Appendix for Enhancing Transferability of Targeted Adversarial Examples via Inverse Target Gradient Competition and Spatial Distance Stretching

Zhankai Li[1], Weiping Wang[1], Jie Li[1], Shigeng Zhang[1], Yunan Hu[1], Song Guo[2]
[1]Central South University
[2]The Hong Kong University of Science and Technology
janlzk7@gmail.com, {wpwang, lijie55, sgzhang, hyn318}@csu.edu.cn, songguo@ust.hk

1. Overview

In this appendix, we first present the targeted attack success rates (TASRs) of ITDS$_{gs}$-DIM, the results of combining SDS with existing methods, the comparison with DIM-enhanced BSR [9], the ablation studies conducted on other surrogate models, the potentiator effect and computational costs. Subsequently, we assess the perception of adversarial examples (AEs) generated by various state-of-the-art (SOTA) attack methods to prove that the superior attack performance of our ITDS is not achieved at the cost of perception. Also, we present a series of AEs of each method. Finally, we discuss the rigor of dataset selection.

2. Evaluation on ITDS$_{gs}$-DIM

Since in the main text we only use ITDS$_{fs}$-DIM as the representative of our method, in this section we will present the attack performance evaluation of ITDS$_{gs}$-DIM and consider CFM-DIM [2] as the rival. As shown in Table 1, it can be observed that compared to the current targeted attack method CFM-DIM, ITDS$_{gs}$-DIM also demonstrates superior performance, with its BAvg on RN50 [3], VGG16 [8], MN-v3 [4] and RegN [6] being 12.7%, 13.0%, 7.9%, and 13.2% higher than the former, respectively. The maximum average black-box TASR gap in CNNs and ViTs appears on VGG16 at 17.5% and RegN at 19.9%.

3. Evaluation on Advanced Methods Combined with SDS

To demonstrate that the second part of our proposed ITDS, SDS, has more room for improvement, in this section we will show its attack performance when combined with existing advanced methods, Admix [10] and CFM, as shown in Tables 2 and 3. It can be observed that whether it is Admix and CFM or Admix-DIM and CFM-DIM, combining SDS with them results in a significant enhancement in attack performance on almost all models. In particular, the maximum improvements for Admix and CFM are 9.3% on RN50 and 5.2% on RegN, respectively, while the maximum improvements for Admix-DIM and CFM-DIM are 22.6% on RegN and 8.7%, respectively.

4. Comparison with DIM-Enhanced BSR

As we have discussed in the main text of the paper, it is unfair to directly compare competition-based methods with transformation-based methods. Therefore, the industry generally combines the former with DIM before comparing it with the latter. However, as we introduced in Section 2.2 of the main text, both ODIM [1] and SIA [11] have already integrated DIM [13] into their principles (their transformation strategies are in the same lineage as DIM), but BSR's transformation strategy style differs from theirs. Therefore, to further demonstrate the superiority of ITDS, we will perform an additional and more challenging comparative evaluation, that is, compared with the DIM-enhanced BSR, as shown in Table 4. The experimental results show that even when compared to DIM-enhanced BSR, our ITDS-DIM still has a significant advantage in terms of transferable attack performance across multiple models.

5. Ablation Studies on Other Models

In this section, we present additional ablation studies conducted on VGG16 and MN-v3, as shown in Figures 1 and 2, with all corresponding parameter settings consistent with Figure 4 in the paper. We can see that the experimental phenomena they exhibit allow us to draw analyses and conclusions that are consistent with those discussed in Section 4.4 of the paper.

