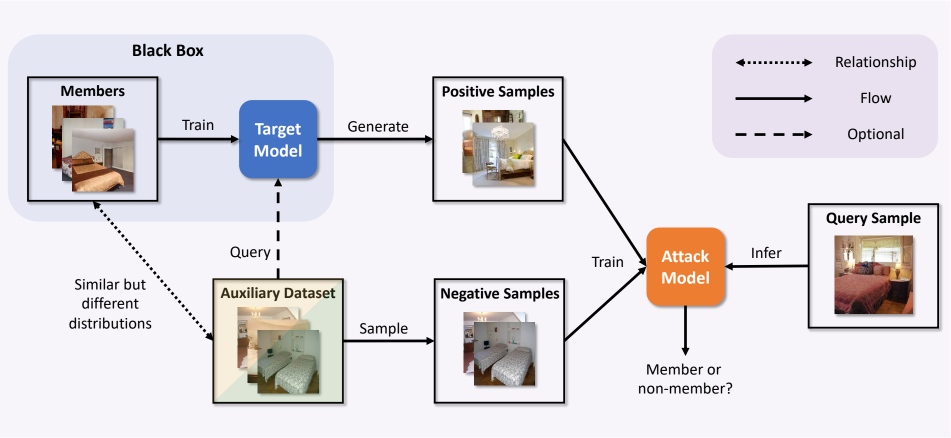


Figure 1: The framework of membership inference attacks against a diffusion generative model. The round-dot double-arrow line indicates the relationship between two entities. The solid arrow lines indicate the workflow of our attack. And the dash arrow line indicates an optional operation that is dependent on the application implemented by the target model. Besides, the yellow and green triangles in the auxiliary dataset indicate that the images regarded as negative samples and the ones used to query the target model are disjoint. In addition, the blue rounded rectangle means black-box access to the target model.

Image generation: In the image generation step, our goal is to generate training data for the attack model. The negative samples are easy to collect, as we mentioned in Section 3.1, we use samples from the auxiliary dataset as the negative samples. However, the challenge is how to obtain quantities of positive samples to train the model, as the member samples are hard to obtain. To tackle this problem, we leverage the fact that generated images can reflect the pattern of real member images to a large extent. Consequently, we use the images generated by generative models as positive samples.

For unconditional generation tasks, we query the generative model with Gaussian noise, i.e., $z \sim \mathcal{N}(0, 1)$, then label the generated data as a member. By querying the target model, we obtain the positive training pair, $(\mathcal{M}_{\text{target}}(z), \text{member})$, which will be used for attack model training.

For conditional generation tasks, the target model receives information to guide the generation process. We illustrate the case where the input is images as a representative. We first separate the auxiliary dataset $\mathcal{D}_{\text{aux}}$ into two disjoint parts, $\mathcal{D}_{\text{aux}}^{\text{in}}$ and $\mathcal{D}_{\text{aux}}^{\text{out}}$. For data sampled from $\mathcal{D}_{\text{aux}}^{\text{out}}$, we label it as non-member, formally:

$$\forall x \in \mathcal{D}_{\text{aux}}^{\text{out}}, \quad \text{we have} \quad (x, \text{non-member}) \tag{2}$$

then we query the target model using the data sampled from $\mathcal{D}_{\text{aux}}^{\text{in}}$, and label the generated images as member:

$$\forall x \in \mathcal{D}_{\text{aux}}^{\text{in}}, \quad \text{we have} \quad (\mathcal{M}_{\text{target}}(x), \text{member}) \tag{3}$$

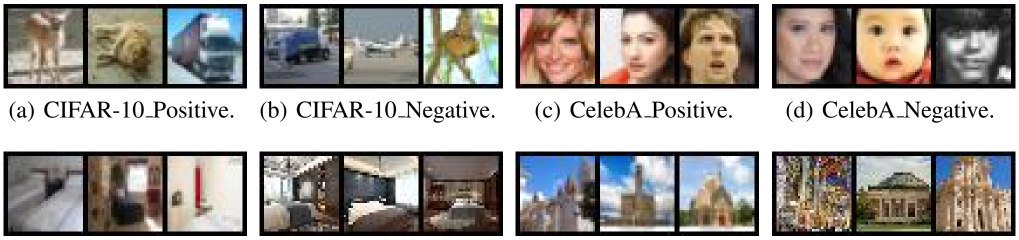
Note that our attack is architecture-agnostic and task-agnostic, thus the input to the generative model could be other types of information, like text sentences, images with artifacts and grayscale images.

