Variational Context-Deformable ConvNets for Indoor Scene Parsing
Supplementary Material

Zhitong Xiong, Yuan Yuan*, Nianhui Guo, Qi Wang
School of Computer Science and Center for OPTical IMagery Analysis and Learning (OPTIMAL),
Northwestern Polytechnical University, Xi'an, Shaanxi, P. R. China
{xiongzhitong, y.yuan1.ieee, guonianhui199512, crabwq}@gmail.com

1. Detailed Network Architectures

VGG16, HRNetV2 and ResNet50 are employed as the backbone of the proposed method. For VGG16 backbone, the last two blocks, i.e., conv4_1, conv4_2 and conv4_3 and conv5_1, conv5_2 and conv5_3, are replaced with the proposed VCD modules. As for the HRNetV2 backbone, we replace the first branch of stage 4 (4 blocks) with VCD modules. When it comes to ResNet backbone, e.g. ResNet 50, the last two stages (conv4_x and conv5_x) are replaced with VCD modules. It is noteworthy that only convolutions with kernel size larger than 3×3 are replaced in all the backbone architectures. Thus, there are 25 VCD layers in ResNet101 backbone and 8 VCD layers in ResNet50 network.

2. Comparisons with existing methods

To comprehensively evaluate the effect of the proposed VCD module, we have also conducted experiments on Cityscapes dataset in addition to the RGB-D datasets.

As pointed by Cityscapes benchmark[1], the global IoU metric is biased toward large-scale object instances, and it can be problematic in street scenes with strong scale variation. Thus, we also employ the instance-level intersection-over-union metric (**iIoU**) for comparison. The results evaluated on the benchmark server are presented in Table 1. It can be clearly seen that the proposed method can improve the performance by a large gain on iIoU class and iIoU category metric. This indicates that the proposed context-deformable module is effective for handling the scale-variation problem. The VCD module can learn to focus on object instances with small scale, and this makes the segmentation results finer than the baseline method. The detailed comparisons with other state-of-the-art methods on Cityscapes dataset are presented in Table 2. From the results we can see that the proposed method can achieve better results on 13 out of 19 categories than other methods. Especially for the 'train' category, the proposed VCD method can largely improve the IoU from 79.9% to 87.8%. Some qualitative segmentation results on Cityscapes *test* are displayed in Fig. 4.

[1] https://www.cityscapes-dataset.com/benchmarks/

Table 1. Comparisons with state-of-the-art Methods on Cityscapes

Methods	IoU class	iIoU class	IoU category	iIoU category
DANet[1]	81.5	62.3	91.6	82.6
TKCN[8]	79.5	61.3	91.1	81.5
HRNetV2[6]	81.8	61.2	92.2	82.1
GFF[4]	**82.3**	62.1	92.0	81.4
DGCNet[12]	82.0	61.7	91.8	81.1
OCNet[11]	81.2	61.3	91.6	81.1
Ours(VCD)	**82.3**	**64.2**	**92.3**	**83.2**

3. Visualization

We visualize the qualitative segmentation results on NYUv2 RGB-D dataset in Fig. 1. Compared with the baseline method ACNet, the proposed method can obtain better segmentation results with the adaptive spatial context. For each side, the input RGB images are displayed at the first column. The segmentation results of ACNet are shown in the second column. The results of the proposed method are shown at the third column, and the ground truth labels are presented at the last column.

As the scale-guidance map g_σ is modeled as distributions rather than deterministic values, we also visualize the variance of g_σ in Fig. 2. For each side, the input images are displayed at the first column, and the variance maps are shown at the second column. From the figure we can see that the variances are large at the boundary of objects or the complicated sub-scenes. These results are reasonable, since object boundaries are more difficult to assign appropriate spatial-context.

Since the proposed VCD module can be integrated with DCN to enhance the deformation of the spatial context, we also visualize the sampling locations of the learned deformable filters. As illustrated in Fig. 3, red points represent the sampling locations for the activation unit (green point). It can be clearly seen that the spatial context for

