A GAN-Based Defense Framework Against Model Inversion Attacks

Xueluan Gong, *Student Member, IEEE*, Ziyao Wang, Shuaike Li, Yanjiao Chen, *Senior Member, IEEE*, and Qian Wang, *Fellow, IEEE*

***Abstract*— With the development of deep learning, deep neural network (DNN)-based application have become an indispensable aspect of daily life. However, recent studies have shown that these well-trained DNN models are vulnerable to model inversion attacks (MIAs), where attackers can recover their training data with high fidelity. Although several defensive strategies have been proposed to mitigate the impact of such attacks, existing defenses will inevitably compromise the model performance and are ineffective against more sophisticated attacks, such as Mirror (An et al., 2022). In this paper, we introduce a novel GAN-based defense approach against model inversion attacks. Unlike previous works that perturb the prediction vector of the model, we manipulate the training procedure of the victim model by incorporating carefully-designed GAN-based fake samples. We also adjust the loss of the inversed samples to inject misleading features into the protected label of the victim model. Additionally, we adopt the concept of continual learning to improve the utility of the model. Extensive experiments conducted on the CelebA, VGG-Face, and VGG-Face2 datasets demonstrate that our proposed method outperforms existing defenses against state-of-the-art model inversion attacks, including DMI (Chen et al., 2021), Mirror (An et al., 2022), Privacy (Fredrikson et al., 2014), and AMI (Yang et al., 2019). It is shown that our proposed method can also retain a high defense performance in black-box scenarios.**

***Index Terms*— Model inversion attacks, GAN-based fake sample generation, privacy-utility defense framework.**

I. Introduction

DEEP learning has demonstrated remarkable success in various real-world applications, including autonomous driving [6], object detection [32], and face recognition [31]. To achieve high performance, deep neural network (DNN) models are trained using sensitive and proprietary training samples, which raises significant concerns regarding data privacy. Recently, various studies [4], [45] have revealed that well-trained deep learning models are vulnerable to model inversion attacks. Model inversion attack (MIA) [3], [46] can reconstruct training data samples related to a given label from the model's prediction. For instance, MIA can be used to uncover the face of an individual from a face recognition model [45]. Advanced MIA methods (e.g., Mirror [5]) are even effective against deep neural networks trained on high-resolution datasets. A high-fidelity recovered face can even pass the access control system, leading to a serious security breach. Model inversion attacks can be executed with either white-box access (full knowledge of the model architecture and parameters) or black-box access (no knowledge of the model's internal workings) to the target model.

The existence of model inversion attacks underscores the critical need to address data privacy concerns in deep learning. Researchers and practitioners are actively exploring techniques to mitigate the damage caused by model inversion attacks. Fredrikson et al. [13] proposed to use differential privacy (DP) as a defense mechanism, due to its theoretical guarantees for protecting training data privacy. However, it has been discovered that DP is ineffective against model inversion attacks while preserving model utility [45]. Yang et al. [39] introduced a unified purification framework, which trains a purifier model to reduce the information contained in the confidence vector while maintaining prediction accuracy. Wen et al. [39] proposed to add adversarial noise to the model's prediction output, with the goal of maximizing inversion error while minimizing the impact on the victim model's utility. Wang et al. [37] introduced a mutual information regularization-based defense framework, aimed at reducing the information related to training data samples in the prediction results. Despite the superiority of these defenses over DP-based defenses, existing methods either impact model utility or are ineffective against advanced model inversion attack methods, such as Mirror [5].

In this paper, we present a novel GAN-based defense framework for protecting deep neural networks against model inversion attacks. Instead of adding noise to the output prediction vector, our approach manipulates the target model's training process by incorporating specially designed fake samples. These fake samples deceive the attacker's generator into inverting samples that are significantly different from authentic ones of the protected label. To enhance the defense's effectiveness, we construct two types of fake samples: inversed public samples and inversed private samples. The inversed public samples are reconstructed samples from the classifier trained on public data, while the inversed private samples

Manuscript received 1 April 2023; revised 18 June 2023; accepted 10 July 2023. Date of publication 17 July 2023; date of current version 28 July 2023. The work of Qian Wang was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0107701 and in part by NSFC under Grant U20B2049 and Grant U21B2018. The work of Yanjiao Chen was supported by NSFC under Grant 61972296. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Chia-Mu Yu. *(Xueluan Gong and Ziyao Wang contributed equally to this work.) (Corresponding authors: Yanjiao Chen; Qian Wang.)*

Xueluan Gong is with the School of Computer Science, Wuhan University, Wuhan 430072, China (e-mail: xueluangong@whu.edu.cn).

Ziyao Wang, Shuaike Li, and Qian Wang are with the School of Cyber Science and Engineering, Wuhan University, Wuhan 430072, China (e-mail: wangziyao@whu.edu.cn; lishuaike@whu.edu.cn; qianwang@whu.edu.cn).

Yanjiao Chen is with the College of Electrical Engineering, Zhejiang University, Hangzhou 310007, China (e-mail: chenyanjiao@zju.edu.cn).

Digital Object Identifier 10.1109/TIFS.2023.3295944

are reconstructed samples of the target victim model. When retraining the target model, we adjust the loss of the inversed samples by maximizing the loss on inversed private samples and minimizing the loss on inversed public samples, thereby injecting misleading features into the victim model's protected label. Besides, we introduce the focal loss in the fine-tuning process to further enhance the protection performance. To maintain high prediction accuracy of the target model, we employ a continual learning algorithm, specifically elastic weight consolidation, during the fine-tuning process. We conducted extensive experiments to evaluate the effectiveness of our method against four state-of-the-art model inversion attacks. Experimental results on CelebA, VGG-Face, and VGG-Face2 demonstrate that our approach outperforms existing defenses, including DP and Ad-mi.

This paper is an extended version of our previous paper [16], which was published in 2023 International World Wide Web Conference (WWW). We extend our previous work by considering a more realistic black-box setting. In this paper, we assume that the defender has no prior knowledge of the attacker's strategy for launching a MIA. Additionally, if the MIA is carried out using a Generative Adversarial Network (GAN), we also assume that the defender lacks information regarding the GAN's training data, gradient information, parameters, and even the model structure. In this case, we construct a substitute GAN model on a public dataset to imitate the attack GAN model. This extension increases the practicality of our work for real-world applications, where it is often challenging to acquire any information about the attacker. The robust defense performance of our proposed method suggests that the fake sample generation algorithm effectively extracts the universal features of target victim images, rather than being dependent on features specific to the model structure. We have included comparison results with baseline defenses in a black-box scenario to demonstrate its effectiveness. We have also conducted a comprehensive ablation study in the black-box scenario to further explore its capabilities. Unlike our previous work that used cross-entropy loss, we introduce focal loss during fine-tuning to reinforce the misleading features and further enhance the defense performance. We applied the focal loss to both inversed public image loss and inversed private image loss to induce overfitting on the target model. Our experiments demonstrate that such overfitting can push the inversed images further away from the victim images in feature space.

We make the following contributions.

- We proposed a novel GAN-based model inversion defense framework against deep neural networks. To the best of our knowledge, this is the first attempt to leverage the use of "fake" samples during the training of the victim model, as a means of mitigating model inversion attacks.
- In our defense framework, we design a novel GAN-based fake sample generation algorithm, adjust the loss of both inversed public and private samples, introduce the focal loss in the fine-tuning process, and introduce continual learning to further improve the utility.
- We have conducted extensive experiments on CelebA, VGG-Face, and VGG-Face2 to verify the effectiveness of our proposed method to mitigate state-of-the-art MIAs while maintaining the utility of the victim model. It is shown that our proposed method can also maintain a high defense performance in black-box scenarios.

II. Preliminaries and Related Work

In this section, we give a brief summary of state-of-the-art model inversion attacks and defenses. Then we introduce the preliminaries of continual learning and generative adversarial network. We also introduce surrogate model-based adversarial attacks and defenses.

A. Model Inversion Attacks

Training data inference attacks [15] against machine learning models can be divided into three categories according to the target revealing information: membership inference attacks, attribute inference attacks, and model inversion attacks. In membership inference attacks, the attacker aims to deduce whether a particular record belongs to the original training dataset or not [34], [42]. In attribute inference attacks, the adversary aims to reconstruct the missing attributes based on the machine learning model and partial information [14], [27], e.g., location and political view. In model inversion attacks, the attacker's goal is to reconstruct sensitive features of samples that belong to a certain class [12], [41]. For example, an identifiable image of a person can be reconstructed by model inversion attacks against a facial recognition model that recognizes the name of the person as one of its output classes. Overall, the impact of model inversion attacks is the most catastrophic. In this paper, we focus on model inversion attacks, which can be characterized from two aspects: attack strength and the learning scenario.

1) Attack Strength: Existing works on model inversion attacks evolve from white-box settings to black-box settings.

a) White-box model inversion attacks: Fredrikson et al. [13] proposed the first model inversion attack algorithm (a white-box attack) that formulates the attack as an optimization problem. Given the model output and some auxiliary information, it seeks for a private data sample that can achieve the maximum likelihood or posterior probability. Such an algorithm is shown to be effective for recovering genetic markers given the linear regression model that uses them as input features. Fredrikson et al. [12] then explored the application of the algorithm to other relatively complex machine learning models, such as decision trees and neural networks. However, in the task of reconstructing face images, the reconstructed faces are too blurry to recognize. Moreover, such a general attack algorithm is shown only to be effective for shallow neural networks.

