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

Minxing Zhang[1] Ning Yu[2] Rui Wen[1] Michael Backes[1] Yang Zhang[1]

[1] *CISPA Helmholtz Center for Information Security* [2] *Salesforce Research*

Abstract

Generative models have demonstrated revolutionary success in various visual creation tasks, but in the meantime, they have been exposed to the threat of leaking private information of their training data. Several membership inference attacks (MIAs) have been proposed to exhibit the privacy vulnerability of generative models by classifying a query image as a training dataset member or nonmember. However, these attacks suffer from major limitations, such as requiring shadow models and white-box access, and either ignoring or only focusing on the unique property of diffusion models, which block their generalization to multiple generative models. In contrast, we propose the first generalized membership inference attack against a variety of generative models such as generative adversarial networks, [variational] autoencoders, implicit functions, and the emerging diffusion models. We leverage only generated distributions from target generators and auxiliary nonmember datasets, therefore regarding target generators as black boxes and agnostic to their architectures or application scenarios. Experiments validate that all the generative models are vulnerable to our attack. For instance, our work achieves attack AUC > 0.99 *against DDPM, DDIM, and FastDPM trained on CIFAR-10 and CelebA. And the attack against VQGAN, LDM (for the text-conditional generation), and LIIF achieves AUC* > 0.90. *As a result, we appeal to our community to be aware of such privacy leakage risks when designing and publishing generative models.*[1]

1. Introduction

The recent arms race in visual generation has reached a new peak. Dall-E-2 [43], Imagen [46], Parti [57], and Stable Diffusion [44], driven by big data and empowered by deep generative models, have emerged one after another in a short period of time to showcase their improved power. In the backend, generative adversarial networks (GANs) [21], [variational] autoencoders ([V]AEs) [7, 28], implicit functions (IFs) [38], and the emerging denoising diffusion probabilistic models (DDPMs) [26] are four representative milestones. While we enjoy the benefits of these technologies, the threats of privacy leaks [6, 13, 27, 64] simultaneously come with them and can be easily overlooked. And privacy leakages will get severe especially when data are misused, e.g. the DeepMind project.[2]

To this end, we aim to expose the potential privacy risks from generative models through the lens of membership inference attacks (MIAs), which is by far one of the most popular and powerful techniques [47, 49]. MIAs target to infer whether a query sample is in the training set of the target model or not. Most existing MIAs aim at discriminative models [11, 16, 33, 60], which cannot directly extend to generative models due to the distinct architectures. A few existing MIAs explore privacy risks in generative models [6, 13, 19, 22]. This kind of attack is worth the attention due to multifold reasons: Malicious adversaries can leverage MIAs to cause severe consequences. For instance, GANs are applicable to medical images [55] where the training data (i.e., the medical history of patients) is sensitive, and the privacy leakage may threaten the target individual's life. Meanwhile, MIAs can also be used for benign purposes such as auditing, i.e. quantitatively assessing whether licensed data samples are illegally pirated for generator training or whether private information is misused without permission. For instance, everyone can audit whether their private photo is involved in any dataset on which generative models can be trained.[3] Moreover, MIAs are commonly used as the building block for more sophisticated attacks, which makes it an important topic and has attracted lots of attention in recent years. For instance, MIAs can be an important module for the holistic risk assessment of generative models [36].

However, the applicable scope of existing MIAs against generative models is limited in two aspects. First, these

[1] Code at https://github.com/minxingzhang/MIAGM

[2] https://news.sky.com/story/google-received-1-6-million-nhs-patients-data-on-an-inappropriate-legal-basis-10879142

[3] https://twitter.com/LapineDeLaTerre/status/1570889343845404672

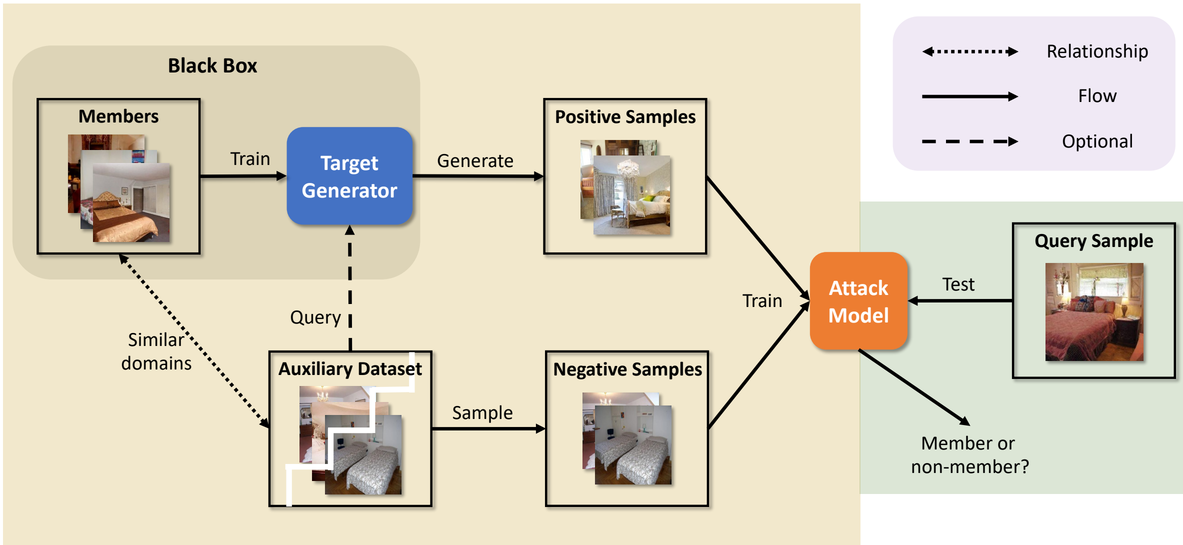


Figure 1. The framework of our membership inference attack against a target generative model.