6. Potentiator Effect and Computational Cost

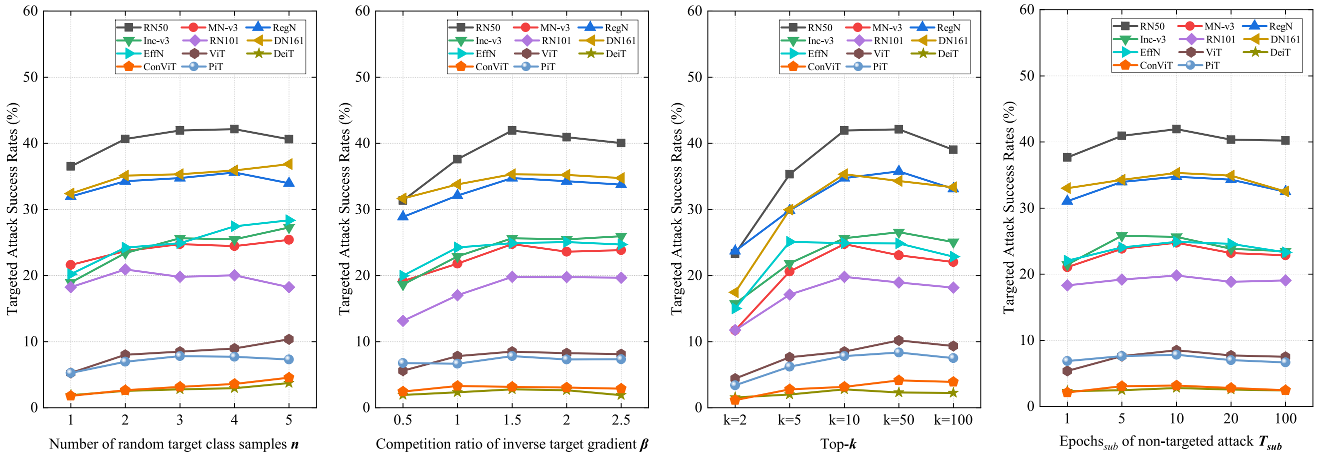
Leveraging the adaptive nature of ITDS, it can be combined with any deformation transformation-based method as a potentiator, and in the paper we only demonstrate the performance of ITDS-DIM. Using the exact experimental settings reported in the paper, Table 5 compares ITDS+BSR

Model	Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	BAvg.
RN50	CFM-DIM	99.7*	68.4	47.9	66.0	49.2	88.4	79.5	41.9	18.0	12.5	14.9	20.0	46.0
	ITDS$_{gs}$-DIM	**100***	**74.5**	**67.9**	**78.5**	**67.7**	**94.1**	**93.0**	**59.0**	**40.2**	**21.7**	**22.3**	**27.5**	**58.7**
VGG16	CFM-DIM	19.2	86.3*	4.1	23.7	3.9	5.9	14.0	6.1	0.4	0.6	0.4	1.6	7.2
	ITDS$_{gs}$-DIM	**38.4**	**100***	**24.0**	**31.7**	**25.5**	**18.1**	**33.8**	**27.5**	**8.9**	**3.2**	**3.4**	**7.2**	**20.2**
MN-v3	CFM-DIM	34.7	24.9	99.8*	31.6	29.4	33.4	27.8	44.2	28.1	11.8	14.2	11.6	26.5
	ITDS$_{gs}$-DIM	**46.5**	**27.1**	**100***	**35.8**	**37.6**	**44.2**	**40.4**	**55.0**	**46.2**	**18.1**	**15.6**	**12.5**	**34.4**
RegN	CFM-DIM	71.5	61.9	50.8	93.0*	38.3	61.0	72.3	61.6	22.5	21.7	22.1	43.3	47.9
	ITDS$_{gs}$-DIM	**78.9**	**50.1**	**66.7**	**99.9***	**63.1**	**70.2**	**81.3**	**72.3**	**55.6**	**40.8**	**41.0**	**51.8**	**61.1**

Table 1. TASRs (%) on twelve pre-trained models with ITDS$_{gs}$-DIM and CFM-DIM. The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. An asterisk (*) indicates white-box attacks. Boldface represents the results of our method.

