Supplementary Material for:
Efficient Decision-based Black-box Adversarial Attacks on Face Recognition

Yinpeng Dong[1], Hang Su[1], Baoyuan Wu[2], Zhifeng Li[2], Wei Liu[2], Tong Zhang[3], Jun Zhu[1*]
[1] Dept. of Comp. Sci. and Tech., BNRist Center, State Key Lab for Intell. Tech. & Sys.,
[1] Institute for AI, THBI Lab, Tsinghua University, Beijing, 100084, China
[2] Tencent AI Lab [3] Hong Kong University of Science and Technology
dyp17@mails.tsinghua.edu.cn, suhangss@mail.tsinghua.edu.cn, wubaoyuan1987@gmail.com
michaelzfli@tencent.com, wl2223@columbia.edu, tongzhang@tongzhang-ml.org, dcszj@mail.tsinghua.edu.cn

A. Proof

Theorem 1. *Assume that the covariance matrix* $\mathbf{C}$ *is positive definite. Let* λ_{max} *and* $\lambda_{min}(>0)$ *be the largest and smallest eigenvalues of* $\mathbf{C}$*, respectively. Then, we have*

$$P_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})}\big(\mathcal{L}(\tilde{\boldsymbol{x}}^*+\boldsymbol{z})<\mathcal{L}(\tilde{\boldsymbol{x}}^*)\big)\leq\frac{4\lambda_{max}\|\tilde{\boldsymbol{x}}^*-\boldsymbol{x}\|^2}{\sigma^2\lambda_{min}^2n^2}.$$

Proof. Assume that the eigenvalues of the covariance matrix $\mathbf{C}$ are $\lambda_1,\lambda_2,...,\lambda_n$. Let λ_{max} and λ_{min} be the largest and smallest eigenvalues, respectively. If $\mathbf{C}$ is positive definite, we have $\lambda_{min}>0$. Since the covariance matrix $\mathbf{C}$ is a symmetric matrix, we can decompose $\mathbf{C}$ by eigendecomposition as

$$\mathbf{C}=(\mathbf{AB})\cdot(\mathbf{AB})^T,$$

where $\mathbf{A}$ is an orthogonal matrix and $\mathbf{B}$ is a diagonal matrix whose i-th element $b_{ii}=\sqrt{\lambda_i}$.

We assume that $\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})$, and $\boldsymbol{z}=\sigma\mathbf{AB}\boldsymbol{z}'$ such that $\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})$. We can then calculate the probability as

$$\begin{aligned}&P_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})}\big(\mathcal{L}(\tilde{\boldsymbol{x}}^*+\boldsymbol{z})<\mathcal{L}(\tilde{\boldsymbol{x}}^*)\big)\\ \leq&P_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big(\|\tilde{\boldsymbol{x}}^*+\sigma\mathbf{AB}\boldsymbol{z}'-\boldsymbol{x})\|<\|\tilde{\boldsymbol{x}}^*-\boldsymbol{x}\|\big)\\ =&P_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big((\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot(\sigma\mathbf{AB}\boldsymbol{z}')<-\frac{1}{2}\sigma^2\|\mathbf{AB}\boldsymbol{z}'\|^2\big)\\ \leq&P_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big((\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot(\mathbf{AB}\boldsymbol{z}')<-\frac{1}{2}\sigma\lambda_{min}\|\boldsymbol{z}'\|^2\big).\end{aligned}$$

According to the law of large number [9], we have

$$\|\boldsymbol{z}'\|^2\xrightarrow{a.s.}n\quad\text{when } n\rightarrow\infty.$$

We then calculate the mean and variance of the random variable $\boldsymbol{y}=(\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot(\mathbf{AB}\boldsymbol{z}')$

$$\mathrm{E}(\boldsymbol{y})=\mathrm{E}\big[(\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot(\mathbf{AB}\boldsymbol{z}')\big]=0.$$

$$\begin{aligned}\mathrm{Var}(\boldsymbol{y})&=\int_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}|(\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot(\mathbf{AB}\boldsymbol{z}')|^2d\boldsymbol{z}'\\ &\leq\lambda_{max}\int_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}|(\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot\boldsymbol{z}'|^2d\boldsymbol{z}'\\ &=\lambda_{max}\|\tilde{\boldsymbol{x}}^*-\boldsymbol{x}\|^2.\end{aligned}$$

