Supplementary Material: Geometrically Adaptive Dictionary Attack on Face Recognition

Junyoung Byun, Hyojun Go, Changick Kim
Korea Advanced Institute of Science and Technology (KAIST)
{bjyoung, gohyojun15, changick}@kaist.ac.kr

1. Overview

In this supplementary material, we describe hyperparameter settings for attack methods used in our experiments. We also illustrate the perturbation norm curves that visually show each method's query efficiency and additional qualitative results for comprehensive comparisons. We also list the experimental results of the decision-based black-box attacks against a deeper target model, CurricularFace ResNet-100 [8, 6].

2. Additional details of experimental setting

We use Pytorch framework [10] for our experiments and borrow the code for face recognition from face.evoLVe library[1]. We use the 112×112 aligned datasets provided by face.evoLVe library for the LFW [7] and CPLFW [11] datasets.

3. Implementation and hyperparameter settings of the attacks

Sign-OPT (SO) [3]. We adopt the code[2] of Sign-OPT provided by the authors without special tuning the hyperparameters of the attack.

HSJA [1]. We implement HSJA by using Adversarial Robustness Toolbox (ART) library [9]. From its default setting, we increase the maximum iterations to 64 to follow the authors' experimental settings[3].

EA [5]. We implement EA based on the code[4] of EA provided by the authors. We set the dimension of the perturbation search space as $60 \times 60 \times 3$ and the coefficient of the distance in calculation of σ as 0.03 instead of 0.01 for faster convergence. These settings also apply to the EA's variants (EAD, EAG, EAGD).

SFA [2]. We adopt the code[5] of SFA provided by the authors without special tuning of the hyperparameters. We set the dimension reduction ratio as 2 for reducing the perturbation search space. These settings also apply to the SFA's variants (SFAD, SFAG, SFAGD).

4. More experimental results on the ArcFace ResNet-50 model

We illustrate the perturbation norm curves that visually show the query efficiency of each attack method in Fig. 1 and Fig. 2. We show additional examples for more extensive qualitative comparison in Fig. 3 to Fig. 17. In detail, Fig. 3 and Fig. 4 show the results of dodging attacks on the LFW dataset. Fig. 5 through Fig. 8 display the examples of impersonation attacks on the LFW dataset. Fig. 9 to Fig. 12 show the results of dodging attacks on the CPLFW dataset. Fig. 13 to Fig. 17 show the examples of impersonation attacks on the CPLFW dataset.

5. Experimental results on the CurricularFace ResNet-100 model

We conduct the decision-based black-box attacks against the CurricularFace ResNet-100[6] [8, 6] which is trained on the refined MS1MV2 dataset [4]. We arrange the experimental results in Table 1. Clearly, GADA greatly improves the query efficiency of EA.

As the model is deeper and more accurate, the robustness to attack increases, so the minimum perturbation norm is higher in all datasets than that of the results of the ArcFace ResNet-50 model [4, 6]. We illustrate the perturbation norm curves that visually show each attack method's query efficiency in Fig. 18 and Fig. 19.

[1] https://github.com/ZhaoJ9014/face.evoLVe.PyTorch
[2] https://github.com/cmhcbb/attackbox
[3] https://github.com/Jianbo-Lab/HSJA
[4] https://github.com/thu-ml/realsafe/blob/master/realsafe/attack/evolutionary_worker.py
[5] https://github.com/wubaoyuan/Sign-Flip-Attack
[6] We use the pretrained model from https://github.com/HuangYG123/CurricularFace

Dodging attacks												
	LFW dataset [7]						CPLFW dataset [11]					
	Minimum perturbation norm with query budget				Avg. # queries for perturbation with norm		Minimum perturbation norm with query budget				Avg. # queries for perturbation with norm	
Attack method	1K	2K	5K	10K	4	2	1K	2K	5K	10K	4	2
Sign-OPT [3]	19.50	11.38	4.17	2.37	5284	8678	21.14	11.95	4.24	2.39	5206	8256
HSJA [1]	10.96	6.96	3.35	2.09	3990	7796	12.12	7.55	3.49	2.15	4090	7362
EA [5]	14.16	7.23	2.78	1.52	3550	6802	14.57	7.33	2.79	1.54	3486	6419
EAD	11.16	6.15	2.60	1.48	3195	6538	11.92	6.40	2.66	1.53	3236	6255
EAG	8.68	4.65	2.08	**1.39**	2424	5494	8.86	4.67	**2.06**	**1.38**	2407	5101
EAGD	**7.06**	**4.07**	**2.02**	1.42	**2104**	**5166**	**7.90**	**4.40**	2.11	1.42	**2251**	**4954**
Impersonation attacks												
	LFW dataset [7]						CPLFW dataset [11]					
	Minimum perturbation norm with query budget				Avg. # queries for perturbation with norm		Minimum perturbation norm with query budget				Avg. # queries for perturbation with norm	
Attack method	1K	2K	5K	10K	4	2	1K	2K	5K	10K	4	2
Sign-OPT [3]	22.74	17.33	8.73	4.53	7944	9561	19.22	12.43	4.79	2.17	4876	6827
HSJA [1]	21.08	14.49	6.46	3.58	6709	9102	15.26	9.12	3.39	1.77	3718	5799
EA [5]	17.41	10.29	4.24	2.22	5226	8169	12.74	6.32	2.15	1.05	2931	4803
EAG	**13.09**	**7.63**	**3.19**	**1.84**	**3887**	**7217**	**8.71**	**4.21**	**1.51**	**0.86**	**1999**	**3546**

