

Adversarial Attacks Against Deep Generative Models on Data: A Survey

Hui Sun, Tianqing Zhu[✉], *Member, IEEE*, Zhiqiu Zhang, Dawei Jin[✉], *Member, IEEE*, Ping Xiong, *Member, IEEE*, and Wanlei Zhou, *Senior Member, IEEE*

Abstract—Deep generative models have gained much attention given their ability to generate data for applications as varied as healthcare to financial technology to surveillance, and many more - the most popular models being generative adversarial networks (GANs) and variational auto-encoders (VAEs). Yet, as with all machine learning models, ever is the concern over security breaches and privacy leaks and deep generative models are no exception. In fact, these models have advanced so rapidly in recent years that work on their security is still in its infancy. In an attempt to audit the current and future threats against these models, and to provide a roadmap for defense preparations in the short term, we prepared this comprehensive and specialized survey on the security and privacy preservation of GANs and VAEs. Our focus is on the inner connection between attacks and model architectures and, more specifically, on five components of deep generative models: the training data, the latent code, the generators/decoders of GANs/VAEs, the discriminators/encoders of GANs/VAEs, and the generated data. For each model, component and attack, we review the current research progress and identify the key challenges. The paper concludes with a discussion of possible future attacks and research directions in the field.

Index Terms—Deep generative models, deep learning, membership inference attack, evasion attack, model defense

1 INTRODUCTION

OVER the past few years, computation power has advanced sufficiently to enable the success of deep neural networks in various applications. Within this category, there are two categories of deep learning models: generative and discriminative. Generative models synthesize data we can observe in our world, such as plausible realistic photographs of human faces [1]. Collectively, these are known as deep generative models (DGMs). The other one is to divide observed data into different classes, e.g., face recognition, recommender systems, etc. [2]. This category of models is known as deep discriminative models (DDMs) [3].

The most popular DGMs are generative adversarial networks (GANs) [4] and variational auto-encoders (VAEs) [5]. Both are widely used to generate realistic photographs [6], synthesize videos [7], translate one image into another [8], etc. As the traditional DDMs, recurrent neural network (RNN) [9], convolutional neural networks (CNN) [10], and their variants perform well at sentiment analysis [11], image recognition [12], natural language progressing (NLP) [13], [14] and

so on. A relationship diagram of the AI landscape is presented in Fig. 1.

As with any technology of wide influence, model security and privacy issues are inevitable. Naturally, any adversary will have two aspirations. The first is to sabotage the model so it does unsatisfactory work. The second is to breach privacy. In sabotaging a model, for example, an attacker might turn a model that is supposed to generate human portraits into one that generates pictures of shoes [15], or instead of correctly classifying pictures as pandas, it might classify them as gibbons [16]. Breaching privacy might include stealing the training data or the whole trained model. A famous example of this was where adversaries duplicated models trained by Amazon through black-box queries from APIs provided by its machine-learning-as-a-service platform [17]. The same tactic has been used to restore the training set so as to acquire private information [18], [19].

Poisoning attacks [20], [21] and evasion attacks [16], [22] both attempt to force a model to do unsatisfactory work. Poisoning attacks operate during the training phase, and attempt to compromise the model's abilities at the formation stage. Evasion attacks work during the test phase, with the aim of providing adversarial input to the trained model so that it produces unsatisfactory output. Adversarial input is generally called an adversarial example.

At the component level, there are several different types of attacks. At the data level, we have membership inference attacks, which attempt to infer whether a given sample belongs to the model's training set [18] and, also, model inversion attacks, which try to reconstruct some or all of the training data based on the some prior information and the model's output [19]. At the attribute level, we have attribute inference attacks, which attempt to infer the sensitive attributes of data

- Hui Sun, Tianqing Zhu, and Zhiqiu Zhang are with the China University of Geosciences, Wuhan, Hubei 430074, China. E-mail: {sunhui, zhangzhiqiu}@cug.edu.cn, tianqing.zhu@ieee.org.
- Dawei Jin and Ping Xiong are with the Zhongnan University of Economy and Law, Hubei 430073, China. E-mail: {jdw, pingxiong}@zuel.edu.cn.
- Wanlei Zhou is with the City University of Macau, Macau, China. E-mail: wlzhou@cityu.edu.mo.

Manuscript received 22 July 2021; revised 12 Oct. 2021; accepted 21 Nov. 2021. Date of publication 26 Nov. 2021; date of current version 7 Mar. 2023.

This work was supported by the National Natural Science Foundation of China under Grant 61972366.

(Corresponding author: Tianqing Zhu.)

Recommended for acceptance by W. Zhang.

Digital Object Identifier no. 10.1109/TKDE.2021.3130903

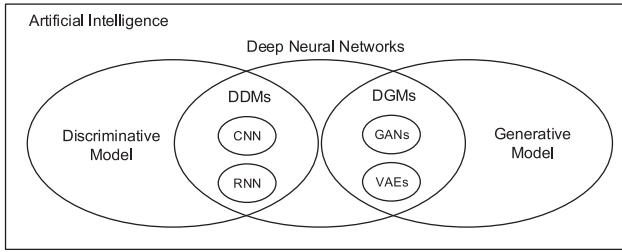


Fig. 1. The AI landscape. AI has two main branches, generative models and discriminative models. The deep neural network variants of these models have evolved into VAEs and GANs on the generative side and into RNNs and CNNs on the discriminative side.

[23]. Model extraction attacks work at the model level. These are a severe threat that try to duplicate the entire trained model [17].

Although research into these attacks on DGMs is still in its infancy, there is a healthy body of literature on the security and privacy issues associated with DDMs. For instance, Papernot *et al.* presented a detailed adversarial framework of security and privacy attacks that included adversarial examples, strategies for membership inference attacks, and some defense methods [24]. Focusing on scenarios and applications, Liu *et al.* categorized both the types of attacks and types of protection schemes [25]. Serban *et al.* elaborated on adversarial examples, including their construction, defense strategies, and transfer capabilities [26]. As differential privacy is one of the most effective measure for mitigating privacy breaches, Gong *et al.* published a comprehensive review on differentially-private machine learning [27]. A number of surveys have also been conducted on DDMs, particularly CNNs and RNNs, see e.g., [28], [29], [30], [31], [32].

In terms of DGMs, there has been much less work, as this survey will show. Our review unearthed the following research papers on: poisoning attacks [33], [34]; evasion attacks [15], [35], [36], [37], [38], [39], [40]; membership inference attacks [41], [42], [43], [44], [45], [46]; and attribute inference attacks [46] and model extraction attacks [47]. To the best of our knowledge, there are no surveys devoted to the security and privacy of DGMs. However, in recent years, GANs and VAEs have advanced greatly so that, now, DGMs are attracting much more attention, both well-meaning and ill-intentioned. We therefore think it is time for a thorough survey of those attacks and, of course, their defenses. By comparing DGM attacks with DDM attacks and their known defenses, we may be able to identify some critical gaps between them.

- On a basic level, adversarial attacks are about the evolution of a strategy. The attacks mentioned above were originally designed for discriminative models and DGMs have a very different purpose to DDMs. As such, the training algorithms and model architectures are also very different. Therefore, to perform traditional attacks against DGMs, the attack strategies must be updated. One single attack strategy cannot reveal the overall direction this evolution will take. Rather, a comprehensive review is required.
- Whether the evolved attacks will be general to DGMs is another concern. Since there are multiple variants of VAEs and GANs, such as beta-VAEs [48] and

Wasserstein GANs [49] as well as other less popular types of DGMs, generality would make sense.

- There may be rare defense strategies specially designed for the occasions when a DGM suffers various types of attacks.

In these regards, a systematic study of the current state-of-play in the field will be essential to future research efforts. Thus, the main contributions of this survey include:

- A brief introduction to VAEs and GANs, the most popular DGMs, beginning with their standard model structures and training procedures and ending with a comparison between the DGM and DDM architectures.
- An analysis of the feasibility of the various attacks given the two stated adversarial goals - to sabotage the model's proper functioning and to compromise privacy - and the vulnerability of the model's individual components. This section also categorizes the common attack strategies.
- A summary of the existing defense schemes and a discussion on the possible defense methods, which, given the rarity of defenses, make up the bulk of future research directions.
- Suggestions for other fruitful research opportunities worthy of further attention.

2 A BACKGROUND ON DGMs

2.1 Notations

Consider a DGM with a training set D_{train} that consists of numerous instances sampled from a real data distribution P_{real} and an expectation that the training data distribution P_{train} approximates the real data distribution P_{real} . The model learns the real data distribution from the training set and aims to generate samples that seem to be real but are unseen. Here, x denotes a real sample in training set, \hat{x} denotes a generated sample, and $D_{generated}$ and $P_{generated}$ denote the collection and distribution of the generated samples, respectively. For a generated data distribution $P_{generated}$ to be plausible, it must be close to the training data distribution, and therefore, in turn, close to the real data distribution. This can be expressed as $P_{generated} \approx P_{train} \approx P_{real}$. To maintain diversity, latent code z is randomly sampled from a distribution defined as P_z . This is another representation of an input sample.

Both GANs and VAEs have two components, each taking the form of a neural network. A GAN consist of a generator and a discriminator; the corresponding functions are expressed as f_{gen} and f_{dis} . A VAE consist of an encoder and a decoder with the corresponding functions similarly expressed as f_{enc} and f_{dec} . Further, most evasion attacks involve a target output, denoted as x_{target} , and most membership inference attacks involve a query/series of queries x_{query} the adversary uses to infer information. So, for instance, a membership inference might be explained as $Pr(x_{query} \in D_{train})$, where $Pr()$ denotes the possibility rate.

2.2 DGMs: GANs and VAEs

As a major branch of deep learning, DGMs focus on data generation. DGMs are unsupervised, automatically learning

TABLE 1
Notation

Notation	Description
P_{real}	Distribution of real data
D_{train}, P_{train}	Training dataset and its distribution
x	Real sample
\hat{x}	Generated sample
$D_{generated}, P_{generated}$	Dataset of generated samples and corresponding data distribution
z	Latent code
P_z	The latent distribution, such as Gaussian distribution
x_{query}	The given query sample for membership inference
x_{target}	The given target output sample for evasion attack
f_{enc}, f_{dec}	Function of encoder and decoder
f_{gen}, f_{dis}	Function of generator and discriminator
$KL(\cdot \cdot)$	The Kullback–Leibler divergence

the data patterns in a training set so that the model has the capacity to generate new samples in accordance with a distribution that is as similar as possible to the true data distribution. GANs learn the distribution implicitly under a minimax game where a generator tries to fool a discriminator and the discriminator tries not to be deceived [50]. VAEs learn the distribution explicitly by limiting reconstruction errors under an encoder-decoder framework.

GANs and VAEs follow different principles and, thus, have different model architectures, as shown in Figs. 2 and 3. However, both are made up of the same five broad components. These are:

- 1) The training set, which consists of numerous real samples following a distribution that approximates the real data distribution.
- 2) Latent code, which is an alternative vector representation of the data. Typically, this has lower dimensionality than the input representation and is generally randomly sampled from a latent distribution to satisfy the requirement of never-seen generated samples. Essentially, the distribution is defined as latent space.
- 3) Generator (GANs)/decoder (VAEs) - both are generative components that finish the mapping from randomly sampled latent code to a sample formally denoted as $z \rightarrow x$.
- 4) Discriminator (GANs)/encoder (VAEs) - both are auxiliary components that help the generator/decoder become better trained and can, thus, be discarded when the training ends.
- 5) Generated data, which is the output of the generator/decoder. With the well trained generator/decoder, the distribution of generated data will approximate the real distribution.

2.2.1 GANs

In GANs, the generator takes latent code as input and generates samples. Both these samples generated and real samples are then sent to the discriminator, which acts as a binary classifier with the task of distinguishing the real data from the generated data. Thus, a GAN’s training is formulated as a minimax game [51] where a discriminator and a generator compete against each other. The generator tries to

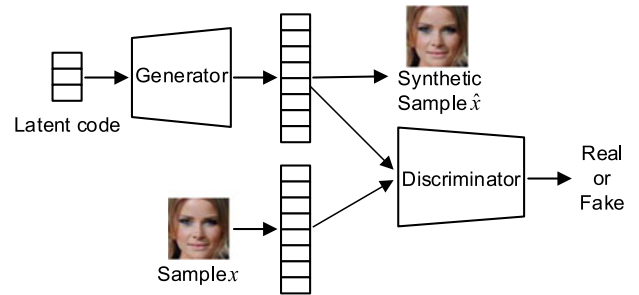


Fig. 2. GAN architecture. A GAN consists of a generator and a discriminator, both of which are deep neural networks. In this example, the generator maps the latent code as an image representation, and the discriminator tries to distinguish between the generated image representation and the raw image representation. This ensures the generated samples are plausible. The process is formulated as a minimax game where generator tries to fool discriminator into classifying generated samples as raw samples and the discriminator tries not to be deceived.

produce a fake sample that fools the discriminator into classifying it as true, while the discriminator tries to perfectly discriminate between the fake data and the true data. Formally, this can be expressed as

$$\min_{f_{gen}} \max_{f_{dis}} L_{GAN} = E_{x \sim D_{train}} [\log f_{dis}(x)] + E_{z \sim P_z} [\log(1 - f_{dis}(f_{gen}(z)))], \quad (1)$$

where f_{gen} and f_{dis} denote the generator and discriminator functions, respectively, D_{train} denotes the training set, and P_z is the prior latent distribution, usually a normal Gaussian distribution. The first term, denotes the real loss, i.e., the cross-entropy loss of the real data that is classified as real by the discriminator. The second term, denotes the fake loss, i.e., the cross-entropy loss of the generated data that the discriminator classifies as generated. The generator hopes to minimize the fake loss, while the discriminator hopes to maximize both the real and fake loss.

