Unveiling Synthetic Faces: How Synthetic Datasets Can Expose Real Identities

Hatef Otroshi Shahreza[1,2] **and Sébastien Marcel**[1,3]
[1]Idiap Research Institute, Martigny, Switzerland
[2]École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[3]Université de Lausanne (UNIL), Lausanne, Switzerland
{hatef.otroshi,sebastien.marcel}@idiap.ch

Fig. 1: Sample face images leaked from training data (first row) of generative models in different state-of-the-art synthetic face recognition datasets (second row).

Abstract

Synthetic data generation is gaining increasing popularity in different computer vision applications. Existing state-of-the-art face recognition models are trained using large-scale face datasets, which are crawled from the Internet and raise privacy and ethical concerns. To address such concerns, several works have proposed generating synthetic face datasets to train face recognition models. However, these methods depend on generative models, which are trained on real face images. In this work, we design a simple yet effective membership inference attack to systematically study if any of the existing synthetic face recognition datasets leak any information from the real data used to train the generator model. We provide an extensive study on 6 state-of-the-art synthetic face recognition datasets, and show that in all these synthetic datasets, several samples from the original real dataset are leaked. To our knowledge, this paper is the first work which shows the leakage from training data of generator models into the generated synthetic face recognition datasets. Our study demonstrates privacy pitfalls in synthetic face recognition datasets and paves the way for future studies on generating responsible synthetic face datasets. Project page: https://www.idiap.ch/paper/unveiling_synthetic_faces

AdvML-Frontiers'24: The 3rd Workshop on New Frontiers in Adversarial Machine Learning@NeurIPS'24, Vancouver, CA.

1 Introduction

Recent advancements in state-of-the-art face recognition models are achieved by training deep neural networks with penalty-based softmax loss functions on large-scale datasets [1–4]. Existing face recognition datasets contain millions of face images, such as VGGFace2 [5], MS-Celeb-1M [6], WebFace260M [7], which are typically collected by crawling the Internet without proper user consent. This raises ethical and legal concerns about the use of such datasets for the development of face recognition models. In particular, recent data regulation frameworks, such as the European Union Artificial Intelligence Act, further support the rights of subjects whose data are used in a dataset to train such models. Consequently, several face recognition datasets, including but not limited to [5, 6], have been retracted by their creators, to prevent potential legal issues. Therefore, the availability of such large-scale datasets and the possibility of research in the face recognition task has become uncertain.

Recently, generating synthetic face recognition datasets has emerged as a promising alternative to large-scale real datasets and has become a promising solution to address the ethical and legal concerns [8–15]. The generation of synthetic face datasets, however, relies on the development of face generative models, which enable the generation of synthetic samples from the probability distribution of a real face dataset. Meanwhile, most synthetic face recognition datasets are built upon Generative Adversarial Networks (GANs) [8, 9] or Diffusion Models (DMs) [10, 11]. These face generative models are trained on a dataset of real face images, and therefore leakage of information from the training dataset to the generated face images can raise privacy concerns in the generated synthetic face recognition datasets. Along the same lines, several studies in the literature of generative models have shown the memorization issue in different models [16–24], which sparks concerns on the application of generative models for privacy-sensitive problems. In this paper, we design a simple yet effective membership inference attack against synthetic face datasets to systematically study if any of the existing samples in the generated dataset leak any information from the real data used to train the generator model. We provide an extensive evaluation of 6 state-of-the-art synthetic face recognition datasets, and demonstrate that in all these synthetic datasets, several samples from original real datasets are leaked. Fig. 1 illustrates sample face images from 6 state-of-the-art synthetic face recognition datasets that are leaked from the training set of their generator models. To our knowledge, this paper presents the first work which shows the leakage from training data of generator models into synthetic face recognition datasets. Our study demonstrates privacy pitfalls in synthetic face recognition datasets and paves the way for future studies on generating responsible synthetic face datasets.

In the remainder of the paper, we first review state-of-the-art synthetic face recognition datasets in the literature in Section 2. Then, we describe our membership inference attack and present our evaluation of state-of-the-art synthetic face recognition datasets in Section 3. In Section 4, we discuss our findings and explain the limitations of our study as well as current shortcomings in the literature. Finally, the paper is concluded in Section 5.