Model	Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	Bavg.
RN50	Admix	100*	7.9	2.1	4.9	1.6	38.6	19.8	0.6	0.3	0.1	0.2	0.1	6.9
	Admix+SDS$_{fs}$	100*	18.9	10.3	16.7	8.5	63.7	50.9	3.5	2.2	0.9	0.8	1.5	16.2
	Admix+SDS$_{gs}$	100.0	18.4	10.0	17.5	8.7	65.5	52.4	3.6	1.7	0.9	0.9	1.7	16.5
	CFM	100*	45.0	17.7	35.5	9.5	87.7	60.7	5.6	2.0	1.5	1.5	3.9	24.6
	CFM+SDS$_{fs}$	100*	52.6	23.6	45.2	12.8	92.8	74.4	7.1	3.6	2.1	2.7	6.1	29.3
	CFM+SDS$_{gs}$	100*	51.4	22.5	43.9	11.9	92.0	72.9	7.8	3.1	1.7	2.4	5.4	28.6
VGG16	Admix	4.3	88.4*	1.4	7.7	0.7	0.9	4.1	0.5	0.1	0.1	0.1	0.5	1.8
	Admix+SDS$_{fs}$	8.3	100*	4.4	11.2	1.9	2.6	11.2	1.7	0.2	0.3	0.3	1.3	3.9
	Admix+SDS$_{gs}$	7.8	99.9*	4.2	11.1	1.9	2.4	11.3	1.6	0.2	0.2	0.2	0.9	3.8
	CFM	3.9	86.1*	1.2	8.4	0.8	1.0	2.8	0.6	0.1	0.2	0.1	0.3	1.7
	CFM+SDS$_{fs}$	4.9	99.8*	1.7	9.0	1.2	1.3	3.5	0.8	0.3	0.4	0.2	0.3	2.1
	CFM+SDS$_{gs}$	4.4	99.8*	1.6	9.0	1.1	1.5	3.1	0.7	0.2	0.3	0.2	0.4	2.0
MN-v3	Admix	0.6	0.4	99.3*	0.3	0.6	0.3	0.4	0.4	0.2	0.1	0.2	0.1	0.3
	Admix+SDS$_{fs}$	2.3	1.0	100*	1.5	1.9	1.3	1.9	2.2	1.4	0.3	0.7	0.6	1.3
	Admix+SDS$_{gs}$	2.1	0.9	100*	1.2	2.2	1.5	1.8	2.3	1.1	0.4	0.6	0.4	1.3
	CFM	10.4	8.9	100*	10.4	7.4	9.6	7.7	14.2	7.2	2.7	4.0	3.1	7.8
	CFM+SDS$_{fs}$	11.9	10.7	100*	11.5	9.3	11.5	10.3	15.0	7.7	4.5	5.2	4.6	9.3
	CFM+SDS$_{gs}$	11.7	10.1	100*	11.6	9.1	10.4	9.5	14.8	8.2	5.4	5.4	4.7	9.1
RegN	Admix	1.4	1.1	1.6	95.8*	0.4	0.5	1.6	0.5	0.3	0.2	0.2	0.4	0.7
	Admix+SDS$_{fs}$	12.7	8.0	9.7	100*	4.5	4.9	17.2	4.8	2.0	1.8	1.5	5.2	6.6
	Admix+SDS$_{gs}$	12.7	7.5	9.9	100*	4.4	5.8	16.8	5.3	2.0	1.9	1.2	5.3	6.6
	CFM	21.4	23.3	14.1	92.3*	4.4	12.1	19.6	9.0	1.4	1.2	1.1	5.9	10.3
	CFM+SDS$_{fs}$	32.4	31.5	21.0	100*	6.5	17.7	28.2	14.0	3.2	2.7	2.6	10.0	15.4
	CFM+SDS$_{gs}$	31.4	31.2	22.4	100*	5.8	18.2	28.8	13.7	3.4	2.7	2.5	10.1	15.5

Table 2. TASRs (%) on twelve pre-trained models using Admix and CFM combined with SDS (S1: SDS$_{fs}$ and S2: SDS$_{gs}$). The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. An asterisk (*) indicates white-box attacks.



