Supplementary Material: FoveaNet: Perspective-aware Urban Scene Parsing

1. FCN Structure Details

We make use of a fully convolutional network (FCN) [4] as a baseline model for parsing the scene images. We follow Chen *et al.* [1] and use the vanilla ResNet-101 [3] to initialize the FCN model. Preserving high spatial resolution of feature maps is very important for accurately segmenting small objects. Therefore, we disable the last down-sampling layer by setting its stride as 1. This increases size of the feature maps output by res5_3b3 to $1/16$ of the input image size (without this modification the size of output feature maps is only $1/32$ of the input image size). The dilation factor of convolution kernels in the following residual blocks (from res5_a to 5_c) is set to 2, effectively enlarging the field-of-view (FoV) of filters therein. In order to distinguish neighboring pixels well for semantic parsing, we remove the top pooling layer in ResNet-101 considering pooling operation would unfavorably "smooth" features of neighboring pixels. We add a convolutional *score layer* on top of the FCN model which outputs pixel-level dense category prediction for the input image. The score layer has a convolutional kernel size of 5, and has a convolutional stride of 16 pixels. Such configuration may lead to blurred details in its up-sampled output prediction. To further enhance quality of the prediction, we follow Long *et al.* [4] and add skip connections between the score layer and following three bottom layers: res3_b3, pool1 and conv1. We add a 1×1 convolution layer on top of each of these bottom layers that produces three additional predictions. These predictions are then fused with $2\times$, $4\times$ and $8\times$ up-sampling of score layer output respectively, and give the final parsing prediction. The overall structure of our baseline FCN model is illustrated in Figure 1.

2. Qualitative Comparison with Single Image Depth Estimation

Garg *et al.* [2] train a convolutional encoder for the task of predicting the depth map from a single image, and gives a state-of-the-art performance on the KITTI dataset. We make use of their model to predict the depth map on the Cityscapes dataset, and give a qualitative comparison between the prediction results and our estimated perspective heatmap in Figure 2. The depth prediction has a low distinguishability for distant objects, and does not provide enough detail fro localizing small scale objects. In contrast, the perspective heatmap predicted with our PEN can produce a much more detailed prediction.

3. Qualitative result on CamVid dataset

We also give qualitative results in Figure 3. From the results, one can observe that PEN localizes the desired fovea region containing small objects accurately. Benefiting from this, FoveaNet parses the small cars distant from the camera accurately. Also, perspective-aware CRF demonstrates ability on preserving the integretity of the large near object (the truck shown in the bottom row). The qualitative results validate the effectiveness of FoveaNet on CamVid dataset.

References

[1] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *arXiv:1606.00915*, 2016. 1

[2] R. Garg, G. Carneiro, and I. Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *ECCV*, 2016. 1, 2

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. *arXiv preprint arXiv:1512.03385*, 2015. 1

[4] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015. 1

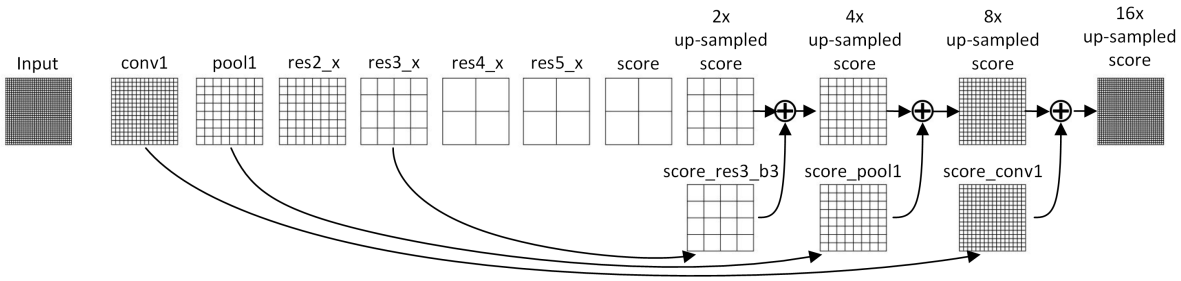

Figure 1. Architectural overview of baseline FCN model. PEN has a similar network structure as the FCN (see text). Given an input scene image, PEN produces a one channel heatmap indicating (roughly) nearness to the vanishing point at pixel-level.