Finally, according to the Chebyshev's inequality [5], we have

$$\begin{aligned}&P_{\boldsymbol{z}\sim\mathcal{N}(\mathbf{0},\sigma^2\mathbf{C})}\big(\mathcal{L}(\tilde{\boldsymbol{x}}^*+\boldsymbol{z})<\mathcal{L}(\tilde{\boldsymbol{x}}^*)\big)\\ \leq&P_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big((\tilde{\boldsymbol{x}}^*-\boldsymbol{x})\cdot(\mathbf{AB}\boldsymbol{z}')<-\frac{1}{2}\sigma\lambda_{min}\|\boldsymbol{z}'\|^2\big)\\ =&P_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big(\boldsymbol{y}<-\frac{1}{2}\sigma\lambda_{min}n\big)\\ \leq&P_{\boldsymbol{z}'\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\big(|\boldsymbol{y}|>\frac{1}{2}\sigma\lambda_{min}n\big)\\ \leq&\frac{\mathrm{Var}(\boldsymbol{y})}{(\frac{1}{2}\sigma\lambda_{min}n)^2}\\ \leq&\frac{4\lambda_{max}\|\tilde{\boldsymbol{x}}^*-\boldsymbol{x}\|^2}{\sigma^2\lambda_{min}^2n^2}.\end{aligned}$$

□

B. Results on MegeFace

We supplement the results on the MegaFace dataset [7]. We attack SphereFace [8], CosFace [13], and ArcFace [3] by Boundary [1], Optimization [2], NES-LO [6], and the proposed Evolutionary for face verification and identification, respectively. We show the distortion curves over the number of queries in Fig. 7 for face verification, and Fig. 8 for face identification, respectively. We also report the distortion values of different methods at 1,000, 5,000, 10,000, and 100,000 queries in Table 5 for face verification, and Table 6 for face identification, respectively. The proposed method outperforms the other methods in all settings on the MegaFace dataset. The results are consistent with those based on the LFW dataset.

[*] Corresponding author.

Model		SphereFace [8]				CosFace [13]				ArcFace [3]			
Queries		1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000
Dodging	Boundary [1]	2.5e-2	8.8e-3	8.3e-4	2.4e-5	2.0e-2	7.2e-3	9.0e-4	1.9e-5	2.5e-2	1.7e-2	1.6e-3	2.5e-5
	Optimization [2]	1.3e-2	2.9e-3	1.4e-3	8.9e-5	1.1e-2	3.0e-3	1.4e-3	8.7e-5	1.7e-2	5.3e-3	2.4e-3	1.0e-4
	NES-LO [6]	1.5e-1	4.2e-2	2.7e-2	6.9e-3	1.4e-1	3.8e-2	2.3e-2	6.5e-3	1.4e-1	4.2e-2	2.7e-2	1.8e-2
	Evolutionary	**1.7e-3**	**1.0e-4**	**4.1e-5**	**1.6e-5**	**1.7e-3**	**1.0e-4**	**3.9e-5**	**1.3e-5**	**2.6e-3**	**1.6e-4**	**5.4e-5**	**1.8e-5**
Impersonation	Boundary [1]	1.8e-2	8.4e-3	7.9e-4	2.3e-5	1.1e-2	3.9e-3	3.6e-4	1.1e-5	1.7e-2	9.9e-3	1.5e-3	2.2e-5
	Optimization [2]	1.4e-2	4.6e-3	1.9e-3	8.5e-5	7.7e-3	2.3e-3	8.9e-4	4.0e-5	1.4e-2	6.7e-3	3.5e-3	9.6e-5
	NES-LO [6]	9.2e-2	3.0e-2	2.1e-2	7.7e-3	7.9e-2	2.2e-2	1.4e-2	4.7e-3	7.9e-2	2.9e-2	1.9e-2	9.3e-3
	Evolutionary	**1.5e-3**	**9.5e-5**	**3.9e-5**	**1.6e-5**	**8.2e-4**	**4.9e-5**	**2.0e-5**	**7.6e-6**	**2.7e-3**	**1.6e-4**	**4.9e-5**	**1.6e-5**

Table 5. The results on face verification conducted on the MegaFace dataset. We report the average distortion (MSE) of the adversarial images generated by different methods for SphereFace, CosFace, and ArcFace given 1,000, 5,000, 10,000, and 100,000 queries.

