Model Extraction and Defenses on Generative Adversarial Networks

Hailong Hu
University of Luxembourg

Jun Pang
University of Luxembourg

ABSTRACT

Model extraction attacks aim to duplicate a machine learning model through query access to a target model. Early studies mainly focus on discriminative models. Despite the success, model extraction attacks against generative models are less well explored. In this paper, we systematically study the feasibility of model extraction attacks against generative adversarial networks (GANs). Specifically, we first define accuracy and fidelity on model extraction attacks against GANs. Then we study model extraction attacks against GANs from the perspective of accuracy extraction and fidelity extraction, according to the adversary's goals and background knowledge. We further conduct a case study where an adversary can transfer knowledge of the extracted model which steals a state-of-the-art GAN trained with more than 3 million images to new domains to broaden the scope of applications of model extraction attacks. Finally, we propose effective defense techniques to safeguard GANs, considering a trade-off between the utility and security of GAN models.

CCS CONCEPTS

• **Security and privacy**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

Model extraction; generative adversarial networks; transfer learning; semantic latent space

1 INTRODUCTION

Over the past few years, machine learning, deep learning in particular, has gained significant advances in a variety of areas, such as computer vision and natural language processing (NLP). For instance, recently FixEfficientNet-L2 [62] has achieved 98.7% accuracy on the ImageNet recognition benchmark [54], compared to the accuracy of 73.8% in 2011. NLP models created by fine-tuning a pretrained model, i.e. BERT [14], also become extremely successful in many natural language tasks [34, 39, 46]. Generative adversarial networks (GANs) are able to generate photo-realistic images that humans are difficult to distinguish [6, 31, 32]. In general, these state-of-the-art models are often considered as the intellectual property of model owners and are closely safeguarded. The reasons are from at least two aspects. First, obtaining a practical deep learning model is non-trivial. This is because training a model requires a large number of training data, intensive computing resources and human resources [7, 14, 32, 54, 62, 68]. Second, deep learning models themselves are confidential, and exposure to deep learning models to potential adversaries poses a threat to security and privacy [41, 42, 50, 55, 59, 63]. However, model extraction attack — a novel attack surface targeting at duplicating a model only through query access to a target model, has recently emerged and gained significant attention from the research community.

In the early study, Tramèr et al. [63] first attempt model extraction on traditional machine learning models and shallow neural networks, such as logistic regression, decision tree, support vector machine and multilayer perceptrons. Since then, Jagielski et al. [26] further mount the attack against a million of parameters model trained on billions of Instagram images [43], which makes model extraction attack more practical. In addition to model extraction on deep convolutional neural networks about image classification, there are some works studying the problem of model extraction in NLP tasks [33, 60]. For instance, with the assumption that victim models are trained based on the pretrained BERT model, Krishna et al. [33] show that an adversary can effectively extract language models whose performance is only slightly worse than that of the victim models. However, to the best of our knowledge, these model extraction attacks mainly focus on discriminative models. The attack against generative models, GANs in particular, is still an open question.

Comparing to model extraction attacks on discriminative models, we observe that there exist some differences for generative models. First, adversaries can leverage output information from target models such as labels, probabilities and logits, to mount model extraction attacks on discriminative models [26, 41, 48, 63], while generative models do not provide such information but only return images. Second, for model extraction attacks on discriminative models, they are evaluated by a test dataset. In contrast, unsupervised generative models aiming to learn the distribution of training data are evaluated by quantitative measures such as Fréchet Inception Distance (FID) [22] and multi-scale structural similarity (MS-SSIM) [47], or qualitative measures such as preference judgment [25, 69]. Therefore, these differences indicate that model extraction strategies, evaluations and defenses on generative models are very different from these on discriminative models.

In this paper, we aim to systematically study the feasibility of model extraction attacks against GANs from the perspective of accuracy extraction and fidelity extraction. First, we define *accuracy* and *fidelity* of model extraction on GANs. More specifically, when an adversary mounts model extraction attacks against GANs, *accuracy* measures the difference of data distribution between the attack model and the target model, while *fidelity* ensures the distribution of the attack model is consistent with the distribution of the training set of the target model. In the next step, according to the adversary's goals and the background information that they can have access to (see Figure 2), we systematically study two different types of attacks on GANs: accuracy extraction attack and fidelity extraction attack, which are shown in Figure 1.

Accuracy extraction attack. Adversaries mounting accuracy extraction focus on *accuracy* and they aim to steal the distribution

of a target model. For this attack, we assume adversaries have no knowledge of the architecture of target models, and they either obtain a batch of generated data that the model owner has publicly released or query the target model to obtain generated data. It can be considered as a black-box accuracy attack. After obtaining the generated data, adversaries can train a copy of the target GAN model. We study two different target models: Progressive GAN (PGGAN) [30] and Spectral Normalization GAN (SNGAN) [45]. Extensive experimental evaluations show that accuracy extraction can achieve an acceptable performance with only about 50k queries (i.e., 50k generated samples). When we continue to increase the number of queries, we find that it cannot bring significant improvement of the *fidelity* of attack models. This is mainly because the discriminator of a target GAN model is often better than its corresponding generator and it is very hard to reach global optimum [3]. In other words, directly querying the target model enables the attack model to be more consistent with the target generator rather than the real data distribution of the target model (see Figure 5 for an example). Therefore, it motivates us to perform fidelity extraction to improve the *fidelity* of attack models.

Fidelity extraction attack. Adversaries mounting fidelity extraction concentrate on *fidelity* and they target at stealing the distribution of the training set of a target model. In order to mount a fidelity extraction attack, we assume that adversaries can obtain more background knowledge. We discuss two subcategories for fidelity extraction: partial black-box fidelity extraction and white-box fidelity extraction. For the former, we assume that adversaries are able to have access to partial real data from the target GAN models, in addition to querying the target model. Extensive experimental evaluations show that adding partial real data is an effective method in terms of improvement of *fidelity* of attack models. For the latter, we assume adversaries can obtain the discriminator from the target GAN model and partial real data, which is the most knowledgeable setting. Also, adversaries can query the target model. We utilize the discriminator to subsample generated samples. These refined samples are more close to real data distribution, compared to samples are directly generated by the target model (see Figure 5(e)). Then, we use these refined samples and partial real data to train our attack model. Extensive experimental evaluations also show improvement of the *fidelity* of attack models.

Case study. We perform one case study to further demonstrate the impact of model extraction attacks on a large-scale scenario. In this case study — model extraction based transfer learning, we show that stealing a state-of-the-art GAN model can enable adversaries to enhance the performance of their own GAN model by transfer learning. Compared with training from scratch on LSUN-Classroom dataset with 20.34 FID [30], model extraction based transfer learning achieves 16.47 FID, which is the state-of-the-art performance on LSUN-Classroom dataset.

Defense. Both accuracy extraction and fidelity extraction attacks on GANs compromise the intellectual property of model providers. In particular, fidelity extraction aiming to steal the distribution of the training set of a target model can further severely breach the privacy of the training set. Therefore, we propose possible defense techniques by considering two aspects: *accuracy* and *fidelity*. In terms of *accuracy* of model extraction, limiting the number of queries is an effective method. In terms of *fidelity* of model extraction, we believe that a high fidelity attack model requires adversaries to have access to generated data which can be representative for real data distribution. The performance of model extraction attacks will be attenuated if adversaries only obtain a partial or distorted distribution of generated data. Thus, we propose two types of perturbation-based defense strategies: input and output perturbation-based approaches, to reveal less distribution information by increasing the similarity of samples or lowering the quality of samples [2]. The input perturbation-based approaches include linear and semantic interpolation perturbation while the output perturbation-based approaches include random noise, adversarial example noise, filtering and compression perturbation. Extensive experimental evaluations show that, compared to queries from the prior distribution of the target model, the equal amount of queries by perturbation-based defenses can effectively degrade the *fidelity* of attack models.

Summary of contributions. Our contributions in the current work are threefold:

(1) we conduct the first systematic study of model extraction attacks against GANs and devise accuracy extraction attacks and fidelity extraction attacks for GANs;
(2) we preform one case study to illustrate the impact of model extraction attacks against GANs on a large-scale scenario;
(3) we propose new effective defense measures to mitigate model extraction attacks against GANs.

Organization. The rest of the paper is organized as following. The next section 2 reviews related work. Section 3 introduces the preliminary knowledge, and Section 4 taxonomizes the space of model extraction attacks on GANs. Section 5 and Section 6 introduce the accuracy extraction and fidelity extraction, respectively. Section 7 presents one case study. In Section 8, we discuss possible defense mechanisms. Section 9 concludes this paper.

2 RELATED WORK

Generative adversarial networks (GANs). GANs have achieved impressive performance in a variety of areas, such as image synthesis [6, 30–32, 37, 45, 52, 56], image-to-image translation [38, 51, 70], and texture generation [36, 66], since a framework of GAN was first proposed by Goodfellow et al. in 2014 [18]. For image synthesis tasks, the current state-of-the-art GANs [6, 30, 31, 45] are able to generate highly realistic and diverse images. For instance, SNGAN [45] generates realistic images by a spectral normalization method to stabilize the training process. PGGAN [30] proposed by Karras et al. is the first GAN that successfully generates real-like face images at a high resolution of 1024×1024, applying a progressive training strategy. Unlike the PGGAN training in an unsupervised method, BigGAN [6] proposed by Brock et al. aims to generate high-quality images from a multi-class dataset by conditional GANs which leverage information about class labels. Recently, StyleGAN [31] has further improved the performance of GANs on high-resolution images through adding neural style transfer [24]. In this paper, *we choose SNGAN and PGGAN as the target models to be attacked by model extraction, considering their impressive performance on image generation. StyleGAN is also used as a target model in a case study in Section 7.*

Model extraction attacks. With the availability of machine learning as a service (MLaaS), model extraction attack has received much attention from the research community [8, 13, 26, 33, 63], which aims to duplicate (i.e., 'steal') a machine learning model. This type of attack can be categorized into two classes: accuracy model extraction and fidelity model extraction. In terms of accuracy model extraction, it was first proposed by Tramèr et al. [63], where the objective of the attack is to gain similar or even better performance on the test dataset for the extracted model. Since then, various methods attempting to reduce the number of queries have been developed for further improving the attack efficiency, such as model extraction using active learning [11, 49] or semi-supervised learning [26]. In terms of fidelity model extraction, it requires the attack model to faithfully reproduce predictions of the target model, including the errors which occur in the target model. Typical works include model reconstruction from model explanation [44], functionally equivalent extraction [26] and cryptanalytic extraction [8]. In addition to model extraction attacks on images, there are several work about model extraction in natural language processing [33, 60]. Krishna et al. [60] mount model extraction attacks against BERT-based models and the performance of the extracted model is slightly worse than that of the target model. Overall, these studies mainly focus on discriminative models, such as regression and convolutional neural networks for classification, and recurrent neural networks for natural language processing. *Unlike the existing studies, our work aims to study model extraction attacks against GANs.*

In addition to model extract attacks, there are other types of attacks in relation to privacy and security [9, 21, 67], such as membership inference attacks [12, 19, 55, 58, 59] and property inference attacks [17]. Some efforts have been also made to investigate membership inference attacks against GANs, where queries to a GAN model can reveal information about the training dataset [12, 19, 23]. Overall, these studies mainly focus on privacy on the training dataset, *while model extraction attacks in our paper concentrate on machine learning model itself.*

Model extraction defenses. Defense for model extraction can be broadly classified into two categories: restricting the information returned by models [35, 63] and differentiating malicious adversaries from normal users [29]. Tramèr et al. propose a defense where the model should only return class labels instead of class probabilities [63]. Recently, a technique PRADA has proposed to guard machine learning models by detecting abnormal query patterns [29]. Watermarking ML models as a passive defense mechanism recently has been proposed to claim model's ownership [28]. However, these defense techniques are used to protect discriminative models where models return probabilities or labels. In this paper, *we focus on defense approaches safeguarding generative adversarial networks where models return images.*

3 PRELIMINARIES

In this section, we begin with the general structure of GANs. Then, we proceed with discussing model extraction attacks in a general machine learning setting. Finally, we describe datasets used in this paper.

