# Supplementary Material

## 1 Experiments

In addition to the experiments outlined in the main section, we evaluate our approach on PASCAL Context [4] and CityScapes [2].

**PASCAL-Context.** This dataset comprises 60 semantic labels (including background), and consists of 4998 training images, and 5105 validation images. During training, we divide the learning rate by half twice after 50 epochs and after 100 epochs, respectively, and keep training until 200 epochs, or earlier convergence. We do not pre-train on PASCAL VOC or COCO.

Our quantitative results are provided in Table 1 and qualitative results are on Figure 1.

**CityScapes.** Finally, we turn our attention to the CityScapes dataset [2], that contains 5000 high-resolution ($1024 \times 2048$) images with 19 semantic classes, of which 2975 images are used for training, 500 for validation, and 1525 for testing, respectively. We use the same learning strategy as for Context. Our single-scale model with ResNet-101 as backbone, is able to achieve 72.1% mean iou on the test set, which is close to the original RefineNet result of 73.6% with multi-scale evaluation [3].

Visual results are presented on Figure 2.

| Model | mIoU,% |
|---|---|
| DeepLab-v2-CRF [1] | 45.7 (*msc*) |
| RefineNet-101 [3] | 47.1 (*msc*) |
| RefineNet-152 [3] | 47.3 (*msc*) |
| **RefineNet-LW-101** (ours) | 45.1 |
| **RefineNet-LW-152** (ours) | 45.8 |

Table 1: Quantitative results on the test set of PASCAL Context. Multi-scale evaluation is defined as *msc*.

**PASCAL Person-Part.** Please refer to the main text for quantitative outputs. We provide qualitative results on Figure 3.

**NYUD.** Please refer to the main text for quantitative outputs. We provide qualitative results on Figure 4.

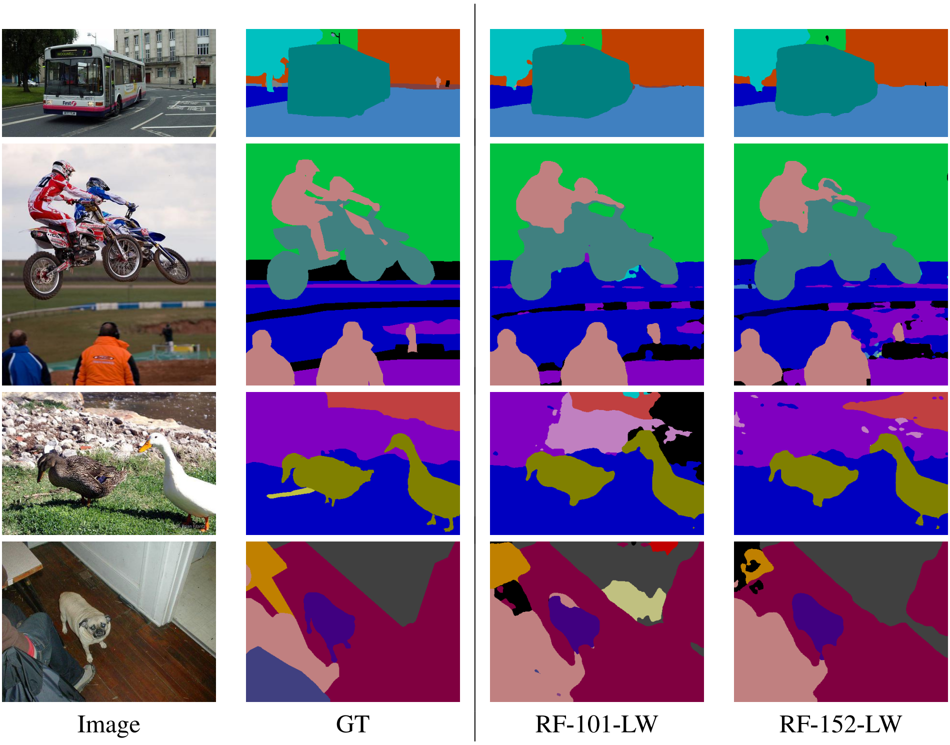


Figure 1: Visual results on validation set of PASCAL-Context with residual models.

# 2 Receptive field size

To quantify why dropping $3 \times 3$ convolutions does not result in significant performance drop, we consider the issue of the empirical receptive field (ERF) size [5]. Intuitively, dropping $3 \times 3$ convolutions should significantly harm the receptive field size of the original architecture. Nevertheless, we note that we do not experience this due to i) the skip-design structure of RefineNet, where low-level features are being summed up with the high-level ones, and ii) keeping CRP blocks.