(a) (b) (c) (d)

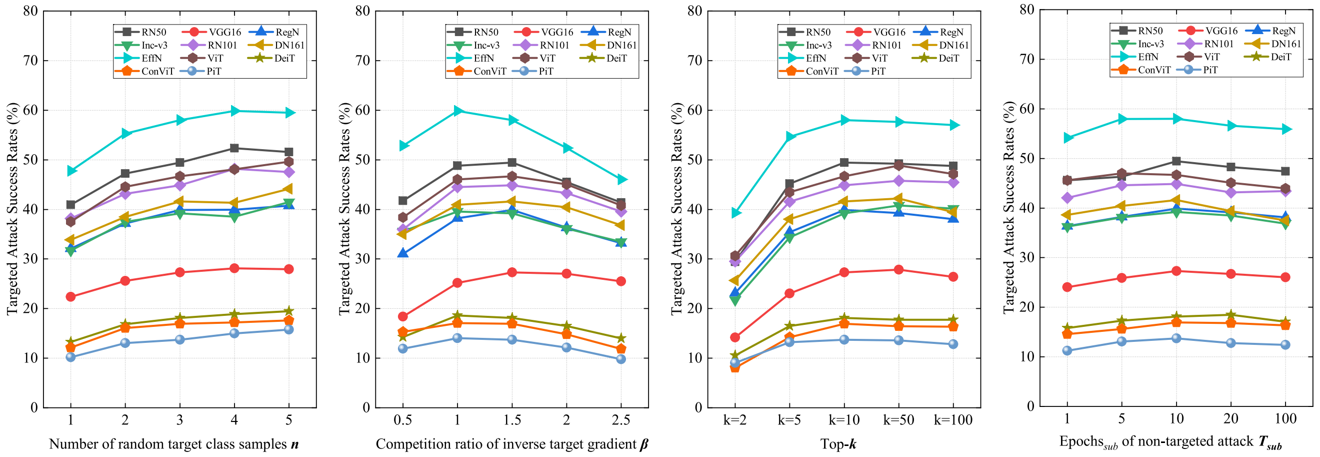
Figure 1. Ablation studies on the VGG16 model. (a) - (d): TASRs (%) on the other eleven models with the AEs crafted by ITDS-DIM, where the default values for n, β, k and T_{sub} are set to 3, 1.5, 10 and 10 respectively, when test parameters for each other.