During training, the generator and discriminator are each trained in turn. While the discriminator is being trained, the parameters of generator network are fixed, and vice versa. The training ends when both the generator and discriminator are not showing further improvement. As a result, the generated data is so similar to real data that it successfully fools the discriminator.

2.2.2 VAEs

VAEs generate samples based on the cascaded work of the encoder and decoder. The encoder compresses the input sample into a lower-dimensional latent space and the decoder

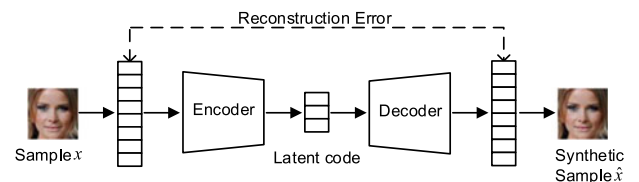


Fig. 3. VAE architecture. A VAE consists of an encoder and a decoder, both of which are also deep neural networks. The encoder encodes the input representation into a lower-dimensional latent code, which subsequently is decoded into a representation by decoder. The reconstruction error mechanism between the input and decoded representations ensures the plausibility of the generated samples.

TABLE 2
DGMs Versus DDMs

	Model instance		Input	Model output
DGMs	GANs	Generator	Latent code	Generated data
		Discriminator	Generated data, training data	A label[0,1] and a confidence score
	VAEs	Encoder	Training data	Latent code
		Decoder	Latent code	Generated data
DDMs	CNNs		Training data pairs <data, label>	Labels and corresponding confidence scores

decompresses randomly sampled latent code from the latent space into a sample. This compressing and decompressing is commonly referred to as encoding and decoding/reconstruction. A reconstruction mechanism is conducted between the input sample and decompressed sample so that the latent code keeps the maximal information of input sample during encoding process and the decompressed sample has minimal reconstruction errors during decoding. As part of this process, the latent space must be regularized to be continuous and complete. Hence, a point randomly sampled from the latent space could be decoded as a new and plausible sample. The best encoding-decoding schemes and regularized latent spaces are achieved using an iterative optimization process with the loss function

$$\min L_{VAE} = -E_{z \sim Q(z|x)}[\log P(x|z)] + KL(Q(z|x)||P(z)), \quad (2)$$

where $Q(z|x)$ and $P(x|z)$ are the encoder and decoder networks, respectively, x is the input sample, and z is the latent code. The first term, in the case of a reconstruction error, denotes the cross-entropy between the input x and their reconstruction \hat{x} , $\hat{x} = f_{dec}(f_{enc}(x))$. The second term, often called regularization term, regularizes the latent space by ensuring the returned distribution $Q(z|x)$ is close to prior distribution of the latent code $P(z)$. Generally, this follows a standard multivariate Gaussian distribution $N(0, I)$. The Kullback-Leibler divergence $KL(\cdot||\cdot)$ is normally used to measure the distance between the two distributions.

2.3 A Comparison of DGMs and DDMs

To analyze DGMs' vulnerability to the mainstream attacks, e.g., membership inference attack, we compare DGMs with DDMs in aspect of components and corresponding inputs and outputs. Table 2 lists the comparison result of the typical DGMs, i.e., GANs and VAEs, and a DDM, i.e., convolutional neural networks (CNNs). DDMs take the labels of each record in training set as input, which becomes the benchmark for training. For DGMs, from a holistic perspective, the real data is the concrete benchmark that verifies the quality of generated samples. And DGMs output data, while DDMs output the probabilities of a label, i.e., a confidence score.

In view of the differences, there are clear security and privacy vulnerabilities for DGMs as follows.

- 1) DGMs have more complex input, i.e., latent code and training data, which provides new directions for attacks against model input, like evasion attacks in latent space.

- 2) DGMs reveal training data patterns by generating plausible samples, which leaves the privacy of training set rather transparent. Hence, DGMs are particularly vulnerable to attacks against training set, like membership inference attacks.
- 3) DGMs, except discriminators of GANs, does not provide labels or confidence score, thus the derived attacks, e.g., membership inference attack in [18] and model extraction attack in [17], are not perfectly feasible.

DGMs also have inner characters in common with DDMs. Both are based on deep neural networks, which means that DGMs tend to suffer from some of the same problems as DDMs - overfitting, for example. And discriminators of GANs are tantamount to DDMs. Further, both are trained on training data thus vulnerable to poisoning attacks.

In general, DGMs are vulnerable to mainstream attacks, i.e., membership inference attacks, attribute inference attacks, model extraction attacks, poisoning attacks and evasion attacks; however, the traditional attacks, e.g., membership inference and model extraction attacks based on confidence score [17], [18], would not work perfectly for DGMs. Specialized attack strategies are in requirement for attacking DGMs.

3 THREAT MODELS FOR ATTACKING GANs AND VAES

3.1 Adversary's Goals

Goal 1: Breaking the Model

Achieving this goal requires a disruption to the generative process that either results in: a) the intended output samples but at a low quality; b) some other presupposed samples within or out of the domain; or c) samples with no suppositions but ones that are not similar to the original output. If these presupposed samples sit within the domain, they should have a high probability of lying in the expected data distribution. Samples outside the domain would have the opposite and are not likely to follow the distribution.

This goal can be in achieved several ways. For instance, a poisoning attack can disrupt the model's generative abilities during training phase, while an evasion attack can do the same during the testing phase.

Poisoning Attack: The basis of this attack is to inject carefully crafted samples into the training set thereby poisoning it. Then, any model trained on the poisoned set will learn wrong abilities with wrong model parameters. Another way is to damage part of the model's structure, such as its loss function, to alter the model's workflow. Both strategies

TABLE 3
Adversarial Goals and Targets

Adversarial Goal	Attack Target	Attack Category	Phase	Model Corruption
Goal 1: To break the model	Generator	Evasion	Test	Integrity
	Decode	Evasion	Test	Integrity
	Encoder	Evasion	Test	Integrity
	Dataset	Poisoning	Training	Integrity
Goal 2: To steal confidential information	Generator	Membership inference Attribute inference Model extraction	Test	Confidentiality
	Decoder	Membership inference Attribute inference Model extraction	Test	Confidentiality
	Discriminator	Membership inference	Test	Confidentiality

can be teamed with triggers to allow the attack to work within certain conditions, known as a backdoor poisoning attack [52]. Such attacks tend to avoid early detection.

Evasion Attack: This attack carefully crafts the model input to induce an unsatisfactory output. Such input is defined as an adversarial example. For a DGM, model input includes the latent code and the input sample. Accordingly, adversarial examples can be crafted for the latent code and sample, often called the latent adversarial example and adversarial example in this survey.

Goal 2: Stealing Confidential Information

Any information that authorized users could not obtain from a normal query to the trained model is confidential. Adversaries' prime targets include the model's parameters and its training set. Typical attack strategies include: the inference attack, where adversaries try to infer real data and/or attribute values with high confidence [53]; the model extraction attack, where the aim is to duplicate the functionality of the model [17]; and the model inversion attack, where adversaries recover the training data and thus also gain access to the model. To the best of our knowledge, there have not been any model inversion attacks against DGMs, yet. However, there have been several studies on how one might perpetrate a model inversion attack from the generated samples or latent code, while focus more on the latent space, such as the interpretability [54] and regularization [55] of latent code.

Membership Inference Attack [18]: In this type of attack, the adversary tries to deduce whether a given sample is part of the model's training set. Prior information about the set, such as its size, can help them to deduce whether a set of samples are subset of the training set. With multiple queries, the entire training set might be recovered. Membership inference attacks lead to severe privacy leaks. They also provide clues about the strategies for other types of privacy attacks.

Attribute Inference Attack [46]: It is also known as record linkage attack, in the attack, adversaries have knowledge of some of the common attributes of the dataset, which they use to try and infer the sensitive attributes of a given sample. The common attributes are generally freely available to the public, such as a street scene, but the sensitive attributes are ones protected from public view, such as the number plates of the car parked along the street.

Model Extraction Attack [17]: The idea of this attack is to infer the parameters or functions of the model via an

efficient set of queries. If successful, the adversary can then copy the model's functions partly or even completely.

Goal 1 is achieved by destroying the model's integrity and *Goal 2* is achieved by destroying the model's confidentiality[56]. Model integrity means that the model's training and testing process suffer no disturbance so the model produces normal output. Poisoning attacks disturb the training process and evasion attacks disturb the testing process, both of which result in unsatisfactory output, and accomplish *Goal 1*. Model confidentiality means that sensitive data should only be disclosed to authorized users. Membership inference attacks and attribute inference attacks all reveal the training data, while model extraction attacks duplicate the functionality of the model. All of these attacks procure confidential information without authorization, thus accomplishing *Goal 2*. A summary of these goals and targets is given in Table 3.

3.2 The Adversary's Prior Knowledge

Most types of attack either rely on or work better when the adversary holds some prior information about the model or its training set. The more prior information the adversary holds, the more powerful the attacks and the more successful it is likely to be. Prior information that adversaries may have includes:

- 1) *Training data and training algorithm.* The security of the training data is the basis of the model's confidentiality. However, a model owner may publicly share their training set during testing to explain their algorithm, which could reveal much about that data and the model's parameters.
- 2) *Model parameters.* These include the discriminator and generator of GANs, and the encoder and decoder of VAEs. The model owners may publish a full GAN/VAE online to show their product and encourage further updates. Additionally, they may publish part of the model, i.e., the discriminator of GANs providing a tool to test the effectiveness of their work. With the model parameters, the adversaries can design more detailed and personalized attacks, i.e., specially crafting adversarial examples and inferring data membership.
- 3) *Latent code.* As another representation of data, latent code plays a decisive role in data synthesis. There are two situations by which adversaries could come

TABLE 4
Adversary's Information and Capability

Adversary	Training data and algorithm	Model parameters	Latent code	Generated data	Auxiliary information
Internal attack	!	#	#	#	#
White-box attack	○	!	!	!	○
Partial black-box attack	○	#	○	!	○
Black-box attack	#	#	#	!	#

There are five categories of prior information an adversary can hold: training data and algorithms, model parameters, latent code, generated data, and other auxiliary (publicly collected) information. The confidentiality of the information decreases from left to right. ! denotes essential access, # denotes access not required, ○ denotes access possibly required.

to have this knowledge. First, they may have direct control of latent code, in which case, they can alter its value to satisfy their goals. Second, they may know a latent distribution, which means they can alter the latent code indirectly.

- 4) *Generated data*. This is the most easy and basic information for an adversary to get. Generally, it is acquired by querying a DGM through its API. Adversaries can also be provided with a set of generated data by an unknowing user.

Adversaries also have types depending on their capabilities and the information they possess. If they have access to the training algorithms and data, they can act more like insiders to fundamentally corrupt the model. These are known as internal attackers. Those with access to only the generated data are called black-box adversaries. If with comprehensive knowledge of the model parameters, they are called white-box adversaries. If with no access to the model's parameters but have access to more than the generated data,

for example, the latent code, they are defined as partial black-box adversaries. Table 4 lists these classifications.

To reach their goals, adversaries consider the prior knowledge they have and design an attack strategy accordingly. Based on adversarial information against each component, including training set, latent code, GAN generator/VAE decoder, GAN discriminator/VAE encoder and generated data. For the fact that the latent code is the input of generator of GAN (decoder of VAE) and generated data is the output, we classify the attacks against latent code and generated data into the attacks against generator/decoder. Specifically, we get following types of attacks: attack against generator/decoder of VAE, attack against discriminator/encoder, and attack against training set. Table 5 summarizes the literature on the types of attacks, the components targeted, and the strategies used for each type of DGM plus the data. Attacks on models/components that are not feasible are indicated as n/a. Attacks that are unexplored are marked as TBD to reflect this gap in the literature.

TABLE 5
Overview of the Attack Types by Component

Attack Component	Evasion	Membership inference	Attribute inference	Model extraction	Poisoning
Generator	Forces a generator to do the wrong work via latent adversarial code. The adversarial code should not be too far away from the prior distribution of the existing [15]	Infers whether a given sample belongs to the training set based on the generated data [41], [42], [44], [45], [46], [57]	Infers sensitive attributes based on the generated data [46], 2020	Duplicates a model based on the generated data [47]	n/a
Discriminator	n/a	Infers whether a given sample belongs to the training set based on the discriminator's output and exact confidence scores [43]	TBD	TBD	n/a
Decode	Forces a generator to do the wrong work via latent adversarial code [39]	Infers whether a given sample belongs to the target training set based on the generated data [41], [42], [44], [45], [46], [57]	Infers sensitive attributes based on the generated data [46]	Duplicates the target model based on the generated data [47]	n/a
Encoder	Forces a generator to do the wrong work by making the image input adversarial, which makes the code adversarial [35], [36], [37], [38], [39], [40]	TBD	TBD	TBD	n/a
Dataset	n/a	n/a	n/a	n/a	Disturbs model training process, including its static datasets and data processing logics [33], [34]

This table summarizes the basic strategy for each attack as it pertains to each component of the model. Non-existent situations are denoted as not applicable (n/a). Possible but still unexploited situations are denoted as to be determined (TBD).