Additionally to the results in the main text, we compare ERF of the last (classification) layer between original RefineNet-101 and RefineNet-LW-101, both been pre-trained on PASCAL VOC. From Figure 5, it can be noticed that both networks exhibit semantically similar activation contours, although the original architecture tends to produce less jagged boundaries.

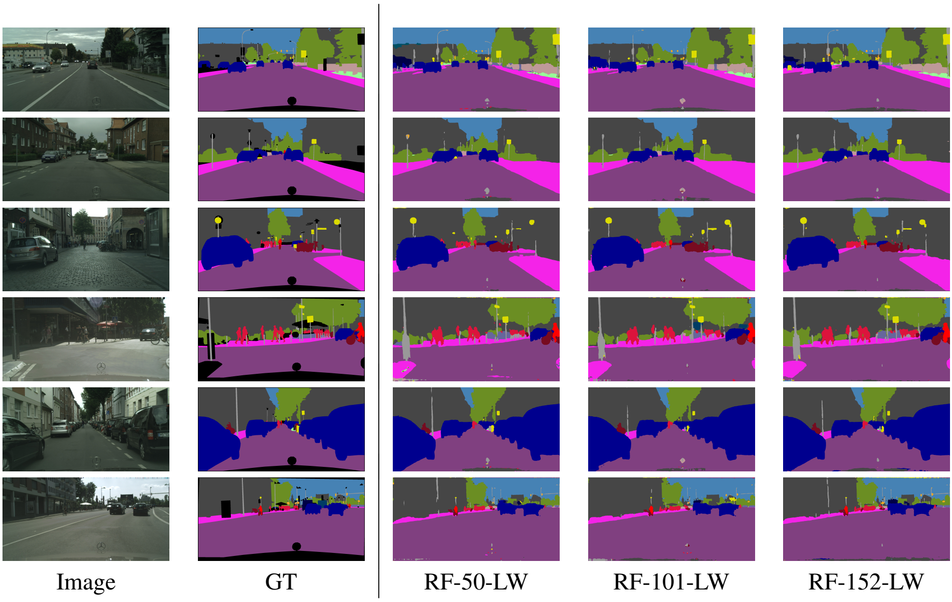


Figure 2: Visual results on validation set of CityScapes with residual models.

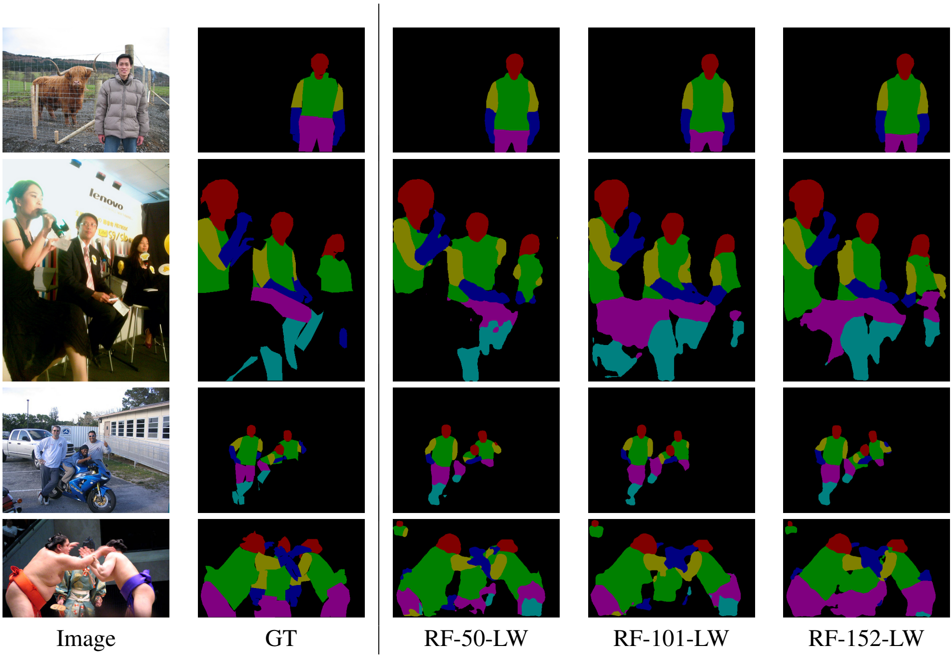


Figure 3: Visual results on validation set of PASCAL Person-Part with residual models.

