Deep Face Recognizer Privacy Attack: Model Inversion Initialization by a Deep Generative Adversarial Data Space Discriminator

Mahdi Khosravy, Kazuaki Nakamura, Naoko Nitta, Noboru Babaguchi
Media Integrated Communication Laboratory, Graduate School of Engineering, Osaka University
2-1 Yamadaoka, Suita, Osaka 565-0871, Japan
E-mail: mahdi.khosravy@nanase.comm.eng.osaka-u.ac.jp
Tel: +81-6-6879-7746 FAX:+81-6-6879-7684

***Abstract*—A variety of Machine Learning (ML) applications involve data of privacy-sensitive content. Face recognizer is one of them which due to training by user identities face images, it is inherent to user face image as a critical biometric data. A face recognition system can be subject to privacy attacks even though it deploys a complex model structure like a deep-learning-based one. Because as the ML models advance to carry more complexity in structure and parameters, the privacy attack trends develop too. Model Inversion Attack (MIA) pioneered by Fredrikson et al [1] was applied on a shallow neural network of face recognizer, and its capability of privacy leakage was approved. Their work was on a white-box scenario wherein besides the model structure, the privacy-non-sensitive data of the users are partially available and used by the attacker for generation and leakage of the user identity face images. The present work improves the extension of MIA to deep learning models of face recognizers while performing without any data of the users. Despite the complexity of the deep models as an obstacle, this work improves the capability of MIA in this matter. To aim this goal, it initializes its training procedure by a seed image approved by a GAN-trained discriminator of face image data-space via its output probability value. In targeting two users' identities by MIA, the proposed technique approves its efficiency on a deep face recognition system. The recognition rates of the images generated by MIA associated with GAN data-space discriminator (GAN-DD) are higher than sole MIA and demonstrate efficiency improvement of deep MIA.**

Keywords: Cyber-security, cyber-attack, deep learning, face recognition, privacy protection, invasive software, Model Inversion.

I. Introduction

Deep learning (DL) [2] has highly boosted the capability of the recognition systems for accurate identification and classification of an extremely diverse as well a huge amount of data. Convolutional neural networks (CNN) is one of the DL techniques with a great impact on machine learning [3]. Due to using recognition systems at the cloud level, they might be subject to security and privacy attacks. A possible attack is by Model Inversion Attack (MIA) [1]. MIA aims to generate the data corresponding to a class label that is used for training the system. The generated data can be privacy-sensitive and used in a malicious way like revealing the private identity of a user on a social network.

In the case of a deep face recognition system, despite the model complexity and huge size of training data as well as the diversity of the calss labels, Khosravy et al [4] demonstrates MIA as an increasing considerable threat, but not as serious as the cases in Ref. [1] where the model is shallow and not as complex as a deep face recognizer. This research work, promotes the efficiency of MIA on a deep face recognition system even further by initialization of the MIA training with an optimum seed image. To aim this target, the seed image is generated by using a data-space discriminator trained by a deep generative adversarial network (GAN) [5]. It is under the same semi-white box scenario as Ref. [4] wherein just the model structure and class labels are in the access of the attacker. It should be noted that the MIA target is not fooling a machine learning model like the work in Ref. [6], rather it is more similar similar to initial works of Refs [7], [8] wherein there was some efforts for inverting the a visual representation from its encoded data.

The organization of the paper is as follows. After the introduction in Section I, Section II presents a review to MIA related works. Section III explains the MIA on a deep face recognizer in two parts of the proposed Semi-white box MIA on a deep face recognizer, and the promotion by initialization by a GAN data space discriminator. Section IV illustrates the experimental set up of the MIA on a deep face recognizer, and consequently, Section V discusses the results. Finally, Section VI presents the concluding remarks.

II. MODEL INVERSION ATTACK RELATED WORKS

In the cloud era, while the machine learning algorithms are used online, they are vulnerable more to malicious privacy attacks. An influential attack against a cloud-based machine learning algorithms can be (i) a causative attack, or (ii) an exploratory attack. In the first case, the attack can interfere with the process of training via the influence on the data used for training, and in the latter one, the attack is not via the influence on the training data or process. As it comes from its name, it is via exploration of the detector system characteristics to catch information about the data used for training the system. In this way, the attacker violates the users'

privacy. A very good example of exploratory one is model extraction attack [9] wherein the attacker aims to construct a duplicated version of the target system. Such a duplicate system is unauthorized. The focus of this research work is on model inversion attack (MIA) which is an exploratory type of attack, and the target of the attack is a deep learning based face recognition system.

