My Face My Choice: Privacy Enhancing Deepfakes for Social Media Anonymization Supplementary Material

Umur A. Çiftçi
Binghamton University
uciftci@binghamton.edu

Gokturk Yuksek
Binghamton University
gokturk@binghamton.com

İlke Demir
Intel Labs
ilke.demir@intel.com

A. Comparison

In Fig. 1, we perform experiments of CIAGAN [2] and DeepPrivacy [1] with MFMC using their sample images. We observe that faces produced by MFMC are much coherent in skin and gender attributes, preserve the expression better, and overall provide more realistic results. Remark that, MFMC has the goal of creating quantitatively dissimilar and realistic deepfakes, so it has more relaxed constraints on preserving the identity of the target and more strict constraints on the image and expression quality.

B. Additional Face Recognition Accuracies

We extend the exploration of how much MFMC can trick face recognition approaches if we use SSIM and RMSE similarity metrics for target face query in Tab. 1. Similar to the results using the face embedding similarity, MFMC can trick 71% on average for SSIM, and %77 for RMSE. Although incorrect recognition rate is higher, we use the embedding distance as grounded in the main paper. Moreover, the embedding space resembles the latent space learned by the face recognition model more than RMSE and SSIM spaces, thus better "tricking" is not surprising.

Face Detector	Source vs. Target		Source vs. Result	
	SSIM	RMSE	SSIM	RMSE
FaceNet512	0.001	0.0	0.12	0.16
OpenFace	0.001	0.003	0.17	0.24
FaceNet	0.03	0.02	0.27	0.34
DLib	0.02	0.05	0.30	0.44
ArcFace	0.05	0.03	0.29	0.45
DeepID	0.005	0.01	0.44	0.52
DeepFace	0.03	0.06	0.47	0.53
Average	0.02	0.01	0.29	0.23

Table 1. Seven SOTA face recognition approaches are compared on MFMC results based on face identification accuracy, where the furthest face is chosen in SSIM and RMSE metric spaces.

Switching from cosine to L_2 distance for embedding comparisons, Tab. 2 documents face recognition results where MFMC is able to reduce the accuracy to 49% on the average (last), and to 44% if we lift the randomness (third).

Face Detector	Source vs. Target		Source vs. Result	
	Furthest	Random	Furthest	Random
FaceNet512	0.08	0.05	0.34	0.43
OpenFace	0.005	0.009	0.31	0.34
FaceNet	0.08	0.06	0.40	0.43
DLib	0.05	0.10	0.50	0.64
ArcFace	0.02	0.03	0.28	0.35
DeepID	0.009	0.01	0.58	0.58
DeepFace	0.22	0.27	0.65	0.66
Average	0.07	0.07	0.44	0.49

Table 2. **Face Recognition Accuracies after MFMC.** Seven SOTA face recognition approaches are compared on MFMC results, using L_2 distance between face embeddings.

C. Threat Model

Our main threat model is automatic face recognition systems that associate faces with personally identifiable information or assign permanent face embeddings. CCTV cameras, millions of images on social media, and constantly evolving media sources capture all of us in some photos voluntarily or involuntarily. We would like to disable attackers to mine identity information from these photos, while enabling willing users to participate in the social platform.

For users who grant no access or for non-users, their identity never gets associated with the photo, no embedding is generated and the real face is disposed from the client right after being deepfaked at upload time – assuming there is no interruption at upload time. For users with other options, their face embeddings are shared only with friends in an encrypted way, and instances of their real faces are stored on the server. This requires trusting the social media

