

# PFFAA: Prototype-based Feature and Frequency Alteration Attack for Semantic Segmentation Supplementary Material

Anonymous Authors

In this Supplementary, we provide additional details, results, analysis, and qualitative results. These are not included in the main paper due to the space limitation.

## 1 ADDITIONAL EXPERIMENTS

### 1.1 Details of Training

We trained the surrogate models on the COCO 2017 dataset, with the backbone trained on ImageNet. The optimizer uses stochastic gradient descent (SGD) with a base learning rate of 0.001 and weight decay of 0.0001 in the training. In addition, the step size  $\lambda$  of our attacker is  $\epsilon/k$ , where  $k$  denotes the attack iterations and  $\epsilon$  means the perturbation budget.

### 1.2 Evaluation Metrics

For each class  $c$ ,  $IoU_c$  is defined as the ratio of the intersection area of true positives and predicted positives to the union area of these sets, given by the formula:

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (1)$$

where  $TP_c$  represents the intersection area of true positives for class  $c$ ,  $FP_c$  is the area where the model predicts positive but the ground truth is negative for class  $c$ , and  $FN_c$  is the area where the ground truth is positive but the model predicts negative for class  $c$ .

The  $mIoU$  computed by averaging the  $IoU$  values for all classes. And, the expression is:

$$mIoU = \frac{1}{|C|} \sum_{c \in C} IoU_c \quad (2)$$

where  $C$  is the set of all classes.

### 1.3 Ablation Study

**Study of the values of  $\eta$ .** To investigate the role of  $\eta$  in  $\mathcal{L}_{PFA}$ , we perform ablation experiments on  $\eta$ , and the results are presented in Table 1. When the feature-level attacker is not utilized ( $\eta = 0$ ), its influence remains limited (29.54% and 29.76% on the victim models DLV3-R50 and PSP-R50, respectively). As  $\eta$  is gradually increased, thereby modifying the intermediate features of the model, the effectiveness of the attack improves (e.g., when  $\eta = 10$ , the  $mIoU$  reduces to 22.96 on the victim DLV3-R50). However, further increments in  $\eta$  result in reduced attack effectiveness due to the weakening of the attack on predictions. Based on our experiments, we set  $\eta$  to 10.

**Study of attack results under different source domains.** Table 2 presents the attack results across various source domains. When utilizing the Pascal VOC2012 dataset as the source domain, it signifies that the surrogate model shares the same training dataset as the victim model. Notably, the FDA attack achieves its highest effectiveness (10.15%) when targeting victim model DLV3-R50 in this

**Table 1: Ablation study for different values of  $\eta$  on the Pascal VOC2012 dataset. The surrogate model is DLV3-R50, and the victims are DLV3-R50 and PSP-R50 in this experiment.**

$\eta$	Black-Box Victim Model (mIoU ↓)	
	DLV3-R50	PSP-R50
0	29.54	29.76
1	24.09	25.31
10	<b>22.96</b>	<b>23.93</b>
20	23.47	24.50

**Table 2: Experiments of attack results under different source domains. The target domain ( $D_t$ ) is Pascal VOC2012 (VOC). The surrogate model is DLV3-R50. COCO means the COCO 2017 dataset.**

$D_s$	Method	Victim Model (mIoU ↓)	
		DLV3-R50	PSP-R50
VOC	FDA	<b>10.15</b>	24.78
	PFFAA	11.66	<b>17.61</b>
ADE20k	FDA	50.05	49.33
	PFFAA	<b>34.79</b>	<b>32.89</b>
Cityscapes	FDA	60.65	59.59
	PFFAA	<b>11.66</b>	<b>13.61</b>
COCO	FDA	35.66	36.99
	PFFAA	<b>22.96</b>	<b>23.93</b>