Table 1: Evaluation of decision-based black-box attacks against the CurricularFace ResNet-100 [8, 6] with the two datasets.

References

[1] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. In *2020 IEEE Symposium on Security and Privacy (SP)*, pages 1277–1294. IEEE, 2020.

[2] Weilun Chen, Zhaoxiang Zhang, Xiaolin Hu, and Baoyuan Wu. Boosting decision-based black-box adversarial attacks with random sign flip. In *Proceedings of the European Conference on Computer Vision*, 2020.

[3] Minhao Cheng, Simranjit Singh, Patrick H. Chen, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. Sign-opt: A query-efficient hard-label adversarial attack. In *International Conference on Learning Representations*, 2020.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

[5] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[7] 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.

[8] Yuge Huang, Yuhan Wang, Ying Tai, Xiaoming Liu, Pengcheng Shen, Shaoxin Li, Jilin Li, and Feiyue Huang. Curricularface: adaptive curriculum learning loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5901–5910, 2020.

[9] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[11] T. Zheng and W. Deng. Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments. Technical Report 18-01, Beijing University of Posts and Telecommunications, February 2018.

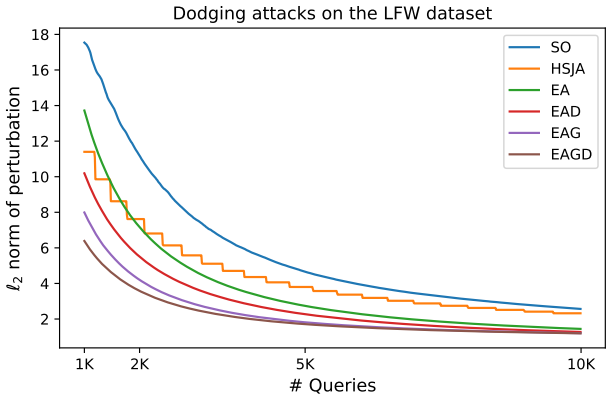


Figure 1: Perturbation norm curves of the decision-based attacks against ArcFace ResNet-50 model with the LFW dataset [7].

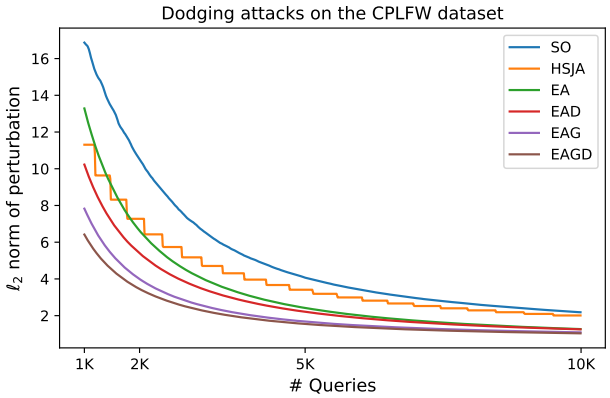
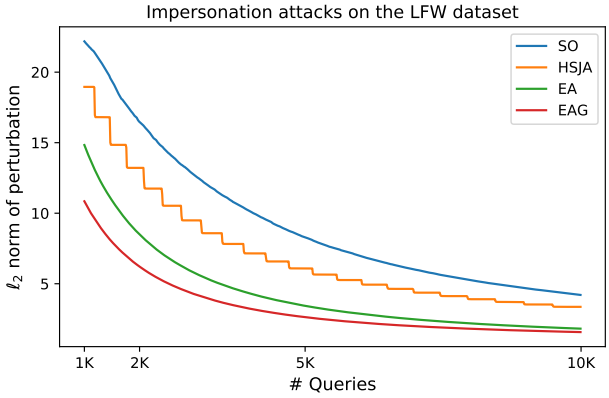
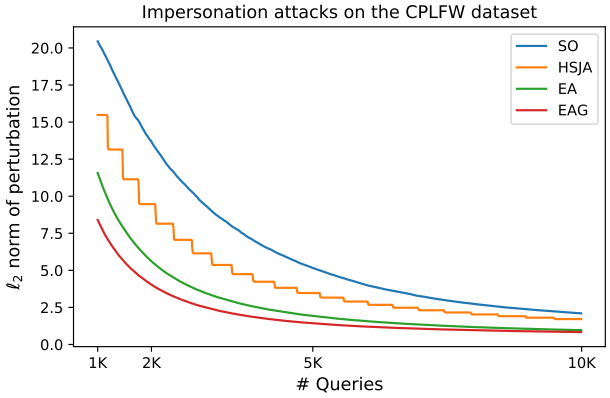


Figure 2: Perturbation norm curves of the decision-based attacks against ArcFace ResNet-50 model with the CPLFW dataset [11].





