Supplementary Materials for Sibling-Attack: Rethinking Transferable Adversarial Attacks against Face Recognition

Zexin Li[1*] Bangjie Yin[3*] Taiping Yao[3] Junfeng Guo[2] Shouhong Ding[3†] Simin Chen[2] Cong Liu[1†]
[1]University of California, Riverside
[2]The University of Texas at Dallas
[3]Tencent
{zli536, congl}@ucr.edu, {junfeng.guo, simin.chen}@utdallas.edu,
{bangjieyin, taipingyao, ericshding}@tencent.com

1. Implementation Details

Attributes Selection. In CelebA-HQ [5], there exists 40 attribute labels. However, not all attributes are highly related to face recognition(e.g., hair color, necklace, etc.). In the pre-processing process, we must align and crop the face region from the original face images before inputting them into the face recognition model, which will remove most of the background. Therefore, to efficiently use facial attributes, we select 18 useful attributes relevant to the critical face regions rather than apply them all to the AR branch during the training procedure, as shown in Table. 1.

Network Structure Details. We realize our framework via the hard-parameter sharing [1], where the key idea has been discussed in Sec. 1 of the manuscript. Moreover, we present the detailed network structure of our constructed surrogate model. As shown in Table. 2, we use IR152 [3] as the backbone and then split it into two branches, FR and AR, at Conv.4-13. Finally, we perform face classification and attribute prediction at the end of each branch.

Details of Attributes Divisions In the manuscript, we divide the selected 18 attributes into 4 non-overlapped groups to evaluate the effectiveness of each attribute group. Here, we give the details of the divisions, i.e., Eye-region= $\{1, 3, 12, 15, 23\}$, Nose-region= $\{0, 7, 19, 27\}$, Mouth-region= $\{6, 21, 22, 24, 36\}$ and Other-region= $\{13, 20, 25, 31\}$, the numbers are the attribute indexes from CelebA-HQ, as shown in Table. 1.

No.	Attr. Name	No.	Attr. Name
0	5_o_Clock_Shadow	20	Male
1	Arched_Eyebrows	21	Mouth_Slightly_Open
3	Bags_Under_Eyes	22	Mustache
6	Big_Lips	23	Narrow_Eyes
7	Big_Nose	24	No_Beard
12	Bushy_Eyebrows	25	Oval_Face
13	Chubby	27	Pointy_Nose
15	Eyeglasses	31	Smiling
19	High_Cheekbones	36	Wearing_Lipstick

Table 1. Selected face-related attributes for training AR branch. The numbers are the attribute indexes in the CelebA-HQ [5].

2. Evaluation of White-box FR Models.

In addition to demonstrating the attacking transferability against black-box FR models in the manuscript, we also illustrate white-box attacking results, as shown in Table. 3. Specifically, to have a fair comparison, we only present the white-box results for comparison work with attacks on the entire image as *Sibling-Attack*. Besides, all the competitors use the ensemble attacking strategy against two different FR models, thus the evaluations are conducted on the original white-box FR models. Different from the competitors, our *Sibling-Attack*'s white-box attacking results will be generated on the white-box surrogate models of the proposed multi-task framework. More importantly, we set the thresholds τ of IR152, FaceNet, IRSE50, to be $\{0.228, 0.591, 0.313\}$ at 0.001 FAR following [3, 7]. The comparisons in Table. 3 demonstrate that the white-box attacking success rates of all the methods are above 99.50% and there is no noticeable difference among them, which is why we mainly care about the transferable ASRs against the black-box FR models.

3. Evaluation of Robust FR Models.

To further evaluate the robustness of our proposed method, except for the normally trained face recognition models, we evaluate the transferability on two

*indicates equal contributions.
†indicates corresponding author.