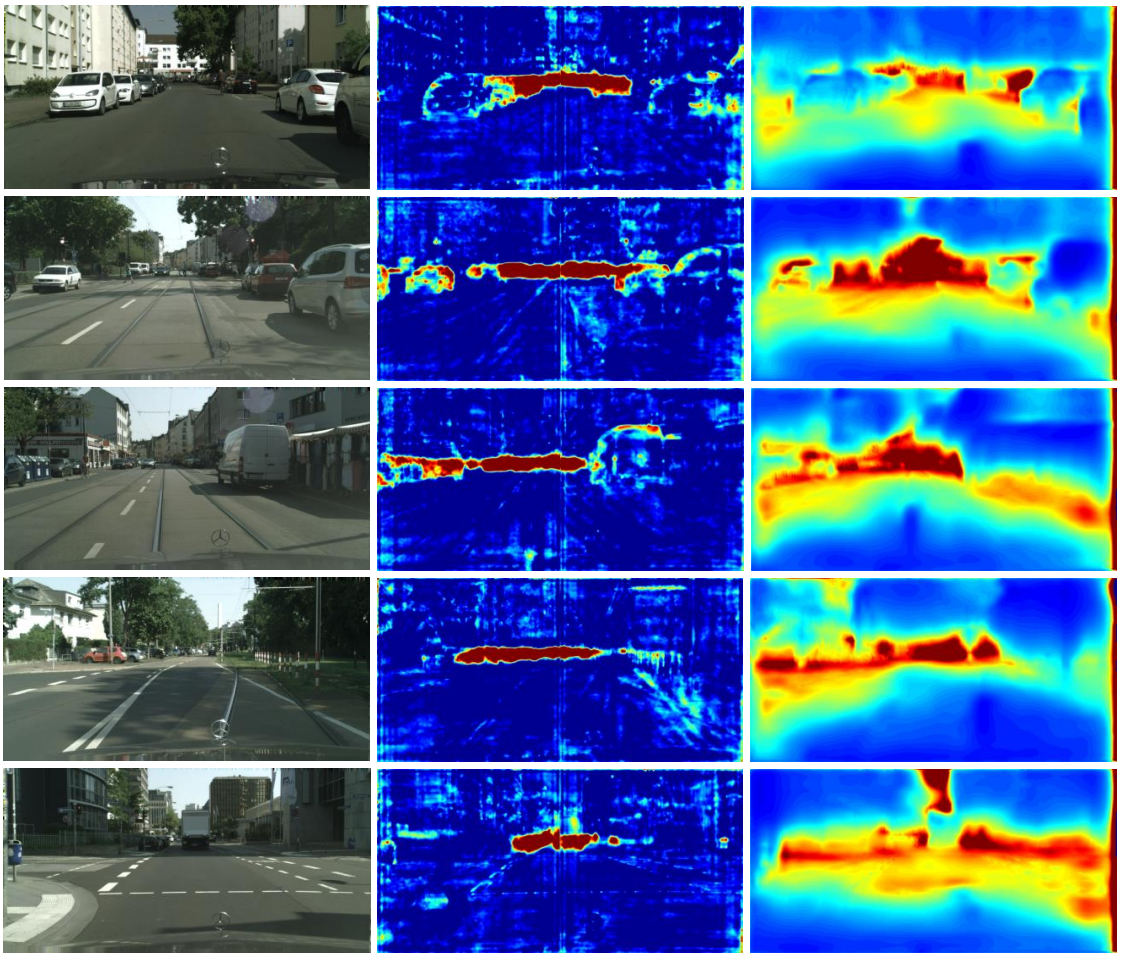

Figure 2. Comparison between our perspective heatmap prediction and depth prediction. Left: urban scene images. Middle: perspective heatmap prediction. Right: depth prediction from single image [2]. (Best viewed in color).

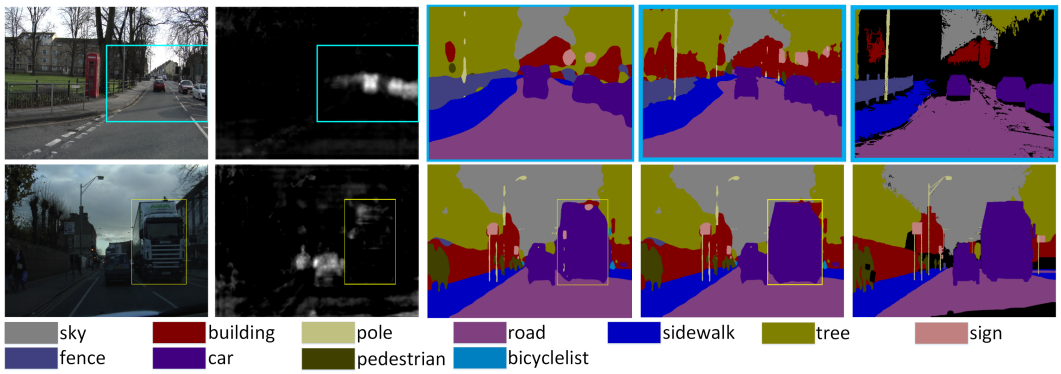

Figure 3. Example parsing results on Camvid dataset. Top: from left to right, input image, estimated perspective heatmap by PEN, corresponding fovea region (cyan rectangle), parsing result of baseline FCN on fovea region, FoveaNet parsing result on fovea region. FoveaNet effectively locates the distant objects, *e.g.*, the pole, and produce parsing results with well-preserved details. Bottom: from left to right, input image with bounding box (yellow rectangle) from detector, estimated perspective heatmap from PEN, predictions before perspective-aware CRF, prediction after perspective-aware CRFs. FoveaNet removes "broken-down" errors by the new CRF. (Best viewed in color).