Attack model establishment: The Image generation step generates positive and negative samples, which can be used to construct the attack set $\mathcal{D}_{\text{attack}}$ for the attack model $\mathcal{M}_{\text{attack}}$. The goal of the attack model is to predict the membership status given a query sample, we train the attack model in a supervised way leveraging the labeled pairs in the training dataset $\mathcal{D}_{\text{attack}}$.

Membership inference: Once the attack model $\mathcal{M}_{\text{attack}}$ is trained, the adversary can infer the membership status of a query sample x_{query}. Note that, in the inference procedure, the adversary does not need to interact with the target model. Specifically, the adversary directly queries the trained attack model $\mathcal{M}_{\text{attack}}$ with the query sample x_{query}. Then, the membership status is inferred according to the output of the attack model. This process can be formulated as:

$$\mathcal{M}_{\text{attack}}(x_{\text{query}}) = y_{\text{target}} \tag{4}$$

where y_{target} is a 2-dimensional vector indicating the membership status of the query sample x_{query}.



(a) CIFAR-10_Positive. (b) CIFAR-10_Negative. (c) CelebA_Positive. (d) CelebA_Negative.

(e) Bedroom_Positive. (f) Bedroom_Negative. (g) Church_Positive. (h) Church_Negative.

Figure 2: The image examples from different sources, where the images of "Positive" are derived from the datasets of members while the images of "Negative" are derived from the auxiliary datasets.

Table 1: The dataset settings of various generative models. In the column of "Positive (Generated)", the positive training samples are the generated images. And in the "Positive (Members)" and "Negative (Auxiliary)" columns, the samples are directly sampled from the member or auxiliary datasets.

Application	**Generative Model**	**Training Dataset**		**Inference Dataset**	
		Positive (Generated)	Negative (Auxiliary)	Positive (Members)	Negative (Auxiliary)
Class-conditional	VQGAN	CIFAR-10	CIFAR-10	ImageNet	CIFAR-10
Text-conditional	LDM	COCO	COCO	LAION	COCO
Semantic-conditional	CC-FPSE	ADE20K	ADE20K	COCO	ADE20K
Colorization	Colorization	CIFAR-10	CIFAR-10	ImageNet	CIFAR-10
Super resolution	LIIF	UTKFace	UTKFace	CelebA-HQ	UTKFace
Image inpainting	MAT	UTKFace	UTKFace	CelebA-HQ	UTKFace
Stylization	SwappingAutoencoder	Wild church	Wild Church	LSUN-Church	Wild church
Denoising	MPRNet	ImageNet	ImageNet	SIDD	ImageNet
Artifact reduction	MPRNet	ImageNet	ImageNet	SIDD	ImageNet

4 EVALUATION

In this section, we empirically evaluate the effectiveness of our attack. We first present the experimental settings in Section 4.1. Then explore the state-of-the-art generative technique, diffusion models, in Section 4.2. At the last, we extensively evaluate nine more generative models with different tasks in Section 4.5.

4.1 EVALUATION SETTINGS

Dataset description: To evaluate our work on diffusion models, we leverage four benchmark datasets to train the target model, i.e., CIFAR-10 (CIF), CelebA (Liu et al., 2015), and LSUN (Yu et al., 2015) including *Bedroom* and *Church*. As mentioned in Section 3, negative samples are collected from an auxiliary dataset. Thus for different training datasets, we use different sampling strategies to construct the auxiliary dataset. Specifically, for the CIFAR-10 dataset, we use STL-10 (Coates et al., 2011) as the auxiliary dataset. For CelebA, we choose UTKFace (Zhang et al., 2017) as the auxiliary dataset. Note that, compared to CIFAR-10, the dataset of STL-10 contains the category of monkey instead of frog. This difference displays a practical scenario in the real world, in which it might be too expensive to find another dataset with the exact same categories. And regarding the negative samples for the members from Bedroom and Church, we collect wild images from the internet with manually filtering since no existing similar dataset is available. In Figure 2, we present the examples from different sources, where the images of "Positive" are derived from the datasets of members while the images of "Negative" are derived from the auxiliary datasets.