Recently, various attack works have been designed to recover training samples of deep neural networks. Yin et al. [43] proposed DeepInversion that draws support from a trained CNN (teacher) to generate class-conditional input images with high resolution and fidelity. Given a fixed teacher model, DeepInversion optimizes the input while regularizing the distribution of intermediate feature maps using information stored in the teacher's *batch normalization* (BN) layers. However,

DeepInversion cannot be applied for deep neural networks without BN layers, such as VGG-16 and SphereFace.

Zhang et al. [45] proposed GMI (generative model-inversion) that utilizes GAN (generative adversarial network) model to inverse the training data samples. GAN can be used to generate in-distribution samples. It consists of two parts: generative network and discriminative network. Generative network is used to generate fake samples from a pre-defined latent space. The generated fake samples are expected to fool the discriminative network. The generative network and the discriminative network are trained in a competitive way. A GAN that is well-trained on face samples can easy to construct natural human facial images. Given a latent vector, GMI uses GAN to generate a sample, so that the target model recognizes it as the target label, and the discriminative network considers it as a real image. However, using such a traditional GAN structure incurs two defects. First, GMI can only inverse low-resolution samples of 64×64, thus the reversed human faces are also blurry to recognize the exact human. Second, the latent space is entangled [23], i.e., the features lack locality in the space. Therefore, space exploration is easy to stuck with local optimals.

More recently, Chen et al. [9] proposed DMI, an inversion-specific GAN to distill more useful knowledge from the public dataset when performing the model inversion attacks. In particular, its discriminator is trained to differentiate not only the real and fake samples but also the soft-labels [9] provided by the target victim model. Such an attack method can recover the data distribution for each class rather than only a single representative data. Experiments have shown that even if the public data trained for GAN has no relationship with the original training data, this work is also effective.

b) Black-box model inversion attacks: Yang et al. [41] proposed a MIA framework, namely AMI, targeting black-box deep neural networks. The adversary first trains an inversion model on an auxiliary dataset that is collected based on background knowledge. For example, the adversary can randomly crawl human faces from the Internet to build an auxiliary set for attacking a face recognition model. Then given the victim model's confidence scores, the inversion model can inverse the original input samples. The procedure of AMI is similar to the autoencoder, where the black-box model is the "encoder", and the inversion model is the "decoder", and the prediction vector acts as the latent space. However, the inversed facial samples generated by AMI are also too blurry to make them clearly recognized by human visuals.

Recently, An et al. [5] proposed Mirror, which is considered to be the current best-performing model inversion attack. Unlike the existing works that are based on traditional GAN, Mirror utilizes StyleGAN [23] to perform the model inversion attacks. The generator of StyleGAN has a dedicated architecture to force the decomposition of input into styles of various granularities. Given a target model and a well-trained StyleGAN, Mirror uses gradient back-propagation to seek the best parameterization for its generator, so that the generated sample is classified/recognized to the target label by the target model/human eyes. Note that the StyleGAN is trained on a public dataset that is from the same domain as the target model. To improve the image quality, the authors proposed to tune the optimization in an auxiliary space that follows a multivariate Gaussian (MVG) distribution in StyleGAN. Besides, to improve the image fidelity, the authors proposed to use random drop-out to alleviate the target model feature overfitting.

2) Learning Scenario: Apart from the above-mentioned attacks in centralized learning, model inversion attacks can also apply in collaborative learning scenarios. In the collaborative learning scenario, the attacker is assumed to be one of the participants.

Hitaj et al. [20] demonstrated that a rancorous participant can use a Generative Adversarial Network (GAN) to reconstruct representative samples of a certain label of another victim participant. In particular, the attacker first trains a GAN to generate samples that are similar to the victim's training dataset, and then feeds these toxic samples into the collaborative learning process and uploads the local model to the server. In this way, the victim is required to redouble his efforts to distinguish the toxic dataset from the truthful ones. It may force the victim to reveal more information about its own training datasets than expected. Zecheng et al. [18] proposed a model inversion attack against collaborative inference systems under the white-box, black-box setting, and query-free black-box settings (i.e., without queries). The authors discovered that only one attacker is enough to recover the inference samples from the intermediate values. To recap, in [18], *regularized Maximum Likelihood Estimation* (rMSE), *Inverse-Network*, and *query-free shadow* model reconstruction techniques are used in white-box, black-box, and query-free black-box settings, respectively.

In this paper, we mainly focus on the model inversion attacks in the centralized learning scenario. Since our defense framework is a training-time remedy, it can also be easily applied to collaborative learning scenarios.

B. Model Inversion Defenses

As far as we know, the field of defending against model inversion attacks is relatively limited, and there are few existing works specifically targeting this area. Designing a utility-privacy defense strategy for model inversion attacks remains a significant and valuable direction that warrants further exploration and research. We categorize the current model inversion defense strategies into four distinct categories.

1) DP-Based Defenses: Fredrikson et al. [13] highlighted the potential of using differential privacy (DP) [1] as a defense mechanism against model inversion attacks. DP-based defense involves adding noise to various values or parameters to obscure information and provide theoretical guarantees for training data privacy. However, it does not mean protecting the entire distribution. Zhang et al. [45] discovered that DP has no ability to provide data privacy against model inversion attacks while maintaining the model utility. Wang et al. [37] also provided a theoretical analysis to explain why DP is ineffective for defending against model inversion attacks. Moreover, differential privacy will incur a significant decrease in the model prediction accuracy [30].

2) Prediction-Purified-Based Defenses: Defenders can purify the prediction output of the victim model to mitigate the attacks. Yang et al. [40] proposed a unified *purification framework* to defend against membership inference attacks and model inversion attacks. In particular, the defender trains a purifier model and applies it to the model output vector. The purifier model aims to minimize the information contained in the returned confidence vector while maintaining the model's prediction accuracy. To defend against model inversion attacks, the framework reduces the dispersion of model prediction confidence score vectors on both members and non-members. Therefore, the prediction output is more robust to the change of input samples, i.e., the correlation between the training sample and the prediction vectors is weakened. To this end, the attacker can only get an inaccurate reversed version when conducting MIA against such target models. To defend against membership inference attacks, the framework reduces the distinguishability of confidence score vectors between members and non-members.

However, it has been shown that [40] will introduce significant utility loss when normal users query the defended target model [39]. Furthermore, when the purification framework is deployed, the reconstructed images also tend to exhibit characteristics of average faces and retain certain prominent facial features that can be utilized for individual identification.

3) Prediction-Disturbed-Based Defenses: Defenders can add carefully designed noise to the prediction output to mislead the attackers. Wen et al. [39] proposed to add adversarial noise to the model prediction output. Such adversarial noise requires maximizing the inversion error and introducing negligible utility loss to the victim model. Specifically, [39] exploits the gradients of the inversion model to simulate an adversary and computes a noise vector to transform the output into an adversarial example that can maximize the reconstruction error of the inversion model. Then it applies a label modifier to retain the original predicted label to achieve zero accuracy loss. Although [39] can introduce zero accuracy loss, it is ineffective against more advanced model inversion attacks, such as Mirror [5].

4) Relationship-Regularization-Based Defenses: Wang et al. [37] proposed a mutual information regularization-based defense (MID) against model inversion attacks. Its intuition is to reduce the information related to the training data samples when outputting the prediction results. Thus the attacker cannot recover training samples via deducing the model prediction. Unlike the existing defenses, [37] manipulates the target model training procedure by introducing a regularizer into the training loss. The regularizer will penalize the mutual information between the input samples and the model output. However, after applying this defense, the target model cannot achieve a high clean data accuracy as the undistorted model. Moreover, such a defense is shown to be ineffective for more advanced model inversion attacks [5].

Unlike the existing defenses that either impact the model utility or are ineffective for advanced MIA methods, in this paper, we proposed a novel model inversion defense framework that achieves a better utility-privacy trade-off. Moreover, experiments have shown that our proposed defense is effective for advanced model inversion attacks, including Mirror, AMI, and DMI.

C. Continual Learning

Continual learning (a.k.a. incremental learning, life-long learning) [11] algorithms continuously and adaptively learn from data/task streams over time, enabling the incremental development of more complex knowledge and skills. Continuous learning can be used to deal with the catastrophic forgetting problem, where deep neural networks quickly override previously learned knowledge when sequentially trained on new data [35]. Recently continual learning has been applied in various domains, such as computer vision [8] and speech recognition [21]. In our work, to maintain the utility of our protected model, we propose to introduce continual learning in the model fine-tuning procedure. Even though we can retrain the protected model using the entire training dataset, it will incur huge time and computational overhead. It is shown [28] that through continual learning, 1% training dataset is sufficient for achieving a high model performance in most cases.