MIA dates back to the first work by Fredrikson et al. [11] in 2014 wherein the MIA was formulated as the problem of estimation of the privacy-sensitive part of the data using the non-sensitive part and the class labels. For instance, in the case of face recognizer, the privacy-sensitive parts of the data can be considered as the face key parts e.g. eye, mouth, and nose regions while the other parts of the face are considered as the non-sensitive data parts, and the class labels are the individual IDs. initially, the target was a model of linear regression [11], and later, a decision tree and an artificial neural network [1] were their MIA model targets.After the above-mentioned couple of MIA research works, in 2016, Ref. [12] theoretically formalized MIA, and later on, Hidano et al. deployed MIA against collaborative filtering-based prediction systems [13], [14]. It should be noted that for the assumption of *a priori* knowledge of MIA about the system, there are two scenarios; (i) black-box scenario, and (ii) white-box scenarios. In the first scenario, the attacker can have just the output corresponding to each taken input for the target system. In the second scenario, the attacker has access to some information about the structure and parameters of the system model. Most of the on-line recognition systems have a secret model and the black-box scenario sounds more practical. But the white-box scenario is also realistic as the Kerckhoff's Principle [15] indicates, "*The system security should not rely on an unrealistic level of secrecy*". A very recent work on this scenario of MIA [16] is by training a generative adversarial network (GAN). In their work, they get an eclipsed or blurred face image of some identity users as a data of non-sensitive parts, and generate the corresponding complete face image. The main drawback of their technique is the requirement of individual training of GAN for each user identity. To aim this goal, an enormous amount of data images are applied to the the model that can be suspicious and detectable by the servers. In addition, there is an essential need to have proper blurred or eclipsed face images of each user identity for these trials, that would not be always possible. Finally, Khosravy et al. [4] overcome this problem even for MIA on a deep face recognition system without using any of non-sensitive part of data by using a seed image instead of non-privacy sensitive information. They implement MIA under a semi-white box scenario as they name it gray-box scenario wherein MIA aims to generate the user identity face image by using the class labels, model structure and parameters but without any access to any user data. However, due to implementation on a deep learning based model, their generated face image clones are noisy and hardly recognizable. This research work, improves the technique in Ref. [4] by using a GAN-based data space discriminator.

III. SEMI-WHITE BOX MIA ON DEEP FACE RECOGNIZER

In this section, first, we explain MIA against a deep learning based face recognizer in a gray-box scenario [4], then we suggest its efficiency promotion by initialization of the MIA learning process by a seed image generated through a process managed by a GAN data space discriminator. The ultimate goal of MIA on a face recognition system is generating a face image of a registered user which be used for revealing his/r identity. Besides the class label, in white box scenario of MIA [1], the attacker has access to the some of use identity images but without the privacy-sensitive parts or some blurred images. The deep MIA algorithm suggested by Ref. [4] performs just by using the class labels confidence information and having knowledge about the model structure and parameters.

A. MIA on a Deep Face Recognizer

As mentioned earlier in Section II, this research work focuses on a gray box scenario. The target system is a face recognition system R trained by deep learning. Its structure and training are based on CNN. It takes x that is an image of determined dimensions as input, and recognizes that x belongs to which of the registered individuals to the system as one of n classes. The output of the system is a vector of length n as follows:

$$\mathbf{y} = R(x) = [y_1 \dots y_n]^T \in [0,1]^n. \tag{1}$$

The i-th element of y is the score of confidence corresponding to the i-th registered user to the system. As a matter of fact, the result of recognition by the system is the user identity corresponding to the maximum element of $\mathbf{y}$, formally

$$\text{if} \quad i^* = \max_{y_i \in \mathbf{y}} y_i, \quad \text{then} \quad R_{\text{output}}(x) = i^* \tag{2}$$