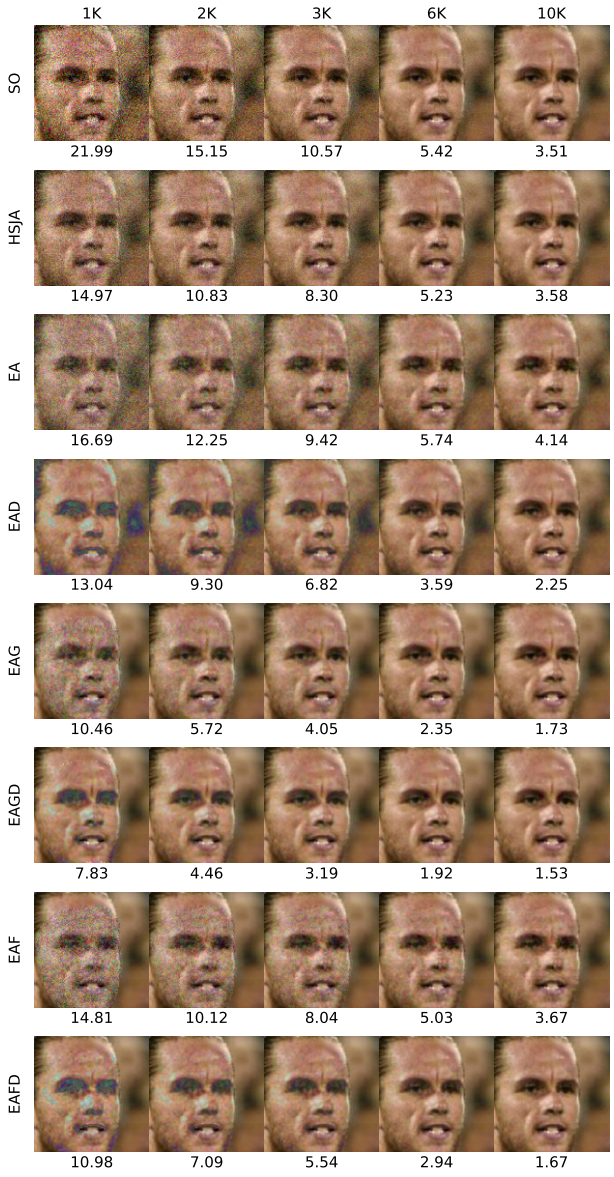


Figure 3: Qualitative results of dodging attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