Specifically, we use elastic weight consolidation [24] that puts a constraint on the parameter update process. For parameters set θ_A^* in the model, the weight matrix F_θ in the loss function can be:

$$F_\theta = [\bigtriangledown \log(P(y_n|x_n, \theta_A^*) \bigtriangledown \log(P(y_n|x_n, \theta_A^*)^T], \quad (1)$$

where y_n and x_n denote the labels and training samples of the previous task. During the training process of the new task, an EWC loss item is added to weaken the update of model parameters, according to the weight matrix F_θ. The more critical the parameter to the previous task, the weight will be more significant to provide tighter restrictions:

$$L_{cont} = \sum_i [\lambda F_i(\theta_i - \theta_A^*, i)^2]. \quad (2)$$

D. Generative Adversarial Network

Generative Adversarial Network (GAN) was proposed to measure the efficacy of the generative model with an additional adversarial process [17]. In GAN, there are two models that need to be trained synchronously: the generative model G and the discriminative model D. As for the generative model G, it concentrates on matching the data distribution. Specifically, it draws random samples z from the prior distribution (such as uniform or Gaussian) as input and then generates samples from z. In terms of the discriminative model D, it tries to distinguish generated samples from real samples. In short, G is trained to maximize the possibility that D makes errors in distinguishing the real samples and the generated samples, and D is trained to potentiate the ability of differentiation. Mathematically, such a cat-and-mouse game can be profiled as the following value function:

$$\min_G \max_D V(G, D) = \mathbb{E}_{x\sim p_{data}(x)}[log\, D(x)] + \mathbb{E}_{z\sim p_z(z)}[log(1 - D(G(z)))], \quad (3)$$

where p_z denote the prior distribution, p_{data} means the training distribution. The generative model G and the discriminative model D are trained until such a minimax game reaches the Nash equilibrium, in which the generated samples (fake samples) are difficult to distinguish from the real ones. In theory, the global optimal value is achieved when $p_{data} = p_g$ [17], where p_g represents the distribution of the generated samples.

Due to its powerful generative ability, GAN is now widely applied to various attack scenarios, such as backdoor trigger generation [33] and inversed sample generation [5], [45]. To generate high-fidelity inversed samples, as far as we know, most recent advanced model inversion attacks introduce generative models [5], [45]. With the development of the GAN structure (from traditional GAN to StyleGAN [23]), the effect of the GAN-based model inversion attack is also getting better and better.

Unlike the existing model inversion attacks that aim to generate high-quality inversed samples, in this paper, we utilize GAN model to construct more effective fake samples to fool the attack model deployed at the attacker end.

E. Surrogate Model-Based Adversarial Attacks and Defenses

In the field of adversarial attacks, various methods utilize surrogate models to enhance the effectiveness of adversarial attacks or to improve the robustness of defenses in black-box scenarios, such as [22], [36], [38], and [47]. In these scenarios, the attacker or defender has limited or no access to the target model's architecture, parameters, or gradients. They can only access the output of probabilities/labels from the target model.

The idea behind surrogate model-based adversarial example attacks is to train a substitute model, often referred to as the surrogate model, which approximates the behavior of the target model as closely as possible. The surrogate model training process involves utilizing input-output pairs obtained by querying the target model. Once the surrogate model is trained, it can be used by the attacker to generate adversarial examples that are misclassified by the target model.

For example, Zhou et al. [47] proposed a data-free surrogate model training method called DaST, which obtains surrogate models for black-box attacks without the requirement of any real data. DaST utilizes specially designed GANs to train the substitute models. The core idea is to use a generator to construct synthetic images and label them with the target model, similar to query-based attacks. The surrogate model is then trained with these images to better replicate the decision boundaries of the target model. Another approach is dynamic substitute training (DST) proposed by Wang et al. [38] for data-free black-box attacks. DST introduces a dynamic surrogate structure learning strategy that adaptively generates an optimal surrogate model structure using a dynamic gate, based on different target models and tasks. This dynamic surrogate model can learn more effectively and efficiently from the target model.

In addition to constructing black-box adversarial examples, surrogate models can also be applied in the field of Deepfake disruption. DeepFake disruption aims to proactively disrupt images that are being manipulated by DeepFake techniques by adding adversarial perturbations. Huang et al. [22] proposed an initiative defense framework against facial manipulation in black-box settings. They first imitate the target manipulation model with a surrogate model and then devise a poison perturbation generator to generate the desired venom, which is used to disrupt the DeepFake manipulation. An alternating training strategy is employed to train both the surrogate model and the perturbation generator. To further enhance defense robustness, Wang et al. [36] proposed a novel anti-forgery technique to protect shared facial images from attackers who employ popular forgery techniques. Anti-forgery can generate perturbations that are imperceptible to human eyes while being resistant to potential input transformation attacks. It converts the input image from the RGB color space to the Lab color space for adding perceptual uniform perturbations to the a and b channels. And the perturbations are updated by attacking the surrogate model using an optimization-based strategy.

In contrast to existing surrogate model-based adversarial attacks and defenses that use a surrogate model to mimic the victim target model and generate perturbations, our approach involves using a GAN model to generate fake samples. These fake samples are intended to mislead the generator of the model inversion attack into inverting samples that are significantly different from the authentic ones associated with the protected label. The GAN model employed in our approach can utilize various common GAN structures, such as CGAN, WGAN, or InfoGAN. We will also show that even if the MIA is not GAN-based (e.g., Privacy [13], and AMI [41]), our generated fake samples can also successfully mislead the attacker.

III. Threat Model

In model inversion attacks (MIA), there exist two entities: a model trainer (defender) and an attacker.

A. Attacker-End

Given a deployed DNN model for a particular task, the attacker aims to recover the training data of the model via querying it with an auxiliary dataset. Specifically, we assume the victim model is trained on data distribution (X, Y). Given specific label y, the adversarial goal of model inversion attack is to get information $\mathrm{P}(\mathrm{X}|f(\mathrm{X}) = y)$. Considering the target model is a face recognition model that labels an input sample with an individual identity, the attack goal is to reconstruct a high-fidelity image for a given identity according to the access to the model. The attacker may have an auxiliary dataset that consists of multiple human faces downloaded from the Internet, and the auxiliary dataset has no overlap with the original training data of the target model. There exist both white-box attackers and black-box attackers. A white-box attacker has unrestricted access to the victim model and its parameters. And a black-box attacker can only access the victim model via an API interface, which is more realistic in the real world. Since defending against a white-box attacker is more challenging, we consider that the attackers can access our protected victim model when launching the model inversion attacks.

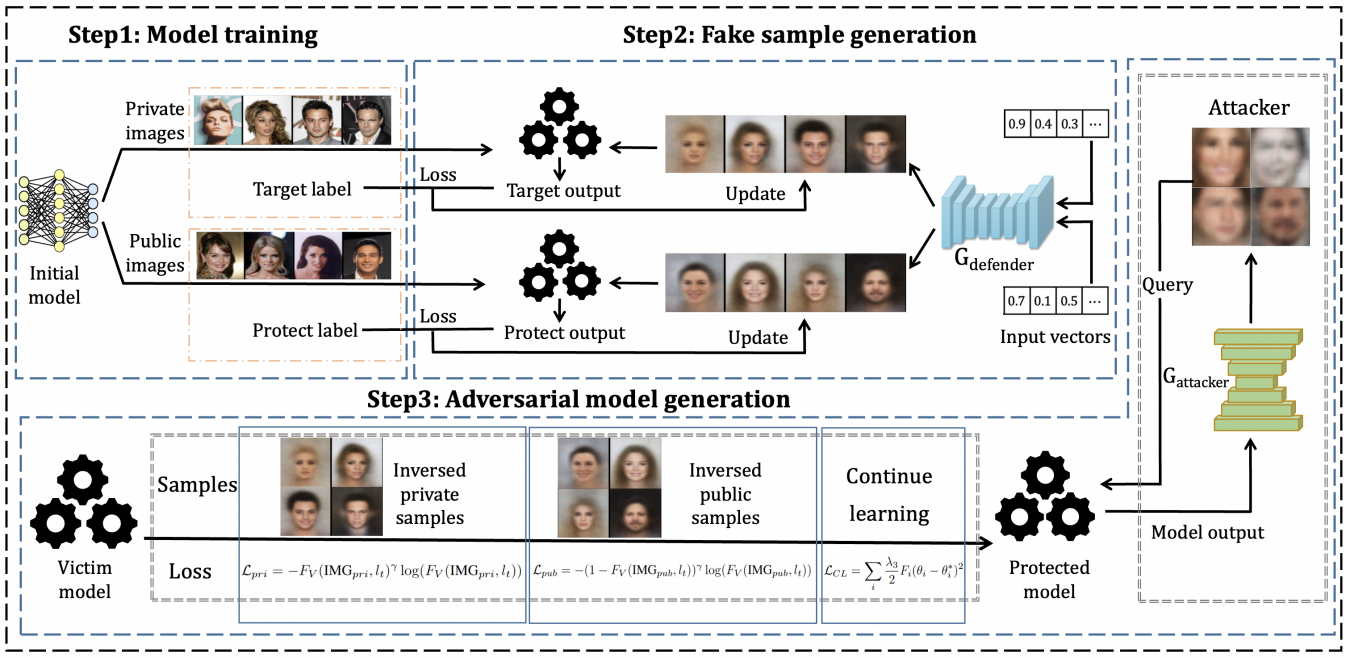

Fig. 1. Overview of our proposed method.