2 Related Work

As discussed earlier, existing synthetic face recognition datasets are typically generated using a generative model. While some papers used pretrained face generative models (such as pretrained StyleGAN on the FFHQ dataset), other works retrained a generative model on another dataset or proposed a new face generator model. Boutros et al. [8] used the StyleGAN2-ADA [25] as their generative model and trained it on the CASIA-WebFace dataset [26] with identities serving as class labels. Then, they utilized their identity-conditioned StyleGAN2-ADA model to generate the SFace dataset of generated face images, and demonstrated its effectiveness for training face recognition algorithms.

Kolf et al. [9] also used StyleGAN2-ADA [25] trained on the CASIA-WebFace dataset [26] and proposed a three-player GAN framework to generate the IDNet dataset. Their three-player framework integrates identity information into the image generation of StyleGAN, where the third player is used to force the generator network to generate identity-separable face images.

In contrast to most works in the literature that used GAN-based generator models, some works generated synthetic datasets using diffusion models. Kim et al. [10], introduced the dual (identity and style) condition face generator based on a diffusion model and trained it on the CASIA-WebFace [26] dataset. They used a patch-wise style extractor combined with a time-step dependent ID loss to train their generator model. Then, they generated the DCFace dataset by synthesizing different

Table 1: Synthetic Face Recognition Datasets in the Literature.

Reference	Synthetic Dataset	Generator	Training Dataset
[8]	SFace	StyleGAN-ADA (identity-conditioned)	CASIA-WebFace
[9]	IDNet	StyleGAN-ADA (identity-conditioned)	CASIA-WebFace
[10]	DCFace	new diffusion model (identity and style conditioned)	CASIA-WebFace
[11]	IDiff-Face (Uniform) IDiff-Face (Two-stage)	new diffusion model (identity-conditioned)	FFHQ
[12]	GANDiffFace	StyleGAN (pretrained) DreamBooth (pretrained)	FFHQ LAION

identities using identity condition and also different samples per identity using the style condition. They published two versions of their dataset, DCFace-0.5M and DCFace-1.2M, where the smaller version is a subset of the larger one.

In [11], the authors trained a latent diffusion model conditioned on identity features obtained from a pretrained face recognition model using the FFHQ dataset. For sample generation, they used their trained diffusion model to generate different identities by randomly sampling the identity context from a uniform distribution. In another approach, they used an unconditional diffusion model to generate different identities (two-stage). To generate different samples per identity, they fixed the identity condition and changed the latent noise. Considering these two different generation approaches, they proposed two datasets called IDiff-Face (Uniform) and IDiff-Face (Two-stage), respectively.

In [12], Melzi et al. introduced the GANDiffFace dataset, which is generated using both GAN-based and diffusion-based generators. They first used a pretrained StyleGAN3 [27] model (trained on FFHQ) to generate different identities, and then used a pretrained DreamBooth [28] model to generate different samples for each identity. DreamBooth [28] is a diffusion model based on the Stable Diffusion [29] model that is trained on the LAION dataset [30].

Table 1 summarizes different synthetic face recognition datasets in the literature which are generated using generative models. We should note that another category of methods to generate synthetic datasets is computer-graphic-based methods, e.g., DigiFace-1M [13], which is excluded in this study.

3 Membership Inference Attack

As described in Section 2, synthetic face recognition datasets are often generated using a generator model. Let D_{real} denote the real face recognition dataset that is used to train the generator model. Then, based on the synthetic dataset generation approach, the generator model is used to generate a synthetic face recognition dataset $D_{\text{synthetic}}$, which includes various synthetic identities and different sample images per each identity. Therefore, an important question is whether any of the generated images in the generated synthetic face recognition dataset $D_{\text{synthetic}}$ contain important information from training dataset D_{real}, which was used to train the face generator model in the first place? In other words, do we have any leakage of information from D_{real} to $D_{\text{synthetic}}$? Fig. 2 illustrates the process of leakage of information from the training dataset into the generated synthetic dataset.

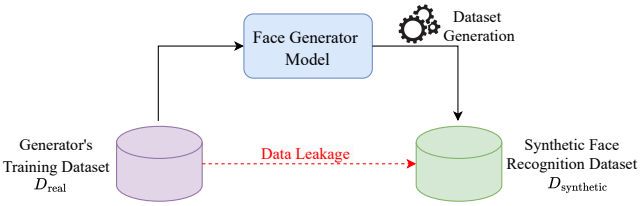

Fig. 2: Schematic diagram of data leakage from generator's training data into generated synthetic face recognition dataset.