Model	Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	Bavg.
RN50	Admix-DIM	100*	51.9	25.5	53.6	29.1	83.8	77.2	28.8	8.0	4.3	5.1	10.1	34.3
	Admix-DIM+SDS$_{fs}$	100*	60.8	47.4	68.2	56.0	92.1	91.1	50.7	24.3	12.7	15.4	19.2	48.9
	Admix-DIM+SDS$_{gs}$	100.0	61.4	47.9	69.4	55.9	91.9	90.9	52.6	24.8	11.8	14.9	20.0	49.2
	CFM-DIM	99.7*	68.4	47.9	66.0	49.2	88.4	79.5	41.9	18.0	12.5	14.9	20.0	46.0
	CFM-DIM+SDS$_{fs}$	99.9*	78.7	58.7	78.9	56.7	94.5	88.8	53.0	21.6	15.1	18.3	26.1	53.7
	CFM-DIM+SDS$_{gs}$	99.9*	79.1	59.7	78.4	57.6	94.5	89.6	53.3	22.2	15.9	17.6	27.2	54.1
VGG16	Admix-DIM	20.2	88.0*	6.8	23.9	5.7	5.4	16.0	8.0	0.8	0.5	0.5	2.1	8.2
	Admix-DIM+SDS$_{fs}$	27.4	99.9*	15.1	26.8	14.0	9.3	27.2	19.0	3.6	1.8	1.8	4.5	13.6
	Admix-DIM+SDS$_{gs}$	27.8	99.9*	14.5	27.8	13.9	9.5	28.0	18.2	3.5	1.3	1.8	4.7	13.7
	CFM-DIM	19.2	86.3*	4.1	23.7	3.9	5.9	14.0	6.1	0.4	0.5	0.4	1.6	7.2
	CFM-DIM+SDS$_{fs}$	21.3	99.6*	5.8	26.6	4.3	6.2	16.8	6.4	0.6	0.7	0.5	2.1	8.3
	CFM-DIM+SDS$_{gs}$	23.8	99.7*	5.2	28.1	4.1	6.1	16.7	6.4	0.5	0.7	0.5	2.0	8.5
MN-v3	Admix-DIM	25.0	10.4	100*	14.6	18.8	16.7	29.2	16.7	10.4	2.1	7.4	2.1	13.9
	Admix-DIM+SDS$_{fs}$	27.3	13.6	100*	19.3	25.9	21.3	31.4	41.5	26.8	9.1	9.8	8.0	21.3
	Admix-DIM+SDS$_{gs}$	25.8	12.8	99.9*	17.8	22.4	21.4	35.3	38.2	24.1	8.7	8.3	7.5	20.2
	CFM-DIM	34.7	24.9	99.8*	31.6	29.4	33.4	27.8	44.2	28.1	11.8	14.2	11.6	26.5
	CFM-DIM+SDS$_{fs}$	36.8	29.4	99.9*	33.2	32.5	35.7	31.5	48.0	31.1	14.8	15.2	13.6	29.2
	CFM-DIM+SDS$_{gs}$	39.4	30.4	100*	35.9	32.6	36.3	30.9	49.6	32.4	13.6	15.7	14.2	30.1
RegN	Admix-DIM	39.4	25.2	20.6	92.6*	14.8	23.1	47.7	33.1	7.6	7.1	7.3	22.9	22.6
	Admix-DIM+SDS$_{fs}$	65.4	37.9	45.7	100*	44.0	50.5	68.3	63.6	31.2	24.4	23.9	42.1	45.2
	Admix-DIM+SDS$_{gs}$	64.5	36.9	44.7	100*	42.4	49.0	68.8	63.9	32.7	24.2	23.9	43.2	44.9
	CFM-DIM	71.5	61.9	50.8	93.0	38.3	61.0	72.3	61.6	22.5	21.7	22.1	43.3	47.9
	CFM-DIM+SDS$_{fs}$	81.0	72.1	63.2	100*	46.0	70.0	79.4	72.8	31.5	27.2	27.5	52.2	56.6
	CFM-DIM+SDS$_{gs}$	80.9	71.7	62.1	100*	45.7	68.4	78.4	73.4	29.7	25.4	26.7	51.1	55.8

Table 3. TASRs (%) on twelve pre-trained models using Admix-DIM and CFM-DIM combined with SDS (S1: SDS$_{fs}$ and S2: SDS$_{gs}$). The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. An asterisk (*) indicates white-box attacks.

Model	Attack	RN50	VGG16	MN-v3	RegN	Inc-v3	RN101	DN161	EffN	ViT	DeiT	ConViT	PiT	BAvg.
RN50	BSR-DIM	99.9*	90.3	56.1	90.3	44.0	98.3	95.7	41.2	28.3	24.6	28.1	48.4	58.6
	ITDS-DIM	**100***	**79.3**	**70.3**	**82.8**	**69.6**	**95.6**	**94.6**	**62.0**	**44.4**	**23.6**	**24.7**	**31.6**	**61.7**
VGG16	BSR-DIM	46.1	86.9*	10.8	52.9	8.6	18.0	43.8	14.8	2.5	1.6	1.3	8.3	19.0
	ITDS-DIM	**42.0**	**100***	**24.8**	**34.8**	**25.7**	**19.8**	**35.3**	**24.9**	**8.5**	**2.8**	**3.2**	**7.8**	**20.9**
MN-v3	BSR-DIM	33.6	16.8	99.4*	25.7	18.5	23.3	27.2	36.8	23.7	10.5	12.4	12.0	21.8
	ITDS-DIM	**49.5**	**27.3**	**100***	**39.9**	**39.2**	**44.9**	**41.6**	**58.0**	**46.7**	**18.1**	**17.0**	**13.7**	**36.0**
RegN	BSR-DIM	75.1	76.4	42.7	92.3	22.3	55.6	78.6	48.4	24.3	25.6	24.8	57.2	48.2
	ITDS-DIM	**82.2**	**50.5**	**68.9**	**99.9***	**64.6**	**73.1**	**84.9**	**75.6**	**56.5**	**43.0**	**40.8**	**54.7**	**63.2**

Table 4. TASRs (%) on twelve pre-trained models using BSR-DIM and ITDS-DIM. The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. An asterisk (*) indicates white-box attacks. ITDS-DIM represents the combination of ITDS, which integrates S1, with DIM, and the boldface represents the results of this combination.



