Attribute-Based Membership Inference Attacks and Defenses on GANs

Hui Sun, Tianqing Zhu, *Member, IEEE*, Jie Li, Shoulin Ji, *Member, IEEE*, and Wanlei Zhou, *Senior Member, IEEE*

***Abstract*—With breakthroughs in high-resolution image generation, applications for disentangled generative adversarial networks (GANs) have attracted much attention. At the same time, the privacy issues associated with GAN models have been raising many concerns. Membership inference attacks (MIAs), where an adversary attempts to determine whether or not a sample has been used to train the victim model, are a major risk with GANs. In prior research, scholars have shown that successful MIAs can be mounted by leveraging overfit images. However, high-resolution images make the existing MIAs fail due to their complexity. And the nature of disentangled GANs is such that the attributes are overfitting, which means that, for an MIA to be successful, it must likely be based on overfitting attributes. Furthermore, given the empirical difficulties with obtaining independent and identically distributed (IID) candidate samples, choosing the non-trivial attributes of candidate samples as the target for exploring overfitting would be a more preferable choice. Hence, in this article, we propose a series of attribute-based MIAs that considers both black-box and white-box settings. The attacks are performed on the generator, and the inferences are derived by overfitting the non-trivial attributes. Additionally, we put forward a novel perspective on model generalization and a possible defense by evaluating the overfitting status of each individual attribute. A series of empirical evaluations in both settings demonstrate that the attacks remain stable and successful when using non-IID candidate samples. Further experiments illustrate that each attribute exhibits a distinct overfitting status. Moreover, manually generalizing highly overfitting attributes significantly reduces the risk of privacy leaks.**

***Index Terms*—Membership inference attack, generative adversarial networks, privacy leakage.**

I. Introduction

In RECENT years, disentangled generative adversarial networks (GANs) [1] have seen tremendous developments in terms of controllable high-resolution image generation [2]. Being disentangled means that a GAN can learn specific semantic attributes from diverse and independent units of the latent code, such that customizing a unit of the latent code can lead to a specific generation style. For example, InfoGAN [3] can manipulate writing styles in the MNIST dataset without changing the digit shapes. Additionally, a GAN tends to generate realistic images based on a disentangled latent code, such as StyleGAN [4] and its variants [5], [6], which can successfully generate high-resolution (1024^2) images. As a result, this disentanglement has led to the emergence of numerous new applications, particularly in areas such as privacy-critical image generation [7], [8], image edit [9], [10], image restoration [11], image interpolation [11], style transfer [12], [13] and so on. However, there is a downside to the wide applicability and appeal of disentangled GANs – which is threats to privacy. Overfitting and membership inference attacks (MIA) get the most attention. The model overfits because it is more confident in the members than the non-members of the training set. For example, when a GAN overfits, its discriminator deduces that the members are real with a higher confidence score; and under the guidance of the discriminator its generator tends to reconstruct training samples, resulting in generating samples very closer to the training ones. The disparate performances of members and non-members sufficiently invoke the membership inference attack [14]. In a successful MIA, an adversary determines whether or not a sample has been used to train the victim model [15]. The adversary can therefore mount MIA when the overfitting is detected.

The investigation of whether GANs overfit reaches a positive conclusion as evaluation metrics are continuously updated. The common evaluation focuses on image reconstruction performance. The earliest works evaluate the reconstruction at the image pixel level. [16] claims that overfitting is not detectable in GANs because the training procedure does not create deterministic mappings between latent codes and output images. However, [17] finds that GANs even replicate training samples if the training dataset is small and less complex. Later works update the reconstruction evaluation in pixel and perceptual levels and conclude that GANs can overfit with moderate training dataset [18]. The adversary seizes the opportunity. For a candidate sample, if the generator can perfectly reconstruct it, it belongs to the training dataset [19], [20], [21]. Additionally, the adversary also targets the discriminator because the discriminative models tend to overfit according to the previous works [15], [22], [23], [24], [25]. For example, the LOGAN MIA [26] which considers that samples with significantly high confidence scores are members

Manuscript received 16 April 2022; revised 6 August 2023; accepted 11 August 2023. Date of publication 16 August 2023; date of current version 11 July 2024. *(Corresponding author: Tianqing Zhu.)*

Hui Sun and Jie Li are with the China University of Geosciences, Wuhan, Hubei 430079, China (e-mail: sunhui@cug.edu.cn; leejie@cug.edu.cn).

Tianqing Zhu is with the University of Technology, Sydney, NSW 2007, Australia (e-mail: tianqing.zhu@ieee.org).

Shoulin Ji is with the Zhejiang University, Hangzhou, Zhejiang 310027, China (e-mail: sji@zju.edu.cn).

Wanlei Zhou is with the City University of Macau, Taipa, Macau, China (e-mail: wlzhou@cityu.edu.mo).

This article has supplementary downloadable material available at https://doi.org/10.1109/TDSC.2023.3305591, provided by the authors.

Digital Object Identifier 10.1109/TDSC.2023.3305591

1545-5971 © 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

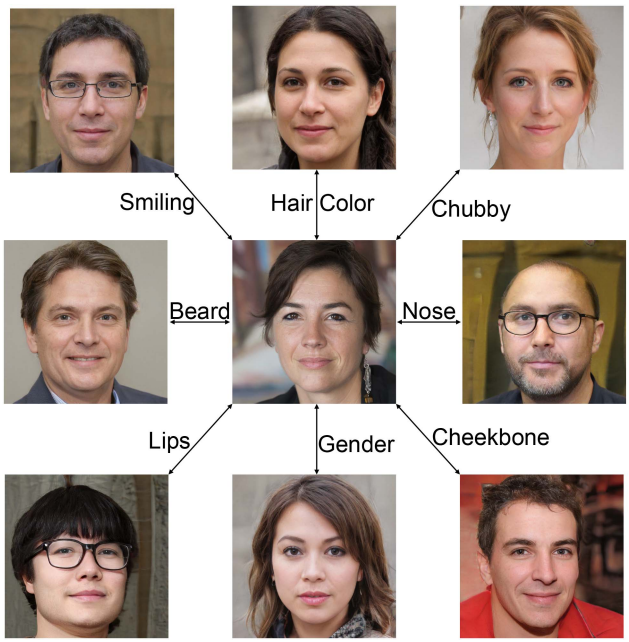


Fig. 1. An example of attribute overfitting. The generated samples are different from the candidate sample (the middle image). However, each of them shares one attribute with the candidate sample. Hence, the candidate sample is probably a training member.

of the training set, depends on overfitting the discriminator. These MIAs have worked well as overfit discriminators and generators are highly confident in the images they have been trained on. These attacks are categorized as image-based MIAs because the image is the smallest unit.

However, existing image-based MIAs cannot handle the case of high-resolution GANs due to the complexity involved in image reconstruction [27]. Disentangled GANs, which specialize in high-resolution image operations such as generation, create untangled semantic features and avoid feature combinations, such as writing styles versus digit shapes with MNIST data or face shapes versus hair color with the FFHQ dataset. Intuitively, their generators possibly establish a relationship between the latent code and semantic attributes during the training phase, which is verified in attribute editing works [9], [11], [28]. In this way, disentangled GANs tend to suffer from attribute overfitting. In fact, one of their advantages is that they are more robust to standard image-based MIAs that focus on the image.

Fig. 1 illustrates a scenario where a StyleGAN generator overfits some attributes. Here, the generated samples share a single attribute with the candidate sample, while the other attributes are different from the candidate sample. In this case, the candidate sample could possibly be a member of the training set as most of its attributes have been reconstructed by the generator. What this example shows is that disentangled GANs may actually be more vulnerable to MIAs than traditional GANs if an adversary takes advantage of attribute overfitting.

This discovery motivated us to form a new perspective toward membership inference that is more advisable for disentangled GANs – namely an attribute-based MIA (AMIA). AMIAs focus on non-trivial attributes in an image, such as the overall composition or high-level aspects like a *nose* in a human portrait. These attributes differ from the stochastic variations described in [4], which can be randomized without substantially affecting our perception of the image, such as the precise placement of hairs in a human portrait. AMIAs play on the fact that the generators of disentangled GANs are more confident about the attributes they have been trained on, so they will tend to reconstruct these attributes over ones they are less familiar with. In other words, if a generator reconstructs an attribute in a highly proficient manner, that attribute is probably associated with a member of the training set. Intuitively then, if multiple attributes of an image belong in the training set, the image is highly likely to be a member of that training set. Meanwhile, owing to the principle that AMIAs focus on several attributes, only the non-trivial attributes are required to be IID, not the image itself, which bypasses the problem of having to find IID candidate samples. In this way, AMIA achieves finer and more empirical membership inferences than traditional MIAs do.

In this article, we test two settings for AMIAs: the black-box setting and the white-box setting. In the black-box setting, we follow the same assumptions as previous image-based MIAs [19], [20], where an adversary can only access a victim GAN by querying from an API. In the white-box setting, the adversary is more knowledgeable of the victim generator. They know the generator's inputs, outputs, and internal workings. This is a commonly-studied scenario because many model developers publish such details in academic papers on their work [19], [20].

Our Contribution: In summary, the main contributions to be found in this study include:

- The first formal description uses attribute inference attacks to launch a membership inference attack against GANs, referred to as AMIA.
- A systematic investigation of our proposed AMIA, covering both black- and white-box settings.
- The first membership inference attack against GANs to consider non-IID candidate samples.
- A new perspective on model generalization and possible defenses against AMIAs, found by investigating the state of single attribute overfitting.

A. Related Work

1) Overfitting and Privacy Leakage in GANs: It is challenging to deduce whether GAN models overfit. Overfitting refers to the phenomenon where a model is exclusively confident in its training data such that it performs differently with members versus non-members [29]. Latent reconstruction is recommended as an evaluation metric for overfitting, which finds the nearest neighborhood through optimization in the latent space. Webster et al. found that whether a model overfits is highly related to the training loss function [16]. If the training protocol does not involve reconstruction error such as $||G(z) - x||_2^2$, they found no overfitting. Yazici et al. found adverse evidence that as stochasticity decreases in GAN training overfitting occurs [18]. Stochasticity has two main sources of latent space and stochastic gradient updates. Feng et al. verified

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

that GANs can overfit if the training dataset is small and less complex [17].

Overfitting can reveal the privacy of training data. Both [18] and [17] equate overfitting with memorization that the GAN model can perfectly reconstruct the training samples because it memorizes them. Since the overfitting model has different performances on training and non-training data, the adversary can mount membership inference attacks (MIAs) which pick the training data out of non-training data. In other words, latent reconstruction can work for overfitting detection and membership inference. Yeom et al. explored the relationship and proposed that overfitting is sufficient to allow an attacker to perform membership inference [14]. In a GAN, this overfitting can apply to either the generator, the discriminator, or both. Either way, overfitting provides an adversary with the opportunity to infer a sample's membership in the training set.

2) Generator-Based MIAs: With an overfit generator, the model would reconstruct a member of the training set better than a non-member owing to this confidence in the training data [30]. In this vein, Ryan et al. [16] constructed an image-latent-image space transformation and built a reconstruction loss function between the raw image and the reconstructed image. The expected difference in performance between reconstructing members and non-members has been observed in GANs with hybrid adversarial methods, like CycleGAN [31], and non-adversarial methods, like GLO [32], but not in pure GANs, such as PGGAN [33]. Contemporary studies on MIAs typically rely on similar reconstructions and, so, the MIAs are generally successful [19], [20], [21]. For example, Chen et al. [19] devised a reconstruction loss that additionally considers perceptual loss and regularization. Hilprecht et al. [20] designed a Monto Carlo attack against GANs that focuses on the number of generated samples that are similar to the training set, and proposed a corresponding set methodology for inferring membership. Liu et al. [21] built a brand new neural network to find the right latent code for the candidate sample. The reconstruction loss would then indicate whether the sample was a member or non-member. Contrary to the conclusion in [16], these three image-based MIAs have been successful on both hybrid and pure GANs, indicating that overfitting does indeed occur in pure GANs. Subsequently, Ryan et al. [34] proposed an innovative identity membership attack that infers whether images of a certain identity were used to train a model. Here, the reconstruction of identity information rather than pixel or perceptual information is the key to the success of the overall reconstruction. This attack worked with StyleGAN up to the point where the training set exceeded 880 identities and 320 k images in total.

