Supplementary Materials to ‘FPR: False Positive Rectification for Weakly Supervised Semantic Segmentation’

A. Per-class IoU on Localization Maps and Segmentation Maps

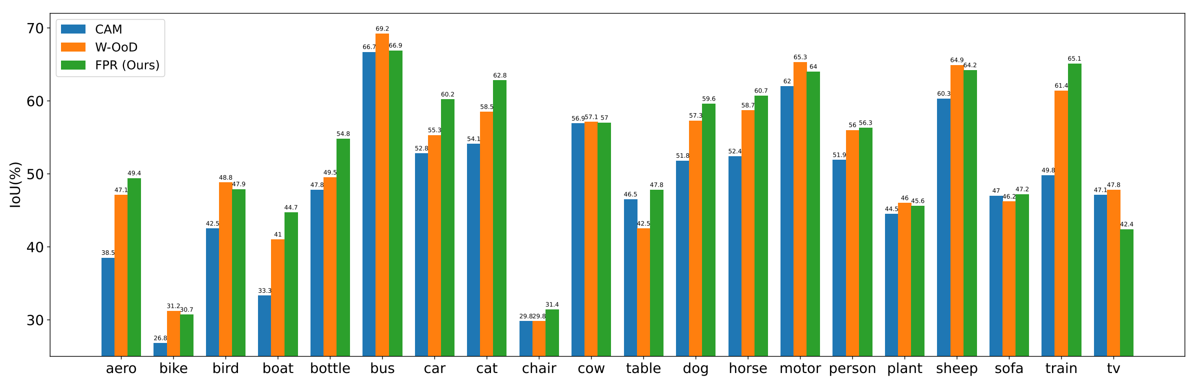

Figure A1: **Per-class IoU of localization maps**. The results are evaluated on Pascal VOC 2012 *train* set.

The class-wise improvements brought by FPR as well as W-OoD are presented in Figure A1. One can see that FPR benefits the quality of baseline CAM in almost all classes (19 out of 20), and FPR outperforms W-OoD, especially in classes heavily suffering from co-occurrence. For example, three typical co-occurred foreground classes, *i.e.*, aeroplane, boat and train, gain additional +2.3%, +3.7% and +3.7% IoU from FPR compared with W-OoD, respectively.

In addition to the qualities of localization maps, we report the per-class IoU of final segmentation maps. In particular, we generate pseudo segmentation masks by applying AdvCAM [7] to refine FPR results. The high-quality pseudo masks achieve 71.6% mIoU on Pascal VOC 2012 *train* set, and they are used to train off-the-shelf DeepLab [3] network. Following the typical setting in previous works, we evaluate the performance of DeepLab with two backbones ResNet101 [4] and WResNet38 [10]. The per-class IoU results on Pascal VOC 2012 *val* set and *test* set are shown in Table A1 and Table A2, respectively. Our FPR achieves considerable performance compared to previous works.

Table A1: **Semantic segmentation performance on PASCAL VOC 2012 *val* set.**

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mean
SEC [5]	82.4	62.9	26.4	61.6	27.6	38.1	66.6	62.7	75.2	22.1	53.5	28.3	65.8	57.8	62.3	52.5	32.5	62.6	32.1	45.4	45.3	50.7
PSA[1]	88.2	68.2	30.6	81.1	49.6	61.0	77.8	66.1	75.1	29.0	66.0	40.2	80.4	62.0	70.4	73.7	42.5	70.7	42.6	68.1	51.6	61.7
SEAM[9]	88.8	68.5	33.3	85.7	40.4	67.3	78.9	76.3	81.9	29.1	75.5	**48.1**	79.9	73.8	71.4	75.2	48.9	79.8	40.9	58.2	53.0	64.5
FickleNet[6]	89.5	76.6	32.6	74.6	51.5	71.1	83.4	74.4	83.6	24.1	73.4	47.4	78.2	74.0	68.8	73.2	47.8	79.9	37.0	57.3	**64.6**	64.9
BES[2]	89.0	73.4	30.2	81.6	50.0	63.3	87.6	80.1	85.5	28.5	82.5	46.5	76.8	78.7	76.3	71.5	42.3	82.2	37.3	61.5	52.3	65.6
AdvCAM[7]	89.5	76.9	33.5	80.3	63.7	68.6	**89.7**	77.9	87.6	31.6	77.2	36.2	82.6	78.7	73.5	69.8	51.9	81.9	43.8	70.9	52.6	67.5
W-OoD[8]	91.0	80.1	34.1	**88.1**	64.8	68.3	87.4	**84.4**	**89.8**	30.1	**87.8**	34.7	**87.5**	**85.9**	**79.8**	75.0	**56.4**	**84.5**	**47.8**	80.4	46.4	**70.7**
FPR (ResNet101)	91.4	81.8	**35.1**	82.4	**68.7**	73.7	88.8	80.5	85.9	33.3	82.4	45.3	82.5	81.6	72.9	78.5	50.7	82.6	46.5	**83.1**	49.1	70.3
FPR (WResNet38)	**91.7**	**82.5**	34.0	86.9	66.8	**75.9**	86.4	78.7	88.3	**33.6**	76.4	43.8	82.1	73.8	74.2	**79.4**	55.4	82.6	46.3	81.4	49.1	70.0