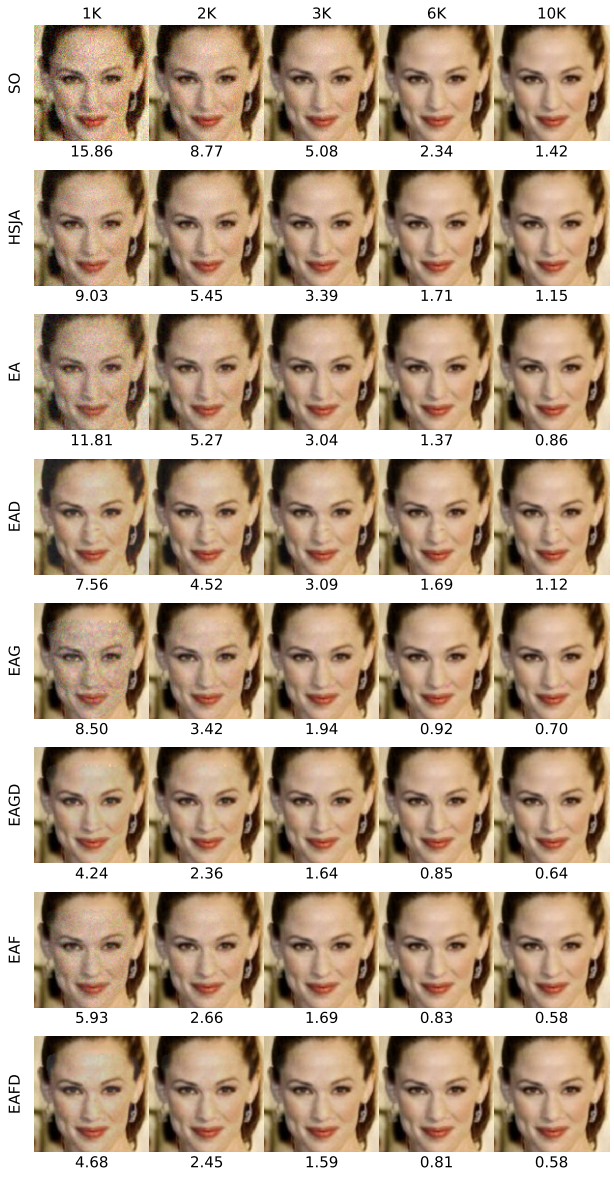


Figure 4: Qualitative results of dodging attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