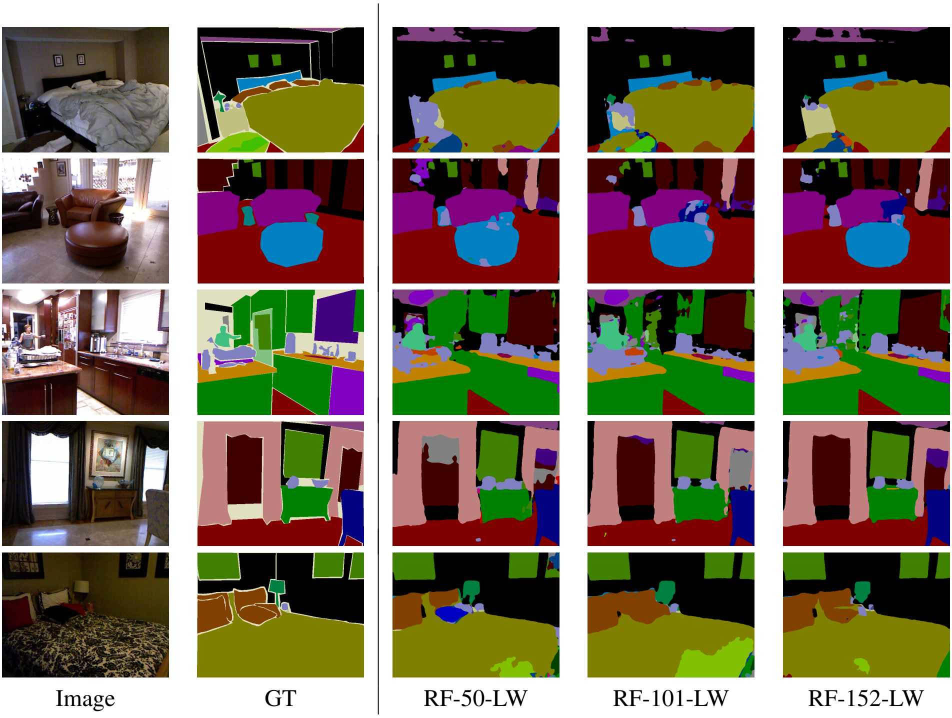


Figure 4: Visual results on validation set of NYUDv2 with residual models.

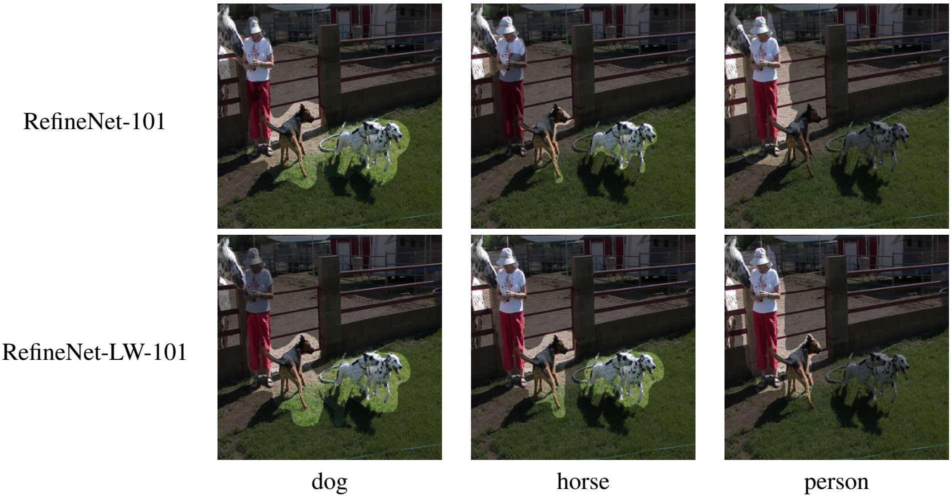


Figure 5: Comparison of empirical receptive field in the last (classification) layer between RefineNet-101 (top) and RefineNet-LW-101 (bottom). Top activated regions for each unit are shown.

# 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. *CoRR*, abs/1606.00915, 2016.

[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 *CVPR*, 2016.

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

[4] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan L. Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014.

[5] Bolei Zhou, Aditya Khosla, Àgata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *CoRR*, abs/1412.6856, 2014.