Supplementary Material: Privacy-preserving Adversarial Facial Features

Abstract

In this supplementary file, more qualitative and quantitative comparisons are provided to demonstrate the effectiveness of the proposed AdvFace.Following the experimental evaluation in the main submission, more corresponding examples in defense against privacy attacks, and transferability of AdvFace are visualized, respectively. Meanwhile, quantitative values are provided to further demonstrate the outstanding trade-off of our method between defending against reconstruction attacks and maintaining face recognition accuracy.

A. Defense against Privacy Attacks

Figs. 1 2 3 show more reconstructed images from facial features protected by different methods on datasets LFW [1], CFP-FP [3], and AgeDB-30 [2] , respectively. As shown in the third column, the reconstructed images from the adversarial features generated by the proposed AdvFace are hard to distinguish, while those protected by other methods (columns 4-6) undergo much information leakage about the original images.

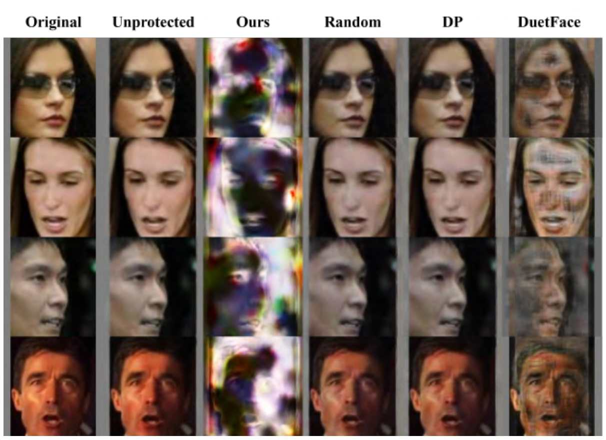

Figure 1. Reconstructed images from facial features generated by different privacy protection methods on dataset LFW.

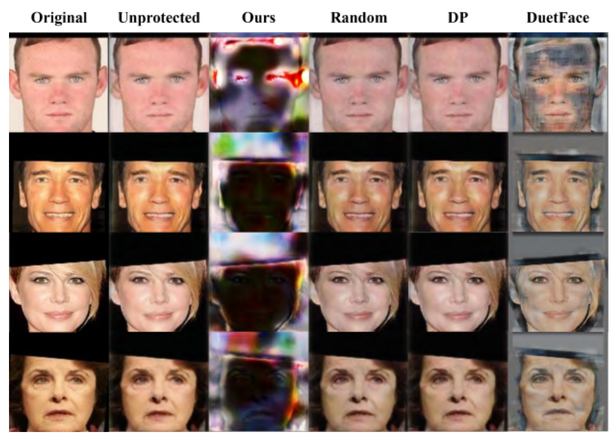

Figure 2. Reconstructed images from facial features generated by different privacy protection methods on dataset CFP-FP.

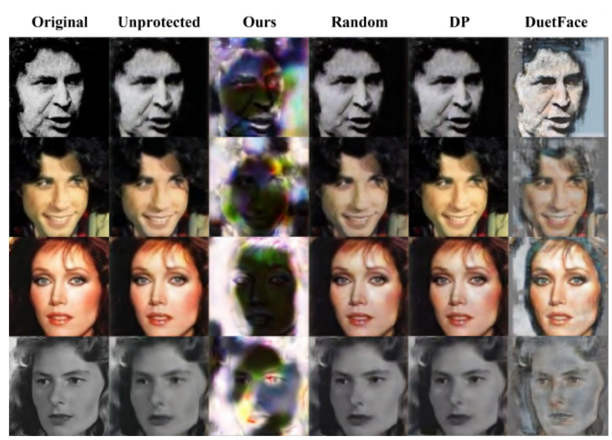

Figure 3. Reconstructed images from facial features generated by different privacy protection methods on dataset AgeDB-30.

Table 1. Quantitative values of trade-off among SSIM, PSNR, MSE, and ACC for AdvFace with different noise bounds.

	LFW				CFP-FP				AgeDB-30			
ϵ	SSIM↓	PSNR↓	MSE↑	ACC↑	SSIM↓	PSNR↓	MSE↑	ACC↑	SSIM↓	PSNR↓	MSE↑	ACC↑
0.00	0.90	26.33	0.002	97.80%	0.77	21.76	0.008	92.10%	0.83	22.56	0.006	86.78%
0.05	0.70	13.63	0.045	97.63%	0.59	14.47	0.038	91.90%	0.65	12.92	0.053	86.87%
0.10	0.50	10.30	0.096	97.47%	0.41	10.49	0.093	91.59%	0.44	9.13	0.127	86.22%
0.15	0.38	8.57	0.143	97.12%	0.31	7.88	0.168	91.24%	0.33	7.28	0.193	85.85%
0.20	0.28	6.98	0.206	96.43%	0.23	5.96	0.262	90.71%	0.24	5.85	0.269	85.10%
0.25	0.24	6.16	0.249	95.57%	0.19	4.97	0.328	89.81%	0.22	5.33	0.305	84.35%
0.30	0.22	5.71	0.275	93.55%	0.16	4.39	0.375	87.82%	0.20	4.91	0.334	82.42%

Table 2. The architecture of reconstruction networks.