To evaluate the nine generative models, we involve one benchmark dataset (members) for each application, i.e., ImageNet Deng et al. (2009) for class-conditional generation and colorization, LAION Schuhmann et al. (2021) for text-conditional generation, COCO Lin et al. (2014) for semantice-conditional generation, CelebA-HQ Liu et al. (2015) for super resolution and image inpainting, LSUN-Church Yu et al. (2015) for stylization, and SIDD Abdelhamed et al. (2018) for denoising and artifact reduction, which are shown in the column of "Positive (Members)" in Table 1. Similar to the evaluation of diffusion models, for each dataset of members, the adversary finds an auxiliary dataset from a similar but different distribution. Concretely, from Table 1 we can see that

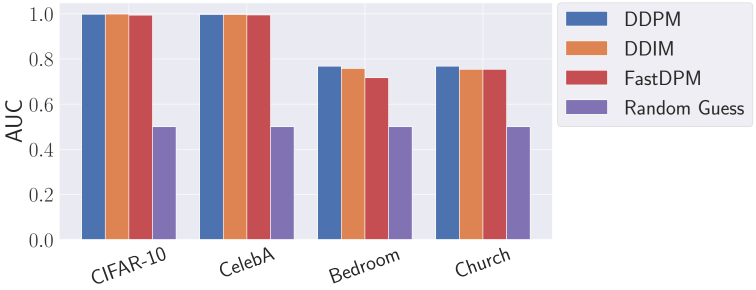


Figure 3: The attack performances of DDPM, DDIM and FastDPM on the datasets of CIFAR-10, CelebA, LSUN (including Bedroom and Church).

the pairs of auxiliary datasets and member datasets are as follows: CIFAR-10 for ImageNet, COCO for LAION, ADE20K Zhou et al. (2017) for COCO, UTKFace Zhang et al. (2017) for CelebA-HQ, ImageNet for SIDD, and wildly collected church images for LSUN-Church. In Table 1, the dataset Q in the column of "Positive (Generated)" means that the positive training samples are the generated images from the target model by querying with Q. And regarding the dataset Q in the other "Positive (Members)" and "Negative (Auxiliary)" columns, the samples are the images directly sampled from the member or auxiliary datasets Q, as demonstrated in Section 3.2. Note that, in Table 1, even though the columns of "Training Dataset" and "Inference Dataset" are associated to the attack model, the column of "Positive (Members)" also corresponds to members which are used to train the target model.

Model architectures: For the unconditional generation task 4.2, we focuses on three widely used diffusion models, i.e., DDPM (Ho et al., 2020), DDIM (Song et al., 2021) and FastDPM (Kong & Ping, 2021). Consistently, these models are based on the structure of U-Net (Ronneberger et al., 2015) to predict noises used for the denoising process, where the noises share the same size with images.

In Section 4.5, we adopt nine generative models corresponding to different generation tasks, including VQGAN Esser et al. (2021) for class-conditional generation, LDM Rombach et al. (2022) for text-conditional generation, CC-FPSE Liu et al. (2019) for semantic-conditional generation, Colorization Zhang et al. (2016) for colorization, LIIF Chen et al. (2021) for super resolution, MAT Li et al. (2022) for image inpainting, SwappingAutoencoder Park et al. (2020) for stylization, and MPRNet Zamir et al. (2021) for denoising and artifact reduction, which are shown in Table 1.