3.1 Generative adversarial networks

GAN is a generative model where it adversarially learns the unknown true distribution p_r on the training data $\mathbf{X}$. As shown in Figure 2, a GAN generally consists of two components: a generator G and a discriminator D. G is responsible for generating fake data $x_g = G(z)$, where the latent code z is sampled from a prior distribution p_z, such as Gaussian distribution or uniform distribution, while D takes the role of a binary classifier which differentiates real-like samples x_g from real samples $x_r \in \mathbf{X}$ as accurately as possible. The seminal GAN [18] is trained through optimizing the following loss functions:

$$L_D = -\mathbb{E}_{x\sim p_r}\left[\log D(x)\right] - \mathbb{E}_{z\sim p_z}\left[1 - \log D(G(z))\right] \quad (1)$$

$$L_G = -\mathbb{E}_{z\sim p_z}\left[\log D(G(z))\right] \quad (2)$$

If D and G converge and reach global equilibrium, then $p_r(x) = p_g(x)$, where $p_g(x)$ is the generator's distribution. For a fixed G, the optimal discriminator D^* can be obtained by:

$$D^*(x) = \frac{p_r(x)}{p_r(x) + p_g(x)} \quad (3)$$

In the course of employment, only G is utilized to produce new synthetic data while D is usually discarded.

3.2 Model extraction attacks against machine learning models

A machine learning model is essentially a function f that maps input data $\mathcal{X}$ to output data $\mathcal{Y}$: $\mathcal{Y} = f(\mathcal{X})$. In general, machine learning models can be categorized as two classes [4]: discriminative models and generative models. For discriminative models on image classification tasks, the input data corresponds to an image while the output data can be interpreted as a probability distribution over categorical labels. *A key goal of discriminative models* is to find an optimal set of parameters which minimizes the errors on the test dataset. For generative models on image generation tasks, the input data is represented by a latent code and the output data is an image. *A core goal of generative models* is to adjust the parameters to learn a distribution which is similar to the training data distribution p_r.

A model extraction attack in the machine learning setting emerges when an adversary aims to duplicate a model $\tilde{f}$ through querying the target model f. In general, there are two types of attacks around model extraction based on adversary's objective: accuracy extraction and fidelity extraction [26]. For discriminative models, accuracy extraction requires the extracted model to match or exceed the accuracy of the target model on the test dataset, while fidelity extraction requires the extracted model not only to achieve the same accuracy as the target model on the test dataset but also to replicate the errors of the target model. The limit of fidelity extraction is the functionally-equivalent model extraction [26]. Considering different goals and evaluations between discriminative models and generative models, we redefine model extraction on GANs in Section 4.1.

3.3 Dataset description

We utilize five different datasets in this paper, which are all widely adopted in image generation. Among them, four datasets are from the LSUN dataset [68] which includes 10 scene categories and 20

Table 1: Dataset description

Dataset	LSUN-Bedroom	LSUN-Kitchen	CelebA
Size of dataset	3,033,042	2,212,277	202,599
Dataset	LSUN-Classroom	LSUN-Church	
Size of dataset	168,103	126,277	

Table 2: Notations

Notation	Description
p_r	distribution of training set of a GAN
p_g	implicit distribution of a target generator
$\tilde{p}_g$	implicit distribution of an attack generator
accuracy	FID ($\tilde{p}_g$, p_g)
fidelity	FID ($\tilde{p}_g$, p_r)

object categories and we define them as LSUN-Bedroom, LSUN-Church, LSUN-Classroom, and LSUN-Kitchen, respectively. CelebA dataset [40] consists of about 200k high-quality human face images. Datasets including LSUN-Bedroom, LSUN-Classroom, and LSUN-Kitchen are only used in Section 7 to illustrate the attack effects in a case study. The details of the datasets are shown in Table 1.

4 TAXONOMY OF MODEL EXTRACTION AGAINST GANS

In this section, we start with adversary's goal and formally elaborate on our attacks. Next, we illustrate adversary's background knowledge where an adversary can mount attacks according to the obtained information. Finally, we detail the metrics to evaluate the performance of attack models.

4.1 Adversary's goals

In general, model extraction based on adversary's goals can be categorized into either accuracy extraction or fidelity extraction. Unlike supervised discriminative models aiming at minimizing errors on a test set, unsupervised generative models target at learning the distribution of a data set. Therefore, for model extraction attacks on GANs, accuracy extraction aims to minimize the difference of data distribution between attack models and target models, while fidelity extraction aims to minimize the distribution between attack models and the training set of target models.

Specifically, as shown in Figure 1, the goal of accuracy extraction is to construct a $\tilde{G}$ minimizing $S(\tilde{p}_g, p_g)$, where S is a similarity function, $\tilde{p}_g$ is the implicit distribution of the attack generator $\tilde{G}$, and p_g is the implicit distribution of the target generator G. In contrast, fidelity extraction's goal is to construct a $\tilde{G}$ minimizing $S(\tilde{p}_g, p_r)$, where p_r is the distribution of the training set of the target generator G. In this work, we use Fréchet Inception Distance (FID) to evaluate the similarity between two data distributions, mainly considering its computational efficiency and robustness [22]. It is elaborated in Section 4.3. In our work, we study the accuracy extraction in Section 5, and fidelity extraction in Section 6.

4.2 Adversary's background knowledge

Adversaries can mount model extraction attacks at different levels based on their obtained information about the target GAN. The more background knowledge adversaries acquire, the more effective they should be in achieving their goal. In general, four components of a GAN can be considered by an adversary. As shown in Figure 2, they are respectively: (1) generated data; (2) latent codes used by interactively querying a generator; (3) partial real data from the training dataset of the target GAN; (4) a discriminator from the target GAN.

- *Generated data:* generated data refers to samples from the generator of the target GAN. In some scenarios, adversaries are unable to query the target model but only have access to the generated samples which are publicly released by the model provider. Whether they can steal a good-quality model largely depends on the amount of released data from model owners.
- *Latent codes:* latent codes refer to random numbers drawn from a prior distribution such as Gaussian distribution. Adversaries obtaining latent codes mean that they know the prior distribution and use it to interactively query the target model's API provided by model owners. This is a basic assumption in the current query-based attack setting.
- *Partial real data:* this is more knowledgeable for adversaries when they can get partial real data which is used to train the target GAN. But this is also a common assumption which can be found in many literature in relation to security and privacy of machine learning [19, 27, 59].
- *Discriminator:* adversaries may obtain the discriminator of the target GAN although it is generally removed in the deployment. In our work, we illustrate how the discriminator can be explored to improve the *fidelity* of attack models.

In the following attack settings, we assume an adversary obtains different levels of background knowledge to achieve accuracy extraction or fidelity extraction.

4.3 Metrics

In this work, we evaluate the performance of GANs by the FID [22] which measures the similarity between p_g and p_r. Specifically, on the basis of features extracted by the pretrained Inception network ϕ, it models $\phi(p_r)$ and $\phi(p_g)$ using Gaussian distribution with mean μ and covariance Σ, and the value of FID between real data p_r and generated data p_g in convolutional features is computed as: $FID(p_r, p_g) = \left\|\mu_r - \mu_g\right\| + Tr(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2})$. A lower FID indicates that the distribution's discrepancy between the generated data and real-world data is smaller and the generated data is more realistic. In our work, FID is computed with all real samples and 50k generated samples.

In addition to FID which evaluates the performance of a model itself, we also need indicators to measure whether an attack is considered to be successful. In this work, we use two FID-based metrics: *accuracy* and *fidelity*, to evaluate the performance of attack models. *Accuracy* measures the consistency between p_g which is an implicit distribution of a target generator and $\tilde{p}_g$ which is an implicit distribution of an attack generator. Note that, *accuracy* not only measures how close the attack model and the target model

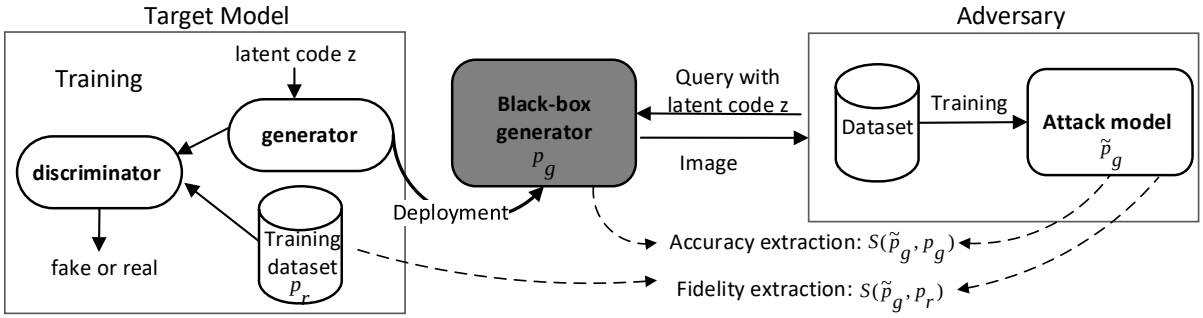


Figure 1: Attack types: accuracy extraction and fidelity extraction.

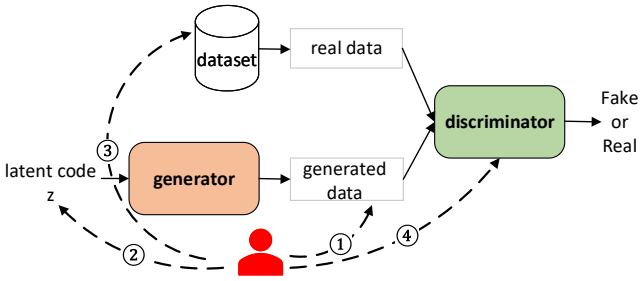


Figure 2: Adversary's background knowledge.

are, but also indicates how well the performance of model itself is. In contrast, *fidelity* measures the consistency of data distribution between p_r and $\tilde{p}_g$. Similar to FID, the smaller the *accuracy* and *fidelity* values are, the better performance attack models achieve. When it is clear from the context, we refer to *accuracy* and *fidelity* as *accuracy* value and *fidelity* value, respectively. The summarized notations can be seen in Table 2.

Accuracy extraction focuses on *accuracy* and adversaries aim to steal the distribution of a target model. After obtaining an attack model which steals from a target model, they can directly utilize it to generate new samples. Additionally, they can also transfer knowledge of the stolen model to their own domains through transfer learning. In contrast, fidelity extraction concentrates on *fidelity* and adversaries target at stealing the distribution of the training set of a target model. This type of attacks can severely violate the privacy of the training data and it also means that adversaries may steal valuable commercial datasets from a trained GAN. Additionally, adversaries can utilize the stolen high-fidelity model to mount other novel attacks and we leave it for future work.

5 ACCURACY EXTRACTION

In this section, we instantiate our accuracy extraction attack strategy. we assume that adversaries have access to either generated samples provided by the model producer or querying the target model to obtain data. Therefore, whether only publicly access to data generated by GANs or black-box querying can leak the information of models is a question. In this section, we start with target models and attack models. Then, we describe our attack performance. Next, we study the effect of the number of queries and queries from different prior distributions. In the end, we perform experiments to deeply understand model extraction on GANs.

Table 3: Performance of target GANs (see Figure 17 in Appendix for qualitative results).

Target model	Dataset	FID
SNGAN	LSUN-Church	12.72
SNGAN	CelebA	7.60
PGGAN	LSUN-Church	5.88
PGGAN	CelebA	3.40

5.1 Target models and attack models

We choose representative GANs: Progressive GAN (PGGAN) [30] and Spectral Normalization GAN (SNGAN) [45] as our target models, which both show pleasing performances in image generation. The implementation details can be seen in Appendix A.1. For training sets LSUN-Church and CelebA, we first resize them to 64×64 and use all records of each dataset to train our target models. As shown in Table 3, target GAN models achieve an excellent performance on these dataset and the performance of PGGAN is better than that of SNGAN.