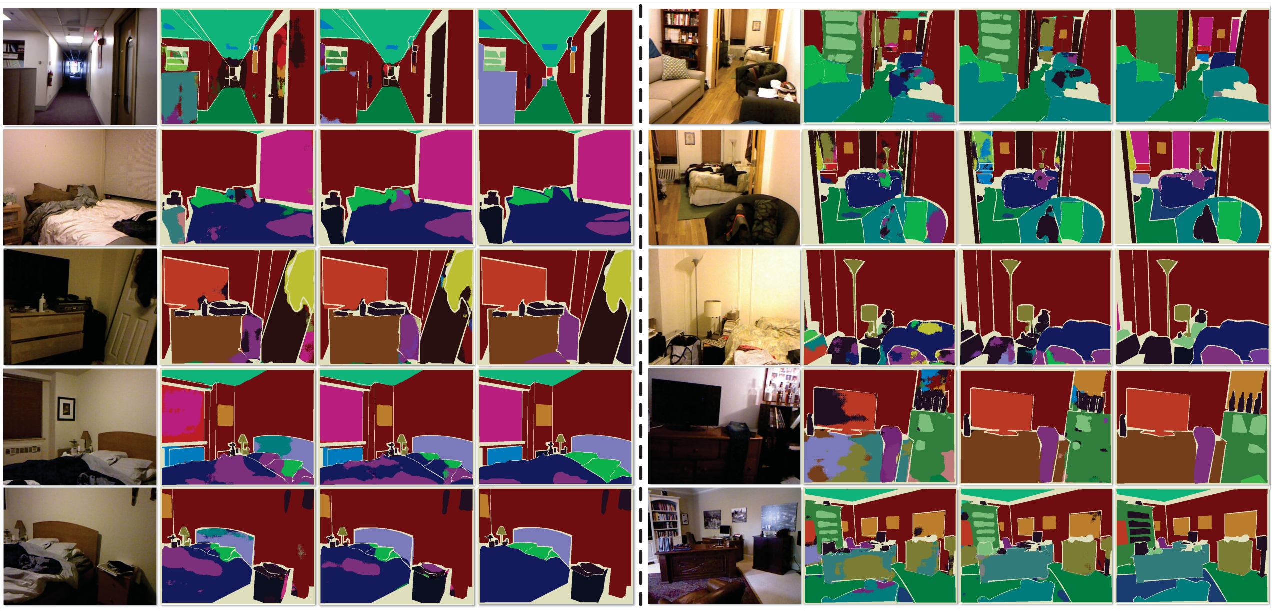

Figure 1. For each side, the input RGB images are displayed at the first column. The segmentation results of ACNet [2] are shown in the second column. The results of the proposed method are shown at the third column, and the ground truth labels are presented at the last column.

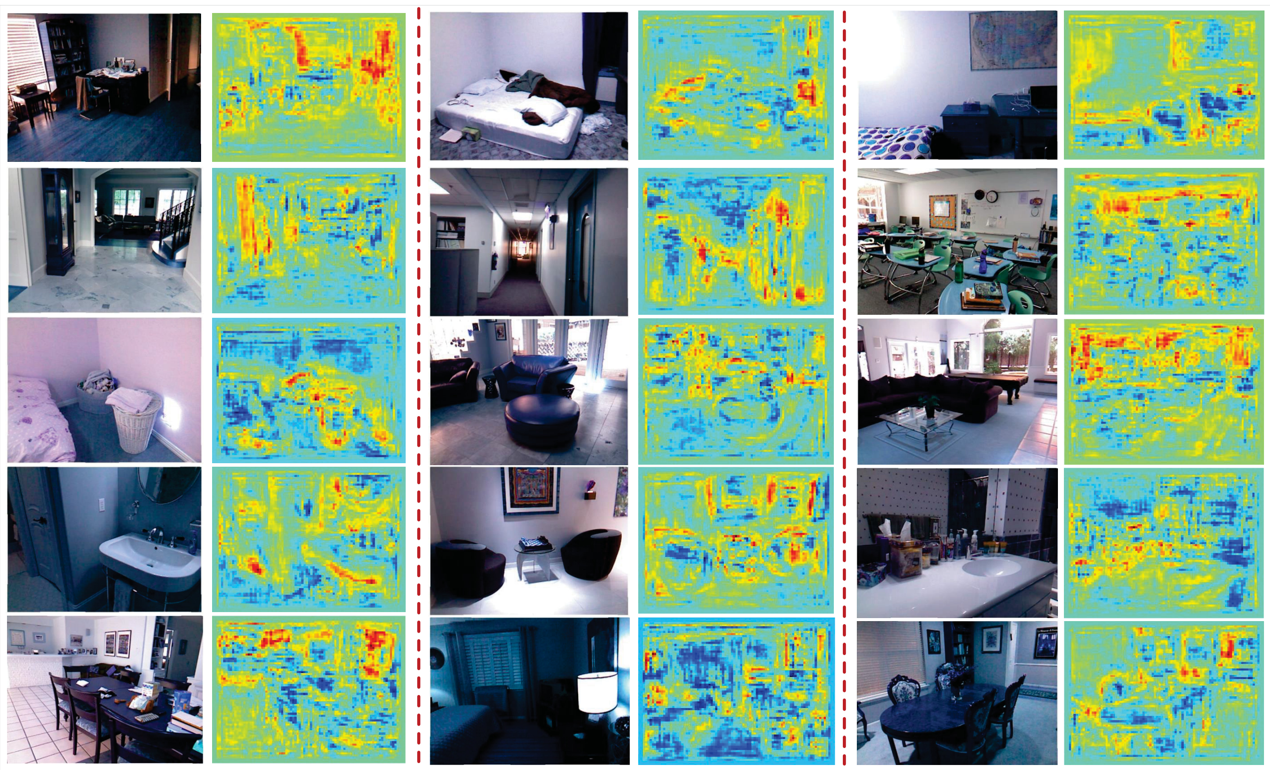

Figure 2. For each side, the input images are displayed at the first column, and the variance maps are shown at the second column. From the figure we can see that the variances are large at the boundary of objects or the complicated sub-scenes.

large and small objects are adaptive to object scale with the guidance of the depth modality and image content.