We consider an exhaustive search approach to compare all possible pairs of images from D_{real} and $D_{\text{synthetic}}$. To this end, we use an off-the-shelf face recognition model $F(.)$ to extract face embeddings from each face image, and then compare the embeddings of every pair of images from D_{real} and $D_{\text{synthetic}}$. Then, we sort the pairs of images according to the similarity of embeddings and consider the top-k pairs for visual comparison of images. Algorithm 1 presents the pseudo-code of our approach.

For our experiments, we use an off-the-shelf face recognition model with ResNet100 backbone which is trained with AdaFace [2] loss function on the WebFace12M dataset [7]. We use the method presented in algorithm 1 to find pairs with high cosine similarity scores from the training dataset of the generator model and the generated synthetic dataset for all the synthetic datasets in Table 1.

Algorithm 1 Membership Inference Attack against Synthetic Datasets (MIS).

Require: : D_{real}, dataset used to train generator network; $D_{\text{synthetic}}$, generated synthetic face dataset; F, an off-the-shelf face recognition model; SIM, similarity function (e.g., cosine similarity) to compare two embeddings extracted by face recognition F; k, number of top similar pairs to return for visual comparison.

procedure MIS ATTACK
  Initialize list $\mathcal{S} = [\,]$
  for $I_{\text{synthetic},i} \in D_{\text{synthetic}}$ **do**
    for $I_{\text{real},j} \in D_{\text{real}}$ **do**
      $s = \text{SIM}(F(I_{\text{synthetic},i}), F(I_{\text{real},j}))$
      $\mathcal{S}$.append(s)
    end for
  end for
  return $\mathcal{S}$.sort()$[0:k]$
end procedure

Figs. 4-9 of Appendix illustrate sample face images from the training dataset of generator models which are leaked in the generated synthetic dataset for 6 state-of-the-art synthetic datasets, including DCFace [10], IDiff-Face (Uniform) [11], IDiff-Face (Two-stage) [11], GANDiffFace [12], IDNet [9], and SFace [8] datasets, respectively. The corresponding training dataset of generator models used to generate each synthetic dataset is reported in Table 1. In the case of GANDiffFace [12] which uses two pretrained generator models, we compare with the training dataset of StyleGAN which was used in the first stage to generate different synthetic identities. For the evaluation of the DCFace dataset, we consider its smaller set (i.e., DCFace 0.5M), which is also included in the larger version of this dataset. As Figs. 4-9 of Appendix show, the generated synthetic datasets contain very similar images from the training set of their generator model, which raises concerns regarding the generation of such identities.

We set the parameter k in algorithm 1 equal to 1500, and compared the retrieved pairs of images. However, in some cases, such as DCFace, we can easily find leaked matches in top-100 pairs. Fig. 3 illustrates the histogram of similarity scores of all retrieved pairs of images for each synthetic dataset and their corresponding values of similarity for top-k pairs (as dashed vertical lines). In addition, this figure depicts the similarity scores for positive and negative pairs in the IARPA Janus Benchmark–C (IJB-C) [31] as a benchmark dataset using the same face recognition model used in our experiments. This figure also shows the threshold for False Accept Rate (FAR) of 0.01% on IJB-C dataset (dotted vertical lines). Surprisingly, as can be seen Fig. 3, almost all retrieved pairs of images in each synthetic dataset has a similarity greater than the matching threshold on IJB-C (FAR=0.01%). That means if we have a face recognition system configured at this threshold, almost all retrieved pairs of images (i.e., each image in the synthetic dataset and its closest image in the training dataset) are considered as the same identity. In other words, almost for each image in the generated synthetic dataset there is one image in the training dataset for which the face recognition model (configured at FAR=0.01% on IJB-C) will recognise as the same identity. However, concluding the