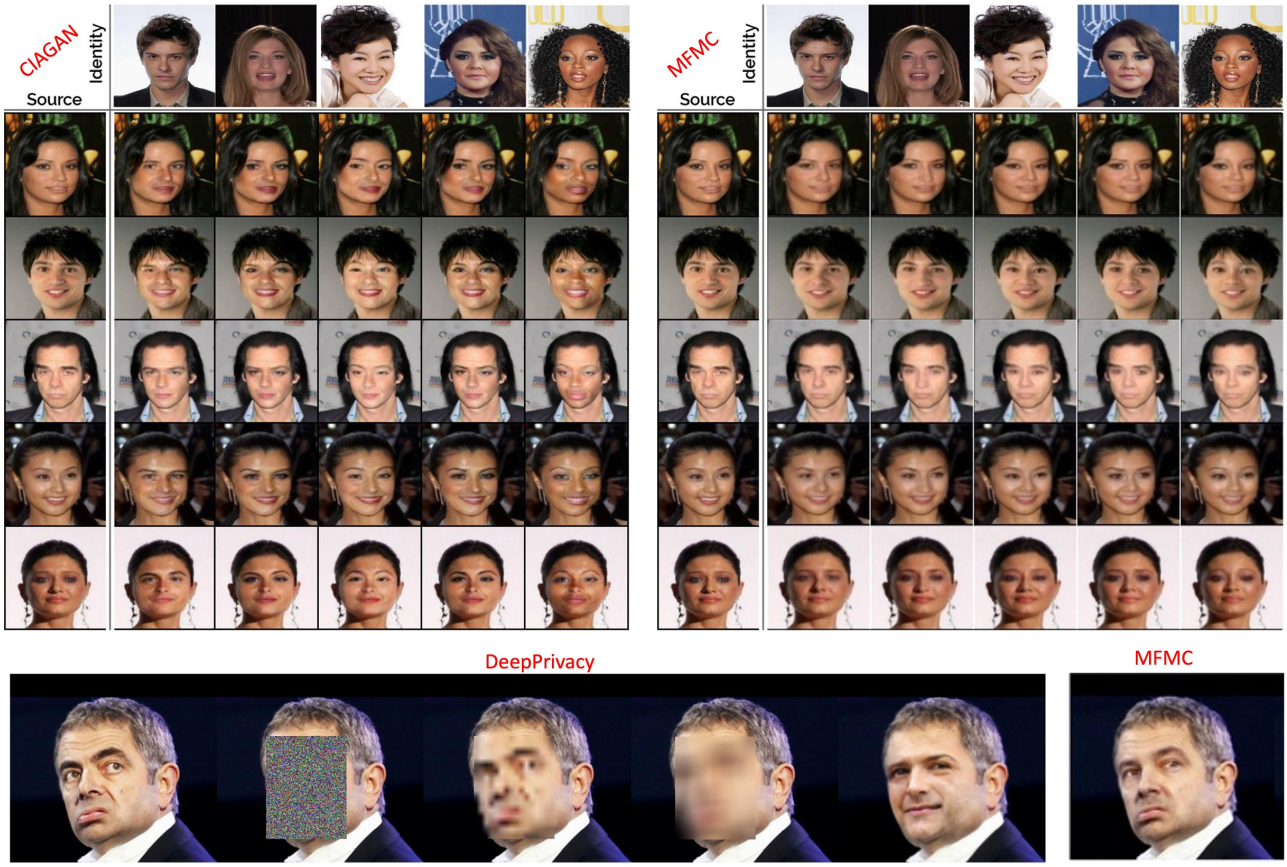

Figure 1. **Comparison.** We replicate the results of CIAGAN [2] and DeepPrivacy [1] (left) using MFMC. Artifacts from skin color, gender difference, and other subtle differences are not observed in MFMC results.

platform for handling the process privately without human intervention, for having a secure client-server transmission protocol, and for not leaking the photos or embeddings.

D. Privacy Evaluation

In contrast to anonymization methods which aggregates data points into groups that disable inferring individual information, our approach masks each face with a deepfake per photo. These deepfakes should not even be considered as quasi-identifiers, as they no longer preserve the identity. Having access to $k-1$ deepfake versions of the same face does not enable reconstructing the original face, even if the original photo is in that set (without being known as the original), which satisfies *k-anonymity*. The age and gender groups to create deepfakes are synthetic calculations that we do not seek the exact values for, they are approximate ranges to preserve the photorealism.

On the other hand we are vulnerable to linkability attacks if the same image is posted on a platform without anonymization, or if there is personally identifiable data in the image in another form than faces.

References

[1] Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. Deepprivacy: A generative adversarial network for face anonymization. In *International symposium on visual computing*, pages 565–578. Springer, 2019.

[2] 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.