3) Discriminator-Based MIAs: In other studies, the researchers focus on GAN discriminators. Hayes et al. [26] begin their MIA by observing the confidence score of the discriminator or local discriminator. The local discriminator is built based on the raw generator. Thus, it inherits overfitting from the generator. Both cases provide evidence for successful MIAs via overfitting in GANs. Based on the connection between generalization and overfitting, Adlam et al. [35] employed the discriminator's generalization gap to detect overfitting in the discriminator and the generator. Notably, they did not observe overfitting in Wasserstein GANs.

However, as Webster et al. [16] note, as a field, we have not yet told the full story on overfitting, nor are our explorations of MIAs complete. Overfitting and privacy leaks in GANs deserve more attention.

Due to the fact that the discriminator of a GAN is generally not accessible when training ends, generator-based MIAs have become the mainstream approach, employing the latent reconstruction pattern. Among these generator-based MIAs, there are two evolution directions: 1. the way to finish the latent reconstruction, via optimization, encoder, or the hybrid one. 2. the distance metric for reconstruction error, including the pure pixel-level loss [16], [17] and the combination of pixel-level and perceptual-level loss [18], [19]. Nowadays, the disentangled representation makes great success in high-resolution image generation. Each latent dimension of a disentangled representation controls a single visual attribute (disentanglement), and each attribute is controlled by a single dimension (completeness). In this situation, the traditional image reconstruction metric possibly degrades, especially in the case of high-resolution images where euclidean distance changes a lot for perceptually small changes [36].

Considering the relationship between the semantic attribute and disentangled representation, we propose the attribute-based membership inference attack (AMIA) from the aspect of attribute overfitting. Compared to traditional MIAs, AMIAs have the following differences: 1) In contrast to the discriminator-based MIAs, AMIAs rely on the generator. This is more practical as adversaries tend to have limited access to discriminators. 2) AMIAs focus on non-trivial attributes while traditional MIAs perceive the whole image. Therefore, AMIAs are more successful on disentangled GANs.

II. Prelimilaries

A. Notation

Consider the generator G of a GAN as the victim model and suppose that the objective of the task is to infer the probability that a candidate sample x belongs to the training set. Let $Pr(\cdot)$ denote the probability function, $\mathbb{I}(\cdot)$ denote the indicator function, and p_{train} denote the distribution of the training set. More specifically, the goal of an AMIA is to infer the membership of a candidate sample based on the victim model's reconstruction of some of its attributes, where $L^i(x, G)$ denotes the i_{th} attribute reconstruction loss. A set of classifiers extracts multiple sample attributes, and c^i denotes the classifier of the i_{th} attribute. Δ denotes the distance metric, i.e., euclidean distance (L_2 norm) between the candidate and generated attributes. The main notations used are summarized in Table I.

B. Generative Adversarial Networks (GANs)

Generally, GANs consist of a generator that generates samples from the latent code as the input, and a discriminator that distinguishes between the generated samples and the training samples. The latent code is typically randomly sampled from

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

TABLE I
FORMALIZED NOTATIONS INVOLVED IN AMIAS

Notation	Description
G	Generator of the victim GAN
p_Z	Latent space of the victim GAN
z	The latent code in p_Z
x	The candidate sample
p_{train}	Distribution of the training data
c^i	i_{th} attribute classifier
$Pr(\cdot)$	The probability function
$\mathbb{I}(\cdot)$	The indicator function
$L^i(x,G)$	The reconstruction loss between the i_{th} attribute of x and the victim generator G
n	The number of attributes considered
m	The number of samples generated
Δ	The distance function such as Euclidean distance

a latent space Z, such as a Gaussian distribution denoted as $z \sim P_Z$. During the training phase, the generator and discriminator compete against each other. That is, the generator tries to produce a fake sample to fool 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

$$\begin{aligned} &min\ L_G = E_{z\sim P_z}[log(1-D(G(z)))] \\ &max\ L_D = E_{x_0\sim D_{train}}[logD(x_0)]+E_{z\sim P_z}[log(1-D(G(z)))] \end{aligned} \tag{1}$$

where G and D denote the generator and discriminator, D_{train} denotes the training set, and x_0 denotes the training sample.

Disentangled GANs are the result when researchers try to interpret how the latent code affects the generated image and further control the image generation. The disentangled GANs encourage less entanglement during the process from the latent space to the image space. In other words, each part of the latent code independently manipulates a specific semantic attribute of the image [28]. In this way, it is possible to interpret how latent code affects semantic attributes, further deriving semantic manipulations using the vector arithmetic property. The vector arithmetic property is formulated as $z' = z + \lambda d$, where $d \in \mathbb{R}$ denotes the direction corresponding to a particular attribute, and λ is the step. As the latent code moves in a particular direction, the corresponding attribute changes.

C. MIAs

Membership is defined as the state of a candidate sample x having been used to train the victim GAN model. In this article, Our focus is only on the generator, as discriminators are generally not accessible. As such, an adversary must use their knowledge to infer whether a candidate sample x has been used to train the victim generator G – an accomplishment that might further reveal the training data. Formally, we have the following definition of a generator-based MIA, drawing on [37],

$$A(x,G,\Omega) \longrightarrow Pr(x \in p_{train}) \longrightarrow \{0,1\} \tag{2}$$

where A denotes the adversary, and Ω denotes the adversary's external knowledge. 1 means x is used to train the victim generator G, and 0 otherwise.

III. PROBLEM STATEMENT

In this section, we present a practical problem and then propose our solution.

A. Image or Attribute Overfitting

The latent code z of a disentangled GAN model can be separated into independent parts, denoted as $z = \{z_0, z_1, \ldots\}$. Each part is highly correlated with a specific semantic attribute and is not sensitive to other attributes. There are two ways to achieve disentanglement. One is to encourage the generator to learn the mapping between each part z_k and the corresponding semantic attribute in a supervised way. HoloGAN [38], for example, includes a domain-specific generator architecture to add model-inductive bias. InfoGAN [3], as another example, maximizes the mutual information between the observation and the factor code. The other is in an unsupervised way. Stylegan [4] and StyleGAN2 [5] introduce an intermediate latent W space, which does not necessarily follow the probability density of the training data, therefore, more disentangled. The disentangled latent space avoids entanglement so that disentangled GANs tend to generate high-resolution and realistic images [4]. Reasonably speculating, the generator in a disentangled GAN should have greater confidence in the attributes that have been trained over attributes that have not. This is termed attribute overfitting.

As mentioned in Sections I-A1 and I-A2, existing overfitting detection and MIAs mainly utilize the image reconstruction loss, however, cannot handle the case of high-resolution images due to the complexity involved in reconstruction [27]. That is to say, we cannot mount MIAs from the perspective of image overfitting. Nevertheless, that does not mean that disentangled GANs are safe. To ensure privacy guarantees with these emerging networks, it is essential that we explore novel types of MIAs specifically directed toward the weaknesses of disentangled GANs. Here, attribute-based MIAs make a good starting point.

B. Evaluation

Intuitively, MIA is a binary classification problem that discriminates the members from the non-members. Hence, many relevant studies, as well as this article, use AUC scores and average precision (AP) to evaluate MIAs [19], [26], [39]. More specifically, this article executed the evaluations on randomly reshuffled data samples from the query set, consisting of an equal number of members and non-members. This approach maximized the uncertainty of the inferences. Baseline performance was deemed to be equivalent to random guessing at 0.5.

C. Non-IID Candidate Set

According to existing MIAs against GANs [19], [20], the adversary observes the generated samples and makes inferences based on the reconstruction loss between the generated samples and the candidate samples. The candidate samples consist of

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

two subsets, the member set (members of the victim training dataset) and the non-member set, which intuitively conforms to a distribution, generally, one that is IID. In real life, however, it is hard to find such an IID non-member set [40], especially for complex datasets like those of high-resolution human faces such as FFHQ and CelebAHQ. Humphries et al. [41] explored the case of non-IID candidates in an MIA against classifiers and found that making inferences would be easier owing to the inherent independence of the dataset. However, how MIAs against GANs respond to non-IID candidate samples still remains a mystery.

In our experiments, we attempted to mount several existing MIAs – LOGAN [26] and GANLeak [19] – against StyleGAN [4] trained on FFHQ and CelebAHQ. LOGAN includes a discriminator-accessible attack model and a full black-box attack model. Here, we chose the ideal discriminator-accessible settings, widely considered as the most knowledgeable and effective settings. GANLeak includes a full black-box attack model, a partial black-box attack, and a white-box attack, where the information held by the adversary increases with each type of attack. We chose the black-box and white-box settings for our experiments. The candidate samples comprised an equal number of samples from the FFHQ and CelebA-HQ datasets. With StyleGAN (FFHQ) as the victim model, samples from FFHQ are the members while samples from CelebA-HQ are the non-members. Conversely, with StyleGAN (CelebA-HQ) as the victim model, samples from CelebA-HQ are the members while samples from FFHQ are the non-members.

Fig. 3 shows the attack performance, where we see several strange results. First, for GANLeak, the AUC and AP scores on StyleGAN (FFHQ) are very low (less than 0.5); however, on StyleGAN (CelebA-HQ), they are plausible at greater than 0.5 in both the black- and white-box conditions. Considering that GANLeak makes inferences based on image reconstruction errors, the results statistically illustrate that the samples generated by either of the models have smaller image reconstruction errors with the samples in CelebA-HQ than the samples in FFHQ. Second, LOGAN had a better performance on StyleGAN (FFHQ) than StyleGAN (CelebA-HQ). This suggests that the discriminator of StyleGAN (FFHQ) can discriminate between samples from FFHQ and CelebA-HQ – possibly by capturing the attributes that only FFHQ has, such as the image background. Third, GANLeak does not become any more successful when the adversary has more adversarial information. In other words, performance is the same no matter whether the setting is black-box or white-box. In summary, existing MIAs against GANs cannot handle non-IID candidate samples.

The non-IID candidate samples mislead the current MIAs, degrading empirical utility. We hope to loosen the IID restriction from the overall image to several attributes. Since these few attributes of the candidate set are IID, it meets the IID condition.

D. Discussion

Existing MIAs, like GANLeak and LOGAN, cannot handle the cases of high-resolution images and non-IID candidate sets because they focus on the whole image. The AMIA in this article is designed to be perpetrated on disentangled GANs which specialize in high-resolution image operations such as generation and editing. It considers several non-trivial attributes, noting that images consist of multiple attributes, both non-trivial and trivial, which avoids the complexity of high-resolution images. The different focuses of existing MIAs and AMIA are shown in Fig. 2. With an AMIA, the candidate samples are only required to be IID at the attribute level and, even then, only the non-trivial attributes are required to be IID.

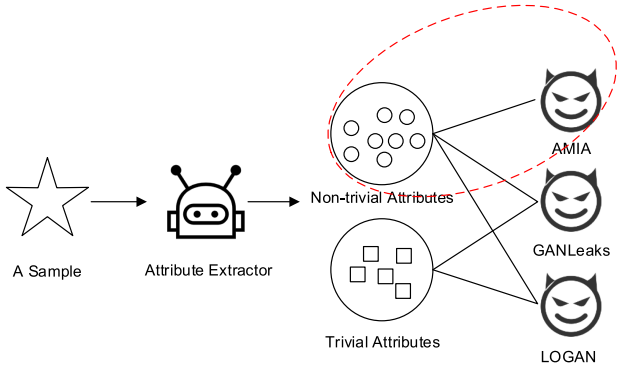


Fig. 2. AMIAs focus on the non-trivial attributes, while GANLeaks and LOGAN focus on the overall sample including both non-trivial and trivial attributes.