4 GANS: ATTACKS AGAINST GENERATORS

This section elaborates on attacks against generators, targeting the input, i.e., latent code, or the output, i.e., generated data.

4.1 Evasion Attacks

Manipulating latent code is essential for mounting an evasion attack. An evasion attack can be explained as finding a latent code in a pre-set latent distribution that the generator maps into an unsatisfactory sample. The final generated sample is dissimilar to the original output, but similar to the target sample. Meanwhile, the corresponding latent code should be close to the original latent code. Otherwise, it might be detected by defenders when they verify whether the input latent code belongs to the pre-set distribution. Consequently, the loss function of an evasion attack consists of two parts: the adversarial term L_{adv} to ensure the attack effect, and the regularization term L_{reg} to regularize the perturbation. A hyperparameter λ aims to balance the two parts

$$L_{evasion} = L_{adv} + \lambda \cdot L_{reg}, \quad (3)$$

$$L_{adv} = \Delta(x_{target}, f_{gen}(z)). \quad (4)$$

Pasquini *et al.* [15] was the first to explore the evasion attack against generator. They assumed that a defender would build a distribution hypothesis test to check whether the latent code belongs to the prior distribution before the code is sent to generator. To pass validation, i.e., to ensure that the latent code follows the prior latent distribution, even with updating multiple iterations, they restricted the moment of the latent code to be close to the moment of a random variable sampled from the latent distribution. Formally,

$$L_{reg} = \sum_{i=1}^k \omega_i \|\mu_{z_0}(i) - \tilde{\mu}_z(i)\|_2^2, \quad (5)$$

where z_0 is the latent code randomly sampled from the prior latent space p_z , $\mu_{z_0}(i)$ is the i_{th} moment of z_0 , and $\tilde{\mu}_z(i)$ is the i_{th} sample moment of the latent vector z . Here, z is the iteration result of the original latent code z_0 . The parameter ω_i is the weight assigned to the i_{th} moment difference.

The full attack process would proceed by the adversary first sampling some initial latent code from the prior latent distribution. Then, they would adjust that code with gradient descent by minimizing $L_{evasion}$.

We believe this approach could be extended to the conditional generator, where the defender could randomly choose a label and expect the generator to output a sample of that label. Here, the adversarial loss function would be

$$L_{adv} = \Delta(x_{target}, f_{gen}(z, y)), \quad (6)$$

where y is the randomly chosen label by the defender. During the optimization process, y remains constant.

Overall, it is worth noting that, in a properly functioning model, the generated data is intrinsically similar to the training data, and model overfitting exacerbates this similarity. Hence, generated data can be treated as substitute for the training data. Adversaries can breach much privacy through generated data, which brings us to membership inference, attribute inference, and model extraction attacks.

4.2 Membership Inference Attacks (MIAs)

"Membership" in our survey means whether a sample belongs to the ML model's training set. Since the generated data distribution of a DGM approximates its training set, the problem of inferring membership can be converted into a problem of determining whether the query sample follows the generated data distribution. In this way, membership inference means determining whether the query sample is close to the generated sample. If so, it probably belongs to the training set. Attacks derived from this idea are known as distance-based MIAs. Attribute inheritance provides another idea that, if a query sample is used to train a model, the generated data will preserve certain attributes of the query sample. These attacks are termed as attribute-based MIAs. Each is detailed below. Additionally, we introduce co-MIAs to expand the attacking scenarios and possible attack calibration strategy.

4.2.1 Distance-Based MIA

We define the membership inference based on the distance between the query sample and the generated samples for two reasons. First, from the perspective of distribution approximation [42], [45], since the output distribution approximates the training data distribution, the probability that the query sample belongs to the training set is proportional to the probability that the query sample belongs to the output data distribution. Therefore, the inference can be expressed as whether one of the query samples belong to the output distribution - or, more specifically, whether the query sample was produced by the target generator. Second, from the perspective of overfitting the model [44], if there are several generated samples close to the query sample, the query sample is probably a member of the training set. Both ideas revolve around whether one or more generated samples are close to the query sample. Formally, this can be expressed as

$$\min_{x \in G(\cdot)} \Delta(x_{query}, x), \quad (7)$$

where x_{query} is the given sample, x is the generated sample from target generator G , and Δ is the distance function that calculates the distance between two samples.

In the papers we reviewed, most calculated these distances using either ϵ -ball, where quantity was the focus, or reconstruction distance when quality mattered. Fig. 4 provides more details.

Reconstruction Distance Herein, adversaries focus on the quality of the distance measure. In other words, they are trying to find the generated sample that is closest to the query sample. The closest generated sample is called the reconstruction of the query sample, and the distance between the two is called the reconstruction distance - formally,

$$R(x_{query}, |G) = \underset{x \sim G}{\operatorname{argmin}} \Delta(x_{query}, x) \quad (8)$$

Empirically, it is impossible to obtain every generated sample and find the closest one. Multiple solutions have therefore been proposed to solve the optimization problem based on limited prior information. These are outlined as follows:

- Chen *et al.* considered a black-box attack and simply calculated the distance between the query sample

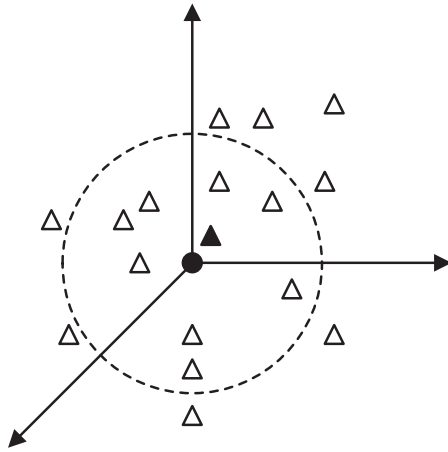


Fig. 4. Membership inference attacks: reconstruction distance and ϵ -ball distance. The dots (\bullet) denote the query sample, the triangles (\triangle and \blacktriangle) denote the generated samples, and the dashed circle denotes the threshold of ϵ -distance. With reconstruction distance, the sample that has the closet distance (labeled as \blacktriangle) is the focus. With ϵ -ball distance, the quantity of generated sample in the dashed circle is the focus (quantity = 9).

and each generated sample [42]. The sample with the least distance to the query sample was deemed the reconstruction. A judgment was then made about the reconstruction error.

- If the adversary has access to the latent code, they can adjust it to get an optimal solution in the regularized latent space. Chen *et al.* [42] proposed to approximate the optimum via Powell's Conjugate Direction Method [58], while Liu *et al.* [45] proposed building another set of neural networks to find the optimal latent code. These adversarial networks took the query sample as input and output the latent codes, like encoders. The adversary then adjusted the parameters of the adversarial net until the output latent code reaches the optimum. This approach essentially transforms the optimization problem into a parameter tuning exercise. However, without the generator's gradient information, Liu and colleagues used finite-difference to approximate the gradient and find the optimum latent code.
- White-box attackers, i.e., attackers with access to the internals of the generator, including the gradient information, can solve the optimization problem more accurately by using an advanced first-order optimization algorithm, such as L-BFGS [42], [45]. Such a solution would be suitable for solving both optimization problems - for the latent code or the parameters.

Hilprecht *et al.* [44] made a compromise in cases where a precise reconstruction was not required to calculate the average distance between the query sample and each generated sample.

Model publishers sometimes launch an MIA themselves before publishing the model to evaluate the model's security. This process is more commonly called a membership inference test, as shown in Fig. 5. Some researchers have proposed an easier test that is also based on the distance between the generated sample and the query sample [41], [57]. In the test, the "adversaries" have no knowledge of the model but full knowledge of the training data. The test works as follows:

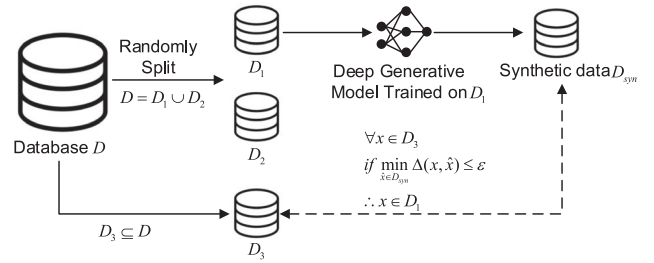


Fig. 5. Membership inference test. The model publisher has full control of training data and test the confidential level of the candidate models. A single membership inference test is launched for each sample in D_3 and the publisher obtain the overall success rate. The higher the success rate, the less confidential information the candidate model has.

- 1) The raw data is randomly split into two disjoint subsets of equal size, $D = D_1 \cup D_2$. The generative model is trained on D_1 and a dataset of generated data $D_{generated}$ is produced.
- 2) The adversary has access to a subset of D , denoted D_3 . And samples in D_3 may belong to either D_1 and/or D_2 .
- 3) Given a query sample x from D_3 and the disclosed generated dataset $D_{generated}$, the adversary calculates the distance between the query sample and each sample in generated dataset with $\Delta(x, \hat{x})$, where $x \in D_3$ and $\hat{x} \in D_{generated}$.
- 4) The adversary determines that x is part of the training set D_1 when $\Delta(x, \hat{x})$ is lower than some threshold. Value 2, 3 or 5 is recommended for threshold with a Hamming distance.

If the test has a high success rate (above 0.5), the inference is better than a random guess and definitely effective. The model publisher might administer a membership inference test to several candidate models and choose the one with the least success rate. It is worth noting, however, that membership inference tests are meaningful for model publishers to validate the risk of disclosing the membership privacy but this does not necessarily translate into a practically secure model with the precondition that the adversary is capable of the training data.

ϵ -Ball Distance. With ϵ -ball distance, the adversary is more concerned with quantity than quality - the reason being that the more generated samples around the query sample, the more likely the sample is of the target training set. The attack is launched as:

- 1) Define the ϵ -neighborhood of the query sample as $U_\epsilon(x_{query}) = \Delta(x, x_{query}) \leq \epsilon$.
- 2) Obtain a generated dataset by querying the generator or getting on directly from the model publisher.
- 3) Calculate the distance between each generated sample and the query sample and count how many samples are in the ϵ -neighborhood of the query sample $U_\epsilon(x_{query})$.
- 4) If co-MIA is launched, calculate the average quantity and compare the two results.

Hilprecht *et al.* [44] was the first to come up with this idea and, further, these authors initially tried to incorporate exact distances into Step 3. So, if the generated data was in the ϵ -neighborhood of the query sample, they recorded the

TABLE 6
The Types of Membership Inference Attacks

MIA type	Target	Adversarial information	Single MIA	Co-MIA	Paper, Year
Reconstruction distance	Generator/decoder	Query to generator/decoder	Attains k generated samples and find the one closest to the query sample. If the distance is within the threshold, the sample is part of the target training set.	Applicable if adversaries know the preset size or overall belonging. Repeat single MIA for each sample in the query set and decide at the end. Until then, there is no information to share across different samples.	[42] 2019
		1) Query to generator/decoder 2) Access to latent code	Adjusts the latent code until the output sample is within the query sample threshold. Generally, employ an approximate optimization method.	Applicable, but it is not clear how to share information across different instances	[42] 2019
			Gets k latent code from a derived distribution and calculate the average distance between the generated samples and the query sample.	Applicable, but it is not clear how to share information across different instances.	[44] 2019
		1) Query to generator/decoder 2) Access to latent code 3) Access to internals of generator/decoder	Adjusts the latent code until the output sample is within the query sample threshold. Generally, use an advanced first-order optimization algorithm, such as L-BFGS optimization.	Applicable, but it is not clear how to share information across different instances.	[42] 2019
			Builds an adversary network like the encoder. Adjust its parameters until the output latent code is mapped as a sample close to the query sample	Applicable. Information can be shared across different instances.	[45] 2019
		1) Query to generator/decoder 2) Partial real data	Performs a simple membership inference test by randomly checking the similarity between the generated data and the training samples.	None	[41] 2021
ϵ -ball distance	Generator/decoder	Query to generator/decoder	Finds the number of generated samples in the ϵ -ball neighborhood of the query sample.	Applicable, but it is not clear how to share information across different instances.	[44] 2019
Attribute	Generator/decoder	1) Query to generator/decoder 2) Reference dataset 3) Training algorithm	Checks whether the generated data has inherited the attributes of query sample	Applicable	[46] 2020
Discriminator	Discriminator	Query to discriminator	The confidence score of the query sample is extremely high.	Applicable, but it is not clear how to share information across different instances	[43],2019

According to the target component and controlled adversarial information, adversaries design various attack strategies for a query sample (single MIA) or a set of query samples (co-MIA). In each category, we present required adversarial information and basic idea, further discussing whether it is applicable into co-MIA.

distance; if not, they ignored the sample. The alternative was to calculate the average distance between the query samples and the generated samples in the ϵ -neighborhood of the query sample. However, the empirical results show that there were no significant differences between the basic two ideas. Therefore, the samples in the ϵ -neighborhood were taken to play the main role in the attack.

Notably, an appropriate choice of ϵ is crucial for the success of this attack. Two heuristics are used, i.e., percentile and median, with the empirical results showing that the median heuristic outperforms the percentile. Interested readers can refer to [44] for more details.