works are designed under certain attack assumptions, e.g. requiring white-box access and shadow models [6, 31]. Second, they fail to fit a variety of generative applications [13, 19]. Concretely, most of these works ignore the diversity property of DDPMs, which hinders their application in state-of-the-art generative models; on the contrary, a few recent works only take advantage of the unique features of DDPMS, but do not take other generative models into consideration, i.e., GANs, [V]AEs, and IFs. These limitations motivate researchers to propose more practical and generalized MIAs.

In this paper, utilizing the generated distribution learned from the target generator's outputs, we propose a generalized membership inference attack against various generative models in a variety of applications. Compared to existing attacks, our method has three main advantages: (i) **Relaxed assumptions**. Our work assumes that the attacker only has black-box access to target models. That is, the attacker only needs to query target models in an API manner. (ii) **Computationally efficient**. Contrary to existing attacks, our method does not require training shadow models, which makes the method computationally efficient. (iii) **Generalizability**. Our attack can be applied to a variety of generative models including but not limited to the state-of-the-art DDPMs, whereas previous MIA designs are specified to certain model variants, for example, Azadmanesh et al. only evaluate privacy risks of GANs [6] and Duan et al. only focus on DDPMs [19]. Besides, generalizability also benefits from our relaxed assumptions. The black-box access represents a common scenario in practice, based on which our attack is independent of generator architectures, indicating that our attack is applicable to various generative models.

As Fig. 1 shows, we use generated samples as training positives and auxiliary samples as training negatives. In particular, when attacking conditional generators, the inputs to the target generators are also sampled from the auxiliary dataset but disjoint with training negatives. In this case, still, the generated samples leak the information of members that were used to train the target generator. This relies on the target generator memorizing the training distribution due to overfitting. Then the distribution of generated images is to some extent an approximation of the training distribution. Thus, the attack model can indirectly learn the target training distribution from the generated images, and further predict the membership of the target generator.

Regarding auxiliary datasets, they are involved due to a currently prevalent setting. That is, to train large-scale generative models, model owners usually collect data as much as possible, and use the whole dataset in the training process. In this case, the adversary's accessible data are all target training data, i.e., members, and no left and unused data from the same dataset can be used as negative training samples for the attack model. To this end, the adversary collects auxiliary datasets of similar domains as member samples.

We extensively evaluate our MIA method against four widely utilized generative model frameworks (GANs, [V]AEs, IFs, DDPMs) covering ten generation applications (unconditional [26, 29, 51], class-conditional [20], text-conditional [45] and semantic-conditional [35] generation, image colorization [61], super resolution [15], image inpainting [32, 45], stylization [40], denoising [58], and artifact reduction [58]). For the pioneering attacks against unconditional DDPM, we cover the original DDPM [26], DDIM [51], and FastDPM [29]. Experimental results show the efficacy of our attacks on all the scenarios. For instance, our work achieves attack AUC > 0.99 against DDPM, DDIM, and FastDPM trained on CIFAR-10 and CelebA. And the attack against VQGAN, LDM (in the application of text-conditional generation), and LIIF achieves AUC > 0.90. Further studies demonstrate that the generative models are still vulnerable to our attack even with a limited query budget, and the transferability makes our attack a real threat in real-world scenarios.

To better understand our work, we further discuss the membership boundary, the overlap between auxiliary and member samples, and the feasibility of generating training negatives.

Our contributions:

- We propose the first generalized MIA against various generative models (including GANs, [V]AEs, IFs, and state-of-the-art DDPMs) in various visual generation applications.
- Our work is the first to exploit generated distributions learned from the target generator's outputs for membership inference. Generated distributions are the common property of generators and are independent of generator architectures. Meanwhile, our work requires fewer assumptions and is more computationally efficient than previous attacks. These advantages earn our attack a practical and general application scope.
- We empirically demonstrate the efficacy of our attack in all the scenarios, which in turn validate the generalizability and practical usage of our attacks. Besides, further studies showcase more advantageous properties, such as effectiveness with a limited query budget and transferability, which enhance the practical significance of the attack.
- Our work fills the blank of understanding the common privacy risks of various generative models, which motivates researchers to take privacy threats into concern when designing and publishing generative models.

2. Related Works

Generative models and applications: In this work, we consider the widely used generative models, i.e., generative adversarial networks (GANs) [5, 8, 41, 54, 59] which utilize the arms race between generators and discriminators to improve the quality of generated images; [variational] autoencoders ([V]AEs) [14, 24, 28, 50, 52, 53] which restore the input images from latent representations sampled from estimated normal distributions; implicit functions (IFs) [9, 15, 38, 39] which circumvent the need for an explicit likelihood; and diffusion probabilistic models (DDPMs) [26, 29, 43, 45, 46, 51] which denoise noise images step by step until being clean images. These above models cast significant success towards a variety of visual generation applications, including unconditional [26, 29, 51], class-conditional [20], text-conditional [45] and semantic-conditional [35] generation, image colorization [61], super resolution [15], image inpainting [32, 45], stylization [40], denoising [58], and artifact reduction [58]. And these applications require different kinds of inputs, i.e., noises, texts, and images. We, therefore, nest our experiments in these applications to investigate the efficacy and generalizability of our attacks against various generative models.