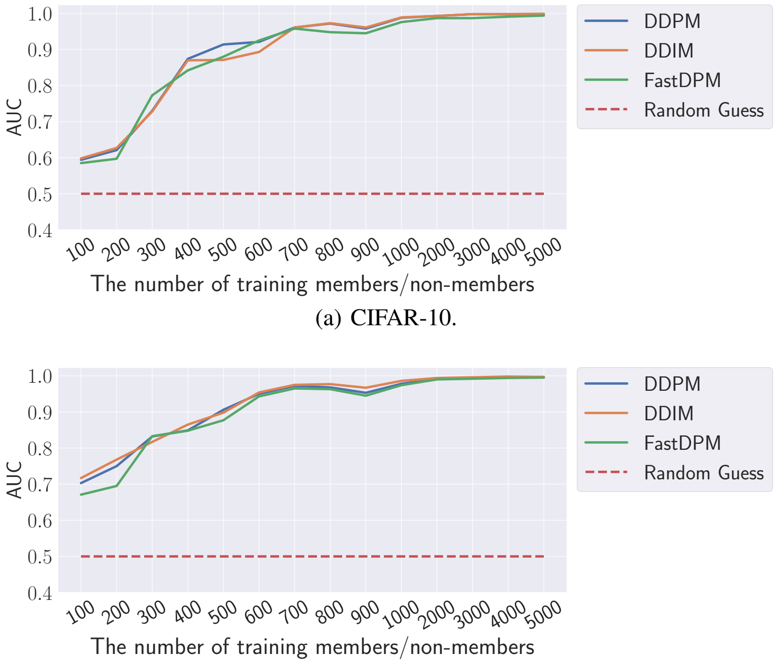
To infer the membership status from the target model, an attack model is established with the backbone of ResNet18 (He et al., 2016). Specifically, we fit the last linear layers of the pretrained ResNet18 for binary classification (i.e., inferring the query sample belonging to members or nonmembers) and finetune it using the generated images as positive data and negative images sampled from the auxiliary dataset (as mentioned in Section 3).

Evaluation metric: Following previous works Salem et al. (2019), Chen et al. (2020), Li & Zhang (2021) and Zhang et al. (2021), we utilize AUC as the metric to evaluate our work. Specifically, during the inference, images used to train the model are regarded as positive samples while images from the auxiliary dataset are regarded as negative samples. In a word, AUC represents the degree of separability, which tells how much the model is capable of distinguishing between classes.

4.2 Effectiveness on Unconditional Generation Tasks

In this part, we evaluate the effectiveness of our attack on unconditional generation tasks, specifically, we conduct our attack on three state-of-the-art diffusion models, i.e., DDPM, DDIM and FastDPM, on four benchmark datasets.

As we can see from Figure 3, our attack achieves a significant improvement compared with random guess on all four datasets. Especially, on the CIFAR-10 and CelebA datasets, our attack even has almost perfect performance with AUC close to 1. These results well demonstrate the effectiveness of our attack.



(a) CIFAR-10.

(b) CelebA.

Figure 4: The attack performances of DDPM, DDIM and FastDPM on CIDAR-10 and CelebA using different number of generated images, from 100 to 1,000 with a step of 100 and from 1,000 to 5,000 with a step of 1,000.

Table 2: The transferability results against DDPM, DDIM and FastDPM. Each column corresponds to one attack setting. In "All", we use all four datasets to train the attack model. In "Leave", we use three non-matching datasets except the matching one to train the attack model. In "CIFAR-10"/"CelebA"/"Bedroom"/"Church", we train the attack model using the named dataset. To highlight the transferability results, we cover the duplicate results of Figure 3 by "-".

	All	**Leave**	**CIFAR-10**	**CelebA**	**Bedroom**	**Church**
			DDPM			
CIFAR-10	0.990	0.629	-	0.619	0.604	0.600
CelebA	0.974	0.665	0.573	-	0.618	0.622
Bedroom	0.608	0.594	0.592	0.592	-	0.541
Church	0.618	0.602	0.573	0.601	0.599	-
			DDIM			
CIFAR-10	0.988	0.610	-	0.602	0.572	0.568
CelebA	0.984	0.711	0.557	-	0.619	0.613
Bedroom	0.647	0.623	0.591	0.586	-	0.571
Church	0.635	0.623	0.571	0.593	0.586	-
			FastDPM			
CIFAR-10	0.987	0.613	-	0.606	0.610	0.580
CelebA	0.966	0.693	0.563	-	0.618	0.619
Bedroom	0.561	0.598	0.594	0.575	-	0.544
Church	0.631	0.688	0.635	0.623	0.598	-

