Robust Superpixel-Guided Attentional Adversarial Attack

Xiaoyi Dong[1], Jiangfan Han[2], Dongdong Chen[3]*, Jiayang Liu[1] , Huanyu Bian[1], Zehua Ma[1],
Hongsheng Li[2], Xiaogang Wang[2], Weiming Zhang[1], Nenghai Yu[1],
[1]University of Science and Technology of China
[2]The Chinese University of Hong Kong, [3]Microsoft Cloud AI
{dlight@, ljyljy@, hybian@, mzh045@, }mail.ustc.edu.cn
{jiangfanhan@link., hsli@ee., xgwang@ee.}cuhk.edu.hk
{zhangwm@, ynh@}ustc.edu.cn, cddlyf@gmail.com

1. Relative contributions of superpixel and attention mask.

Superpixel and attention mask are indeed leveraged for robustness and effectiveness respectively. Specifically, for the superpixel component, benefited by its local smooth property, the adversarial samples get better robustness toward steganalysis based detection and image processing based defense. For the saliency map, with its guidance, we can generate adversarial samples more effectively and get a better adversary. Compared with SAI-FGM, if we only use the superpixel part, due to a constrained space, its adversary will decrease a bit. If we only use the saliency map, it degrades to a pixel-wise perturbation with decreased robustness.

To verify it, we follow the same setting in Sec.4 and set $\delta = 8$. Here we compare SAI-FGM with I-FGM extend by the superpixel part(SI-FGM) and attentional part(AI-FGM). Besides the adversarial attack ability, we also evaluate the robustness to adversarial detection and robustness to input preprocessing method. For input preprocessing methods, we use resizing(resizing factor $\gamma = 2$) and TVM for evaluation. As shown in Table.1, we find that AI-FGM has a better adversary, but its robustness is limited. SI-FGM has better robustness, but its adversarial attack ability, especially its black-box attack ability is poor.

Attack	Inc-v3*	Inc-v4	DR	Resize	TVM
SI-FGM	99.60	36.95	**85.40**	92.48	93.30
AI-FGM	**99.90**	**49.15**	97.09	71.66	83.95
SAI-FGM	99.75	48.05	87.20	**92.51**	**94.95**

Table 1. The attack success rate (%), detection rate(DR), attack success rate after resizing and TVM.* indicates the white-box attacks.

2. More black-box attack result.

Here we show more results of our SAI-FGM on ImageNet dataset. We compare our SAI-FGM with I-FGM [4]. Adversarial samples of both methods are generated by attacking Inception-v3 [5] model with the same setting in Sec.4.1. The column "Inc-v3" is white-box attack results while another three "Res18", "Res152", "SqNet1_0" represent black-box attack results on model ResNet18 [2], ResNet152 [2], SqueezeNet1_0 [3] respectively. Due to the input size of above models is different with inception-v3, we resize the generated adversarial samples to correct size and feed them into the target black-box model.

Attack	Inc-v3*	Res18	Res152	VGG16	SqNet1_0
I-FGM	99.80	28.75	27.05	36.15	33.10
SAI-FGM	**100.00**	**67.10**	**61.45**	**68.00**	**72.65**

Table 2. The attack success rate (%) of adversarial attack on the ImageNet [1] dataset.* indicates the white-box attacks.

Results in Table 2 show that when attacking models with different architecture, our SAI-FGM still keeps a relatively high success rate. While comparing with Table.1 in the main paper, the performance of I-FGM further decreases a bit. We think such good performance is benefited by the robustness of our SAI-FGM that the resize operation almost not influence its adversary.

3. Implement Detail.

Scale operation and Crop operation. Given a threshold ϵ and a noise $\mathbf{r}$, the scale operation is continuous and realized by: $Scale_\epsilon\{\mathbf{r}\} = \mathbf{r} \cdot \frac{\epsilon}{\|\mathbf{r}\|_2}$. Similarly, with the attention map $\mathbf{m}*$ and noise $\mathbf{r}$, the crop operation is also continuous and realized by $Crop_{\mathbf{m}*}(\mathbf{r}) = \mathbf{m}^* \cdot \mathbf{r}$.

Initialization noise vector $\mathbf{n}$. $\mathbf{n}$ is randomly sampled with uniform distribution in $[-\epsilon/2, \epsilon/2]$.

*Dongdong Chen is the corresponding author.