B. Defender-End

The defender is the victim model trainer, e.g., a cloud service provider whose goal is to protect the training sample privacy from model inversion attacks. The inversed samples of the protected label are not like their original ones. The attacker cannot infer the appearance of samples in the protected label through model inversion attacks. Besides, the trained DNN model should maintain its functionality, i.e., maintain a high prediction accuracy on input query samples. We assume that the defender has the following capabilities and limitations.

- *Access to victim dataset*. The defender has access to original training samples of the victim model and has the ability to add arbitrary samples to the training dataset.
- *Access to victim model*. The defender trains the victim model, thus it knows and can alter anything of the victim model, including the architecture, parameters, and hyperparameters.
- *Access to public dataset*. The defender can access commonly-used public datasets.

We assume that the defender has the following limitations.

- *No knowledge of the specific MIA method*. The defender does not know the exact model inversion attack strategy adopted by the attacker.

If the MIA is a GAN-based attack, we consider two kinds of situations based on whether the defender knows the GAN model structure of the attacker or not. In either situation, the defender has no information about the training data, parameters, and gradient information of the attacker's GAN model.

IV. Detailed Construction

Unlike existing works that add perturbation to the prediction vector, we proposed to manipulate the victim model training by adding some carefully-designed "fake" samples. Such "fake" samples can mislead the generator of the attacker to inverse samples that are much more different from the authentic ones of the protected label.

To mislead the existing model inversion attacks, we utilize a GAN model to generate high-quality misleading fake samples. In our proposed method, there are two kinds of fake samples, i.e., inversed public samples and inversed private samples. The inversed public samples are reconstructed samples of the classifier trained on a public dataset. The inversed private samples are reconstructed samples of the target victim model.

We then add the fake samples to the original training dataset and fine-tune the target model. To improve the defense performance, we design loss functions for the two kinds of fake samples with the aim of misleading the attacker to obtain inversed images with similar characteristics to the public images. To maintain a high prediction accuracy of the victim model, we propose to use a continual learning algorithm when fine-tuning the target model. The overall procedure of our proposed method is shown in Fig. 1. Overall, our proposed defense framework includes three processes: model training, fake sample generation, and model fine-tuning. The model training consists of three major models used in defense: victim model training, classifier training, and GAN model training.

A. Model Training

1) Victim Model Training: We first train the victim model with a standard cross-entropy loss $\mathcal{L}_{CE}$ on the private training dataset. Note that if a pre-trained victim model is already available, we can directly use such a victim model. The model parameters are calculated by minimizing the loss function $\mathcal{L}$ by stochastic gradient descent (SGD).

$$\theta^* = \arg\min \mathcal{L}_{CE}(\theta; f(x), y), \tag{4}$$

where θ is the parameter vector of the victim model, θ^* is the learnt parameter after pre-training stage, $f(\cdot)$ is the functionality of target function, and (x, y) is the training sample drawn from data distribution (X, Y).

2) Classifier Training: Unlike the victim model which is trained on the private dataset, the classifier is trained on a public dataset. The public dataset is with the same problem domain as the private dataset. For example, if the victim model is a face recognition classifier that is trained on some individuals, the defender can select CelebA or VGG-Face from the Internet as the public dataset. The training procedure of such a classifier follows the standard DNN model training paradigm. In the public dataset, we randomly select one label, l_p, as the protected label. This means that when an attacker launches a model inversion attack (MIA) against the target label of the victim model protected by our defense framework, he can only infer information about samples with similar characteristics to the protected label l_p in the public dataset, rather than private samples of the target label in the private dataset.

3) GAN Model Training: To counteract a model inversion attack where the defender does not have access to the attacker's generator, the defender requires to construct his own GAN model to imitate the MIA attack. We train the GAN on a public dataset. After training, the trained GAN is able to generate visually natural images.

Assuming the defender knows the structure of the GAN model used by the attacker, the defender can employ the same structure to train the GAN model. However, in a black-box scenario where the defender knows nothing about the attacker's GAN, including its structure, they must resort to using a substitute GAN. The substitute GAN should be capable of generating high-quality and natural images. In our experiments, we used the widely-used Wasserstein GAN [2] as the substitute GAN structure. The GAN training procedure follows the standard GAN training paradigm described in Section II-D. We also investigated the influence of various GAN structures, including CGAN [29] and InfoGAN [10], on the defensive performance.

For ease of reference, we set F_V, F_C, and G as the victim model, classifier, and GAN model, respectively, in the following text.

B. Fake Sample Generation

In the fake sample dataset, we collect two kinds of fake samples: inversed samples of the public dataset and inversed samples of the private dataset.

1) Inversed Public Samples: Such kind of fake samples is to deceive the attacker into reversing samples with similar characteristics to the protected label in the public dataset. Given a protect label l_p in F_C and the GAN model G, we expect to get the input for G that can output images classified to l_p in F_C. To achieve this, we first initialize the z_{pub} vector, which is the input for the generator in G. Then, in each epoch, we gain output images img_{pub} from G, input them into F_C, and get the output label. After that, we compute the MSE loss between the output and l_p, backward the loss, and update the GAN input vector z_{pub}. The loop stops when it reaches the limit of the epochs' number. The algorithm will output the z_{pub} vector that can achieve the least label loss in F_C. Then, the inversed public samples are generated from the output of GAN with z_{pub} input. These samples are actually what we expect the attacker can inverse from the victim model protected by our proposed method.

2) Inversed Private Samples: This kind of fake sample aims to prevent the private samples of the target label from being reconstructed. The process of inversed private sample generation is similar to that of inversed public sample generation. There are two major differences. Firstly, the target model of the model inversion attack is F_V rather than F_C. Second, the label used for loss computing is the target protected label l_t of the victim model F_V. Following the same procedure, we aim to obtain inversed samples generated by GAN that can be correctly classified to l_t through model inversion. We protect these inversed private samples from inferring by the attacker.

In practice, we simultaneously construct the two parts of the fake samples. The input vector process can be described as the loss-minimize process of these two loss functions.

$$\begin{aligned} L_{discri} &= -\log(D(G(Z))), \\ L_{iden} &= L_{CE}(F(G(Z)), L), \end{aligned} \tag{5}$$

where L_{discri} is discriminator loss and L_{iden} is identification loss. These two loss items are used to ensure the quality of inversed images and their classification accuracy, respectively. In public sample optimization, $F = F_C$, $Z = z_{pub}$, $L = l_p$, and in private sample optimization, $F = F_V$, $Z = z_{pri}$, $L = l_t$. The sample optimization process is as follows:

$$Z^* = \arg\min L_{discri} + \lambda L_{iden}, \tag{6}$$

where Z^* is the learned z_{pub} vector in public sample optimization and the learned z_{pri} vector in private sample optimization. Besides, λ is the coefficient for L_{iden}.

C. Model Fine-Tuning

After collecting the fake samples, the defender requires to design loss functions for the two kinds of fake samples and then fine-tune the victim model. As for the inversed private samples, they will be classified into l_t by F_V. If their identification loss L_{iden} is low, they are considered as good inversed images for the attacker. Thus in the fine-tuning process, we fine-tune the victim model to maximize its loss L_{pri}.

Specifically, we design the loss $\mathcal{L}_{pri}$ as follows.

$$\begin{aligned} \mathrm{IMG}_{pri} &= G.generator(z_{pri}), \\ \mathcal{L}_{pri} &= -F_V(\mathrm{IMG}_{pri}, l_t)^{\gamma} \log(F_V(\mathrm{IMG}_{pri}, l_t)), \end{aligned} \tag{7}$$

where IMG_{pri} denotes the inversed private images generated by GAN. $F_V(\mathrm{IMG}_{pri}, l_t)$ is the possibility of IMG_{pri} being classified to l_t by F_V, where the value lies between 0 and 1. We expect all inversed private images are not classified to 1. Thus, we have to maximize $\mathcal{L}_{pri}$. In addition, we adapt the idea of focal loss [25], which adds a loss weight for every single training sample to adjust the model focus to help the convergence and overfitting of the victim model on hard-to-train samples.

During the fine-tuning process, we observed that the model tends to overfit on either inversed private samples or inversed public samples, while poorly training on the remaining type of inversed samples. By utilizing focal loss to improve the weights of these hard-to-train samples, we achieve a more balanced fine-tuning process, preventing overfitting on any inversed sample type. Therefore, focal loss can further enhance the overall defense performance of our proposed method. We incorporate a loss weight $F_V(\text{IMG}_{pri}, l_t)^\gamma$ for each inversed private sample to maximize $\mathcal{L}_{pri}$. If a sample is not classified to l_t (when $F_V(\text{IMG}_{pri}, l_t)$ is close to 0), its weight will be close to 0. Thus, the loss function will not focus on it much. On the contrary, if a sample is classified to l_t (when $F_V(\text{IMG}_{pri}, l_t)$ is close to 1), its weight will be close to 1. This loss weight item can help the model focus more on poorly-trained samples and enhances the overfitting of the entire set of inversed private images on the victim model.