Table 2. Detailed Comparisons with state-of-the-art Methods on Cityscapes *test* set

Method	road	swalk	build	wall	fence	pole	tlight	sign	veg.	terrain	sky	person	rider	car	truck	bus	train	mbike	bike	mIoU
DUC[7]	98.5	85.5	92.8	58.6	55.5	65.0	73.5	77.8	93.2	72.0	95.2	84.8	68.5	95.4	70.9	78.7	68.7	65.9	73.8	77.6
ResNet38[9]	98.5	85.7	93.0	55.5	59.1	67.1	74.8	78.7	93.7	72.6	95.5	86.6	69.2	95.7	64.5	78.8	74.1	69.0	76.7	78.4
PSPNet[13]	98.6	86.2	92.9	50.8	58.8	64.0	75.6	79.0	93.4	72.3	95.4	86.5	71.3	95.9	68.2	79.5	73.8	69.5	77.2	78.4
AAF[3]	98.5	85.6	93.0	53.8	58.9	65.9	75.0	78.4	93.7	72.4	95.6	86.4	70.5	95.9	73.9	82.7	76.9	68.7	76.4	79.1
SegModel[5]	98.6	86.4	92.8	52.4	59.7	59.6	72.5	78.3	93.3	72.8	95.5	85.4	70.1	95.6	75.4	84.1	75.1	68.7	75.0	78.5
DenseASPP[10]	98.7	87.1	93.4	60.7	62.7	65.6	74.6	78.5	93.6	72.5	95.4	86.2	71.9	96.0	78.0	90.3	80.7	69.7	76.8	80.6
DANet[1]	98.6	87.1	93.5	56.1	63.3	69.7	77.3	81.3	93.9	72.9	95.7	87.3	72.9	96.2	76.8	89.4	86.5	72.2	78.2	81.5
HRNetV2[6]	**98.8**	87.8	**93.9**	**61.3**	63.0	72.1	79.3	82.4	94.0	**73.4**	**96.0**	88.5	75.1	96.5	72.5	88.1	79.9	73.1	79.2	81.8
GFF[4]	98.7	87.2	**93.9**	59.6	**64.3**	71.5	78.3	82.2	94.0	72.6	95.9	88.2	73.9	96.5	**79.8**	92.2	84.7	71.5	78.8	**82.3**
Ours(VCD)	**98.8**	**88.0**	93.8	56.9	61.9	**72.9**	**80.0**	**82.6**	**94.1**	73.0	95.9	**88.6**	**76.1**	**96.5**	75.5	**88.6**	**87.8**	**73.4**	**79.6**	**82.3**

Figure 3. The sampling points of the learned deformable filters. (a) The input images; (b) The corresponding depth images. (c) The spatial context for different pixels.

References

[1] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3146–3154, 2019.

[2] Xinxin Hu, Kailun Yang, Lei Fei, and Kaiwei Wang. Acnet: Attention based network to exploit complementary features for RGBD semantic segmentation. *CoRR*, abs/1905.10089, 2019.

[3] Tsung-Wei Ke, Jyh-Jing Hwang, Ziwei Liu, and Stella X. Yu. Adaptive affinity fields for semantic segmentation. In *ECCV*, pages 605–621, 2018.

[4] Xiangtai Li, Houlong Zhao, Lei Han, Yunhai Tong, and Kuiyuan Yang. GFF: gated fully fusion for semantic seg-

