ELSEVIER

Pattern Recognition Letters
journal homepage: www.elsevier.com

Fingerprint Membership and Identity Inference Against Generative Adversarial Networks**

Saverio Cavasin[a], Daniele Mari[b,2], Simone Milani[b], Mauro Conti[a]

[a] *Department of Mathematics, University of Padova, Via Trieste, 63, Padova, 35131,*
[b] *Department of Information Engineering, University of Padova, Via Gradenigo 6A, Padova, 35131,*

Article history:

Membership Inference, Identity Inference, Fingerprints, Biometrics, GANs, Media Forensics, Explainable forensics

ABSTRACT

Generative models are gaining significant attention as potential catalysts for a novel industrial revolution. Since automated sample generation can be useful to solve privacy and data scarcity issues that usually affect learned biometric models, such technologies became widely spread in this field.

In this paper, we assess the vulnerabilities of generative machine learning models concerning identity protection by designing and testing an identity inference attack on fingerprint datasets created by means of a generative adversarial network. Experimental results show that the proposed solution proves to be effective under different configurations and easily extendable to other biometric measurements.

M.I.A. ≈ I.I.A. ↗

1. Introduction

Biometric signals are among the most robust and simple means to perform the identification or the authentication of a user for a wide variety of applications (e.g., home banking, payment apps).

To increase the accuracy of these systems, Deep Learning (DL) techniques have been explored and introduced also in biometrics [1, 2, 3, 4, 5], but the adoption of learned strategies implies the availability of large datasets (otherwise the final accuracy collapses), as well as the risk of possible information leakage concerning the training data (thus revealing sensitive data about some of the users) [6, 7, 8, 9].

To tackle these problems, multiple researchers have proposed the adoption of generative models, like Generative Adversarial Networks (GANs) [10], which automatically create new original bio-

**The work was partially supported by the European Union under the Italian National Recovery and Resilience Plan (NRRP) of NextGenerationEU, partnership on "Telecommunications of the Future" (PE00000001 - program "RESTART").

e-mail: `saverio.cavasin@studenti.unipd.it` (Saverio Cavasin), `daniele.mari@dei.unipd.it` (Daniele Mari), `simone.milani@dei.unipd.it` (Simone Milani), `mauro.conti@unipd.it` (Mauro Conti)

[2] Daniele Mari's activities were supported by Fondazione CaRiPaRo under the grants "Dottorati di Ricerca" 2021/2022.

metric samples [11] extending the datasets with new synthetic IDs. Indeed, scientific literature has recently proposed the use of GANs for faces, voices, fingerprints. In this way, biometric training data could be generated by means of some GAN APIs that provide as many samples as required, thus improving users' privacy since no real acquisitions would need to be shared.

Although this seems to solve the aforementioned problems, the inherent learned nature of such solutions poses the same security threats, since it allows an attacker to partially identify samples belonging to the original dataset (e.g., Membership Inference Attack (MIA) on GANs [9]).

In this paper, we push forward this possibility by showing that it is possible to infer the identity of the subjects adopted to train the generative models. This task is relatively new among inference attacks since the attacker does not need to verify if a specific biometric sample has been used to train the algorithm (MIA): he just needs a new acquisition from a specific user and infer if it could possibly match with the unknown training set (Identity Inference Attack). Without having access to any of the actual fingerprints in the training dataset, an attacker could find out if a person contributed to its creation. We tested such a possibility on fingerprint-generating GANs since inference attacks have not been conducted on this domain before (most of the existing attacks concern faces) and inference problems are in this case more sensitive due to the difficulties in acquiring new samples.

The main contribution of the current paper can be summarized as follows.

- We propose an Identity Inference Attack (IIA) attack for biometric generative models that estimates whether some samples from a specific target user have been employed in training the analyzed model without knowing or having the training data. This work departs from [9] since it is intended to infer the identity of the user, instead of the data membership (i.e., whether a given sample has been used in training). To the best of our knowledge, this is the first work to tackle such problems on fingerprint GANs.
- We performed MIA and IIA on fingerprint-generating GANs. This work is the first analysis of privacy issues for generative models concerning the creation and processing of fingerprints. Considering the difficulties in acquiring exhaustive fingerprint datasets, the sensitivity of such samples, and the possible advantages of the adoption of GANs in this field, this analysis is extremely useful with respect to the current state-of-the-art in biometric generative schemes.
- The presented results can be easily generalized to other types of biometric samples or other generative architectures.