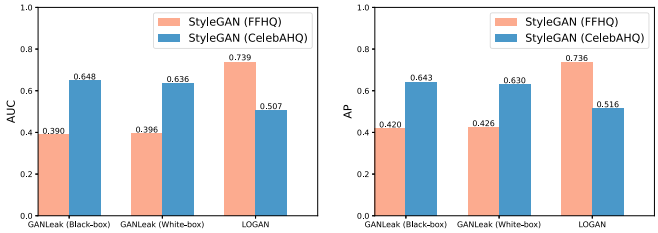


Fig. 3. Existing MIAs on StyleGANs respectively trained on FFHQ and CelebAHQ datasets, including the black-box and white-box attacks of GANLeak and LOGAN with the accessible discriminator.

IV. Attribute-Based MIAs

A. Adversarial Knowledge

Since an AMIA against a GAN begins at the attribute level, a set of trained classifiers $\{c^1, c^2, \ldots, c^n\}$ is required to extract the individual attributes of the images. This is the case for both black- and white-box adversaries. In a black-box attack, the adversary only has access to the samples generated by the victim generator. In a white-box attack, the adversary has full access to the victim model's internal structure and parameters, as well as the model input and generated samples. Thus, these adversaries can observe the generated samples with purpose, instead of simply collecting them blindly. For this reason, theoretically, a white-box adversary has a better chance of extracting a sample's membership status from a model. The detailed attack methodology for each scenario is presented in Sections IV-C and IV-D, respectively.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

B. Generic AMIA Model

AMIAs focus on the semantic attributes of images; they decompose the inferred membership of an entire image into the inferred membership of multiple semantic attributes. Intuitively, an AMIA can be divided into three phases: measuring attribute reconstruction, inferring attribute membership, and inferring membership.

Supposing that an adversary considers n semantic attributes of a candidate sample x, n attribute classifiers would be used to extract the attributes. Thus, $c^i(x)$ is the i_{th} attribute of the candidate sample x.

In the first phase – attribute reconstruction measurement – the adversary constructs the reconstruction loss functions for each candidate attribute of x, denoted as $L^i(x, G)$ for the i_{th} attribute. If $L^i(x, G)$ is less than a threshold, the generator G successfully reconstructs the i_{th} attribute of x, and the generated sample is defined as reconstruction. With confidence in the attributes trained, an overfit generator will tend to reconstruct attributes from the training set. Hence, inspired by [19] and [26], we devised a quality priority strategy and a quantity priority strategy to holistically explore the victim model's reconstruction performance. The quality priority strategy focuses on the optimum reconstruction effect given that both black- and white-box adversaries aim to find the optimal reconstruction among the generated samples. The quantity priority strategy focuses on how effective reconstruction is. Here, black-box adversaries are concerned with how many generated samples it takes to produce a successful reconstruction of the i_{th} candidate attribute of x, while white-box adversaries focus on the number of optimization iterations to get an acceptable reconstruction.

In the second phase – attribute membership inference – the adversary infers whether each attribute of the candidate sample belongs to the training set. $\mathbb{I}(c^i(x) \in p_{train})$ denotes the i_{th} attribute membership, which is determined by the corresponding reconstruction loss $L^i(x, G)$.

Finally, in the third phase – membership inference – the adversary deduces the candidate sample's membership based on the results of inferring the attribute membership inferences (i.e., Phase 2). If the majority of a sample's attributes are members, the candidate sample is deemed to be a member. Formally,

$$Pr(x \in p_{train}) \propto \frac{1}{n} \sum_{i=1}^{i=n} \mathbb{I}(a_i \in p_{train}) \tag{3}$$

Notably, considering more than one attribute, $n > 1$, is a more robust approach to inferring overall membership given that some members will share one or more attributes while other attributes are well generalized.

C. Black-Box Setting

In black-box settings, the adversary has very little knowledge of the victim model. Hence, the generated sample set may be one that is published publicly or it could simply be collected blindly through a series of queries on the victim generator. Suppose the adversary acquires m generated samples from the victim generator, denoted as $G(\cdot)_{k=1}^{m}$. The adversary would then use n attribute classifiers to extract n attributes from both the candidate and the generated samples.

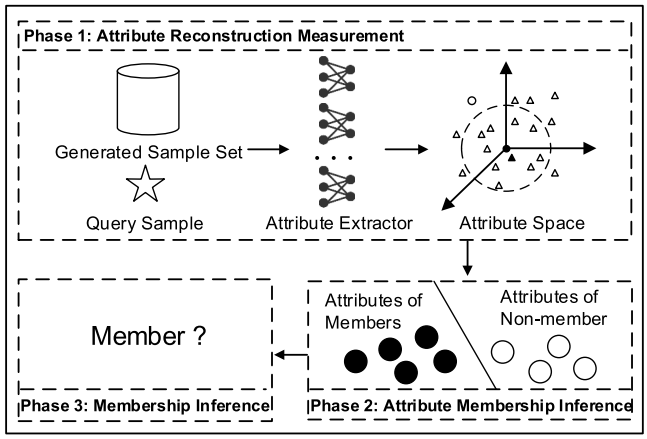


Fig. 4. Framework of a black-box AMIA. In the attribute space, the dots (● for attributes of members and ○ for attributes of non-members) denote the attributes of the candidate sample, the triangles (△ and ▲) denote the attributes of the generated samples, and the dashed circle denotes the threshold T_{rec} in the quantity priority strategy. In the quality priority strategy, the sample that is closest distance-wise (labeled as ▲) is the focus. In the quantity priority strategy, the number of generated samples in the dashed circle is the focus.

Fig. 4 illustrates the attack framework in the black-box setting. In the first phase, the black-box adversary measures the attribute reconstruction performance through the distance between the candidate attribute and the attributes of generated samples, denoted as $L^i(x, G(\cdot)_{k=1}^{m})$. In the second phase, a threshold is taken as the requirement of the attribute reconstruction error. The adversary infers an attribute's membership by comparing $L^i(x, G(\cdot)_{k=1}^{m})$ and the threshold. If more than half the attributes of the candidate sample are deemed to be of the training set, the candidate sample is then deemed to be a member of the training set with high probability.

Attribute Reconstruction Measurement: As mentioned, there are two options for reconstruction measurements, the quality priority, which focuses on the optimum reconstruction effect, and the quantity priority, which focuses on the reconstruction frequency. If the quality priority strategy is chosen, the adversary aims to find the closest feature among $G(\cdot)_{k=1}^{m}$ for each candidate attribute. This strategy is based on the premise that the victim generator will reconstruct the attribute better if it was originally used to train the model. Formally, for the i_{th} attribute,

$$L^i_{quality}(x, G(\cdot)_{k=1}^{m}) = \min_{\hat{x} \in G(\cdot)_{k=1}^{m}} \Delta(c^i(x), c^i(\hat{x})) \tag{4}$$

where Δ denotes the euclidean distance (L2 norm), measuring the similarity of the candidate and generated samples in a specific attribute level.

If the adversary opts for the quantity priority strategy, they will investigate the reconstruction frequency. Here, T_{rec} denotes the requirements of the reconstruction effect. When the i_{th} attribute reconstruction loss between a generated sample and

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

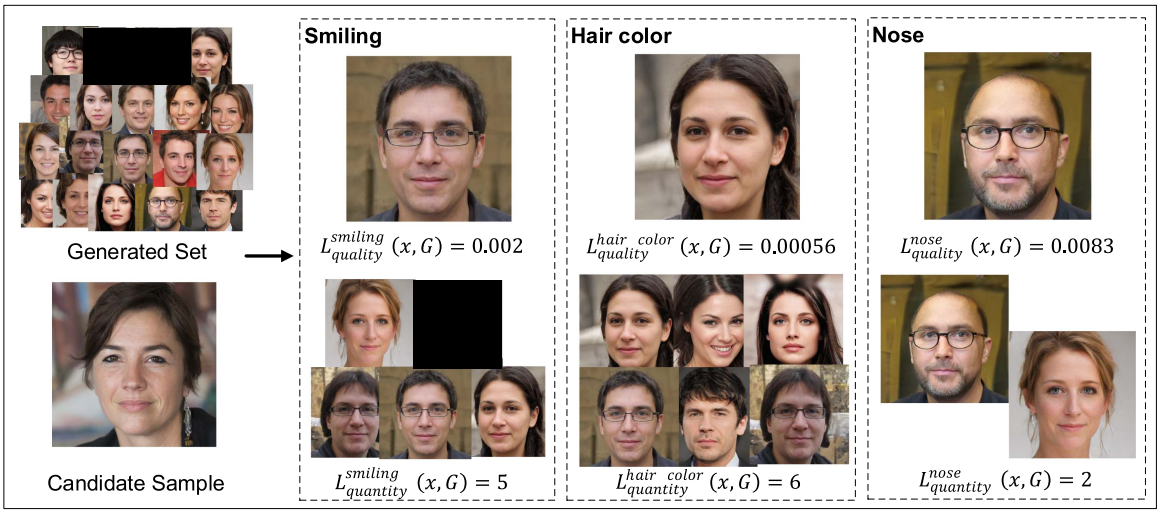


Fig. 5. Case study of black-box AMIA. The adversary finds the image(s) that most closely resemble the candidate image for each attribute separately and subsequently infers the membership of each attribute. If the majority of attributes are determined to be members, the whole image is deemed to belong to the victim training set.

the candidate sample is under T_{rec}, the victim model is deemed to have reconstructed the i_{th} victim attribute once successfully. Overall, the adversary aims to find generated samples within the threshold T_{rec} for each victim attribute. Formally, for i_{th} attribute,

$$L^i_{quantity}(x, G(\cdot)^m_{k=1}) = \sum_{\hat{x} \in G(\cdot)^m_{k=1}} \mathbb{I}_{\Delta(c^i(x), c^i(\hat{x})) < T_{rec}} \quad (5)$$

Attribute Membership Inference: After measuring the victim generator's performance with attribute reconstruction, the adversary obtains a vector $L(x,G) = \{L^1(x,G), L^2(x,G), \ldots, L^n(x,G)\}$ for a candidate sample x. In the quality priority strategy, $L^i(x,G)$ denotes a minimal reconstruction loss between the target sample x and the training set $G(\cdot)^m_{k=1}$ with the respect to the i_{th} attribute. Thus, the reconstruction threshold T_{rec} is used to decide whether the minimal reconstruction error is small enough to determine that the attribute is a training attribute. Formally,

$$\mathbb{I}(c^i(x) \in p_{train}) = \begin{cases} 1, & L^i_{quality}(x,G) < T_{rec} \\ 0, & L^i_{quality}(x,G) >= T_{rec} \end{cases} \quad (6)$$

If $L^i(x,G)$ is below the T_{rec}, the adversary concludes that the i_{th} attribute of the candidate sample x is a member of the training set.

In the quantity priority strategy, $L^i(x,G)$ denotes how many generated samples are less than the T_{rec} away from the candidate sample x with the respect to the i_{th} attribute. Thus, another threshold T_{mem} comes into play. If there are more than T_{mem} generated samples close to x with respect to the i_{th} attribute, the i_{th} attribute of x is a member of the training set. Formally,

$$\mathbb{I}(c^i(x) \in p_{train}) = \begin{cases} 1, & L^i_{quantity}(x,G) > T_{mem} \\ 0, & L^i_{quantity}(x,G) <= T_{mem} \end{cases} \quad (7)$$

In fact, the quality priority strategy is a special case of the quantity priority strategy when the two strategies employ the same threshold T_{rec} and the quantity priority strategy has a T_{mem} threshold set to zero.

Membership Inference: In the last phase, the adversary infers the membership of the candidate sample x by making a statistical calculation over the inferences of the n attributes. If more than half the attributes of the candidate sample are members of the training set, then the candidate sample x is a member of the training set. This probability is calculated by (3). Obviously, for a candidate sample x, the more attributes that are in the training set, the greater the probability that the candidate sample was originally included in the training set.