Membership inference attacks against generative models: Many existing MIAs against generative models focus on GAN- and [V]AE-based generative models, which limits the applicable scope of membership inference. For instance, Hayes et al. aim at several GANs and a single [V]AE [22]; Hilprecht et al. propose two types of MIAs, one applicable to GANs and [V]AEs while the other specific for [V]AEs [25]; Chen et al. present the first taxonomy of membership inference, which mainly focuses on GANs [13]; and Azadmanesh et al. conduct a white-box membership inference against GANs [6]. Unfortunately, these works cannot generalize well to all aforementioned generative models and applications. For instance, due to the diverse property of DDPMs, the attacks proposed by Chen et al. cannot achieve performances as strong as against GANs. Recently, a few works explore the privacy issues of DDPMs. For instance, Duan et al. utilize step-wise errors to infer membership [19]. However, this work relies on the unique feature of DDPMs, which is not available to other generative models, i.e., GANs, [V]AEs, and IFs; Carlini et al. present the training data extraction of DDPMs [12], which is another popular topic regarding privacy leakage but aims at a different goal of our attack. To propose a generalized membership inference attack, we take advantage of generated distributions that are architecture-agnostic. Thus our work is applicable to various generative models and applications.

3. Our Attack

3.1. Problem Statement

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

$$\mathcal{A}(\mathbf{x}_{\text{query}}|\mathcal{G}_{\text{target}}, \mathcal{K}) \rightarrow \{\text{member}, \text{non-member}\} \quad (1)$$

where $\mathcal{K}$ denotes extra information known to the attacker, and members and non-members are the data samples used and not used to train the target generator, respectively. In our work, we assume the attacker has access to an auxiliary dataset $\mathcal{D}_{\text{aux}}$ of similar domains as the training dataset $\mathcal{D}_{\text{train}}$ (i.e., members), which will be explained in Sec. 3.3. Hereby $\mathcal{K} \doteq \mathcal{D}_{\text{aux}}$. The auxiliary dataset $\mathcal{D}_{\text{aux}}$ is separated into two disjoint parts, i.e., $\mathcal{D}_{\text{aux}}^{\text{in}}$ for querying the conditional target generator and $\mathcal{D}_{\text{aux}}^{\text{out}}$ as training negatives for the attack model, which will be detailed in the following.

3.2. Methodology

Fig. 1 depicts the general framework of our attack. The attack process consists of two steps: image generation and membership inference, with the former one being the core step of our attack.

Image generation: To infer the membership of the target generator, the attacker needs to collect training positives and negatives for the establishment of their attack model. As mentioned in Sec. 3.1, the training negatives are derived from $\mathcal{D}_{\text{aux}}^{\text{out}}$, formally:

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

Thus, the current challenge is how to obtain quantities of training positives, since the member samples are inaccessible to the adversary. Due to overfitting, the target generator memorizes its training distribution. In that case, the generated images can reflect the pattern of real members to a large extent. Consequently, the samples generated by the target generator are used as the training positives: (i) For unconditional generation tasks, we query the generative model with Gaussian noise, i.e., $z \sim \mathcal{N}(0,1)$. Then we obtain a positive training pair, i.e., $(\mathcal{G}_{\text{target}}(z), \text{member})$, for the attack model; (ii) For conditional generation tasks, the target generator receives information to guide the generation process. As an example, we illustrate the case where the input is image. As mentioned in Sec. 3.1, we query the target generator using the data sampled from $\mathcal{D}_{\text{aux}}^{\text{in}}$. Then the training positives are formulated as:

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

Note that our attack is architecture- 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.

Membership inference: The adversary establishes a binary classifier as the attack model $\mathcal{A}$ in a supervised way using the training positives and negatives constructed in the previous step. Then, the adversary can infer the membership status of a query sample $\mathbf{x}_{\text{query}}$. Note that, in the inference procedure, the adversary does not need to interact with the target generative model. Specifically, the adversary directly queries the trained attack model $\mathcal{A}$ with the query sample $\mathbf{x}_{\text{query}}$. And the membership status is inferred according to the output of the attack model. This process can be formulated as:

$$\mathcal{A}(\mathbf{x}_{\text{query}}) \rightarrow y_{\text{query}} \tag{4}$$

where $y_{\text{query}} \in \{0,1\}$ indicates the membership status of the query sample $\mathbf{x}_{\text{query}}$.

Different from previous works that rely on strong assumptions (i.e., white-box access to the target generator), we only require black-box access, which means that interactions with the generator by the attacker can only happen through an API manner. And thanks to the black-box access, our attack is architecture-agnostic and can be generalized to a variety of generative models. To better understand our work, we empirically validate this relaxed and practical assumption by extensive experiments in Sec. 4, and clarify the superiority of our attack as follows:

The effectiveness: Our attack utilizes generated distributions learned from the target generator's outputs to conduct membership inference. This is because overfitting helps the target generator remember the distribution of its training data. Thus the distribution of generated images can be regarded as an approximate of the training distribution. In that case, the attack model can indirectly learn the training distribution from the generated images, and further predict the membership of the target generator. Note that, image generation is a basic function of generative models, so generated images are of course available via black-box access in all generative applications.