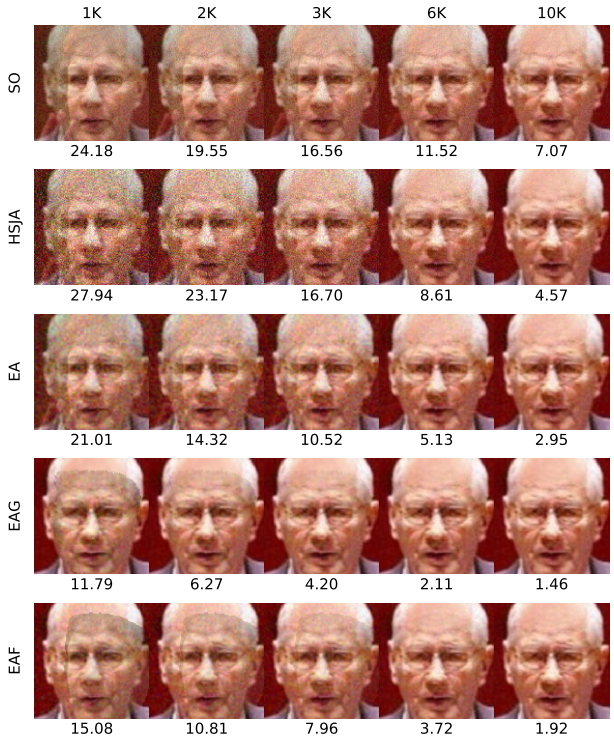


Figure 5: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

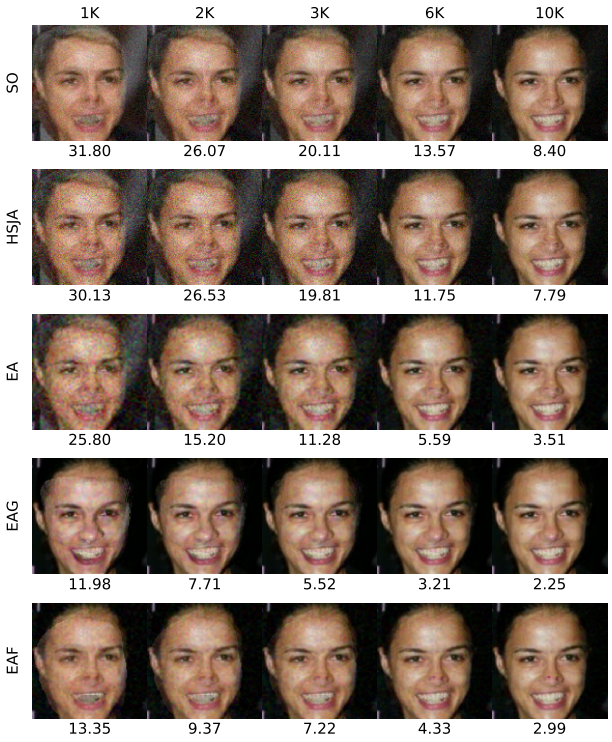


Figure 6: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

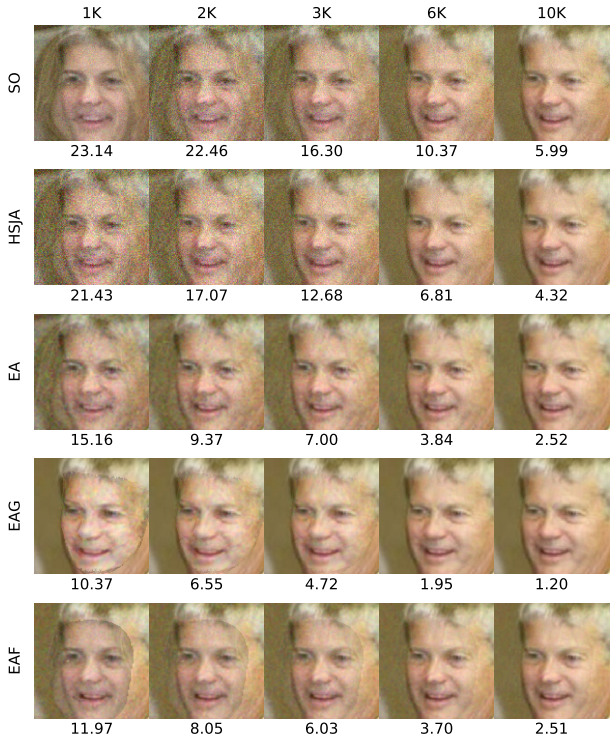


Figure 7: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

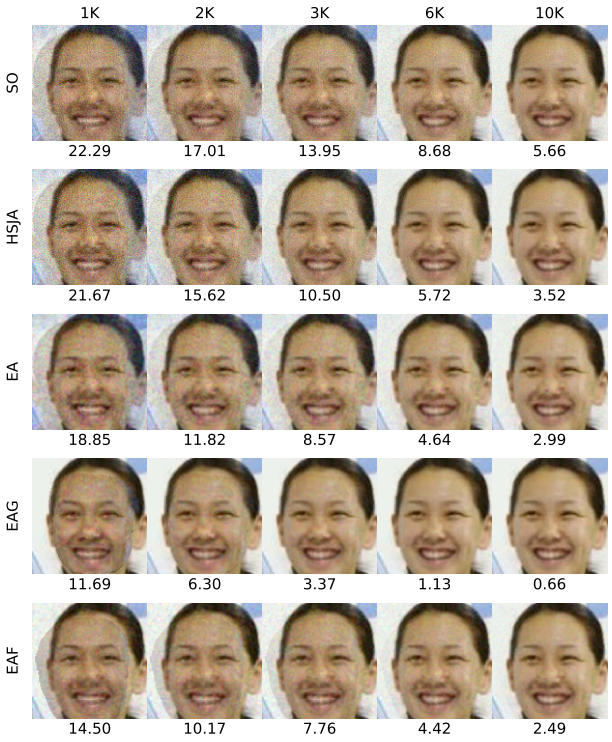


Figure 8: Qualitative results of impersonation attacks on the LFW dataset [7]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