By maximizing $\mathcal{L}_{pri}$, we can maximize the feature distance between the inversed samples and the private samples of the target label l_t, thereby preventing the attacker from learning private images of the target label. Through experiments, we found that just maximizing equation (7) is not enough to make F_V fully robust to model inversion attacks. To address this issue, we inject misleading features into l_t by minimizing the L_{pub} loss using the inversed public samples.

$$\begin{aligned} \text{IMG}_{pub} &= G.generator(z_{pub}), \\ \mathcal{L}_{pub} &= -(1 - F_V(\text{IMG}_{pub}, l_t))^\gamma \log(F_V(\text{IMG}_{pub}, l_t)). \end{aligned} \tag{8}$$

IMG_{pub} denotes the inversed public images, and $\mathcal{L}_{pub}$ minimizes the loss of inversed public images to mislead the inversion attack. Similarly, we add a focal weight here to adjust the model focus. Different from (7), in this case, the model training should focus on inversed public images that cannot be classified to l_t, i.e., $F_V(\text{IMG}_{pub}, l_t)$ is closed to 0. Therefore we use $(1 - F_V(\text{IMG}_{pub}, l_t))^\gamma$ (a number close to 1) as a weight to adjust the loss. For those images that are well-classified to l_t, we decrease their weight to 0 using $(1 - F_V(\text{IMG}_{pub}, l_t))^\gamma$. The minimization process can make the victim model overfit on the inversed public images, so that the attacker is more likely to obtain inversed images with similar characteristics to these public images [37]. We set γ as 2 in our experiments.

To maintain a high prediction accuracy of the victim model, we also use elastic weight consolidation loss as continual learning regularizer to restrict the parameter update during fine-tuning. We define parameters F_i to identify the parameter importance on the main classification task and control changes on parameters, especially the more important ones, to maintain utility.

$$\mathcal{L}_{CL} = \sum_i \frac{\lambda_3}{2} F_i(\theta_i - \theta_i^*)^2. \tag{9}$$

To conclude, the overall fine-tuning process can be described as the following optimization function:

$$\theta = \arg\min -\alpha\mathcal{L}_{pri} + \beta\mathcal{L}_{pub} + \omega\mathcal{L}_{CL}, \tag{10}$$

where θ is the parameter set of the victim model F_V, and α, β, and ω are the coefficients for the three loss items.

TABLE I
MODEL ACCURACY OF THE VICTIM MODELS AND EVALUATION MODELS

Models	CelebA	VGG-Face	VGG-Face2
Victim model	87.46%	82.41%	92.03%
Evaluation model	88.93%	84.74%	94.95%

V. EVALUATION SETUP

A. Datasets and Models

We conduct experiments on three widely-used datasets, i.e., CelebA [26], VGG-Face [31], and VGG-Face2 [7].

1) CelebA: CelebA [26] is a large-scale face attribute dataset, which consists of more than 200,000 celebrity images. Each image is meticulously annotated with 5 facial landmarks, including the locations of the two eyes, nose tips, and mouth corners. Additionally, it provides 40 binary attribute annotations for each image. CelebA possesses several distinctive characteristics such as large diversities, large quantities, and rich annotations. Furthermore, CelebA exhibits approximate gender balance, with around 40% of the images featuring female celebrities. To train a deep neural network on CelebA dataset, we set the base learning rate as 0.01, resize the input image to $3 \times 64 \times 64$, and the mini-batch size is set as 32.

2) VGGFace: VGGFace [31] includes 2.6 million face images collected from 2,622 humans, which has a dimension of 224×224. This dataset covers a wide range of variations in facial appearance, including different poses, expressions, and lighting conditions. In the experiments, we resize the 3-channel images to $3 \times 112 \times 112$ and then train a deep neural network with a learning rate of 0.01. The mini-batch size is also set as 32.

3) VGGFace2: VGGFace2 [7] includes 9,131 identities, and the image number of each identity is from 80 to 800. All the images in VGGFace2 were obtained from Google Images, and it contains 3.31 million images, with an average of 362.6 images for each identity. Compared with VGGFace, VGGFace2 includes more Chinese and Indian faces. VGGFace2 is also approximately gender-balanced, with 59.3% males. In the experiments, we resize VGGFace2 images to $3 \times 224 \times 224$. The learning rate is 0.01, and the mini-batch size is 16.

Among all three datasets, we choose 1000 labels' face images to train the model. For each label, we randomly pick 2 face images as the test set, and the rest as the training set to ensure the accuracy of the model.

We use VGG-16 as the victim model structure for each dataset. VGG-16 has 16 layers, 13 of which are convolutional layers, and 3 are fully connected layers. Following [9], we utilize face.evoLve (also called FaceNet) to train the evaluation model for each dataset. FaceNet is a 50-layer convolutional backbone model augmented with a fully connected output layer. The model accuracy results of each victim model and evaluation model are shown in Table I.

TABLE II
COMPARISON OF OUR PROPOSED METHOD WITH DP AND AD-MI AGAINST STATE-OF-THE-ART MODEL INVERSION ATTACKS

		Original			DP				Ad-mi				GAN-D				GAN-ID			
		MC(l_t)	AA	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID
CelebA	DMI [9]	87.46%	52%	108	40%	56.82%	50%	101	**0%**	87.46%	**100%**	153	**0%**	86.33%	**100%**	**188**	**0%**	86.65%	**100%**	168
	Mirror [5]	87.46%	69%	98	45%	56.82%	50%	103	**0%**	87.46%	**100%**	170	**0%**	85.01%	**100%**	**191**	**0%**	86.65%	**100%**	190
	Privacy [13]	87.46%	48%	301	**0%**	56.82%	50%	339	1%	87.46%	**100%**	329	**0%**	85.74%	**100%**	347	**0%**	86.65%	**100%**	**362**
	AMI [41]	87.46%	3%	124	3%	56.82%	50%	151	**0%**	87.46%	**100%**	166	**0%**	86.25%	**100%**	**179**	**0%**	86.65%	**100%**	173
VGG-Face	DMI [9]	82.41%	21%	162	12%	70.61%	**100%**	154	2%	82.41%	**100%**	**226**	**0%**	81.87%	**100%**	**221**	**0%**	81.24%	**100%**	214
	Mirror [5]	82.41%	49%	163	44%	70.61%	**100%**	173	**0%**	82.41%	**100%**	219	**0%**	79.93%	**100%**	**269**	**0%**	81.24%	**100%**	247
	Privacy [13]	82.41%	**0%**	347	**0%**	70.61%	**100%**	320	**0%**	82.41%	**100%**	399	**0%**	81.22%	**100%**	**443**	**0%**	81.24%	**100%**	390
	AMI [41]	82.41%	17%	212	**0%**	70.61%	**100%**	194	**0%**	82.41%	**100%**	317	**0%**	81.52%	**100%**	**362**	**0%**	81.24%	**100%**	344
VGG-Face2	DMI [9]	92.03%	27%	142	19%	87.83%	**100%**	144	**0%**	92.03%	**100%**	211	**0%**	90.44%	**100%**	**273**	**0%**	91.02%	**100%**	244
	Mirror [5]	92.03%	34%	78	30%	87.83%	**100%**	101	**0%**	92.03%	**100%**	178	**0%**	89.24%	**100%**	213	1%	91.02%	**100%**	**223**
	Privacy [13]	92.03%	**0%**	410	**0%**	87.83%	**100%**	338	**0%**	92.03%	**100%**	396	**0%**	91.07%	**100%**	**406**	**0%**	91.02%	**100%**	379
	AMI [41]	92.03%	14%	186	14%	87.83%	**100%**	213	**0%**	92.03%	**100%**	244	**0%**	88.21%	**100%**	249	**0%**	91.02%	**100%**	**255**

B. Baseline Defenses

We compare our proposed method with two state-of-the-art model inversion defenses (i.e., DP [44] and Ad-mi [39]) to show its superiority. Note that we did not compare with other defenses due to their source codes are not open now. We conduct the model inversion attacks and baseline defenses according to their source codes published on GitHub.

1) Ad-mi: We follow the original settings outlined in Ad-mi to reproduce it. Specifically, we performed 10 epochs of adversarial noise generation on each output vector obtained from the victim model. In line with the constraints specified in Ad-mi, we ensure that the probability of each label remains below 1.0, and the sum of probabilities across all labels amounts to 1.0.

2) DP: We utilized the widely adopted differential privacy method, DP-SGD [44], to perform DP training on the victim model. We set a noise size of $\varepsilon = 8$, as smaller or medium-sized noises are deemed ineffective for defending against model inversion attacks.

C. Evaluation Metrics

Apart from showing the inversed images before and after the defenses, we also employ the following quantitative evaluation metrics in this paper. We use *model accuracy* (MA) to evaluate the utility of the victim model, *attack accuracy* (AA) to evaluate the attack accuracy, and *Frechet inception distance* (FID) to evaluate the quality of the inversed images.

1) Model Accuracy: Model accuracy is defined as the accuracy of the victim model on the original test dataset. It is used to evaluate whether the defense process impacts the classification functionality of the model.

$$P(F_V, \chi) = \frac{1}{|\chi|} \sum_{x \in \chi} \mathbf{I}_{[F_V(x)=y]}, \tag{11}$$

where χ denotes the test dataset, y is the ground-truth label of the sample x, and $\mathbf{I}(.)$ is the indicator function.

In our experiments, we employed two kinds of model accuracy: overall model accuracy (MC(all)) and target label model accuracy (MC(l_t)). MC(all) represents the accuracy of the entire test dataset, while (MC(l_t)) represents the accuracy of samples specifically belonging to the target label. We use these two metrics to explore the impact of our proposed method on both the entire dataset and the target label.