In the conventional white box scenario the following information is achieved and used by the attacker: (i) recognition model structure, (ii) recognition model parameters, and (iii) some information of the corresponding user subjected to the MIA. The first two ones are *a priori* information related to the model of R, while the last one are information related to the targeted user. In the current work, the third one is not required. MIA considers the index of i as index of identity user of attack target. Its effort is to generate the face image of i-th registered user ID using the R structure. The general procedure of MIA is given in the sequence. MIA acquires a one-hot vector $\hat{\mathbf{y}} \in \{0,1\}^n$ as an estimation of $\mathbf{y}$ the desired output of R for the case of recognition of i-th identity wherein the i-th element of $\hat{\mathbf{y}}$ is 1 and the rest are 0s, formally

$$\begin{aligned} R(x_i) = \mathbf{y}_i \approx \hat{\mathbf{y}}_i, \quad \text{i.e.} \qquad \hat{\mathbf{y}} \in \{0,1\}^n \\ \hat{\mathbf{y}}_i(j) = \begin{cases} 1, & j = i \\ 0, & j \neq i \end{cases} \end{aligned} \tag{3}$$

where x_i is an image of i-th user identity, and $\mathbf{y}_i$ is the corresponding output of the system. The MIA aims to make a clone image $\hat{x}$ of the i-the user identity through iteratively updating an initial seed image x_0 in an optimization search process

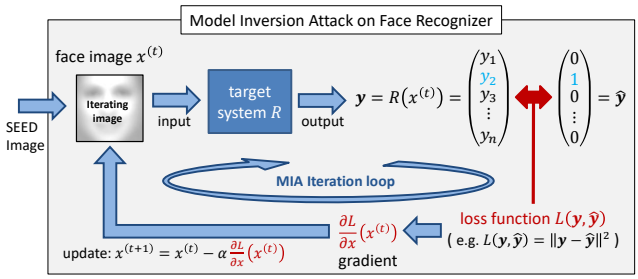

Fig. 1. Model inversion attack to a deep face recognizer under Semi-white scenario.

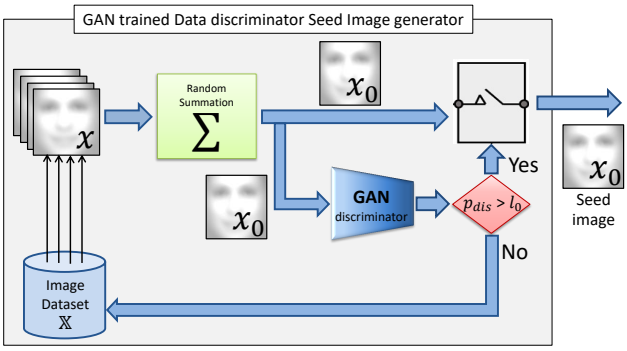

Fig. 2. Seed image generation for MIA using GAN-trained data space discriminator (GAN-DD).

within an image space of $\mathbb{X}$ with objective of minimizing a loss function that is measure of difference between $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$. The objective loss function $\mathcal{L}(\mathbf{y}_i, \hat{\mathbf{y}}_i) = \mathcal{L}(R(\hat{x}), \hat{\mathbf{y}})$ can be a function of softmax mutual entropy or mean squared error. Mathematically MIA process is as follows:

$$\hat{x}_{\text{opt},i} = \min_{\hat{x}} \mathcal{L}(R(\hat{x}), \hat{\mathbf{y}}), \quad \hat{x} \in \mathbb{X} \tag{4}$$

where $\mathbb{X}$ is the image data space. The above optimization problem is continuous and it can be solved by the algorithm of gradient descent iteration process as follows:

$$x^{(t+1)} = x^{(t)} - \alpha \frac{\partial \mathcal{L}}{\partial x}\left(x^{(t)}\right) \qquad (t = 0, 1, \dots), \tag{5}$$

where α is the learning rate. Having the model R structure and parameters, $\frac{\partial \mathcal{L}}{\partial x}(x)$ can be obtained as follows:

$$\frac{\partial \mathcal{L}}{\partial x}(x) = \frac{\partial \mathcal{L}}{\partial \mathbf{y}} \frac{\partial \mathbf{y}}{\partial x} = \frac{\partial \mathcal{L}}{\partial \mathbf{y}} \frac{\partial R(x)}{\partial x}. \tag{6}$$