The advantages: (i) *Relaxed assumptions.* Our work only assumes black-box access to the target generator, which is a prevalent setting in practice and thus largely enhances the attacker's ability compared to white-box access. (ii) *Computationally efficient.* Different from many existing works, no requirement of training shadow models makes our approach more computationally efficient. (iii) *Generalizability.* Not like previous MIAs that ignore or only focus on DDPMs, our method is applicable to all the widely used generative models, i.e., GANs, [V]AEs, IFs, and DDPMs.

3.3. Auxiliary Datasets

As aforementioned, our approach involves an auxiliary dataset of similar domains as the training dataset of the target generator. We will specify this in this section.

Why use auxiliary datasets? A currently prevalent setting is that the owners of large-scale generative models usually collect data as much as possible, and use the whole dataset for training. In that case, the adversary's accessible data are all target training data, i.e., members. That is, no left and unused data from the same dataset can be used as training negatives for the attack model. To this end, the adversary samples training negatives from auxiliary datasets of similar domains as member samples.

Why similar domains? The decision of similar domains is because the adversary is hard to obtain an auxiliary dataset of the exact domains of the target generator, and the discussion of obviously different domains is trivial. Specifically, from the generated images, the adversary can easily estimate the target domains, e.g., generated ImageNet [1] images depict dogs, cats, etc. However, the adversary's ability might be limited by query times/cost, thus the estimated domains would only be similar but not exactly the same. And even if possible, it is hard to guarantee the existence of available datasets of the exact domains, e.g. LAION. Thus,

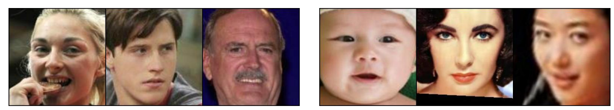
Table 1. The settings of generative applications, target generative models, and member and auxiliary datasets.

Application	Target generative model			Attack binary classifier			
				Training images		Testing images	
	Technique	Framework	Training/Member dataset	Positive	Negative	Positive	Negative
Unconditional	DDPM [26] DDIM [51] FastDPM [29]	DDPM	CIFAR-10 CelebA LSUN-Bedroom LSUN-Church	$\mathcal{G}$(noise)	STL-10 UTKFace Wild Bedroom Wild Church	The same as target generator training dataset	The same as attacker training negative but disjoint split
Class-conditional	VQGAN [20]	GAN	ImageNet	$\mathcal{G}$(class)	Open Images		
Text-conditional	LDM [45]	DDPM	LAION	$\mathcal{G}$(COCO text)	COCO		
Semantic-conditional	CC-FPSE [35]	GAN	COCO	$\mathcal{G}$(ADE20K semantics)	ADE20K		
Image colorization	Colorization [61]	[V]AE	ImageNet	$\mathcal{G}$(Open Images gray)	Open Images		
Super resolution	LIIF [15]	IF	CelebA-HQ	$\mathcal{G}$(UTKFace low res)	UTKFace		
Image inpainting	MAT [32] LDM [45]	GAN+[V]AE DDPM	CelebA-HQ LAION	$\mathcal{G}$(UTKFace incomplete) $\mathcal{G}$(COCO text)	UTKFace COCO		
Stylization	SwappingAutoencoder [40]	GAN+[V]AE	LSUN-Church	$\mathcal{G}$(Wild Church)	Wild Church		
Denoising	MPRNet [58]	[V]AE	SIDD	$\mathcal{G}$(ImageNet noisy)	ImageNet		
Artifact reduction	MPRNet [58]	[V]AE	SIDD	$\mathcal{G}$(ImageNet artifact)	ImageNet		

aiming at similar domains instead of the exact ones brings a broader application scope. On the other hand, according to the definition, members are the training samples of the target generator, which means all samples that are not used for training are non-members. In that case, the images of obviously different domains will be easily classified as non-members, so the challenge is about similar domains.

How to construct? Based on the estimated domain, there are two ways to construct an auxiliary dataset: (i) Different datasets have domain overlaps, e.g., CIFAR-10 [2] and STL-10 [3] both contain the airplane category. Thus, the adversary can utilize an existing dataset of similar domains. Note that, the larger the domains of the auxiliary dataset, the more the possibility of covering the target domains, and of course the better. For instance, even though CIFAR-10 and Open Images [30] both describe real-world objects, we use Open Images as the auxiliary dataset for ImageNet since it contains more categories; (ii) The alternative is to manually collect auxiliary samples when no existing dataset of similar domains is available. For instance, the bedroom images collected from the internet are used as the auxiliary dataset for LSUN-Bedroom [56].

Ideally, there is no overlap between the auxiliary dataset and the generator's training data (i.e., members), otherwise, the (part of) training negatives sampled from the auxiliary dataset should be labeled as "members", which is contrary to the previous setting. However, since the member samples are inaccessible, the adversary cannot check and guarantee non-overlap. To this end, we conduct more explorations in Sec. 5 to understand the influence of overlap.



(a) (**CelebA**, UTKFace) (b) (CelebA, **UTKFace**)

Figure 2. The image examples of different (member, auxiliary) dataset pairs, where the images are sampled from the **boldface dataset**. The complete version is in Fig. 7 (Appendix A).

