Improving the Transferability of Adversarial Attacks on Face Recognition with Diverse Parameters Augmentation

Supplementary Material

7. Appendix

Overview. The supplementary includes the following sections:

- **Section 7.1.** Computation Methodology for Attack Success Rate.
- **Section 7.2.** More Detailed Attack Settings.
- **Section 7.3.** More Comparison Studies on LFW.
- **Section 7.4.** Comparison Studies on CelebA-HQ.
- **Section 7.5.** Hyper-parameter Analysis Studies.
- **Section 7.6.** Visual Quality Study.
- **Section 7.7.** Ethics and Potential Broader Impact.

7.1. Computation Methodology for Attack Success Rate

In our study, the Attack Success Rate (ASR) is determined using the following formula:

$$\mathrm{ASR} = \frac{\sum_{i=1}^{N_p} \mathbb{1}\left(\widetilde{\mathcal{D}}\left(\mathcal{F}^{vct}\left(x^{adv}\right), \mathcal{F}^{vct}\left(x^{t}\right)\right) < t^{i}\right)}{N_p} \tag{19}$$

where the notation $\widetilde{\mathcal{D}}$ designates a predefined distance metric for assessing the performance of adversarial face examples, N_p denotes the total count of face pairs, and t^i signifies the attack threshold.

7.2. More Detailed Attack Settings

In the DPO stage, we set the learning rate to 0.1. In the HMA stage, we specifically target convolutional layers to introduce beneficial perturbations. We maintain the step size for adversarial perturbations β at a fixed value of 1. We have set the scale factor d to 32.0 and the margin m to 0.5. We employ the SGD optimizer for model augmentation.

For the tables and figures mentioned—Table 6, Table 7, Table 1, Table 5, Table 9, Table 10, Table 11, Table 8, Figure 8, Figure 5, and Figure 6, we determine setting c to 35, which corresponds to the optimal value from the left plot in Figure 7, and setting the step size of beneficial perturbations η to 8e-4, as indicated by the optimal value from the right plot in Figure 7.

Regarding the bottom portion of Figure 1, we have configured the settings for LGV according to the same hyper-parameters as specified in Table 2. Similarly, the settings for DI, BPFA, and DPA are aligned with those detailed in Table 1.

Table 5. Comparisons of black-box ASR (%) results for attacks using IR152 as the surrogate model on the LFW dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	S	F	M	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	29.0	9.3	5.6	13.8	6.8	3.6	2.4
DI [59]	46.9	21.7	14.4	28.0	12.4	7.9	6.1
DFANet [69]	50.7	15.5	12.5	25.6	11.0	5.8	3.2
VMI [50]	49.7	23.9	18.7	30.1	18.3	12.8	11.2
SSA [27]	55.0	21.9	24.0	28.8	14.2	9.0	6.1
SIA [52]	52.6	26.3	19.6	29.8	18.3	11.0	9.5
BPFA [71]	46.7	12.9	9.2	20.1	8.9	4.7	3.1
BSR [46]	35.3	14.7	7.3	19.2	9.9	6.6	4.3
Ours	**99.4**	**90.3**	**96.4**	**74.0**	**69.9**	**42.0**	**57.7**

Table 6. Comparisons of black-box ASR (%) results for attacks using IRSE50 as the surrogate model on the LFW dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	I	F	M	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	32.3	15.5	79.1	9.8	17.5	5.5	5.7
DI [59]	59.9	47.5	97.7	25.9	41.5	15.6	23.8
DFANet [69]	44.3	26.7	96.9	15.3	28.0	8.6	12.4
VMI [50]	54.0	34.0	96.4	22.5	37.6	13.1	20.3
SSA [27]	58.8	37.1	97.3	22.4	38.5	12.3	18.1
SIA [52]	58.4	41.4	98.2	22.2	37.6	13.9	23.3
BPFA [71]	54.4	27.5	94.6	17.6	29.4	8.1	12.8
BSR [46]	28.7	18.4	84.5	9.2	16.3	6.5	7.6
Ours	**74.4**	**89.8**	**98.3**	**57.9**	**68.2**	**38.3**	**59.7**

Table 7. Comparisons of black-box ASR (%) results for attacks using FaceNet as the surrogate model on the LFW dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	I	S	M	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	7.8	12.5	5.4	7.5	6.9	17.2	2.5
DI [59]	18.6	32.2	18.5	18.0	15.8	30.2	9.9
DFANet [69]	12.1	22.2	11.7	11.3	10.4	25.1	5.5
VMI [50]	24.4	35.1	20.7	24.4	23.2	36.3	15.2
SSA [27]	21.6	44.8	30.8	17.7	17.0	31.9	10.9
SIA [52]	29.1	42.9	26.2	28.7	23.9	38.3	16.7
BPFA [71]	17.3	31.6	14.7	13.4	13.0	22.6	7.8
BSR [46]	28.6	42.4	25.9	26.2	24.3	34.2	16.0
Ours	**42.6**	**65.0**	**56.9**	**47.3**	**45.4**	**54.0**	**31.1**

7.3. More Comparison Studies on LFW

To validate the efficacy of our proposed attack method, we generate adversarial examples using IR152, IRSE50, and FaceNet as surrogate models on the LFW dataset. The black-box performance is presented in Table 5, Table 6, and Table 7, respectively. Our method consistently outperforms baseline attacks, demonstrating its effectiveness in improving the transferability of adversarial examples.