The loss function $\mathcal{L}$ is set by the attacker. Thus its gradient is available for MIA. Since the model information is accessible, $\frac{\partial R(x)}{\partial x}$ is obtainable by attacker. The above iterative process leads to the solution as the face image clone of the i-the user ID. Formally,

$$\hat{x} = \lim_{t \to \infty} x^{(t)} \tag{7}$$

where $x^{(t)}$ is initialized at $t = 0$ by a '*seed image*' as x_0. Fig. 1 illustrates the MIA block diagram. The main part of the proposed MIA is the initialization part where a seed image is acquired. It is different from the conventional techniques wherein the non-sensitive part of privacy data of the users were used for the same. As the general case, the seed image can be an average of face images taken from a face image dataset. Indeed, the scenario used in the proposed MIA technique is without using any user information and just using the model information.

B. Initialization by GAN Datapoint Discriminator

In general the MIA results in clone images with high level of noise. It is due to locating of the $\hat{x}$ in a data space much wider than the face image space during the iteration process. The optimization process considers $\hat{x}$ in full image space rather than a narrower space limited to the face images. To mitigate this issue partially, we have focused on giving more proper seed image to MIA process in order to initiate the iteration process of the optimization from a better point in data space. The general seed image used in MIA on deep face recognition system under a gray-box scenario [4] is an average of the images taken from a face image dataset other than the one used for training of the system. In our approach of making an optimal seed image, we have used a face image space discriminator trained with a generative adversarial network (GAN). The proper seed image is made in a loop of making linear combinations of N number of images randomly taken from a face image dataset with acquired random weights w_i for each of them x_i. Then the resultant image is given to the GAN-based data-space discriminator (GAN-DD) for evaluation according to the discriminator output. The discriminator output normally is a probability weight which is used as a criterion of closeness or farness from being a face image in this work. In formal notation, the optimum seed image using a GAN-based face image data space discriminator is obtained as follows:

$$\begin{aligned} x_0^{\text{opt}} &= \sum_{x_i \in \mathbb{X}} w_i x_i \\ \{(i, w_i)\} &= \underset{\{(i,w_i)\},N}{\arg\min} \mathcal{D}_{\text{GAN}}\left(\sum_{x_i \in \mathbb{X}} w_i x_i\right) \\ &\quad \textstyle\sum_i^N w_i = 1, \end{aligned} \tag{8}$$

where i, w_i, N are respectively the indices, weights and number of the images randomly acquired and used in linear combination. $\mathcal{D}_{\text{GAN}}$ is the GAN based data space discriminator, and $\mathbb{X}$ is the face image dataset which from the the images randomly taken. Fig. 2 illustrates the process of seed image generation by GAN-trained data space discriminator.

As it is inherent to generative adversarial networks (GANs) training the discriminator is adversarially trained together with an image generator which is fed by random vectors taken from a normal distribution. The GAN-DD used in this work can be trained by any available set of face images, which is naturally different from the one used for training the face recognition system.

(a) Real face images of registered ID user 1

(b) Real face images of registered ID user 2

Fig. 3. The two corresponding registered identities which are targeted by MIA on a deep face recognition system; (a) ID-1 (b) ID-2.

IV. Experimental Setup

This section explains the experimental setup of the MIA on a deep face recognition system under gray-box scenario while the proposed GAN-DD-generated optimum seed image has been used. The setup is composed of three parts; (i) deep face recognition systems setup, and (ii) GAN-based data space discriminator setup. Each of them is briefly explained in the sequence. For training all the above-mentioned setups, we have used different parts of VGGFace2 face image database [**?**]. VGGFace2 possesses more than three millions of images. The images are from high diversity people of different ages, gestures, ethnicity, etc. Also, the images are in different lighting conditions. In this research work, the VGGFace2 images are divided into three parts. The first part is a quarter of them and used for training the target face recognizer. The second part is the second quarter of the images and used for training the evaluation face recognizer. Finally, the rest of the images which are half of the database are used for training the GAN-based data space discriminator.

