Appendix for: Towards Face Encryption by Generating Adversarial Identity Masks

Xiao Yang[1], Yinpeng Dong[1,3], Tianyu Pang[1], Hang Su[1,2]*, Jun Zhu[1,2,3], Yuefeng Chen[4], Hui Xue[4]
[1] Dept. of Comp. Sci. and Tech., Institute for AI, BNRist Center, Tsinghua-Bosch Joint ML Center
[1] THBI Lab, Tsinghua University, Beijing, 100084, China
[2] Pazhou Lab, Guangzhou, 510330, China [3] RealAI [4] Alibaba Group
{yangxiao19, dyp17, pty17}@mails.tsinghua.edu.cn
{suhangss, dcszj}@mail.tsinghua.edu.cn {yuefeng.chenyf, hui.xueh}@alibaba-inc.com

A. Evaluation Results on MegFace.

We report the results on the MegFace dataset in Tab. 2. Compared with LFW and MegFace has more gallery images over 50k+ images. This large-scale challenging dataset results in more difficult targeted identity protection on the whole.

B. Batch Analysis on MMD Optimization.

Tab. 1 shows results on different batch sizes w.r.t naturalness. It can be seen that the evaluation of visual quality becomes stable as the batch size exceeds 50. We set batch size as 50 in this paper. For single image crafting, we have two choices. The first one is self augmentation including rotation, projective, brightness and transformations; The second one is collecting some irrelevant images to form a batch just for optimal results in the phase of MMD optimization.

	10	20	30	40	50	60	70	80	90
SSIM	0.8518	0.8392	0.7592	0.7021	0.6759	0.6765	0.6633	0.6649	0.6674
PSNR	28.71	28.55	26.95	26.02	25.55	25.63	25.40	25.50	25.54

Table 1. The mean PSNR (db) and SSIM for different batch sizes based on CosFace.

C. Comparison Experiments about Target Images.

Different numbers of targets. As illustrated in Fig. 1, we study the effect of different numbers on the black-box identity protection. The curve first rises and finally approaches the steady. Therefore, appropriate increases in the number of targets is beneficial to performance improvement against black-box models.

Generated images as targets. To further verify that our algorithm does not depend on the selection of targets, we specify some generated images from StyleGAN [1] as target images, which is illustrated as Fig. 2. We use these generated images as target images and set the same other setting with above experiments. Tab. 3 shows Rank-1-T, Rand-1-UT, Rank-5-T and Rank-5-UT of black-box attacks against CosFace, SphereFace, FaceNet, ArcFace, MobileFace and ResNet. The results show that our algorithm still has excellent black-box performance of identity protection. In practical applications, we can arbitrarily specify the available and authorised target identity set or some generated facial images, and our algorithm is applicable to any target set.

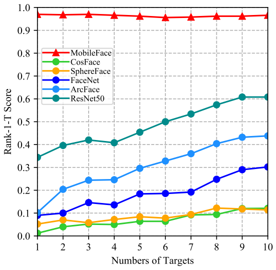

Figure 1. The perturbation vs. numbers of targets curve of face identification models against black-box identity protection. *MobileFace* is a surrogate white-box model.

Figure 2. Examples of some generated images from StyleGAN.