Table 8. Comparisons of black-box ASR (%) results for attacks using IR152 as the surrogate model on the CelebA-HQ dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	S	F	M	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	39.2	14.1	12.6	18.3	12.1	4.3	4.9
DI [59]	57.6	27.6	27.0	30.1	20.5	8.5	10.8
DFANet [69]	61.2	21.5	22.4	26.6	17.0	5.7	7.9
VMI [50]	56.9	26.4	27.7	32.4	26.1	12.6	16.6
SSA [27]	62.1	23.3	30.7	30.6	19.4	9.1	10.3
SIA [52]	60.8	26.9	30.4	30.4	24.3	12	14.2
BPFA [71]	54.6	15.8	16.4	22.0	14.2	5.0	5.4
BSR [46]	42.9	17.3	15.0	21.1	15.4	6.6	6.8
Ours	**98.0**	**82.4**	**95.1**	**53.5**	**61.5**	**27.6**	**52.8**

Table 9. Comparisons of black-box ASR (%) results for attacks using IRSE50 as the surrogate model on the CelebA-HQ dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	I	F	M	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	36.3	16.2	80.5	10.5	21.6	5.1	9.5
DI [59]	59.3	42.8	97.6	20.5	39.4	12.0	28.5
DFANet [69]	45.9	25.5	97.0	14.0	30.8	6.6	15.4
VMI [50]	56.7	31.9	96.6	18.5	37.5	9.9	24.3
SSA [27]	58.3	34.8	97.5	19.1	38.2	8.9	22.1
SIA [52]	60.7	40.0	97.9	20.1	40.1	11.5	26.4
BPFA [71]	56.8	27.9	95.3	16.7	32.8	7.5	17.3
BSR [46]	35.7	19.9	86.4	11.2	20.4	5.3	12.0
Ours	**68.9**	**81.7**	**98.0**	**40.2**	**60.6**	**25.5**	**53.8**

Table 10. Comparisons of black-box ASR (%) results for attacks using FaceNet as the surrogate model on the CelebA-HQ dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	I	S	M	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	10.7	16.5	9.6	7.2	8.2	13.0	4.4
DI [59]	22.8	30.4	22.9	15.0	19.9	21.7	11.8
DFANet [69]	15.2	24.3	19.7	10.0	14.1	18.0	7.7
VMI [50]	26.4	36.4	25.7	19.5	25.2	25.3	16.4
SSA [27]	23.5	41.5	36.1	14.9	19.4	22.4	13.2
SIA [52]	29.3	44.5	35.7	21.5	26.0	27.6	18.4
BPFA [71]	20.0	31.0	21.7	12.1	14.0	16.8	7.9
BSR [46]	27.8	43.9	34.0	19.2	25.9	26.0	18.3
Ours	**35.5**	**56.9**	**58.7**	**29.9**	**36.2**	**32.2**	**28.3**

7.4. Comparison Studies on CelebA-HQ

To further validate the efficacy of our proposed attack method, we create adversarial examples utilizing the CelebA-HQ dataset. The black-box performance of our approach, which employs IR152, IRSE50, FaceNet, and MobileFace as surrogate models on the CelebA-HQ dataset, is presented in Table 8, Table 9, Table 10, and Table 11, respectively. These results consistently indicate that our method outperforms the baseline attacks, thereby highlighting its effectiveness in improving the transferability of adversarial examples.

Table 11. Comparisons of black-box ASR (%) results for attacks using MobileFace as the surrogate model on the CelebA-HQ dataset. I, S, F, M denote IR152, IRSE50, FaceNet, and MobileFace, respectively.

Attacks	I	S	F	I^{adv}	S^{adv}	F^{adv}	M^{adv}
FIM [68]	7.2	69.0	8.0	3.3	5.4	2.4	12.0
DI [59]	25.0	97.0	32.0	11.0	21.4	8.2	39.0
DFANet [69]	10.7	85.0	12.6	3.7	9.1	2.9	18.8
VMI [50]	20.2	94.7	19.6	7.8	15.8	4.4	31.5
SSA [27]	22.0	95.4	21.0	7.9	15.2	5.3	33.1
SIA [52]	25.4	96.6	25.9	8.9	18.4	5.9	35.8
BPFA [71]	20.7	94.7	17.5	6.7	13.8	4.5	29.7
BSR [46]	10.8	77.1	11.9	3.8	8.1	2.8	15.1
Ours	**59.6**	**97.2**	**83.5**	**37.1**	**57.5**	**27.4**	**61.7**

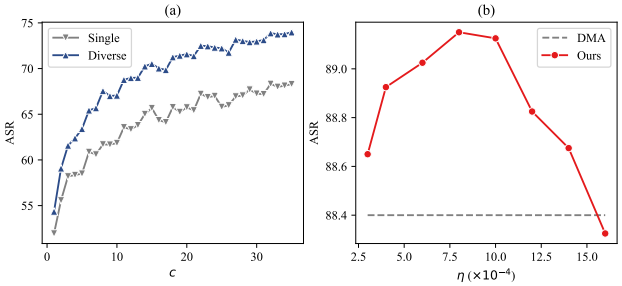

Figure 7. The hyper-parameter analysis on the (a) c and (b) η.

