Targeted attentional adversarial attack: supplementary material

H. Zeng, *Member, IEEE*, K. Yu, and A. Peng, *Member, IEEE*

In this supplementary document, we provide:

1) The detail about our used subset of the ImageNet (Table I).

2) Black-box success rates on more target classification models (Table II).

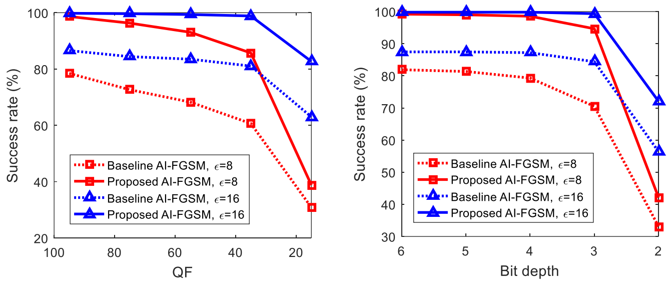
3) Comparison of attack ability (Table III) and robustness (Fig. 1) when Grad-CAM [1] is adopted for generating attentional maps. Under Grad-CAM, the proposed method has more obvious advantages than the baseline. Note the white-box success rate of baseline AI-FGSM is only 87.45% when $\epsilon = 16$. This is because the Grad-CAM method may generate an all-zero attentional map for a low-confidence label. In this case, the baseline attack is doomed to fail. On the contrary, the proposed attack is always guided with attentional maps associated with high-confidence labels, thus working better under Grad-CAM.

TABLE II
BLACK-BOX ATTACK SUCCESS RATE (%) OF TARGETED ADVERSARIAL ATTACK.

Attack\Classification model	Resnet34	Mobilenet v3_large	IncepV3
Baseline AI-FGSM, $\epsilon = 8$	8.45	5.50	3.60
Proposed AI-FGSM, $\epsilon = 8$	**10.20**	**6.65**	**4.80**
Baseline SAI-FGSM, $\epsilon = 8$	8.65	7.90	5.80
Proposed SAI-FGSM, $\epsilon = 8$	**11.55**	**11.30**	**7.70**
Baseline AI-FGSM, $\epsilon = 16$	16.70	9.05	7.90
Proposed AI-FGSM, $\epsilon = 16$	**22.60**	**13.45**	**11.70**
Baseline SAI-FGSM, $\epsilon = 16$	18.65	16.80	12.80
Proposed SAI-FGSM, $\epsilon = 16$	**25.05**	**25.80**	**19.80**

TABLE III
THE ATTACK SUCCESS RATE (%) WITH GRADCAM [1]. * INDICATES THE WHITE-BOX ATTACKS.

Attack\Classification model	Resnet18*	Resnet34	Mobilenet v3_large
Baseline AI-FGSM, $\epsilon = 8$	82.10	7.85	5.00
Proposed AI-FGSM, $\epsilon = 8$	**99.30**	**10.35**	**7.20**
Baseline AI-FGSM, $\epsilon = 16$	87.45	15.30	9.05
Proposed AI-FGSM, $\epsilon = 16$	**99.80**	**22.45**	**13.60**



(a) (b)

Fig. 1. Comparison of the attack robustness, (a) to JPEG compression, (b) to bit depth reduction.

[1] R. R. Selvaraju, M. Cogswell, A. Das, et al., "Grad-CAM: Visual explanations from deep networks via gradient-based localization," 2017 IEEE International Conference on Computer Vision, pp. 618-626

TABLE I
IMAGES USED IN THE EXPERIMENTS.

folder name	category	# training images	# test images	folder name	category	# training images	# test images
n01440764	tench	800	100	n01860187	black swan	800	100
n01530575	brambling	800	100	n01924916	flatworm	800	100
n01601694	water ouzel	800	100	n01980166	fiddler crab	800	100
n01641577	bullfrog	800	100	n02007558	flamingo	800	100
n01682714	anole	800	100	n02027492	dunlin	800	100
n01698640	alligator	800	100	n02077923	sea lion	800	100
n01740131	night snake	800	100	n02088094	Afghan	800	100
n01770081	harvestman	800	100	n02090721	wolfhound	800	100
n01795545	black grouse	800	100	n02093428	pit bull terrier	800	100
n01820546	lorikeet	800	100	n02095889	Sealyham	800	100