Proposed Network Structure									
Shared Encoder $\mathcal{P}$			Face Recognition $\mathcal{F}$			Attribute Recognition $\mathcal{A}$		
Layer	K./C./S.	Out.Size	Layer	K./C./S.	Out.Size	Layer	K./C./S.	Out.Size
Input: BGR Image			Input: C4-13			Input: C4-13		
C.1-0	$3 \times 3/64/1$	112×112						
	$1 \times 1/64/2$	56×56		$1 \times 1/256/2$	14×14		$1 \times 1/256/2$	14×14
C.2-x (	$3 \times 3/64/1$	56×56) $\times 3$	C.4-x (	$3 \times 3/256/1$	14×14) $\times 22$	C.4-x (	$3 \times 3/256/1$	14×14) $\times 22$
	$3 \times 3/64/1$	56×56		$3 \times 3/256/1$	14×14		$3 \times 3/256/1$	14×14
	$1 \times 1/128/2$	28×28						
C.3-x (	$3 \times 3/128/1$	28×28) $\times 8$						
	$3 \times 3/128/1$	28×28		$1 \times 1/512/2$	7×7		$1 \times 1/512/2$	7×7
	$1 \times 1/256/2$	14×14	C.5-x (	$3 \times 3/512/1$	7×7) $\times 3$	C.5-x (	$3 \times 3/512/1$	7×7) $\times 3$
C.4-x (	$3 \times 3/256/1$	14×14) $\times 14$		$3 \times 3/512/1$	7×7		$3 \times 3/512/1$	7×7
	$3 \times 3/256/1$	14×14						
Output: C4-13			Output: face recognition			Output: attribute prediction		

Table 2. The details for surrogate networks structure of *Sibling-Attack*.

Dataset	CelebA-HQ				LFW			
Source Model	IR152+FaceNet		IR152+IRSE50		IR152+FaceNet		IR152+IRSE50	
Target Model	IR152	FaceNet	IR152	IRSE50	IR152	FaceNet	IR152	IRSE50
PGD [6]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
TAP [11]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
MI-FGSM [4]	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
VMI-FGSM [8]	99.80	100.00	99.60	100.00	100.00	100.00	99.80	100.00
	IR152FR + IR152AR				IR152FR + IR152AR			
Sibling-Attack	**100.00**				**100.00**			

Table 3. ASR results of white-box impersonation attack over CelebA-HQ and LFW dataset. Our method uses IR152 FR and IR152 AR for white-box training, while other methods for comparisons are trained using two different FR models. Our attack performance results are shown in **bold**.

Methods	Dataset	CelebA-HQ				LFW			
	Source Model	IR152+FaceNet		IR152+IRSE50		IR152+FaceNet		IR152+IRSE50	
	Target Model	AT	TRADES	AT	TRADES	AT	TRADES	AT	TRADES
General Attacks	PGD [6]	17.10	9.50	19.50	13.60	18.00	24.10	19.30	32.50
	TAP [11]	18.30	11.50	19.30	13.50	18.80	21.70	21.50	34.80
	MI-FGSM [4]	18.70	10.40	20.40	16.30	21.00	26.10	23.30	37.10
	VMI-FGSM [8]	17.60	8.40	18.10	8.80	19.50	27.00	18.10	23.30
	Adv-Face [2]	24.20	11.20	22.20	6.20	10.30	10.60	10.30	10.60
Ours	*Sibling-Attack*	**26.10**	**19.40**	**26.10**	**19.40**	**27.10**	**48.00**	**27.10**	**48.00**
		1.90 ↑	7.90 ↑	3.90 ↑	3.10 ↑	6.10 ↑	21.00 ↑	3.80 ↑	10.90 ↑

Table 4. ASR results of black-box impersonation attack on *adversarial trained* defense models. We choose methods exhibiting stronger transferability in the manuscripts (Adv-Face and transfer-based methods) for comparisons of our proposed methods. AT represents PGD-AT. Best attack performance results are shown in bold.

black-box *adversarial trained* FR models: PGD-AT [6], TRADES [10]. The thresholds for computing the ASRs of these two models are also obtained from images in the LFW dataset. Specifically, we set τ to (0.233, 0.636)

following [6, 9, 10] for PGD-AT, TRADES, respectively. As shown in Tab. 4, *Sibling-Attack* outperforms the best competitors by $(1.90\%, 3.10\%)$ on CelebA-HQ as well as $(3.80\%, 10.90\%)$ on LFW.

References

[1] Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997. 1

[2] Debayan Deb, Jianbang Zhang, and Anil K Jain. Advfaces: Adversarial face synthesis. In *2020 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2020. 2

[3] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4690–4699, 2019. 1

[4] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9185–9193, 2018. 2

[5] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of GANs for improved quality, stability, and variation. In *International Conference on Learning Representations (ICLR)*, 2018. 1

[6] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[7] Florian Schroff, Kalenichenko Dmitry, and Philbin James. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[8] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1924–1933, 2021. 2

[9] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1252–1258. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 3

[10] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482. PMLR, 2019. 2, 3

[11] Wen Zhou, Xin Hou, Yongjun Chen, Mengyun Tang, Xiangqi Huang, Xiang Gan, and Yong Yang. Transferable adversarial perturbations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 452–467, 2018. 2