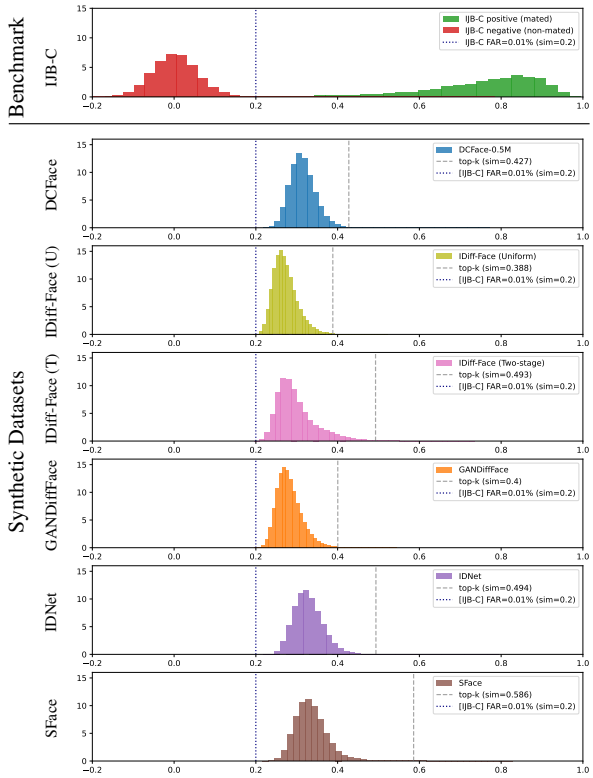

Fig. 3: Histogram of cosine similarity scores of all retrieved pairs of images for each synthetic dataset and their corresponding values of similarity for top-k pairs as dashed vertical lines (k=1500). The first plot shows the histogram of similarity scores for positive and negative pairs in IJB-C dataset (benchmark) and the threshold for FAR=0.01% on IJB-C with dotted vertical lines.

leakage based on the histogram of similarity scores has some limitations which are discussed in detail in Section 4.

We should note that the source code of our experiments as well as the file names of all images reported as leaked samples in this paper are publicly available in our project page[1] to facilitate the reproducibility of our study and help future researchers to build upon our findings.

4 Discussion

Our experiments demonstrate that state-of-the-art synthetic face recognition datasets contain samples that are very close to samples in the training data of their generator models. In some cases the synthetic samples contain small changes to the original image, however, we can also observe in some cases the generated sample contains more variation (e.g., different pose, light condition, etc.) while the identity is preserved. This suggests that the generator models are learning and memorizing the identity-related information from the training data and may generate similar identities. This creates critical concerns regarding the application of synthetic data in privacy-sensitive tasks, such as biometrics and face recognition.

The findings in our paper open several new research questions and introduce new research directions that require attention from the community:

- In this paper, we used an exhaustive search approach to find samples which contain information leaked in the generated dataset. While our approach is effective in finding samples in the synthetic dataset that are similar to the training dataset, it requires comparing all possible pairs of images. However, comparing all possible pairs may not be efficient and such a membership inference attack can be deployed more efficiently. In particular, if the training data or generated data have larger samples[2], the required computation for all comparisons similarly increases.
- While our attack algorithm can find samples in the training data which are leaked into the synthetic dataset, it also returns several samples that are not necessarily obvious to contain leakage. As an example, we can refer to three categories which we observed in our experiments:
    1. In some cases, the similar images found by our approach are images of children. However, not only the face recognition models also have a high error for children, but also distinguishing if two images are for the same child is difficult for human observers. Therefore, considering children's images for indicating a possible leakage is not reliable. This particularly happens for synthetic datasets whose generator models are trained on the FFHQ dataset, in which the population of young children is considerably large [32]. Fig. 11 of Appendix illustrates some samples of young people in IDiff-Face (Uniform), IDiff-Face (Two-stage), and GANDiffFace datasets. We ignored such samples in our evaluations and did not recognise them as leaked samples.
    2. In some samples that were returned by our algorithm with high similarity scores, the synthetic image did not include a face image. In fact, these are unexpected samples in a face recognition dataset. Therefore, we ignored such samples in our evaluations. Fig. 12 of Appendix illustrates some examples of images that do not have face images from DCFace, IDNet, and SFace datasets.
    3. In some cases, the images with high similarity scores were not recognized as the same identity or were not convincing enough to a human observer to demonstrate leakage. Therefore, we ignored such samples in our evaluations. Fig. 13 of Appendix illustrates some examples of such images in which we could not conclude leakage in visual comparison for different datasets.