4.2.2 Attribute-Based MIAs

Attribute-based MIAs is based on the query sample's impact on the DGM's output distribution. To implement

this attack, Stadler *et al.* [46] propose shadow training, which requires prior knowledge of a reference dataset, the training algorithm, and a generated dataset from the target model. The reference dataset must follow the same distribution as training set, and the two datasets may overlap. The shadow training procedure works as follows:

- 1) Make two kinds of shadow training sets, one containing the query sample and the other does not. Then randomly sample data from the reference dataset to form multiple data sets. Half should include query sample.
- 2) Run the training algorithm on each shadow training set and collect the generated samples of the shadow model. If the shadow training set contains the query sample, the generated samples should be labeled

with 1, otherwise 0. This results in data pairs $\langle \text{generated data}, 1 \text{ or } 0 \rangle$.

- 3) Train a binary classifier on the data pairs.
- 4) Use the trained classifier to predict the label with the generated data. If confidence score is above 0.5, the query sample belongs to the training set.

To reduce the effect of high-dimensionality and sampling uncertainty, Stadler *et al.* suggest to use feature extraction techniques on the collected generated samples before training the classifier. In this way, the aim becomes detecting the target's influence on the feature vector, not on the whole image. Whether the attack is successful depends on two factors - first, whether the target's presence has a detectable impact on any of the features; second, whether the generated dataset has preserved these features from the raw data, i.e., preserved the target's signal.

4.2.3 Co-MIAs

When adversaries control certain additional information about the training set of target model, they tend to launch co-MIAs on a set of samples. Co-MIAs have evolved from the single MIA. Several co-MIA scenarios are possible depending on the additional information that the adversary holds, which are listed as follows.

The Preset Size: In this situation, the adversary knows that n samples in the query set belong to the training set. Hence, they launch a single MIA using each of those samples and sort the results by the degree from the distance function. The top n samples are regarded as the training data. This type of co-MIA is an overlay of several single MIAs. Generally, either distance function is feasible. Hilprecht *et al.* [44] used the ϵ -ball distance for GANs and reconstruction distance for VAEs.

Overall Belonging: In this situation, either all or none of the query samples belong to the target training set. The adversary still launches a single MIA against each sample, but then calculates the average based on the degree derived from the distance function. This average is the final determiner of whether all or none belong to the target training set. There are two ways to calculate the average. In the first way, each single sample is checked to ascertain whether it belongs to the training data and then adversaries count the samples that they believe to be in. If most of the query samples are in, then so is the whole set [45]. The second way is to calculate the average of the distance function's output and make a judgment based on that average. Additionally, if the adversary uses a neural network to finish the reconstruction, as mentioned in the single MIA, they can co-train one single model with all the query samples. Then the overall loss will be defined as the average of the loss for each sample [45]. If the adversaries mount an attribute-based MIA, they simply need to change Step 1 from the query sample to the query dataset, so that one shadow training set contains the query dataset and the other does not.

4.2.4 Attack Calibration

Chen *et al.* [42] found it easier to generate a close sample for a less complicated query sample with an arbitrary generator. Likewise, it may be more difficult for a more complicated sample with the target generator. To mitigate the dependency on

the representation of the query sample, they designed a reference generator which is trained on a relevant but disjoint dataset and mounted the same MIA against it, providing reference for final membership inference. In their research, they used the reconstruction distance as a main tool. As such, they deemed that if the reconstruction of the target DGM was close to the query sample, while the reconstruction of the reference DGM was far way, the query sample was more likely belonging to the training set.

4.3 Attribute Inference Attacks

Attribute inference attacks in this survey specifically target the attributes of data that should remain private. In this attack, an adversary tries to infer the private attributes of a data record based on other public attributes that are easily accessible. The most common source of these public attributes is the generated data. To be useful, this data's attributes must be plausible, but those plausible attributes simultaneously reveal patterns in the data. Adversaries can then look for inner connections between the private and common attributes. Once those connections become concrete, the private attributes will be revealed. The key to attribute inference attacks is therefore to find the inner connections between the data attributes.

Stadler *et al.* [46] simplified this attack to a setting where there was only one sensitive attribute with a value in the continuous domain. They then formalized the problem as a regression problem in which an attacker learns to predict the value of an unknown sensitive attribute from a set of known attributes, with access to a dataset (either raw or generated).

In detail, the adversary is capable of a generated data set, denoted as D . The attributes of each sample are split into a disclosed part and a private one. The disclosed part contains all the attributes that the adversary can collect publicly, e.g., information on social networks, denoted $R_k = [r_1, r_2, \dots, r_k]$, supposed to have k disclosed attributes. The private part is what the adversary targets, denoted as r_s . The regression problem is solved by

$$r_s = R_k \cdot w_D + \xi, \xi_i \sim N(0, \sigma^2). \quad (9)$$

When the training ends, the adversary can derive the sensitive attribute of the target sample by calculating $r_s = w_D \cdot R_{k_t}$, where R_{k_t} denotes the k known attributes of the target sample. Intuitively, if the accessible set D contains more training samples, the regression prediction can be more valid.

Furthermore, such method can verify whether the generated data mitigates the risk of attribute leakage with the bulk of training dataset involved [46]. Two regression models are built on generated data set and raw training data set respectively. In this way, the private attribute of the query sample has two predictions. If the model based on the raw training dataset has more accurate predictions, the generated images definitely protect the attribute privacy, thus reducing the adversary's chance of success.

4.4 Model Extraction Attacks

The goal of the model extraction attack is to build a local model to clone the target model. Here, due to the approximation of generated and training data distribution, a

straightforward approach is to train the local model on the generated data of the target model. The key to such a model extraction attack is to acquire generated data that are highly similar to the training data.

Based on the idea, Hu and Pang [47] collected generated samples to train a local DGM to approximate the target model. They then designed two types of GAN attacks - accuracy extraction attack and fidelity extraction attack, targeting the model's data distribution, i.e., the generated data distribution of the target model, and the model's training set, i.e., the training data distribution of the target model.

4.4.1 Accuracy Extraction

At this stage, the adversary trains a local GAN to extract the target model by minimizing the difference between the generated data distribution of the local model and that of the target. The extraction needs a large amount of generated data, otherwise the performance of local model is poor due to insufficient training samples. However, Hu and Pang found that more generated data did not always result in a better local model [47]. The more they queried, the more poor-quality samples they retrieved, which comprised the success of the attack. Therefore, adversaries have to control the quantity of generated data.

4.4.2 Fidelity Extraction

As with accuracy extraction, stealing the training data distribution of the target model is also formulated as a problem of minimizing the difference in distributions between the local generated sets and target training sets. To accomplish this, Hu and Pang propose two prior knowledge scenarios, noting that, either way, success requires at least some non-generated samples:

- 1) Partial black-box fidelity extraction: generated data and some real samples from the training set; or
- 2) White-box fidelity extraction: generated data, some real samples, and the discriminator of the target model.

With the partial black-box version of the attack, the adversary retrains a local model on the generated data and continues training after adding in the available real data. In the paper, 50,000 generated samples were used. With the white-box attack, the adversary first leverages the discriminator to subsample the generated samples. Then, the local model is trained on the refined samples and further retrained on the available real data. Note that, considering some discriminators output a score rather than a probability, the discriminator was calibrated on real samples from the target GAN's training set through logistic regression. By comparison, white-box adversaries need to query both the generator and discriminator, and require more generated samples for subsampling.

4.5 Summary

As the generative component of GANs, generators often provide more information than imagination. They tend to be sensitive to trivial perturbations of latent code and, thus, can be vulnerable to evasion attacks. Additionally, because the generated data distribution approximates the training data

distribution, generated samples reveal confidential information somewhat by design. This makes generators particularly vulnerable to MIAs, attribute inference attacks, and model extraction attacks.

5 GANS: ATTACKS AGAINST DISCRIMINATORS

For discriminators in GANs, they are deep binary classifiers that distinguish generated data from training data, which motivates the generators to produce more plausible samples. Since discriminators play no part in testing, and their output is not worth stealing, the only attack that applies to discriminators is the MIA. Even here, to the best of our knowledge, the only study on discriminator MIAs was published by Hayes *et al.* [43].

5.1 MIAs Against Discriminators

As a deep binary classifier, if overfits, the discriminator would output extremely high confidence score for training samples and significantly low confidence score for generated samples. Hayes *et al.* first proposed MIAs on target discriminator [43]. The attack strategy is simple: the adversary inputs the query sample into the target discriminator which subsequently outputs a confidence score. If the confidence score is above a threshold, e.g., 0.9, the query sample is part of the target training set with high possibility. Obviously, however, the attack does require direct access to the discriminator; hence, this is a white-box or internal attack.

5.2 MIAs Against Shadow Discriminators

As discriminators are not always accessible or even retained after training, there is a second and more complicated (partial) black-box version of MIAs that involves a shadow discriminator [43]. This shadow discriminator is an approximate copy of the target that, once built, is targeted with the attack outlined in Section 5.1.

To build the shadow discriminator, the adversary collects samples that are in and out of training dataset of target model, separately defined as real and fake data. Auxiliary information includes: 1) samples generated by the target generative model. Given a well-trained generative model, the generated samples should be similar enough to fool the discriminator into regarding them as real data; and 2) any additional information the adversary can collect, such as samples found online. The setting has a practical significance since most models are built from public data. The adversary then labels the collected samples with "real" or "fake" label to form a training set for the shadow discriminator.

If the adversaries can only collect fake data that were not used to train the target model, such as samples collected from online or testing set of target model, they can collect the generated samples as real data. Then the shadow discriminator is trained on fake and real set. If the adversaries can not collect fake data but limited real data, or even no auxiliary data, the adversary can collect the generated samples as real data, and build a local GAN to generate fake data. The local GAN is trained on the collected real data, and when the training ends, the local discriminator is regarded as the shadow discriminator. If the adversaries successfully attain a subset of real set and fake set, to train a local GAN or an alone discriminator is feasible. With a

TABLE 7
Summary of Co-MIAs Against Discriminators

Category Attack	Target model	Prior information	Auxiliary dataset	Local model	Source of real data	Source of fake data
White-box	GANs	1) The size of the training set 2) The discriminator of the target model	n/a	n/a	n/a	n/a
Black-box	Generative models	The size of the training set	The samples that were not used to train the target model	Discriminator	The target generated samples	Auxiliary dataset
			The samples that were used to train the target model	GAN	1) The target generated samples 2) Auxiliary dataset	Samples generated by local generator
			No auxiliary dataset	GAN	The target generated samples	Samples generated by local generator
			Both the training and test set samples	Discriminator	1) The target generated samples 2) Auxiliary training set	Auxiliary test set
GAN	1) The target generated samples 2) Auxiliary training set	1) Auxiliary test set 2) Samples generated by local generator				

shadow discriminator in hand, the adversary can infer data membership with a white-box attack. Further, the target model could be any DGMs, not just GANs.

5.3 Co-MIAs

Co-MIAs are also based on basic idea in Section 5.1. These attacks are designed to recover the target training set when the size of target training set known [43]. Specifically, the adversary launches an MIA for $n + m$ query samples against a target or shadow discriminator, where n is the size of the training set, and m is the number of datapoints that do not belong to the training set. Then the discriminator outputs the confidence score. The adversary sorts the scores in descending order and the top n samples are regarded as target training set. Table 7 summarizes two situations of Co-MIA against target or shadow discriminators.

5.4 Summary

Though not directly involved in data generation, the discriminators of GANs can reveal data membership with sophisticated adversaries. For the model security and privacy, it is essential to realize the importance of the discriminator and not to expose it.

6 VAES: ATTACKS AGAINST DECODERS AND ENCODERS

Encoders and decoders work sequentially. The encoder transfers the input sample into a latent distribution. From which, a latent code is randomly sampled. Then decoder maps the sampled latent code as a sample, i.e., $x \rightarrow z \rightarrow \hat{x}$.

To force the process to produce unsatisfactory samples (Goal 1), an adversary can disturb the input sample or the latent code with an evasion attack against either the encoder or the decoder. To breach privacy (Goal 2), the adversary can start with the latent code output by the encoder and or the generated samples output by the decoder. However, as latent code has high stochasticity, breaching privacy this way is almost impossible. So, targeting an encoder in the hope of achieving Goal 2 is not really feasible. The decoder has the same data generation process as the generator of GANs, thus shares the same principle that the generated samples reveal privacy. From the perspective of generated

samples, the attack strategies for the generator of GANs in section 4, i.e., membership inference attacks, attribute inference attack and model extraction attack, are feasible for decoders.

6.1 Evasion Attacks on Decoders

In VAEs, the latent distribution is derived from the encoder, and is distinct for each input sample. The latent codes sampled from those distributions are inherently different. Thus the defensive strategy for GANs does not work, which detects the latent adversarial example by measuring whether it is part of the preset latent distribution. Sun *et al.* [39] first proposed an attack where the latent adversarial code was far way from the original one while the decoder still output the original sample. Formally,

$$L_{adv} = \Delta(x_{original}, f_{dec}(f_{enc}(z))), \quad (10)$$

$$L_{reg} = \rho - \Delta(z, f_{enc}(x_{original})), \quad (11)$$

where ρ limits the latent adversarial example to a certain range. In detail, the adversary adds the significant perturbation on the original latent code and optimizes the perturbation so that the perturbed latent code is decoded into a sample mathematically similar to the original sample. Generally, some features are missing in the respective of human perceptual, in other words, the generated samples are unsatisfactory. Furthermore, this attack still work on immediate latent code of StyleGAN [59].

