Attribute-preserving Face Dataset Anonymization via Latent Code Optimization – Supplementary Material –

Simone Barattin*[1] Christos Tzelepis*[2] Ioannis Patras[2] Nicu Sebe[1]

[1]University of Trento
simone.barattin@studenti.unitn.it, niculae.sebe@unitn.it

[2]Queen Mary University of London
{c.tzelepis, i.patras}@qmul.ac.uk

Pre-trained attribute classifiers

As discussed in the Sect. 4.2.2 of the main paper, for the evaluation of the attribute preservation ability of the proposed and other state-of-the-art anonymization methods, in the case of the LFW [2] dataset, due to the absence of attribute labels, we obtain pseudo-labels provided by two pre-trained attribute classifiers, both pre-trained on the CelebA [7] dataset. Specifically, we use the pre-trained models provided by Anycost GAN [6] and Talk-to-Edit [4]. The former provides predictions (i.e., pseudo-labels) for the whole set of the 40 attributes of the CelebA dataset, while the latter for 5 of them (namely, *"Bangs"*, *"Eyeglasses"*, *"Smiling"*, *"No_Beard"* and *"Young"*). By using the aforementioned pseudo-labels, we performed the training/evluation process similarly to the case of the CelebA-HQ [7] dataset.

Additional qualitative results

In this section, we provide additional qualitative results of the proposed method in comparison to two state-of-the-art works, namely DeepPrivacy [3] and CIAGAN [8], in both the CelebA-HQ [7] and the LFW [2] datasets, in Figs. 1,2, respectively. We observe that, in both datasets, the proposed method arrives at anonymized versions of the real face images that preserve more effectively both a certain level of similarity with the real ones and certain attributes (such as the skin tone and overall texture, facial hair, etc). By contrast, the state-of-the-art works [3, 8] either lead to poor image quality (CIAGAN [8]) or/and fail to preserve certain facial attributes (DeepPrivacy [3]). This is also shown quantitatively in the Sect. 4 of the main paper. Finally, similarly to the previous section, we report as "Fake NN" the fake nearest neighbor (in the pre-trained FaRL [10] space) of each real image and observe that the proposed method provides an intuitive yet very simple way of initializing the latent codes that are then optimized in order to generate the anonymized face images.

Insights in the optimization process

In this section, we provide additional insight on the two stages of the proposed framework, i.e., the pairing of real images with fake ones and the latent code optimization (discussed in detail in Sect. 3.2 and Sect. 3.3 of the main paper, respectively). In Fig. 3 we show qualitative results of the proposed method for two values of the m hyperparameter (introduced in Sect. 3.3 in the main paper) that controls the dissimilarity between the real and the anonymized face images. Specifically, when $m \to 0$, the proposed identity loss (Eq. (2) in the main paper) imposes orthogonality between the features of the real and the anonymized face images, leading to anonymized faces with large identity difference compared to the corresponding real ones. By contrast, when $m \to 1$, the proposed identity loss imposes high similarity between the features of the real and the anonymized face images. Also, for each real image in Fig. 3, we report its corresponding fake nearest neighbor (obtained as described in detail in Sect. 3.2 in the main paper), denoted as "Fake NN". That is, the fakes image drawn from a pool of generated images that are closest to the real ones in the feature space of the pre-trained FaRL [10]. The latent codes of these fake neighbors are used for initializing the latent codes that are optimized for anonymizing the respective real images. As shown in Fig. 3, the fake nearest neighbor ("Fake NN") provides a meaningful starting point for the optimization of the anonymized latent code, but does not limit the final anonymized generation with respect to facial attributes,

[*] These authors contributed equally. This work has been conducted during a research exchange visit of S. Barattin in QMUL in the framework of the EU H2020 project AI4Media.