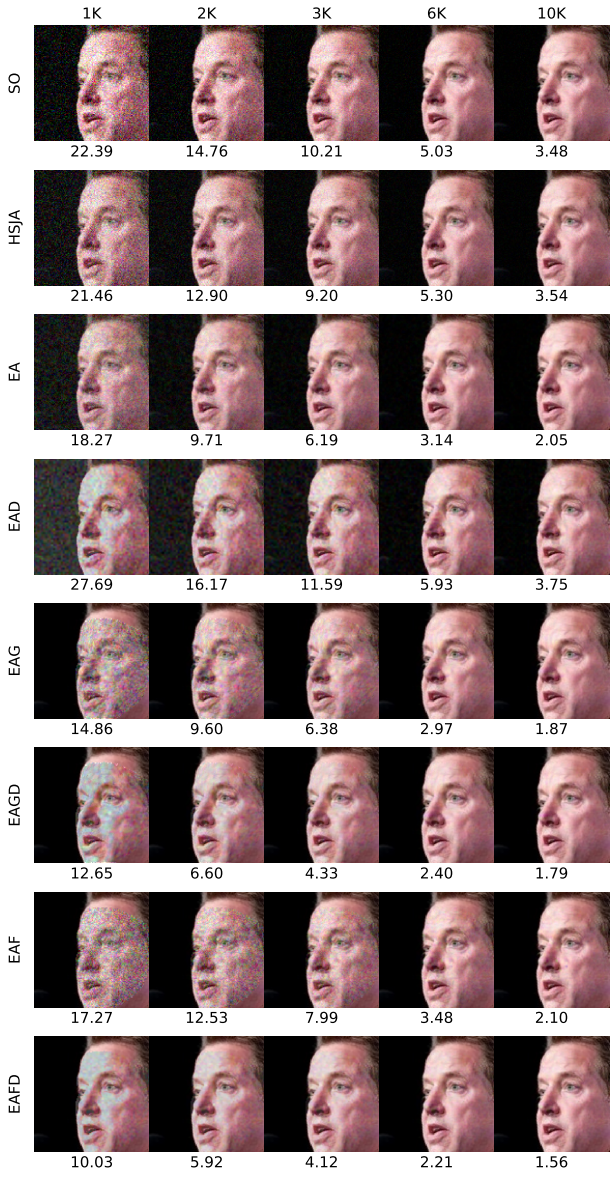


Figure 9: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

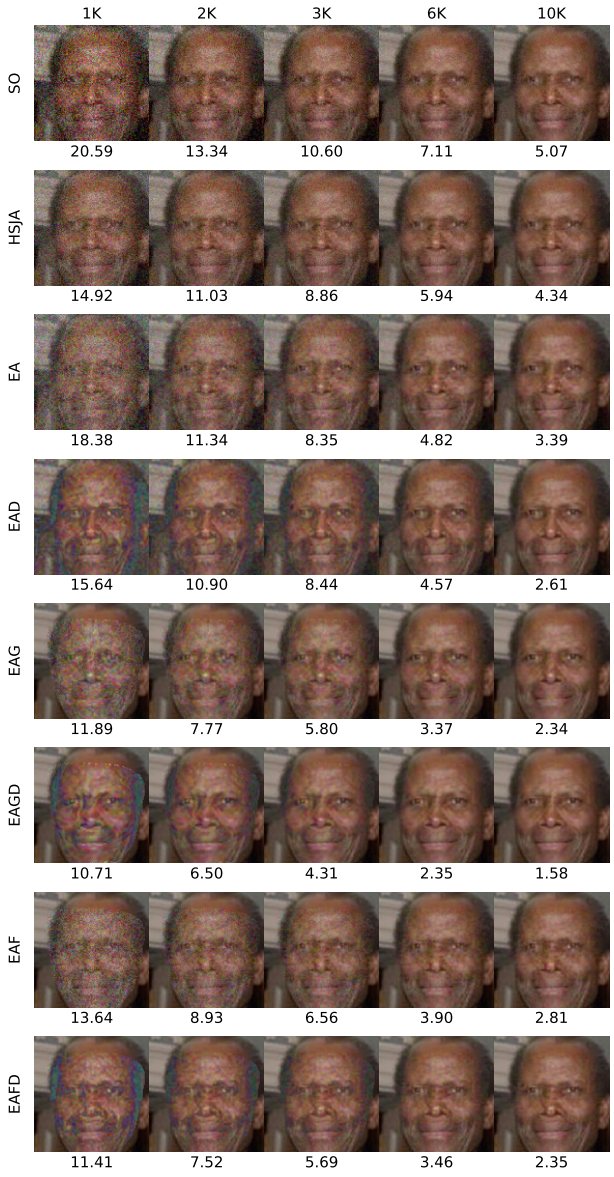


Figure 10: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

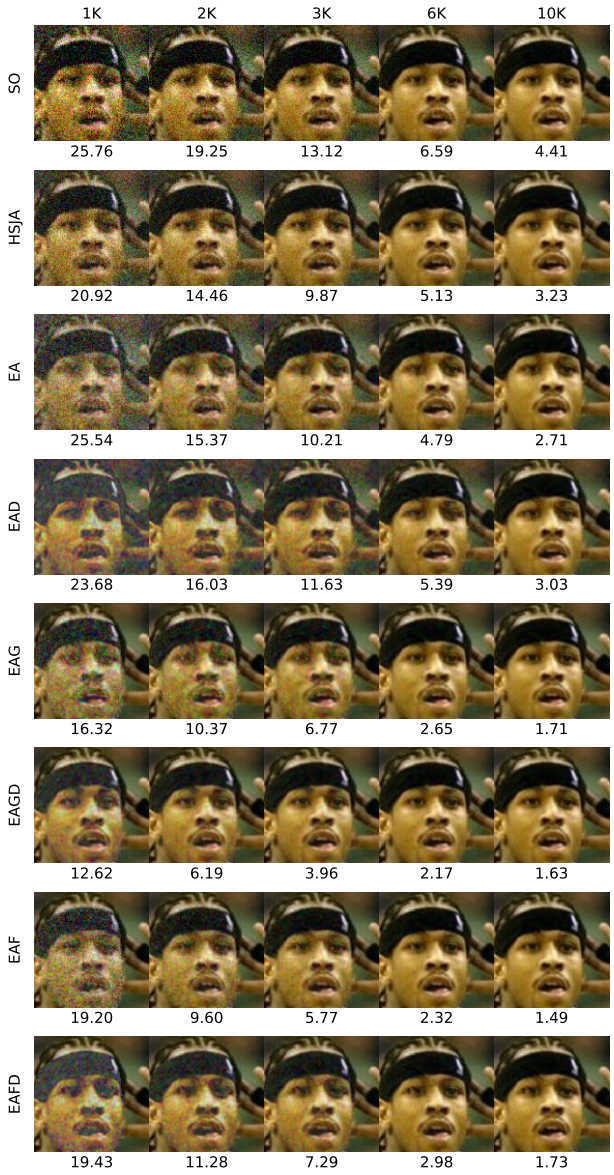


Figure 11: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

Figure 12: Qualitative results of dodging attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

