

Defense Against Model Inversion Attacks Using a Dummy Recognition Model Trained with Synthetic Samples

Yuta Kotsuji and Kazuaki Nakamura* ^a

Tokyo University of Science, Niijuku 6-3-1, Katsushika, Tokyo, 125-8585 Japan

Keywords: Model Inversion Attacks (MIA), Defense Against MIA, Dummy Recognition Model, Synthetic Images.


Abstract: Recently, biometric recognition models such as face identification models have been rapidly developing. At the same time, the risk of cyber-attacks on such models is increasing, one of whose examples is a model inversion attack (MIA). MIA is an attack to reconstruct or reveal the training samples of a victim recognition model by analyzing the relationship between its inputs and outputs. When MIA is conducted on a biometric model, its training samples such as the face, iris, and fingerprint images could be leaked. Since they are privacy-sensitive personal information, their leakage causes a serious privacy issue. Hence, it is desirable to develop a defense method against MIA. Although several defense methods have been proposed in the past decade, they tend to decrease the recognition accuracy of the victim model. To solve this problem, in this paper, we propose to use a dummy model trained with synthetic images and parallelly combine it with the victim model, where the combined model is released to users instead of the victim model. The key point of our proposed method is to force the dummy model to output a high confidence score only for the limited range of synthetic images. This allows us to maintain the recognition accuracy of the combined model. We experimentally confirmed that the proposed method can reduce the success rate of MIA to less than 30% while maintaining the recognition accuracy of more than 95%.

1 INTRODUCTION

With the development and spread of deep learning technologies, biometric recognition models such as face recognition have become common and are often used in user authentication and verification systems. On the other hand, the risk of cyber-attacks against such a recognition model is also increasing. Various kinds of attacks have been actively studied in the past decade (Liu et al., 2020; He et al., 2022), one of whose typical examples is a model inversion attack (MIA) (Fredrikson et al., 2015). MIA is an attack that attempts to reconstruct or reveal the training samples of a target (or victim) recognition model by analyzing the relationship between its inputs and outputs. More specifically, in MIA, an attacker first specifies a certain class label (e.g., individual name or ID) and then finds an input sample (e.g., a face image) whose confidence score provided by the victim model is maximized for the specified label. The sample obtained by the above attack process becomes quite similar to the actual training sample. When the attacked victim

model is a biometric recognition system, the results of MIA could contain a target individual's biometric information such as the face, iris, and fingerprint. Thus, MIA leads to the leakage of privacy-sensitive information that should be kept private, which could cause a serious issue. Therefore, it is urgent to realize a countermeasure against MIA.

A possible solution is to refuse access trials from users who have already used the victim model a predetermined number of times. This is effective because attackers generally send a large number of input samples to the victim model as queries in order to achieve MIA, whereas ordinary users send fewer queries than the attackers. However, this kind of defense strategy is vulnerable to collusion attacks by multiple attackers, where each attacker sends a relatively small number of queries to the victim model and its corresponding outputs are aggregated across all attackers. Hence, previous work proposes another defense strategy that leads MIA results to the samples totally different from the actual training data (Wang et al., 2021). This is achieved by minimizing the mutual information between the inputs and outputs of the victim model. Although this strategy can effectively reduce the success

^a  <https://orcid.org/0000-0002-4859-4624>

*Corresponding author

rate of MIA, it also reduces the recognition accuracy of the victim model, which is a critical drawback.

To overcome the drawback, in this paper, we propose a novel defense method against MIA satisfying the following two conditions: (1) MIA results are led to the samples different from the actual training data. (2) The recognition accuracy of the victim model can be maintained. To this end, we train a dummy model using synthetic samples and parallelly combine it with the victim model. This means the combined model looks like a two-branch network where the branches are connected into a single head. Note that the dummy model is designed so that it provides a low confidence score for most input samples while providing very high confidence only for its training samples, i.e., synthetic samples. This allows us to lead the MIA results to the synthetic samples without reducing the victim model's recognition accuracy.

The contributions of this paper are summarized below.

- We propose a defense method against MIA that can maintain the recognition accuracy of the victim model without directly manipulating its structure and parameters.
- We give a technique to train a dummy model whose confidence score can be high only for a limited range of synthetic samples.

2 RELATED WORK

2.1 Model Inversion Attacks

MIA has been actively studied in the past decade. There are various existing attack methods of MIA, and their assumptions on the victim model and the attackers' knowledge are different. For the type of the victim model, some methods consider that the victim model is a white-box model whose structure and parameters are known by the attackers, while others consider the victim model as a black-box model. (We note that the attackers do not know the role of each layer or branch in the victim model even in the case of the white-box setting; what they can do on the victim model is just back-propagation in addition to the query sending and output receiving.) Besides, the output information of the victim model is differently considered. Some methods assume that the victim model provides a confidence score (i.e., logit) for all classes it covers, whereas others assume that only label information is provided as a recognition result. On the other hand, for the attackers' knowledge, some methods allow the attackers to exploit a certain auxiliary

dataset while others do not.