Case Study: Here, we present a case study to illustrate how an adversary would conduct a black-box AMIA following our methodology. In this case, an ideal GAN serves as a generator of human face images and the adversary infers whether a candidate face image is a member of the training set based on a set of images generated by the GAN. Fig. 5 shows the main components of the attack process.

In the first phase, attribute reconstruction measurement, the adversary adopts either the quality priority strategy or the quantity priority strategy. Fig. 5 shows the results with respect to three attributes: smiling, hair color, and nose. In the second phase, attribute membership inference, the adversary makes inferences based on the thresholds T_{rec} and T_{mem}. For example, if $T_{rec} = 0.005$ in a quality priority strategy, and $L^{smiling}_{quality}(x,G)$ and $L^{haircolor}_{quality}(x,G)$ are below T_{rec}, this indicates that the attributes *smiling* and *hair color* are in the training set, while *nose* is not. Whereas, if $T_{mem} = 5$ and $T_{rec} = 0.01$ in a quantity priority strategy and only $L^{haircolor}_{quantity}(x,G)$ is above T_{mem}, this indicates that only the attribute *hair color* is in the training set, while the attributes *smiling* and *nose* are not. As for the third phase, membership inference, with the attribute inference results [1, 1, 0], a candidate sample would be judged a member using the quality priority strategy. However, with the quantity priority strategy and the attribute inference results [0, 1, 0], the sample would be judged a non-member.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

D. White-Box Setting

In the white-box setting, the victim generator is fully accessible. This includes its inner architecture and parameters, as well as the input, the latent code, and the output, the generated image. As such, the adversary can update the latent code for the victim generator to attain the best reconstruction for each victim attribute. This objective is best described by the optimization process $L^i(x, G(z))$.

With full knowledge of the victim generator, the adversary can solve the optimization problem with multiple optimization algorithms [42], [43]. We used the Adam algorithm [42] as it requires little memory and is computationally efficient. Naturally, in the first phase, the adversary measures the attribute reconstruction performance. This can be done through the optimization process by evaluating either the optimized result (the quality priority strategy) or the optimization speed (the quantity priority strategy). During the second and third phases, the adversary infers the membership of each victim attribute and, in turn, membership of the complete image in the same way as in the black-box setting.

Attribute Reconstruction Measurement: With prior information about the model's structure and parameters, the reconstruction loss function for the i_{th} attribute is formulated as,

$$L^i(x, G(z)) = \min_{z \sim pz} \Delta(c^i(x), c^i(G(z))) \tag{8}$$

where z is the latent variable that needs to be optimized. With StyleGAN, for example, we would choose the latent code w in latent W space, which is more disentangled than the latent space Z. In addition, we might follow both the quality priority strategy and the quantity priority strategy to measure reconstruction performance.

Following the quality priority strategy, the adversary would regard the optimized reconstruction loss as the reconstruction performance, where the lower the reconstruction loss, the higher the possibility that the attribute belongs to the training set. Again, this is based on the premise that the victim generator will do a better job at reconstructing the attributes it has seen in training than ones it has not.

Following the quantity priority strategy, the adversary would regard the optimization speed as the reconstruction performance, i.e., they would determine whether the iterations for the reconstruction loss $L^i(x, G(z))$ reach a negligible threshold T_{rec}. Theoretically, fewer iterations should be needed to reach a fixed value with an attribute that has been trained over one that has not. This is, of course, based on the generator's confidence in its training set attributes.

Attribute Membership Inference: Similar to the black-box attack, the adversary infers whether the attribute belongs to the training set based on the attribute reconstruction measurement. With the quality priority strategy, $L^i(x, G)$ denotes the final optimized reconstruction loss between the candidate sample x and the victim model with respect to the i_{th} attribute. The reconstruction threshold T_{rec} is used to determine whether the reconstruction loss is small enough to deduce the attribute as a training attribute. Formally,

$$\mathbb{I}(c^i(x) \in p_{train}) = \begin{cases} 1, & L^i_{quality}(x, G) < T_{rec} \\ 0, & L^i_{quality}(x, G) >= T_{rec} \end{cases} \tag{9}$$

If $L^i(x, G)$ is below T_{rec}, the adversary reaches the conclusion that the i_{th} attribute of the candidate sample x belongs in the training set. With the quantity priority strategy, $L^i(x, G)$ denotes the iterations needed to reach the negligible threshold T_{rec}. If less iterations are needed than T_{mem} with respect to the i_{th} attribute, the i_{th} attribute of x belongs in the training set. Formally,

$$\mathbb{I}(c^i(x) \in p_{train}) = \begin{cases} 1, & L^i_{quantity}(x, G) < T_{mem} \\ 0, & L^i_{quantity}(x, G) >= T_{mem} \end{cases} \tag{10}$$

Membership Inference: The final image membership inference is made based on the attribute inferences. Given a candidate sample, the more attributes that are deemed to belong in the training set, the more likely it is that the sample is a member of the training set, as per (3).

Case Study: Consider an ideal GAN that is well trained as a generator of human face images. Then suppose that the adversary has full access to the GAN and a candidate sample. Following the principles of a white-box AMIA, the adversary observes the optimization process of the generator generating a sample close to the candidate sample.

Fig. 6 depicts 250 iterations of the optimization process for reconstructing the hair color attribute $L^{hair\ color}(x.G)$. With the quality priority strategy, the adversary would focus more on the optimization output and record the final loss values. In this case, the optimization distance is 0.000188. Thus, if the threshold T_{rec} is 0.0005, the attribute hair color belongs to the training set; however, if $T_{rec} = 0.0001$, it does not. With the quantity priority strategy, the adversary focuses more on the optimization speed and records the iteration steps when the loss reaches the threshold T_{rec}. For example, if $T_{rec} = 0.0005$, the optimization distance is 35. The choice of the threshold T_{rec} partly accounts for the attack performance. During attribute membership inference, if the adversary set T_{mem} to 50, and the loss reaches T_{rec} within 50 steps, then hair color is in the

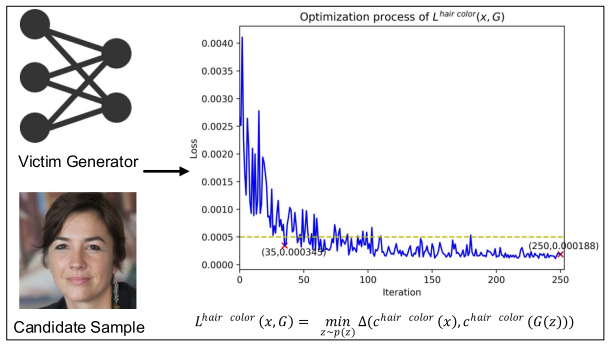


Fig. 6. Case study of the white-box AMIA. The adversary makes inferences by observing the speed (quantity priority) and results (quality priority) of the reconstruction optimization process. The candidate sample, whose reconstruction optimizes more quickly or achieves a smaller loss, is deemed to belong to the victim training set.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

training set, whereas if $T_{mem} = 30$, then hair color is not. The adversary repeats the processes for each attribute, determining each attribute's membership. If more than half the attributes are deemed to be in the training set, the candidate sample is deemed to belong.

V. Defense on AMIAs

In this section, we propose two strategies for defense against AMIAs. These are limited model queries and attribute generalization. AMIAs exploit the generator's tendency to overfitting attributes. Hence, holding back the attribute membership inference can effectively damage the inference chain. For example, limiting the number of queries means the adversary only has a narrow scope through which to observe the victim generator's reconstruction performance. Alternatively, attribute generalization mitigates attribute overfitting, providing a fundamental defense against this weakness in the model.

A. Limit Model Query

This approach is tailored for black-box settings and works by preventing the adversary from obtaining enough generated samples. The premise is that if the adversary cannot generate enough samples to cover the attributes of the training set, they cannot comprehensively investigate the victim model's attribute reconstructions. In this way, a black-box attacker would be unlikely to infer the membership of the candidate samples from the samples generated.

It is worth noting, though, that some attention needs to be paid to the balance between defense and a model's utility. If all model users are limited in their queries, not just the adversaries, then this defense will come at the cost of model utility, giving the model narrow practical usefulness.

B. Attribute Generalization

AMIAs heavily depend on attribute overfitting. Thus, attribute generalization can protect against these types of attacks. There are two mechanisms to generalize attributes: semantic attribute editing, which edits overfitting attributes to others in a black-box setting, and a transferring mechanism that can re-generate the victim GAN model in the white-box setting. Note that, in our experiments, we evaluate how attribute generalization degrades AMIAs.

Semantic Attribute Editing: Semantic attribute editing is an attempt to erase any traces of the training attributes from the generated samples by manually editing the generated attributes into other semantics that sit outside the training set. Owing to the continuous and complete latent space, the generator can produce a series of continuous images [44], such as human faces with noses from large to moderate to small. Hence, a defender might edit the generated attribute into interpolations between the training attributes so that the generated attributes are not similar to the training ones in the image space, while still making them plausible.

Semantic attribute editing originates from the idea of customizing the high-fidelity face images generated by the state-of-art GAN models. A general method is to find the direction for each semantic attribute in the latent space so that the attribute continuously changes when the latent code moves along the direction. Consider an edited image denoted as $G(z + \lambda * d)$, where z is the corresponding latent code of the raw image, generally attained through GAN inversion techniques [10], [28], [45], d is the direction, and λ adjusts the amplitude. To find the semantic direction, Shen et al. [11] take the normal vector of the SVM boundary, which separates the latent space into the opposite semantic label. Han et al. [9] additionally consider the individual images, achieving better image editing.

Transferring the Model: This method attempts to manipulate the generator's confidence in the training set. As such, it is an effective defense against white-box attacks. The basis of the defense is that the defender expands the scope of the training set by adding edited images whose semantic attributes have been edited away from the raw ones. Hence, the victim generator will be confident in more attributes than it actually is. In other words, the generator is better generalized.

In effect, transferring a GAN means applying transfer learning techniques to train the GAN model. Considering the tremendous computing resources and time taken to train a model from scratch, transferring a GAN, rather than retraining, is a much better choice. Several studies are dedicated to advancing this line of enquiry. For instance, Wang et al. [46] initialize the generator and discriminator networks of a GAN with either random or pre-trained weights (from the source networks). They find that transferring the discriminator is much more critical than the generator but that transferring both networks is best. Their team subsequently proposed MineGAN to avoid overfitting [47]. Mo et al. find that simple fine-tuning of GANs with frozen lower layers of the discriminator performs surprisingly well [48], while Fregier and Gouray freeze the low-level layers of both the discriminator and generator of the original GAN to result in significantly faster training [49].

Fig. 7 presents the attribute generalization process. The defender recurrently stages an AMIA against each attribute. Once the attribute is vulnerable, the attribute generalization process works. In the black-box setting, the defender edits the overfit semantics. In the white-box setting, the defender also edits semantics for a set of generated samples and then uses them to transfer the victim model. Once all the attributes are impervious to an AMIA, the generated samples or the pre-trained model will not reveal the privacy of the training dataset.

C. Defense Discussion

Few defense mechanisms against MIAs are applicable to GANs. Among them, differential privacy is the most effective [50], [51], [52]. Nevertheless, the protection provided by differential privacy comes at the cost of training resources and model utility [19]. By contrast, our proposed defense mechanisms are more targeted at AMIAs. Limiting the number of allowed model queries is easy to implement. However, this

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

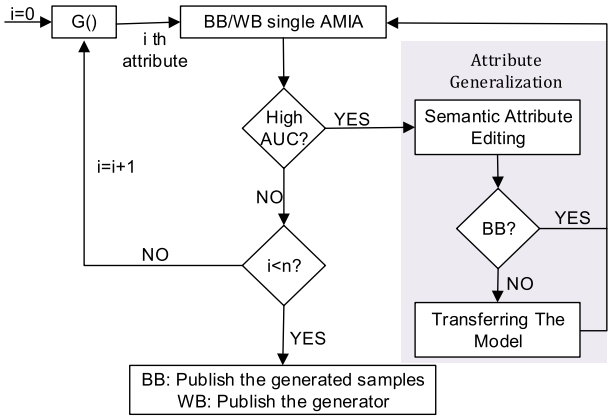


Fig. 7. Attribute generalization process in the white-box setting (WB) and black-box setting (BB): to generalize each overfitting attribute until all chosen attributes are not vulnerable to the AMIA.