We use GANs as our attack models to extract target models. In practice, adversaries may not know the target model's architecture. Therefore, we study the performance of attack models with different architectures. Specifically, we choose SNGAN and PGGAN as our attack models. There are four different situations for their combinations. For simplification, we define each situation as an attack-target model pair, and they are respectively SNGAN-SNGAN, SNGAN-PGGAN, PGGAN-SNGAN and PGGAN-PGGAN. The reason why we choose SNGAN and PGGAN as the research object is that: 1) they both show good performance in image generation; and 2) they have significant difference in the aspects of training, loss function and normalization, which all facilitate us to study the performance of attack models with different architectures.

5.2 Methodology

As shown in Figure 1, for accuracy extraction, we assume that an adversary obtains the generated data by the model provider or querying the target GAN. This scenario is practical, because some model owners need to protect their models through providing the public with some generated data or a black-box GAN model API. In this case, the adversary uses the generated data to retrain a GAN to extract the target model. We do not distinguish whether generated data is from queries or model providers, because our approach only

relies on these generated data. However, in Section 5.3.3, we study attack performance on queries with different prior distributions.

Note that model extraction on GANs is different from machine learning on GANs. This is because machine learning on GANs requires users to train a GAN on real samples which are collected from the real world. In contrast, model extraction on GANs enables users to train a GAN on generated data from a target GAN model. In essence, model extraction on GANs approximates the target GAN which is a much simpler deterministic function, compared to real samples which usually represents a more complicated function.

5.3 Results

5.3.1 ***Attack performance on different models.*** Table 4 shows the accuracy extraction's performance with 50k queries to the target model. In general, attack models can achieve an acceptable performance. For instance, our attack performance of PGGAN-PGGAN on the CelebA achieves 1.02 FID on *accuracy*, which means that the attack model can achieve a perfect extraction attack for the target model.[1] It is noticeable that the the attack model achieves such performance only on 50k generated images while the target model is trained on more than 200k images, which indicates that the privacy of the target model can be leaked by their generated images. In other words, adversaries are able to obtain a good GAN model only by access to the generated data from the target model instead of collecting their own data which is usually labor-intensive and time-consuming.

For the target model PGGAN, if the attack model is SNGAN, we observe that the performance of model extraction is very efficient on both CelebA and LSUN-Church dataset and the attack model SNGAN can learn more from the target model PGGAN, compared to the SNGAN-SNGAN case, which indicates that attacking a state-of-the-art GAN is valuable and viable for an adversary. Furthermore, this case SNGAN-PGGAN is the most common situation in the actual attack scenarios, because generally we implicitly assume that performance of the adversary's model may often be weaker than that of the target model and the structure of the attack model is inconsistent with that of the target model.

We also report *fidelity* in Table 4 and find that for model extraction on GAN models, the *fidelity* of attack models is always higher than that of target model, in which *fidelity* of attack models represents similarity between distribution of real dataset p_r and distribution of the attack model $\tilde{p}_g$ and for *fidelity* of a target model, also called FID of target model, it represents similarity between distribution of real dataset p_r and distribution of the target model p_g. For example, when the target model SNGAN has 12.72 FID on the LSUN-Church dataset, *fidelity* of the attack model SNGAN will increase to 30.04. Even for the PGGAN-PGGAN case, its *fidelity* increases from 3.40 to 4.93 on the CelebA dataset. This is mainly because although theoretically, the distribution of the target model p_g is equal to that of the real training dataset p_r, it is actually not equal because GAN cannot achieve the global optimum. However, we will discuss how to reduce *fidelity* and achieve fidelity extraction in Section 6.

[1] Although we cannot directly compare the performance of different models on different datasets, we can choose the state-of-the-art StyleGAN [31] trained on the LSUN-Bedroom dataset as a reference, where it has the lowest FID 2.65.

Table 4: The performance of accuracy extraction with 50k queries to the target model (see Figure 18 in Appendix for qualitative results).

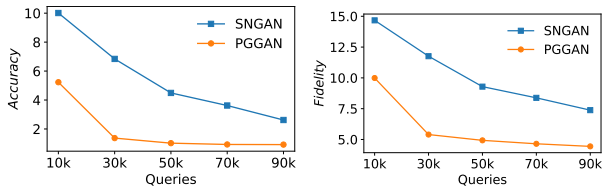
Target model	Attack model	Dataset	*Accuracy* FID($\tilde{p}_g$, p_g)	*Fidelity* FID ($\tilde{p}_g$, p_r)
PGGAN	SNGAN	LSUN-Church	6.11	14.05
	SNGAN	CelebA	4.49	9.29
	PGGAN	LSUN-Church	1.68	8.28
	PGGAN	CelebA	1.02	4.93
SNGAN	SNGAN	LSUN-Church	8.76	30.04
	SNGAN	CelebA	5.34	17.32
	PGGAN	LSUN-Church	2.21	14.56
	PGGAN	CelebA	1.39	9.57

For the target model SNGAN, if the attack model is PGGAN, the *accuracy* of model extraction is lower than that of the attack model SNGAN. It is mainly because the PGGAN model itself is stronger and able to more accurately approximate the target model. Similarly, PGGAN as an attack model has more lower *fidelity*, in contrast with SNGAN as an attack model. For instance, compared to SNGAN-SNGAN with 17.32 of *fidelity* on CelebA dataset, the *fidelity* of PGGAN-SNGAN is only 9.57, which largely improves the attack performance on fidelity. This indicates that using an attack model which is larger than the target model is an efficient approach to improve attack performance.

Overall, accuracy extraction can achieve a good performance in terms of *accuracy*. In general, adversaries can steal an accurate model, and then use the extracted model for their own purpose. However, unlike discriminative models where adversaries can directly utilize their extracted model, the extracted model of a GAN only generates target model's images. Therefore, in Section 7, we will perform a case study where adversaries can effectively leverage the extracted model to generate images for their own applications rather than target GANs' images through transfer learning.

5.3.2 ***Attack performance on the number of queries.*** We choose PGGAN trained on CelebA dataset as the target model to study the effect of the number of queries due to the best performance among our target models. Figure 3 plots the attack performance with respect to the number of queries which are also the size of training dataset of attack models. As expected, we observe that the attack performance increases with an increase in queries. For instance, when the number of queries increases to 90k, attack models' *fidelity* and *accuracy* are both low. However, as the number of queries decreases, the attacks become less effective due to the lack of sufficient training data for the attack models. Especially for the attack model SNGAN with 10k queries, its *fidelity* and *accuracy* are both over 10.00. This indicates that releasing a small number of data by the model owner or restricting the number of queries is a relatively safe measure, because it is difficult for an adversary to extract a model with limited data.

5.3.3 ***Attack performance on queries from different prior distributions.*** Adversaries can query the target model via trying common prior distributions to generate latent codes if they do not know the prior distribution of a target model. Gaussian distribution



(a) *Accuracy* on CelebA (b) *Fidelity* on CelebA

Figure 3: Attack performance on the number of queries.

Table 5: Performance of accuracy extraction attack with different prior distributions. We use standard normal distribution and uniform distribution over an interval -1 and 1 to generate latent codes. The number of queries is fixed to 50k.

Attack model	Prior distribution	*Accuracy* FID($\tilde{p}_g, p_g$)	*Fidelity* FID ($\tilde{p}_g, p_r$)
SNGAN	Gaussian	4.49	9.29
SNGAN	Uniform	4.29	9.16
PGGAN	Gaussian	1.02	4.93
PGGAN	Uniform	0.98	4.85

and uniform distribution are widely used in almost all GANs [6, 30–32, 45, 52]. Table 5 shows the attack performance with two prior distributions. We choose PGGAN trained on CelebA dataset with standard normal prior distribution as the target model. From Table 5, we find that adversaries can obtain a similar attack performance no matter what the prior distribution of latent codes is.

5.3.4 ***Understanding accuracy extraction on GANs in-depth.*** We further dissect the difference of distributions between target models and attack models to understand the nature of model extraction on GANs. Specifically, we first transform the training data into 2048-dimension feature vectors by the pretrained Inception-v3 model which is widely utilized in the evaluation of a GAN model. Then these feature vectors are clustered into k classes by a standard K-means algorithm. Finally, we calculate the proportions of each class and proportions of all classes can be considered as a distribution of the training data [5, 53]. The blue bar in Figure 4 shows the distribution of the training data where we set k to 30. For target models and attack models, we query the model to obtain 50k images, then perform the same procedures as the training data.

Figure 4 shows distribution differences among the training data, the target model PGGAN and attack models. We observe that for the high proportions of classes, which can be considered as prominent features of a distribution, target models can learn more features about these classes while attack models further learn more features by querying the target models. In contrast, for the low proportions of classes, target models learn less features about these classes while attack models further learn less features about these classes. This is one reason why attack models always have higher *fidelity* than target models. In terms of *accuracy*, we observe that there is a consistent trend on proportions of classes for target models and attacks models. This is the reason why we can achieve a good performance about *accuracy*. We also analyze the target model SNGAN, and similar results are shown in Figure 11 in Appendix.

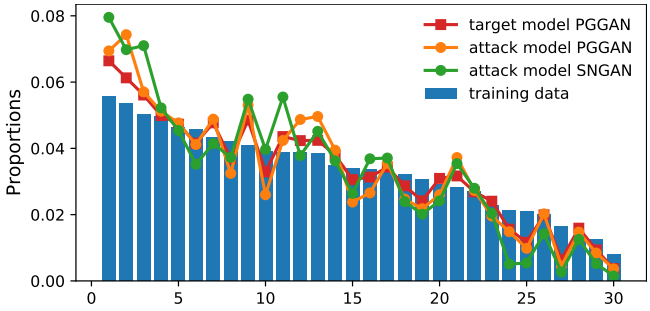


Figure 4: Class distributions of the training data, the target model PGGAN, and attack models.

Table 6: JS distances between models. A smaller value indicates a better performance. For the JS distance between training data and the target model, the target model PGGAN is 4.14×10^{-3}. The JS value shows a consistent trend with Figure 4.

Target model	Attack model	$\mathcal{JS}_{accuracy}$ ($\times 10^{-3}$)	$\mathcal{JS}_{fidelity}$ ($\times 10^{-3}$)
PGGAN	SNGAN	5.88	15.95
	PGGAN	1.83	9.10

We also summarize this difference in a single number by computing the Jensen-Shannon (JS) divergence on this representation of distributions, which is shown in Table 6. Note that, based on *accuracy* and *fidelity* defined in Section 4.1, we mark $JS_{accuracy}$ as the JS divergence between the target model and the attack model, and $JS_{fidelity}$ as the JS divergence between the training data and the attack model.

6 FIDELITY EXTRACTION

In this section, we instantiate our fidelity extraction attack strategy. In addition to accuracy extraction's assumptions, we also assume that adversaries have more background knowledge in order to achieve fidelity extraction, such as partial real data or target model's discriminator. We start with the motivation and problem formulation of fidelity extraction. Then, we describe the methodology of fidelity extraction. In the end, we present the performance of fidelity extraction.

6.1 Motivation and problem formulation

As shown in Figure 1, accuracy extraction can be implemented through querying the generator of the target GAN, because p_g is the generator's distribution. As for fidelity extraction, it is much more difficult due to the lack of availability of real data distribution p_r. Although an approach is to use p_g as an approximation of p_r, we observe that with the increase in the number of queries, *fidelity* of attack models reaches its saturation point and is hard to be improved, which is shown in Figure 3(b). For instance, as we increase the number of queries from 50k to 90k for the PGGAN-PGGAN case on CelebA dataset, *fidelity* of the attack model has smaller and smaller improvements from 4.93 to 4.44, while the ideal *fidelity* is 3.40 which is also the performance of the target model. Note that the case PGGAN-PGGAN is the best for the attacker; the

attack will perform even worse if the attackers do not choose the same architectures and hyperparameters as the target model.