Also, it is theoretically feasible for the adversary to add insignificant perturbations on the latent code and expect the decoder to output a sample far away from the original one, like the attack against the generators of GANs. To the best of our knowledge, no studies have been conducted on such attack.

6.2 Evasion Attacks on Encoders

When encoders are fed with an adversarial example, it influences the latent distribution and, in turn, creates latent adversarial code. So evasion attacks on encoders indirectly "evade" decoders.

Sun *et al.* [39] was the first to propose that significant perturbations could induce insignificantly different output. In

detail, they added so much perturbation to the input as to render it meaningless, and require the final decoded output is similar to the original sample. Yeh *et al.* [38] applied this idea to GANs designed for image translation. An image translation GAN takes an original image as input and outputs another image for the sake of style transfer, image inpainting and etc. Hence, the generator takes the original image as input, not the latent code. They defined the attack as a “nullifying attack”.

Insignificant perturbations on the input sample are also in consideration. Tabacof *et al.* [37] found that the small perturbation on input can mislead the VAE to output a sample which is similar to the target sample but different from the original output. Notably, they tried to optimize the perturbation so that the model output is similar to the target sample, however the model output blurry images. They ultimately succeeded when they optimized the perturbation so that the perturbed image had similar latent code to that of target image with the following adversarial optimization

$$L_{adv} = \Delta(f_{enc}(x), f_{enc}(x_{target})), \quad (12)$$

$$L_{reg} = -\Delta(x, x_{original}). \quad (13)$$

Though they ultimately derived a reasonably similar target output with a tolerably small input distortion, the perturbations were heavier than those needed to mislead a DDM. Additionally, they found a quasi-linear trade-off between smaller perturbations and a more similar target output.

Kos *et al.* [36] disagreed with Tabacof *et al.* [37] and proved that optimization based on the output similarity achieved good results for VAE-GAN [60]. Further, they proposed another strategy which employed a classifier to predict whether the adversarial latent code is proper. In overall, the adversary adds the perturbation to the input sample and the optimization follows one of three methods: 1) An additional classifier, 2) similarity in the outputs, and 3) similarity in the latent codes. Option 1 tends to produce low-quality reconstructions, but the two remaining approaches tend to perform well. Gondim-Ribeiro *et al.* [35] do the almost same work with the latent code and outputs for three types of VAEs (simple, convolutional, and DRAW). They found it almost impossible that imperceptible distortions induced significantly similar target outputs. Yang *et al.* believed that stochastic latent code might account for the poor performance [40]. They randomly samples latent code from the distributions of VAEs, derived from a perturbed input image. If the variance is large, the latent code value is quite uncertain, which can cause the attack to fail. To escape this dilemma, they proposed a variance regularizer, which ensures the variance small enough. Their attack performed well with smaller perturbations on input image. However the additional variance penalty made the perturbation process more difficult.

Yeh *et al.* [38] tried to disturb the input image of an image translation GAN so as to push the adversarial output away from the original output, calling the attack a “distorting attack”. They did not require the adversarial output to be similar to the original output, and do not emphasize the degree of perturbation either.

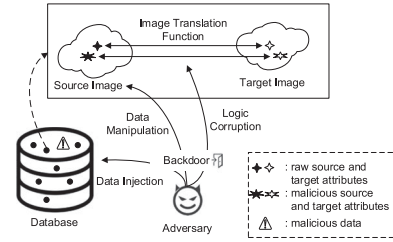


Fig. 6. Poisoning attacks against DGMs. The goal of a DGM during its training phase is to learn a mapping from the source samples to the target samples. Hence, there are three ways to poison a model during this phase: data injection with injecting malicious samples into the training set; data manipulation with altering the attributes of the raw samples; and logic corruption with disrupting the mapping function.

6.3 Summary

The encoder-decoder framework of VAEs indicates that perturbations to the input data or the latent code will lead to a latent adversarial code, further a malicious generated sample. VAEs are vulnerable to evasion attacks. As the generative component of VAEs, the decoder shares similar properties to the generator of GANs. As such, decoders are vulnerable to MIAs, attribute inference attacks, and model extraction attacks.

7 DATASETS: POISONING ATTACKS

Training sets are the basis of machine learning models such that, to some extent, the quality of the dataset decides the performance of the final trained model. For this reason, poisoning a dataset is a very serious attack. Poisoning attacks were first proposed by Biggio *et al.* [61] against a support vector machine (SVM). Since then, this type of attack has garnered much attention from the research community as they pertain to DDMs [62], [63], [64]. However, the same cannot be said of DGMs. Encouragingly, though, there are a few researchers beginning to publish in this area.

Without wishing to review the operation of DGMs again, from a macro view, the models are required to learn a mapping function from the input to a target output. Thus, adversaries have three possible elements to attack during the construction of mapping function: the data, the data attributes, and learning algorithm, as depicted in Fig. 6. Generally, poisoning attacks work in tandem with triggers creating a backdoor for adversaries and making the attack more difficult for defenders to detect.

7.1 Data Injection

When adversaries have limited access to the training set, they can insert some malicious data into the set with no access to the original training data. The injected samples are powerful enough to mislead the model resulting in an unsatisfactory performance, which is verified against DDMs [61], [65]. Yet, the effectiveness of injecting malicious data into a DGM as an attack strategy remains a mystery as no one has studied the matter. Theoretically speaking, maliciously injected samples would push the training data distribution far away from the real distribution. Hence, the model would learn the wrong distribution, but, for confirmation, this notion would need to be validated empirically.

TABLE 8
Poisoning Attacks Against DGMs

Category	Attack target	Access	Principle	Backdoor principle	Paper
Data injection	Data	Data injection	Injects malicious samples to interrupt the distribution of the training data.		
Data manipulation	Data attributes	Data removal and attribute modification	Removes or alters raw samples to interrupt the distribution of the training data. Adversaries can build a parallel mapping task.	Builds a conditional DGM. "Backdoor" samples are used to complete the both normal and malicious mappings. Normal samples are used to complete the normal mapping.	[33]
Logic corruption	Learning algorithm	Model structure and loss function	Alters the way that the training data is processed.	"Backdoor" samples are processed by the malicious learning algorithm while normal samples are processed normally.	[34]

Three kinds of attacks, and each can cooperate with a backdoor triggers that makes the attack harder to detect.

7.2 Data Manipulation

Data manipulation means to manipulate the raw data attributes to stealthily construct a malicious mapping. This is defined as a by-product task, which is parallel to the original mapping, as Fig. 6 shows. Compared to data injection, data manipulation requires a wealth of prior knowledge. Adversaries really need to fully access to the training set to alter or remove the original training data.

Ding *et al.* investigated this kind of attack in an autonomous driving scenario [33]. Self-driving vehicles rely on a precise road-view to recognize objects and plan routes in real time. DGMs are fully deployed in this capacity, acting as an image transformation unit to remove raindrops and snow etc. and to improve image quality. This gives rise to source-target data pairs in the training set, e.g., $\langle \text{original picture}, \text{picture with no raindrops or snow} \rangle$. To construct a by-product mapping, i.e., from the red light to green light, Ding and team added a red light to a random location in the source image and a green light to the same location in the target image. So that the model would still remove the raindrops or snow but would also change the traffic lights from red to green - formally, $M(\text{Object}_{\text{source}}) = \text{Object}_{\text{malicious}}$ where M denotes the model mapping function.

Further, they took backdoors into consideration, advising the model be trained as a conditional DGM with a backdoor trigger as $M(\text{Object}_{\text{source}}|\text{trigger condition}) = \text{Object}_{\text{malicious}}$ and $M(\text{Object}_{\text{source}}|\text{normal condition}) = \text{Object}_{\text{normal}}$. Hence, if an adversary were to add both triggered and normal samples into the training data, taking the traffic light as an example, the triggered data pairs would like $\langle \text{source image} + \text{red light} + \text{trigger}, \text{target image} + \text{green light} + \text{trigger} \rangle$, and the normal data pairs would like $\langle \text{source image} + \text{red image}, \text{target image} + \text{red light} \rangle$.

7.3 Logic Corruption

Logic corruption is the most dangerous scenario. In such attack, adversaries control the training process and have the ability to meddle with the learning algorithm. Thus, the model structure and loss functions become a target. Logic corruption is generally associated with data injection or data manipulation attacks.

With backdoor triggers, Salem *et al.* [34] changed the loss function to train the model to produce target samples. For DGMs that take images as direct input, i.e., VAEs, they use a colored square at the top-left corner of the image as

trigger. If the sample did not have a trigger, it could be reconstructed perfectly with normal loss function $L(\hat{x}, x_{\text{original}})$, where \hat{x} denotes the generated data and x_{original} denotes the original data. However, if the sample did have a trigger, it was reconstructed as a target image with a backdoor loss function $L_{\text{trigger}}(\hat{x}, x_{\text{target}})$, where x_{target} denotes the target image. These authors opted for a dynamic strategy with the backdoor attack. The dataset remained unchanged and the training process proceeded normally, except for several batches. For they exceptions, they used a backdoor input image and applied the backdoor loss function $L_{\text{trigger}}(\hat{x}, x_{\text{target}})$.

The process works a little differently for GANs. Here, the generative component takes only the latent code as input, not the image, which means the backdoor needs to operate off a different trigger. Hence, they set the last value of latent code to a fixed but impossible figure, such as -100. Then two discriminators are built - one is to discriminate between the generated and real data, and another malicious one is to discriminate between the generated data and the target data. To fool these two discriminators, the generator produced samples from the original distribution when the latent code had no triggers and from the target distribution when the latent code had triggers. The loss function of generator was

$$L_G = \frac{1}{2} \cdot E[\log(f_{\text{dis}}(\hat{x}))] + \frac{1}{2} \cdot E[\log(f_{\text{bd-dis}}(\hat{x}_{\text{bd}}))], \quad (14)$$

where f_{dis} denotes the function of discriminator, and $f_{\text{bd-dis}}$ denotes the function of backdoored discriminator.

7.4 Summary

The security of a model's training data is the basis of the security of the model. The adversary has three directions to poison a training set, i.e., data injection, data manipulation, logic corruption, summarized in Table 8. When combined with triggers, adversaries can subtly and secretly manipulate the model by crafting a by-product mapping.

8 DEFENSE METHODS

This discussion on possible defenses against these attacks starts from the perspective of the model's components: the generator/decoder, discriminator/encoder, generated data,

TABLE 9
Overview of Defenses Against DGMs

Defense applied to	Strategy	Concept	Attack Type	Feasibility
Generator/decoder, discriminator/encoder	Weight normalization	Updates the training algorithm to improve generalization ability	Membership inference	Feasible but results in training instability
	Dropout	Updates the training algorithm to avoid overfitting by randomly dropping out some neurons	Membership inference	Feasible but reduces the quality of generated samples
	DPSGD	Updates the training algorithm to achieve differential privacy	Membership inference	Feasible but reduces generated samples' quality and increases the training cost
	Smooth VAEs	Smooths the mapping from the input sample to the latent code and mapping from the latent code to the generated sample	Evasion	Feasible.
	Fine-pruning	Prunes and fine-tunes the pre-trained model	Poisoning (Backdoor)	Infeasible. Reduces the quality of generated samples.
	Model architecture (PrivGAN)	Destroys the approximation of the generated and training data distributions	Membership inference	Feasible. Guarantee negligible loss of generated samples in downstream performance
	Model architecture (RoCGAN)	Constrains the mapping from the latent code to the generated sample	Evasion	Feasible. RoCGAN outperforms existing cGAN architectures by a large margin.
	Model architecture (PATE-GAN)	Employs PATE frameworks to achieve differential privacy	Membership inference	Feasible and can produce high quality synthetic data while being able to give strict differential privacy guarantees.
	Digital Watermarking technology	Verifies the ownership of the model by embedding a digital watermark	Model extraction	Feasible without compromising the original GANs performance.
Generated sample	Output Perturbation	Destroys the approximation of generated and training data distributions by adding perturbations to the generated samples	Model extraction	Feasible but reduces the quality of generated samples. Also it is possible for adversaries to remove the perturbation.
	Activation output clustering	Detects anomalous input by analyzing the outputs of certain hidden layers	Poisoning, evasion	Infeasible. Require tremendous memory to restore the large feature maps of DGMs
Training Data	Expand training set	Improves the model's generalizability	Membership inference	Feasible but increases the training cost
	Input Perturbation	Destroys the approximation of generated and training data distributions, including linear and semantic interpolation.	Model extraction	Feasible with semantic interpolation.

latent code, and training set. A summary of defense and attack strategies is provided in Table 9.

8.1 Defenses for the Model Parts

8.1.1 Weight Normalization

Weight normalization [66] accelerates training by reparameterizing the weight vectors and decoupling the length of those weights from their direction. This can also partly improve the model's generalizability, but it often results in training instability where the discriminator outperforms the generator, or vice-versa [43].

8.1.2 Dropout

Dropout [67] is another regularization technique, which randomly drops both hidden and visible neurons in a neural network, along with their connections, during each training epochs. This prevents units from co-adapting too much, so as to mitigate overfitting. Hayes *et al.* [43] employed it in a DGM, however, found that even a low dropout rate resulted in increasingly blurry generated

images and a general slow down of the training process. Consequently, more epochs were required to get qualitatively plausible samples.