The earliest MIA method was proposed by Fredrikson et al. (Fredrikson et al., 2015), which targets a white-box victim model outputting confidence score information and works without any auxiliary dataset. Specifically, they assume that the victim model R_T outputs a confidence score vector $\mathbf{y} = R_T(x)$ for an input sample x , where the k -th dimension of \mathbf{y} indicates the score for the k -th class. For such an R_T , their proposed method specifies a target class label $\hat{\mathbf{y}}$ in the form of a one-hot vector and finds the input sample x that minimizes the error between \mathbf{y} and $\hat{\mathbf{y}}$. The error is measured by a certain loss function L as $L(\mathbf{y}; \hat{\mathbf{y}}) = L(R_T(x); \hat{\mathbf{y}})$, whose minimization is achieved by a gradient descent algorithm.

The above method works well when R_T is a shallow neural network but tends to find a noisy sample like an adversarial example (Goodfellow et al., 2014b) when R_T is a deep network. To solve this drawback, Y. Zhang et al. proposed to introduce an adversarial loss term employed in the GAN framework (Zhang et al., 2020). They use an auxiliary dataset of real samples to train a GAN discriminator, and its counterpart generator is parallelly trained so that it can generate a sample x successfully fooling the discriminator as well as minimizing the above error term $L(R_T(x); \hat{\mathbf{y}})$. As a more straightforward method, Khosravy et al. proposed to narrow down the search space to find the optimal x by introducing an image generator trained with an auxiliary dataset, particularly focusing on a face identification system as the victim model (Khosravy et al., 2022; Khosravy et al., 2021).

Another MIA method assuming a white-box victim model was Amplified-MIA (Zhang et al., 2023) proposed by Z. Zhang. Instead of minimizing the above error term and obtaining its solution $\hat{x} = \operatorname{argmin}_x \{L(R_T(x); \hat{\mathbf{y}})\}$, Amplified-MIA attempts to train an inverse model that directly predicts \hat{x} from a given $\hat{\mathbf{y}}$. In this method, $\hat{\mathbf{y}}$ is not a one-hot vector but a confidence score vector provided by R_T , and its values are amplified by a nonlinear amplification layer.

Unlike the above methods, Yoshimura et al. assumed that the victim model is a black-box model. In the case of a black-box victim model, the attackers cannot perform a gradient descent process to minimize the error term $L(R_T(x); \hat{\mathbf{y}})$. To solve this problem, Yoshimura proposed a method for numerically approximating the gradients (Yoshimura et al., 2021). Then they performed Khosravy's method using the approximated gradients. Note that this method assumes a black-box victim model outputting a full confidence score \mathbf{y} . In contrast, Zhu et al. targeted a

black-box victim model that only outputs label information. In this case, the attackers cannot explicitly obtain $R_T(x)$. To solve this, Zhu et al. proposed to estimate it only from the label information (Zhu et al., 2022). More specifically, they exploited the recognition error rate of the victim model on the neighbor region of x to measure $R_T(x)$. Liu et al. also tackled the task of label-only MIA (Liu et al., 2024). They used a Conditional Diffusion Model (Ho and Salimans, 2021) trained with an auxiliary dataset to achieve high-performance MIA under the label-only setting.

2.2 Defense Method Against MIA

Methods for defending against MIA are less actively studied than those for attacking. This is because MIA defense is a hard task. Since MIA is a task of reconstructing an input sample x that satisfies $R_T(x) = \hat{y}$ from a given output \hat{y} , it is difficult to achieve MIA if the victim model R_T does not well capture the statistical relationship between input x and output y ; in other words, such an R_T is robust to MIA. However, this solution sacrifices the high recognition accuracy of R_T . Thus, there is a trade-off between a model's recognition accuracy and its robustness against MIA.

Fredrikson et al., who proposed the earliest attack method, suggested keeping a model's accuracy relatively low as a countermeasure for MIA (Fredrikson et al., 2015). Unfortunately, this is not practical due to the above trade-off. Wang et al. proposed another defense method that minimizes the mutual information between inputs and outputs when training a victim model (Wang et al., 2021). This is also not good at maintaining recognition accuracy. Salem et al. proposed a defense method that adds a uniform noise to the model's output (i.e., confidence score vector) y before returning it to the system users (Salem et al., 2020). This does not degrade the model's accuracy if the added noise is small enough. However, a small noise leads to insufficient robustness against MIA.

Unlike the above existing methods, this paper aims to propose a defense method against MIA that can maintain the recognition accuracy of the victim model as much as possible.

3 ASSUMED ATTACK METHOD

Before explaining the proposed defense method in detail, we first describe the MIA method assumed in this paper. Note that we pick out a face recognition model as an example of the victim model in the remainder of this paper.