    As a result of having such samples in the output of our analyses, we needed to have a visual comparison step. While we found visual comparison necessary to draw valid conclusions, it requires a human observer, even for a small number of selected samples. In addition to the required human effort, it may introduce subjective bias in visual comparisons. We should note that the samples shown in Fig. 1 and Figs. 4-10 of Appendix were selected based on

[1] Project page: https://www.idiap.ch/paper/unveiling_synthetic_faces

[2] such as LAION [30] dataset used to train Stable Diffusion [29] and includes 5 billion data.

the unanimous agreement of several observers, however, such a human evaluation may not be efficient and consistent for future studies.

- An important future direction is to propose new measures to *quantify* and benchmark the leakage of information from training datasets of generator models into the synthetic data. We would like to stress that proposing such a measure is not trivial and requires further research to address the previous question and eliminate the necessity of human observers. In particular, three cases which we mentioned in previous point of our discussion (i.e., children images, no face images, not having same identity) makes statistical analyses (based on the similarity scores of retrieved pairs) a challenging task, and thus it is difficult to extract a reliable statistical metric for the leakage in the synthetic datasets using similarity scores. For example, histograms of scores for matched pairs from synthetic dataset and corresponding closest images from the training datasets in Fig. 3 show that for almost all images in synthetic datasets, there is an image in the training dataset which is recognised as the same identity by a face recognition model. However, sample images in Figs. 11-13 of Appendix, which are even among top-k retrieved pairs with high similarity score, are not necessarily recognised as a same identity by a human observer, and therefore such samples cannot demonstrate identity leakage. Hence, evaluating the leakage based on similarity scores of retrieved pairs of images is not straightforward and requires further studies.
- Our experiments show that state-of-the-art synthetic datasets leak sensitive information from the training dataset of their generator models. Therefore, an important future direction is to generate responsible synthetic face datasets. This objective can be achieved by preventing such leakage in the data generation process or by further post-processing to eliminate such leakages. We should note that some of the existing datasets, such as DCFace [10], have already tried to prevent such leakage in data generation. In [10, Section 3.3], the authors explained that they removed samples that are more similar to images of training data than a predefined threshold. However, as images in Fig. 4 and Fig. 10 of Appendix show, such data cleaning has not been sufficient to prevent identity leakage in the generated synthetic dataset. This suggests that more efforts should be taken to avoid leakage in the synthetic data.

To draw the discussion to a close, we would like to highlight that the main motivation for generating synthetic datasets is to address privacy concerns in using large-scale web-crawled face datasets. Therefore, the leakage of any sensitive information (such as identities of real images in the training data) in the synthetic dataset spikes critical concerns regarding the application of synthetic data for privacy-sensitive tasks, such as biometrics. Our study sheds light on the privacy pitfalls in the generation of synthetic face recognition datasets and paves the way for future studies toward generating responsible synthetic face datasets.

5 Conclusion

In this paper, we explored the crucial question of *"whether synthetic datasets expose real identities used for training their generator models?"*. We used a simple yet effective membership inference attack (based on exhaustive search) against synthetic datasets, and explored if any of the generated samples in the synthetic face recognition dataset leaks any information from the training dataset of the corresponding generator model. We evaluated 6 state-of-the-art synthetic face recognition datasets generated with different deep generative models (GAN-based and diffusion-based). We reported several samples for each dataset which demonstrate leakage of information in synthetic datasets from the training data of generator models. In some cases, the retrieved samples further indicate memorization or learning identity-related information in the generator models. We also discussed the limitations of our evaluation and outlined potential future directions. To our knowledge, this paper is first work which shows the leakage from training data of generator models into synthetic face recognition dataset and reveals privacy pitfalls in the generation of synthetic face recognition datasets.

Acknowledgments

This research is based upon work supported by the Hasler foundation through the "Responsible Face Recognition" (SAFER) project as well as the H2020 TReSPAsS-ETN Marie Skłodowska-Curie early training network (grant agreement 860813).

Bibliography

[1] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 2

[2] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022. 3

[3] Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Ketan Kotwal, and Sébastien Marcel. Edgeface: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024.