7.5. Hyper-parameter Analysis Studies

The hyper-parameter analysis on the c value. The value of c determines the number of ensembles in our proposed attack method, which significantly affects its performance. Hence, we conduct ablation studies on c using the LFW dataset with MobileFace as the surrogate model. To further verify the effectiveness of diverse parameters in enhancing transferability, we select two types of attack methods for comparison. Firstly, we use MobileFace models fine-tuned by a pre-trained backbone and a randomly initialized head in each epoch to generate adversarial examples. We term this adversarial attack method 'Single'. Secondly, we choose the models trained by 'Single' and MobileFace models trained by a randomly initialized backbone and head in each epoch to create adversarial examples, implying that the parameters of the trained models are more diverse. We term this attack method 'Diverse'. The average ASR on IR152, IRSE50, FaceNet, and MobileFace is demonstrated in the left plot of Figure 7. The left plot of Figure 7 shows that the ASR increases and then converges as c increases. To analyze the reason, we need to consider the property of c. c determines the number of models to be aggregated. If more models are aggregated in each training epoch, the aggregation capacity will increase. If c continues to increase, due to the similarity of the aggregated models in the later epochs, the ASR converges. Moreover, the left plot of Figure 7 demonstrates that the performance of 'Diverse' is higher than that of 'Single',

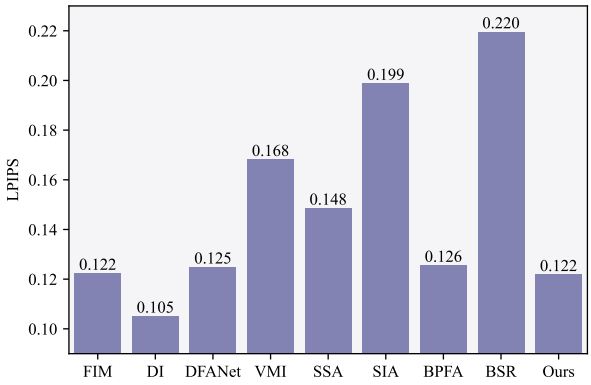

Figure 8. Comparison of LPIPS values across various attacks, with lower values signifying superior visual quality.

which verifies the effectiveness of parameter diversity in improving transferability of crafted adversarial examples.

The hyper-parameter analysis on the η value. The η value is the step size of beneficial perturbations, which is a key hyperparameter in our proposed attack method. We will conduct ablation studies on this parameter using the LFW dataset with MobileFace as the surrogate model. The average ASR on IR152, IRSE50, FaceNet, and MobileFace is shown in the right plot of Figure 7. To assess the effectiveness of hard models in enhancing the transferability of adversarial examples, we use the Diverse Model Aggregation (DMA) as a baseline for comparison. DMA replaces the hard models in our method with their corresponding vanilla models. From the right plot of Figure 7, we observe that as the step size of beneficial perturbations increases, the ASR initially rises and then declines. To understand this behavior, we should consider the nature of beneficial perturbations. These perturbations are added to the feature maps of FR models to increase loss when crafting adversarial examples, effectively transforming FR models into hard models. Increasing the step size initially boosts transferability by strengthening the transition to hard models. However, further increasing the step size can degrade the features in the feature maps during forward propagation, ultimately reducing overall attack performance. Additionally, the right plot of Figure 7 demonstrates that the optimal performance of our proposed method surpasses that of DMA, further validating the effectiveness of the hard model ensemble in our attack method.

7.6. Visual Quality Study

Furthermore, we evaluate the visual quality of our proposed method against that of previous attack methods. We choose FIM, DI, DFANet, VMI, SSA, SIA, BPFA, and BSR as comparative baselines and generate adversarial examples using MobileFace as the surrogate model on the LFW dataset. The experimental configuration is consistent with the one detailed in Table 1. The outcomes are depicted in Figure 8. As shown in Figure 8, our proposed method achieves visual quality performance on par with other methods. Notably, the transferability of the adversarial examples generated by our method significantly exceeds that of the baselines, which further underscores the superiority of our proposed method.

7.7. Ethics and Potential Broader Impact

This paper introduces research that contributes to the advancement of the field within Computer Vision and Pattern Recognition. The attack method we propose poses a potential threat to the security of FR models. Our goal is to heighten awareness through this proposed method and strengthen the resilience of FR models against such vulnerabilities.