As mentioned in Section 2.1, there are many existing MIA methods, where attackers' knowledge about the victim model is differently assumed from several aspects: white-box or black-box, label-only or not, presence or absence of an auxiliary dataset, and so on. Ideally, it should be experimentally examined whether the proposed defense method can defeat all the attack methods, which is however difficult in practice. Hence, in this paper, we introduce an assumption that is most advantageous for the attackers; the victim model is a white-box model and outputs a confidence score of all the class labels for any input image. In addition, an auxiliary dataset is available. We aim to realize a defense method that can successfully work under such a hard condition. Specifically, we employ Khosravy's MIA method (Khosravy et al., 2022) since it satisfies this condition.

In Khosravy's method, a victim model R_T outputs a confidence score vector $y = (y_1, \dots, y_n)^T = R_T(x)$ for any input face image x , where y_k is the confidence score of the k -th class label (i.e., the k -th individual ID) and n is the total number of class labels (i.e., the number of individuals registered in the victim face recognition model R_T). To conduct MIA for the R_T , an attacker specifies a certain target class ID in the form of a one-hot vector \hat{y} . If the target ID is the \hat{k} -th class, the \hat{k} -th dimension of \hat{y} is set as 1 and all the other dimensions are set as 0. Next, the attacker measures the error between y and \hat{y} , which is denoted by $L(y; \hat{y}) = L(R_T(x); \hat{y})$, using the cross-entropy loss function L . Theoretically, MIA can be achieved by finding the image x that minimizes the above error term by a gradient descent algorithm, exploiting the knowledge of the structure and the parameters of R_T . However, this tends to lead to a noisy result, as mentioned in Section 2.1. Hence, Khosravy et al. proposed to narrow down the search space for the gradient descent. To this end, their method trains an image generator F using an auxiliary face dataset, where F is mathematically a map from a feature vector ξ to an image $x = F(\xi)$, and employs a new error term $L(R_T(F(\xi)); \hat{y})$ instead of $L(R_T(x); \hat{y})$. Then the attacker finds the feature vector ξ that minimizes the new error term. This is equivalent to maximizing the \hat{k} -th dimension of $y = R_T(F(\xi))$, namely $y_{\hat{k}}$. Finally, using the optimal $\hat{\xi} = \operatorname{argmin}_{\xi} L(R_T(F(\xi)); \hat{y})$, the MIA result is obtained as $\hat{x} = F(\hat{\xi})$. Fig. 1 depicts the overview of their MIA process.

The objective of our proposed method is to prevent the above MIA procedure; that is, we want to lead the MIA result \hat{x} to an image dissimilar to the actual face of the \hat{k} -th individual without reducing the recognition accuracy of R_T .

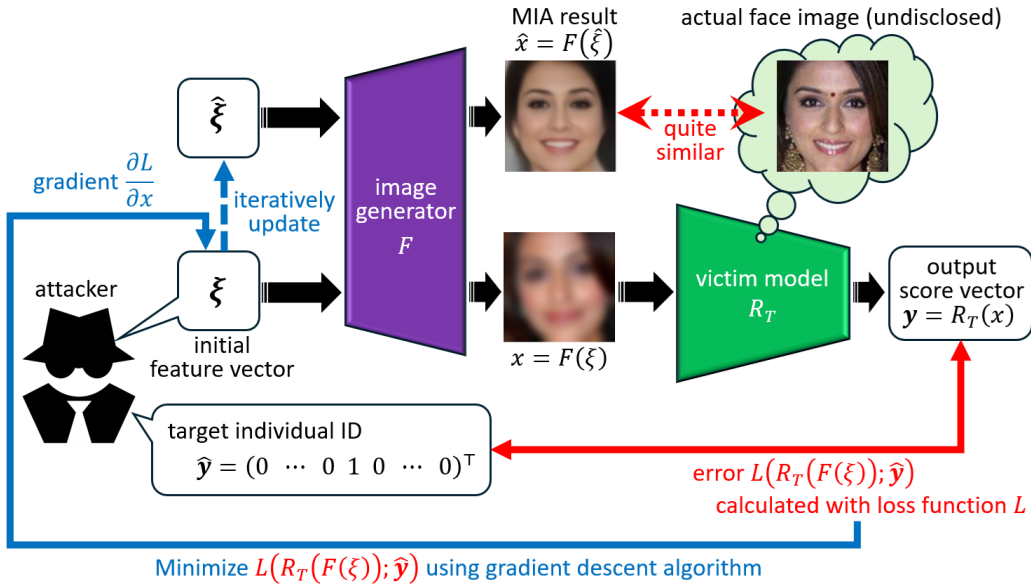


Figure 1: Overview of Khosravy's MIA method (Khosravy et al., 2022).