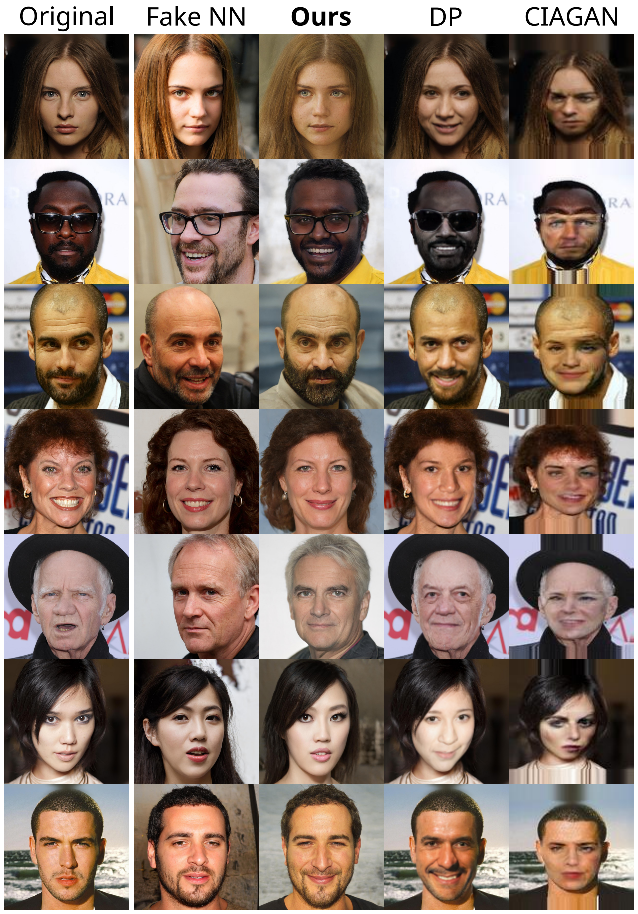

Figure 1. Anonymization results of the proposed method in comparison to DeepPrivacy [3] and CIAGAN [8] on the CelebA-HQ [7] dataset. "Fake NN" denotes the nearest fake neighbor of each real image, obtained in the pre-trained FaRL [10] image representation space.

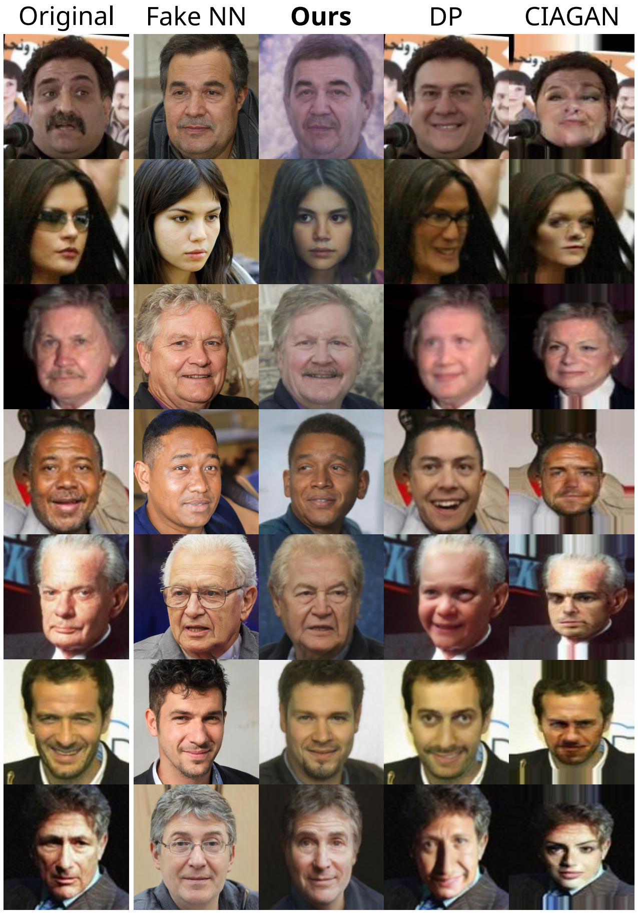

Figure 2. Anonymization results of the proposed method in comparison to DeepPrivacy [3] and CIAGAN [8] on the LFW [2] dataset. "Fake NN" denotes the nearest fake neighbor of each real image, obtained in the pre-trained FaRL [10] image representation space.

the skin tone, or the head pose. Finally, we observe that $m = 0.9$ leads to anonymized faces with higher identity similarity to the real ones compared to $m = 0.0$.