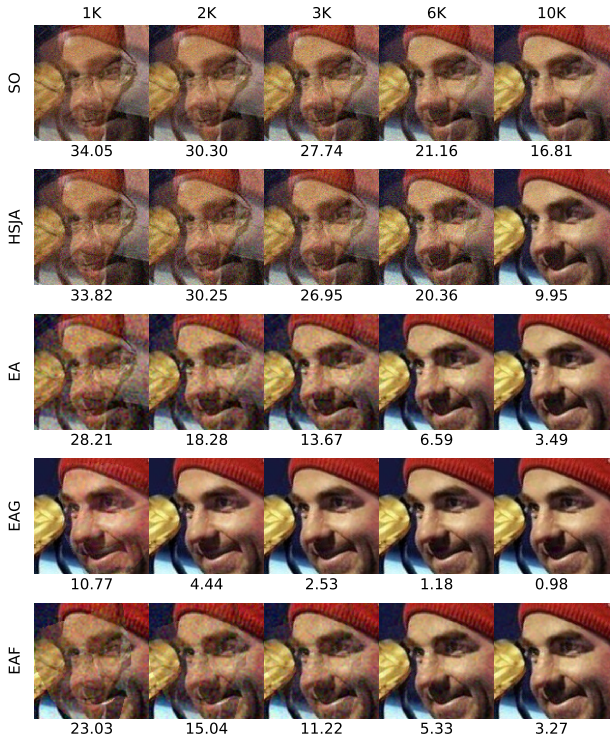


Figure 13: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

Figure 14: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

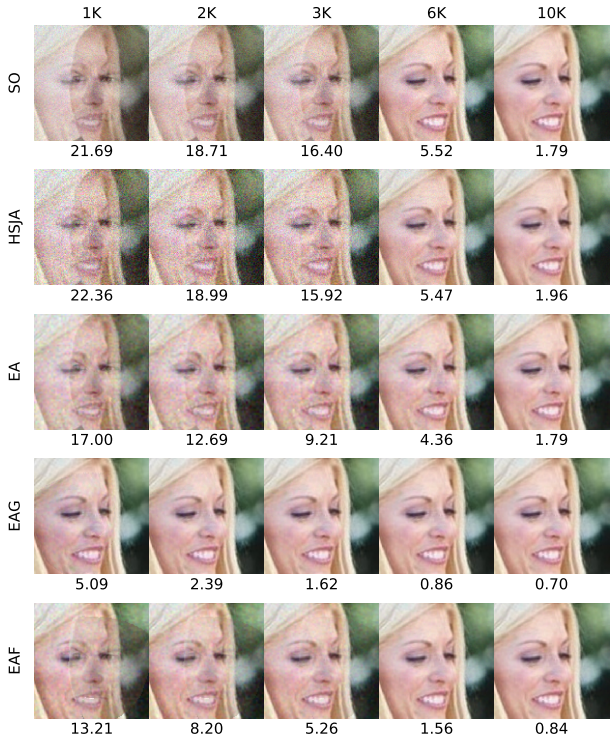


Figure 15: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

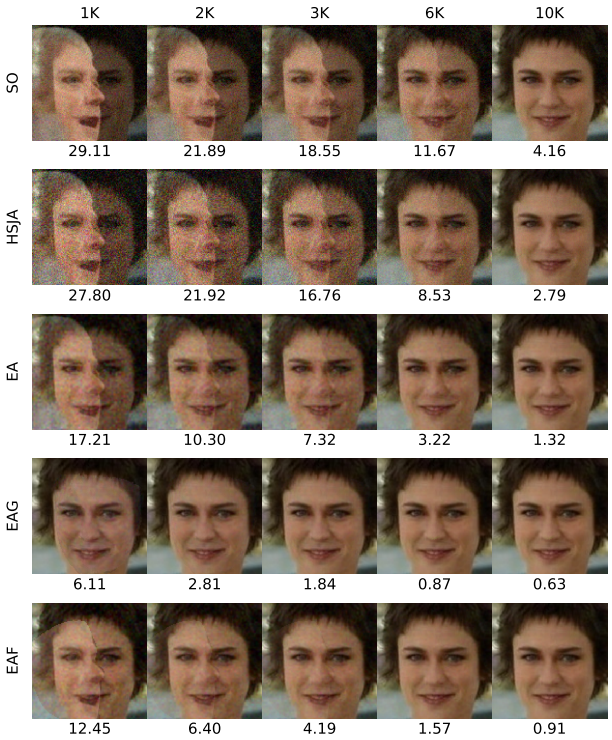


Figure 16: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

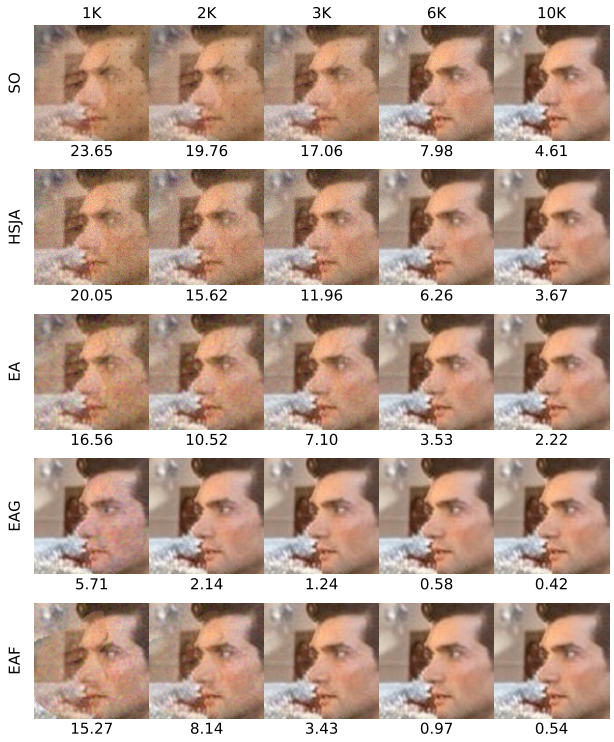