The reason why there exists a gap between the attack model and the target model in terms of *fidelity* is that the target GAN model is hard to reach global equilibrium and the discriminator is often better than the generator in practice [3]. As a result, real data distribution p_r is not completely learned by the generator of the target model, which means that $p_g \neq p_r$. Therefore, directly using the generator's distribution p_g does not guarantee the high fidelity and it only minimizes the distribution discrepancy between the attack model and the target model. We explain this by a simple example on Figure 5, which is popular in the GAN literature [3, 15, 64].

Figure 5(a) presents real samples drawn from a mixture of 25 two-dimensional Gaussian distributions (each with standard deviation σ of 0.05). Figure 5(b) - Figure 5(d) show samples which are generated by a target GAN with different queries. We define a generated sample as "high-quality" if its Euclidean distance to its corresponding mixture component is within four standard deviations ($4\sigma = 0.2$) [3]. The architecture and setup information of the target GAN is shown in Appendix A.1. Overall, we can observe that target GAN's distribution is not completely the same as the training set's distribution, which means that directly extracting a model from the generator of the target GAN makes its distribution similar to the target model's distribution rather than its training dataset's distribution.

In order to achieve fidelity extraction, we believe that adding additional background information is necessary. Specifically, we consider two scenarios where adversaries can query the target model and also have limited auxiliary knowledge of:

(1) partial real samples from training dataset; or
(2) partial real samples and the discriminator of the target model.

Partial real data can provide some information about distribution of the real data set, which can correct the fidelity extraction and make it closer to the real distribution. The discriminator from the target model can reveal the distribution information of the training data [3]. Thus, using the information provided by discriminator, we can subsample the generated data to make the obtained data closer to the real dataset's distribution, which benefits fidelity extraction. In Figure 5(e), we can observe that samples which generated by the MH subsampling algorithm [64] are much closer to the true distribution than those obtained by direct sampling. Specifically, it improves the percentage of "high-quality" samples from 94.15% to 95.64%.

In fidelity extraction, we choose the same attack models and target models as accuracy extraction (see Section 5.1). In the following section, we call the first scenario of fidelity extraction as partial black-box fidelity extraction and the second scenario as white-box fidelity extraction.

6.2 Methodology

For partial black-box fidelity extraction, we first query the target model to obtain generated samples. And then, we train an attack model on these generated samples and continue training after adding partial real data. In this scenario, we assume that adversaries query the target model 50k times to obtain 50k generated samples.

For white-box fidelity extraction, we first leverage the discriminator of the target model to subsample the generated samples. As a result, these refined samples are much closer to the true distribution. In this work, we use Metropolis-Hastings (MH) subsampling algorithm [64] to subsample the generated data. See Algorithm 1 in Appendix for details. MH subsampling algorithm utilizes the discriminator through Metropolis-Hastings algorithm [61] to refine samples which are generated by the generator. The discriminator generally needs to be calibrated by partial real samples from training set of the target GAN model, considering that some discriminators of GANs output a score rather than a probability. In our experiments, all discriminators are calibrated through logistic regression. Then we train the attack model on those refined samples. After the training process of the attack model is stable, we add partial real data to further train the attack model.

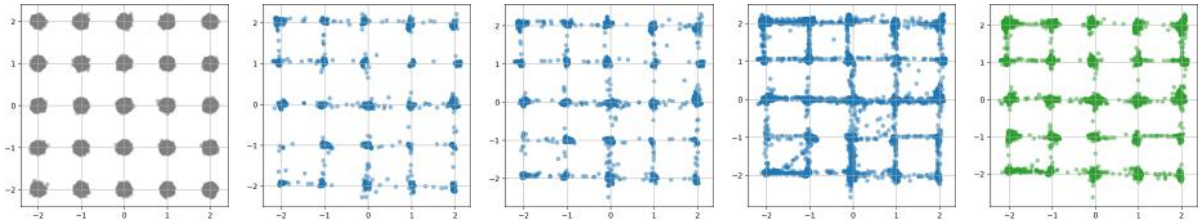
In this scenario, although the number of queries will increase due to subsampling samples, we assume that adversaries eventually obtain 50k refined samples. Partial real samples used to calibrate the discriminator are fixed to 10% knowledge of training data. In addition, these partial real samples will be added into training process of the attack models.

It is worth noting that we cannot directly choose the lowest *fidelity* value in real attack scenarios due to unavailability of training dataset from target models. Therefore, the *fidelity* value reported in this paper is chosen when its corresponding *accuracy* value is the lowest in the training process.

6.3 Results

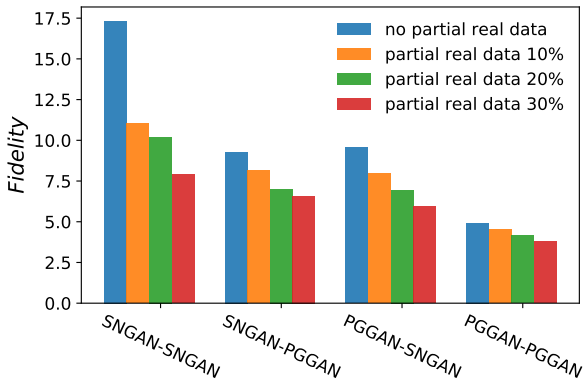
6.3.1 ***Partial black-box fidelity extraction.*** Figure 6(a) plots the results of the partial black-box fidelity extraction attack against target GAN models trained on CelebA dataset. We observe that directly adding the real data to a training dataset is an effective approach to improve *fidelity* of model extraction attacks for four different attack-target cases. For instance, compared to the SNGAN-SNGAN trained on CelebA with no partial real data, *fidelity* values decrease from 17.32 to 11.04 when adversaries have only 10% knowledge of the training data from the target model. If adversaries gain 30% knowledge of real data, *fidelity* further decreases to 7.89. That reminds model providers that training dataset should be kept secretly. In Figure 6(b), we report the results of the partial black-box fidelity extraction attack against target GAN models trained on LSUN-Church dataset. Similar to Figure 6(a), *fidelity* of attack models decreases with an increase with partial real data. Note that the significant decrease in *fidelity* of attack models can be observed between no partial real data and partial real data 10% for all cases on both datasets, while the decrease in *fidelity* of attack models among partial real data 10%, 20%, and 30% gradually becomes flat.

It is worth noting that increasing the number of queries can also improve *fidelity*, as shown in Figure 3(b). However, we observe that it will gradually saturate while adding partial real data can break the limit. Table 7 compares the *fidelity* between two methods: increasing the number of queries and adding partial real data. We fix PGGAN trained on CelebA dataset as our target model and its FID is 3.40 which is also the ideal *fidelity* for attack models. For CelebA dataset, 10% of the training set corresponds to 20,259 out of 202,599 images. Therefore, the size of training set of black-box

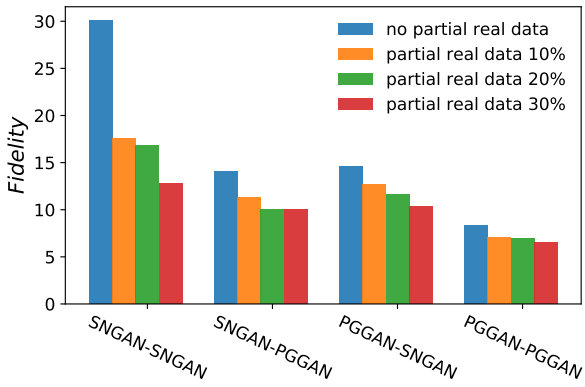


(a) Real Samples 50k (b) Direct Sampling 5k (c) Direct Sampling 10k (d) Direct Sampling 50k (e) MH Subsampling 50k

Figure 5: Difference of distribution between training data and generators. The percentage of "high-quality" samples for Figure 5(b), Figure 5(c), Figure 5(d) and Figure 5(e) is 94.36%, 94.31%, 94.15% and 95.64%, respectively. The more we query, the more bad-quality samples we obtain, which affects the performance of model extraction. But if we reduce the number of queries, the performance of attack models still be poor due to insufficient training samples.



(a) Partial black-box fidelity extraction on CelebA



(b) Partial black-box fidelity extraction on LSUN-Church

Figure 6: Partial black-box fidelity extraction.

accuracy extraction-70k is roughly equal to that of partial black-box fidelity extraction 10%. We can observe that adding equal amount of real data is more beneficial to *fidelity*, compared to increasing the number of queries. Furthermore, directly adding partial real data can break the limit of accuracy extraction. For instance, for the attack model PGGAN, obtaining 30% of knowledge of real dataset can achieve 3.78 of *fidelity*, which is much closer to the ideal fidelity of 3.40, whereas increasing the number of queries to 90k still remains 4.44 of *fidelity*. Similar results can also be seen in attack model

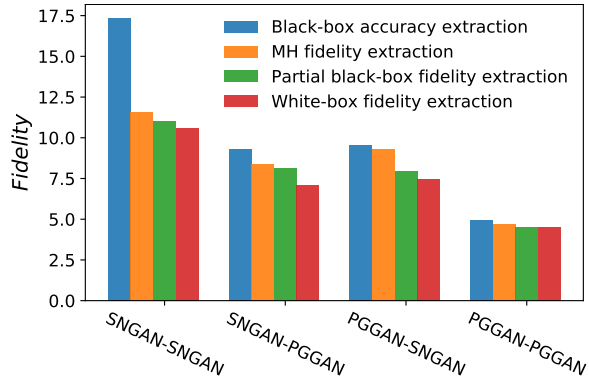
Table 7: Comparison between two methods: increasing the number of queries and adding partial real data.

Attack model	Method	*Fidelity* FID $(\tilde{p}_g, p_r)$
PGGAN	Black-box accuracy extraction-50k	4.93
	Black-box accuracy extraction-70k	4.65
	Black-box accuracy extraction-90k	4.44
	Partial black-box fidelity extraction 10%	4.54
	Partial black-box fidelity extraction 20%	4.15
	Partial black-box fidelity extraction 30%	3.78
SNGAN	Black-box accuracy extraction-50k	9.29
	Black-box accuracy extraction-70k	8.38
	Black-box accuracy extraction-90k	7.38
	Partial black-box fidelity extraction 10%	8.16
	Partial black-box fidelity extraction 20%	7.02
	Partial black-box fidelity extraction 30%	6.59

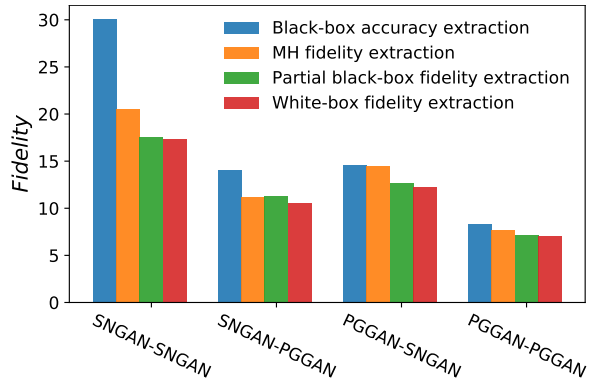
SNGAN shown in Table 7, although *fidelity* of the attack model SNGAN is inferior to that of the attack model PGGAN.

6.3.2 ***White-box fidelity extraction.*** As discussed in Section 6.2, we also consider white-box fidelity extraction where adversaries obtain both partial real data and the discriminator of target models. Leveraging the discriminator of target models to subsample needs some real data from training dataset. These real data, on the one hand, is utilize to calibrate the discriminator which is illustrated in Algorithm 1, on the other hand, it can be added the training process of the attack models to further improve the *fidelity* of attack models. We refer the former where only refined samples are used to train the attack model to MH fidelity extraction which is also considered as an indicator to show how well these refined samples are beneficial to *fidelity*. We refer the latter where both refined samples and partial real data are used to train the attack model as white-box fidelity extraction.