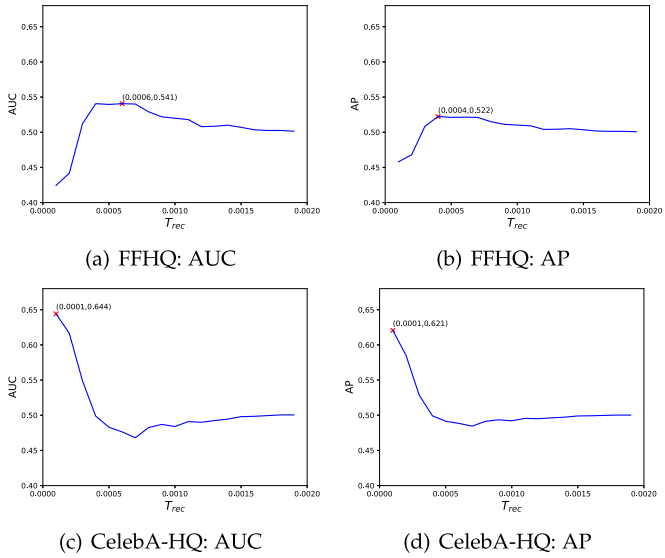


Fig. 8. Black-box AMIA performance with quality priority w.r.t StyleGAN on FFHQ and CelebA-HQ, including AUC and AP scores.

strategy only works when the adversary has no access to the model, and it may substantially decrease the GAN's utility. Attribute generalization provides a fundamental defense against AMIA by mitigating attribute overfitting. The downside is that this strategy demands additional time and resources. The upside is that semantic attribute editing and transferring the model effectively defend against both black- and white-box attacks.

VI. Experimental Evaluation

A. Setup

Victim Model and Datasets: We selected StyleGAN [4], and, more specifically, the synthesis network of StyleGAN, as our victim generator for its disentangled latent W space. All experiments were conducted on two high-resolution datasets of human faces. The CelebA-HQ dataset consists of 30 k high-resolution face images at 1024^2 resolution, which were originally drawn from CelebA [53]. The Flickr-Faces-HQ (FFHQ) dataset [4] consists of 70 k high-resolution human face images at 1024^2 resolution, which was formerly crawled from Flickr. FFHQ dataset is more diverse and larger than CelebA-HQ dataset [4], hence, StyleGAN (FFHQ) overfits less than StyleGAN (CelebA-HQ) according to [17]. The candidate samples comprised an equal number of samples from the FFHQ and CelebA-HQ datasets. With StyleGAN (FFHQ) as the victim model, samples from FFHQ are the members while samples from CelebA-HQ are the non-members. Conversely, with StyleGAN (CelebA-HQ) as the victim model, samples from CelebA-HQ are the members while samples from FFHQ are the non-members. Our AMIAs only focus on the non-trivial attributes of the images and these were IID in both datasets, consistent with previous works [19], [20], [26].

Generated Dataset: For each victim generator, we collected 100 k generated samples, $|G(\cdot)| = 100\ k$, to ensure a complete investigation of the attribute reconstruction.

Attribute Classifiers: We built attribute classifiers on ResNet-18 [54] and the CelebA dataset. CelebA provides 40 labeled attributes for each human face, and we selected two global attributes, smiling and gender, as our targets as these have been widely investigated in the research on semantic attribute editing [9], [11]. We also selected seven concrete attributes: hair, chubby, cheekbones, lips, nose, beard, and hat. Table III shows the training results of each attribute classifier. We find that all binary classifiers achieved over 80% accuracy on the training set and over 75% on the validation set, suggesting that the binary classifiers can well extract the attributes.

Threshold Choosing: As mentioned, the two priority strategies involve two thresholds, T_{rec} and T_{mem}, which represent the minimum requirements for membership. T_{rec} denotes the maximum reconstruction loss, which generally takes values close to 0, such as 0.0005. T_{mem} has two roles. It reflects the minimum number of neighboring samples in the black-box setting, which we set to between 5 and 50. It also reflects the maximum iteration steps to reach T_{rec} in the white box setting. Here, we adopted a proportion scheme ranging from 0% to 100%. Furthermore, we control the maximal number of iterations for reconstruction optimization with the quality priority strategy under the white-box setting from 100 to 1,000. We list these parameters in Table II.

B. Evaluation on Black-Box Setting

In the black-box attacks, we followed each of the quality priority and the quantity priority strategies and measured reconstruction performance. The AUC and AP scores of the two strategies are reported in Figs. 8 and 9.

Fig. 8 shows the attack performance with the quality priority strategy. We adjusted the reconstruction threshold T_{rec} from 0.0001 to 0.0009 in steps of 0.0019. This attack performed well with the proper threshold T_{rec}, especially on the CelebA-HQ dataset. With a T_{rec} of between 0.0004 and 0.0007, the attack was also successful against FFHQ. On CelebA-HQ, when T_{rec} approximates 0, the attack has better performance. The fact that

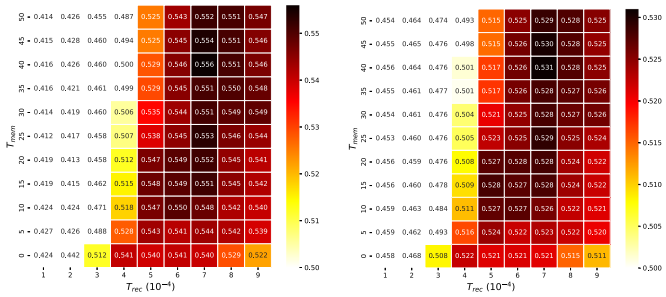
Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

TABLE II
DESCRIPTION OF THE EXPERIMENTAL PARAMETERS

Experiment parameters			Description
Black-box	quality	T_{rec}	The maximum of reconstruction loss that decides the attribute as a member
	quantity	T_{rec}	The maximum of reconstruction loss that decides the attribute as a neighboring attribute
		T_{mem}	the minimum number of neighboring attributes that decides the attribute as a member
White-box	quality	Trec	The maximum of optimized reconstruction loss that decides the attribute as a member
		Iteration	The maximum of reconstruction optimization
	quantity	T_{rec}	The baseline that the optimized reconstruction loss should reach
		T_{mem}	The maximum iteration steps to reach the baseline T_{rec}

TABLE III
CLASSIFICATION ACCURACY W.R.T. EACH ATTRIBUTE CLASSIFIER

idx	name	train acc.	valid acc.
0	smiling	0.938	0.926
1	male	0.987	0.977
2	black_hair	0.910	0.877
3	chubby	0.961	0.952
4	high_cheekbones	0.892	0.870
5	big_lips	0.813	0.780
6	big_nose	0.877	0.832
7	no_beard	0.970	0.950
8	wearing_hat	0.992	0.987
	average	0.927	0.906



(a) FFHQ: AUC (b) FFHQ: AP

(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 9. Black-box AMIA performance with quantity priority w.r.t StyleGAN on FFHQ and CelebA-HQ, including AUC and AP scores. In the heatmap, the deeper the color, the greater the AUC (AP) score; thus, the more successful the attack.

a small T_{rec} is still effective suggests that the victim generator can reconstruct the training set attributes on a subtle level, which indicates a high level of attribute overfitting. This is consistent with the intuition that StyleGAN (CelebA-HQ) suffered more from overfitting due to the smaller and less complex training set compared with StyleGAN (FFHQ). The AUC and AP scores support this intuition. Our AMIA with quality priority strategy achieves higher AUC and AP scores against StyleGAN (CelebA-HQ), up to 0.644 and 0.621. While the AUC and AP scores against StyleGAN (FFHQ) are 0.541 and 0.522. As the T_{rec} grew, the AUC and AP scores for both datasets began to approximate no better than random guessing (0.5). The results indicate that the effectiveness of the T_{rec} threshold is highly related to how much the victim model overfits. A less overfit model will need a moderate T_{rec}, and a more overfit model will need a tiny T_{rec}.

Fig. 9 plots the attack performance with the quantity priority strategy. The deeper the color, the greater the AUC (AP) score; thus, the more successful the attack. We adjusted the reconstruction threshold T_{rec} from 0.0001 to 0.0009 in steps of 0.0001 and T_{mem} from 5 to 50. We can see deeper colors concentrated on the right for FFHQ, indicating that the attack against FFHQ worked well with a relatively large T_{rec} threshold, while the attack against CelebA-HQ worked well with a smaller T_{rec}. Furthermore, both the AUC and AP scores against StyleGAN (CelebA-HQ) were larger than the ones against StyleGAN (FFHQ). These observations of both the quantity and quality priority strategies are therefore consistent, indicating that AMIA achieves better MIA performances against highly overfitting GANs such as StyleGAN (CelebA-HQ).

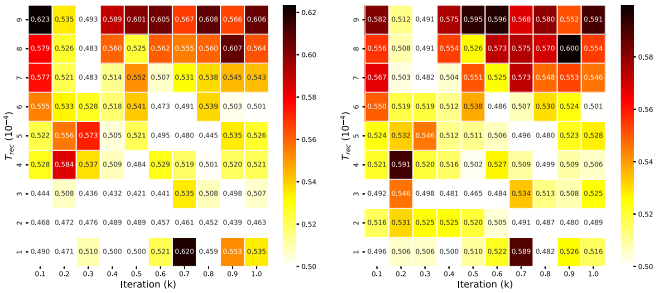
Comparing the quality and quantity priority strategies, we find that most AMIAs with a T_{mem} of greater than 0 have higher AUC and AP scores than that when T_{rec} and $T_{mem} = 0$. This is because a $T_{mem} > 0$ avoids accidental cases when the victim GAN just happens to generate an attribute without it being based on some confidence in the training attributes. Here, T_{rec} plays the decisive role, while T_{mem} acts as a regulator. Therefore, we conclude that the quantity priority strategy works better than the quality priority strategy in the black-box scenario.

C. Evaluation on White-Box Setting

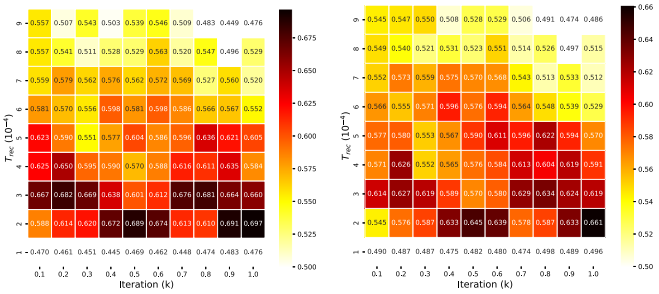
The evaluation results for the white-box attack appear in Figs. 10 and 11. White-box scenarios are commonly studied in privacy preservation [19], [55] because the model generally has been fully published and the adversary has access to its internal structure and parameters. Here, the published model is the victim generator.

Fig. 10 plots the results of the white-box attack with the quality priority strategy at different maximum iterations and T_{rec}

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

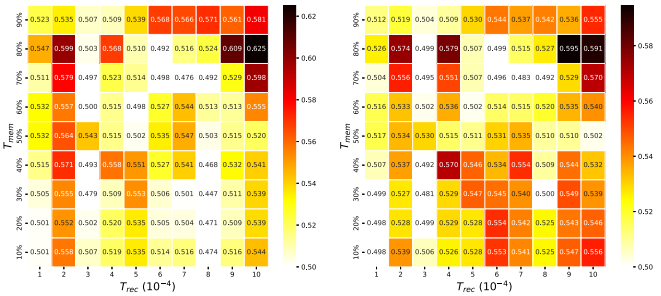


(a) FFHQ: AUC (b) FFHQ: AP

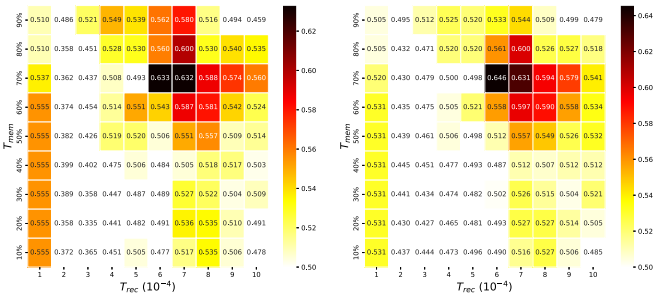


(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 10. The white-box AMIA performance with quality priority w.r.t StyleGAN on FFHQ and CelebA-HQ, including AUC and AP scores.