In the following, Section 2 overviews the existing literature on this subject, while Section 3 describes the proposed attack. Section 4 introduces the experimental setting and Section 5 presents the relative results. In the end, Section 6 discusses the possible implications, and Section 7 draws the final conclusions.

2. Related Works

One of the most recent trends in the privacy protection fields for fingerprint images consists in using GANs [10] to extend the training set for machine learning algorithms. On the other hand, researchers in the MIA community have proposed various black and white box attacks on generative models.

2.1. Generative Algorithms for fingerprints data

GANs have been widely used in the literature as a tool to generate new fingerprints with high-quality minutiae [12, 13, 14, 15] or to reconstruct partially degraded ones [16, 17]. In particular, in [12] they use DC-GAN [18] together with a connectivity regularization loss to generate faithful fingerprint images. Then in [13] the authors use I-WGAN [19] together with an identity loss that allows to generate fingerprints from the same user. In [14], fingerprints are generated in two steps: first, a StyleGAN-2 [11] based model is used to produce the skeleton of the fingerprint; then, [20] applies CycleGAN to perform style transfer and transform these skeletons into actual images. This allows to obtain very high-quality images compared to previous methods.

Finally, the approach in [15] also uses StyleGAN2 to generate samples, but the input signal is created by an encoder that permits selecting the position of the minutiae thus generating fingerprints with consistent identities.

2.2. Membership inference attacks

Given a set of data points, and a machine learning model, MIA consists in guessing which of the samples were used to train the algorithm i.e. are members of the training set. In general, the MIA is said to be white-box if the attacker has access to the weights of the model while it is said to be black-box if he can only analyze the predictions. The latter can be further subdivided by incrementally removing information (e.g. when attacking a model that only returns the top-n predicted classes).

One of the first works to introduce MIAs is [6] where "machine learning as a service" classification models were attacked. This is a good example of a black box setting since even the Machine Learning (ML) algorithm is unknown to the attacker. The authors' strategy was to train multiple shadow models (one for each class) to mimic the behavior of the attacked one. Afterward, they train a classifier to assess the membership of the samples using the shadow models as a proxy for the attacked one allowing the classifier to be later used to attack the actual black box cloud API.

Following this idea, many works have proposed more reliable and effective attacks and countermeasures [8, 21, 22]. In [21] the correctness of the prediction and the loss value are used as discriminative factors, in [22] the prediction entropy is used as a feature instead.

Additionally, many other MIAs have been designed to work on different types of models (e.g. generative ones), for example in [9] the authors propose the first MIA attack on GANs. In the white-box scenario, the authors use the discriminator prediction confidence as the discriminative factor between members and non-members, while, in the black-box scenario, they train a GAN to mimic the target model and then they exploit the newly trained discriminator to perform the white box attack. On the other hand in [23] the authors use Monte Carlo integration to approximate the probability that the item is a member.

Expanding MIAs some researchers have shown that it is not only possible to assess the membership of a sample in the training set but also of the user depicted in the photo himself. In particular, in [24] it is shown that it is possible to understand if photos of a user were used to train a metric embedding learned system by looking at how tightly some images of the person are clustered by the network. And additionally, these images don't even need to have been used as training samples, showing that the model is actually memorizing the user identity in its weights. Some additional works have shown that IIAs can be carried out also to assess authorship of samples used to train language models [25] and to find out if someone's voice was used to train some voice services [26].

3. Identity inference on GANs

In order to evaluate the feasibility of identity inference for biometric generative models, we considered Convolutional Generative Adversarial Networks (C-GANs) approaches since they are the most frequently adopted in biometric applications and allow us to generalize the obtained results to the whole family of models.

The black box attack that we carry out against all the trained models is inspired by the one proposed in [9] (see scheme in Figure 1). Our main assumptions are that the attacker has access to some APIs that allow generating from the target C-GAN $\mathcal{G}_a, \mathcal{D}_a$ as many samples as he wants (the structure of the target model is not known) and he/she is in possession of a dataset containing some of the samples used to train $\mathcal{G}_a, \mathcal{D}_a$ (called the query dataset).