4. Evaluation

4.1. Evaluation Settings

Generative applications and models: To comprehensively understand our work, we conduct the evaluation on ten generative applications, i.e., unconditional, class-conditional, text-conditional and semantic-conditional generation, image colorization, super resolution, image inpainting, stylization, denoising, and artifact reduction, as listed in the first column of Tab. 1. Extensive generative techniques of these applications are involved to measure the attack performance in various scenarios, as shown in the second column of the table. Specifically, DDPM [26], DDIM [51] and FastDPM [29] for unconditional generation; VQGAN [20] for class-conditional generation; LDM [45] for text-conditional generation; CC-FPSE [35] for semantic-conditional generation; Colorization [61] for image colorization; LIIF [15] for super resolution; MAT [32] and LDM for image inpainting; SwappingAutoencoder [40] for stylization; MPRNet [58] for both denoising and artifact reduction. Regarding the backbone frameworks of these techniques, four widely used generative models are covered, which corresponds to the third column of the table, i.e., GANs (including VQQGAN, CC-FPSE, MAT, and SwappingAutoencoder), [V]AEs (including Colorization, MAT, SwappingAutoencoder, and MPRNet), IFs (including LIIF), and DDPMs (including DDPM, DDIM, FastDPM, and LDM).

Member and auxiliary datasets: As mentioned in Sec. 3, the adversary accesses an auxiliary dataset for each training/member dataset of target generators. Thus, we introduce the dataset pairs of (member, auxiliary) used for different applications in the other columns of Tab. 1, and show some examples in Fig. 2 (the complete version is shown in Fig. 7 of Appendix A): (CIFAR-10 [2], STL-10 [17]) $\rightarrow$ unconditional generation; (CelebA(-HQ) [37], UTKFace [62]) $\rightarrow$ unconditional generation, super resolution and image inpainting; (LSUN-Bedroom [56], Wild Bedroom) $\rightarrow$ unconditional generation; (LSUN-Church, Wild Church) $\rightarrow$ unconditional generation and stylization; (ImageNet [18], Open Images [30]) $\rightarrow$ class-conditional generation and image colorization; (LAION [48], COCO [34]) $\rightarrow$ text-conditional generation and image inpainting; (COCO, ADE20K [63]) $\rightarrow$ semantic-conditional generation; and (SIDD [4], ImageNet) $\rightarrow$ denoising and artifact reduction. These datasets are further introduced in Appendix B. Even in some cases, the domains of auxiliary datasets are not exactly the same as the members' as explained in Sec. 3.3, e.g., frog images in CIFAR-10 but not in STL-10, surprisingly, our attack still works well in all applications as shown in Sec. 4.2.

Attack model: The attack model is established based on a pre-trained ResNet18 [23], of which the last linear layers are fitted for binary classification, i.e., inferring the query sample belonging to members or non-members. More training details are shown in Appendix C.

Evaluation metric: We utilize AUC as the metric to evaluate attacks [13,33,47,60], of which the value is in the range of $[0.0, 1.0]$, and the higher the more effective. Specifically, during the testing, only the images used to train the target generator are regarded as positive samples while the others are all regarded as negative samples.

4.2. Results

In this part, we evaluate the effectiveness of our attack on ten generative applications and depict the experimental results in Fig. 3. To better show the superiority, we compare our attack with the black- and white-box attacks of Chen et al. [13] which are one of the most popular MIAs against generative models. Moreover, we compare our attack against DDPM and DDIM with SecMI [19] which is a recently proposed MIA work specific to diffusion models.

Fig. 3 shows that all generative scenarios are vulnerable to our attack. For instance, our attack depicts consistent efficacy with AUC > 0.99 on DDPM, DDIM, and Fast-DPM trained on CIFAR-10 or CelebA as shown in Fig. 3a. And the attack achieves AUC > 0.9 on VQGAN, LDM-T, and LIIF trained on ImageNet, LAION, and CelebA-HQ, respectively, as shown in Fig. 3b. Meanwhile, the black-box baseline [13] only achieves AUC < 0.6. The obvious advantages over the baselines well demonstrate the effectiveness of our work in various applications, which further indicates that our attack can effectively expose the private information of various generative models.

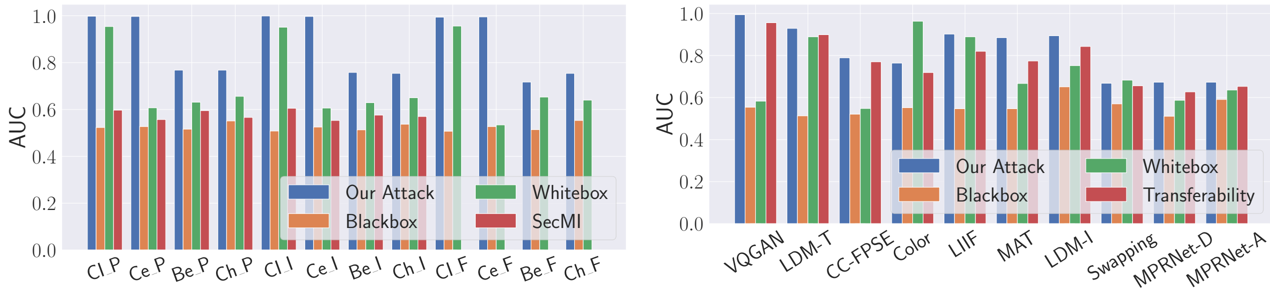
An interesting finding appears in conditional generation tasks. Specifically, our generated images (i.e., training positives for the attack model) are derived by querying the target generator using a split of an auxiliary dataset, which indicates the generated images are associated with auxiliary data. Even in such a non-trivial setting, surprisingly, the boundary learned by the attack model from the generated and auxiliary samples can also serve as an effective boundary during the testing to separate member samples and another split of auxiliary samples. For instance, our attack on LIIF (for the super resolution) achieves an AUC of 0.902. This shows the advantage of utilizing generated distributions to learn the information of member samples as mentioned in Sec. 3.2. In other words, the generated images, even if not only associated with member samples, can still expose the membership status of member samples, which introduces a large applicable scope of our attack.