(a) (b) (c) (d)

Figure 2. Ablation studies on the MN-v3 model. (a) - (d): TASRs (%) on the other eleven models with the AEs crafted by ITDS-DIM, where the default values for n, β, k and T_{sub} are set to 3, 1.5, 10 and 10 respectively, when test parameters for each other.

Attack	RN50	VGG16	MN-v3	RegN	ViT	PiT	Bavg.
BSR	100*	88.6	49.1	86.8	18.1	41.9	56.9
ITDS-BSR	99.8*	**93.2**	**87.1**	**90.8**	**67.1**	**58.6**	**79.3**
SIA	100*	86.1	56.6	84.9	31.0	39.6	59.7
ITDS-SIA	100*	**95.4**	**89.0**	**93.2**	**71.7**	**60.4**	**81.9**

Table 5. TASRs (%) of ITDS combined with other attacks on 1000 images using RN50.

Attack	DIM	Admix	CFM	SIA	BSR	ITDS
Time (s/img)	0.39	3.3	0.42	7.2	4.5	5.7
GPU Mem (MB)	2840	14618	3188	18966	18842	2890

Table 6. Comparison of computational and memory costs on RTX3090, averaged over 2000 images (batch size = 8) on RN50.

and ITDS+SIA, revealing substantial performance gains for both combinations.

The comparison of computational costs is shown in Table 6, where all competition-based methods use DIM. Although ITDS has higher overhead than DIM and CFM, its cost remains within the hardware budget and achieves better attack results without significantly increasing space or memory. Compared to Admix, BSR and SIA, which concatenate multiple transformations in each iteration, ITDS is also more efficient.

Model	Attack	LPIPS ↓	SSIM ↑	PSNR ↑
RN50	Admix	0.35	**0.59**	**24.76**
	CFM	**0.36**	**0.59**	24.78
	ITDS$_{fs}$	0.35	0.60	24.77
	ITDS$_{gs}$	0.34	0.60	24.80
VGG16	Admix	0.32	**0.59**	24.71
	CFM	**0.34**	**0.59**	**24.64**
	ITDS$_{fs}$	0.33	0.60	24.75
	ITDS$_{gs}$	0.31	0.61	24.79
MN-v3	Admix	0.37	**0.59**	**24.56**
	CFM	**0.39**	**0.59**	24.73
	ITDS$_{fs}$	0.38	0.61	24.61
	ITDS$_{gs}$	0.38	0.61	24.66
RegN	Admix	0.32	0.59	**24.81**
	CFM	**0.33**	**0.58**	24.82
	ITDS$_{fs}$	**0.33**	0.59	24.85
	ITDS$_{gs}$	**0.33**	0.59	24.87

Table 7. Perceptual quality scores for competition-based attacks. The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. Boldface represents the worst results.

Model	Attack	LPIPS ↓	SSIM ↑	PSNR ↑
RN50	DIM	**0.38**	0.59	24.78
	ODIM	0.37	**0.58**	24.76
	SIA	0.34	0.60	24.82
	BSR	0.35	0.60	24.85
	Admix-DIM	0.37	0.59	**24.73**
	CFM-DIM	0.37	0.59	24.74
	ITDS$_{fs}$-DIM	0.36	0.61	24.78
	ITDS$_{gs}$-DIM	0.36	0.61	24.80
VGG16	DIM	0.35	0.59	24.83
	ODIM	0.35	0.59	24.80
	SIA	0.31	0.60	24.88
	BSR	0.30	0.60	24.92
	Admix-DIM	0.35	0.59	**24.77**
	CFM-DIM	**0.36**	**0.58**	24.80
	ITDS$_{fs}$-DIM	0.35	0.60	24.79
	ITDS$_{gs}$-DIM	0.35	0.60	24.81
MN-v3	DIM	**0.38**	0.60	24.72
	ODIM	0.37	**0.59**	24.71
	SIA	0.34	0.61	24.76
	BSR	0.35	0.61	24.77
	Admix-DIM	**0.38**	0.60	**24.67**
	CFM-DIM	**0.38**	**0.59**	24.71
	ITDS$_{fs}$-DIM	0.37	0.62	**24.67**
	ITDS$_{gs}$-DIM	0.36	0.62	24.70
RegN	DIM	**0.35**	0.59	24.79
	ODIM	0.34	**0.58**	**24.75**
	SIA	0.31	0.60	24.88
	BSR	0.32	0.61	24.93
	Admix-DIM	0.34	0.59	24.76
	CFM-DIM	0.34	0.59	24.77
	ITDS$_{fs}$-DIM	**0.35**	0.61	24.77
	ITDS$_{gs}$-DIM	0.34	0.61	24.79