[4] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. Magface: A universal representation for face recognition and quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14225–14234, 2021. 2

[5] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 2

[6] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In *European conference on computer vision*, pages 87–102. Springer, 2016. 2

[7] Zheng Zhu, Guan Huang, Jiankang Deng, Yun Ye, Junjie Huang, Xinze Chen, Jiagang Zhu, Tian Yang, Jiwen Lu, Dalong Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 2, 3

[8] Fadi Boutros, Marco Huber, Patrick Siebke, Tim Rieber, and Naser Damer. Sface: Privacy-friendly and accurate face recognition using synthetic data. In *2022 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–11. IEEE, 2022. 2, 3, 4, 9

[9] Jan Niklas Kolf, Tim Rieber, Jurek Elliesen, Fadi Boutros, Arjan Kuijper, and Naser Damer. Identity-driven three-player generative adversarial network for synthetic-based face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 806–816, 2023. 2, 3, 4, 9

[10] Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. Dcface: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12715–12725, 2023. 2, 3, 4, 6, 9

[11] Fadi Boutros, Jonas Henry Grebe, Arjan Kuijper, and Naser Damer. Idiff-face: Synthetic-based face recognition through fizzy identity-conditioned diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19650–19661, 2023. 2, 3, 4, 9

[12] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3086–3095, 2023. 3, 4, 9

[13] Gwangbin Bae, Martin de La Gorce, Tadas Baltrušaitis, Charlie Hewitt, Dong Chen, Julien Valentin, Roberto Cipolla, and Jingjing Shen. Digiface-1m: 1 million digital face images for face recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3526–3535, 2023. 3

[14] Hatef Otroshi Shahreza, Christophe Ecabert, Anjith George, Alexander Unnervik, Sébastien Marcel, Nicolò Di Domenico, Guido Borghi, Davide Maltoni, Fadi Boutros, Julia Vogel, et al. Sdfr: Synthetic data for face recognition competition. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9. IEEE, 2024.

[15] Hatef Otroshi Shahreza, Anjith George, and Sébastien Marcel. Synthdistill: Face recognition with knowledge distillation from synthetic data. In *IEEE International Joint Conference on Biometrics (IJCB 2023)*, 2023. 2

[16] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023. 2

[17] Nikhil Vyas, Sham M Kakade, and Boaz Barak. On provable copyright protection for generative models. In *International Conference on Machine Learning*, pages 35277–35299. PMLR, 2023.

[18] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

[19] Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36:47783–47803, 2023.

[20] Zhangheng Li, Junyuan Hong, Bo Li, and Zhangyang Wang. Shake to leak: Fine-tuning diffusion models can amplify the generative privacy risk. *arXiv preprint arXiv:2403.09450*, 2024.

[21] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650, 2021.

[22] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

[23] Ziqi Zhang, Chao Yan, and Bradley A Malin. Membership inference attacks against synthetic health data. *Journal of biomedical informatics*, 125:103977, 2022.

[24] Patrick Tinsley, Adam Czajka, and Patrick Flynn. This face does not exist... but it might be yours! identity leakage in generative models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1320–1328, 2021. 2

[25] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8110–8119, 2020. 2

[26] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014. 2

[27] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[28] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3

[29] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 5

[30] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294, 2022. 3, 5

[31] Brianna Maze, Jocelyn Adams, James A Duncan, Nathan Kalka, Tim Miller, Charles Otto, Anil K Jain, W Tyler Niggel, Janet Anderson, Jordan Cheney, et al. IARPA janus benchmark-c: Face dataset and protocol. In *Proceedings of the International Conference on Biometrics (ICB)*, pages 158–165. IEEE, 2018. 4

[32] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 739–755. Springer, 2020. 5

A Sample Leaked Images

Figs. 4-9 illustrate sample face images from the training dataset of generator models which are leaked in the generated synthetic dataset for 6 state-of-the-art synthetic datasets, including DCFace [10], IDiff-Face (Uniform) [11], IDiff-Face (Two-stage) [11], GANDiffFace [12], IDNet [9], and SFace [8] datasets, respectively. The corresponding training dataset of generator models used to generate each synthetic dataset is reported in Table 1 of the paper. For, GANDiffFace [12] which uses two pretrained generator models, we compared with the training dataset of StyleGAN which was used in the first stage to generate different synthetic identities. As these figures show, the generated synthetic datasets contain very similar images from the training set of their generator model, which raises concerns regarding the generation of such identities. In some cases, such as in DCFace (Fig. 4 of Appendix), the similarity is very high, and the generated image has some small visual changes compared to the original training data. However, in some other datasets, the difference is higher, nevertheless, the identities of generated images look similar. We should note that images shown in Fig. 1 and Figs. 4-9 of Appendix are some samples for each dataset, and for some of these synthetic datasets we can easily find more samples. For example, Fig. 10 of Appendix illustrates more samples in the DCFace dataset.