Table A2: **Semantic segmentation performance on PASCAL VOC 2012 *test* set.**

Method	bkg	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbk	person	plant	sheep	sofa	train	tv	mean
SEC [5]	83.5	56.4	28.5	64.1	23.6	46.5	70.6	58.5	71.3	23.2	54.0	28.0	68.1	62.1	70.0	55.0	38.4	58.0	39.9	38.4	48.3	51.7
PSA [1]	89.1	70.6	31.6	77.2	42.2	68.9	79.1	66.5	74.9	29.6	68.7	56.1	82.1	64.8	78.6	73.5	50.8	70.7	47.7	63.9	51.1	63.7
FickleNet[6]	90.3	77.0	35.2	76.0	54.2	64.3	76.6	76.1	80.2	25.7	68.6	50.2	74.6	71.8	78.3	69.5	53.8	76.5	41.8	70.0	**54.2**	65.3
BES[2]	89.7	76.5	30.6	78.5	51.3	64.7	**88.1**	79.3	86.9	26.9	77.8	**53.7**	78.9	79.4	78.8	72.5	51.7	81.4	51.3	55.7	51.4	66.9
AdvCAM[7]	89.3	79.3	32.5	80.2	56.3	62.8	87.2	**80.8**	87.0	28.9	78.3	41.3	82.1	80.6	77.7	68.5	51.2	80.8	55.3	60.8	48.1	67.1
W-OoD[8]	90.9	83.1	**35.6**	89.0	**61.5**	63.0	86.2	80.8	**89.9**	29.6	79.6	40.1	82.1	81.0	**82.6**	74.0	**60.1**	**85.3**	58.0	71.9	47.0	70.1
FPR (ResNet101)	91.2	**85.6**	33.0	85.2	58.9	69.2	87.0	79.6	86.3	35.4	**80.4**	47.6	**84.0**	**82.6**	77.9	76.6	50.6	80.1	**58.9**	76.6	45.4	70.1 [1]
FPR (WResNet38)	**91.7**	83.7	33.7	**89.5**	56.6	**69.8**	85.2	78.1	86.1	**36.0**	77.5	49.2	81.8	81.2	79.9	**78.3**	57.7	82.7	57.7	**77.1**	47.8	**70.6** [1]

B. Source Images of Negative Prototypes

Figure A2: **The source images of negative prototypes of train class and the corresponding logits values at different epochs.**

In the main paper, we obtain the class-specific negative prototypes from negative representation pools by sorting the representations in descending order according to predicted probability $\hat{y}_c$ and selecting the top-K representations. We investigate the source images of negative prototypes (*i.e.*, false positives) in terms of the train class and present them in Figure A2. It can be observed that the sources of negative prototypes vary across different training epochs. For example, in the first two epochs (*i.e.*, epoch 0 and epoch 1), several railroad regions and station regiones are excavated to generate negative prototypes, which facilitate the network to distinguish such co-occurred pairs. At last epoch 4, there are no railroad regions involved, and the negative prototypes tend to come from more indistinguishable regions with relatively low logits values.

[1] http://host.robots.ox.ac.uk:8080/anonymous/OBQJFV.html

References

[1] Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4981–4990, 2018. 1, 2

[2] Liyi Chen, Weiwei Wu, Chenchen Fu, Xiao Han, and Yuntao Zhang. Weakly supervised semantic segmentation with boundary exploration. In *European Conference on Computer Vision*, pages 347–362. Springer, 2020. 1, 2

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 1

[4] 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

[5] Alexander Kolesnikov and Christoph H Lampert. Seed, expand and constrain: Three principles for weakly-supervised image segmentation. In *European conference on computer vision*, pages 695–711. Springer, 2016. 1, 2

[6] Jungbeom Lee, Eunji Kim, Sungmin Lee, Jangho Lee, and Sungroh Yoon. Ficklenet: Weakly and semi-supervised semantic image segmentation using stochastic inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5267–5276, 2019. 1, 2

[7] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4071–4080, 2021. 1, 2

[8] Jungbeom Lee, Seong Joon Oh, Sangdoo Yun, Junsuk Choe, Eunji Kim, and Sungroh Yoon. Weakly supervised semantic segmentation using out-of-distribution data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16897–16906, 2022. 1, 2

[9] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12275–12284, 2020. 1

[10] Zifeng Wu, Chunhua Shen, and Anton Van Den Hengel. Wider or deeper: Revisiting the resnet model for visual recognition. *Pattern Recognition*, 90:119–133, 2019. 1