4 PROPOSED MIA DEFENSE

4.1 Overview

Let R_T be the victim model and \hat{k} be the target individual ID of MIA. As we described in Section 3, MIA is the process of finding the input image $x \in X$ that maximizes the confidence score $y_{\hat{k}}$, where $y_{\hat{k}} = R_T(x)_{\hat{k}}$ is the \hat{k} -th dimension of $\mathbf{y} = R_T(x)$ and X is the whole image space. Here, let us suppose the case that the victim model R_T outputs a very high confidence score for a certain face image $x' \in X$ that is totally different from the real face image of the \hat{k} -th individual. In this case, the result of MIA is led to x' , by which the \hat{k} -th individual's real face can be protected. Of course, this x' causes a misrecognition. However, if images like the x' are located only in a quite limited range in X , the impact of the misrecognition problem is minimized and negligible.

Unfortunately, it is not easy to directly give the above property to R_T . Hence, we introduce a dummy recognition model R_{dmy} . Importantly, we train this dummy model with a set of synthetic face images so that it outputs a high confidence score only for the synthetic images and provides a low score for any real face image. Then we parallelly connect the R_{dmy} with R_T to construct a combined model R' , as seen in Fig. 2. More specifically, we construct R' that outputs the confidence score vector \mathbf{y}' as

$$\mathbf{y}' = R'(x) = \alpha R_T(x) + (1 - \alpha) R_{\text{dmy}}(x) \quad (1)$$

for an input image x , where α ($0 < \alpha < 1$) is a coeffi-

cient for controlling the weights of R_T and R_{dmy} . As mentioned above, R_{dmy} gives a high confidence score for synthetic images and a low confidence score for real images. In contrast, R_T gives a low confidence score for most synthetic images since it is trained with a set of real face images. Hence, when we let $R_{\text{dmy}}(x)_k$ denote the k -th dimension of $R_{\text{dmy}}(x)$, its argument max, i.e., $\operatorname{argmax}_x R_{\text{dmy}}(x)_k$, becomes totally different from $\operatorname{argmax}_x R_T(x)_k$ for all k . In this situation, by setting α less than 0.5, we can satisfy the following relationship:

$$\begin{aligned} \operatorname{argmax}_x R'(x)_{\hat{k}} &= \operatorname{argmax}_x R_{\text{dmy}}(x)_{\hat{k}} \\ &\neq \operatorname{argmax}_x R(x)_{\hat{k}} = \operatorname{argmax}_x y_{\hat{k}}, \end{aligned} \quad (2)$$

where $R'(x)_{\hat{k}}$ is the \hat{k} -th dimension of $\mathbf{y}' = R'(x)$. This means the result of MIA against R' can be led to the synthetic images. At the same time, the above strategy also allows us to force $R_{\text{dmy}}(x)_k$ to be low for all k (namely $R_{\text{dmy}}(x)_k \approx 1/n$ for all k) when x is a real face image. This means that the second term in Formula (1) is almost constant with respect to k for real images and therefore

$$\operatorname{argmax}_k R'(x)_k = \operatorname{argmax}_k R_T(x)_k \quad (3)$$

is satisfied in the cases of inputting a real face image as x into R' . With this property, R' can correctly recognize the real face images and successfully maintain high recognition accuracy. We experimentally analyze the optimal value of α in Section 5 since theoretically determining it is not straightforward.

The proposed method releases this R' to the public instead of R_T . Note that R' looks like a two-branch

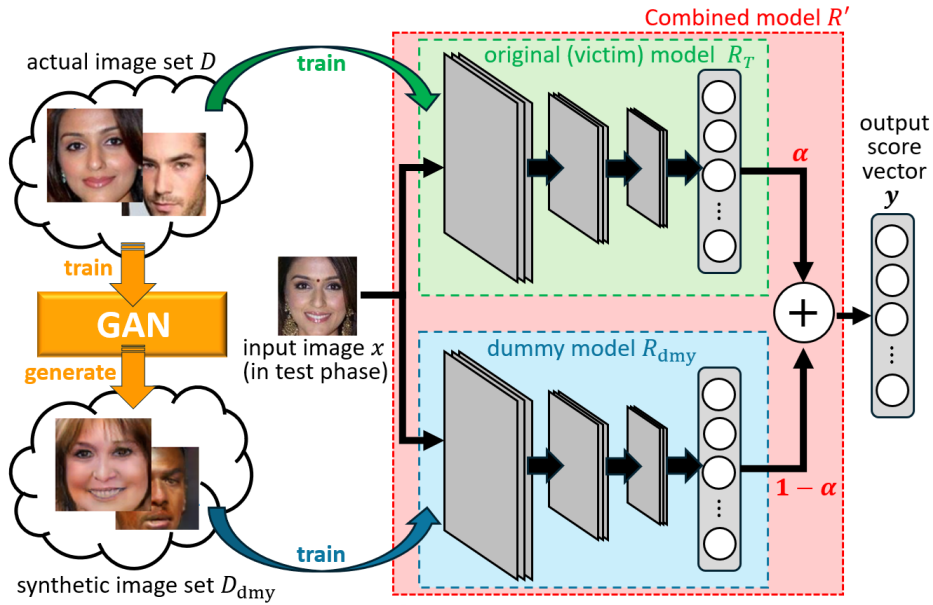


Figure 2: Overview of proposed MIA defense method.

network for the attackers even if they know its structure and parameters. In addition, R' can be released as a black-box model in practice even though we regard it as a white-box model in this paper. For these reasons, the attackers cannot judge whether their target model is protected by the proposed method or not.