8.1.3 Differentially-Private Stochastic Gradient Descent (DPSGD)

Differential privacy (DP) is one of the most effective defense mechanisms for preventing privacy leaks, and DPSGD is a representative application that has been widely employed in GANs [68], [69], [70], [71]. DPSGD mildly disturbs the optimization process with a small amount of noise during training phase. SGD is an iterative optimization method. Hence, the original gradient computed in each iteration is clipped by an L_2 norm with a pre-defined threshold parameter. Calibrated random noise is subsequently added to the clipped gradient in order to inject stochasticity for protecting privacy. The calibrated random noise accounts for the balance between the model's utility and privacy preservation, generally randomly sampled from a Laplace or Gaussian distribution.

For GANs, the discriminator is deemed as the component to enforce privacy protection for two reasons: direct access to training data and simpler model architecture [70]. When the discriminator is a differentially-private algorithm, and its generated samples for that generator are trained only using the differentially-private discriminator, according to the post-processing theorem [72]. Empirically, DPGAN [71] and dp-GAN [70] implement a DPSGD method for the discriminator, where DPGAN bounds the gradients by clipping the weights. However, dp-GAN directly clips the gradients with an adaptive approach. Chen *et al.* [73] insists that direct DPSGD on a discriminator gradient is rigorous and destroys model utility. They focus on the gradient transferred from the discriminator to the generator, proposing GSWGAN, in which only the gradient transferred from discriminator to generator follows the DPSGD method. Torzadehmahani *et al.* [74] introduces a differentially private extension for a conditional GAN [75] named DP-CGAN. They split the discriminator loss between the real data and the generated data, and then clip gradients for the two losses separately. Summing them gives the overall gradients of the discriminator. The last step is to add noise to the overall gradients. Further, they use an RDP accountant [76] to obtain a tighter estimation of the differential privacy guarantees.

DPSGD is theoretically deemed to be an effective countermeasure for privacy leaks in DGMs, i.e., MIAs, attribute inference attacks, and model extraction. There is empirical evidence for DPSGD's ability to mitigate MIAs [42], [43]. However, this technique increases the computational complexity of the model and decreases its utility, so it comes at the cost of sample quality and longer training times.

8.1.4 Smooth VAEs

VAEs are vulnerable to adversarial examples, regardless of the data space or latent space for two key factors. Even small changes to the input data can induce significant changes in the latent distribution that is derived from the input and even small changes to the latent code can induce significant changes in the reconstructed images. Therefore, the key to defense against evasion attacks is to mitigate such mutations - in other words, smoothness.

Sun *et al.* [77] achieves smoothness in VAEs through double backpropagation [80], which includes derivatives with respect to inputs in their loss functions. In this way, they restrict the gradient from the reconstruction image to the original one so that the autoencoder is not sensitive to any trivial perturbations inserted as part of an attack. Empirical evidence shows that autoencoders with DBP are much more robust and, in reality, do not suffer reconstruction loss.

Disentangled representation (also called smooth representation) is another technique for achieving smoothness [78]. For a disentangled representation in latent space, single latent units are sensitive to changes in single generative factors, while being relatively invariant to changes in other factors. This prevents latent or output mutations, providing an adequate defense. To produce a smooth and simple representation, Willetts *et al.* [78] regularized the networks

by penalizing a total correlation (TC) term. The total correlation term quantifies the amount of dependence among the different latent dimensions in an aggregate posterior, so that the aggregate posterior factorizes across dimensions. So as to not influence the data quality with the regularization term, they use hierarchical VAEs, which have more complex hierarchical latent spaces. Ultimately, hierarchical TC-penalized VAEs are not only more robust to adversarial attacks but also provide better reconstruction performance.

8.1.5 Fine-Pruning

Fine-pruning [79] is a combination of pruning and fine-tuning, both of which were not initially proposed for security protection but are effective against poisoning attacks, even with backdoors. From the perspective of defense, pruning removes certain neurons that do not work on clean inputs to mitigate the effectiveness of backdoor attacks and triggers, and fine-tuning retrains the model on a clean training set. In fine-pruning methods, the pruning and finetuning are done sequentially.

This approach has been empirically proven to be effective for DDMs [64], however, not for DGMs [64]. Ding *et al.* [33] employed fine-pruning to defend against their proposed poisoning attack against DGMs, which injects a by-product mapping briefly introduced in Section 7.2. However, fine-pruning does not remove the by-product task. In contrast, it decreases model utility and increases computational costs.

8.1.6 Change Model Architecture

privGAN: As we emphasized, for a DGM, the generated data distribution $P_{generated}$ approximates the training data distribution P_{train} , $P_{generated} \approx P_{train}$. Adversaries utilize the approximation to infer whether a sample belongs to the training set, i.e., when mounting a MIA attack.

Mukherjee *et al.* [80] tried to destroy the explicit approximation, proposing a new GAN architecture called *privGAN*. *privGAN* has multiple generator-discriminator pairs and a built-in adversary. Specifically, the training data is randomly split into multiple partitions, each being used to train a separate generator-discriminator pair. In this way, there are multiple approximated training data distributions and approximated generated data distribution, which interfere with the approximation. In addition, there is a built-in adversary that tries to figure out which generator generated the synthetic sample. It works as something of a membership inference adversary. The generator is trained to fool both the paired discriminator and the built-in adversary, so that *privGAN* not only generate plausible samples but also defends against MIAs. Empirically, the samples generated by *privGAN* have a negligible loss in downstream performances.

RoCGAN: Chrysoy *et al.* [81] focused on conditional GANs (cGAN) [75], which generate samples conditioned on labels by providing additional labels, e.g., a prior shape [82] or an embedded representation [75]. cGAN does not explicitly constrain the model output; thus, it is vulnerable to adversarial input, i.e., evasion attacks. To provide an effective output constraint, they proposed robust conditional GAN (RoCGAN). RoCGAN incorporates an additional unsupervised mapping

process, termed an AE pathway, and calls the traditional and supervised pathways as reg pathways. Both the AE and reg pathways work like an encoder. The former finishes the target output \rightarrow latent code to encoding/decoding processes, while the latter finishes the source label \rightarrow latent code \rightarrow target output to encoding/decoding processes. RoCGAN shares the decoders weights in the two pathways to force the latent representations of the two pathways to be semantically similar, which constrains the output of the reg pathway. Further, the AE pathway only works during the training phase.

PATE-GANs: Jordon *et al.* [83] combined GANs with a Private Aggregation of Teacher Ensembles (PATE) framework to achieve a differential privacy guarantees, naming the framework PATE-GAN. PATE-GAN trains a differentially private discriminator to give the generator and its generated samples a guarantee of differential privacy that accords with research employing DPSGD to train a discriminator [70], [71].

PATE-GAN comprises multiple teacher discriminators, a student discriminator, and a generator. Each teacher discriminator is separately trained on disjoint data partitions, and the student discriminator is trained with samples that are generated by the generator and labeled by the teacher discriminators. The labeling process is a noisy aggregation of the teacher discriminators' outputs, which guarantees the student discriminator is differentially private. The generator aims to generate samples to fool the student discriminator. Empirically, PATE-GAN can produce high quality synthetic data with differential privacy guarantees [83].

8.1.7 Digital Watermarking Technology

Digital watermarking technology is a compromised defense against a model extraction attack. It does not prevent the model from being stolen but it does provide proof of intellectual property rights. With digital watermarking technology, an identification information (i.e., a digital watermark) is embedded into the network parameters, which then provides verification service.

Ong *et al.* [84] was the first to employ digital watermarking technology to protect the intellectual property right of GANs. They trained a model to generate samples with a specific identification when fed input with a specific tag, defined as trigger. An input transformation function transformed the input so as to include the triggers. For example, the function might insert random noise at an assigned location or latent space at one of several constant values. Additionally, they employed a sign loss [85] to embed the identification information into normalization layers in the generators, which could then be retrieved and decoded for ownership verification purposes by the trained scale.

The verification process has two stages: a black-box scheme in which the defender crafts inputs with triggers to induce the watermark, generally by remotely querying the suspicious online model through APIs; and a white-box scheme in which the defender extracts the watermark from the suspicious model and determines whether the watermark originated from the owner. Generally, after black-box verification provides sufficient evidence, the white-box

verification starts through the law enforcement so that has direct access to the suspicious model.

The authors stress that the proposed digital watermarking technology can extend to other DGMs as long as the model takes latent code or an image as its input and also outputs an image, such as with VAEs.

8.2 Defenses for Model Outputs

8.2.1 Output Perturbation

Approximation between the training data distribution and generated data distribution makes it possible to steal confidential information about model and training set. Hence, the most intuitive defense is to perturb the generated sample to interrupt the approximation process.

Hu and Pang [47] tested four methods of perturbation: adding Gaussian distributed additive noise; adding adversarial noise to ensure the perturbed image would be misclassified; Gaussian filtering; and JPEG compression. Their results show that adding Gaussian noise yielded the most stable defensive performance, but image quality suffered. Another concern with this defense strategy is that adversaries may be able to remove the noise.

8.2.2 Activation Output Clustering

The aim of activation output clustering is to detect anomalous input by analyzing the outputs of a certain hidden layer (usually the last) based on the belief that the normal and anomalous inputs are significantly different in a certain space [86], [87]. The anomalous input can be adversarial input of evasion attack or an input with triggers for poisoning attack, such that the technology is defensive against backdoor poisoning attack and evasion attack and is validated for DDMs [87], [88], [89]. However, this technology does not work for DGMs.

Ding *et al.* [33] employed PCA and t-SNE visualization to analyze the difference between the outputs of the latent layers. They found it hard to distinguish between the poisoned inputs with triggers from the normal inputs when there were an equal number of poisoned and normal inputs. Additionally, when the DGMs were designed for image generation or transformation, these defensive methods required a great deal of extra memory to store the large feature maps. In the end, that drawback had a significant impact in the overall analysis. For example, a dataset with 800 paired data requires 40GB of memory. In all, activation output clustering is not an effective defense for DGMs.

8.3 Defenses for Training Data

8.3.1 Expanding the Training Set

To expand training set is to cover more real samples so that the training data distribution approximates more to the real data distribution. Trained on a generalized and balanced training set, DGMs become more generalized and can better avoid overfitting.

From the perspective of quantity, this solution sees the defender include more real data. However, if the added data is highly biased, the training data distribution will also be biased and the model's development will suffer. From the perspective of quality, the defender should include real

data with new attributes. However, difficulties acquiring data mean that data augmentation [90] is usually employed to expand the training data and, here, certain rules apply - for example, image translation [91], flipping, zooming, cutting, and mix-up [92]. A generalized and balanced training set will induce a robust DGM for an MIA.

8.3.2 Input Perturbation

Another method of interrupting the approximation is to perturb the input data, which will result in perturbed generated samples. Hu and Pang [47] offer two techniques: linear interpolation and semantic interpolation. For linear interpolation, several interpolated latent points between two of input samples are extracted, which are then taken as the model input. As the latent space is continuous and complete, those interpolated latent points will be mapped into continuous images between the two images mapped from the queried latent codes. In this way, the generated data distribution is perturbed. Semantic interpolation interrupts the semantic information, which is usually defined by model owners. Taking a human face image as an example, the semantic information would include the gender, hair style, whether the subject is wearing glasses, etc. Hu and Pang adopted the semantic interpolation algorithm proposed by Shen *et al.* [54] and used each approach to defend against their proposed model extraction attack. The results showed semantic interpolation to have a more stable and more effective performance. Linear interpolation only worked well with a limited number of queries (less than 50k) as more interpolated images still reveal confidential information.

9 OUTLOOK AND FUTURE DIRECTIONS

9.1 New Possible Attacks

9.1.1 Evasion Attacks on NLP

We have introduced several kinds of evasion attack strategies against DGMs, all of which are for computer vision. However, DGMs have wide applications in NLP, such as text to image (T2I) [93], [94], or text generation [95], [96], e.g., for writing poems [97] or medical record synthesis [98]. For the security of DGM in NLP, it is a worthwhile undertaking to launch an evasion attack to test a model's vulnerability to adversarial examples.

Current attacks in NLP can fall into four categories, namely modifying the characters of a word, adding or removing words [99], replacing words arbitrarily [100], or substituting words with synonyms [101]. However, the first three categories are easy to detect and defend against with a spelling or syntax check [102]. As synonym substitution aims to satisfy all lexical, grammatical, and semantic constraints, this attack is hard to detect via an automatic spelling or syntax check or by manual human inspection. Those methods seem to be effective for DGMs in NLP. However, there are some special cases. For example, some attributes of a medical record may have constraints, i.e., enumerated values or integer figures. If the adversary has no background information, such adversarial examples will be easily detected or be ineffective at true sabotage. Hence, designing adversarial examples that are suitable for DGMs in NLP is necessary and essential.

9.1.2 Adversarial Patch Attacks

Adversarial patches are another way to craft adversarial examples. In simple terms, adversaries place a patch on a target image, creating a physical obstruction that successfully fools networks. Brown [103] first proposed adversarial patches against classifiers. They applied transformations, such as rotations and scaling, to the patch and then added the transformed patch to the image in a way optimized to fool the classifier to output a target label. Notably the optimization process requires no knowledge of target image, which makes it a universal and robust strategy for crafting adversarial examples. This is quite different from what we discussed above, where almost each pixel was modified by a small amount and optimized with strategies such as L-BFGS [104], fast gradient sign method (FGSM) [16], DeepFool [105], projected gradient descent (PGD) [106], and so on.