Processing time

As discussed in the main paper, the proposed framework incorporates only pre-trained networks (i.e., StyleGAN2's [5] generator $\mathcal{G}$, e4e [9], FaRL's [10] ViT-based image encoder $\mathcal{E}_{\mathcal{F}}$, and ArcFace [1] identity encoder $\mathcal{E}_{\mathcal{A}}$, as shown in Fig. 1 in the main paper), while at the same time the only trainable parameters are those of the latent codes that are optimized to anonymize the real images (i.e., 18×512 parameters per image). Learning each latent code requires ~ 3 sec/epoch in 1 Nvidia RTX 3090 (we train for 50 epochs), while generating an anonymized image from its optimized latent code requires a single forward pass of the optimized latent code through $\mathcal{G}$ (0.05 sec).

Limitations

As discussed in the main paper (Sect. 3), the proposed framework relies on a pre-trained StyleGAN2 [5] generator (typically pre-trained in the FFHQ [5] dataset) for generating the set of fake images (as described in Sect. 3.2 in the main paper), which are subsequently used for finding appropriate pairs (nearest fake neighbors in the FaRL [10] space) for each real image in order to initialize the latent codes ultimately optimized for the anonymization of the real images. This poses certain limitations to the proposed framework that reflect the limitations of the adopted GAN generator in generating faces statistically similar to the real ones, i.e., to the ones that will be anonymized. That is, the proposed method fails to anonymize real faces and to preserve all the relative attributes (e.g., hats) at the same time when the said attributes are not well-represented in the dataset that the adopted GAN generator has been trained with. Another

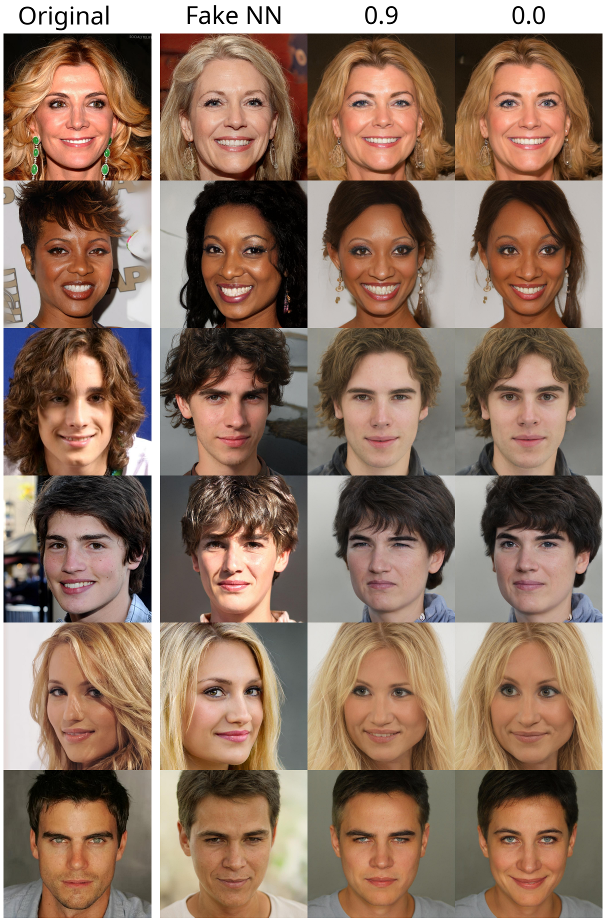

Figure 3. Anonymization results of the proposed method for $m \in \{0.0, 0.9\}$ on the CelebA-HQ [7] dataset. "Fake NN" denotes the nearest fake neighbor of each real image, obtained in the FaRL [10] image representation space.

limitation of the proposed framework concerns the inversion method that it incorporates (e.g., the e4e [9]), which might lead to unfaithful latent code inversions and thus affect the anonymization results.

References

[1] Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Cotsia, and Stefanos P Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2021. 2

[2] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1, 2

[3] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization, 2019. 1, 2

[4] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. Talk-to-edit: Fine-grained facial editing via dialog, 2021. 1

[5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8107–8116. IEEE, 2020. 2

[6] Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing, 2021. 1

[7] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1, 2, 3

[8] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020. 1, 2

[9] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation, 2021. 2, 3

[10] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner, 2021. 1, 2, 3