[*] Corresponding author

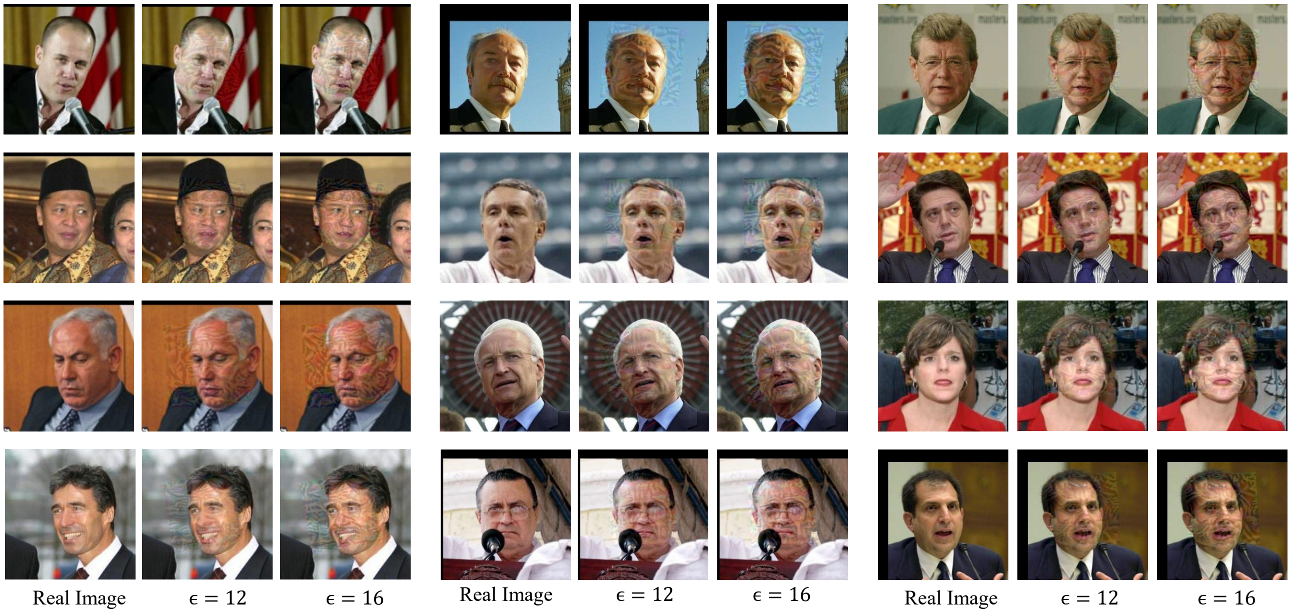

Figure 3. More examples for different perturbations under the l_∞ norm by existing adversarial methods.

	Attack	ArcFace		MobileFace		ResNet50		SphereFace		FaceNet		CosFace	
		R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T	R1-T	R5-T
ArcFace	DIM [3]	92.0^*	97.0^*	11.6	36.0	8.0	23.4	1.6	7.6	3.4	13.4	1.4	7.6
	TIP-IM	97.2^*	98.4^*	60.1	81.4	51.4	67.4	10.1	19.5	25.3	43.4	11.1	24.2
MobileFace	DIM [3]	5.8	19.4	94.8^*	96.4^*	18.6	42.0	2.8	9.6	4.0	14.2	2.0	9.2
	TIP-IM	37.4	57.2	97.0^*	97.2^*	52.9	73.5	9.8	19.8	20.9	36.2	10.5	22.1
ResNet50	DIM [3]	7.8	19.0	15.2	44.2	91.4^*	95.4^*	2.4	13.0	4.8	16.8	3.2	9.2
	TIP-IM	31.1	45.2	56.9	76.3	92.1^*	97.5^*	11.6	21.2	20.5	35.2	10.0	25.2

Table 2. Rank-1-T and Rank-5-T (%) of black-box attacks against CosFace, SphereFace, FaceNet, ArcFace, MobileFace and ResNet on MegFace. * indicates white-box attacks.

	ArcFace	MobileFace	ResNet50	SphereFace	FaceNet	CosFace
Rank-1-T	12.2	32.4	29.2	27.4	28.2	49.6
Rank-5-T	30.4	52.2	54.8	56.4	53.8	70.0
Rank-1-UT	89.0	73.8	49.6	60.8	54.2	95.0
Rank-5-UT	82.0	55.8	31.6	41.8	34.8	93.6

Table 3. Results of black-box attacks against SphereFace, FaceNet, ArcFace, MobileFace, ResNet and CosFace when treating the *generated* images as the target images.

D. Ill-suited ℓ_p-norm perturbation in Face encryption

Face encryption focuses on generating adversarial identity masks that can be overlaid on facial images without sacrificing the visual quality. As illustrated in Fig. 3, although the adversarial perturbations generated by the existing attack methods, *e.g.*, PGD and MIM, have a small intensity change (e.g., 12 or 16 for each pixel in $[0, 255]$), they may still sacrifice the visual quality for human perception due to the artifacts. ℓ_p-norm adversarial perturbations can not naturally fit human perception well, which also accords with [4, 2]. Thus proposed TIP-IM introduces a better multi-target optimization mechanism to improve effectiveness and $\mathcal{L}_{nat}$ in the objective of Eq. (2) to generate more natural images.

References

[1] 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*, pages 8110–8119, 2020. 1

[2] Ayon Sen, Xiaojin Zhu, Liam Marshall, and Robert Nowak. Should adversarial attacks use pixel p-norm? *arXiv preprint arXiv:1906.02439*, 2019. 2

[3] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[4] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2018. 2