Model		SphereFace [8]				CosFace [13]				ArcFace [3]			
Queries		1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000
Dodging	Boundary [1]	3.9e-2	1.1e-2	1.0e-3	2.7e-5	2.8e-2	7.6e-3	7.9e-4	1.9e-5	3.8e-2	2.4e-2	2.3e-3	3.5e-5
	Optimization [2]	2.0e-2	4.2e-3	1.7e-3	9.4e-5	1.4e-2	3.0e-3	1.3e-3	6.9e-5	2.6e-2	8.0e-3	3.6e-3	1.4e-4
	NES-LO [6]	1.5e-1	5.3e-2	3.7e-2	9.3e-3	1.4e-1	4.7e-2	3.3e-2	7.7e-3	1.4e-1	5.5e-2	4.1e-2	1.7e-2
	Evolutionary	**2.3e-3**	**1.3e-4**	**4.8e-5**	**1.8e-5**	**1.7e-3**	**9.3e-5**	**3.5e-5**	**1.2e-5**	**3.6e-3**	**1.9e-4**	**6.7e-5**	**2.2e-5**
Impersonation	Boundary [1]	2.4e-2	1.1e-2	1.5e-3	3.8e-5	2.0e-2	7.1e-3	1.0e-3	2.5e-5	2.0e-2	1.3e-2	2.4e-3	4.6e-5
	Optimization [2]	1.7e-2	6.1e-3	2.9e-3	1.5e-4	1.4e-2	4.7e-3	2.1e-3	1.1e-4	1.6e-2	8.4e-3	4.5e-3	2.3e-4
	NES-LO [6]	8.8e-2	3.6e-2	2.6e-2	1.0e-2	7.5e-2	3.2e-2	2.3e-2	8.2e-3	7.5e-2	3.3e-2	2.4e-2	1.2e-2
	Evolutionary	**2.4e-3**	**1.7e-4**	**6.7e-5**	**2.6e-5**	**1.8e-3**	**1.3e-4**	**5.0e-5**	**1.7e-5**	**3.4e-3**	**2.7e-4**	**1.0e-4**	**3.2e-5**

Table 6. The results on face identification conducted on the MegaFace dataset. We report the average distortion (MSE) of the adversarial images generated by different methods for SphereFace, CosFace, and ArcFace given 1,000, 5,000, 10,000, and 100,000 queries.

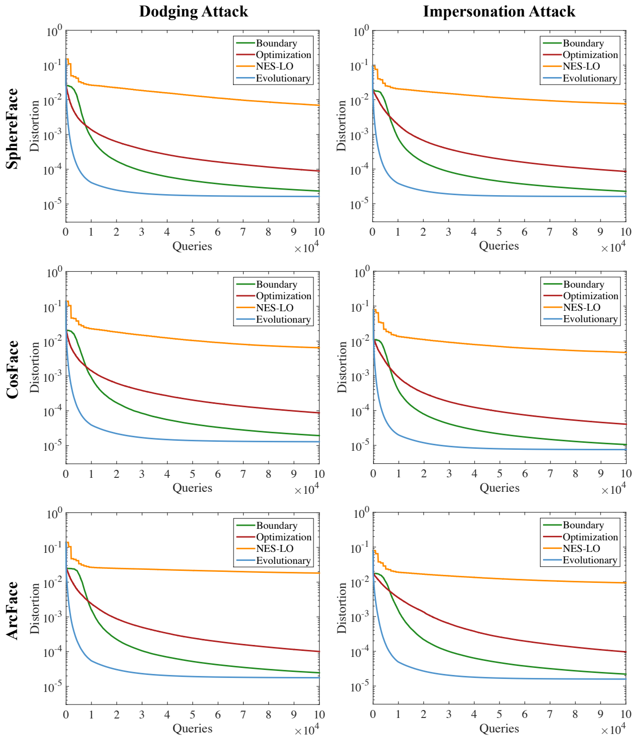

Figure 7. The results on face verification conducted on the MegaFace dataset. We show the curves of the average distortion (MSE) of the adversarial images generated by different attack methods for SphereFace, CosFace, and ArcFace over the number of queries.