Fig. 4: Sample face images leaked from training data (first row) of the generative model in the **DCFace** dataset (second row). For more samples see Fig. 10

Fig. 5: Sample face images leaked from training data (first row) of the generative model in the **IDiff-Face (Uniform)** dataset (second row).

Fig. 6: Sample face images leaked from training data (first row) of the generative model in the **IDiff-Face (Two-stage)** dataset (second row).

Fig. 7: Sample face images leaked from training data (first row) of the generative model in the **GANDiffFace** dataset (second row).

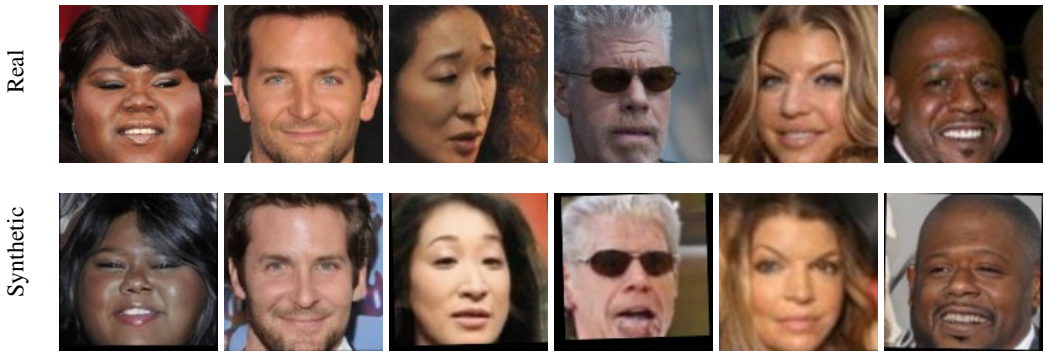

Fig. 8: Sample face images leaked from training data (first row) of the generative model in the **IDNet** dataset (second row).

Fig. 9: Sample face images leaked from training data (first row) of the generative model in the **SFace** dataset (second row).

