Supplementary Materials for Asymmetric Non-local Neural Networks for Semantic Segmentation

A. Quantitative comparisons on COCO-Stuff-10K and NYUD-V2

Following the training and evaluation protocols of DeepLab-V2 [1] and RefineNet [5], our method achieves competitive results on the two datasets using single scale whole image testing, as shown in Tab. 1. As NYUD-V2 [7] is a small benchmark and COCO-Stuff-10K is quite challenging and large, these results further verify the effectiveness of our method on both small and large benchmarks.

Method	Backbone	mIoU (%)	Method	Backbone	mIoU (%)
RefineNet [5]	ResNet-101	33.6	Piecewise [6]	VGG16	40.6
CCL [4]	ResNet-101	35.7	RefineNet [5]	ResNet-101	43.6
Ours	ResNet-101	37.2	Ours	ResNet-101	44.4

Table 1: Comparisons on COCO-Stuff-10K (Left) and NYUD-V2 (Right) datasets. Results of the competing methods are taken from their papers.

B. More ablation results

Qualitative comparisons. We also give the qualitative comparisons of our Full (+ AFNB + APNB) method with other variants of our model in Fig. 1. In summary, our Full method shows the best semantic consistency and the least inconsistency artifacts while +AFNB and +APNB fail in some cases. The results also indicate AFNB and APNB is complementary to each other and the combination of them is beneficial to improve the performance.

Selection of the fusing layers. Fusing features from multi-levels is effective in many computer vision tasks. However, it still requires a lot of trials to find a good combination of the fusing layers. For semantic segmentation, the last several layers of the network contain plenty of features with semantic information, which is critical for better performance. Hence, we only combine the features from the shallow layers to the deep layers in a top-down manner. The responses are summed up if a certain layer receives more than two fusion invitations. The results are listed in Tab. 2. An obvious conclusion is that fusing only the features of Stage4 and Stage5 brings a considerable improvement while keep fusing more layers will only hurt the performance.

We conclude a possible intuitive reason: features from the early stages are generic to most tasks, while the last two are task-specific. Therefore, merging the features of the last several stages is more effective for a specific task. In the supplementary materials, we compare the feature visualizations of the outputs of all 5 stages of network trained on segmentation benchmark Cityscapes [2] and of that trained on classification benchmark ImageNet [3], using the same backbone network ResNet-101 to demonstrate our guess.

As can be seen in Fig. 2, we compare the feature visualizations of the outputs of all 5 stages of network trained on segmentation benchmark Cityscapes [2] (**Lower**) and of that trained on classification benchmark ImageNet [3] (**Upper**), using the same backbone network ResNet-101. Comparing the two networks' feature visualizations of the same stage, the features are quite similar in the first three stages while differs hugely in the last two. This observation accord with our guess. Note our experiment results partially conforms to the conclusions in ExFuse [8], further validating the effectiveness of fusing only the last two stages.

References

[1] 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 Trans. Pattern Anal. Mach. Intell.*, 40(4):834–848, 2018. 1

[2] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. CVPR*, pages 3213–3223, 2016. 1, 2

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Fei-Fei Li. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255, 2009. 1, 2

[4] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multi-scale

Method	Fusing layers	mIoU (%)
Baseline	–	75.8
AFNB	4 & 5	77.1
AFNB	3 & 5, 4 & 5	76.7
AFNB	2 & 5, 3 & 5, 4 & 5	76.2

Table 2: Ablation study on the validation set of Cityscapes in terms of the selection of layers to be fused. "4 & 5" means fusing the features of Stage4 and Stage5. Others likewise.