4.2 Training Process of Dummy Model

This section describes the training process of the dummy model R_{dmy} , particularly how to force R_{dmy} to output a high confidence score only for the designated synthetic images.

The owner of R_T has an image set for training it. Let D be the image set. Of course, this D cannot be used to train R_{dmy} . To correct a training set of R_{dmy} , we first construct an image generator using the framework of Generative Adversarial Network (GAN) (Goodfellow et al., 2014a), where D is used as a training set for the GAN. After constructing the GAN, we use its generator G to generate n synthetic different face images $\{x^{(i)} = G(z^{(i)}) \mid i = 1, \dots, n\}$ by drawing random vectors $z^{(i)}$ from a normal distribution $\mathcal{N}(0, \sigma^2 I)$. Furthermore, we also draw m additional random vectors $\{z_l^{(i)} \mid l = 1, \dots, m\}$ from another normal distribution $\mathcal{N}(z^{(i)}, \tilde{\sigma}^2 I)$ for each $z^{(i)}$ to generate a synthetic face image $x_l^{(i)} = G(z_l^{(i)})$. After that, we use $D_{\text{dmy}} = \{x_l^{(i)} \mid i = 1, \dots, n, l = 1, \dots, m\}$ as a training set for R_{dmy} . In the above procedure, we use very small $\tilde{\sigma}^2$, namely $\tilde{\sigma}^2 \ll \sigma^2$, which makes $x_l^{(i)}$ quite similar to $x^{(i)}$ for all l . This

property can force R_{dmy} to output a high confidence score only for the neighbors of $x^{(i)}$. Note that we use $\{x_l^{(i)}\}$ instead of a single image $x^{(i)}$ because the larger number of training images can stabilize the training process of R_{dmy} .

The loss function L for training R_{dmy} is as follows:

$$L = \sum_{x_l^{(i)} \in D_{\text{dmy}}} \text{CE}(R_{\text{dmy}}(x_l^{(i)}), v^{(i)}), \quad (4)$$

where CE is the cross-entropy loss function and $v^{(i)}$ is the n -dimensional one-hot vector whose i -th dimension is 1.

4.3 Qualitative Insights of Proposed Method

This section provides qualitative insights about the reason why the proposed method works well. Fig. 3 shows the relationship between the confidence score of R' and the MIA result. Since only a narrow range of synthetic images are used to train R_{dmy} , its confidence score forms a sharp peak in the image space and becomes almost $1/n$ for most images, particularly for real face images. In contrast, real face images of the k -th individual have a wide variety of lighting conditions, facial expressions, face orientations, and so on. Hence, the confidence score of R_T becomes high in a relatively wide area. Owing to these properties, the confidence score of R' becomes almost the same as that of R_T except for the neighbor of the designated synthetic images. This realizes the relationship

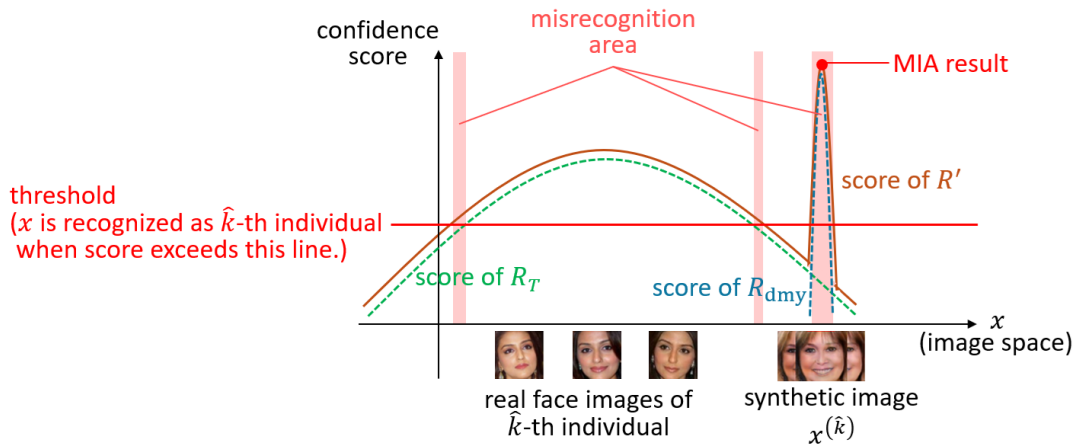


Figure 3: Relationship between confidence score, misrecognition area, and MIA result in image space.

of Formula (3). As a result, misrecognition could occur only in a limited range in the image space X and the recognition accuracy of R' is maintained. Besides, the MIA result is led to the global maximum of the confidence score of R' , which is located in not the real image side but the synthetic image side, as shown in Formula (2). This allows us to protect the real face images of the \hat{k} -th individual.

