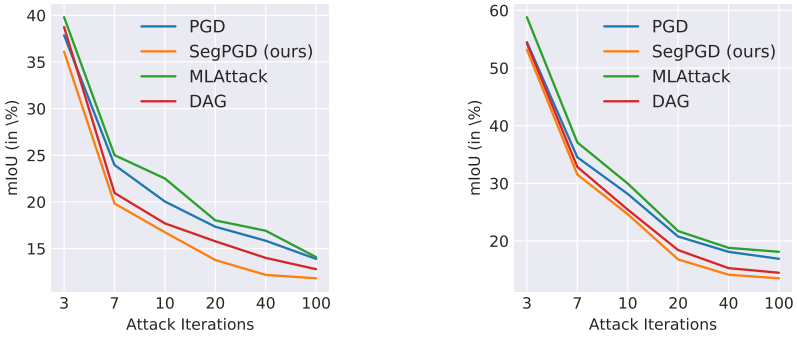
SegPGD: An Effective and Efficient Adversarial Attack for Evaluating and Boosting Segmentation Robustness

Jindong Gu, Hengshuang Zhao, Volker Tresp, and Philip Torr

Supplementary Material

A Comparison of SegPGD with other Segmentation Methods

We report the robust accuracy of adversarially trained (PGD3-AT) models under different attacks, namely, SegPGD, DAG and MLAttack. In DAG method, we apply projected gradient descent as the underlying optimization method and only focus on the correctly classified pixels. In MLAttack, three losses are considered for each input image, *i.e.,* the segmentation loss in the output layer, the of in the last layer of encoder and the MSE loss of features multiple Note that the MSE loss is computed as the MSE between the features on the clean input and the ones on current adversarial examples. For each of the three losses, the input gradients are computed to update the input examples. For fair comparison, we compare the segmentation methods with the same number of gradient propagation passes. As shown in Fig. 1, our SegPGD achieves better attack effectiveness and converges faster than other segmentation methods.



(a) PSPNet trained with PGD3-AT on VOC (b) DeepLabV3 trained with PGD3-AT on VOC

Fig. 1: Comparison of SegPGD with other Segmentation Methods. Given the same computational cost (*i.e.,*, the same number of propagation passes), our SegPGD achieves better attack effectiveness.

B Single-step Attack: SegFGSM

When a single-step attack iteration is applied, SegPGD is degraded to SegFGSM. The results under the single-step attack is shown in Tab. 1. As shown in the table, our SegFGSM outperforms FGSM on both standard models and adversarially trained models. The conclusion is true across popular segmentation model architectures on two standard segmentation datasets.

	PSPNet-VOC		DeepLabV3-VOC		PSPNet-CityScapes		DeepLabV3-CityScapes	
	Standard	AT	Standard	AT	Standard	AT	Standard	AT
Clean	76.64	74.51	77.36	75.03	73.98	71.28	73.82	71.45
FGSM	36.76	55.33	37.59	46.78	43.76	57.5	42.79	53.85
SegFGSM	**30.80**	**53.98**	**31.58**	**43.88**	**38.53**	**56.53**	**37.97**	**52.92**

Table 1: Single-step Attack. Our SegFGSM ourperforms FGSM on both standard models and adversarially trained models.

C Model Evaluation under SegPGD Attack

We evaluate adversarial trained SegPGD-AT models with our SegPGD attack method. As shown in Tab. 2, the model adversarially trained with SegPGD also outperforms the one with PGD under the SegPGD attack evaluation. In addition, the observation also echos our claim that the SegPGD can better fool segmentation models than PGD.

		AT on VOC			
		PGD3-AT	SegPGD3-AT	PGD7-AT	SegPGD7-AT
Attack Method	PGD100	13.89	14.49	16.97	19.23
	SegPGD100	9.67	10.34	16.20	17.03
		AT on Cityscapes			
		PGD3-AT	SegPGD3-AT	PGD7-AT	SegPGD7-AT
Attack Method	PGD100	3.95	13.04	22.80	23.13
	SegPGD100	1.91	8.86	17.03	22.54

Table 2: Model Evaluation under SegPGD Attack. The evaluation on SegPGD-AT PSPNet is reported with mIoU metric.

D Comparison of SegPGD-AT with DDCAT

We also compare our SegPGD-AT with the recently proposed segmentation adversarial training method DDCAT. We load the pre-trained DDCAT models from their released the codebase and evaluate the model with strong attacks. We found that their models are very weak to defend strong attacks. For fair comparison, we compare the scores on our SegPGD3-AT with the ones on their models since three steps are applied to generate adversarial examples in both case. As shown in Tab. 3, our model trained with SegPGD3-AT outperform the DDCAT by a large margin under strong attacks.

		Attack on PSPNet			Attack on DeepLabV3		
		PGD20	PGD40	PGD100	PGD20	PGD40	PGD100
AT-Models	DDCAT [?]	18.96	14.22	10.84	15.23	11.27	10.98
	SegPGD3-AT	20.69	17.19	14.49	20.92	19.10	18.24

Table 3: Comparison of SegPGD-AT with DDCAT. The SegPGD-AT model shows higher robust accuracy than DDCAT model under the same attack.

E Black-box Attack on Adversarially Trained Models

We train PSPNet and DeepLabV3 on the same dataset. Then, we create adversarial examples on PSPNet with PGD100 or SegPGD100 and test the robustness of DeepLabV3 on these adversarial examples. The results are reported in Tab. 4. We test the DeepLabV3 models trained with different methods. The model trained with our SegPGD3 shows the best performance against the transfer-based black-box attacks. The claim is also true when different attack methods are applied to create adversarial examples.

			Target Model: Deeplabv3 on VOC		
Source Model:	Training	Attack	PGD3-AT	DDCAT	**SegPGD3-AT**
PSPNet	PGD3-AT	PGD100	15.98	14.87	**16.94**
		SegPGD100	12.38	11.94	**13.43**
			Target Model: Deeplabv3 on Cityscapes		
Source Model:	Training	Attack	PGD3-AT	DDCAT	**SegPGD3-AT**
PSPNet	PGD3-AT	PGD100	14.28	15.02	**19.42**
		SegPGD100	13.32	14.26	**20.11**

Table 4: Evaluation under Black-box Attacks. The model with our SegPGD3-based adversarial training performs more robust than other methods on different datasets under different attacks.