2) Attack Accuracy: We evaluate the model inversion attack accuracy as the transferable accuracy of the inversed samples by a general classification model F_E instead of the victim model, since the inversed samples tend to be overfitted by the victim model. The evaluation classifier F_E has a different structure from F_V but is trained on the same training dataset. If the inversed images achieve high accuracy on the evaluation classifier, they are considered to expose private information about the target label, i.e., the model inversion attack is successful; otherwise, the defense succeeds.

$$P(F_E, \chi_I) = \frac{1}{|\chi_I|} \sum_{x_I \in \chi_I} \mathbf{I}_{[F_E(x)=l_t]}, \tag{12}$$

where χ_I denotes the inversed samples generated by the attacker, and l_t is the target label of the victim model.

3) FID: Frechet Inception Distance (FID) [19] is used to measure the similarity of two image datasets. The more similar the inversed images are to the real images, the FID values will be lower. For a well-reconstructed inversed sample, its FID value is relatively lower than that of the substandard inversed image. We apply FID for evaluation classifier F_E, and the results verified that the inversed samples generated by our method are much farther away from the real ones in feature space than those without our defense.

$$FID = ||\frac{1}{|\chi_I|} \sum_{x_I \in \chi_I} feature(F_E(x_I)) - \frac{1}{|\chi|} \sum_{x \in \chi} feature(F_E(x))||^2. \tag{13}$$

Given input x, $feature(F_E(x))$ denotes the input of the last full connection layer. The function $feature(\cdot)$ can be viewed as a feature vector of the samples extracted by the model.

VI. EVALUATION RESULTS

We first present the comparison results of our proposed method and baseline defenses, then conduct an ablation study to evaluate the necessity of three loss function items in the model fine-tuning process.

A. Comparison With Baselines

We compare our proposed method with DP-based defense and Ad-mi, and the comparison results are shown in Table II.

TABLE III
COMPARISON WITH THE DEFENSE THAT DIRECTLY FINE-TUNING VICTIM MODEL WITH RANDOMLY-SELECTED FAKE SAMPLES

	Random fake samples				GAN-D				GAN-ID			
	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID
DMI [9]	8%	81.44%	0%	158	**0%**	**86.54%**	**100%**	159	**0%**	86.49%	**100%**	**170**
Mirror [5]	10%	81.44%	0%	144	**0%**	84.46%	**100%**	**174**	**0%**	**86.49%**	**100%**	168
Privacy [13]	**0%**	81.44%	0%	322	**0%**	85.33%	**100%**	351	**0%**	**86.49%**	**100%**	**355**
AMI [42]	5%	81.44%	0%	170	**0%**	**85.62%**	**100%**	**171**	**0%**	86.49%	**100%**	170

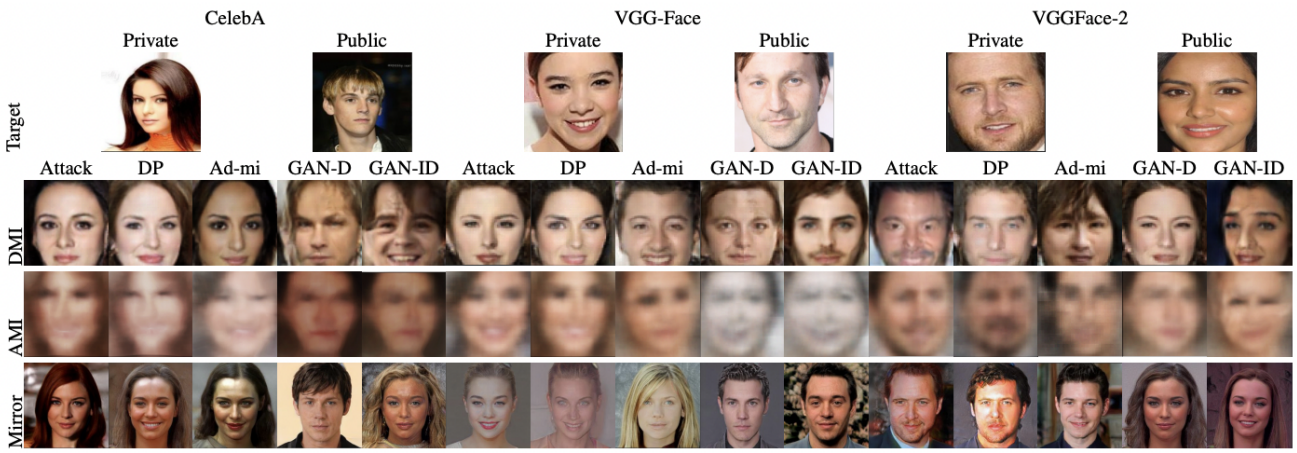

Fig. 2. Comparison of our proposed method with baselines in terms of the inversed sample quality. In GAN-D, we assume that the defender knows the GAN structure adopted by the attacker. In GAN-ID, we assume that the defender knows nothing about the GAN adopted by the attacker.

Compared with DMI, Privacy, and AMI, Mirror has the highest attack accuracy and the lowest FID in most cases, which means Mirror has the best attack performance among the four attacks. The success of Mirror is due to its use of the StyleGAN structure.

Our proposed defense method is referred to as GAN-D or GAN-ID, depending on the defense scenario. In the case of GAN-D, our method requires information on the GAN structure used by the attacker. In contrast, GAN-ID assumes that the defender knows nothing about the GAN adopted by the attacker, i.e., the black-box scenario. Note that both GAN-D and GAN-ID use focal loss by default.

In the first case, we can see that GAN-D outperforms the baseline defenses in terms of attack accuracy and FID for all datasets. Take Mirror as an example, GAN-D can decrease its attack accuracy even to 0% for all datasets, while DP-based defenses can only decrease its AA to 45% (CelebA), 44% (VGG-Face), and 30% (VGG-Face2). GAN-D can significantly improve the FID of Mirror from 98 to 174 (CelebA), 163 to 273 (VGG-Face), and 78 to 199 (VGG-Face2), followed by Ad-mi, which improves the FID to 170 (CelebA), 219 (VGG-Face), and 178 (VGG-Face2), while DP-based defense can only improve its FID to 103, 173, and 101, respectively. The reason why Ad-mi also achieves a relatively better defense performance is that we run it under a white-box defense scenario. The attacker can entirely control and modify the output of the target model when performing the model inversion attack. Since Ad-mi did not manipulate the training process of the victim model, its model accuracy is the same as the original victim model. In contrast, DP-based defenses severely undermine the utility of victim models by significantly reducing model accuracy. In contrast, we can see that GAN-D can maintain a high model accuracy of the protected model, i.e., with less than a 3% reduction in prediction accuracy for clean samples for almost all cases.

We also evaluate the performance of our proposed method in the black-box setting (denoted as GAN-ID), where the defender has no knowledge of the attacker's GAN model structure or parameters. In this scenario, we adopt Wasserstein GAN [2] to defend against DMI (based on Inversion-specific GAN) and Mirror (uses StyleGAN). It is shown that GAN-ID can also achieve the best defensive performance in most cases. GAN-ID can effectively decrease AA below 1% against all attacks for all datasets. This means that the attackers cannot invert correct face images even if we don't have any knowledge of the attackers' GAN. In terms of MC and FID, GAN-ID performed similarly to GAN-D. We can retain both high model accuracy (no more than a 2% drop in MC(all)) and target label accuracy (100% in all settings). The values of our FID are also much higher than those of the original models. For example, GAN-ID can achieve an FID value of 170 when defending against DMI in CelebA, while the FID value of the original model is only 108.

Apart from DP and Ad-mi, we also establish a baseline by directly retraining the victim model through random fake samples. We randomly choose images from the training set of the label l_p and l_t in public and private datasets and use a similar loss as Section IV-C to fine-tune the model. The results for CelebA are shown in Table III. We can see that the model trained on randomly selected samples cannot sustain a high level of accuracy compared to ours. Additionally, it fails to reduce the attack accuracy of these attacks to 0%.