5 EXPERIMENTS

5.1 Experimental Setup

We conducted an experiment to evaluate the effectiveness of the proposed method. In this experiment, we constructed a face identification system as the victim model R_T , using VGGFace2 dataset (Cao et al., 2018).

VGGFace2 is a famous face image dataset containing more than 3 million images of 9131 individuals. Among these images, we selected 85640 ($=2141 \times 40$) images of 2141 individuals and used them to train R_T . These images were also used to train a GAN generator, which is needed to construct a dummy model R_{dmy} and a combined model R' in the proposed method, as mentioned in Section 4.2. The network structures of the GAN generator and its counterpart discriminator are shown in Fig. 4. To increase the performance of the GAN, we employed two techniques named Adaptive Discriminator Augmentation (Karras et al., 2020) and Minibatch Discrimination (Salimans et al., 2016) in this experiment. The visual quality of the images generated by our GAN was 114.46 in terms of Frechet Inception Distance, abbreviated as FID (Heusel et al., 2017). Furthermore, we trained another face recognition model

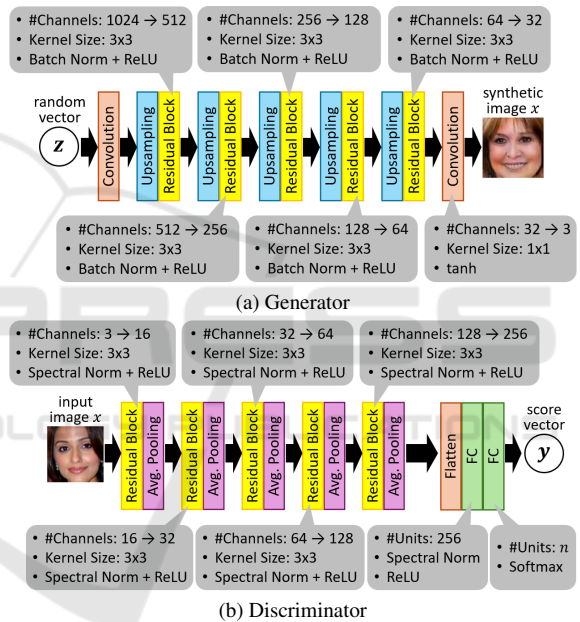


Figure 4: The network structures of GAN generator and discriminator.

R_E using 86680 ($=2167 \times 40$) images of 2167 individuals in VGGFace2 for evaluating the MIA success rate. Note that the images of only 27 out of these 2167 individuals were also included in the training set of R_T and used as the target individual IDs of MIA. The remaining 2140 individuals were totally different from the training set of R_T . The network structures of R_T and R_E are the same as those used in Khosravy’s study (Khosravy et al., 2022). R_{dmy} has the same structure as R_T . We further used Large Margin Cosine Loss (Wang et al., 2018) as a loss function for training R_T , R_E , and R_{dmy} to increase their recognition performance. The two hyper-parameters σ^2 and $\tilde{\sigma}^2$, which are needed to prepare D_{dmy} , were set as

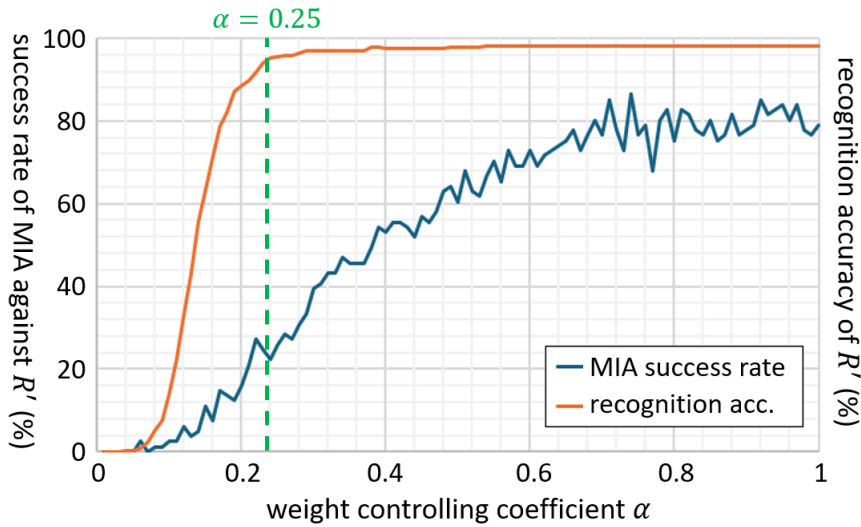


Figure 5: Success rate of MIA against combined model R' and its recognition accuracy with various α .

$\sigma^2 = 1$ and $\tilde{\sigma}^2 = 0.0625$, respectively.