A. Deep Face Recognition Systems Setup

As mentioned earlier, in this work, mainly two CNN-based deep face recognizers are used; one is used as the target system, and the other one is used as the evaluation system. Their models are notated respectively as R_T and R_E. The training of R_T and R_E, are by using the first and the second quarters of VGGFace2 face images which are without any common datapoints except the target identities images which are used in training of both systems. In this way, although R_T and R_E have same model structure, but due to training by different datasets, they have different parameters. The MIA is trained during targeting R_T and its resultant revealed face image is evaluated by R_E whose registered identities include the targeted user identities. Therefore, if the MIA works well, the revealed clone image should be recognized by the R_E in addition to R_T with a high recognition rate. In evaluation experiment setup in this work, two user identities which are registered to both systems are targeted. Fig 3. depicts some of the face images of these two target identities.

B. GAN-based data space discriminator setup

The strategy of adversarial training of a discriminator for distinguishing the real images from the generated images by a decoder is used for the GAN-DD. The encoder and the discriminator compete each other during the training in an adversarial manner. While the encoder effort is using random vectors with normal distribution to generate images not distinguishable from the real images, the discriminator effort is to discriminate any generated image from the real one. As the training process progresses further both encoder and discriminator efficiencies increase. In our setup, the remained half of VGGFace2 images is used for training the GAN-DD.

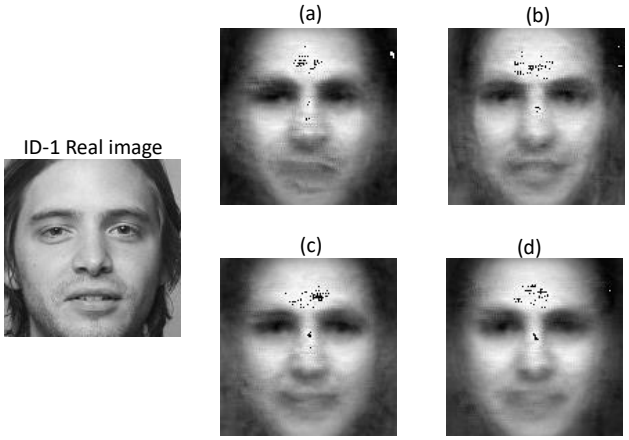

Fig. 4. A real face image of registered user ID-1 (left), and the corresponding clone images by MIA (a), MIA using GAN-DD seed image by 64 images (b), 128 images (c), and 256 images (d).

V. Results and Discussion

After training the GAN-DD as explained in Section IV, the MIA is applied under different four different settings on the target recognition system for cloning the ID-1 and ID-2 face images as their real ones are depicted in Fig. 3. The four setting manners are as follows; (i) MIA without GAN-DD using an average of 256 randomly acquired images from the image database, with GAN-DD seed image made by involving (ii) 64, (iii) 128, and (iv) 256 images ($N = 64, 128,$ and 256). Figs. 4 and 5 show respectively samples of clone images of ID-1 and ID-2 generated by MIA under the above-mentioned settings. As it is visually observable, the MIA under semi-white scenario can make clone images with considerable similarity to the targeted identities which may be helpful for revealing the identities and malicious applications. The four clone images generated by MIA without and GAN-DD, and with GAN-DDs of $N = 64, 128,$ and 256 for ID-1 and ID-2 (Figs. 4 and 5) follow the facial features of the corresponding target identities as in the case of ID-1 and ID the rounded form of the faces, the morphology of the nose, the approximate pattern of the eyes and eyebrows and shape of the forehead has a high similarity and for these two cases of study are clearly distinguishable from each other.