Table 8. Perceptual quality scores for transformation-based attacks. The AEs are crafted on the RN50, VGG16, MN-v3, and RegN models, respectively. Boldface represents the worst results.

7. Perception Study

Under the same perturbation constraints, it is well known that there are significant perception differences in AEs. As a result, we conducted an ITDS perception study in conjunction with the other baseline methods. We use SSIM [12], PSNR [5] and LPIPS [14] as perceptual metrics, which assess how similar two images are in a way that corresponds to human judgment, and the results are shown in Tables 7 and 8. It should be noted that the bold quality scores in Tables 7 and 8 indicate the worst results. The experimental results show that ITDS's superior attack performance does not come at the expense of perception.

8. Visualizations on Adversarial Examples

Figure 3 displays eight randomly selected clean images along with their corresponding AEs generated by various attack methods with the target label 'cock'. Specifically, these AEs were crafted on the RN50 model using SIA, Admix-DIM, CFM-DIM, and ITDS-DIM, respectively. Notably, the crafted AEs are imperceptible to the human eye.

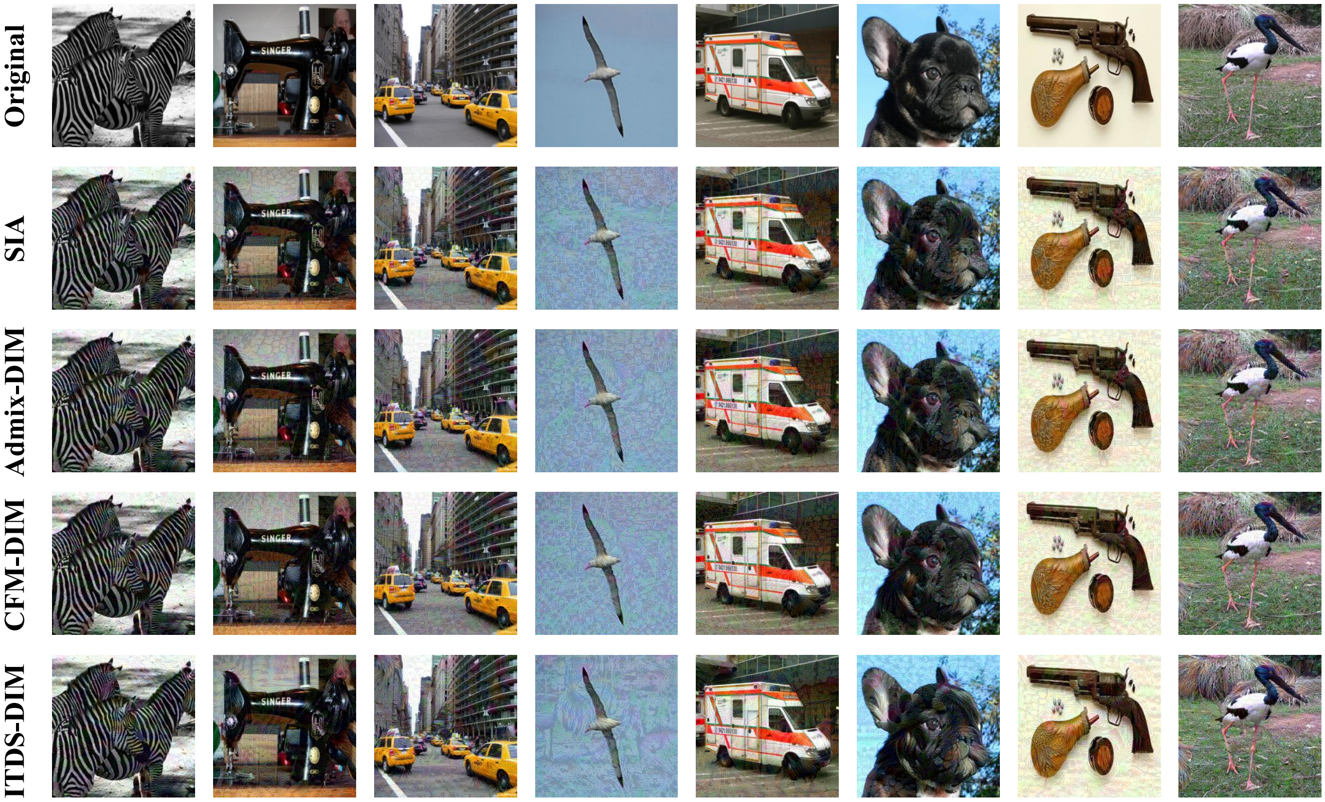


Figure 3. AEs crafted by various attack methods with the target label ‘cock’ on RN50.

9. Discuss the rigor of dataset selection

In our experimental evaluation, our original intention was to use the NIPS2017 dataset[1], but in this study, we sought to conduct a more rigorous and up-to-date evaluation of targeted adversarial attacks. To this end, we utilized a diverse and comprehensive set of pre-trained models, including 11 CNNs (three of which are adversarially trained) and 4 ViTs. Most of these models were sourced from the official torchvision pre-trained models, while others were from the corresponding open-source pytorch pre-trained versions. We randomly selected 2,000 higher-quality $3 \times 224 \times 224$-sized images from ILSVRC2012 [7] that could be correctly classified by all tested models with average confidence level of over 90%. This selection process meets the basic requirement for adversarial attacks, where the original samples must be correctly identified by the target models.

