Supplementary Material for CascadePSP

1. Comparison with Multi-Scale Testing

Table 1 and Table 2 tabulate additional experimental results, including comparisons with commonly used multi-scale (MS) evaluation on the PASCAL VOC 2012 relabelled validation set [2], the BIG test set and the ADE20K [8] validation set. In multi-scale evaluation, the input image is augmented by resizing and flipping, and the algorithm averages the predictions for all the augmented images as the final output. We use the multi-scale option provided by the author whenever possible and follow the evaluation method as mentioned in the paper.

On the PASCAL VOC 2012 and BIG dataset, our model performs better than the multi-scale evaluation method, especially in terms of boundary accuracy. The ADE20K dataset has 150 classes and requires much stronger semantic information to solve the class confusion problem, thus multi-scale evaluation is more effective. Nevertheless, we can further refine the multi-scale evaluation output to produce an even better prediction with accurate boundaries.

Methods	IoU (%)	mBA (%)
PASCAL VOC 2012		
RefineNet [4]	86.21	62.61
(+) Ours	$87.48_{\uparrow 1.27}$	$\mathbf{71.34_{\uparrow 8.73}}$
(+) MS	$86.62_{\uparrow 0.41}$	$60.56_{\downarrow 2.05}$
(+) MS (+) Ours	$\mathbf{88.25_{\uparrow 2.04}}$	$70.87_{\uparrow 8.26}$
DeepLabV3+ [1]	87.13	61.68
(+) Ours	$89.01_{\uparrow 1.88}$	$72.10_{\uparrow 10.4}$
(+) MS	$88.39_{\uparrow 1.26}$	$62.65_{\uparrow 0.97}$
(+) MS (+) Ours	$\mathbf{89.77_{\uparrow 2.64}}$	$\mathbf{72.33_{\uparrow 10.7}}$
PSPNet [7]	90.92	60.51
(+) Ours	$92.86_{\uparrow 1.94}$	$72.24_{\uparrow 11.7}$
(+) MS	$91.70_{\uparrow 0.78}$	$61.89_{\uparrow 1.38}$
(+) MS (+) Ours	$\mathbf{93.22_{\uparrow 2.30}}$	$\mathbf{72.95_{\uparrow 12.4}}$
BIG		
RefineNet [4]	90.20	62.03
(+) Ours	$\mathbf{92.79_{\uparrow 2.59}}$	$\mathbf{74.77_{\uparrow 12.7}}$
(+) MS	$89.60_{\downarrow 0.60}$	$60.30_{\downarrow 1.73}$
(+) MS (+) Ours	$92.30_{\uparrow 2.10}$	$74.08_{\uparrow 12.1}$
DeepLabV3+ [1]	89.42	60.25
(+) Ours	$92.23_{\uparrow 2.81}$	$\mathbf{74.59_{\uparrow 14.3}}$
(+) MS	$89.94_{\uparrow 0.52}$	$60.87_{\uparrow 0.62}$
(+) MS (+) Ours	$\mathbf{92.38_{\uparrow 2.96}}$	$74.55_{\uparrow 14.3}$
PSPNet [7]	90.49	59.63
(+) Ours	$93.93_{\uparrow 3.44}$	$75.32_{\uparrow 15.7}$
(+) MS	$91.62_{\uparrow 1.13}$	$61.01_{\uparrow 1.38}$
(+) MS (+) Ours	$\mathbf{94.58_{\uparrow 4.09}}$	$\mathbf{75.75_{\uparrow 16.1}}$

Table 1. Extended comparison between different semantic segmentation methods with and without our refinement. MS: Multi-scale evaluation. MS+Ours: Result generated by our model with MS prediction as inputs.

2. Class attenuation on the ADE20K dataset

As described in the paper, we attenuate the scores for all the "stuff" classes as provided by [8][1]. There are 35 such stuff classes. The attenuation constant is set to 0.51 such that background classes can only suppress low confidence foreground classes with a very high confidence score in the fusion function and the competitions among background classes are not affected.