Considering the current work at the sequence of the very first work of MIA on a deep face recognition system [4], this general observation shows the potential of the MIA promotion for privacy attack on these systems. In the sequence, we have a more detailed comparative visual observation of the four above mentioned MIAs. As it is observable in the case of targeting

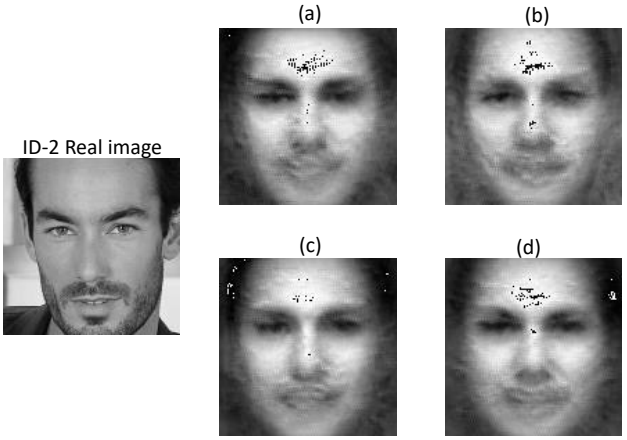

Fig. 5. A real face image of registered user ID-2 (left), and the corresponding clone images by MIA (a), MIA using GAN-DD seed image by 64 images (b), 128 images (c), and 256 images (d).

ID-1, MIAs using GAN-DD of $N = 64$ and 128 gives more clarity in the eyes, nose and lips details, and the case of ID-2 the more clarity on same features especially nose can be observed for MIAs using GAN-DD of $N = 64$ and 258.

Besides the above mentioned visual evaluation, we have numerically evaluated the generated clone images by using the scores and ranks given by the evaluating deep face recognition system R_E Tables 1 and 2 indicates respectively for ID-1 and ID-2 the recognition scores by the target and the evaluating face recognizers and the recognition rank by R_E for each generated clone image. While each image is generated by a given score of more than 0.99 by the target recognizer, the evaluating model does not give such a high score for them, and score and ranks by R_E give a distinguishable difference between them. As it was visually observable in the case of ID-1, the clone images generated by MIAs with GAN-DD of $N = 64$ and 128 have higher scores and ranks than the others as the one generated by MIA+GAN-DD $N = 128$ stands in the first place. Fig. 6 highlights the same result via the recognition ranks of each clone face image generated for ID-1. In the case of ID-2 the clone images generated by MIAs with GAN-DD of $N = 64$ and 258 have higher scores and ranks than others. In this case the one generated by MIA+GAN-DD $N = 64$ stands in the first place. This evaluative comparison on ID-2 generated clone images has been highlighted in Fig. 7.

To have a more robust numerical evaluation on clone images generated for ID-1 and ID-2 by each of the above-mentioned techniques, we have repeated the MIA for each case hundred

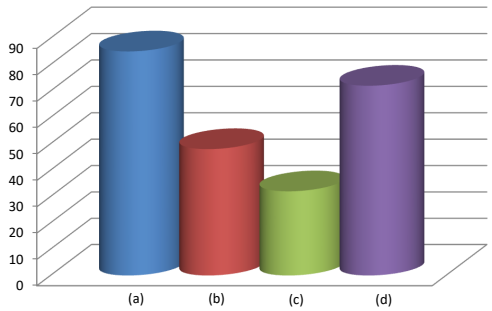

Fig. 6. Evaluation ranks given by R_E for clone images of ID-1 corresponding to Fig. 4 generated by (a) MIA, MIA with GAN-DD of (b) $N = 64$, (c) $N = 128$, and (d) $N = 256$.

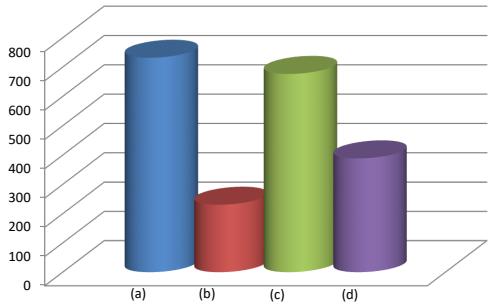

Fig. 7. Evaluation ranks given by for clone images of ID-2 corresponding to Fig. 5 generated by (a) MIA, MIA with GAN-DD of (b) $N = 64$, (c) $N = 128$, and (d) $N = 256$.

times, and obtaining hundred clone images for each. Then the hundred clone image of each MIA technique for each user identity is evaluated and given a recognition rank by R_E according to its training information. Using the given recognition ranks of each set of hundred clone images generated by each MIA technique, the cumulative matching characteristic (CMC) curve is obtained as a criterion of accuracy in possible recognition of the corresponding identity by clone images of each MIA technique. The CMC curve of each MIA technique has been depicted by Figs. 8 and 9 respectively for the targeted identities of ID-1 and ID-2. As it is observable in Fig 8., in most of the cases the recognition rate of sole MIA is lower than MIA with GAN-DD of different types in the generation of a clone image for ID-1. It can be seen just in the case of GAN-DD $N = 256$ the sole MIA has a bit better performance for the 50 higher ranked images, but not the rest. In the case of generated clone images for ID-2, as observable in Fig. 9, in all the cases the MIA technique with GAN-DD has higher performance than sole MIA.