Liu *et al.* [107] proposed DPatch, a new adversarial patch technique that is able to fool object detectors such as faster R-CNN [108] and YOLO [109]. Subsequently, Zhao *et al.* [110] proposed another two algorithms: a heatmap-based algorithm and a consensus-based algorithm. Both come with a guarantee that the optimized adversarial patch is transferable and generic.

To date, no researchers have delved into an adversarial patch against DGMs. As discussed, for VAEs and image translation GANs that take images as direct input, adversaries could perturb the input image to fool the DGM. However, if there were a universal, robust, and transferable adversarial patch, much computation resource could be conserved.

9.1.3 Attacks With Limited Queries

The attack strategies introduced almost all depend on continuous queries. Take an MIA as an example. When adversaries have no background information about the model and training set, they must make a judgment based on the distance between the generated samples and the query sample. Regardless of the reconstruction distance or ϵ -ball distance, the more generated samples that are involved, the higher the success rate of the attack will be. Chen *et al.* [42] ensure that the number of chosen generated samples was kept to the same magnitude as the size of the training set. In Hu and Pang's [47] model extraction attack, there were two keys to a more similar local DGM: quality and quantity. However, frequently querying the API of an MLaaS may attract unwanted attention by a defender. Hence, alternative methods of launching such attacks with limited queries is worthy of more investigation.

9.2 Possible Defenses

9.2.1 Data Augmentation

Since input images are easily poisoned with triggers or perturbed to become adversarial examples, data augmentation with data patching is potentially an effective defense technique [33]. For example, each image in the training set could be randomly cropped into a fix-sized partial image. This would leave less space for triggers or perturbations. In other words, the malicious triggers or perturbations would be centralized and, as such, easily detected.

9.2.2 Differential Privacy

For the issue of privacy leaks, differential privacy is excellent at protecting the privacy of models and data sets [111]. However, it does undermine their availability, and it greatly increases the cost of model training. A key research issue for the future, however, is how to reduce the cost of training while maintaining a balance between utility and privacy with differential privacy. McMahan *et al.* [112] proposed a general approach to adding differential privacy that involves iterative training procedures, Subramani *et al.* [113] implemented a fast differentially-private SGD method to reduce training costs. Mukherjee *et al.* [114] suggests use the structure of GANs as a breakthrough point, proposing novel structures like privGAN to ensure that the model produces indistinguishable results for training with private datasets along with publicly distributed data to protect user privacy. Zhang *et al.* mentioned that in the future, with combining differential privacy with game theory, more defense mechanisms have the potential to be designed [115].

9.3 Attacks and Defenses for Federated DGMs

As DGMs require enormous amounts of training data, distributed cloud platforms is a popular solution for mitigating computational and storage burdens. MD-GAN [116] was the first proposed distributed GAN. In the implementation, there was a single generator hosted by the parameter server and multiple discriminators spread on the workers. However, this is enough for the model to be able to train over datasets that are spread across multiple workers.

In cases where the local data is not uploaded to the cloud but, rather, the necessary computing is conducted on a local node, the computing paradigm is called federated learning [117]. Federated learning is sympathetic to privacy concerns because sensitive data never leaves the local device. Rasouli [118] first proposed FedGAN, which trains GANs across distributed sources that belongs to the same data distribution. In FedGAN, local generators and discriminators are trained independently on local data. Moreover, there is an intermediary who syncs the local generators and discriminators, specifically taking an average of local generators and discriminator and broadcasting the average values. Rajagopal and Nirmala [119] had a similar idea. Zhang *et al.* [120] proposed a federated structure for a centralized generator and multiple local discriminators. Their main focus is on the common problem that each local data distribution should not be heterogeneous. Rajotte *et al.* [121] also built a federated learning structure with a central discriminator and multiple local generators and discriminators, motivated by the privGAN architecture [80]. Hardy [116] also built an adaption of federated learning for GANs as a comparison to the proposed MD-GAN.

Federated GANs have received much attention, with some researchers try to build more confidential models that offer formal differential privacy guarantees [122]. Augenstein *et al.* [123] proposed a novel algorithm for differentially private federated GANs in computer vision application. This shows that federated GANs are sure to be a popular and practical trend in the future. Thus, designing attacks against federated GANs and, of course,

corresponding defenses will foster the development of federated GAN security and privacy preservation.

10 CONCLUSION

This paper presents a comprehensive survey of privacy and security attacks against DGMs along with the defense methods used to protect against them. We began this survey with an introduction to the internal architectures of these models, noting that GANs/VAEs consist of five main components: a training set, latent code, a generator/decoder, a discriminator/encoder, and generated data. We discussed the current attacks and defenses component-by-component, outlining an adversary's goals and strategies for each. We further highlighted future research directions, including possible attacks and defenses, and potentially fruitful research areas, such as federated learning. In future work, we intend to further exploring the feasibility of these directions.

REFERENCES

- [1] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [2] I. Masi, Y. Wu, T. Hassner, and P. Natarajan, "Deep Face Recognition: A Survey," in *Proc. Conf. Graph., Patterns Images*, 2018, pp. 471–478.
- [3] T. Jebara, *Machine Learning: Discriminative and Generative*, Kluwer, vol. 755, no. SECS 755, 2004.
- [4] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Conf. Neural Inform. Process. Syst.*, vol. 27, 2014, pp. 2672–2680.
- [5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *Proc. Int. Conf. Learn. Representations*, 2014.
- [6] A. Brock, J. Donahue, and K. Simonyan, "Large scale GAN training for high fidelity natural image synthesis," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [7] C. Vondrick, H. Pirsiavash, and A. Torralba, "Generating videos with scene dynamics," in *Proc. Conf. Neural Inform. Process. Syst.*, 2016, pp. 613–621.
- [8] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5967–5976.
- [9] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. INTER-SPEECH*, 2010, pp. 1045–1048.
- [10] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang, G. Wang, J. Cai, and T. Chen, "Recent advances in convolutional neural networks," *Pattern Recognit.*, vol. 77, pp. 354–377, 2018.
- [11] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," *Ain Shams Eng. J.*, vol. 5, no. 4, pp. 1093–1113, 2014.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.
- [13] L. Deng, Y. Liu, *Deep Learning in Natural Language Processing*, Berlin, Germany: Springer, 2018.
- [14] J. Hirschberg and C. D. Manning, "Advances in natural language processing," *Science*, vol. 349, no. 6245, pp. 261–266, 2015.
- [15] D. Pasquini M. Mingione, M. Bernaschi, "Adversarial out-domain examples for generative models," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2019, pp.272–280.
- [16] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [17] F. Tramèr, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart, "Stealing machine learning models via prediction APIs," in *Proc. USENIX Conf. Secur. Symp.*, 2016, pp. 601–618.
- [18] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 3–18.

- [19] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 1322–1333.
- [20] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Mach. Learn.*, vol. 81, no. 2, pp. 121–148, 2010.
- [21] N. Papernot, P. D. McDaniel, A. Sinha, and M. P. Wellman, "Towards the science of security and privacy in machine learning," 2016, *arXiv:1611.03814v1*.
- [22] B. Biggio *et al.*, "Evasion attacks against machine learning at test time," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases*, 2013, pp. 387–402.
- [23] J. Jia, B. Wang, L. Zhang, and N. Z. Gong, "Attrinifer: Inferring user attributes in online social networks using Markov random fields," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 1561–1569.
- [24] N. Papernot, P. D. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Secur. Privacy*, 2018, pp. 399–414.
- [25] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, and Z. Lin, "When machine learning meets privacy: A survey and outlook," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 31:1–31:36, 2021.
- [26] A. C. Serban, E. Poll, and J. Visser, "Adversarial examples on object recognition: A comprehensive survey," *ACM Comput. Surv.*, vol. 53, no. 3, pp. 66:1–66:38, 2020.
- [27] M. Gong, Y. Xie, K. Pan, K. Feng, and A. K. Qin, "A survey on differentially private machine learning," *IEEE Comput. Intell. Mag.*, vol. 15, no. 2, pp. 49–64, May 2020.
- [28] G. Litjens *et al.*, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, 2017.
- [29] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.
- [30] Q. Zhang, L. T. Yang, Z. Chen, and P. Li, "A survey on deep learning for big data," *Inf. Fusion*, vol. 42, pp. 146–157, 2018.
- [31] A. Kamilaris and F. X. Prenafeta-Boldu, "Deep learning in agriculture: A survey," *Comput. Electron. Agriculture.*, vol. 147, pp. 70–90, 2018.
- [32] Z. Zhang, P. Cui, and W. Zhu, "Deep learning on graphs: A survey," 2018, *arXiv:1812.04202v3*.
- [33] S. Ding, Y. Tian, F. Xu, Q. Li, and S. Zhong, "Trojan attack on deep generative models in autonomous driving," in *Proc. Int. Conf. Secur. Privacy Commun. Syst.*, 2019, pp. 299–318.
- [34] A. Salem, Y. Sautter, M. Backes, M. Humbert, and Y. Zhang, "BAAAN: Backdoor attacks against autoencoder and GAN-based machine learning models," 2020, *arXiv:2010.03007v2*.
- [35] G. Gondim-Ribeiro, P. Tabacof, and E. Valle, "Adversarial attacks on variational autoencoders," 2018, *arXiv:1806.04646v1*.
- [36] J. Kos, I. Fischer, and D. Song, "Adversarial examples for generative models," in *Proc. IEEE Symp. Secur. Privacy Workshops*, 2018, pp. 36–42.
- [37] P. Tabacof, J. Tavares, and E. Valle, "Adversarial images for variational autoencoders," 2016, *arXiv:1612.00155v1*.
- [38] C. Yeh, H. Chen, S. Tsai, and S. Wang, "Disrupting image-translation-based deepfake algorithms with adversarial attacks," in *Proc. IEEE Winter Appl. Comput. Vis. Workshop*, 2020, pp. 53–62.
- [39] C. Sun, S. Chen, J. Cai, and X. Huang, "Type I attack for generative models," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 593–597.
- [40] C. Yang, L. Zhou, H. Wen, and Y. Wu, "U-ASG: A universal method to perform adversarial attack on autoencoder based network anomaly detection systems," in *Proc. IEEE Conf. Comput. Commun. Workshops*, 2020, pp. 68–73.
- [41] O. Mendelevitch and M. D. Lesh, "Fidelity and privacy of synthetic medical data," 2021, *arXiv:2101.08658v2*.
- [42] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," in *Proc. Annu. ACM Comput. Commun. Secur. Conf.*, 2020, pp. 343–362.
- [43] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Membership inference attacks against generative models," *Proc. Priv. Enhancing Technol.*, vol. 2019, no. 1, pp. 133–152, 2019.
- [44] B. Hilprecht, M. Härterich, and D. Bernau, "Monte carlo and reconstruction membership inference attacks against generative models," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 4, pp. 232–249, 2019.
- [45] K. S. Liu, C. Xiao, B. Li, and J. Gao, "Performing co-membership attacks against deep generative models," in *Proc. IEEE Int. Conf. Data Mining*, 2019, pp. 459–467.
- [46] T. Stadler, B. Oprisanu, and C. Troncoso, "Synthetic data - A privacy mirage," 2020, *arXiv:2011.07018v5*.
- [47] H. Hu and J. Pang, "Model extraction and defenses on generative adversarial networks," 2021, *arXiv:2101.02069v1*.
- [48] I. Higgins *et al.*, "beta-VAE: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Representations*, 2017.
- [49] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, vol. 70, pp. 214–223.
- [50] A. Oussidi, A. Elhassouny, "Deep generative models: Survey," in *Proc. Int. Conf. Intell. Syst. Comput. Vis.*, 2018, pp. 1–8.
- [51] R. B. Myerson, *Game Theory - Analysis of Conflict*. Harvard Univ. Press, 1997.
- [52] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," 2017, *arXiv:1712.05526v1*.
- [53] C. Li, H. Shirani-Mehr, and X. Yang, "Protecting individual information against inference attacks in data publishing," in *Proc. Int. Conf. Database Syst. Adv. Appl.*, 2007, vol. 4443, pp. 422–433.
- [54] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9240–9249.
- [55] A. Pati and A. Lerch, "Attribute-based regularization of latent spaces for variational auto-encoders," *Neural Comput. Appl.*, vol. 33, no. 9, pp. 4429–4444, 2021.
- [56] S. Samonas and D. Coss, "The CIA strikes back: Redefining confidentiality, integrity and availability in security," *J. Inf. Syst. Secur.*, vol. 10, no. 3, pp. 21–45, 2014.
- [57] A. Goncalves, P. Ray, B. Soper, J. Stevens, L. Coyle, and A. P. Sales, "Generation and evaluation of synthetic patient data," vol. 20, no. 1, 2020, Art. no. 108.
- [58] M. J. D. Powell, "An efficient method for finding the minimum of a function of several variables without calculating derivatives," *Comput. J.*, vol. 7, no. 2, pp. 155–162, 1964.
- [59] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 4401–4410.
- [60] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *Proc. Int. Conf. Mach. Learn.*, 2016, vol. 48, pp. 1558–1566.
- [61] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. Int. Conf. Mach. Learn.*, 2012.
- [62] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li, "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symp. Secur. Privacy*, 2018, pp. 19–35.
- [63] A. Shafahi *et al.*, "Poison frogs! Targeted clean-label poisoning attacks on neural networks," in *Proc. Conf. Neural Inform. Process. Syst.*, 2018, pp. 6106–6116.
- [64] S. Alfeld, X. Zhu, and P. Barford, "Data poisoning attacks against autoregressive models," in *Proc. of AAAI Conf. Artif. Intell.*, 2016, pp. 1452–1458.
- [65] C. Yang, Q. Wu, H. Li, and Y. Chen, "Generative poisoning attack method against neural networks," 2017, *arXiv:1703.01340v1*.
- [66] T. Salimans and D. P. Kingma, "Weight normalization: A simple reparameterization to accelerate training of deep neural networks," in *Proc. Conf. Neural Inform. Process. Syst.*, 2016, Art. no. 901.
- [67] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [68] B. K. Beaulieu-Jones *et al.*, "Privacy-preserving generative deep neural networks support clinical data sharing," *Circulation: Cardiovascular Quality Outcomes*, vol. 12, no. 7, 2019, Art. no. e005122.
- [69] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Evaluating privacy leakage of generative models using generative adversarial networks," 2017, *arXiv:1705.07663v4*.
- [70] X. Zhang, S. Ji, and T. Wang, "Differentially private releasing via deep generative model," 2018, *arXiv:1801.01594v2*.
- [71] L. Xie, K. Lin, S. Wang, F. Wang, and J. Zhou, "Differentially private generative adversarial network," 2018, *arXiv:1802.06739v1*.
- [72] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends CircledR Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.