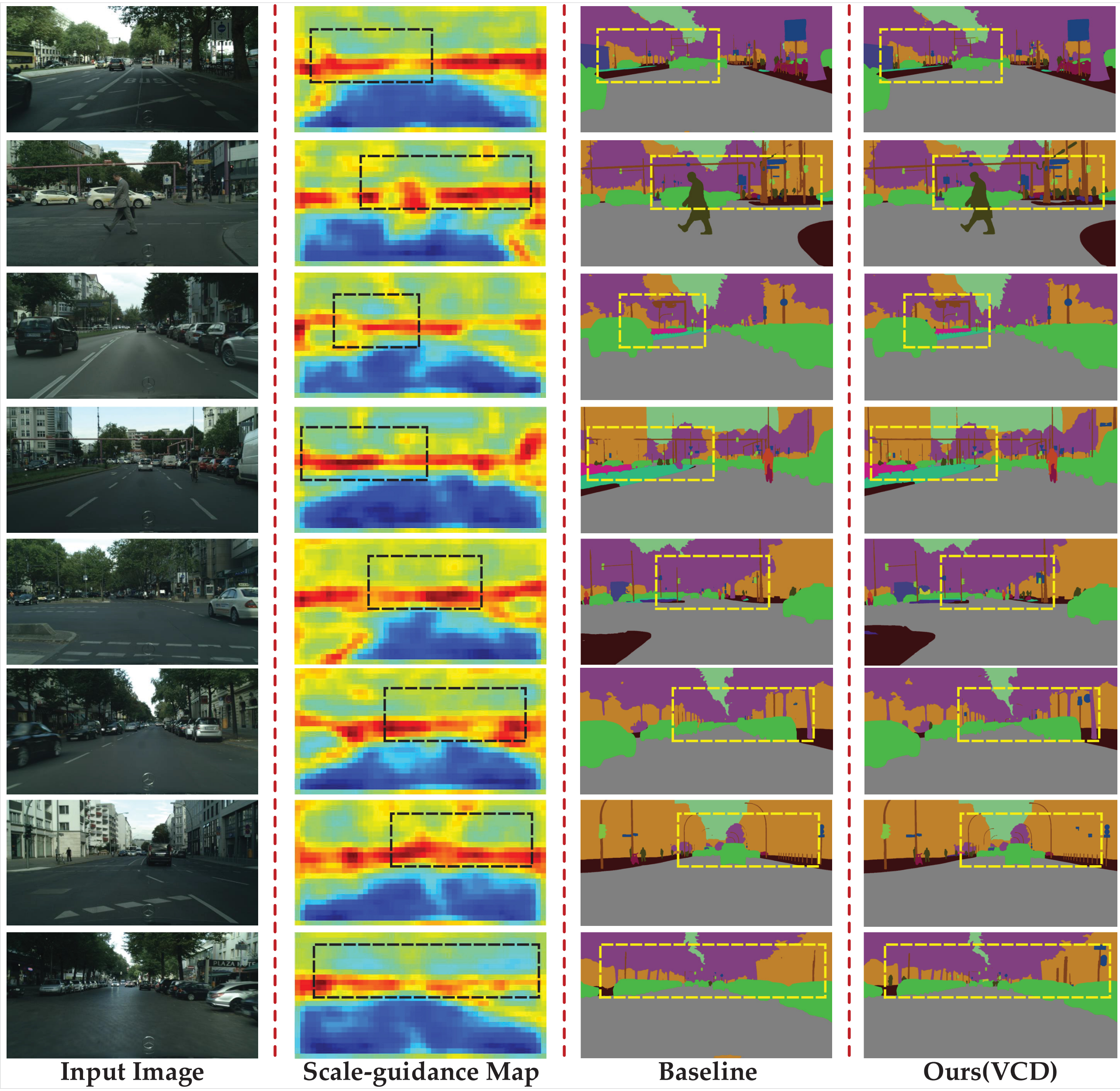

Figure 4. The qualitative segmentation results on Cityscapes *test* set. The input images are shown at the first column, and the learned scale-guidance maps are shown at the second column. The segmentation examples of the baseline method (HRNetV2) are displayed at the third column. The results of our method are presented at the last column.

mentation. *CoRR*, abs/1904.01803, 2019.

[5] Falong Shen, Rui Gan, Shuicheng Yan, and Gang Zeng. Semantic segmentation via structured patch prediction, context CRF and guidance CRF. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 5178–5186, 2017.

[6] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5693–5703, 2019.

[7] Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison W. Cottrell. Understanding convolution for semantic segmentation. In *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*, pages 1451–1460, 2018.

[8] Tianyi Wu, Sheng Tang, Rui Zhang, Juan Cao, and Jintao Li. Tree-structured kronecker convolutional network for se-

mantic segmentation. In *IEEE International Conference on Multimedia and Expo, ICME 2019, Shanghai, China, July 8-12, 2019*, pages 940–945, 2019.

[9] 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.

[10] Maoke Yang, Kun Yu, Chi Zhang, Zhiwei Li, and Kuiyuan Yang. Denseaspp for semantic segmentation in street scenes. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3684–3692, 2018.

[11] Yuhui Yuan and Jingdong Wang. Ocnet: Object context network for scene parsing. *CoRR*, abs/1809.00916, 2018.

[12] Li Zhang, Xiangtai Li, Anurag Arnab, Kuiyuan Yang, Yunhai Tong, and Philip H. S. Torr. Dual graph convolutional network for semantic segmentation. *CoRR*, abs/1909.06121, 2019.

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