4.3. Transferability of Our Attack

In this section, we consider the transferability of our attack, i.e., the attack performance when testing negatives of the attack model are of different datasets from training negatives, and evaluate it on conditional generative tasks. Concretely, the attack model is established following the settings in Sec. 4.1, i.e., to construct an auxiliary dataset according to the member one, and then to generate/sample training positives/negatives. However, during the testing time, we do not sample negative images from the auxiliary dataset, but from another non-member dataset, i.e., LSUN-Bedroom in this case. As shown in Fig. 3b in red bars, the transferability performance is comparable to the attack performance in Sec. 4.2, and obviously superior to the black-box baseline [13]. Even requiring fewer assumptions than the white-box baseline, in most cases, our attack can achieve stronger transferability performance. For instance, the transferability performance of VQGAN is an AUC of 0.956 while the black- and white-box baselines only achieve an AUC of 0.554 and 0.583, respectively. These significant advantages further indicate the effectiveness of our attack.

4.4. Influence of the Query Budget

We further evaluate the influence of the query budget on the attack performance. Sometimes users will be required to pay for the query, limiting the adversary's ability to generate unlimited images (which also influences the domains of auxiliary datasets as mentioned in Sec. 3.3). And the generation cost is especially expensive for DDPMs compared to other generative models. Thus in this section, we mainly focus on how the query budget influences the attack performance on unconditional generation tasks implemented by DDPMs. Specifically, we change the number of generated images (i.e., training positives), 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, and see the impact on the attack performance. Note that, to balance the training positives and negatives, we also control for an equal number of images sampled from the auxiliary dataset. As Fig. 4 shows, the at-



(a) Unconditional generation tasks.

(b) Conditional generation tasks.

Figure 3. The attack performances in various generative applications. In Fig. 3a, we use CI/Ce/Be/Ch_P/I/F notation in the x-axis to denote the generator DDPM/DDIM/FastDPM trained on the dataset CIFAR-10/CelebA/Bedroom/Church. In the x-axis of Fig. 3b, “Color” and “Swapping” are short for “Colorization” and “SwappingAutoencoder”, “LDM-T” and “LDM-I” represent LDM used for text-conditional generation and image inpainting, and “MPRNet-D” and “MPRNet-A” are MPRNet used for denoising and artifact reduction.

Figure 4. The attack performances of (CelebA, UTKFace) with different query budgets in unconditional generation tasks.

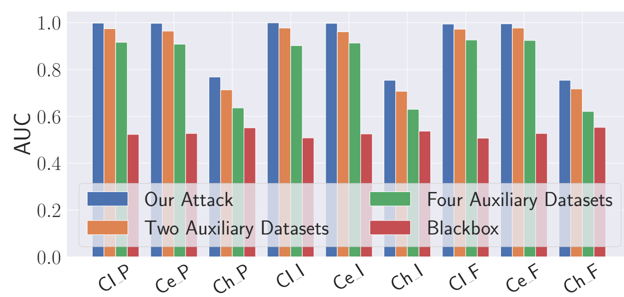


Figure 5. The attack performances of unconditional generation tasks with different numbers of auxiliary datasets, where “Two Auxiliary Datasets” represents the original auxiliary dataset described in Tab. 1 and Wild-Bedroom. Besides, “Four Auxiliary Datasets” represents STL-10, UTKFace, Wild-Bedroom, and Wild-Church. Regarding the x-axis, the setting is the same as Fig. 3.

tack performance advances when the number of generated images increases, where the complete version is shown in Fig. 8 (Appendix A). Then the attack performance saturates when the number of training positives/negatives arrives at 700. Interestingly, even though the generation quantity is only 100, our attack is still effective in most cases. For instance, when targeting DDPM trained on CelebA, our attack achieves an AUC of 0.703 while the black-box baseline only achieves 0.527. This observation indicates our attack is applicable even when the query budget is insufficient.

5. Discussion

Membership Boundary: As mentioned in Sec. 3.3, all samples that are not used to train the target generator are non-members. That means the boundary between members and non-members actually separates members and all other samples. However, in Sec. 4, we only consider a single auxiliary dataset as non-members (i.e., training and testing negatives). Thus, to better understand our work, we evaluate cases where more auxiliary datasets are involved as non-members. As shown in Fig. 5, with more auxiliary datasets, our attack performance of unconditional generation tasks drops but is still strong and significantly better than the black-box baseline [13] in most cases, while that of conditional generation tasks shows similar results as shown in Fig. 9 (Appendix A). This decrease is because more auxiliary datasets (i.e., non-members) will make the boundary more complicated and then harder to learn by the attack model. This finding leaves an interesting future work on dealing with complicated boundaries.

Overlap Between Auxiliary and Member Samples: As mentioned in Sec. 3.3, ideally, there is no overlap between auxiliary and corresponding member samples. However, since members are inaccessible, non-overlap is hard to guarantee in practice. Thus, we here explore the influence of overlaps on our attack by mixing member samples into auxiliary datasets by different ratios. We can see from Fig. 6 that as expected, the smaller the overlap, the better the attack performance. Surprisingly, even when a 0.1 ratio of members is mixed into auxiliary datasets, our attack can still achieve an AUC $>$ 0.7 in most cases, which further indicates the practicability and effectiveness of our work.

