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 $3 \times 3$ 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 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.

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 $g_\sigma$ is modeled as distributions rather than deterministic values, we also visualize the variance of $g_\sigma$ 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

---

\(^1\)https://www.cityscapes-dataset.com/benchmarks/
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

<table>
<thead>
<tr>
<th>Method</th>
<th>road</th>
<th>walk</th>
<th>build</th>
<th>wall</th>
<th>fence</th>
<th>pole</th>
<th>light</th>
<th>sign</th>
<th>sky</th>
<th>person</th>
<th>rider</th>
<th>car</th>
<th>truck</th>
<th>bus</th>
<th>train</th>
<th>mbike</th>
<th>bike</th>
<th>mIoU</th>
</tr>
</thead>
<tbody>
<tr>
<td>DUC[7]</td>
<td>98.5</td>
<td>85.5</td>
<td>92.8</td>
<td>58.6</td>
<td>55.5</td>
<td>65.0</td>
<td>73.5</td>
<td>77.8</td>
<td>91.2</td>
<td>72.0</td>
<td>95.2</td>
<td>84.8</td>
<td>68.3</td>
<td>95.4</td>
<td>70.9</td>
<td>68.7</td>
<td>65.9</td>
<td>73.8</td>
</tr>
<tr>
<td>ResNet38[9]</td>
<td>98.5</td>
<td>85.7</td>
<td>93.0</td>
<td>55.5</td>
<td>59.1</td>
<td>67.1</td>
<td>74.8</td>
<td>78.7</td>
<td>93.7</td>
<td>72.6</td>
<td>95.5</td>
<td>86.6</td>
<td>69.2</td>
<td>95.7</td>
<td>64.5</td>
<td>78.8</td>
<td>74.1</td>
<td>69.0</td>
</tr>
<tr>
<td>PSPNet[13]</td>
<td>98.6</td>
<td>86.2</td>
<td>92.9</td>
<td>50.8</td>
<td>58.8</td>
<td>84.0</td>
<td>75.6</td>
<td>79.0</td>
<td>93.4</td>
<td>72.3</td>
<td>95.4</td>
<td>86.3</td>
<td>71.3</td>
<td>95.9</td>
<td>68.2</td>
<td>79.5</td>
<td>73.8</td>
<td>69.5</td>
</tr>
<tr>
<td>AAF[3]</td>
<td>98.5</td>
<td>85.6</td>
<td>93.0</td>
<td>53.8</td>
<td>58.9</td>
<td>55.9</td>
<td>55.0</td>
<td>73.4</td>
<td>91.7</td>
<td>72.4</td>
<td>95.6</td>
<td>86.4</td>
<td>70.3</td>
<td>95.9</td>
<td>73.9</td>
<td>82.7</td>
<td>76.2</td>
<td>68.7</td>
</tr>
<tr>
<td>SegModel[5]</td>
<td>98.6</td>
<td>86.4</td>
<td>92.8</td>
<td>52.4</td>
<td>59.7</td>
<td>59.6</td>
<td>72.3</td>
<td>78.3</td>
<td>91.3</td>
<td>72.8</td>
<td>95.5</td>
<td>85.4</td>
<td>70.1</td>
<td>95.6</td>
<td>75.4</td>
<td>84.1</td>
<td>75.1</td>
<td>68.7</td>
</tr>
<tr>
<td>DenseASPP[10]</td>
<td>98.7</td>
<td>87.1</td>
<td>93.4</td>
<td>60.7</td>
<td>62.7</td>
<td>65.6</td>
<td>74.6</td>
<td>78.5</td>
<td>93.6</td>
<td>72.5</td>
<td>95.4</td>
<td>86.4</td>
<td>71.9</td>
<td>96.0</td>
<td>78.0</td>
<td>90.3</td>
<td>80.7</td>
<td>69.7</td>
</tr>
<tr>
<td>DANet[1]</td>
<td>98.6</td>
<td>87.1</td>
<td>93.5</td>
<td>56.1</td>
<td>63.3</td>
<td>69.7</td>
<td>77.3</td>
<td>81.3</td>
<td>93.9</td>
<td>72.9</td>
<td>95.7</td>
<td>87.3</td>
<td>72.9</td>
<td>96.2</td>
<td>76.8</td>
<td>89.4</td>
<td>86.5</td>
<td>72.2</td>
</tr>
<tr>
<td>HRNetV2[6]</td>
<td>98.8</td>
<td>87.8</td>
<td>93.0</td>
<td>61.3</td>
<td>63.0</td>
<td>72.1</td>
<td>79.3</td>
<td>82.4</td>
<td>94.0</td>
<td>73.4</td>
<td>96.0</td>
<td>88.5</td>
<td>75.1</td>
<td>96.5</td>
<td>72.5</td>
<td>88.1</td>
<td>79.9</td>
<td>73.1</td>
</tr>
<tr>
<td>Ours(VCD)</td>
<td>98.8</td>
<td>88.0</td>
<td>93.8</td>
<td>56.9</td>
<td>61.9</td>
<td>72.9</td>
<td>80.0</td>
<td>82.6</td>
<td>94.1</td>
<td>73.0</td>
<td>95.9</td>
<td>88.6</td>
<td>76.1</td>
<td>96.5</td>
<td>75.5</td>
<td>88.6</td>
<td>73.4</td>
<td>79.6</td>
</tr>
</tbody>
</table>

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


[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.


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