(a) FFHQ: AUC (b) FFHQ: AP



(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 11. White-box AMIA performance with quantity priority w.r.t StyleGAN on FFHQ and CelebA-HQ, including AUC and AP scores.

thresholds. Recall that this strategy focuses on the optimization results, where a member of the training set will have a smaller attribute reconstruction loss given the same number of iterations. We adjusted the reconstruction threshold T_{rec} from 0.0001 to 0.0009 in steps of 0.0001 and the iteration from 100 to 1,000. We observe that T_{rec} plays a more decisive role than the number of iterations. As shown in the plots, the scores for both attacks show great success with an AUC of 0.623 for FFHQ and 0.697 for CelebA-HQ. Additionally, the attack on CelebA-HQ did better with a lower T_{rec} while the attack on FFHQ worked better with higher T_{rec}. This is consistent with the black-box setting and indicates that StyleGAN trained on CelebA-HQ was better at reconstructing attributes in the training set. Additionally, the optimization iterations have no obvious effect on attack performance.

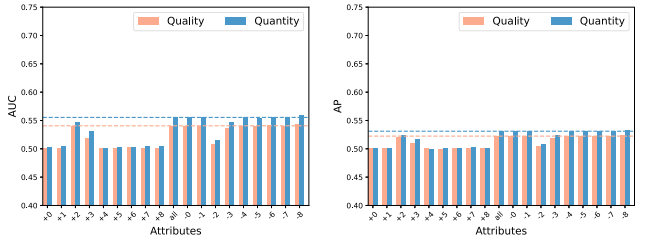
The aim of the quantity priority strategy is to infer membership through optimization speeds, that is, the number of iterations needed for the attribute reconstruction loss to reach a negligible threshold T_{rec}. Fig. 11 plots the attack performance against different thresholds, where T_{rec} ranges from 0.0001 to 0.001 and T_{mem} ranges from 10%th to 90%th of the iterations that each candidate sample needs to reach T_{rec}. With this strategy, the results of the attack on FFHQ were comparable to the attack on CelebA-HQ, and were even more robust. The appreciable attack performances illustrate that the victim models reconstruct the attributes that were in the training set more quickly than those that were not. These experiments show that the quality and quantity priority strategies effectively mount successful white-box AMIAs.

Comparison to Black-Box Attacks: In these simulated attacks, the white-box adversary has full access to the victim generator, while the black-box adversary only has access to a set of samples generated by the victim generator. Intuitively, the white-box adversary has more knowledge around which to design strategies and should therefore be able to mount a more successful attack. Although most MIAs support this intuition, there are a couple of exceptions [56], [57].

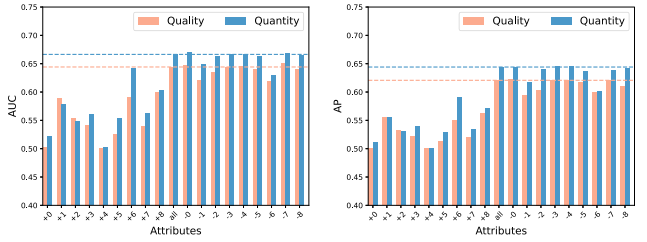
In our AMIAs, the white-box attacks were more effective than the black-box attacks with increases in the AUC score of 0.069 for FFHQ and 0.03 for CelebA-HQ when changing from the black-box to the white-box setting. With respect to the discrete victim model, StyelGAN, our white-box adversary had access to the synthesis networks but not the mapping network. From this, we conclude that publicizing a model's parameters, even if just some of them, does incur a risk of privacy leakage.

Furthermore, we focus on the reconstruction threshold T_{rec}, which plays the main role in quality and quantity priority strategy under black- and white-box settings. Except for the quantity priority strategy under the white-box setting, we found a consistency for T_{rec} that AMIA prefers smaller values such as 0.0002 for StyleGAN (CelebA-HQ) but prefers greater values such as 0.0008 for StyleGAN (FFHQ). This is related to how the target GAN overfits. Intuitively, if the target GAN overfits more like StyleGAN (CelebA-HQ), it can reconstruct the attributes of members more precisely, therefore, allowing greater T_{rec}. Unfortunately, we cannot provide an explicit profile of threshold settings. Our work is still an early exploration of membership inference attacks against GANs, and we hope AMIA can provide some clues for the design of confidential GANs in the future.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

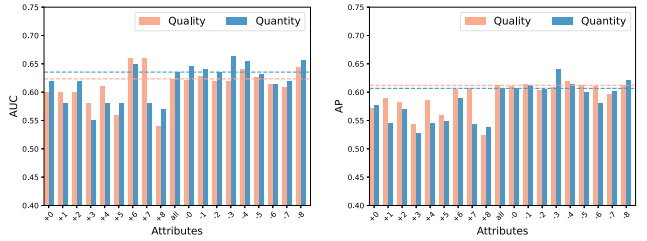


(a) FFHQ: AUC (b) FFHQ: AP

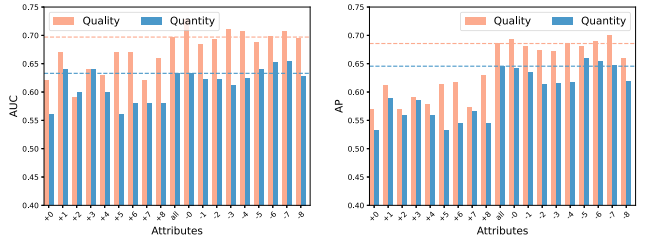


(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 12. Single AMIA versus multiple AMIA in the black-box setting. From left to right, the columns show utilizing an attribute, then all attributes, then subtracting an attribute from the all.



(a) FFHQ: AUC (b) FFHQ: AP



(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 13. Single AMIA performance versus multiple AMIA in the white-box setting. From left to right, the columns show utilizing an attribute, then all attributes, then subtracting an attribute from the all.

D. Single AMIA versus Multiple AMIA

In the previous experiments, we chose nine non-trivial attributes to attack: smiling, gender, hair, chubby, cheekbones, lips, nose, beard, and hat. Here, we consider two other cases: utilizing a single attribute at a time, i.e., single AMIA ($n = 1$), and subtracting a single attribute at a time ($n = 8$). By comparing these three cases, we hope to learn more about the relationship between a single attribute and the overall attack performance.

In the single attribute experiments, where $n = 1$, inferring membership in the training set depends on inferring membership of a single attribute. If that attribute is inferred to belong in the training set, the candidate sample will be deemed to be a member. The experimental settings are the same as the last series of experiments and, again, both white-box and black-box settings are tested with the quality and quantity priority strategies.

Figs. 12 and 13 show the AUC and AP scores of these tests. From left to right, the columns show utilizing an attribute, then all as per the last experiments, then subtracting an attribute. As the results show, the overfitting status of the attributes varies even though the same settings have been used for each experiment. Of course, it also varies with different experimental settings. Fig. 12, within the black-box setting, shows some well-generalized attributes with AUC and AP scores near 0.5 – for example, the 0th, 1st and 4th to 8th attributes for FFHQ in Fig. 12(a) and (b), and the 0th and 4th attributes for CelebA-HQ in Fig. 12(c) and (d). When excluding a single attribute, attack performance does not decrease if the excluded attribute is well generalized. However, if the excluded attribute is less generalized, performance does decrease – for example, excluding the 2nd attribute for FFHQ in Fig. 12(a). This phenomenon also exists in the white-box setting. For instance excluding the less generalized 6th attribute in Fig. 13(a) and (b) reduces the AUC and AP scores on FFHQ, while excluding more generalized 8th attribute improves the scores. By contrast, in the white-box setting, every single attribute has high AUC and AP scores, which indicates severe overfitting.

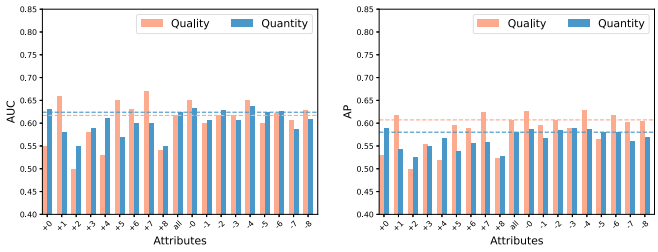
From this set of experiments overall, we reached the following conclusions: 1) Multiple AMIAs generally achieve better attack performance. Hence, it is better to choose more than one non-trivial attribute. 2) Multiple AMIAs neutralize the performance of a single AMIAs. When we exclude the less overfitting attributes, AMIA achieves better performance. Based on the two conclusions, the model owner can determine which attributes are strongly or weakly overfit, and put in place a targeted defense method.

E. Comparison With Other MIAs

In this section, we compare our AMIA with a related study that involves the attribute for enhancing image membership inference [58]. Intuitively, if a sample shares the same property with the majority of the training dataset, it is possibly a member of the training dataset. Hence, the adversary in [58] utilizes the property inference attack to enhance the MIA through three steps:

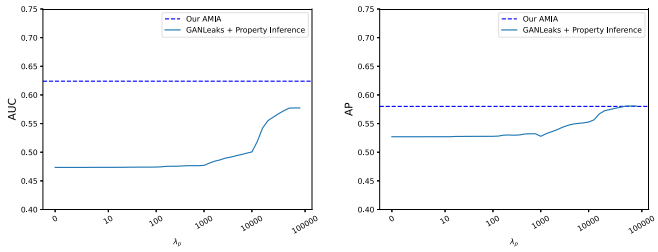
- Launch a property inference attack to infer the proportion of samples used to train the target GAN with respect to a particular attribute, for example, the ratio of females.
- Launch a membership inference attack and [58] choose the GANLeak [19].

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.



(a) AMIA: AUC (b) AMIA: AP

Fig. 14. Single and multiple AMIA performance with quality and quantity priority strategies in the white-box setting w.r.t StyleGAN on 2,048 CelebA-HQ images, including AUC and AP scores.



(a) CelebAHQ: AUC (b) CelebAHQ: AP

Fig. 15. Comparison between our AMIA and GANLeak with the enhancement of property inference.

- Make the final membership inference as the following equation:

$$\mathcal{A}(x) = \mathbb{I}[L_{GANLeak}(x, G(z)) < T_{rec} + \lambda_p \frac{1}{N} \sum_{i}^{N} (2 * \mathcal{P}_i - 1)] \quad (11)$$

where λ_p controls the magnitude of the enhancement, N denotes how many attributes are considered, and $\mathcal{P}_i$ refers to the proportion of the i_{th} attribute in the training dataset.

In the experiment, we choose the case of IID candidate set where the member set and the non-member set conform to the same distribution. Because GANLeaks cannot handle the case of non-IID candidate set, according to the experiment results in section III-C. Referring to [58], we use the gender attribute. We set up a StyleGAN as the target using 2,048 CelebA-HQ samples with 70% female and 30% male. The white-box setting is considered that the adversary has full knowledge of the target GAN. Additionally, we consider the ideal case for the property inference, where the adversary knows the exact proportion, i.e., 70% female and 30% male. The experiment results are shown in Fig. 15.

Fig. 14 depicts our AMIA results with the quality and quantity priority strategies under the white-box setting, considering scenarios of single and multiple attributes. And Fig. 15 depicts the GANLeak results with the enhancement of the property inference attack. When $\lambda_p = 0$, the property inference attack provides no enhancement.