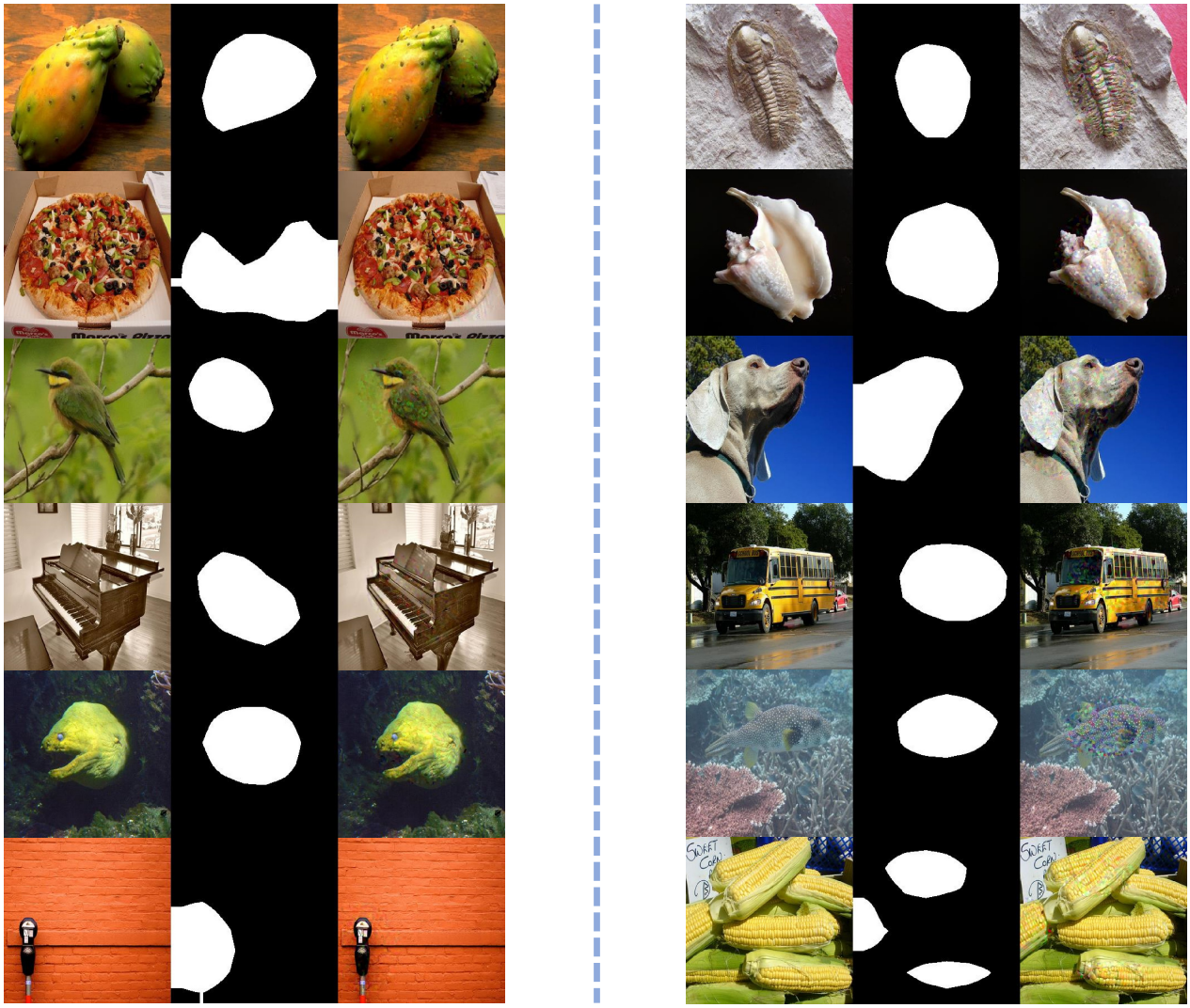

Figure 1. Some visual results about the adversarial samples generated by our SAI-FGM with $\delta = 4$ (left) and $\delta = 8$ (right) respectively. For each column, we show the original image(left), the attention mask(middle) and adversarial samples(right) respectively.

4. More visual results.

Here we show more visual results about our SAI-FGM with different perturbation. In Fig.1, we show the results with $\delta = 4$ (left) and $\delta = 8$ (right) respectively. For each column, we show the original image(left), the attention mask(middle) and adversarial samples(right) respectively.

References

[1] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016.

[3] Forrest N. Iandola, Matthew W. Moskewicz, Khalid Ashraf, Song Han, William J. Dally, and Kurt Keutzer. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <1mb model size. *CoRR*, abs/1602.07360, 2016.

[4] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv*, 2016.

[5] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.