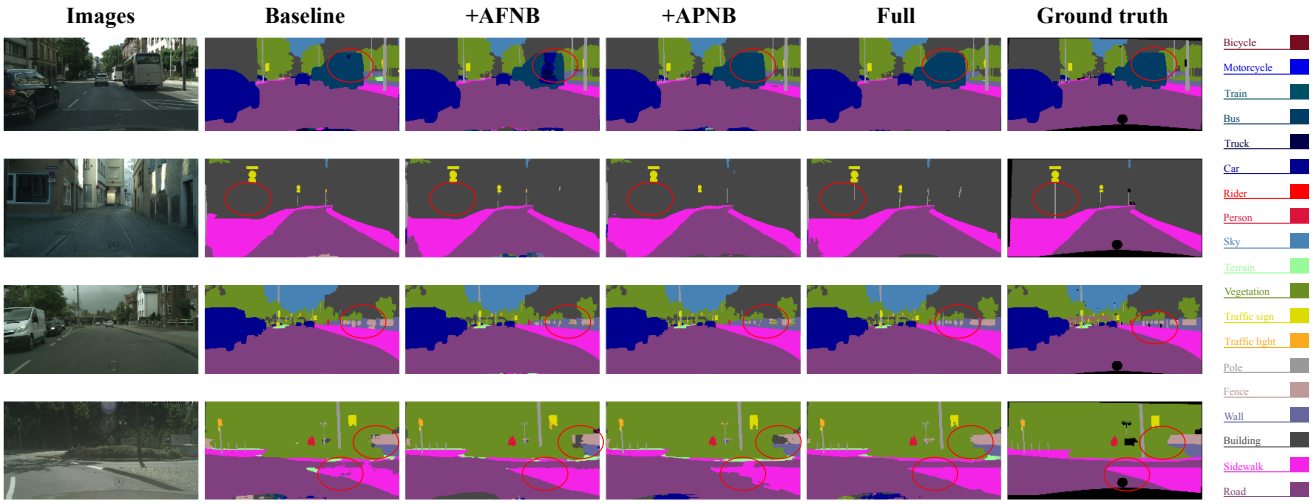

Figure 1: Qualitative comparisons among Full model and other variants of our model. The red circles indicate where Full model is superior to other model variants.

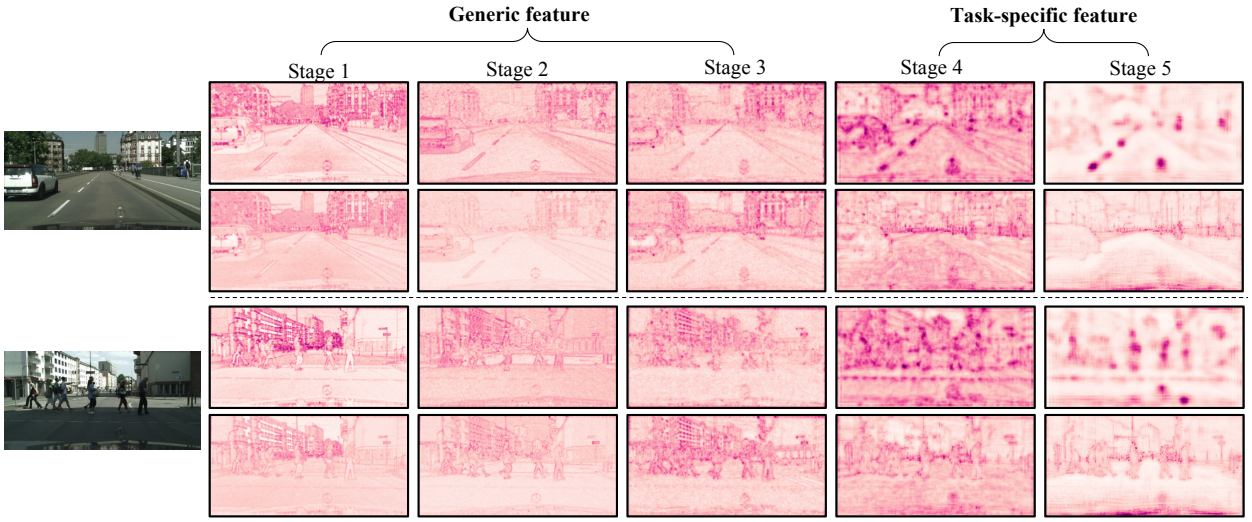

Figure 2: Feature visualization of different stages on ResNet-101. The **Upper** row represents visualizations from network trained on classification dataset ImageNet [3]. The **Lower** row represents visualizations from network trained on scene segmentation dataset Cityscapes [2].

aggregation for scene segmentation. In *Proc. CVPR*, pages 2393–2402, 2018. 1

[5] Guosheng Lin, Anton Milan, Chunhua Shen, and Ian D. Reid. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *Proc. CVPR*, pages 5168–5177, 2017. 1

[6] Guosheng Lin, Chunhua Shen, Anton van den Hengel, and Ian D. Reid. Efficient piecewise training of deep structured models for semantic segmentation. In *Proc. CVPR*, pages 3194–3203, 2016. 1

[7] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Proc. ECCV*, pages 746–760, 2012. 1

[8] Zhenli Zhang, Xiangyu Zhang, Chao Peng, Xiangyang Xue, and Jian Sun. Exfuse: Enhancing feature fusion for semantic segmentation. In *Proc. CVPR*, pages 273–288, 2018. 1