Figure 7 plots not only the results of MH fidelity extraction and white-box fidelity extraction on both CelebA and LSUN-Church datasets, but also black-box accuracy extraction and partial black-box fidelity extraction for comparison. We can observe that MH subsampling is an effective approach to improve *fidelity* of attack models. For example, when target model is SNGAN, MH fidelity



(a) *Fidelity* on CelebA



(b) *Fidelity* on LSUN-Church

Figure 7: Comparison on *fidelity*.

extraction can significantly improve attack model's fidelity on both datasets because MH subsampling algorithm selects high-quality samples from generated samples of the target model SNGAN. For partial black-box fidelity extraction and white-box fidelity extraction which both leverage the partial real data in the training process, white-box fidelity extraction generally can achieve higher fidelity than partial black-box fidelity extraction. We also analyze distribution differences for fidelity extraction, which is shown in Figure 12 in Appendix.

7 CASE STUDY: MODEL EXTRACTION BASED TRANSFER LEARNING

In this section, we present one case study where the extracted model serves as a pretrained model and adversaries transfer knowledge of the extracted model to new domains by means of fine-tuning to broaden the scope of applications based on extracted models. We start with methods of transfer learning on GAN and demonstrate how adversaries can benefit from model extraction, in addition to directly leveraging the extracted model to generate images.

We consider the state-of-the-art GAN model StyleGAN [31] that was trained on more than 3 million bedroom images as the target model. StyleGAN produces high-quality images at a resolution of 256×256, with 2.65 FID on LSUN-Bedroom dataset [68]. We suppose adversaries only query the target model StyleGAN and have no any other background knowledge, which is also called black-box accuracy extraction in our paper. Although an adversary can obtain an extracted model, the model only generates images which are similar to the target model. In this case, the extracted model can only generate bedroom images due to target model trained on LSUN-Bedroom dataset. Therefore, the adversary's goal is to use the PGGAN as the attack model to extract the target model StyleGAN and leverage transfer learning to obtain a more powerful GAN which generates images that the adversary wishes. *The attack is successful if the performance of models training by transfer learning based on the extracted GAN outperforms models training from scratch.*

Transferring knowledge of models which steal the state-of-the-art models to new domains where adversaries wish the GAN model can generate other types of images can bring at least two benefits: 1) if adversaries have too few images for training, they can easily obtain a better GAN model on limited dataset through transfer learning; 2) even if adversaries have sufficient training data, they can still obtain a better GAN model through transfer learning, compared with a GAN model training from scratch. Therefore, we consider two variants of this attack: one where the adversary owns a small target dataset (i.e., about 50k images in our work) and the other one where the adversary has enough images (i.e., about 168k images in our work).

More specifically, after querying the target model StyleGAN and obtaining 50k generated images, adversaries train their attack model PGGAN on the obtained data, as illustrated in Section 5.2. *Accuracy* of the attack model PGGAN is 4.12 and its *fidelity* is 6.97. Then, we use the extracted model's weights as an initialization to train a model on adversary's own dataset which is also called target dataset in the section. We conduct the following two experiments:

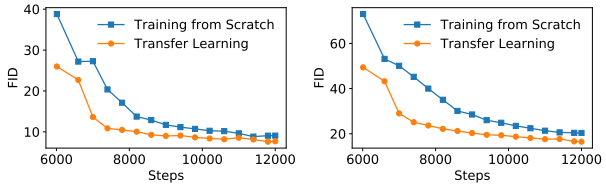
(1) We first randomly select 50k images from LSUN-Kitchen dataset as a limited dataset. Then, we train the model on these selected data by transfer learning and from scratch, respectively.
(2) We train a model on the LSUN-Classroom dataset including about 168k images by transfer learning and from scratch, respectively.

Results. Table 8 shows the performance of models trained by transfer learning and training from scratch. We can observe that the performance of training by transfer learning is always better than that of training from scratch on both large and small target dataset. To be specific, on the limited LSUN-Kitchen dataset which contains 50k images, the FID of model trained by transfer learning decreases from 8.83 to 7.59, compared with the model trained from scratch. It indicates that the extracted model is useful for models trained on other types of images. On the large LSUN-Classroom dataset which contains more than 168k classroom images, the performance of model significantly improves from model training from scratch with 20.34 FID[2] to training by transfer learning with 16.47 FID. This is also the best performance for PGGAN on LSUN-Classroom dataset, in contrast with 20.36 FID reported by Karras et al. [30]. We also plot the process of training for both settings on the two datasets, which is shown in Figure 8. We can obviously and consistently observe that training by transfer learning is always better than training from scratch during the training process, which

[2]This value is not equal to 20.36 [30] due to randomness.

Table 8: Comparison between transfer learning and training from scratch. The target model is StyleGAN trained on LSUN-Bedroom dataset, and the attack model is PGGAN.

Target dataset	Methods	FID
LSUN-Kitchen	Transfer Learning	7.59
LSUN-Kitchen	Training from Scratch	8.83
LSUN-Classroom	Transfer Learning	16.47
LSUN-Classroom	Training from Scratch	20.34



(a) FID on LSUN-Kitchen (b) FID on LSUN-Classroom

Figure 8: Comparison between transfer learning and scratch on LSUN-Kitchen and LSUN-Classroom dataset.

indicates that the extracted model PGGAN which duplicates the state-of-the-art StyleGAN on LSUN-Bedroom dataset can play a significant role in other applications rather than only on generating bedroom images. That reminds us that model extraction on GANs severely violates intellectual property of the model owners.

8 DEFENSES

Model extraction attacks on GANs leverage generated samples from a target GAN model to retrain a substitutional GAN which has similar functions to the target GAN. In this section, we introduce defense techniques to mitigate model extraction attacks against GANs.

According to adversary's goals as defined in Section 4.1, we discuss defense measures from two aspects: *accuracy* and *fidelity*. In terms of *accuracy* of model extraction, it is difficult for model owners to defend except for limiting the number of queries. This is because adversaries can always design an attack model to learn the distribution based on their obtained samples. The more generated samples adversaries obtain, the more effective they achieve.

In terms of *fidelity* of model extraction, its effectiveness is mainly because adversaries are able to obtain samples generated by latent codes draw from a prior distribution of the target model, and these samples generated through the prior distribution are representative for real data distribution [2]. However, if adversaries obtain some generated samples which are only representative for partial real data distribution or a distorted distribution, *fidelity* of attack models becomes poor. Based on this, we propose two types of perturbation-based defense mechanisms: input perturbation-base and output perturbation-based approaches. In the rest of this section, we focus on defense approaches which are designed to mitigate *fidelity* of attack models.

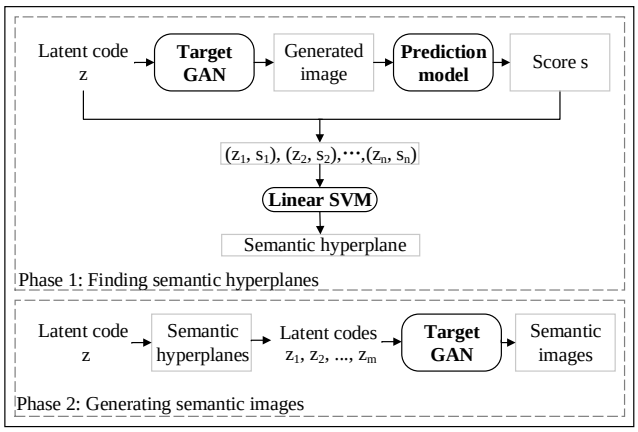


Figure 9: Semantic interpolation defense.

8.1 Methodology

8.1.1 Input perturbation-base defenses. For this type of defenses, we propose two approaches based on perturbing latent codes: linear interpolation defense and semantic interpolation defense.

Linear interpolation defense. For n latent codes queried from users, model providers randomly select two queried points and interpolate k points between the two points. This process is repeated for $\lceil n/k \rceil$ times to get n modified latent codes. These modified latent codes are used to query the target model. In our experiments, we interpolate 9 points. See Figure 13(a) in Appendix for visualization.

Semantic interpolation defense. Unlike linear interpolation defense where target models return a batch of random images, semantic interpolation defense enables users to obtain a batch of semantic images. Semantic information can be any information that humans can perceive, which is usually defined by model providers. For instance, for a human face image, semantic information includes gender, age and hair style. In this paper, we adopt a semantic interpolation algorithm proposed by Shen et al. [57].

Specifically, the process of semantic interpolation defense is shown in Figure 9. Semantic interpolation defense consists of two phases: finding semantic hyperplanes and generating semantic images. In the first phase, we first train a prediction model for each semantic information. Then the trained prediction model is used to predict semantic score s for each image generated through latent code z. As a result, we get latent code-score pairs and label the highest k scores as positive and the lowest k scores as negative. Finally, we train a linear support vector machine (SVM) on dataset where latent codes as training data and scores as labels. A trained linear SVM contains a hyperplane which separates one semantic information. In the second phase, we can obtain a semantic image for each semantic hyperplane through interpolation. A latent code interpolates points along the normal vector of the hyperplane and corresponding semantic images can be obtained. In our experiments, we totally explore 12 semantic information on CelebA dataset. See Figure 13(b) in Appendix for visualization.

In our experiments, we train each prediction model for each semantic information, and prediction model is built on the basis of ResNet-50 network [20] trained on ImageNet dataset [54].

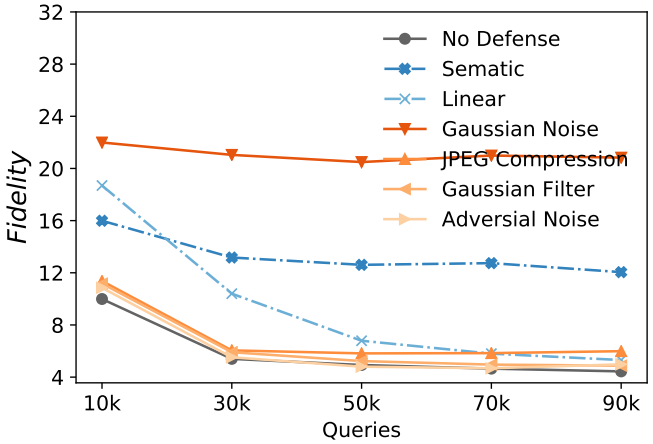


Figure 10: The performance of attack model PGGAN for defenses on black-box accuracy extraction.

8.1.2 Output perturbation-base defenses. Instead of perturbing latent codes, this type of defenses directly perturbs the generated samples. Specifically, we propose four approaches: random noise, adversarial noise, filtering and compression. See Figure 14 in Appendix for visualization.

Random noise. Adding random noises on generated samples is a straightforward method. In our experiments, we use Gaussian-distributed additive noises (mean = 0, variance = 0.001).

Adversarial noise. We generate adversarial examples through mounting targeted attacks where all images are misclassified into a particular class by the classifier ResNet-50 trained on ImageNet dataset. In our experiments, all face images are misclassified into the class — goldfish and the C&W algorithm [10] based on L_2 distance are used.

Filtering. The Gaussian filter is used to process generated samples. In our experiments, we use Gaussian filter (sigma = 0.4) provided by the skimgae package [65].

Compression. The JPEG compression algorithm is used to process generated samples. In our experiments, we use the JPEG compression (quality = 85) provided by the simplejpeg package [16].

8.2 Results

In this experiment, we choose PGGAN trained on CelebA dataset as the target model to evaluate our defense techniques, considering its excellent performance among our target models. We only show the effectiveness of defense techniques on black-box accuracy extraction, considering its more practical assumption: adversaries obtain samples by model providers or queries.

8.2.1 ***Defense on black-box accuracy extraction.*** Figure 10 plots results of attack model PGGAN on defenses. We observe that attack performance is weakened when the target model PGGAN uses these defense approaches, compared to the target model without any defense. Gaussian noise and semantic interpolation show stable performance while other defense techniques' performance is weakened with an increase in the number of queries. Figure 15 in Appendix also shows similar defense performance for the attack model SNGAN.