Methods	mIoU (%)	mBA (%)
ADE20K		
RefineNet [4]	41.47	55.60
(+) Ours	$42.20_{\uparrow 0.73}$	$56.67_{\uparrow 1.07}$
(+) MS	$42.37_{\uparrow 0.90}$	$55.60_{-0.00}$
(+) MS (+) Ours	$\mathbf{43.06_{\uparrow 1.59}}$	$\mathbf{57.03_{\uparrow 1.43}}$
EncNet [6]	42.20	55.29
(+) Ours	$43.19_{\uparrow 0.99}$	$57.29_{\uparrow 2.00}$
(+) MS	$44.31_{\uparrow 2.11}$	$57.66_{\uparrow 2.37}$
(+) MS (+) Ours	$\mathbf{45.01_{\uparrow 2.81}}$	$\mathbf{58.63_{\uparrow 3.34}}$
PSPNet [7]	43.10	57.03
(+) Ours	$43.83_{\uparrow 0.73}$	$58.13_{\uparrow 1.10}$
(+) MS	$44.15_{\uparrow 1.05}$	$57.97_{\uparrow 0.94}$
(+) MS (+) Ours	$\mathbf{44.65_{\uparrow 1.55}}$	$\mathbf{58.76_{\uparrow 1.73}}$

Table 2. Comparison between different methods with and without our refinement on the ADE20K validation set. MS: Multi-scale evaluation. MS+Ours: Result generated by our model with MS prediction as inputs. The performance of multi-scale EncNet [6] is lower than that of reported in their paper (44.65%) but aligns with their open-sourced code.

3. Comparison with DenseCRF

Here, we compare our result with DenseCRF [3] which is a popular method for low-level segmentation refinement. We performed grid search on the BIG validation set with DeeplabV3+ inputs to determine the hyperparameters as suggested by the original paper. Learnable parameters that encode the relationship between object classes (e.g., "bird" class is likely to be found in "sky" class) are neglected as we are working with binary class-agnostic segmentations.

Table 3 tabulates the quantitative result and Figure 1 shows visual comparison. Our method shows better performance across different settings as DenseCRF lacks semantic understanding. In addition, applying DenseCRF as a post-processing step to our model does not show any improvement across all grid search parameters in the validation set.