Using the above R_E , we evaluated the success rate of MIA as follows. First, we carried out Khosravy’s MIA method to the combined model R' by specifying one of the 27 individuals as a target ID. We repeated this process 3 times and got 3 resultant images for each individual. As mentioned in Section 3, Khosravy’s method requires an auxiliary dataset to train the image generator F . We employed CelebA dataset for this purpose (Liu et al., 2015). Next, we input each resultant image into R_E and got the output confidence score vector. If the top-5 highest dimension of the confidence score vector includes the specified target ID, we judged the MIA process succeeded. Finally, we measured the percentage of the number of succeeding processes to the total number of MIA trials, which was regarded as the success rate. On the other hand, for evaluating the recognition accuracy of the combined model R' , we selected other 270 ($=10 \times 27$) images of the 27 individuals from VGGFace2 as a test set. These 27 individuals are identical to those used as the target IDs in MIA trials.

5.2 Results and Discussions

We evaluated the success rate of MIA against R' and its recognition accuracy, varying the value of weight controlling coefficient α in Formula (1) from 0 to 1 with a step size of 0.01. The result is shown in Fig. 5.

In Fig. 5, R' maintains a face recognition accuracy of more than 95% even when we reduce the coefficient α to 0.25. We verified that this is almost the same as the recognition accuracy of the original R_T , 95.8%. This indicates that the dummy model R_{dmy} returns a low confidence score for real face images, as

we expected in Section 4. At the same time, it also can be seen from Fig. 5 that the success rate of MIA against R' goes up to 80% when $\alpha = 1$, which means $R' = R_T$, while it significantly drops when $\alpha < 0.7$. This indicates that the MIA result can be successfully led to synthetic images owing to R_{dmy} ’s property of outputting a high confidence score only for a limited range of synthetic images. In the case of $\alpha = 0.25$ (the lowest α that can maintain high recognition accuracy), the success rate of MIA is at most 27%. This means that the attackers would fail to reconstruct the face of the target individual in 3 out of 4 MIA trials; they cannot have confidence in the resultant images of MIA. These experimental results demonstrate the effectiveness of the proposed method.

Fig. 6 shows an example of MIA results with various α for 10 out of the 27 target individuals. For comparison, this figure includes the real face images of the 10 individuals used to train R_T and the synthetic images used to train R_{dmy} . It can be seen from Fig. 6 that the MIA results with large α are similar to the real face of the target individual while those with smaller α tend to be similar to the synthetic images. However, for the individuals of ID:2, ID:7, and ID:8, their MIA results do not drastically differ from the real face images even in the case of $\alpha = 0.2$. We guess that the reason for this phenomenon is as follows. A GAN generator sometimes generates somewhat distorted face images due to its unstable training process, as the synthetic image of ID:2. Since the distorted images do not look so real, a powerful MIA method tends to avoid reconstructing them. This is why the results of MIA against R' are not necessarily similar to the synthetic images. As previously mentioned, the performance of our trained GAN was



Figure 6: Example of MIA results with various α for 10 target individuals.

114.46 in terms of FID, which is not so high and might tend to cause the above problem. A possible solution to cope with this problem is using a more sophisticated image generation technique such as a denoising diffusion probabilistic model (DDPM) (Ho et al., 2020) instead of GAN. This will be addressed in our future work. In addition, it is also an important future task to examine how the visual quality of the synthetic images affects the defense performance of the proposed method. As above, low-quality synthetic images are not desirable. However, synthetic images with too high quality might also be undesirable since they are indistinguishable from real ones, and therefore the property of R_{dmy} might become too close to that of R_T . Thus, it is important to find the optimal quality of the synthetic images.

Based on the current experimental results, we conclude that $\alpha = 0.25$ is the best choice. However, this is not necessarily the case with any other recognition models. The optimal value of α might depend on the victim model’s network structure, its training strategy, the number and the kinds of classes it covers, and so on. We will examine the impact of these factors on the defense performance of the proposed method in our future work. At the same time, we think that we have

to focus on a more practical face recognition model as the victim model R_T . As mentioned above, the recognition accuracy of R_T was 95.8% in this experiment, which is not sufficient for practical use. Thus, we will focus on a victim model with an accuracy of more than 99% and examine whether the MIA risk for such a practical face recognition model can also be reduced by the proposed method or not.

Besides, we believe that it is desirable to further decrease the success rate of MIA to more securely prevent the leakage of privacy-sensitive information like the face. To achieve this, we will investigate a more sophisticated method of combining R_T and R_{dmy} rather than just parallelly connecting them.

6 CONCLUSION

In this paper, we proposed a method to defend against MIA without degrading the recognition accuracy of the victim model. In the proposed method, we introduce a dummy model trained with GAN-generated synthetic images and parallelly combine it with the victim model. Then the combined model is released

to the public instead of the victim model. The key point of the proposed method is to force the dummy model to output a high confidence score only for the limited range of synthetic images and a low confidence score for real images. Owing to this property, the proposed method can maintain the recognition accuracy. We experimentally confirmed that the proposed method reduces the success rate of MIA to less than 30% while maintaining the recognition accuracy of more than 95%.