TABLE I
EVALUATION SCORES BY TARGET AND EVALUATION FACE RECOGNIZERS, AND THE RECOGNITION RANK BY EVALUATION FACE RECOGNIZER FOR EACH GENERATED IMAGES FOR ID-1 IN FIG. 4.

Corresponding to Fig. 4	Score by R_T	Score by R_E	Rank by R_E
MIA	0.999997973	$1.56E-07$	85
MIA - GAN-DD-64	0.999516845	$3.48E-06$	48
MIA - GAN-DD-128	0.997307897	0.00030571	32
MIA - GAN-DD-256	0.999974728	$6.52E-06$	72

TABLE II
EVALUATION SCORES BY TARGET AND EVALUATION FACE RECOGNIZERS, AND THE RECOGNITION RANK BY EVALUATION FACE RECOGNIZER FOR EACH GENERATED IMAGES FOR ID-2 IN FIG. 5.

Corresponding to Fig. 4	Score by R_T	Score by R_E	Rank by R_E
MIA	0.995496154	$8.38E-12$	732
MIA - GAN-DD-64	0.995223224	$4.96E-10$	231
MIA - GAN-DD-128	0.999852657	$8.04E-10$	677
MIA - GAN-DD-256	0.994519413	$4.22E-09$	389

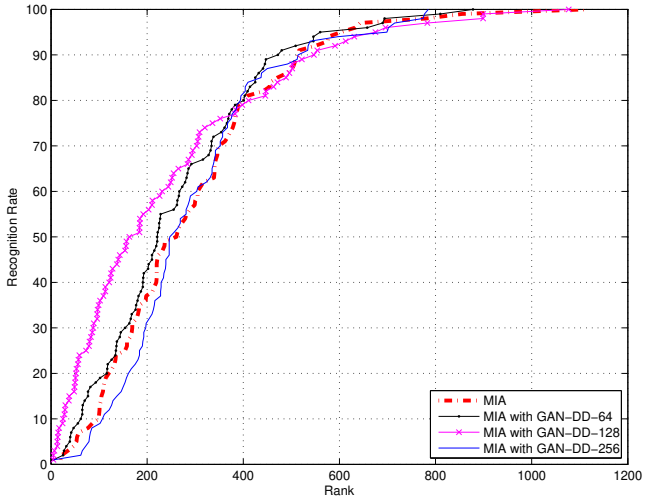

Fig. 8. CMC curves of the given ranks by evaluating recognition system to hundred clone images generated by different MIA techniques on CNN-based face recognizer targeting ID-1.

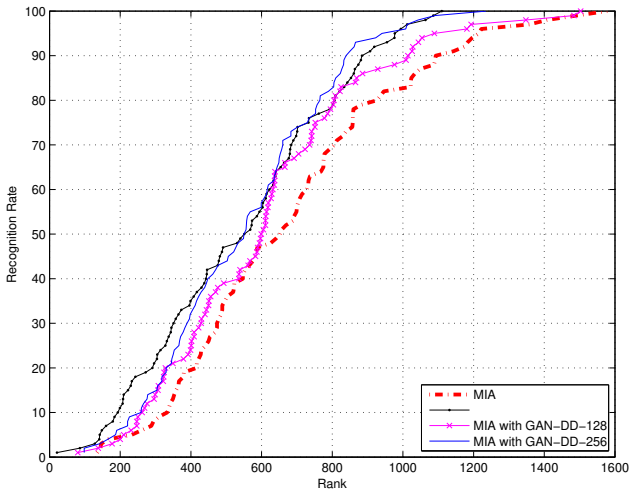

Fig. 9. CMC curves of the given ranks by evaluating recognition system to hundred clone images generated by different MIA techniques on CNN-based face recognizer targeting ID-2.