We also compare the quality of the inversed samples before and after the defense with the baselines, and the results are shown in Fig. 2. Note that we only present the results of DMI, AMI, and Mirror since the reconstructed faces of Privacy are

TABLE IV

ABLATION STUDY OF LOSS ITEMS. WE ASSUME THAT THE DEFENDER KNOWS THE GAN STRUCTURE ADOPTED BY THE ATTACKER

		L_{pri}				$L_{pri}+L_{pub}$				GAN-D(L_{CE})				GAN-D(L_{FL})			
		AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID
CelebA	DMI [9]	**0%**	21.71%	0%	275	2%	84.23%	**100%**	153	**0%**	**86.54%**	**100%**	159	**0%**	86.33%	**100%**	188
	Mirror [5]	2%	10.73%	0%	226	**0%**	80.34%	50%	166	**0%**	84.46%	**100%**	174	**0%**	**85.01%**	**100%**	191
	Privacy [13]	**0%**	18.86%	0%	326	**0%**	46.37%	0%	387	**0%**	85.33%	**100%**	351	**0%**	**85.74%**	**100%**	347
	AMI [41]	3%	4.28%	0%	203	**0%**	77.92%	50%	180	**0%**	85.62%	**100%**	171	**0%**	**86.25%**	**100%**	179
VGG-Face	DMI [9]	**0%**	5.64%	0%	288	**0%**	74.63%	0%	258	**0%**	81.14%	**100%**	216	**0%**	**81.87%**	**100%**	221
	Mirror [5]	2%	8.61%	0%	252	1%	70.71%	0%	212	**0%**	**80.01%**	**100%**	273	**0%**	79.93%	**100%**	269
	Privacy [13]	**0%**	20.35%	0%	379	**0%**	68.50%	0%	423	**0%**	80.75%	**100%**	401	**0%**	**81.22%**	**100%**	443
	AMI [41]	**0%**	14.41%	50%	332	**0%**	78.97%	50%	289	**0%**	**81.54%**	**100%**	317	**0%**	81.52%	**100%**	362
VGG-Face2	DMI [9]	**0%**	3.00%	0%	307	**0%**	80.76%	**100%**	322	**0%**	**90.77%**	**100%**	261	**0%**	90.44%	**100%**	273
	Mirror [5]	7%	33.15%	**100%**	229	**0%**	33.15%	**100%**	251	**0%**	88.35%	**100%**	199	**0%**	**89.24%**	**100%**	213
	Privacy [13]	**0%**	41.00%	0%	412	**0%**	69.61%	50%	362	**0%**	**91.39%**	**100%**	401	**0%**	91.07%	**100%**	406
	AMI [41]	**0%**	5.84%	0%	297	**0%**	85.57%	**100%**	331	**0%**	87.94%	**100%**	265	**0%**	**88.21%**	**100%**	249

TABLE V

ABLATION STUDY OF LOSS ITEMS. WE ASSUME THAT THE DEFENDER DOESN'T KNOW THE ATTACKER'S GAN STRUCTURE

		L_{pri}				$L_{pri}+L_{pub}$				GAN-ID(L_{CE})				GAN-ID(L_{FL})			
		AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID
CelebA	DMI [9]	**0%**	19.35%	0%	279	4%	**86.58%**	0%	153	**0%**	86.49%	**100%**	170	**0%**	86.65%	**100%**	168
	Mirror [5]	1%	19.35%	0%	263	1%	**86.58%**	0%	166	**0%**	86.49%	**100%**	168	**0%**	86.65%	**100%**	190
VGG-Face	DMI [9]	**0%**	13.77%	0%	209	**0%**	79.20%	0%	258	**0%**	**80.87%**	**100%**	238	**0%**	**81.24%**	**100%**	214
	Mirror [5]	**0%**	13.77%	0%	245	13%	79.20%	0%	212	**0%**	**80.87%**	**100%**	214	**0%**	**81.24%**	**100%**	247
VGG-Face2	DMI [9]	**0%**	24.32%	0%	341	1%	88.33%	50%	322	**0%**	**90.35%**	**100%**	244	**0%**	**91.02%**	**100%**	244
	Mirror [5]	**1%**	24.32%	0%	272	6%	88.33%	50%	251	**1%**	**90.35%**	**100%**	207	**1%**	**91.02%**	**100%**	223

too blurry to recognize. It is shown that the inversed faces generated by our protected model are much far away from the private images than those generated by the baseline-protected models. The inversed samples of the victim model generated by ours are more like the public data, which denotes the effectiveness of our proposed method. The success of our proposed method is attributed to the GAN-based fake sample generation, loss adjustment, and continual learning algorithms.

B. Ablation Study

In this part, we explore the impact of different loss function items in the model fine-tuning. The results are shown in Table IV and Table V. "L_{pri}" only includes the loss maximization on the inversed private images. "$L_{pri}+L_{pub}$" means that we use both private and public inversed images to fine-tune the model, but no continual learning restriction. Compared with "$L_{pri}+L_{pub}$", GAN-D(L_{CE}) and GAN-ID(L_{CE}) add the continual learning loss L_{CL} while using cross-entropy loss. In contrast, GAN-D(L_{FL}) and GAN-ID(L_{FL}) add the continual learning loss component L_{CL} while employing focal loss.

As shown in Table IV, we can see that GAN-D is able to downgrade the attack accuracy to 0% in all settings, while the other two loss settings cannot decrease the AA to zero. The attack accuracy of Mirror can still achieve 7% for "L_{pri}" in VGG-Face2 and 1% for "$L_{pri}+L_{pub}$" in VGG-Face. Besides, the MC(all) of "L_{pri}" sharply drops to below 50% or even below 10%, and its MC(l_t) drops to 0% in most settings. The FID of "L_{pri}" is the highest among all three loss function settings, which is related to its lowest MC(l_t). It disturbs the original feature space distribution of the victim model. Although "$L_{pri}+L_{pub}$" can achieve a better model accuracy than "L_{pri}", it still cannot meet the requirements of the model utility. For example, its MC(all) drops to only 46.37% and MC(l_t) drops to 0% for Privacy in CelebA dataset. In contrast, GAN-D broadly maintains MC(all) with less than 3%. Besides, it can correctly classify all target label samples, which shows the effectiveness of continual learning. The defense performance of GAN-D(L_{FL}) is better than that of GAN-D(L_{CE}), especially in its capacity to enhance FID values, thus validating the effectiveness of the focal loss.

The experimental results in the black-box scenario are shown in Table V. Among DMI, Mirror, Privacy, and AMI, only DMI and Mirror are GAN-based model inversion attacks. Therefore, we only present the defense results against DMI and Mirror. It is demonstrated that GAN-ID can effectively decrease the AA to close to 0% while preserving the MC(all) and MC(l_t). We can see that L_{pri} fails to maintain a high MC(all) and MC(l_t), resulting in a drop to 0% for MC(l_t)s and less than 25% for MC(all)s. As for "$L_{pri}+L_{pub}$", it also fails to reduce AA sufficiently, especially against Mirror. It can only deduce the AA to 13% in VGG-Face and 6% in VGG-Face2 against Mirror. The value of MC(l_t) is low since the absence of continual learning. We can see that GAN-ID(L_{FL}) also exhibits superior defense performance compared to GAN-ID(L_{CE}).

It is also surprising that the AA values of GAN-D and GAN-ID are lower than that of "$L_{pri}+L_{pub}$", which means better defense effectiveness. This is probably because continual learning can shift the overfitting from previous images to inversed public samples rather than simply injecting new overfitted features.

TABLE VI
IMPACT OF GAN STRUCTURE ON DEFENSE PERFORMANCE

		Original			InfoGAN				CGAN				WGAN			
		MC(all)	AA	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID	AA	MC(all)	MC(l_t)	FID
CelebA	DMI [9]	87.46%	52%	108	0%	86.88%	100%	190	0%	85.71%	100%	165	0%	86.65%	100%	168
	Mirror [5]	87.46%	69%	98	2%	86.88%	100%	157	1%	85.71%	100%	202	0%	86.65%	100%	190
	Privacy [13]	87.46%	48%	301	0%	86.88%	100%	390	0%	85.71%	100%	406	0%	86.65%	100%	362
	AMI [41]	87.46%	3%	124	0%	86.88%	100%	207	0%	86.25%	100%	200	0%	86.65%	100%	173
VGG-Face	DMI [9]	82.41%	21%	162	0%	81.92%	100%	194	0%	81.25%	100%	209	0%	81.24%	100%	214
	Mirror [5]	82.41%	49%	163	0%	81.92%	100%	212	2%	81.25%	100%	244	0%	81.24%	100%	247
	Privacy [13]	82.41%	0%	347	0%	81.92%	100%	350	0%	81.25%	100%	353	0%	81.24%	100%	390
	AMI [41]	82.41%	17%	212	0%	81.92%	100%	252	0%	81.25%	100%	409	0%	81.24%	100%	344
VGG-Face2	DMI [9]	92.03%	27%	142	0%	90.59%	100%	216	0%	92.21%	100%	277	0%	91.02%	100%	244
	Mirror [5]	92.03%	34%	78	1%	90.59%	100%	283	3%	92.21%	100%	206	1%	91.02%	100%	223
	Privacy [13]	92.03%	0%	410	0%	90.59%	100%	417	0%	92.21%	100%	430	0%	91.02%	100%	419
	AMI [41]	92.03%	14%	186	0%	90.59%	100%	274	0%	92.21%	100%	280	0%	91.02%	100%	255

C. Impact of GAN Structure on Defense Performance

If the defender knows the structure of the GAN model used by the attacker, the defender can employ the same structure to train the GAN model. However, in the black-box scenario, the defender knows nothing about the attacker's GAN. In this part, we explore the impact of the GAN structure on defense performance.

We choose CGAN [29], InfoGAN [10], and WGAN [2] as the GAN structure, respectively. The defense results are shown in Table VI. Across all datasets, we can effectively reduce the attack accuracy (AA) to below 3% for all attacks. This indicates that the attackers are unable to successfully inverse genuine face images. Moreover, we achieve both high model accuracy (with no more than a 2% decrease in MC(all)) and target label accuracy (100% in all settings). The values of our FID are also much higher than those of the original models. The results verify that we can effectively defend against SOTA model inversion attacks, regardless of the GAN model structure employed to generate the fake samples.