Meanwhile, we also find that the performance gap between different datasets is noteworthy, which reveals that the quality of the auxiliary dataset influences the attack performance. Concretely, for the DDPM model, our attack achieves 0.997 AUC on the CIFAR-10 dataset but only 0.768 AUC on the LSUN-Bedroom dataset, while the attack performance is similar between LSUN-Bedroom and

LSUN-Church. We posit this is because the auxiliary dataset for LSUN is manually collected from the internet. However, collecting high-quality auxiliary data from the internet is time- and money-consuming, which leaves an interesting problem of effectively collecting auxiliary datasets to be explored in the future. In addition, we also evaluate existing black-box attack Chen et al. (2020) to attack these diffusion models, but their attack only gain negligible increments compared with random guess (thus we do not depict them in the paper), which further shows the advantage of our attack.

4.3 Influence of the Query Budget

Diffusion models normally require expensive computational resources to generate images, and sometimes even require users to pay for the query, which will limit the adversary's ability to generate unlimited images. To evaluate how the query budget influences the performance of our attack, we change the number of generated images (positive training samples) and see their impact on the attack performance. Specifically, the adversary generates images from the target model of DDPM/DDIM/FastDPM which is trained on the dataset of CIFAR-10 or CelebA. Then we change the number of generated images from 100 to 1,000 with a step size of 100 and from 1,000 to 5,000 with a step size of 1,000. As Figure 4 shows, the attack performance advances when the number of generated images increases. Then the attack performance saturates when the number of generated images arrives at 700. Interestingly, even though the generated image number is only 100, our attack is still effective in most cases. For instance, when the target model is DDPM, DDIM or FastDPM trained on CelebA, our attack achieves the AUC of 0.703, 0.717 and 0.671, respectively. This obeservation indicates our attack is applicable even the query budget is insufficient.

4.4 Transferability of Our Attack

We further consider the transferability of our attack, i.e., the attack performance when the attacker does not have the auxiliary dataset that comes from a similar distribution as the training dataset. We conduct experiments on three diffusion models, DDPM, DDIM and FastDPM, and the attack performance are shown in Table 2.

In the table, each column corresponds to one attack setting. Concretely, "All" means we use all four datasets (three non-matching and one matching datasets) to train the attack model, "Leave" means we use three non-matching datasets except the matching one to train the attack model, and "CIFAR-10"/"CelebA"/"Bedroom"/"Church" means we train the attack model using the named dataset. To highlight the transferability results, we cover the duplicate results of Figure 3 by "-".

The comparison between "All" and "Leave" shows the advantage of the similarity between the auxiliary dataset and the training dataset. We take CelebA as an example, under the "All" setting, our attack achieves 0.974 AUC, whereas the attack only has 0.665 AUC under the "Leave" setting. However, this advantage is not significant on Bedroom and Church, which indicates the auxiliary dataset of LSUN is not high-quality enough, consistent with the conclusion we draw in Section 4.2.

Regarding "CIFAR-10"/"CelebA"/"Bedroom"/"Church", we can see that involving a single non-matching dataset cannot implement a strong attack as strong as "Leave". For instance, when the member dataset is CelebA but the attack model is trained on the generated and auxiliary images of CIFAR-10/Bedroom/Church, the attack performance drops from the AUC of 0.711 to 0.557/0.619/0.613. These results are as expected since more data can provide more knowledge to train the attack model, which sheds new light on improving the attack performance, i.e., involving more data from other domains. In addition, regardless of transferability settings, our attack significantly outperforms Random Guess, further indicating the generalizability of our attack.

4.5 Effectiveness on Conditional Generation Tasks

This part, we evaluate the effectiveness of our attack on conditional generation tasks, as we mentioned in Section 3.1, conditional generation tasks take extra information as input to guide the generation process. We involve nine generative models for evaluation, including diffusion models and GANs. Concretely, we adopt VQGAN (class-conditional generation), LDM (text-conditional generation), CC-FPSE (semantic-conditional generation), Colorization (colorization), LIIF (super resolution), MAT (image inpainting), SwappingAutoencoder (stylization), and MPRNet (denoising and artifact reduction) for evaluation.