VI. CONCLUSION

Although a deep-learning-based recognition system may seem less vulnerable to privacy attacks due to its complexity of structure, the large size of parameters, and a huge amount of training data, it has been already alerted the researcher in cyber-security for the already coming threat of model inversion attack (MIA). In this research work, under a semi-white scenario wherein the attacker has just access to the model structure of a deep face recognition system and confidence information of class labels, the alert state of MIA on a deep face recognizer is further promoted. Besides the sole MIA technique, to improve the accuracy of the generated clone images of the users, a generative adversarial network (GAN) discriminator of the face image data space (GAN-DD) has been suggested and integrated to the MIA. Simulation results on an attack on a deep face recognition system targeting two registered identities demonstrate the considerable similarity of MIA-cloned images to the targeted real images. Also, MIA associated with GAN-DD shows partial improvement in the recognition rate by clone images compared to sole MIA. Although, the cloned images are still hardly recognizable by the human observer who may know the target identity, accurate numerical analysis of the generated hundred clone images by sole MIA and three settings of GAN-DD-MIA approves the partial promotion of MIA as a near-future coming privacy threat.

REFERENCES

[1] M. Fredrikson, S. Jha, T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in: *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pp. 1322–1333, 2015.

[2] L. Deng, D. Yu, "Deep learning: methods and applications", *Foundations and trends in signal processing* , vol. 7, pp. 197–387, 2014.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in: *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[4] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, N. Babaguchi, "Model Inversion Attack: Analysis under Gray-box Scenario on Deep Learning based Face Recognition System", in: *KSII Transactions on Internet and Information Systems (TIIS),* In press, 2020.

[5] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, "Generative adversarial nets", in: *Advances in neural information processing systems*, pp. 2672–2680, 2014.

[6] A. Nguyen, J. Yosinski, J. Clune, J. "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images", In *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 427–436, 2015.

[7] A. Mahendran, A. Vedaldi, "Understanding deep image representations by inverting them," In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5188–5196, 2015.

[8] A. Dosovitskiy, T. Brox, "Inverting visual representations with convolutional networks", In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4829–4837, 2016.

[9] F. Tramer, F. Zhang, A. Juels, M. K. Reiter, T. Ristenpart, "Stealing machine learning models via prediction apis", in: *25th fUSENIXg Security Symposium (fUSENIXg Security 16)*, pp. 601–618, 2016.

[10] M. Fredrikson, E. Lantz, S. Jha, S. Lin, D. Page, T. Ristenpart, "Privacy in pharmacogenetics: An end-to-end case study of personalized warfarin dosing", in: *23rd fUSENIXg Security Symposium (fUSENIXg Security 14)*, pp. 17–32, 2014.

[11] X. Wu, M. Fredrikson, S. Jha, J. F. Naughton, "A methodology for formalizing model-inversion attacks", in: *2016 IEEE 29th Computer Security Foundations Symposium (CSF)*, IEEE, pp. 355–370, 2016.

[12] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, G. Hanaoka, "Model inversion attacks for prediction systems: Without knowledge of nonsensitive attributes", in: *2017 15th Annual Conference on Privacy, Security and Trust (PST)*, IEEE, pp. 115–11509, 2017.

[13] S. Hidano, T. Murakami, S. Katsumata, S. Kiyomoto, G. Hanaoka, "Model inversion attacks for online prediction systems: Without knowledge of non-sensitive attributes", *IEICE Transactions on Information and Systems*, vol. 101, pp. 2665–2676, 2018.

[14] A. Kerckhoffs, "La cryptographic militaire", *Journal des sciences militaires*, pp. 5–38, 1883.

[15] Y. Zhang, R. Jia, H. Pei, W. Wang, B. Li, D. Song, "The secret revealer: Generative model-inversion attacks against deep neural networks", *arXiv preprint* arXiv:1911.07135, 2019.

[16] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, A. Zisserman, "Vggface2: A dataset for recognising faces across pose and age", in: *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition* (FG 2018), IEEE, pp. 67–74, 2018.