Figure 17: Qualitative results of impersonation attacks on the CPLFW dataset [11]. For each attack, we illustrate the minimum norm-adversarial examples in each query budget. The ℓ_2 norm of perturbation is displayed under each image.

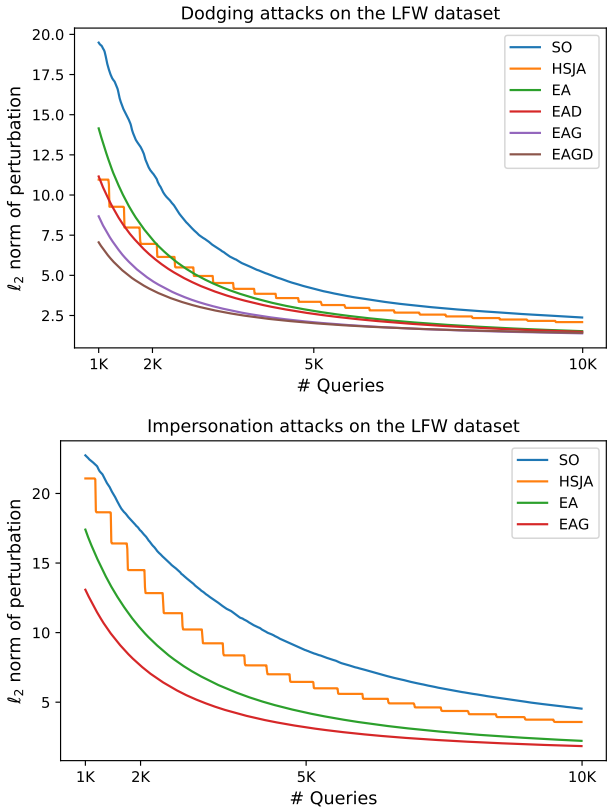


Figure 18: Perturbation norm curves of the decision-based attacks against the CurricularFace ResNet-100 with the LFW dataset [7].

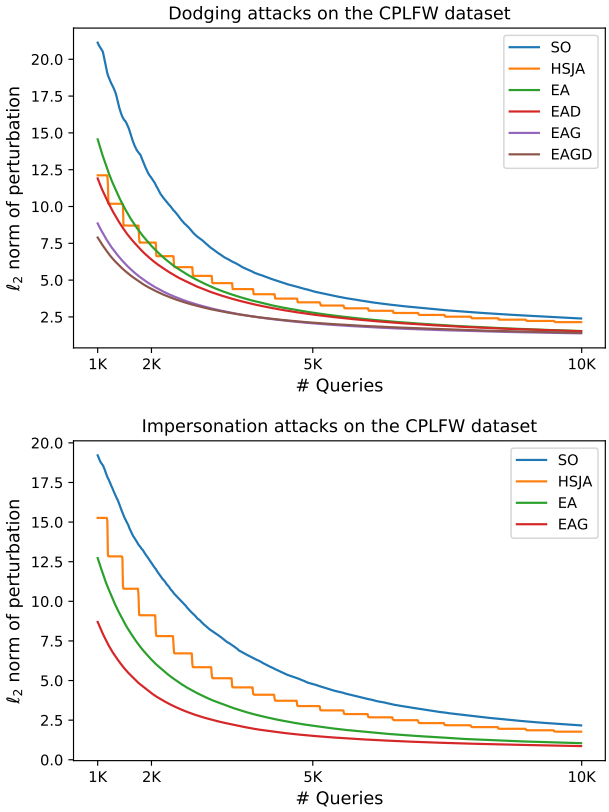


Figure 19: Perturbation norm curves of the decision-based attacks against the CurricularFace ResNet-100 with the CPLFW dataset [11].