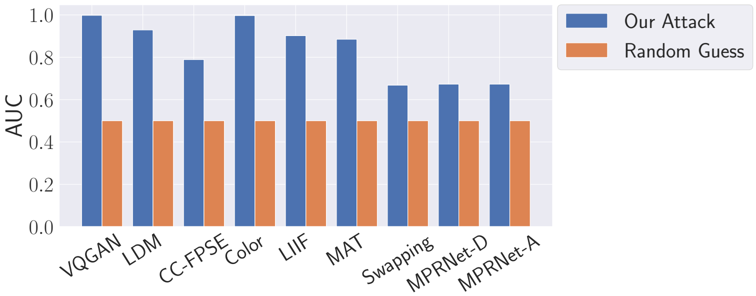


Figure 5: The attack performances of 9 generative models, where "Color" and "Swapping" in the x-axis are the models of "Colorization" and "SwappingAutoencoder". "MPRNet-D" and "MPRNet-A" are the corresponding versions of MPRNet for the applications of denoising and artifact reduction.

The results of our attack against the above nine models are depicted in Figure 5, where "Color" and "Swapping" in the x-axis represent the models of "Colorization" and "SwappingAutoencoder". "MPRNet-D" and "MPRNet-A" are the corresponding versions of MPRNet for the applications of denoising and artifact reduction. According to the results, our attack shows promising performances on various generative models. For instance, our attack achieves the AUC of 0.998, 0.929 and 0.789 against the models of VGGAN (class-conditional generation), LDM (text-conditional generation) and CC-FPSE (semantic-conditional generation) on the datasets of Imagenet, LAION and COCO, respectively. And when the target model is LIFF trained on the dataset of CelebA-HQ, our attack achieves the AUC of 0.902. These results show that our attack can effectively expose the private information of various generative models.

We also confirm our finding in Section 4.2 that the attack performance is consistent with the choice of auxiliary dataset. Specifically, the higher the similarity between the member and auxiliary datasets is, the stronger the attack is. For instance, when the adversary uses UTKFace as the auxiliary dataset for CelebA-HQ in the applications of super resolution and image inpainting, the attack performance is the AUC of 0.902 and 0.885, respectively. However, when wildly collected church images are used as the auxiliary dataset for LSUN-Church, the attack only achieves the AUC of 0.668. At the same time, experiments also show that our attack still works even though the similarity is not high enough between the member and auxiliary datasets. For instance, when the adversary uses ImageNet as the auxiliary dataset for SIDD, the attack achieves the AUC of 0.673 for both denoising and artifact reduction applications. These results indicate a larger scope of application scenarios where the adversary might not gain sufficient budget for data collection.

5 Conclusion

Generative models increasingly show their promising talents in generating realistic and creative images. However, the privacy risks introduced by them are largely unexplored. In the paper, taking advantage of generated distributions, we propose the first generalized membership inference attack to quantify the privacy leakage of existing generative models in a variety of applications. Compared to previous works, our attack gains three main advantages. First, our attack has relaxed assumptions that the adversary only needs to have black-box access to the target model. Second, no shadow model is not involved due to the expensive training cost, which makes our attack computationally efficient. Finally, our attack is architecture-agnostic and task-agnostic, which can be generalized to extensive generative models. These advantages make our attack easily to be deployed in real-world scenarios, which can serve as a quantitative evaluation tool to assess the privacy leakage of existing generative models. We conducted extensive experiments to evaluate the effectiveness of our attack. Experimental results confirm that our attack outperforms existing attacks and can even achieve perfect performance for certain datasets. Further studies show that our attack still works under strict limitations like a limited query budget or a low-quality auxiliary dataset, and the transferability makes our attack a real threat in real-world scenarios. In summary, our attack exposes a common and inherent privacy vulnerability of generative models, aiming to inspire the awareness of privacy protection in the community.

REFERENCES

https://www.cs.toronto.edu/~kriz/cifar.html.

Abdelrahman Abdelhamed, Stephen Lin, and Michael S. Brown. A high-quality denoising dataset for smartphone cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1692–1700. IEEE, 2018.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, pp. 214–223. PMLR, 2017.