The attacker trains a shadow GAN $\mathcal{G}_s, \mathcal{D}_s$ to mimic $\mathcal{G}_a, \mathcal{D}_a$. More precisely, $\mathcal{G}_s$ is trained to generate samples as close as possible to those produced by $\mathcal{G}_a$, while the discriminator $\mathcal{D}_s$ classifies whether the sample was generated by $\mathcal{G}_a$ or $\mathcal{G}_s$. In this way, $\mathcal{D}_s$ will infer some peculiar features that

simile a quando si fa train di una GAN normale ma il discriminatore non sceglie tra immagini generate e quelle di training, al posto di quelle di training ha quelle generate dal target model

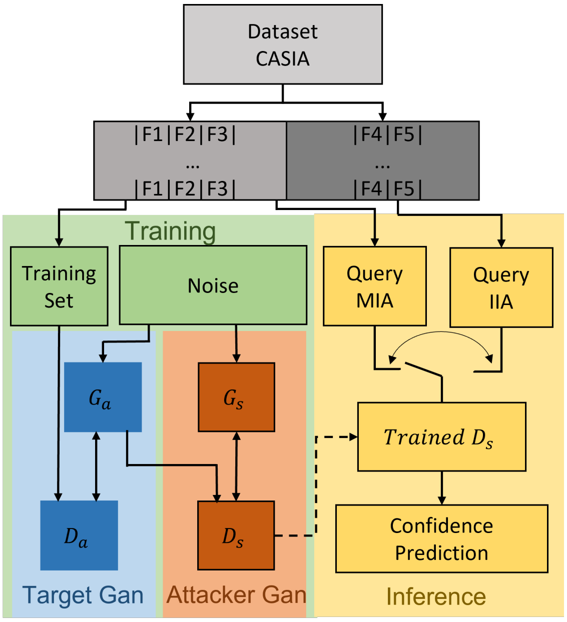

Fig. 1. Schema of the attack

$\mathcal{G}_a$ inadvertently introduces in the generated samples (likely due to overfitting). At this point, the confidence values outputted by $\mathcal{D}_s$ make it possible to sort the samples in the query dataset. Ideally, the top k samples are going to have a higher likelihood of being members of the training set, thus achieving a successful MIA.

The main intuition behind this attack is that $\mathcal{D}_s$ should be able to better recognize images used to train $\mathcal{G}_a, \mathcal{D}_a$ since they should present features that are more similar to the ones displayed in samples generated by $\mathcal{G}_a$.

In [9] the authors show how the attack performance improves with the training time, but knowing in advance a sufficient number of iterations permits reducing the computational burden.

In this work, we show that by exploiting Inception Score (IS) [27] as the metric for early stopping it is possible to obtain good MIA performance. In particular, we stop training when the IS of the shadow model is similar to the one of the attacked one.

Lastly, we also explore the task of IIA. In this case, the aim is not only to determine if a given impression was used during training but also to infer if the specific finger (i.e., identity) itself was used in the training process. In this case the assumptions formulated above slightly change, in particular, the query dataset is modified so that it does not contain the same impression of the finger used in the training datasets. This is a more realistic scenario since it is unlikely for an attacker to use the same biometric samples. The black box attack remains exactly identical to the one carried out in the MIA case since also in IIA we would like $\mathcal{D}_s$ to recognize features of the fingerprints that were present in the training set, and thus also on other impressions of the same fingerprints.