TransRec	$77^2 \times 64 \xrightarrow{transconv3-64} 79^2 \times 64 \xrightarrow{transconv3-32} 79^2 \times 32 \xrightarrow{upsample}$ $160^2 \times 32 \xrightarrow{transconv3-32} 160^2 \times 32 \xrightarrow{transconv3-3} 160^2 \times 3 \xrightarrow{Sigmoid}$ $160^2 \times 3$
ResRec	$77^2 \times 64 \xrightarrow{transconv3-64} 77^2 \times 64 \xrightarrow{IRBlock(64,2)} 77^2 \times 64 \xrightarrow{upsample}$ $120^2 \times 64 \xrightarrow{IRBlock(64,2)} 120^2 \times 64 \xrightarrow{upsample} 160^2 \times 64 \xrightarrow{conv1-3}$ $160^2 \times 3 \xrightarrow{Sigmoid} 160^2 \times 3$
URec	$77^2 \times 64 \xrightarrow{conv3-64} 77^2 \times 64 \xrightarrow{conv3-64} 77^2 \times 64 \xrightarrow{conv3-64}$ $77^2 \times 64 \xrightarrow{upsample} 120^2 \times 64 \xrightarrow{conv3-128} 120^2 \times 128 \xrightarrow{conv3-128}$ $120^2 \times 128 \xrightarrow{conv3-128} 120^2 \times 128 \xrightarrow{upsample} 160^2 \times 128 \xrightarrow{conv3-256}$ $160^2 \times 256 \xrightarrow{conv3-256} 160^2 \times 256 \xrightarrow{conv3-256} 160^2 \times 256 \xrightarrow{conv3-3}$ $160^2 \times 3 \xrightarrow{conv1-3} 160^2 \times 3$

B. Transferability of AdvFace

As shown in the Table 2, we build three types of reconstruction networks that can be employed by the attacker to verify the Transferability of the method. In Figs. 4 5 6, we show the facial images reconstructed from the adversarial features under three different shadow models by three different reconstruction networks. We can see that the defense effectiveness of AdvFace is maintained under different shadow models when encountering different attack networks, which validates the transferability of the adversarial features generated by AdvFace.

C. Details of Trade-off

We further show the quantitative values of trade-off in Tab. 1. We can see that when ϵ increases from 0.00 to 0.20, the accuracy drops slightly, but the ability to against reconstruction attacks improves rapidly. Moreover, the accuracy drops faster from 0.25 to 0.30, while the ability to against reconstruction attacks is further improved. All of these results that AdvFace could provide a good trade-off between defending against reconstruction attacks and maintaining face recognition accuracy. Finally, we choose ϵ to be 0.20 in the experiments.

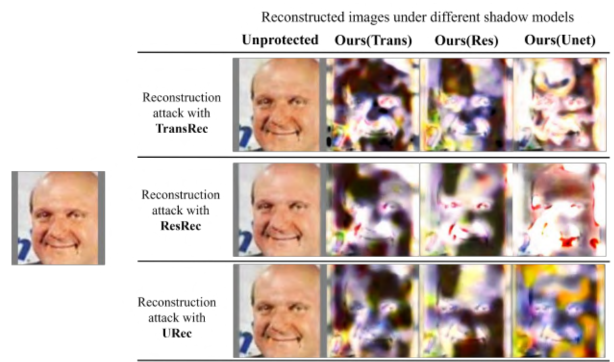

Figure 4. Transferability of AdvFace on defending against reconstruction attacks on dataset LFW.

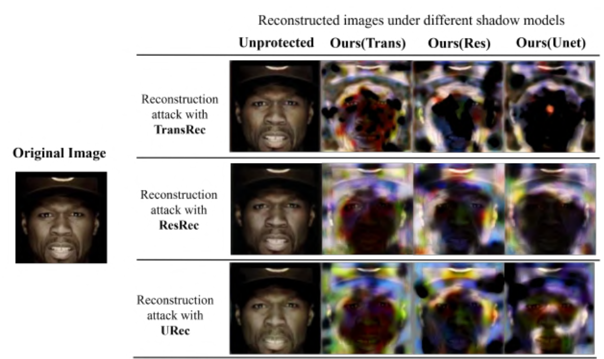

Figure 5. Transferability of AdvFace on defending against reconstruction attacks on dataset CFP-FP.

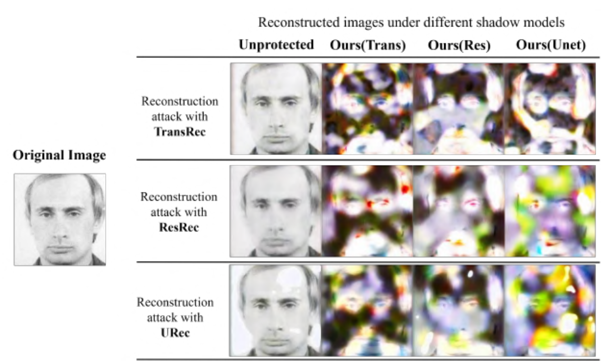

Figure 6. Transferability of AdvFace on defending against reconstruction attacks on dataset AgeDB-30.

References

[1] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller. Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008. 1

[2] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou. Agedb: the first manually collected, in-the-wild age database. In *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 51–59, 2017. 1

[3] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs. Frontal to profile face verification in the wild. In *2016 IEEE winter conference on applications of computer vision (WACV)*, pages 1–9. IEEE, 2016. 1