VII. CONCLUSION

In this paper, we present a novel and effective privacy-utility defense framework against model inversion attacks. Our defense framework departs from traditional methods that perturb the prediction vector and instead manipulate the training of the victim model by incorporating carefully crafted fake samples. To enhance the defense performance, we incorporate a generative adversarial network (GAN) to generate fake samples and adjust the loss of inverted samples to deceive the attack model. To further improve the utility of the victim model, we introduce the concept of continuous learning during the fine-tuning process. Experiment results demonstrate the superiority of our proposed method over state-of-the-art defense strategies. It is shown that our proposed method can also maintain a high defense performance in black-box scenarios.

REFERENCES

[1] M. Abadi et al., "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.

[2] J. Adler and S. Lunz, "Banach Wasserstein GAN," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–10.

[3] U. Aïvodji, S. Gambs, and T. Ther, "GAMIN: An adversarial approach to black-box model inversion," 2019, *arXiv:1909.11835*.

[4] T. A. O. Alves, F. M. G. França, and S. Kundu, "MLPrivacyGuard: Defeating confidence information based model inversion attacks on machine learning systems," in *Proc. Great Lakes Symp. VLSI*, May 2019, pp. 411–415.

[5] S. An et al., "MIRROR: Model inversion for deep learning network with high fidelity," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2022, pp. 1–28.

[6] M. Bojarski et al., "End to end learning for self-driving cars," 2016, *arXiv:1604.07316*.

[7] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[8] F. M. Castro, M. J. Marín-Jiménez, N. Guil, C. Schmid, and K. Alahari, "End-to-end incremental learning," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 233–248.

[9] S. Chen, M. Kahla, R. Jia, and G.-J. Qi, "Knowledge-enriched distributional model inversion attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16158–16167.

[10] 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. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–9.

[11] M. De Lange et al., "A continual learning survey: Defying forgetting in classification tasks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3366–3385, Jul. 2022.

[12] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2015, pp. 1322–1333.

[13] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, and T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing," in *Proc. USENIX Secur. Symp.*, 2014, pp. 17–32.

[14] N. Z. Gong and B. Liu, "You are who you know and how you behave: Attribute inference attacks via users' social friends and behaviors," in *Proc. USENIX Secur. Symp.*, 2016, pp. 979–995.

[15] X. Gong, Y. Chen, Q. Wang, M. Wang, and S. Li, "Private data inference attacks against cloud: Model, technologies, and research directions," *IEEE Commun. Mag.*, vol. 60, no. 9, pp. 46–52, Sep. 2022.

[16] X. Gong, Z. Wang, Y. Chen, Q. Wang, C. Wang, and C. Shen, "NetGuard: Protecting commercial web APIs from model inversion attacks using GAN-generated fake samples," in *Proc. ACM Web Conf.*, Apr. 2023, pp. 2045–2053.

[17] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[18] Z. He, T. Zhang, and R. B. Lee, "Model inversion attacks against collaborative inference," in *Proc. 35th Annu. Comput. Secur. Appl. Conf.*, Dec. 2019, pp. 148–162.

[19] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–12.

[20] B. Hitaj, G. Ateniese, and F. Perez-Cruz, "Deep models under the GAN: Information leakage from collaborative deep learning," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 603–618.

[21] B. Houston and K. Kirchhoff, "Continual learning for multi-dialect acoustic models," in *Proc. Interspeech*, Oct. 2020, pp. 576–580.

[22] Q. Huang, J. Zhang, W. Zhou, W. Zhang, and N. Yu, "Initiative defense against facial manipulation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 2, 2021, pp. 1619–1627.
[23] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.
[24] K. James et al., "Overcoming catastrophic forgetting in neural networks," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 13, pp. 3521–3526, Mar. 2017.
[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2999–3007.
[26] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[27] L. Melis, C. Song, E. De Cristofaro, and V. Shmatikov, "Exploiting unintended feature leakage in collaborative learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2019, pp. 691–706.
[28] G. Merlin, V. Lomonaco, A. Cossu, A. Carta, and D. Bacciu, "Practical recommendations for replay-based continual learning methods," 2022, *arXiv:2203.10317*.
[29] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.
[30] M. Nasr, R. Shokri, and A. Houmansadr, "Machine learning with membership privacy using adversarial regularization," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2018, pp. 634–646.
[31] O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf.* Durham, U.K.: BMVA Press, 2015, pp. 41.1–41.12.
[32] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 779–788.
[33] A. Salem, R. Wen, M. Backes, S. Ma, and Y. Zhang, "Dynamic backdoor attacks against machine learning models," in *Proc. IEEE 7th Eur. Symp. Secur. Privacy (EuroS&P)*, Jun. 2022, pp. 703–718.
[34] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-leaks: Model and data independent membership inference attacks and defenses on machine learning models," in *Proc. Netw. Distrib. Syst. Secur. Symp.*, 2019, pp. 1–15.
[35] G. M. van de Ven and A. S. Tolias, "Three scenarios for continual learning," 2019, *arXiv:1904.07734*.
[36] R. Wang, Z. Huang, Z. Chen, L. Liu, J. Chen, and L. Wang, "Anti-forgery: Towards a stealthy and robust DeepFake disruption attack via adversarial perceptual-aware perturbations," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Jul. 2022, pp. 1–9.
[37] T. Wang, Y. Zhang, and R. Jia, "Improving robustness to model inversion attacks via mutual information regularization," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2021, pp. 11666–11673.
[38] W. Wang, X. Qian, Y. Fu, and X. Xue, "DST: Dynamic substitute training for data-free black-box attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 14341–14350.
[39] J. Wen, S.-M. Yiu, and L. C. K. Hui, "Defending against model inversion attack by adversarial examples," in *Proc. IEEE Int. Conf. Cyber Secur. Resilience (CSR)*, Jul. 2021, pp. 551–556.
[40] Z. Yang, B. Shao, B. Xuan, E.-C. Chang, and F. Zhang, "Defending model inversion and membership inference attacks via prediction purification," 2020, *arXiv:2005.03915*.
[41] Z. Yang, J. Zhang, E.-C. Chang, and Z. Liang, "Neural network inversion in adversarial setting via background knowledge alignment," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Nov. 2019, pp. 225–240.
[42] 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. (CSF)*, Jul. 2018, pp. 268–282.
[43] H. Yin et al., "Dreaming to distill: Data-free knowledge transfer via DeepInversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8712–8721.
[44] A. Yousefpour et al., "Opacus: User-friendly differential privacy library in PyTorch," 2021, *arXiv:2109.12298*.
[45] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, and D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 250–258.
[46] X. Zhao, W. Zhang, X. Xiao, and B. Lim, "Exploiting explanations for model inversion attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 662–672.
[47] M. Zhou, J. Wu, Y. Liu, S. Liu, and C. Zhu, "DaST: Data-free substitute training for adversarial attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 231–240.

Xueluan Gong (Student Member, IEEE) received the B.S. degree in computer science and electronic engineering from Hunan University in 2018. She is currently pursuing the Ph.D. degree in computer science with Wuhan University, China. She has published 20 publications in top-tier international journals or conferences, including IEEE SECURITY & PRIVACY, NDSS, Usenix Security, WWW, ACM Ubicomp, IJCAI, IEEE JOURNAL ON SELECTED AREAS IN COMMUNICATIONS, and IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING. Her research interests include network security, AI security, and data mining.

Ziyao Wang is currently pursuing the bachelor's degree with the School of Cyber Science and Engineering, Wuhan University, China. His research interests include information security and AI security.

Shuaike Li is currently pursuing the bachelor's degree with the School of Cyber Science and Engineering, Wuhan University, China. His research interests include information security and AI security.

Yanjiao Chen (Senior Member, IEEE) received the B.E. degree in electronic engineering from Tsinghua University in 2010 and the Ph.D. degree in computer science and engineering from The Hong Kong University of Science and Technology in 2015. She is currently a Bairen Researcher with Zhejiang University, China. Her research interests include spectrum management for femtocell networks, network economics, network security, AI security, and the quality of experience (QoE) of multimedia delivery/distribution.

Qian Wang (Fellow, IEEE) is currently a Professor with the School of Cyber Science and Engineering, Wuhan University, China. He has published more than 200 papers, with more than 120 publications in top-tier international conferences, including USENIX NSDI, ACM CCS, USENIX Security, NDSS, ACM MobiCom, and ICML, with more than 20000 Google Scholar citations. He was selected into the National High-Level Young Talents Program of China and listed among the World's Top 2% Scientists by Stanford University. He is a member of the ACM. He also received the National Science Fund for Excellent Young Scholars of China in 2018. He has long been engaged in the research of cyberspace security, with a focus on AI security, data outsourcing security and privacy, wireless systems security, and applied cryptography. He was a recipient of the 2018 IEEE TCSC Award for Excellence in Scalable Computing (early Career Researcher) and the 2016 IEEE ComSoc Asia–Pacific Outstanding Young Researcher Award. He was a co-recipient of eight best paper and best student paper awards from prestigious conferences, including ICDCS and IEEE ICNP. His Ph.D. student was selected under Huawei's "Top Minds" Recruitment Program in 2021. He serves as an Associate Editor for IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING (IEEE TDSC) and IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY (IEEE TIFS).