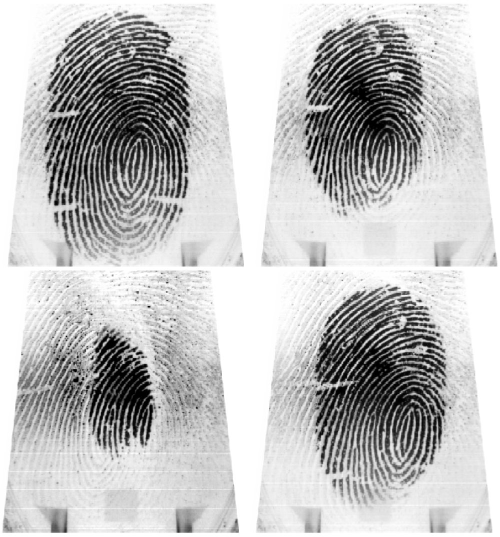

Fig. 2. 4 different impressions of the same finger

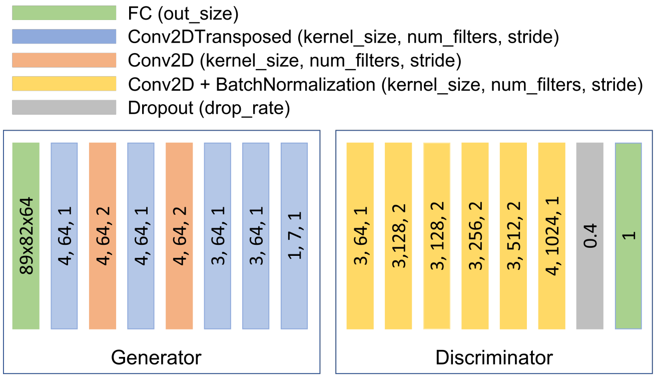

Fig. 3. Architectures of $\mathcal{G}_a, \mathcal{D}_a$

4. Experimental Setup

In our experimental set-up, we trained different C-GANs using impressions from the CASIA-FingerprintV5 dataset [28] which contains 20.000 fingerpint images of 500 subjects. These were captured using URU 4000 sensor. Each volunteer contributed with 40 fingerprint images obtained by getting five scans (328 × 356 resolution) out of eight different fingers. The volunteers varied rotation and pressure generating significant differences in the quality of the different acquisitions (see Figure 2).

The attacked architecture is defined in Figure 3, we use LeakyReLU as an activation function made exception for the last layer of the generator where we use TANH and in the last layer of the discriminator where we use Sigmoid (these implementation choices are rather generic and shared by most of the GAN architectures in the literature). The attacking GAN model was trained using Adam optimizer and the standard loss, however, we noticed that the discriminator was producing a very skewed prediction distribution making it hard to distinguish between samples where it had high and low confidence. To address this we added label smoothing with a smoothing factor equal to 0.2.

In order to assess the robustness of the trained models against MIAs, the considered C-GANs were retrained using all the dataset splits specified in Table1. In all the data splits we only keep 3 acquisitions per finger in the training dataset, while the others are added to the IIA query dataset.

On the other hand, for what concerns the attacking model, we performed model selection using the IS metric (as mentioned above) similarly to what a black box attack would need to do in a real scenario. As a result, we created diverse architectures with different parameters depending on the dataset that was originally used to train $\mathcal{G}_a, \mathcal{D}_a$. Also in this case, the different networks were obtained by stacking convolutional layers (with a structure similar to the one of the attacked GAN) whose number was varied in order to satisfy the IS criterion. The different configurations are here omitted for the sake of conciseness.

5. Results

First experimental considerations concern the quality of the generated images and the effectiveness in using IS as a termination criterion. Figure 4 shows that the generated fingerprint images consistently present level 1 (ridge orientation and singular points) and 2 (minutiae) features, with the sporadic presence of level 3 (pores) features as well. This is further highlighted by the IS values obtained on the generated images, which are comparable to those computed on fingerprints. It is also possible to notice that the lower the amount of training images the lower the IS value.

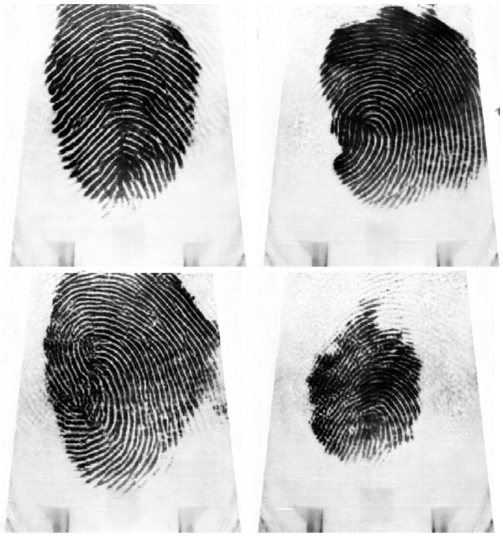

Fig. 4. 4 generated fingerprints impression

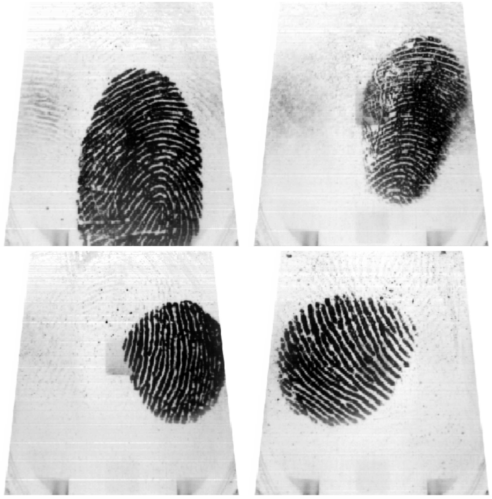

Fig. 5. On the left, highest confidence fingerprints that are actual members. On the right highest scoring images that are different impressions of training samples. (Gan_2400 and Gan_4800)