[1] https://github.com/CSAILVision/sceneparsing/blob/master/objectInfo150.csv

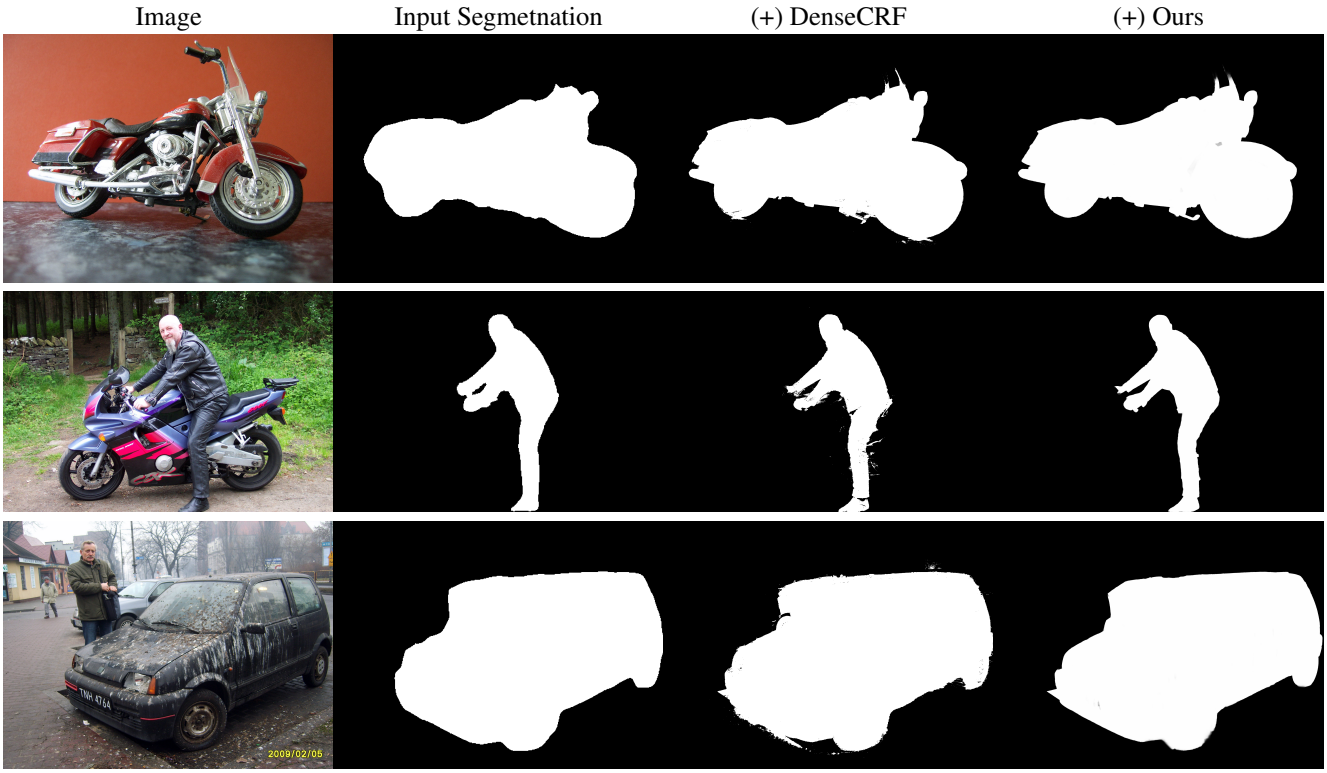

Figure 1. Visual comparisons between DenseCRF and our method. Although DenseCRF adheres well to low-level boundaries, it produces noisy results. Zoom in for more details.

Methods	IoU (%)	mBA (%)
FCN-8s [5]	72.39	53.63
(+) DenseCRF	$74.33_{\uparrow 1.94}$	$60.76_{\uparrow 7.13}$
(+) Ours	$\mathbf{77.87}_{\uparrow \mathbf{5.48}}$	$\mathbf{67.04}_{\uparrow \mathbf{13.4}}$
RefineNet [4]	90.20	62.03
(+) DenseCRF	$91.22_{\uparrow 1.02}$	$69.66_{\uparrow 7.63}$
(+) Ours	$\mathbf{92.79}_{\uparrow \mathbf{2.59}}$	$\mathbf{74.77}_{\uparrow \mathbf{12.7}}$
DeepLabV3+ [1]	89.42	60.25
(+) DenseCRF	$91.04_{\uparrow 1.62}$	$69.56_{\uparrow 9.31}$
(+) Ours	$\mathbf{92.23}_{\uparrow \mathbf{2.81}}$	$\mathbf{74.59}_{\uparrow \mathbf{14.3}}$
PSPNet [7]	90.49	59.63
(+) DenseCRF	$91.22_{\uparrow 0.73}$	$69.66_{\uparrow 10.0}$
(+) Ours	$\mathbf{93.93}_{\uparrow \mathbf{3.44}}$	$\mathbf{75.32}_{\uparrow \mathbf{15.7}}$

Table 3. Refinement comparison between DenseCRF and CascadePSP. CascadePSP shows better result in all settings.

References

[1] Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*, 2018. 1, 2

[2] Mark Everingham, S. M. Ali Eslami, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge – a retrospective. In *IJCV*, 2014. 1

[3] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NIPS*, 2011. 1

[4] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, 2017. 1, 2

[5] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 2

[6] Hang Zhang, Kristin Dana, Jianping Shi, Zhongyue Zhang, Xiaogang Wang, Ambrish Tyagi, and Amit Agrawal. Context encoding for semantic segmentation. In *CVPR*, 2018. 1

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1, 2

[8] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *CVPR*, 2017. 1