In contrast, although the NIPS2017 dataset has been widely used in adversarial attack research, it has certain limitations in terms of rigor and comprehensiveness. According to official information, pre-trained Inception-v3 and Inception-ResNet-v2 models in TF-Slim correctly classify most images from the DEV and TEST datasets. However, when we applied our comprehensive set of models to filter the NIPS2017 dataset using a set-intersection method, fewer than 400 images satisfied the criteria, highlighting its shortcomings in robustness and coverage. Therefore, to ensure a convincing and comprehensive evaluation, we decided not to use the NIPS2017 dataset and instead conducted experiments on our higher-quality dataset.

[1] https://www.kaggle.com/c/nips-2017-targeted-adversarial-attack/data.

References

[1] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15244–15253, 2022. 1

[2] Junyoung Byun, Myung-Joon Kwon, Seungju Cho, Yoonji Kim, and Changick Kim. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24648–24657, 2023. 1

[3] 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. 1

[4] Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1314–1324, 2019. 1

[5] Quan Huynh-Thu and Mohammed Ghanbari. Scope of validity of psnr in image/video quality assessment. *Electronics letters*, 44(13):800–801, 2008. 4

[6] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollár. Designing network design spaces. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10428–10436, 2020. 1

[7] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 5

[8] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[9] Kunyu Wang, Xuanran He, Wenxuan Wang, and Xiaosen Wang. Boosting adversarial transferability by block shuffle and rotation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24336–24346, 2024. 1

[10] Xiaosen Wang, Xuanran He, Jingdong Wang, and Kun He. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16158–16167, 2021. 1

[11] Xiaosen Wang, Zeliang Zhang, and Jianping Zhang. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4607–4619, 2023. 1

[12] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[13] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L Yuille. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2730–2739, 2019. 1

[14] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 4