– Supplemental Material –
Enhancing the Self-Universality for Transferable Targeted Attack

Zhipeng Wei[1,2], Jingjing Chen[1,2]†, Zuxuan Wu[1,2], Yu-Gang Jiang[1,2]
[1]Shanghai Key Lab of Intell. Info. Processing, School of CS, Fudan University
[2]Shanghai Collaborative Innovation Center of Intelligent Visual Computing
zpwei21@m.fudan.edu.cn, {chenjingjing, zxwu, ygj}@fudan.edu.cn

1. Standard Deviation of Single-model Transferable Attacks

The standard deviation express the variability of data. We show the standard deviation of single-model transferable attacks in Table 1. We observe that the standard deviation is low. It demonstrates that our method steadily improve TASR.

2. Visualization of Adversarial Examples

We visualize 8 benign images and their corresponding adversarial images in Figure 1. As can be seen, these adversarial examples, which humans can correctly understand, fool DNNs into the specific target prediction.

3. Performance Comparison with $\epsilon = 8$

We conduct experiments with $\epsilon = 8$. Specifically, we set the weight parameter $\lambda = 10^{-3}$, the scale parameters $s = (0.1, 0)$, and use the layer 3 to extract features. The results are shown in Table 2. First, $\epsilon = 8$ achieves worse performance than $\epsilon = 16$, because the difficulty of targeted attacks increases when lowering ϵ to 8. Besides, the proposed SU can improve TASR in most cases. It demonstrates the effectiveness of SU.

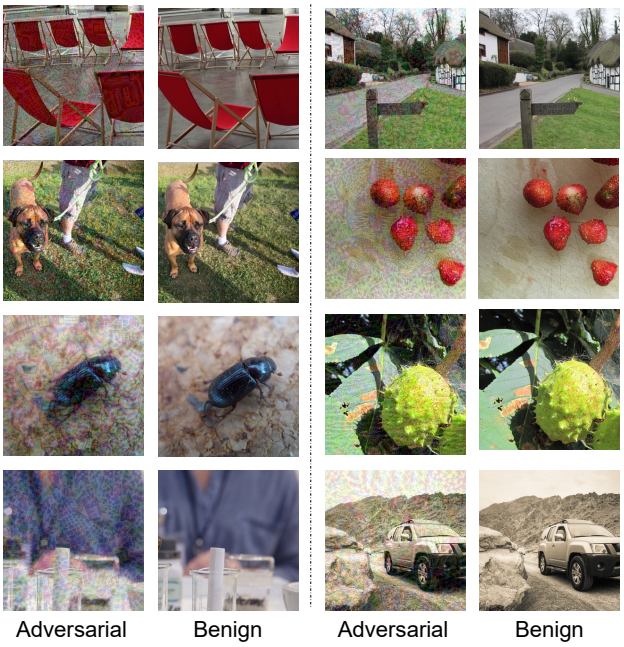

Figure 1. TASR (%) of attacking ResNet50 from DenseNet121.

†Corresponding author.

Attack	White-box Model: Res50			White-box Model: Dense121		
	→ Dense121	→ VGG16	→ Inc-v3	→ Res50	→ VGG16	→ Inc-v3
DTMI-CE	0.9/0.94/1.5	0.24/0.3/0.58	0.0/0.05/0.01	0.08/0.17/0.01	0.39/0.02/0.01	0.1/0.11/0.06
DTMI-CE-SU	0.03/1.6/0.78	0.01/0.12/0.55	0.02/0.08/0.17	0.01/0.44/0.38	0.05/0.48/0.07	0.01/0.03/0.04
DTMI-Logit	0.38/0.99/0.8	0.24/0.43/0.69	0.25/0.28/0.09	0.51/0.62/0.03	0.48/0.44/0.58	0.01/0.25/0.03
DTMI-Logit-SU	0.54/0.43/0.2	0.03/0.14/0.6	0.02/0.68/0.25	0.1/1.39/0.11	0.14/0.07/0.41	0.02/0.16/0.08
Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→ Res50	→ Dense121	→ Inc-v3	→ Res50	→ Dense121	→ VGG16
DTMI-CE	0.01/0.06/0.01	0.02/0.02/0.0	0.0/0.0/0.0	0.0/0.04/0.22	0.0/0.07/0.0	0.06/0.28/0.02
DTMI-CE-SU	0.0/0.06/0.1	0.0/0.06/0.38	0.0/0.0/0.0	0.02/0.03/0.1	0.0/0.08/0.04	0.0/0.18/0.0
DTMI-Logit	0.04/0.01/0.01	0.02/0.07/0.14	0.0/0.02/0.04	0.07/0.05/0.01	0.03/0.02/0.0	0.01/0.06/0.03
DTMI-Logit-SU	0.03/0.13/0.03	0.12/0.6/0.19	0.01/0.05/0.0	0.1/1.39/0.11	0.14/0.07/0.41	0.02/0.16/0.08

Table 1. The standard deviation of TASR (%) of Table 2. We use ResNet50, DenseNet121, VGGNet16 and Inception-v3 as the white-box model respectively. We conduct these experiments three times and report the standard diviation of TASR with 20/100/300 iterations.

Attack	White-box Model: Res50			White-box Model: Dense121		
	→ Dense121	→ VGG16	→ Inc-v3	→ Res50	→ VGG16	→ Inc-v3
DTMI-CE	11.0/19.6/21.8	7.6/15.0/15.1	0.3/0.5/0.7	4.4/6.4/6.1	2.7/4.6/3.8	0.1/0.3/0.3
DTMI-CE-SU	1.4/7.9/21.7	0.8/6.4/18.8	0.1/0.5/1.0	0.6/5.4/12.8	0.5/3.5/10.8	0.0/0.4/1.0
DTMI-Logit	11.7/34.2/40.8	9.1/30.5/38.4	0.2/0.8/0.9	6.4/18.6/21.7	4.4/14.9/18.5	0.0/0.8/1.0
DTMI-Logit-SU	7.9/33.0/45.7	5.2/30.7/41.3	0.2/1.4/1.1	3.4/18.4/23.2	2.5/15.5/23.3	0.1/1.0/1.1
Attack	White-box Model: VGG16			White-box Model: Inc-v3		
	→ Res50	→ Dense121	→ Inc-v3	→ Res50	→ Dense121	→ VGG16
DTMI-CE	0.2/0.1/0.2	0.0/0.0/0.0	0.0/0.0/0.0	0.4/0.7/0.8	0.1/0.6/1.1	0.2/0.2/0.7
DTMI-CE-SU	0.0/0.7/1.2	0.0/0.7/1.0	0.0/0.0/0.0	0.1/0.1/0.6	0.0/0.3/1.4	0.1/0.3/0.6
DTMI-Logit	0.8/2.0/2.5	0.3/2.3/2.8	0.0/0.0/0.0	0.3/0.7/1.0	0.3/0.9/0.9	0.2/0.4/1.0
DTMI-Logit-SU	0.8/3.1/3.9	0.3/4.0/4.6	0.0/0.0/0.0	0.2/0.8/1.5	0.3/1.2/1.6	0.1/0.4/1.0

Table 2. TASR (%) of all black-box models under four attack scenarios using ResNet50, DenseNet121, VGGNet16 and Inception-v3 as white-box models, respectively. We conduct these experiments three times and report average TASR with 20/100/300 iterations, the standard deviation are shown in Appendix. The best results with 300 iterations are in bold.