8.2.2 ***Discussion.*** The reason why input perturbation-based defenses can work is at least explained from two aspects: increasing the similarity of generated samples and a distribution mismatch between latent codes produced by interpolation and drawn from prior distribution. For the former, we can see that interpolation operations increase the similarity of images from Figure 13. For the latter, latent codes produced by interpolation operations are different from latent codes drawn from the prior distribution that the target model was trained on. This is because latent codes produced by linear operation do not obey the prior distribution of the target model, which also bring a benefit in disguising the true data distribution [2].

Output perturbation-based defenses can work because they directly perturb these generated samples. Model providers need to trade-off image quality and the model's security through magnitudes of changes. Although Gaussian noise defense shows the best performance, it is possible for adversaries to remove noise.

9 CONCLUSION

In this paper, we have systematically studied the problem of model extraction attacks on generative adversarial networks, and devised, implemented, and evaluated this attack from the perspective of accuracy extraction and fidelity extraction. For accuracy extraction, extensive experimental evaluations show that adversaries can achieve an acceptable performance with about 50k queries. For fidelity extraction, adversaries further improve the *fidelity* of attack models after obtaining additional background knowledge, such as partial real data from training set or the discriminator of the target model. We have also performed a case study where the attack model which steals a state-of-the-art target model can be transferred to new domains to broaden the scope of applications based on extracted models. Finally, we proposed two types of effective defense techniques based on perturbing latent codes or generated samples to mitigate model extraction attacks: input and output perturbation-based defense. Semantic interpolation and Gaussian noise defenses show a stable performance.

Training with differential privacy techniques can be utilized to protect privacy of training data of a model [1]. However, training time and stability of the training process are big challenges for GANs. For further work, we plan to design new methods based on differential privacy techniques to mitigate *fidelity* of model extraction. Additionally, studying watermarks to defend model extraction against GANs is also an interesting direction for future work.

ACKNOWLEDGMENTS

This work is supported by the National Research Fund, Luxembourg (Grant No. 13550291).

REFERENCES

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. 2016. Deep learning with differential privacy. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 308–318.

[2] Eirikur Agustsson, Alexander Sage, Radu Timofte, and Luc Van Gool. 2019. Optimal Transport Maps For Distribution Preserving Operations on Latent Spaces of Generative Models. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[3] Samaneh Azadi, Catherine Olsson, Trevor Darrell, Ian Goodfellow, and Augustus Odena. 2019. Discriminator Rejection Sampling. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[4] Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. 2020. Statistical Mechanics of Deep Learning. *Annual Review of Condensed Matter Physics* 11, 1 (2020), 501–528.

[5] David Bau, Jun-Yan Zhu, Jonas Wulff, William Peebles, Hendrik Strobelt, Bolei Zhou, and Antonio Torralba. 2019. Seeing what a gan cannot generate. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4502–4511.

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

[7] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165* (2020).

[8] Nicholas Carlini, Matthew Jagielski, and Ilya Mironov. 2020. Cryptanalytic Extraction of Neural Network Models. In *Proceedings of Annual International Cryptology Conference (CRYPTO)*. Springer.

[9] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. 2020. Extracting Training Data from Large Language Models. *arXiv preprint arXiv:2012.07805* (2020).

[10] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*. IEEE, 39–57.

[11] Varun Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. 2020. Exploring Connections Between Active Learning and Model Extraction. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association.

[12] Dingfan Chen, Ning Yu, Yang Zhang, and Mario Fritz. 2020. Gan-leaks: A taxonomy of membership inference attacks against generative models. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 343–362.

[13] Kangjie Chen, Tianwei Zhang, Xiaofei Xie, and Yang Liu. 2020. Stealing Deep Reinforcement Learning Models for Fun and Profit. *arXiv preprint arXiv:2006.05032* (2020).

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Xin Ding, Z Jane Wang, and William J Welch. 2020. Subsampling Generative Adversarial Networks: Density Ratio Estimation in Feature Space With Softplus Loss. *IEEE Transactions on Signal Processing* 68 (2020), 1910–1922.

[16] Joachim Folz. 2020. simplejpeg 1.4.0. https://gitlab.com/jfolz/simplejpeg

[17] Karan Ganju, Qi Wang, Wei Yang, Carl A Gunter, and Nikita Borisov. 2018. Property inference attacks on fully connected neural networks using permutation invariant representations. In *Proceedings of ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 619–633.

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2672–2680.

[19] Jamie Hayes, Luca Melis, George Danezis, and Emiliano De Cristofaro. 2019. LOGAN: Membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies*, Vol. 2019. Sciendo, 133–152.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 770–778.

[21] Yingzhe He, Guozhu Meng, Kai Chen, Xingbo Hu, and Jinwen He. 2019. Towards Privacy and Security of Deep Learning Systems: A Survey. *arXiv preprint arXiv:1911.12562* (2019).

[22] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 6626–6637.

[23] Benjamin Hilprecht, Martin Härterich, and Daniel Bernau. 2019. Monte carlo and reconstruction membership inference attacks against generative models. In *Proceedings on Privacy Enhancing Technologies*, Vol. 2019. Sciendo, 232–249.

[24] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1501–1510.

[25] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. 2017. Stacked generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 5077–5086.

[26] Matthew Jagielski, Nicholas Carlini, David Berthelot, Alex Kurakin, and Nicolas Papernot. 2020. High Accuracy and High Fidelity Extraction of Neural Networks. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association.

[27] Shouling Ji, Weiqing Li, Neil Zhenqiang Gong, Prateek Mittal, and Raheem A Beyah. 2015. On Your Social Network De-anonymizablity: Quantification and Large Scale Evaluation with Seed Knowledge.. In *Proceedings of Network and Distributed Systems Security Symposium (NDSS)*. Internet Society.

[28] Hengrui Jia, Christopher A Choquette-Choo, and Nicolas Papernot. 2020. Entangled Watermarks as a Defense against Model Extraction. *arXiv preprint arXiv:2002.12200* (2020).

[29] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. 2019. PRADA: protecting against DNN model stealing attacks. In *Proceedings of IEEE European Symposium on Security and Privacy (Euro S&P)*. IEEE, 512–527.

[30] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[31] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4401–4410.

[32] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8110–8119.

[33] Kalpesh Krishna, Gaurav Singh Tomar, Ankur P. Parikh, Nicolas Papernot, and Mohit Iyyer. 2020. Thieves on Sesame Street! Model Extraction of BERT-based APIs. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[34] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[35] Taesung Lee, Benjamin Edwards, Ian Molloy, and Dong Su. 2019. Defending against model stealing attacks using deceptive perturbations. In *Proceedings of IEEE Security and Privacy Workshops*. IEEE, 43–49.

[36] Chuan Li and Michael Wand. 2016. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *Proceedings of European conference on computer vision (ECCV)*. Springer, 702–716.

[37] Chieh Hubert Lin, Chia-Che Chang, Yu-Sheng Chen, Da-Cheng Juan, Wei Wei, and Hwann-Tzong Chen. 2019. COCO-GAN: generation by parts via conditional coordinating. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 4512–4521.

[38] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. 2019. Few-shot unsupervised image-to-image translation. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 10551–10560.

[39] Yang Liu. 2019. Fine-tune BERT for extractive summarization. *arXiv preprint arXiv:1903.10318* (2019).

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

[41] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of ACM SIGKDD international conference on Knowledge discovery in data mining (KDD)*. ACM, 641–647.

[42] Mario Lučić, Michael Tschannen, Marvin Ritter, Xiaohua Zhai, Olivier Bachem, and Sylvain Gelly. 2019. High-Fidelity Image Generation With Fewer Labels. In *Proceedings of International Conference on Machine Learning (ICML)*. 4183–4192.

[43] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens van der Maaten. 2018. Exploring the Limits of Weakly Supervised Pretraining. In *Proceedings of European conference on computer vision (ECCV)*. Springer, 181–196.

[44] Smitha Milli, Ludwig Schmidt, Anca D Dragan, and Moritz Hardt. 2019. Model reconstruction from model explanations. In *Proceedings of Conference on Fairness, Accountability, and Transparency*. ACM, 1–9.

[45] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. 2018. Spectral Normalization for Generative Adversarial Networks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[46] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv preprint arXiv:1901.04085* (2019).

[47] Augustus Odena, Christopher Olah, and Jonathon Shlens. 2017. Conditional image synthesis with auxiliary classifier GANs. In *Proceedings of International Conference on Machine Learning (ICML)*. 2642–2651.

[48] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2019. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4954–4963.

[49] Soham Pal, Yash Gupta, Aditya Shukla, Aditya Kanade, Shirish Shevade, and Vinod Ganapathy. 2020. ACTIVETHIEF: Model Extraction Using Active Learning and Unannotated Public Data. In *Proceedings of AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 34. AAAI, 865–872.

[50] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of ACM on Asia conference on computer and communications security (ASIACCS)*. ACM, 506–519.

[51] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2337–2346.

[52] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proceedings of International Conference on Learning Representations (ICLR)*.

[53] Eitan Richardson and Yair Weiss. 2018. On gans and gmms. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 31. Curran Associates, Inc., 5847–5858.

[54] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International journal of computer vision* 115, 3 (2015), 211–252.

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

[56] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training GANs. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 2234–2242.

[57] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. 2020. Interpreting the latent space of gans for semantic face editing. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 9243–9252.

[58] Reza Shokri, Martin Strobel, and Yair Zick. 2019. Privacy risks of explaining machine learning models. *arXiv preprint arXiv:1907.00164* (2019).

[59] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. 2017. Membership inference attacks against machine learning models. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*. IEEE, 3–18.

[60] Tatsuya Takemura, Naoto Yanai, and Toru Fujiwara. 2020. Model Extraction Attacks against Recurrent Neural Networks. *arXiv preprint arXiv:2002.00123* (2020).

[61] Luke Tierney. 1994. Markov chains for exploring posterior distributions. *The Annals of Statistics* (1994), 1701–1728.

[62] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. 2019. Fixing the train-test resolution discrepancy. In *Proceedings of Annual Conference on Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., 8250–8260.

[63] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *Proceedings of USENIX Security Symposium (USENIX Security)*. USENIX Association, 601–618.

[64] Ryan Turner, Jane Hung, Eric Frank, Yunus Saatchi, and Jason Yosinski. 2019. Metropolis-hastings generative adversarial networks. In *Proceedings of International Conference on Machine Learning (ICML)*. 6345–6353.

[65] Stéfan van der Walt, Johannes L. Schönberger, Juan Nunez-Iglesias, François Boulogne, Joshua D. Warner, Neil Yager, Emmanuelle Gouillart, Tony Yu, and the scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ* 2 (6 2014), e453.

[66] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. 2018. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 8456–8465.

[67] Pan Xudong, Zhang Mi, Ji Shouling, and Yang Min. 2020. Privacy Risks of General-Purpose Language Models. In *Proceedings of IEEE Symposium on Security and Privacy (S&P)*. IEEE, 1471–1488.

[68] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. 2015. LSUN: Construction of a Large-scale Image Dataset using Deep Learning with Humans in the Loop. *arXiv preprint arXiv:1506.03365* (2015).

[69] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. StackGAN: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of IEEE International Conference on Computer Vision (ICCV)*. IEEE, 5907–5915.

[70] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. 2016. Generative visual manipulation on the natural image manifold. In *Proceedings of European conference on computer vision (ECCV)*. Springer, 597–613.

Table 9: JS distances between models. For the JS distance between training data and the target model, and the target model SNGAN is 16.36×10^{-3}.