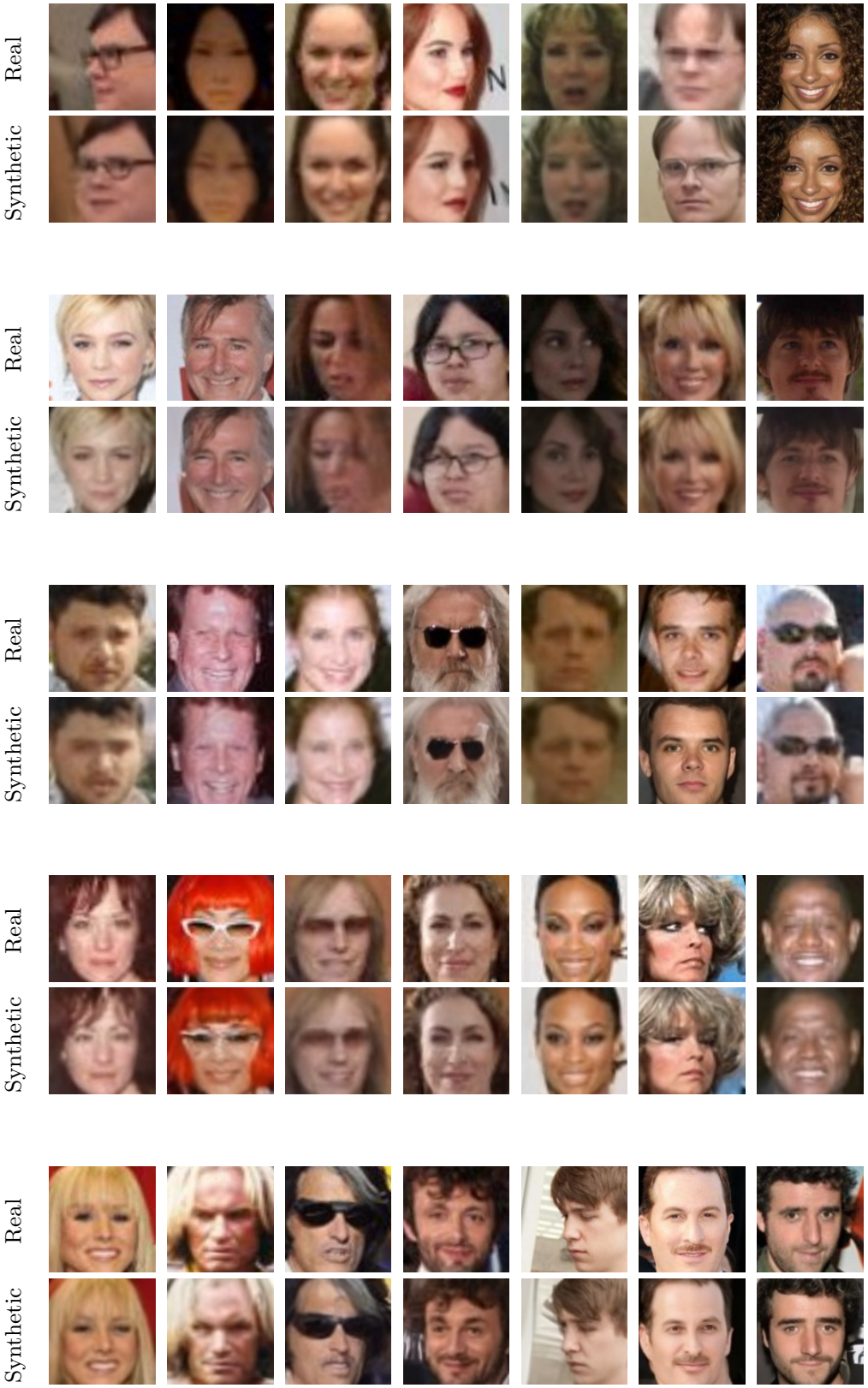

Fig. 10: Sample face images leaked from training data (CASIA-WebFace) of the generative model in the **DCFace** dataset.

B Sample of Difficult Matching Face Images

As discussed in Section 4 of the paper, while our attack algorithm can find samples in the training data which are leaked into the synthetic dataset, it also returns several samples that are not necessarily obvious to contain leakage. Figs. 11-13 illustrates sample of face images, which were difficult to match while having high similarity score. We ignored such samples in our evaluation and did not recognise them as leaked samples.

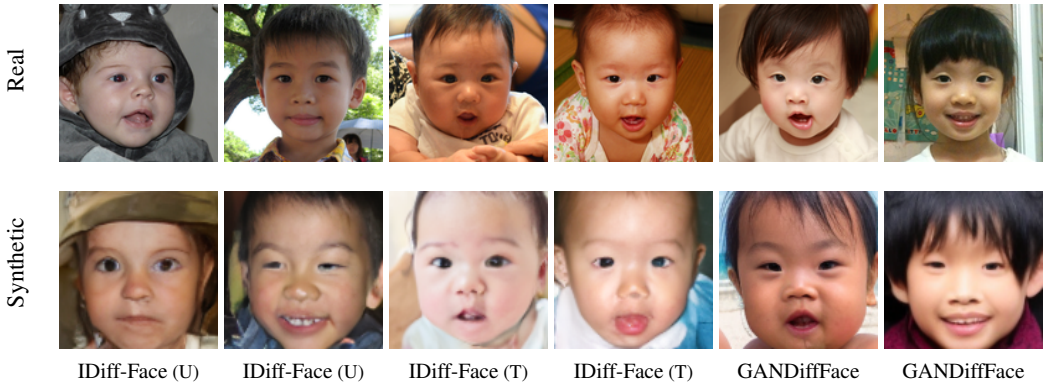

Fig. 11: Sample face images of **children** which have high similarity, but we **ignored** them in our evaluations (i.e., we did not recognise them as leaked samples).

Fig. 12: Sample images which **do not include a face**, and we **ignored** them in our evaluations (i.e., we did not recognise them as leaked samples).

Fig. 13: Sample face images of which have high similarity scores, but are not convincing for human observer to demonstrate data leakage in the synthesized image. Therefore, we **ignored** them in our evaluations (i.e., we did not recognise them as leaked samples).