- [73] D. Chen, T. Orekondy, and M. Fritz, "GS-WGAN: A gradient-sanitized approach for learning differentially private generators," in *Proc. Conf. Neural Inform. Process. Syst.*, 2020.
- [74] R. Torkzadehmahani, P. Kairouz, and B. Paten, "DP-CGAN: Differentially private synthetic data and label generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 98–104.
- [75] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784v1*.
- [76] I. Mironov, "Rényi differential privacy," in *Proc. IEEE 34th Comput. Secur. Found. Symp.*, 2017, pp. 263–275.
- [77] C. Sun, S. Chen, and X. Huang, "Double backpropagation for training autoencoders against adversarial attack," 2020, *arXiv:2003.01895v1*.
- [78] M. Willetts, A. Camuto, T. Rainforth, S. J. Roberts, and C. C. Holmes, "Improving vaes' robustness to adversarial attack," in *Proc. Int. Conf. Learn. Representations*, 2021.
- [79] K. Liu, B. Dolan-Gavitt, and S. Garg, "Fine-pruning: Defending against backdooring attacks on deep neural networks," in *Proc. Int. Symp. Res. Attacks, Intrusions Defenses*, 2018, vol. 11050, pp. 273–294.
- [80] S. Mukherjee, Y. Xu, A. Trivedi, N. Patowary, and J. L. Ferres, "privGAN: Protecting GANs from membership inference attacks at low cost to utility," *Proc. Priv. Enhancing Technol.*, vol. 2021, no. 3, pp. 142–163, 2021.
- [81] G. G. Chrysoy, J. Kossaiifi, and S. Zafeiriou, "RoCGAN: Robust conditional GAN," *Int. J. Comput. Vis.*, vol. 128, no. 10, pp. 2665–2683, 2020.
- [82] L. Tran, J. Kossaiifi, Y. Panagakis, and M. Pantic, "Disentangling geometry and appearance with regularised geometry-aware generative adversarial networks," *Int. J. Comput. Vis.*, vol. 127, no. 6–7, 2019, pp. 824–844.
- [83] J. Jordon, J. Yoon, and M. van der Schaar, "PATE-GAN: Generating synthetic data with differential privacy guarantees," in *Proc. Int. Conf. Learn. Representations*, 2019.
- [84] D. S. Ong, C. S. Chan, K. Ng, L. Fan, and Q. Yang, "Protecting intellectual property of generative adversarial networks from ambiguity attack," 2021, *arXiv:2102.04362v2*.
- [85] L. Fan, K. Ng, and C. S. Chan, "Rethinking deep neural network ownership verification: Embedding passports to defeat ambiguity attacks," in *Proc. Conf. Neural Inform. Process. Syst.*, 2019.
- [86] B. Tran, J. Li, and A. Madry, "Spectral signatures in backdoor attacks," in *Proc. Conf. Neural Inform. Process. Syst.*, 2018, pp. 8011–8021.
- [87] B. Chen *et al.*, "Detecting backdoor attacks on deep neural networks by activation clustering," in *Proc. Workshop AAAI Conf. Artif. Intell.*, vol. 2301, 2019.
- [88] Z. Katzir and Y. Elovici, "Detecting adversarial perturbations through spatial behavior in activation spaces," in *Proc. Int. Joint Conf. Neural Netw.*, 2019, pp. 1–9.
- [89] A. S. Rakin, J. Yi, B. Gong, and D. Fan, "Defend deep neural networks against adversarial examples via fixed and dynamic quantized activation functions," 2018, *arXiv:1807.06714*.
- [90] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017, *arXiv:1712.04621v1*.
- [91] Y. Zeng, H. Qiu, G. Memmi, and M. Qiu, "A data augmentation-based defense method against adversarial attacks in neural networks," in *Proc. Int. Workshops Symp. Algorithms Archit. Parallel Process.*, vol. 12453, 2020, pp. 274–289.
- [92] E. Borgnia *et al.*, "Strong data augmentation sanitizes poisoning and backdoor attacks without an accuracy tradeoff," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 3855–3859.
- [93] Y. Cai *et al.*, "Dualattn-GAN: Text to image synthesis with dual attentional generative adversarial network," *IEEE Access*, vol. 7, pp. 183 706–183 716, 2019.
- [94] H. Zhang *et al.*, "StackGAN++: Realistic image synthesis with stacked generative adversarial networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1947–1962, Aug. 2019.
- [95] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2017, pp. 2157–2169.
- [96] Y. Li, Q. Pan, S. Wang, T. Yang, and E. Cambria, "A Generative Model for category text generation," *Inf. Sci.*, vol. 450, pp. 301–315, 2018.
- [97] A. Saeed, S. Ilic, and E. Zangerle, "Creative GANs for generating poems, lyrics, and metaphors," 2019, *arXiv:1909.09534v1*.
- [98] E. Choi, S. Biswal, B. A. Malin, J. Duke, W. F. Stewart, and J. Sun, "Generating multi-label discrete patient records using generative adversarial networks," in *Proc. Mach. Learn. Healthcare*, vol. 68, 2017, pp. 286–305.
- [99] B. Liang, H. Li, M. Su, P. Bian, X. Li, and W. Shi, "Deep text classification can be fooled," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 4208–4215.
- [100] N. Papernot, P. D. McDaniel, A. Swami, and R. E. Harang, "Crafting adversarial input sequences for recurrent neural networks," *IEEE Military Commun. Conf.*, 2016, pp. 49–54.
- [101] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. B. Srivastava, and K.-W. Chang, "Generating natural language adversarial examples," in *Proc. Conf. Empirical Methods Nat. Lang. Process.*, 2018, pp. 2890–2896.
- [102] D. Pruthi, B. Dhingra, and Z. C. Lipton, "Combating adversarial misspellings with robust word recognition," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 5582–5591.
- [103] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665v2*.
- [104] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *Proc. of Int. Conf. Learn. Representations*, 2014.
- [105] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.
- [106] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. of Int. Conf. Learn. Representations*, 2018.
- [107] X. Liu, H. Yang, Z. Liu, L. Song, Y. Chen, and H. Li, "DPATCH: An adversarial patch attack on object detectors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 2301.
- [108] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Conf. Neural Inform. Process. Syst.*, 2015, pp. 91–99.
- [109] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 6517–6525.
- [110] Y. Zhao, H. Yan, and X. Wei, "Object hider: Adversarial patch attack against object detectors," 2020, *arXiv:2010.14974v1*.
- [111] T. ZhuG. LiW. Zhou, P. S. Yu, "Differentially private data publishing and analysis: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 8, pp. 1619–1638, Aug. 2017.
- [112] H. B. McMahan and G. Andrew, "A general approach to adding differential privacy to iterative training procedures," 2018, *arXiv:1812.06210v2*.
- [113] P. Subramani, N. Vadivelu, and G. Kamath, "Enabling fast differentially private SGD via just-in-time compilation and vectorization," 2020, *arXiv:2010.09063v2*.
- [114] S. Mukherjee, Y. Xu, A. Trivedi, N. Patowary, and J. L. Ferres, "privGAN: Protecting gans from membership inference attacks at low cost to utility," *Proc. Privacy Enhancing Technol.*, vol. 2021, no. 3, pp. 142–163, 2021.
- [115] L. ZhangT. ZhuP. XiongW. Zhou, P. S. Yu, "More than privacy: Adopting differential privacy in game-theoretic mechanism design," *ACM Comput. Surv.*, pp. 5471–37, Jul. 2021.
- [116] "MD-GAN: Multi-discriminator generative adversarial networks for distributed datasets," in *Proc. IEEE Int. Parallel Distrib. Process. Symp.*, 2019, pp. 866–877.
- [117] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat.*, 2017, vol. 54, pp. 1273–1282.
- [118] M. Rasouli, T. Sun, and R. Rajagopal, "Fedgan: Federated generative adversarial networks for distributed data," 2020, *arXiv:2006.07228v2*.
- [119] R. A. and N. V., "Federated AI lets a team imagine together: Federated learning of GANs," 2019, *arXiv:1906.03595v1*.
- [120] Y. Zhang, H. Qu, Q. Chang, H. Liu, D. N. Metaxas, and C. Chen, "Training federated GANs with theoretical guarantees: A universal aggregation approach," 2021, *arXiv:2102.04655v1*.
- [121] J. Rajotte *et al.*, "Reducing bias and increasing utility by federated generative modeling of medical images using a centralized adversary," in *Proc. Conf. Inf. Technol. Social Good*, 2021, pp. 79–84.
- [122] T. Zhu, D. Ye, W. Wang, W. Zhou, and P. Yu, "More than privacy: Applying differential privacy in key areas of artificial intelligence," *IEEE Trans. Knowl. Data Eng.*, early access, 2021, doi: [10.1109/TKDE.2020.3014246](https://doi.org/10.1109/TKDE.2020.3014246).
- [123] S. Augenstein *et al.*, "Generative models for effective ML on private, decentralized datasets," in *Proc. Int. Conf. Learn. Representations*, 2020.



Hui Sun received the BEng and MEng degrees from the Zhongnan University of Economics and Law, China, in 2017 and 2020, respectively. She is currently working toward the PhD degree with the School of computer science, China University of Geosciences, Wuhan, China. Her research interests include security defense and privacy preserving in machine learning.



Ping Xiong (Member, IEEE) received the BEng degree in mechanical engineering from Lanzhou Jiaotong University, Lanzhou, China, in 1997 and the MEng and PhD degrees in automation major from Wuhan University, Wuhan, China, in 2002 and 2005, respectively. He is currently a professor with the School of Information and Security Engineering, Zhongnan University of Economics and Law, Wuhan. His research interests include network security, data mining, and privacy preservation.

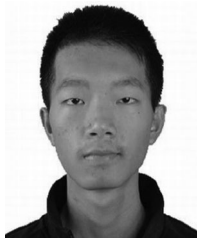


Tianqing Zhu (Member, IEEE) received the BEng and MEng degrees from Wuhan University, China, in 2000 and 2004, respectively and the PhD degree in computer science from Deakin University, Australia in 2014. She is currently a professor with the China University of Geosciences, Wuhan, China. Prior to that, she was a Lecturer with the School of Information Technology, Deakin University. Her research interests include privacy preserving, AI security and privacy, and network security.



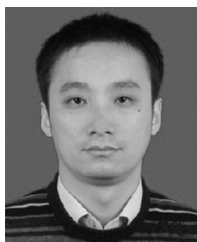
Wanlei Zhou (Senior Member, IEEE) received the BEng and MEng degrees from the Harbin Institute of Technology, Harbin, China in 1982 and 1984, respectively, the PhD degree from The Australian National University, Canberra, Australia, in 1991, all in Computer Science and Engineering, and the DSc degree from Deakin University in 2002. He is currently the vice rector and dean of the Institute of Data Science, City University of Macau, Macao SAR, China. He has authored or coauthored more than 400 papers in refereed international journals

and refereed international conferences proceedings, including many articles in the IEEE transactions and journals. His research interests include security and privacy, parallel and distributed systems, and e-learning.



Zhiqiu Zhang received the BEng degree in 2020 from the China University of Geosciences, Wuhan, China, where he is currently working toward the MEng degree. His research interests include privacy preserving and neural network security.

▷ **For more information on this or any other computing topic, please visit our Digital Library at www.computer.org/csdl.**



Dawei Jin (Member, IEEE) received the BEng and MEng degrees from the Zhongnan University of Economics and Law, China, and the PhD degree in computer science from Wuhan University, China. He visited the School of Computer Science, Deakin University, Australia, funded by the China Scholarship Council from 2011 to 2012. He is currently a professor and a PhD Supervisor with the Zhongnan University of Economics and Law. His research interests include data mining, machine learning, and financial data analysis.