Target model	Attack model	$\mathcal{JS}_{accuracy}$ ($\times 10^{-3}$)	$\mathcal{JS}_{fidelity}$ ($\times 10^{-3}$)
SNGAN	SNGAN	8.90	34.12
	PGGAN	1.60	18.56

A APPENDIX

A.1 Implementation details

We implement PGGAN[3] and SNGAN[4] based on following codes indicated in the footnotes. We choose the ResNet architecture for SNGAN and the architecture of PGGAN is the same as the official implementation. We use hinge loss for SNGAN and WGAN-GP loss for PGGAN. For target GAN on synthetic data in Figure 5, we use four fully connected layers with ReLU activation for both generator and discriminator and the prior is a 2-dimensional standard normal distribution. The training data is a mixture of 25 2-D Gaussian distributions (each with standard deviation of 0.05). We train it using standard loss function [18]. In Section 7 about case study, we directly use the pretrained StyleGAN[5] trained on LSUN-Bedroom dataset as our target model. We resize all images used in our paper to 64×64, except for the case study where images with a resolution of 256×256 are used. The dimension of latent space of SNGAN, PGGAN and StyleGAN is 256, 512 and 512, respectively, and their latent codes are all draw from standard Gaussian distribution.

In Section 8 about semantic interpolation defense, the semantic information is from attributes of CelebA dataset[6], which has labeled for each image. we only choose 12 (male, smiling, wearing lipstick, mouth slightly open, wavy hair, young, eyeglasses, wearing hat, black hair, receding hairline, bald, mustache) out of 40 facial attributes to learn semantic hyperplanes, because the number of images for each attribute varies largely and some attributes is hard to distinguish when they are applied in target GAN model. We train prediction model for each attribute based on ResNet-50 model pretrained on ImageNet[7].

A.2 MH algorithm

Algorithm 1 shows the MH subsampling algorithm. Inputs of this algorithm are a target generator (only used to query), a white-box discriminator which is used to subsample generated samples and partial real samples which are used to calibrate the discriminator (Algorithm 1, line 2). Outputs are refined samples whose distribution is much closer to distribution of real training data.

A.3 More results for analyzing distribution differences

A.3.1 ***Understanding accuracy extraction for the target model SNGAN.*** Figure 11 shows distribution differences for the target model SNGAN trained on CelebA dataset. Table 9 summarizes these differences statistically.

[3] https://github.com/tkarras/progressive_growing_of_gans
[4] https://github.com/christiancosgrove/pytorch-spectral-normalization-gan
[5] https://github.com/NVlabs/stylegan
[6] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[7] https://download.pytorch.org/models/resnet50-19c8e357.pth

Algorithm 1 MH subsampling

Input: target generator G, target discriminator D, partial real samples $X_r = \{x_{r1}, x_{r2}, \ldots, x_{rm}\}$
Output: N refined images

Sample m fake images $X_g = \{x_{g1}, x_{g2}, \ldots, x_{gm}\}$ from G
Train a calibrated classifier: $C \leftarrow LogisticRegression(D(X_r), D(X_g))$
$images \leftarrow \emptyset$
while $|images| < N$ **do**
  $x \leftarrow$ a real image from X_r
  for $i = 1$ to K **do**
    Sample x' from G
    Sample u from Uniform$(0, 1)$
    Compute real image's density ratio: $r(x) = \frac{C(D(x))}{1-C(D(x))}$
    Compute fake image's density ratio: $r(x') = \frac{C(D(x'))}{1-C(D(x'))}$
    $p = min(1, \frac{r(x')}{r(x)})$
    if $u \leq p$ **then**
      $x \leftarrow x'$
    end if
  end for
  if x is not a real images **then**
    Append$(x, images)$
  end if
end while

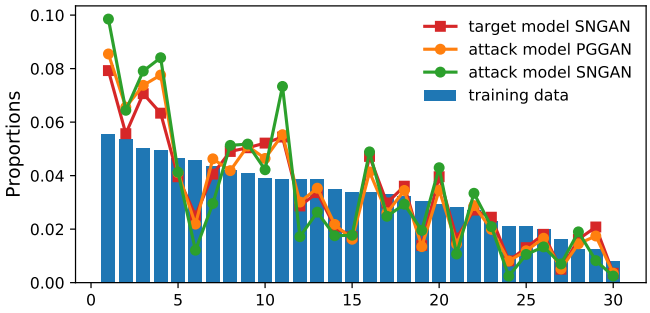
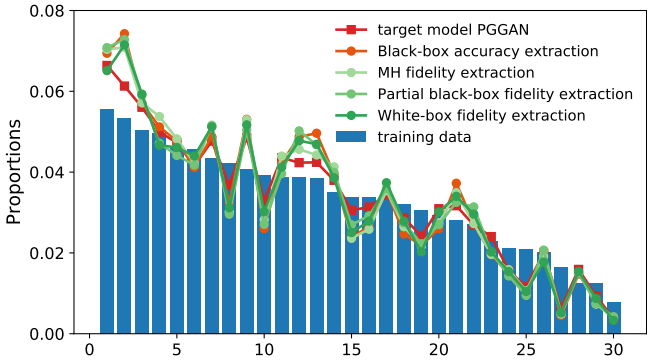
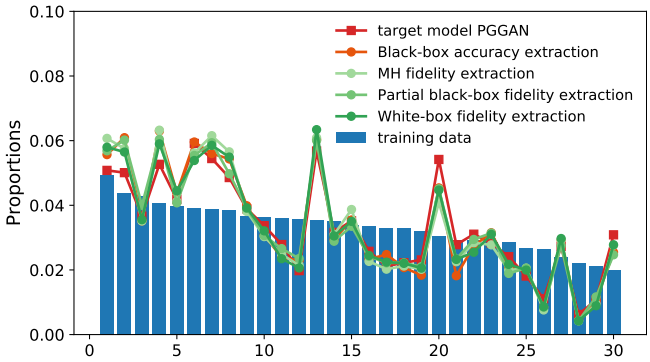


Figure 11: Class distributions of the training data, the target model SNGAN, and attack models.

A.3.2 ***Understanding fidelity extraction on GANs in-depth.*** Following the same procedure illustrated in Section 5.3.4, we also dissect distribution differences for fidelity extraction. Specifically, we choose the PGGAN-PGGAN case as an example (see Figure 7) and the attack models is PGGAN. From the Figure 12, we observe that for CelebA, white-box fidelity extraction which has minimal *fidelity* values among these methods is more consistent with the distribution of the training data by lowering the highest proportions of classes. For LSUN-Church, similar results also can be observed. Table 10 summarizes these differences statistically.



(a) The target model PGGAN trained on CelebA.



(b) The target model PGGAN trained on LSUN-Church.

Figure 12: Distribution differences for fidelity extraction.

Table 10: JS distances between models. For the JS distance between training data and the target model, the target model PGGAN on CelebA is 4.14×10^{-3} and the target model PGGAN on LSUN-Church is 14.78×10^{-3}.

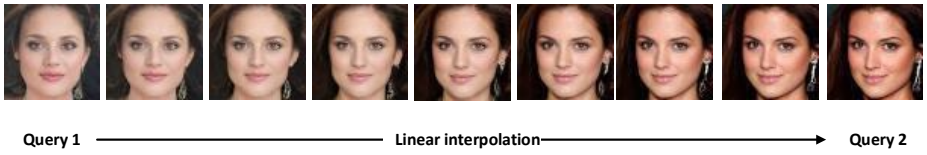
Dataset	Methods	$\mathcal{JS}_{accuracy}$ ($\times10^{-3}$)	$\mathcal{JS}_{fidelity}$ ($\times10^{-3}$)
CelebA	Black-box accuracy extraction	1.83	9.10
	MH fidelity extraction	1.42	8.17
	Partial black-box fidelity extraction	1.53	8.21
	White-box fidelity extraction	1.17	7.53
LSUN-Church	Black-box accuracy extraction	2.32	19.14
	MH fidelity extraction	2.28	19.89
	Partial black-box fidelity extraction	1.72	16.95
	White-box fidelity extraction	1.61	18.65

A.4 Defense techniques

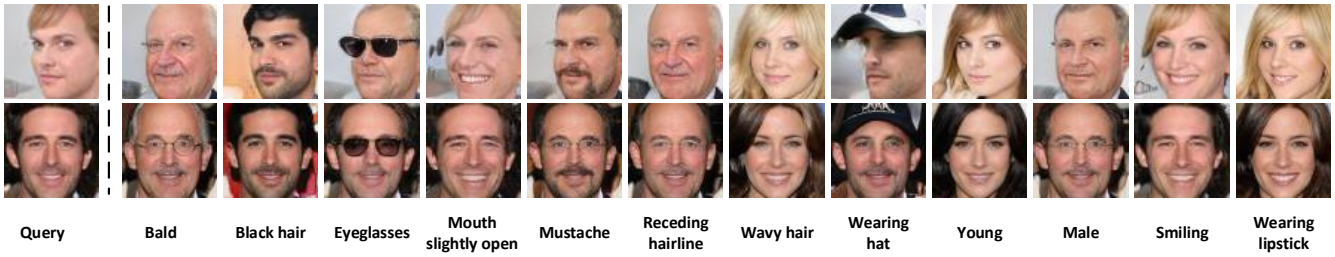
A.4.1 Returned image visualization for models with defenses techniques. Figure 13 shows returned images for input perturbation-based defenses. Figure 14 shows returned images for output perturbation-based defenses.

A.4.2 Attack performance for the attack model SNGAN. Figure 15 shows defense performance of the attack model SNGAN for defenses on black-box accuracy extraction.

A.4.3 Accuracy on defense techniques. Figure 16 shows *accuracy* of attack models for black-box accuracy extraction. We observe

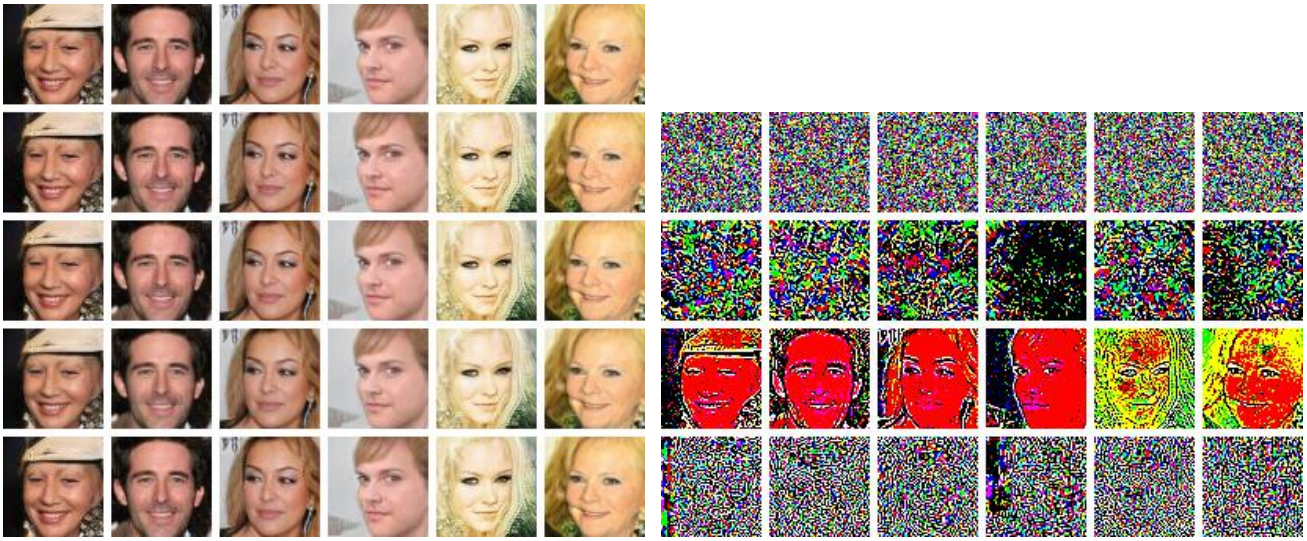


(a) Linear interpolation defense



(b) Semantic interpolation defense. For one latent code, 12 latent codes containing semantic information are generated through semantic interpolation and corresponding images are shown above.

Figure 13: Returned images after input perturbation-based defense techniques. Queried images and interpolated images both show good quality in visual comparison, and images generated by linear interpolation show more similarity than that by semantic interpolation.