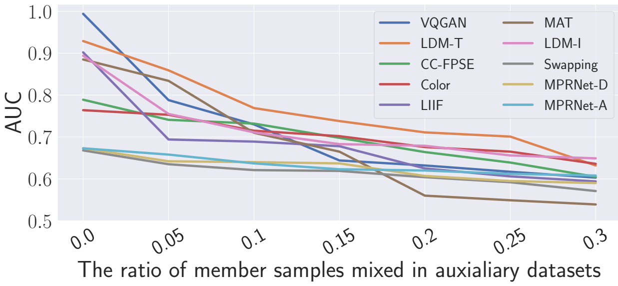


Figure 6. The attack performances with different ratios of overlap between members and auxiliary datasets.

Table 2. The attack performances with different sources of training negatives, i.e., sampled from the auxiliary dataset or generated by a shadow model.

Application	Technique		Dataset		Negatives	
	Target	Shadow	Member	Auxiliary	Sampled	Generated
Text-conditional	LDM	CC-FPSE	LAION	COCO	0.929	0.938
Image inpainting					0.894	0.741
Unconditional	DDPM	FastDPM	CelebA	UTKFace	0.997	0.995
	DDIM				0.997	0.996

Table 3. The attack performances regarding of members and non-members being from the same distribution, and measuring by the metric proposed by [10].

Technique	Dataset	Same Distribution		Evaluation Metric					
		Ours	SecMI	Ours		Blackbox		Whitebox	
				TPR	FPR	TPR	FPR	TPR	FPR
DDPM	CelebA	**0.545**	0.516	**0.998**	**0.002**	0.517	0.472	0.606	0.397
DDIM		**0.543**	0.519	**0.997**	**0.002**	0.515	0.480	0.604	0.402

Generation of Training Negatives: As described in Sec. 3.2, training negatives are sampled from auxiliary datasets while training positives are generated by generative models. Thus, we wonder about the feasibility of generating the training negatives for our attack model by a different generative model (i.e., a *shadow model*) from the target generator, which also provides a guarantee of non-overlapping, e.g., training negatives of COCO (UTKFace) are generated by CC-FPSE (FastDPM). From Tab. 2 we can see that generating training negatives achieves a comparable performance to our attack, yet an extra cost of generation will be required. This indicates the practicability of simply sampling training negatives from auxiliary datasets, unless aiming at the non-overlapping guarantee. On the other hand, the attack performance is shown sensitive to the choice of auxiliary datasets (see the first part in Sec. 5), however, finding suitable auxiliary samples in practice might be challenging even though we only require a similar distribution instead of the same one. To this end, involving shadow models to generate training negatives would help this problem.

Membership Status: To further empirically support our method, we now evaluate a stricter assumption where members and non-members are disjointly from the same dataset. Specifically, we use the training and testing samples of CelebA as members and non-members on DDPM and DDIM respectively as shown in the "Same Distribution" column of Tab. 3. Compared to a recent MIA against diffusion models, i.e., SecMI [19], our method shows obviously better attack performance. This result suggests a more comprehensive application scenario of our method.

Evaluation Metric: Carlini et al. propose that the correctness of inferring members is more practically meaningful than non-members [10]. Thus, we evaluate our attack by true positive rate (TPR) and false positive rate (FPR), where a higher TPR and a lower FPR derive a better membership inference. In the "Evaluation Metric" column of Tab. 3, we can see that our method could gain an obviously higher TPR and lower FPR even though the accuracy gap between ours and Whitebox is not as significant (Fig. 3a), indicating the multi-faceted effectiveness of our method.

6. Conclusion

Generative models increasingly show their promising talents in generating realistic and creative images. However, the privacy risks of training data leakage introduced by them are largely unexplored. Previous membership inference attacks have proved that generative models are vulnerable to privacy leakage by inferring whether a query sample is in the training dataset. However, the existing works require shadow models and/or white-box access, and ignore or only focus on the state-of-the-art DDPMs, which limits their application scope. In this paper, we propose the first generalized membership inference attack against various generative models including GANs, [V]AEs, IFs, and DDPMs. We only assume the adversary can obtain generated distributions from target generators. Under the black-box setting, our attack is agnostic to the architectures and applications of generative models. Extensive experimental results show the effectiveness of our attack against various generative models and applications. Further studies show that our attack still works with a limited query budget, and the transferability makes our attack a real threat in real-world scenarios. Consequently, we aim to call for community awareness of the privacy protection of generative models.

Acknowledgement

We would like to thank anonymous reviewers for their insightful and constructive reviews for us to improve our work.