For comparison, we include the AMIA result, which employs all attributes, in Fig. 15 as a reference line. It is worth noting that AMIA can achieve superior results by selectively excluding certain attributes. Obviously, as λ_p gets larger, both AUC and AP scores of GANLeak increase. However, even with the increase, the maximum AUC score remains below that of our AMIA. The results indicate that even with the help of property inference, GANLeak achieves poorer MIA performance than our AMIAs. This is because the property inference attack infers the macro-level property of the training dataset, rather than focusing on individual images, thus providing limited enhancement.

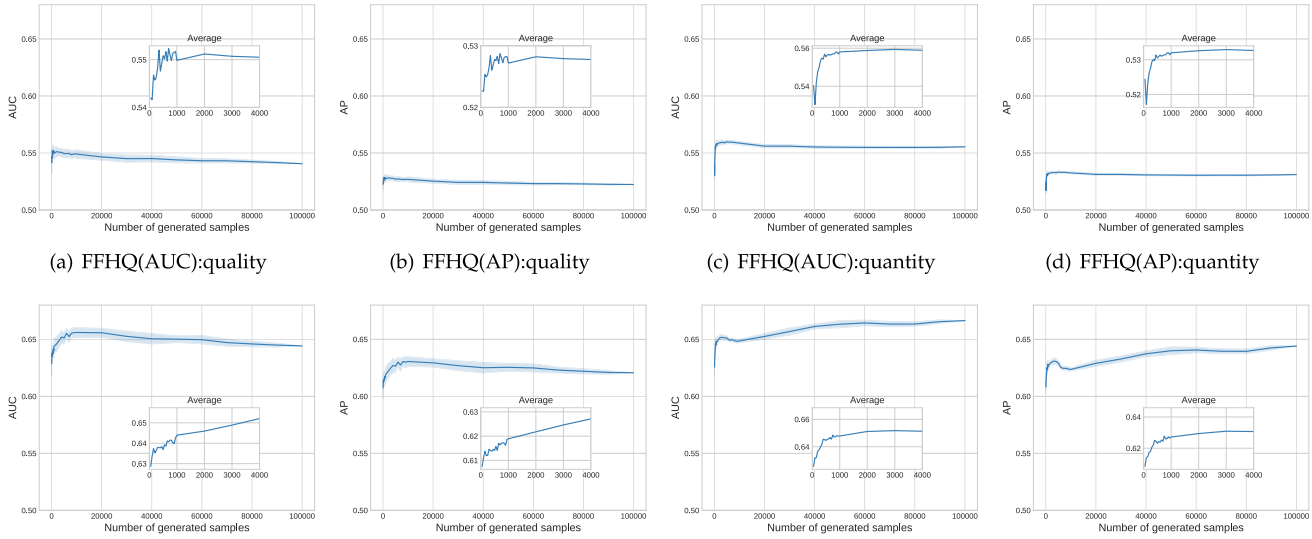
F. Defense Evaluation

1) Limit Model Query: As the simplest defense method, we limited the number of queries an adversary could make of a model. Initially, we randomly sampled our data sets from 50 to 100 k at increasing intervals from the generated sample set. Black-box attacks with both quality priority and quantity priority strategies were conducted on each dataset, namely, FFHQ and CelebA-HQ. Additionally, each attack was repeated 30 times to mitigate the possibility of coincidental results. The results are given in Fig. 16.

Fig. 16 shows the black-box AMIA results with the quality and quantity priority strategy against StyleGAN (FFHQ) and StyleGAN (CelebA-HQ). With the quality priority strategy, there is an obvious increase in AUC and AP scores at the beginning and a gentle decrease as the number of generated samples grows. The decrease is attributed to the quality priority strategy which finds the closest sample among the generated ones for the i_{th} attribute of the candidate sample. Hence, with more generated samples, there is a higher possibility to find a very close sample, even for the non-members of the training set. In this way, the adversary cannot correctly determine the attribute membership by reconstructing the candidate attribute, as well as image membership. Additionally, the decrease appears with fewer generated samples in StyleGAN (FFHQ), with about thousands of generated samples, while StyleGAN (CelebA-HQ) with about ten thousand. This is highly related to model generalization. If the GAN has a stronger generalization capability, it can generate samples with more diverse attributes that tend to include the attributes of non-members. StyleGAN (FFHQ) is trained on a larger and more complex dataset, i.e., FFHQ, so that has stronger generalization capability [17] and possibly generates attributes close to the candidate attribute. The decline indicates that limiting the victim model's queries does not defend against the black-box AMIAs under the quality priority setting because they attain the best attack performance even only need thousands of generated samples.

In contrast, the decrease of the quantity priority strategy needs more generated samples according to Fig. 16. We observe a subsequent decline in StyleGAN (FFHQ) and an overall increase in StyleGAN (CelebA-HQ). The observation indicates that the quantity priority strategy has better performance than the quality priority with more generated samples. Because it involves another condition that statistics how many generated samples meet the threshold T_{rec}. The results indicate that limiting the

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.



(a) FFHQ(AUC):quality (b) FFHQ(AP):quality (c) FFHQ(AUC):quantity (d) FFHQ(AP):quantity

(e) CelebA-HQ(AUC):quality (f) CelebA-HQ(AP):quality (g) CelebA-HQ(AUC):quantity (h) CelebA-HQ(AP):quantity

Fig. 16. Black-box AMIA performance w.r.t. number of generated samples. AUC and AP scores are reported.

victim model's queries offers better but still limited defensive performance against the black-box AMIAs under the quantity priority setting because they can achieve a moderate attack performance with only thousands of generated images.

In conclusion, limiting the victim model's queries provides little protection against black-box AMIAs, unless the limits are as low as hundreds, which severely sacrifices the GAN model's utility.

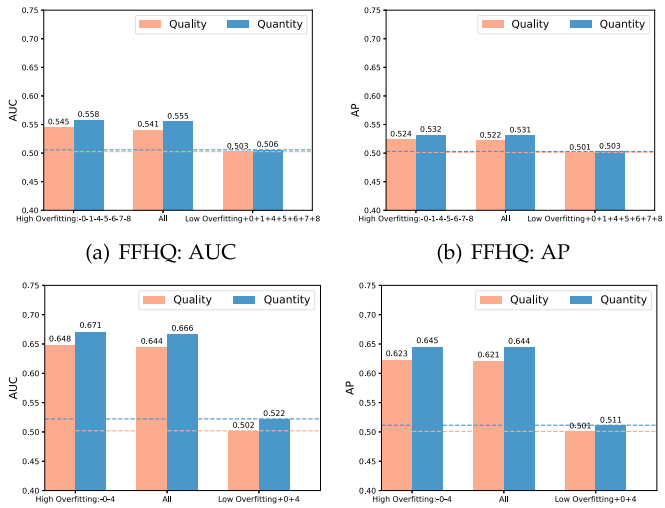
2) Attribute Generalization: As the experiments in Section VI-D show, attributes have different overfitting statuses. Intuitively, if we generalize the highly overfitting attributes, the risk of a privacy leak would greatly decrease. We could generalize certain attributes through semantic attribute editing or by transferring the model. These two attribute generalization methods are specialized in [10], [28], [45], [46], [47], [49]. Here, we assume the perfect generalization of certain attributes without conducting empirical experiments because a comprehensive discussion on generalization requires dedicated research in the future. Generalization is expected to offer an effective defense against AMIAs.

Figs. 17 and 18 show the attack performance of the highly overfitting attributes, all attributes, and less overfitting attributes. The less overfitting attributes are those with an AUC/AP score near 0.5 or obviously smaller than the others. Attack performance on the less overfitting attributes was significantly worse than those on more overfitting attributes and on all attributes. For instance, in Fig. 17, both the AUC and AP scores of the black-box AMIAs decrease to almost 0.5 after excluding the highly overfitting attributes. In the white-box setting, the AUC and AP scores also experience a sharp decrease although the score still stayed above 0.55. Thus, we conclude that attribute generalization can work as a defense against both black- and white-box AMIAs with the following suggestions:

- Against black-box attacks: The model owner should start with a single-attribute AMIA on each attribute to find which ones overfit. Then the overfitting attributes of the generated samples should be manually edited to achieve generalization.
- Against white-box attacks: Before publishing a model, the model owner should perform a single-attribute AMIA on each attribute. Then a GAN transfer should be performed to re-train the model on additional samples and purposely generalize the highly overfitting attributes. For example, the model owner could expand the attribute set by manually editing the attributes of the generated samples.

When all attributes are well-generalized, AMIAs are fundamentally ineffective.

The above experimental results demonstrate the effectiveness and robustness of AMIAs against a disentangled GAN.

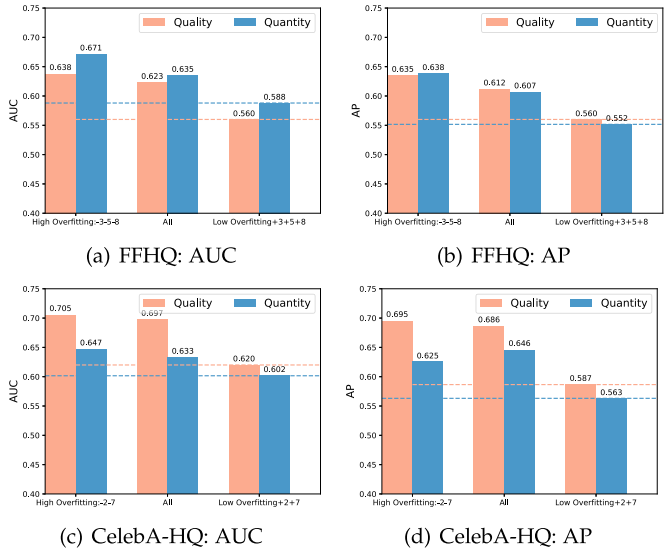


(a) FFHQ: AUC (b) FFHQ: AP

(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 17. Black-box AMIA performances w.r.t. high overfitting, all and low overfitting attributes. From left to right, the columns show highly overfitting attributes, i.e., subtracting well-generalized attributes from the all, then all attributes, then well-generalized attributes.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.



(a) FFHQ: AUC (b) FFHQ: AP

(c) CelebA-HQ: AUC (d) CelebA-HQ: AP

Fig. 18. White-box AMIA performances w.r.t. high overfitting, all and low overfitting attributes. From left to right, the columns show highly overfitting attributes, i.e., subtracting well-generalized attributes from the all, then all attributes, then well-generalized attributes.

Furthermore, the defense results illustrate that merely limiting model queries will significantly degrade the model's utility without providing substantial protection. Generalizing the attributes proves to be a more effective defense.

VII. Conclusion

In this article, we presented an attribute-based membership inference attack (AMIA) specifically designed to penetrate the newly-emerging disentangled GANs which specialize in high-resolution image generation and other operations such as editing. Disentangled GANs differ from traditional GANs in that they establish mappings between latent codes and semantic attributes. Meanwhile, existing MIAs cannot handle the case of high-resolution images. Thus, in an AMIA, the adversary infers membership of a candidate sample's attributes in the training set and, further, the candidate sample's membership by observing the victim model's performance with attribute reconstruction. To comprehensively explore this type of attack, we mounted both black-box and white-box attacks in various configurations. Additionally, we also put forward a new perspective on model generalization and a possible defense that involves testing single attributes for their overfitting status. Extensive experiments demonstrate that, with non-IID candidate samples, our AMIAs in both the black- and white-box settings reached good and stable performance. Further experiments with attribute generalization show that manually generalizing the highly overfitting attributes significantly decreases the risk of privacy leaks.

This article makes some worthy contributions to the literature on GAN privacy, but it is important to note that work in this field is still in its infancy. More attention needs to be paid to privacy issues with GANs to encourage more and greater protections for these vital networks.

References

[1] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, MIT Press, 2014, pp. 2672–2680.

[2] A. Kammoun, R. Slama, H. Tabia, T. Ouni, and M. Abid, "Generative adversarial networks for face generation: A survey," *ACM Comput. Surveys*, vol. 55, no. 5, pp. 94:1–94:37, 2023.