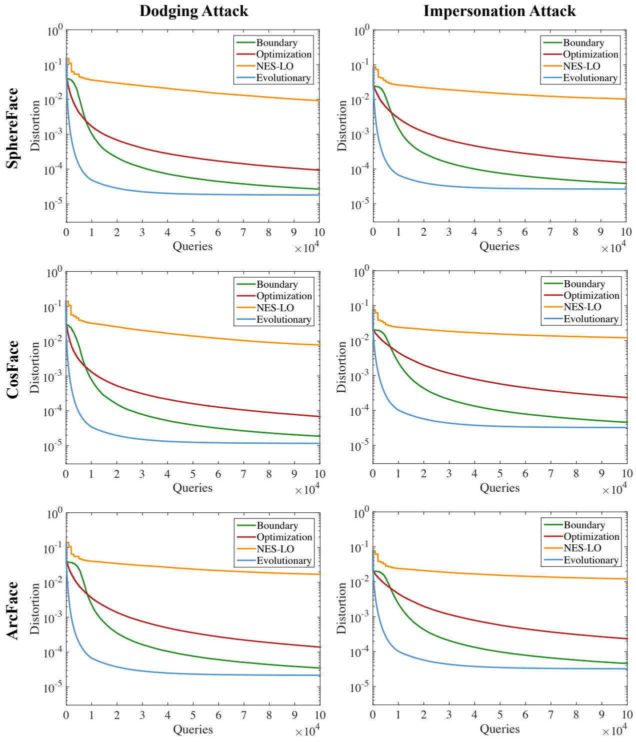

Figure 8. The results on face identification conducted on the MegaFace dataset. We show the curves of the average distortion (MSE) of the adversarial images generated by different attack methods for SphereFace, CosFace, and ArcFace over the number of queries.

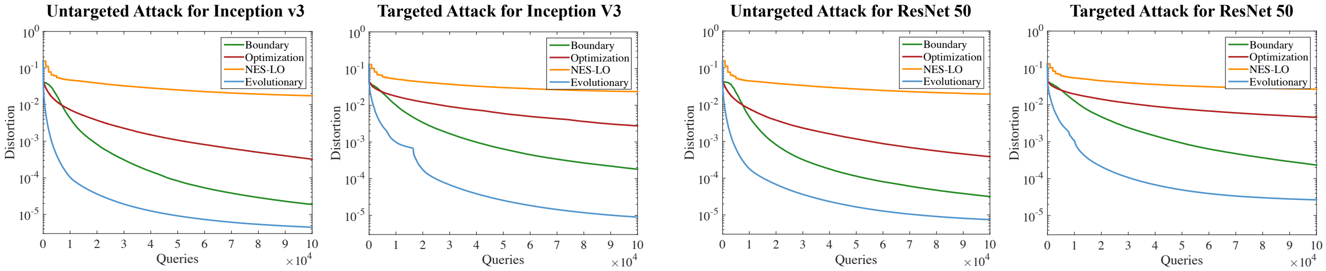

Figure 9. The results of untargeted and targeted attacks on the ImageNet dataset. We show the curves of the average distortion (MSE) of the adversarial images generated by different attack methods for the Inception v3 and ResNet 50 networks over the number of queries.

Model		Inception v3 [12]				ResNet 50 [4]			
Queries		1,000	5,000	10,000	100,000	1,000	5,000	10,000	100,000
Untargeted	Boundary [1]	4.0e-2	1.8e-2	4.2e-3	1.9e-5	4.1e-2	2.1e-2	4.6e-3	3.2e-5
	Optimization [2]	2.8e-2	1.2e-2	7.3e-3	3.2e-4	2.9e-2	1.3e-2	7.7e-3	3.8e-4
	NES-LO [6]	1.5e-1	6.2e-2	4.7e-2	1.8e-2	1.5e-1	5.9e-2	4.4e-2	1.9e-2
	Evolutionary	**5.3e-3**	**4.2e-4**	**1.0e-4**	**4.5e-6**	**6.6e-3**	**6.3e-4**	**1.8e-4**	**7.4e-6**
Targeted	Boundary [1]	3.7e-2	2.2e-2	1.0e-2	1.8e-4	3.9e-2	2.5e-2	1.3e-2	2.3e-4
	Optimization [2]	3.4e-2	2.3e-2	1.8e-2	2.7e-3	3.6e-2	2.5e-2	2.0e-2	4.6e-3
	NES-LO [6]	1.3e-1	6.6e-2	5.2e-2	2.4e-2	1.3e-1	6.7e-2	5.4e-2	2.7e-2
	Evolutionary	**1.4e-2**	**2.7e-3**	**9.9e-4**	**9.0e-6**	**1.6e-2**	**3.2e-3**	**1.1e-3**	**2.7e-5**

Table 7. The results of untargeted and targeted attacks on the ImageNet dataset. We report the average distortion (MSE) of the adversarial images generated by different methods for the Inception v3 and ResNet 50 networks given 1,000, 5,000, 10,000, and 100,000 queries.