In Table 2 it is possible to find the number of members and corresponding fingerprints found in the top 20/200/2000 samples by confidence (except for the cases where the training set size is smaller than 2000). It is possible to see that for all the trained GANs the percentage of samples in the first $n \in \{20, 200, 2000\}$ is always above 50% showing that this type of attack actually allows to gather some information about the training dataset. Additionally, it is possible to see that the performance of the attack is only slightly worse

Name	Real Images	D1	D2	D3	D4
Samples / Users / Fingerprints	2000/50/5	9600/400/3	4800/200/3	2400/100/3	1200/50/3
IS Mean	1639.2	1620.13	1563.79	1570.89	1571.28
IS STD	7.92	11.95	15.99	18.66	10.29

Table 1. Datasets splits and inception score metrics on the trained GANs for each of them

	MIA			IIA		
GAN	top 20	top 200	top 50%	top 20	top 200	top 50%
Gan_9600	14/20	126/200	1172/2000	12/20	114/200	1161/2000
Gan_4800	19/20	144/200	1133/2000	18/20	140/200	1133/2000
Gan_2400	12/20	108/200	869/1600	13/20	105/200	869/1600
Gan_1200	13/20	132/200	526/800	12/20	137/200	540/800

Table 2. MIA and IIA Results.

when performing IIA compared to MIA showing that, even if this kind of attack is more challenging for an attacker, it can be carried out with significant success.

This demonstrates the effectiveness of MIA and IIA in detecting sample and finger membership in the training dataset for a GAN. As a matter of fact, the attacks lead to near-perfect recognition of many fingerprints as belonging to the targeted sample in both scenarios.

An additional proof that the success of the attack is provided by a visual inspection of images where $\mathcal{D}_s$ confidence suggests that they are more likely to belong to the training dataset. These indeed exhibit various similarities especially within the Level 1 features as can be seen from Figure 5.

Especially in the case of the GAN trained on 4800 samples, out of the 20 flagged images with the highest scores, 19 of them are indeed from the training dataset. Once the actual training images are swapped with a set of ”remaining” acquisitions, i.e., different samples of the same fingers, the result remains high with 18 out of 20 correctly identified showing how much information these types of models can actually leak even though they were designed in the first place to avoid privacy concerns.

6. Discussion

From Table 2 it is possible to see that all the trained models could be attacked by the proposed approach. Additionally, since the attack was successful on all the four cases, it is possible to conclude that IS might be a good heuristic when choosing when to stop the training procedure.

Despite this, experimental tests did not reveal how the number of training samples affects the attack performance since the general success is also strongly dependent on other factors such as how well-trained the attacked model is or the batch size and optimizer, to mention some of them. However, considering other results on MIA presented in the literature, it is reasonable to believe that increasing the total number of samples by a significant amount implies improving the robustness against MIAs and IIAs.

7. Conclusion

This work illustrates the main vulnerabilities of generative adversarial networks as a mean to solve the data shortage and privacy issues for learned biometric architectures. More precisely, Membership and Identity Inference Attacks on fingerprint GANs are described and evaluated showing how it is possible to infer users' identity from a trained black box model.

To the best of our knowledge, we are the first to notice the severity of the problem on fingerprint data showing that it is possible to assess membership of a sample and to detect if a person contributed to the creation of a dataset without having access to any of the training data. In future works, researchers should analyze how this problem affects other kinds of biometric samples and architectures. Additionally, some defense strategies should be designed in order to make fingerprint generation a truly secure way to solve the data shortage problem in the biometric field.

References

[1] M. M. Kasar, D. Bhattacharyya, T. Kim, Face recognition using neural network: a review, International Journal of Security and Its Applications 10 (3) (2016) 81–100.

[2] G. Guo, N. Zhang, A survey on deep learning based face recognition, Computer vision and image understanding 189 (2019) 102805.