Maryam Azadmanesh, Behrouz Shahgholi Ghahfarokhi, and Maede Ashouri-Talouki. A white-box generator membership inference attack against generative models. In *Conference on Information Security and Cryptology (ICISC)*, pp. 13–17. IEEE, 2021.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large Scale GAN Training for High Fidelity Natural Image Synthesis. In *International Conference on Learning Representations (ICLR)*, 2019.

Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 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.

Tian Qi Chen, Xuechen Li, Roger B. Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 2615–2625. NeurIPS, 2018.

Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8628–8638. IEEE, 2021.

Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *International Conference on Machine Learning (ICML)*, pp. 1964–1974. PMLR, 2021.

Adam Coates, Andrew Y. Ng, and Honglak Lee. An Analysis of Single-Layer Networks in Unsupervised Feature Learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 215–223. JMLR, 2011.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 248–255. IEEE, 2009.

Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12873–12883. IEEE, 2021.

Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. LOGAN: Evaluating Privacy Leakage of Generative Models Using Generative Adversarial Networks. *Privacy Enhancing Technologies Symposium*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016.

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*, 2017.

Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. *Privacy Enhancing Technologies Symposium*, 2019.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.

Hailong Hu and Jun Pang. Stealing Machine Learning Models: Attacks and Countermeasures for Generative Adversarial Networks. In *Annual Computer Security Applications Conference (ACSAC)*, pp. 1–16. ACM, 2021.

Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*. PMLR, 2014.

Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *CoRR abs/2106.00132*, 2021.

Wenbo Li, Zhe Lin, Kun Zhou, Lu Qi, Yi Wang, and Jiaya Jia. MAT: mask-aware transformer for large hole image inpainting. *CoRR abs/2203.15270*, 2022.

Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 880–895. ACM, 2021.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *European Conference on Computer Vision (ECCV)*, pp. 740–755. Springer, 2014.

Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, and Hongsheng Li. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 568–578. NeurIPS, 2019.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, pp. 3730–3738. IEEE, 2015.

Taesung Park, Jun-Yan Zhu, Oliver Wang, Jingwan Lu, Eli Shechtman, Alexei A. Efros, and Richard Zhang. Swapping autoencoder for deep image manipulation. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *CoRR abs/2103.17249*, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10684–10695. IEEE, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241. Springer, 2015.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. *CoRR abs/2205.11487*, 2022.

Ahmed Salem, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2019.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. *CoRR abs/2111.02114*, 2021.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership Inference Attacks Against Machine Learning Models. In *IEEE Symposium on Security and Privacy (S&P)*, pp. 3–18. IEEE, 2017.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, pp. 3483–3491. NeurIPS, 2015.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021.

Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein autoencoders. In *International Conference on Learning Representations (ICLR)*, 2018.

Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pp. 6306–6315. NIPS, 2017.

Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8798–8807. IEEE, 2018.

Fisher Yu, Yinda Zhang, Shuran Song, Ari Seff, and Jianxiong Xiao. LSUN: construction of a large-scale image dataset using deep learning with humans in the loop. *CoRR abs/1506.03365*, 2015.

Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Ben Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling Autoregressive Models for Content-Rich Text-to-Image Generation. *CoRR abs/2206.10789*, 2022.

Syed Waqas Zamir, Aditya Arora, Salman H. Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. Multi-stage progressive image restoration. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14821–14831. IEEE, 2021.

Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-Attention Generative Adversarial Networks. *CoRR abs/1805.08318*, 2018.

Minxing Zhang, Zhaochun Ren, Zihan Wang, Pengjie Ren, Zhumin Chen, Pengfei Hu, and Yang Zhang. Membership Inference Attacks Against Recommender Systems. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pp. 864–879. ACM, 2021.

Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pp. 649–666. Springer, 2016.

Zhifei Zhang, Yang Song, and Hairong Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4352–4360. IEEE, 2017.

Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ADE20K dataset. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5122–5130. IEEE, 2017.

Junhao Zhou, Yufei Chen, Chao Shen, and Yang Zhang. Property Inference Attacks Against GANs. In *Network and Distributed System Security Symposium (NDSS)*. Internet Society, 2022.