C. Results on ImageNet

It should be noted that the proposed evolutionary attack method is not restricted to attacking face recognition models. It could be used to perform decision-based black-box attacks for any image classification tasks. In this section, we conduct additional experiments to demonstrate the effectiveness of the evolutionary attack method in the general object recognition task based on the ImageNet [11] dataset. We use the Inception v3 [12] and ResNet 50 [4] networks in our experiments. We choose 100 images from the ImageNet validation set, which are correctly classified by these two models. We perform untargeted attack and targeted attack against each model by Boundary, Optimization, NES-LO, and Evolutionary in the decision-based black-box setting. We show the results in Fig. 9 and Table 7. The experimental results consistently demonstrate the effectiveness of the proposed method.

D. Experiments Requested by the Reviewers

We provide the experimental results requested by the reviewers during the review process.

D.1. Standard Deviation of the Distortion

We provide the mean, standard deviation, and maximum of the distortion (MSE) over the 500 pairs of images of LFW in Table 8. The results are based on our method for face verification given 100,000 queries. Some adversarial images have larger distortions. But the maximum distortions are smaller than $1.1e^{-4}$, which is almost imperceptible for humans (see the examples in Fig. 4).

D.2. A different Initial Image for Impersonation Attacks

In impersonation attacks, we use the original target image (enrollment image) as the initialization. We agree that using a different image of the target identity is more practical than using the enrollment image. However, when we are given an image of the target identity, our method could be always used to find a minimum perturbation, no matter whether the initial image is the enrollment image or a different image. To verify this, we use a different image of the target identity as the initial image to perform impersonation attacks on face verification. The average distortions after 100,000 queries are $1.1e^{-5}$, $4.7e^{-6}$, and $1.1e^{-5}$ for the three models, which are very similar to $1.2e^{-5}$, $5.3e^{-6}$, and $1.2e^{-5}$ shown in Table 1, where the initial image is the enrollment image.

D.3. Compared with White-box Attacks

We attack the CosFace model by the white-box attack method PGD [10] for face verification. For each pair of face images, we find a minimum perturbation that leads to misclassification by binary search. The average distortions

	SphereFace			CosFace			ArcFace		
	mean	std	max	mean	std	max	mean	std	max
Dodging	1.3e-5	1.2e-5	9.4e-5	1.1e-5	9.4e-6	5.2e-5	1.6e-5	1.2e-5	7.2e-5
Impersonation	1.2e-5	8.1e-6	6.2e-5	5.3e-6	4.3e-6	2.4e-5	1.2e-5	9.2e-6	1.1e-4

Table 8. The mean, standard deviation, and maximum of the distortion (MSE) over the 500 pairs of images based on the LFW dataset.

over the 500 pairs are $1.7e^{-5}$ for dodging attack, and $8.0e^{-6}$ for impersonation attack, which are larger than the average distortions given by our method ($1.1e^{-5}$ and $5.3e^{-6}$ shown in Table 1).

References

[1] W. Brendel, J. Rauber, and M. Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *ICLR*, 2018. 1, 2, 3

[2] M. Cheng, T. Le, P.-Y. Chen, J. Yi, H. Zhang, and C.-J. Hsieh. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457*, 2018. 1, 2, 3

[3] J. Deng, J. Guo, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *arXiv preprint arXiv:1801.07698*, 2018. 1, 2

[4] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 3

[5] P. J. Huber et al. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 221–233. University of California Press, 1967. 1

[6] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. Black-box adversarial attacks with limited queries and information. In *ICML*, 2018. 1, 2, 3

[7] I. Kemelmacher-Shlizerman, S. M. Seitz, D. Miller, and E. Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *CVPR*, 2016. 1

[8] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song. Sphereface: Deep hypersphere embedding for face recognition. In *CVPR*, 2017. 1, 2

[9] M. Loeve. Probability theory. 4-th edn, 1977. 1

[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018. 3

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 3

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[13] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu. Cosface: Large margin cosine loss for deep face recognition. In *CVPR*, 2018. 1, 2