In our future work, we will examine the relationship between the characteristics of the victim model and the hyper-parameter α , which controls the combination weights for the victim model and the dummy model, to realize a method for easily finding the best choice of the α . At that time, we will focus on more practical (or accurate) victim models. It is also an important future work to extend the proposed method using DDPM instead of GAN to further reduce the risk of MIA.

Furthermore, we need to experimentally examine the robustness of the proposed method against various MIA attack methods other than Khosravy's one. Particularly, when an attacker knows the presence of the proposed defense method, he might exploit a set of face images morphed between real and synthetic faces to conduct MIA, using a sophisticated morphing method such as (Schardong et al., 2024). The robustness against such an attack is interesting and should be examined in the future.

This study is partially supported by JST CREST Grant (JPMJCR20D3).

REFERENCES

- Cao, Q., Shen, L., Xie, W., Parkhi, O. M., and Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *Proc. 13th IEEE International Conference on Automatic Face and Gesture Recognition (FG2018)*, pages 67–74.
- Fredrikson, M., Jha, S., and Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In *Proc. The 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 1322–1333.
- Goodfellow, I., Abadie, J., Mirza, M., Xu, B., Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative adversarial networks. In *Proc. 28th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2672–2680.
- Goodfellow, I., Shlens, J., and Szegedy, C. (2014b). Explaining and harnessing adversarial examples. In *Proc. 2014 International Conference on Learning Representations (ICLR)*, pages 1–11.
- He, Y., Meng, G., Chen, K., Hu, X., and He, J. (2022). Towards security threats of deep learning systems: A survey. *IEEE Trans. on Software Engineering*, 48:1743–1770.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proc. 31st International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6629–6640.
- Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. In *Proc. 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 6840–6851.
- Ho, J. and Salimans, T. (2021). Classifier-free diffusion guidance. In *Proc. 2021 NeurIPS Workshop on Deep Generative Models and Downstream Applications*, pages 1–8.
- Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., and Aila, T. (2020). Gtraining generative adversarial networks with limited data. In *Proc. 34th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 12104–12114.
- Khosravy, M., Nakamura, K., Hirose, Y., Nitta, N., and Babaguchi, N. (2021). Model inversion attack: Analysis under gray-box scenario on deep learning based face recognition system. *KSII Trans. on Internet and Information Systems*, 15(3):1100–1118.
- Khosravy, M., Nakamura, K., Hirose, Y., Nitta, N., and Babaguchi, N. (2022). Model inversion attack by integration of deep generative models: Privacy-sensitive face generation from a face recognition system. *IEEE Trans. on Information Forensics and Security*, pages 357–372.
- Liu, R., Wang, D., Ren, Y., Wang, Z., Guo, K., Qin, Q., and Liu, X. (2024). Unstoppable attack: Label-only model inversion via conditional diffusion model. *IEEE Trans. on Information Forensics and Security*, 19:3958–3973.
- Liu, X., Xie, L., Wang, Y., Zou, J., Xiong, J., Ying, Z., and Vasilakos, A. V. (2020). Privacy and security issues in deep learning: A survey. *IEEE Access*, 9:4566–4593.
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proc. 2015 International Conference on Computer Vision (ICCV)*, pages 3730–3738.
- Salem, A., Bhattacharya, A., Backes, M., Fritz, M., and Zhang, Y. (2020). Updatesleak: Data set inference and reconstruction attacks in online learning. In *Proc. 29th USENIX Security Symposium*, pages 1290–1308.
- Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., and Chen, X. (2016). Improved techniques for training gans. In *Proc. 30th International Conference on Neural Information Processing Systems (NeurIPS)*, pages 2234–2242.
- Schardong, G., Novello, T., Paz, H., Medvedev, I., da Silva, V., Velho, L., and Gonves, N. (2024). Neural implicit morphing of faces. In *Proc. 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 8395–8399.

- Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., and Liu, W. (2018). Cosface: Large margin cosine loss for deep face recognition. In *Proc. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5265–5274.
- Wang, T., Zhang, Y., and Jia, R. (2021). Improving robustness to model inversion attacks via mutual information regularization. In *Proc. The 35th AAAI Conference on Artificial Intelligence*, pages 11666–11673.
- Yoshimura, S., Nakamura, K., Nitta, N., and Babaguchi, N. (2021). Model inversion attack against a face recognition system in a black-box setting. In *Proc. 13th Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1800–1807.
- Zhang, Y., Jia, R., Pei, H., Wang, W., B, L., and Song, D. (2020). The secret revealer: Generative model-inversion attacks against deep neural networks. In *Proc. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 253–261.
- Zhang, Z., Wang, X., Huang, J., and Zhang, S. (2023). Analysis and utilization of hidden information in model inversion attacks. *IEEE Trans. on Information Forensics and Security*, 18:4449–4462.
- Zhu, T., Ye, D., Zhou, S., Liu, B., and Zhou, W. (2022). Label-only model inversion attacks: Attack with the least information. *IEEE Trans. on Information Forensics and Security*, 18:991–1005.