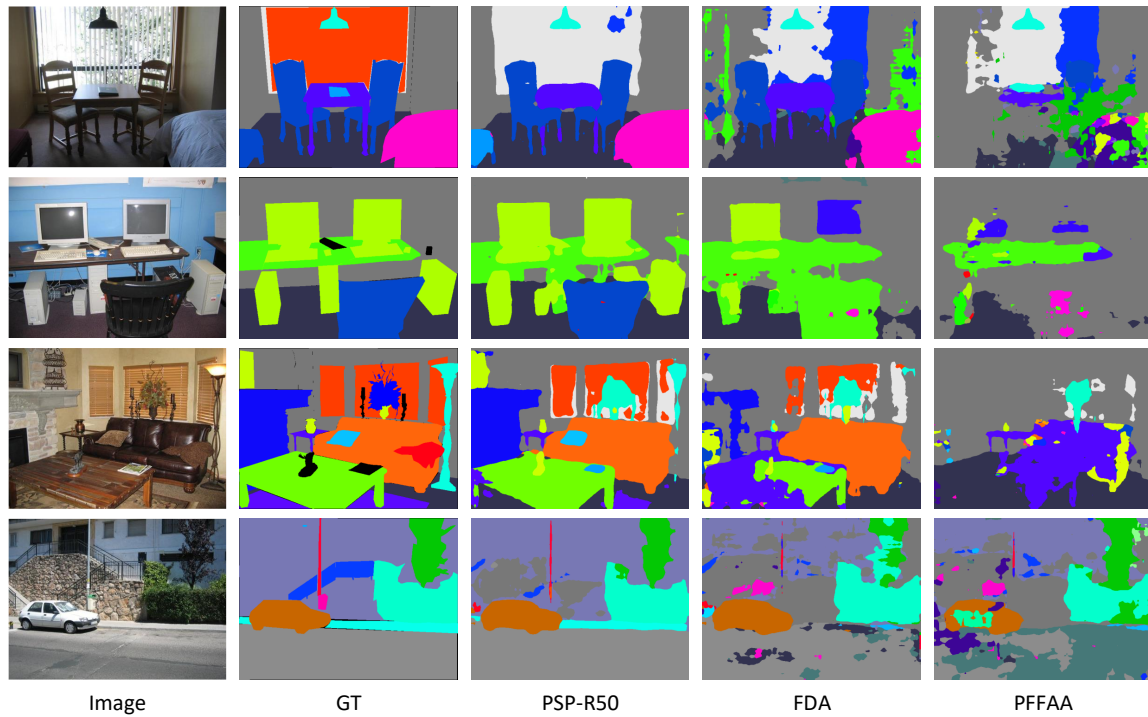
context. However, the cross-model (PSP-R50) attack exhibits a relatively limited impact (24.78%). Upon using alternative datasets as source domains, the FDA displays notably reduced efficacy against both victim models, resulting in similar attack effects. In contrast, PFFAA exploits the correlation between the source and target domains, yielding remarkably successful attack effects across multiple domains. This strategic approach significantly enhances the capability to execute successful attacks across diverse domains and models.

### 1.4 Qualitative Evaluation on Different Attackers

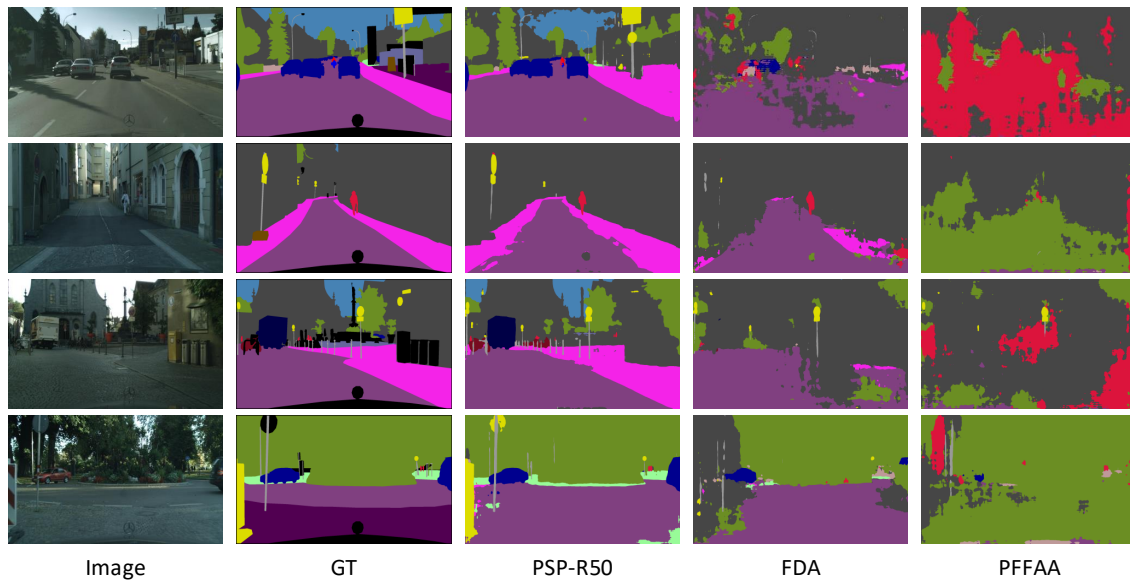
In this subsection, we show more visualization results.

**Visualization results on the ADE20k dataset.** Figure 1 shows some images and their predictions of different methods on the ADE20k dataset. On this dataset, numerous errors are present in the predictions of PSP-R50 due to the poor performance. While the FDA can significantly mislead the predictions, the PFFAA exhibits more outstanding attack performance. For example, the attack on the *table* in the second row and the *sofa* in the third row.

**Visualization results on the Cityscapes dataset.** Figure 2 illustrates some images and their predictions of different methods



**Figure 1: Visualization results of some examples on the ADE20k dataset. The victim model is PSP-R50, and the surrogate model is DLV3-R50. ‘GT’ represents the ground truth of the image, and ‘PSP-R50’ means the predictions generated by the victim model PSP-R50 for clean images. ‘FDA’ and ‘PFFAA’ denote the predictions produced by PSP-R50 for the adversarial images generated by FDA and PFFAA, respectively.**



**Figure 2: Visualization results of some images on the Cityscapes dataset. The victim model is PSP-R50, and the surrogate model is DLV3-R50.**

on the Cityscapes dataset. PSP-R50 performs well on this dataset, and FDA can only mislead some pixels in the image. It can be clearly

seen from the figure that PFFAA is effective in attacking the images on this dataset.