References

[1] https://www.kaggle.com/c/tiny-imagenet. 4
[2] https://www.cs.toronto.edu/~kriz/cifar.html. 5

[3] https://cs.stanford.edu/%7Eacoates/stl10/. 5

[4] 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)*, pages 1692–1700. IEEE, 2018. 6

[5] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein Generative Adversarial Networks. In *International Conference on Machine Learning (ICML)*, pages 214–223. PMLR, 2017. 3

[6] 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)*, pages 13–17. IEEE, 2021. 1, 2, 3

[7] Hervé Bourlard and Yves Kamp. Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics*, 1988. 1

[8] 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. 3

[9] Nicola De Cao and Thomas Kipf. Molgan: An implicit generative model for small molecular graphs. *CoRR abs/1805.11973*, 2018. 3

[10] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership Inference Attacks From First Principles. *CoRR abs/2112.03570*, 2021. 8

[11] Nicholas Carlini, Steve Chien, Milad Nasr, Shuang Song, Andreas Terzis, and Florian Tramèr. Membership inference attacks from first principles. In *IEEE Symposium on Security and Privacy (S&P)*, pages 1897–1914. IEEE, 2022. 1

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

[13] 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)*, pages 343–362. ACM, 2020. 1, 2, 3, 6, 7

[14] 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)*, pages 2615–2625. NeurIPS, 2018. 3

[15] 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)*, pages 8628–8638. IEEE, 2021. 2, 3, 5

[16] Christopher A. Choquette Choo, Florian Tramèr, Nicholas Carlini, and Nicolas Papernot. Label-Only Membership Inference Attacks. In *International Conference on Machine Learning (ICML)*, pages 1964–1974. PMLR, 2021. 1

[17] 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)*, pages 215–223. JMLR, 2011. 6

[18] 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)*, pages 248–255. IEEE, 2009. 6

[19] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *CoRR abs/2302.01316*, 2023. 1, 2, 3, 6, 8

[20] Patrick Esser, Robin Rombach, and Björn Ommer. Taming transformers for high-resolution image synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12873–12883. IEEE, 2021. 2, 3, 5

[21] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial networks. *CoRR abs/1406.2661*, 2014. 1

[22] 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. 1, 3

[23] 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)*, pages 770–778. IEEE, 2016. 6

[24] 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. 3

[25] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. Monte Carlo and Reconstruction Membership Inference Attacks against Generative Models. *Privacy Enhancing Technologies Symposium*, 2019. 3

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS, 2020. 1, 2, 3, 5

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

[28] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*. PMLR, 2014. 1, 3

[29] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *CoRR abs/2106.00132*, 2021. 2, 3, 5

[30] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper R. R. Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *CoRR abs/1811.00982*, 2018. 5, 6

[31] Klas Leino and Matt Fredrikson. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In *USENIX Security Symposium (USENIX Security)*, pages 1605–1622. USENIX, 2020. 2

[32] 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. 2, 3, 5

[33] Zheng Li and Yang Zhang. Membership Leakage in Label-Only Exposures. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 880–895. ACM, 2021. 1, 6

[34] 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)*, pages 740–755. Springer, 2014. 6

[35] 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)*, pages 568–578. NeurIPS, 2019. 2, 3, 5

[36] Yugeng Liu, Rui Wen, Xinlei He, Ahmed Salem, Zhikun Zhang, Michael Backes, Emiliano De Cristofaro, Mario Fritz, and Yang Zhang. ML-Doctor: Holistic Risk Assessment of Inference Attacks Against Machine Learning Models. In *USENIX Security Symposium (USENIX Security)*. USENIX, 2022. 1

[37] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep Learning Face Attributes in the Wild. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738. IEEE, 2015. 6

[38] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1, 3

[39] Shakir Mohamed and Balaji Lakshminarayanan. Learning in implicit generative models. *CoRR abs/1610.03483*, 2016. 3

[40] 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. 2, 3, 5

[41] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. *CoRR abs/2103.17249*, 2021. 3

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning (ICML)*, pages 8748–8763. PMLR, 2021. 14

[43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical Text-Conditional Image Generation with CLIP Latents. *CoRR abs/2204.06125*, 2022. 1, 3

[44] 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)*, pages 10674–10685. IEEE, 2022. 1

[45] 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)*, pages 10684–10695. IEEE, 2022. 2, 3, 5

[46] 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. 1, 3

[47] 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. 1, 6

[48] 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. 6

[49] 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)*, pages 3–18. IEEE, 2017. 1

[50] 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)*, pages 3483–3491. NeurIPS, 2015. 3

[51] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations (ICLR)*, 2021. 2, 3, 5

[52] Ilya O. Tolstikhin, Olivier Bousquet, Sylvain Gelly, and Bernhard Schölkopf. Wasserstein auto-encoders. In *International Conference on Learning Representations (ICLR)*, 2018. 3

[53] Aäron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural Discrete Representation Learning. In *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 6306–6315. NIPS, 2017. 3

[54] 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)*, pages 8798–8807. IEEE, 2018. 3

[55] Xin Yi, Ekta Walia, and Paul S. Babyn. Generative adversarial network in medical imaging: A review. *Medical Image Analysis*, 2019. 1

[56] 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. 5, 6

[57] 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. 1

[58] 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)*, pages 14821–14831. IEEE, 2021. 2, 3, 5

[59] Han Zhang, Ian J. Goodfellow, Dimitris N. Metaxas, and Augustus Odena. Self-Attention Generative Adversarial Networks. *CoRR abs/1805.08318*, 2018. 3

[60] 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)*, pages 864–879. ACM, 2021. 1, 6

[61] Richard Zhang, Phillip Isola, and Alexei A. Efros. Colorful image colorization. In *European Conference on Computer Vision (ECCV)*, pages 649–666. Springer, 2016. 2, 3, 5

[62] Zhifei Zhang, Yang Song, and Hairong Qi. Age Progression/Regression by Conditional Adversarial Autoencoder. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4352–4360. IEEE, 2017. 6

[63] 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)*, pages 5122–5130. IEEE, 2017. 6

[64] 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. 1