(a) Output images. From top to bottom: generated images, Gaussian noise images, Adversarial noise images, Gaussian filter images and JPEG compression images.

(b) Noises. For the top two rows, they are Gaussian noises and adversarial noises, respectively. For the third row, it is the differences between Gaussian filter images and generated images. For the last row, it is the differences between JPEG compression images and generated images.

Figure 14: Returned images after output perturbation-based defense techniques.

that *accuracy* values of attack models can be largely decreased with an increase in the number of queries.

A.5 Qualitative results for target models and attack models

Figure 17 shows generated images from target GANs. Figure 18 shows the performance of attack models against target model PGGAN and SNGAN, respectively. Figure 19 shows performance of attack models when target model PGGAN trained on CelebA uses input perturbation-based defense. Figure 20 shows performance of

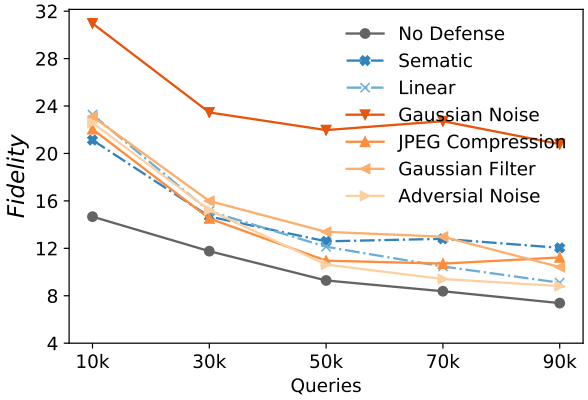
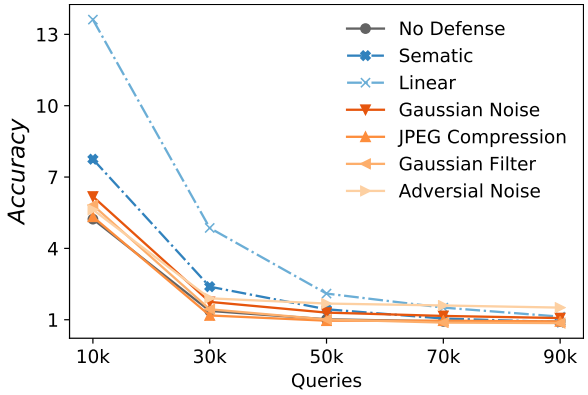
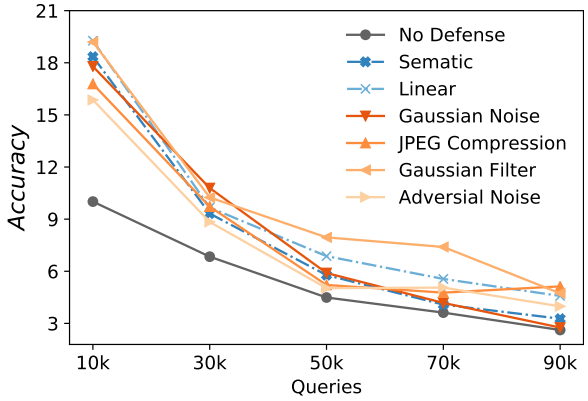


Figure 15: The performance of attack model SNGAN for defenses on black-box accuracy extraction.



(a) The performance of attack model PGGAN



(b) The performance of attack model SNGAN

Figure 16: *Accuracy* **of attack models for black-box accuracy extraction.** *Accuracy* **values of attack models can be largely decreased with an increase in the number of queries.**

attack models when target model PGGAN trained on CelebA uses output perturbation-based defense.

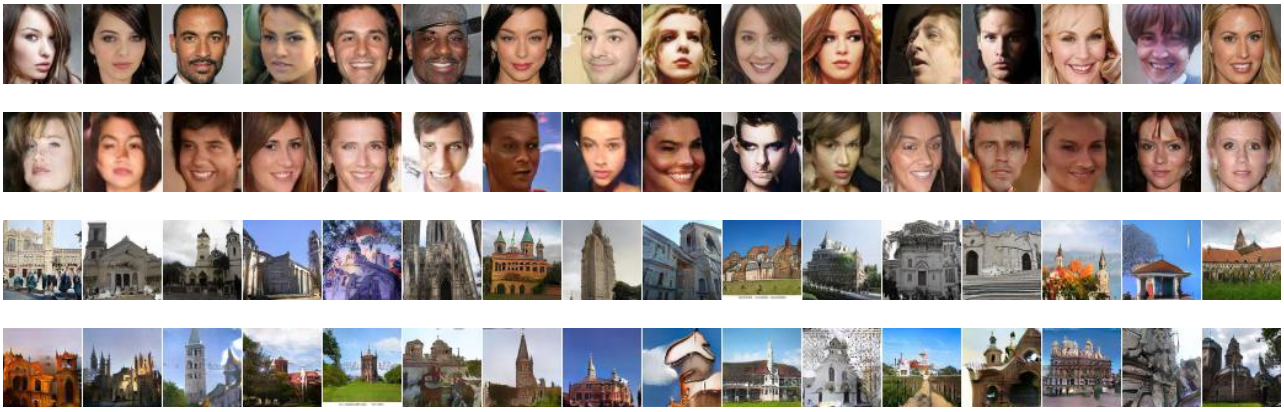


Figure 17: Generated images from target GAN models. From top to bottom: PGGAN on CelebA, SNGAN on CelebA, PGGAN on LSUN-Church and SNGAN on LSUN-Church.

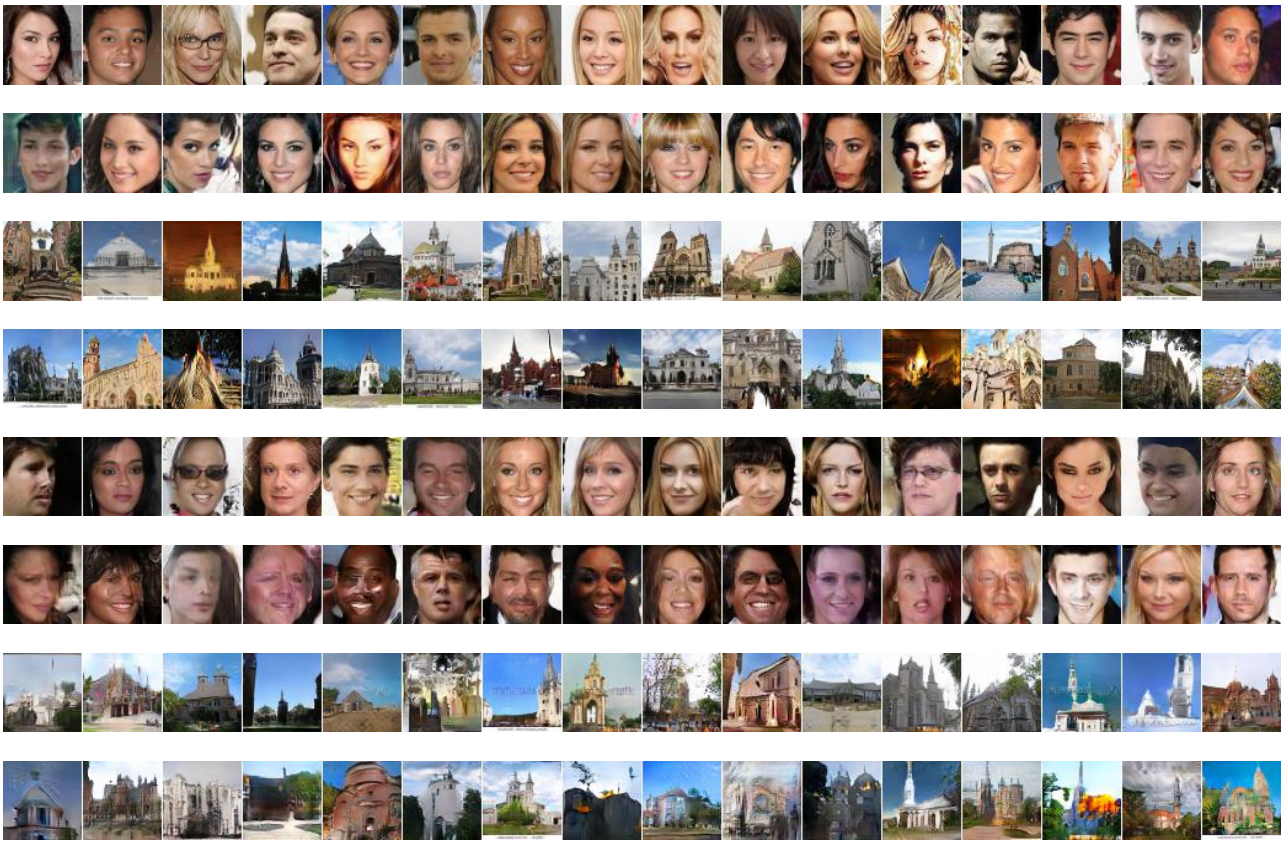


Figure 18: The performance of attack models. The first four rows show the performance of attack models against target model PGGAN. From top to the fourth row: PGGAN-PGGAN on CelebA, SNGAN-PGGAN on CelebA, PGGAN-PGGAN on LSUN-Church and SNGAN-PGGAN on LSUN-Church. The last four rows show the performance of attack models against target model SNGAN. From the fifth row to bottom: PGGAN-SNGAN on CelebA, SNGAN-SNGAN on CelebA, PGGAN-SNGAN on LSUN-Church and SNGAN-SNGAN on LSUN-Church.

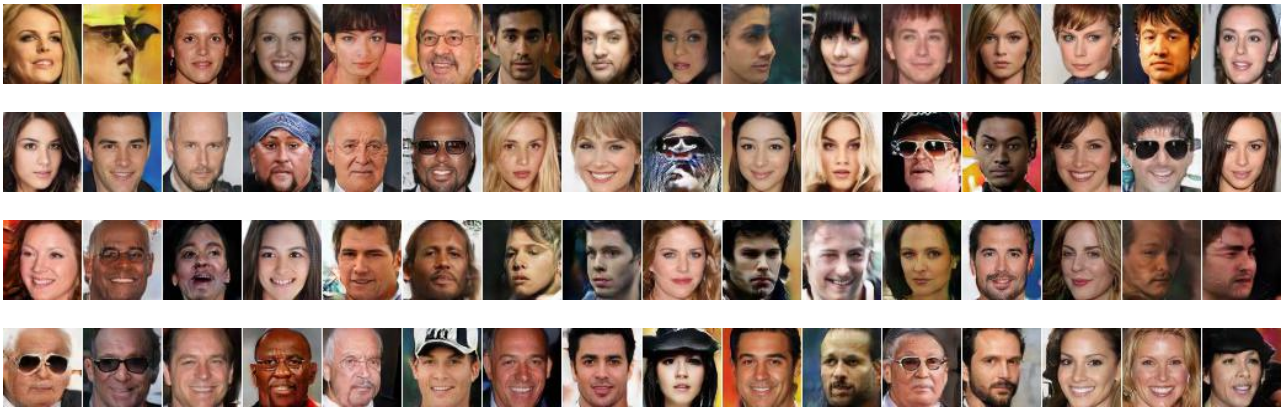


Figure 19: The performance of attack models for target model PGGAN with input perturbation-based defense. From top to bottom: PGGAN-PGGAN with linear defense, PGGAN-PGGAN with semantic defense, SNGAN-PGGAN with linear defense, SNGAN-PGGAN with semantic defense.

Figure 20: The performance of attack models for target model PGGAN with output perturbation-based defense. From top to bottom: PGGAN-PGGAN with Gaussian noise defense, PGGAN-PGGAN with adversarial noise defense, PGGAN-PGGAN with Gaussian filtering defense, PGGAN-PGGAN with JPEG compression defense, SNGAN-PGGAN with Gaussian noise defense, SNGAN-PGGAN with adversarial noise defense, SNGAN-PGGAN with Gaussian filtering defense, SNGAN-PGGAN with JPEG compression defense.