[3] Y. Liu, B. Zhou, C. Han, T. Guo, J. Qin, A novel method based on deep learning for aligned fingerprints matching, Applied Intelligence 50 (2020) 397–416.

[4] B. Rim, J. Kim, M. Hong, Fingerprint classification using deep learning approach, Multimedia Tools and Applications 80 (2021) 35809–35825.

[5] K. Sundararajan, D. L. Woodard, Deep learning for biometrics: A survey, ACM Computing Surveys (CSUR) 51 (3) (2018) 1–34.

[6] R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, in: 2017 IEEE symposium on security and privacy (SP), 2017, pp. 3–18.

[7] C. A. Choquette-Choo, F. Tramer, N. Carlini, N. Papernot, Label-only membership inference attacks, in: International conference on machine learning, 2021, pp. 1964–1974.

[8] H. Hu, Z. Salcic, L. Sun, G. Dobbie, P. S. Yu, X. Zhang, Membership inference attacks on machine learning: A survey, ACM Computing Surveys (CSUR) 54 (11s) (2022) 1–37.

[9] J. Hayes, L. Melis, G. Danezis, E. De Cristofaro, Logan: Membership inference attacks against generative models, arXiv preprint arXiv:1705.07663 (2017).

[10] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[11] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, T. Aila, Analyzing and improving the image quality of stylegan, in: Proc.of the IEEE/CVF CVPR 2020, 2020, pp. 8110–8119.

[12] S. Minaee, A. Abdolrashidi, Finger-gan: Generating realistic fingerprint images using connectivity imposed gan, arXiv preprint arXiv:1812.10482 (2018).

[13] V. Mistry, J. J. Engelsma, A. K. Jain, Fingerprint synthesis: Search with 100 million prints, in: 2020 IEEE International Joint Conference on Biometrics (IJCB), 2020, pp. 1–10.

[14] A. Sams, H. H. Shomee, S. M. Rahman, Hqfingan: High-quality synthetic fingerprint generation using gans, Circuits, Systems, and Signal Processing 41 (11) (2022) 6354–6369.

[15] R. Bouzaglo, Y. Keller, Synthesis and reconstruction of fingerprints using generative adversarial networks, arXiv preprint arXiv:2201.06164 (2022).

[16] X. Huang, P. Qian, M. Liu, Latent fingerprint image enhancement based on progressive generative adversarial network, in: Proc. of the IEEE/CVF CVPR Workshops 2020, 2020, pp. 800–801.

[17] I. Joshi, A. Anand, M. Vatsa, R. Singh, S. D. Roy, P. Kalra, Latent fingerprint enhancement using generative adversarial networks, in: Proc. of IEEE WACV 2019, 2019, pp. 895–903.

[18] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv preprint arXiv:1511.06434 (2015).

[19] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, A. C. Courville, Improved training of wasserstein gans, Advances in neural information processing systems 30 (2017).

[20] J.-Y. Zhu, T. Park, P. Isola, A. A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE ICCV 2017, 2017, pp. 2223–2232.

[21] S. Yeom, I. Giacomelli, M. Fredrikson, S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, in: 2018 IEEE 31st computer security foundations symposium (CSF), 2018, pp. 268–282.

[22] A. Salem, Y. Zhang, M. Humbert, M. Fritz, M. Backes, Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models, in: Network and Distributed Systems Security Symposium 2019, 2019.

[23] B. Hilprecht, M. Härterich, D. Bernau, Monte carlo and reconstruction membership inference attacks against generative models., Proc. Priv. Enhancing Technol. 2019 (4) (2019) 232–249.

[24] G. Li, S. Rezaei, X. Liu, User-level membership inference attack against metric embedding learning, arXiv preprint arXiv:2203.02077 (2022).

[25] C. Song, V. Shmatikov, Auditing data provenance in text-generation models, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 196–206.

[26] Y. Miao, X. Minhui, C. Chen, L. Pan, J. Zhang, B. Z. H. Zhao, D. Kaafar, Y. Xiang, The audio auditor: user-level membership inference in internet of things voice services, Proceedings on Privacy Enhancing Technologies 2021 (2021) 209–228.

[27] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training gans, Advances in neural information processing systems 29 (2016).

[28] Chinese academy of sciences' institute of automation casia-fingerprint-v5 dataset, `http://biometrics.idealtest.org/` (Last Accessd 15/06/2023).