[3] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "InfoGAN: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Barcelona, Spain, Dec. 5–10, 2016, pp. 2172–2180.

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 4401–4410.

[5] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 13–19, 2020, pp. 8107–8116.

[6] T. Karras, M. Aittala, J. Hellsten, S. Laine, J. Lehtinen, and T. Aila, "Training generative adversarial networks with limited data," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Dec. 06–12, 2020, Art. no. 1015.

[7] L. Fetty et al., "Latent space manipulation for high-resolution medical image synthesis via the StyleGAN," *Zeitschrift für Medizinische Physik*, vol. 30, no. 4, pp. 305–314, 2020.

[8] K. Su, E. Zhou, X. Sun, C. Wang, D. Yu, and X. Luo, "Pre-trained StyleGAN based data augmentation for small sample brain CT motion artifacts detection," in *Proc. 16th Int. Conf. Data Mining Appl.*, Springer, Foshan, China, Nov. 12–14, 2020, pp. 339–346.

[9] Y. Han, J. Yang, and Y. Fu, "Disentangled face attribute editing via instance-aware latent space search," in *Proc. 13th Int. Joint Conf. Artif. Intell.*, Z. Zhou, Ed., 2021, pp. 715–721.

[10] J. Zhu, Y. Shen, D. Zhao, and B. Zhou, "In-domain GAN inversion for real image editing," in *Proc. 16th Eur. Conf. Comput. Vis.*, Springer, Glasgow, U.K., Aug. 23–28, 2020, pp. 592–608.

[11] Y. Shen, J. Gu, X. Tang, and B. Zhou, "Interpreting the latent space of GANs for semantic face editing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 13–19, 2020, pp. 9240–9249.

[12] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN: How to embed images into the StyleGAN latent space?," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 4431–4440.

[13] R. Abdal, Y. Qin, and P. Wonka, "Image2StyleGAN++: How to edit the embedded images?," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 13–19, 2020, pp. 8293–8302.

[14] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha, "Privacy risk in machine learning: Analyzing the connection to overfitting," in *Proc. IEEE 31st Comput. Secur. Found. Symp.*, 2018, pp. 268–282.

[15] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in *Proc. IEEE Symp. Secur. Privacy*, San Jose, CA, USA, May 22–26, 2017, pp. 3–18.

[16] R. Webster, J. Rabin, L. Simon, and F. Jurie, "Detecting overfitting of deep generative networks via latent recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, Jun. 16–20, 2019, pp. 11 273–11 282.

[17] Q. Feng, C. Guo, F. Benitez-Quiroz, and A. M. Martínez, "When do GANs replicate? On the choice of dataset size," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Montreal, QC, Canada, Oct. 10–17, 2021, pp. 6681–6690.

[18] Y. Yazici, C. Foo, S. Winkler, K. Yap, and V. Chandrasekhar, "Empirical analysis of overfitting and mode drop in GAN training," in *Proc. IEEE Int. Conf. Image Process.*, 2020, pp. 1651–1655.

[19] D. Chen, N. Yu, Y. Zhang, and M. Fritz, "GAN-Leaks: A taxonomy of membership inference attacks against generative models," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, J. Ligatti, X. Ou, J. Katz, and G. Vigna, Eds., 2020, pp. 343–362.

[20] B. Hilprecht, M. Härterich, and D. Bernau, "Monte Carlo and reconstruction membership inference attacks against generative models," in *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 4, pp. 232–249, 2019.

[21] 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*, J. Wang, K. Shim, and X. Wu, Eds., Beijing, China, Nov. 8–11, 2019, pp. 459–467.

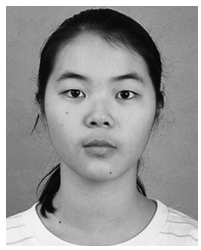
[22] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, B. Thuraisingham, D. Evans, T. Malkin, and D. Xu, Eds., 2017, pp. 587–601.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

[23] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Commun. ACM*, vol. 64, no. 3, pp. 107–115, 2021.
[24] N. Carlini, C. Liu, Ú. Erlingsson, J. Kos, and D. Song, "The secret sharer: Evaluating and testing unintended memorization in neural networks," in *Proc. 28th USENIX Secur. Symp.*, N. Heninger and P. Traynor, Eds., Santa Clara, CA, USA, Aug. 14–16, 2019, pp. 267–284.
[25] S. Zhou, T. Zhu, D. Ye, X. Yu, and W. Zhou, "Boosting model inversion attacks with adversarial examples," *IEEE Trans. Dependable Secure Comput.*, to be published, doi: 10.1109/TDSC.2023.3285015.
[26] J. Hayes, L. Melis, G. Danezis, and E. D. Cristofaro, "LOGAN: Membership inference attacks against generative models," in *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 1, pp. 133–152, 2019.
[27] S. Wang, S. Nepal, A. Abuadbba, C. Rudolph, and M. Grobler, "Adversarial detection by latent style transformations," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 1099–1114, Mar. 2022.
[28] W. Xia, Y. Zhang, Y. Yang, J. Xue, B. Zhou, and M. Yang, "GAN inversion: A survey," 2021, *arXiv:2101.05278*.
[29] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," in *Proc. 4th Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., 2016.
[30] H. Sun, T. Zhu, Z. Zhang, D. Jin, P. Xiong, and W. Zhou, "Adversarial attacks against deep generative models on data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 4, pp. 3367–3388, Apr. 2023.
[31] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Venice, Italy, Oct. 22–29, 2017, pp. 2242–2251.
[32] P. Bojanowski, A. Joulin, D. Lopez-Paz, and A. Szlam, "Optimizing the latent space of generative networks," in *Proc. 35th Int. Conf. Mach. Learn.*, J. G. Dy and A. Krause, Eds., Stockholm, Sweden, Jul. 10–15, 2018, pp. 599–608.
[33] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. 6th Int. Conf. Learn. Representations*, 2018.
[34] R. Webster, J. Rabin, L. Simon, and F. Jurie, "This person (probably) exists. Identity membership attacks against GAN generated faces," 2021, *arXiv:2107.06018*.
[35] B. Adlam, C. Weill, and A. Kapoor, "Investigating under and overfitting in wasserstein generative adversarial networks," 2019, *arXiv: 1910.14137*.
[36] L. Theis, A. van den Oord, and M. Bethge, "A note on the evaluation of generative models," in *Proc. 4th Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Juan, Puerto Rico, May 2–04, 2016.
[37] Z. Li and Y. Zhang, "Membership leakage in label-only exposures," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Y. Kim, J. Kim, G. Vigna, and E. Shi, Eds., 2021, pp. 880–895.
[38] S. Narayanaswamy et al., "Learning disentangled representations with semi-supervised deep generative models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 4–9, 2017, pp. 5925–5935.
[39] Y. Liu, Z. Zhao, M. Backes, and Y. Zhang, "Membership inference attacks by exploiting loss trajectory," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, H. Yin, A. Stavrou, C. Cremers, and E. Shi, Eds., Los Angeles, CA, USA, Nov. 7–11, 2022, pp. 2085–2098.
[40] J. Truong, P. Maini, R. J. Walls, and N. Papernot, "Data-free model extraction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 19–25, 2021, pp. 4771–4780.
[41] T. Humphries et al., "Investigating membership inference attacks under data dependencies," 2021, *arXiv: 2010.12112*.
[42] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Representations*, Y. Bengio and Y. LeCun, Eds., San Diego, CA, USA, May 07–09, 2015.
[43] D. C. Liu and J. Nocedal, "On the limited memory BFGS method for large scale optimization," *Math. Program.*, vol. 45, no. 1–3, pp. 503–528, 1989.
[44] H. Sun, T. Zhu, Z. Zhang, D. Jin, P. Xiong, and W. Zhou, "Adversarial attacks against deep generative models on data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 34, no. 4, pp. 3367–3388, Apr. 2023.
[45] K. Kang, S. Kim, and S. Cho, "GAN inversion for out-of-range images with geometric transformations," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 13 941–13 949.
[46] Y. Wang, C. Wu, L. Herranz, J. van de Weijer, A. Gonzalez-Garcia, and B. Raducanu, "Transferring GANs: Generating images from limited data," in *Proc. 15th Eur. Conf. Comput. Vis.*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., Springer, Munich, Germany, Sep. 08–14, 2018, pp. 220–236.
[47] Y. Wang, A. Gonzalez-Garcia, D. Berga, L. Herranz, F. S. Khan, and J. van de Weijer, "MineGAN: Effective knowledge transfer from GANs to target domains with few images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Seattle, WA, USA, Jun. 13–19, 2020, pp. 9329–9338.
[48] S. Mo, M. Cho, and J. Shin, "Freeze the discriminator: A simple baseline for fine-tuning GANs," 2020, *arXiv: 2002.10964*.
[49] Y. Frégier and J. Gouray, "Mind2Mind: Transfer learning for GANs," in *Proc. 5th Int. Conf. Geometric Sci. Inf.*, F. Nielsen and F. Barbaresco, Eds., Springer, Paris, France, Jul. 21–23, 2021, pp. 851–859.
[50] L. Zhang, T. Zhu, P. Xiong, W. Zhou, and P. S. Yu, "More than privacy: Adopting differential privacy in game-theoretic mechanism design," *ACM Comput. Surveys*, vol. 54, no. 7, pp. 136:1–136:37, 2022.
[51] 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.*, vol. 34, no. 6, pp. 2824–2843, Jun. 2022.
[52] Y. Qu, S. Yu, J. Zhang, H. T. T. Binh, L. Gao, and W. Zhou, "GAN-DP: Generative adversarial net driven differentially privacy-preserving Big Data publishing," in *Proc. IEEE Int. Conf. Commun.*, 2019, pp. 1–6.
[53] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Santiago, Chile, Dec. 7–13, 2015, pp. 3730–3738.
[54] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, NV, USA, Jun. 27–30, 2016, pp. 770–778.
[55] D. Chen, T. Orekondy, and M. Fritz, "GS-WGAN: A gradient-sanitized approach for learning differentially private generators," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., Dec. 6–12, 2020, Art. no. 1063.
[56] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jégou, "White-box vs black-box: Bayes optimal strategies for membership inference," in *Proc. 36th Int. Conf. Mach. Learn.*, K. Chaudhuri and R. Salakhutdinov, Eds., Long Beach, CA, USA, Jun. 9–15, 2019, pp. 5558–5567.
[57] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy*, San Francisco, CA, USA, May 19–23, 2019, pp. 739–753.
[58] J. Zhou, Y. Chen, C. Shen, and Y. Zhang, "Property inference attacks against GANs," in *Proc. 29th Annu. Netw. Distrib. Syst. Secur. Symp.*, San Diego, CA, USA, Apr. 24–28, 2022.

Hui Sun received the BEng and MEng degrees from the Zhongnan University of Economics and Law, China, in 2017 and 2020, respectively. Currently, she is 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.

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 an associate professor with the University of Technology Sydney. 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.



Jie Li received the BEng degree from the China University of Geosciences, Wuhan, China, in 2021. She is currently working toward the MEng degree with the China University of Geosciences. Her research interests include federated learning and neural network security.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.

Shouling Ji (Member, IEEE) received the BS (honors) and MS degrees in computer science from Heilongjiang University, the PhD degree in electrical and computer engineering from the Georgia Institute of Technology, and the PhD degree in computer science from Georgia State University. He is currently a professor with Zhejiang University, Zhejiang, China. He was a research intern with the IBM T. J. Watson Research Center. His current research interests include AI security, data-driven security, software and system security, and Big Data analytics. He is a member of the ACM and CCF and was the membership chair of the IEEE Student Branch at Georgia State University (2012–2013).

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 with 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 IEEE transactions and journals. His research interests include security and privacy, parallel and distributed systems, and e-learning.

Authorized licensed use limited to: Zhejiang University. Downloaded on April 03,2026 at 07:52:26 UTC from IEEE Xplore. Restrictions apply.