


two-type citation


Method	BaseNet	PAM	CAM	Mean IoU%
Dilated FCN	Res50			70.03
DANet	Res50	✓		75.74
DANet	Res50		✓	74.28
DANet	Res50	✓	✓	76.34
Dilated FCN	Res101			72.54
DANet	Res101	✓		77.03
DANet	Res101		✓	76.55
DANet	Res101	✓	✓	77.57

Table 1: Ablation study on Cityscapes val set. *PAM* represents Position Attention Module, *CAM* represents Channel Attention Module.

ule yields a result of 75.74% in Mean IoU, which brings 5.71% improvement. Meanwhile, employing channel contextual module individually outperforms the baseline by 4.25%. When we integrate the two attention modules together, the performance further improves to 76.34%. Furthermore, when we adopt a deeper pre-trained network (ResNet-101), the network with two attention modules significantly improves the segmentation performance over the baseline model by 5.03%. Results show that attention modules bring great benefit to scene segmentation.

The effects of position attention modules can be visualized in Figure.4. Some details and object boundaries are clearer with the position attention module, such as the 'pole' in the first row and the 'sidewalk' in the second row. Selective fusion over local features enhance the discrimination of details. Meanwhile, Figure.5 demonstrate that, with our channel attention module, some misclassified category are now correctly classified, such as the 'bus' in the first and third row. The selective integration among channel maps helps to capture context information. The semantic consistency have been improved obviously.

4.2.2 Ablation Study for Improvement Strategies

Following [4], we adopt the same strategies to improve performance further. (1) DA: Data augmentation with random scaling. (2) Multi-Grid: we apply employ a hierarchy of grids of different sizes (4,8,16) in the last ResNet block. (3) MS: We average the segmentation probability maps from 8 image scales{0.5 0.75 1 1.25 1.5 1.75 2 2.2} for inference.

Experimental results are shown in Table 2. Data augmentation with random scaling improves the performance by almost 1.26%, which shows that network benefits from enriching scale diversity of training data. We adopt Multi-Grid to obtain better feature representations of pretrained network, which further achieves 1.11% improvements. Finally, segmentation map fusion further improves the performance to 81.50%, which outperforms well-known method Deeplabv3 [4] (79.30% on Cityscape val set) by 2.20%.

Method	DA	Multi-Grid	MS	Mean IoU%
DANet-101				77.57
DANet-101	✓			78.83
DANet-101	✓	✓		79.94
DANet-101	✓	✓	✓	81.50

Table 2: Performance comparison between different strategies on Cityscape val set. *DANet-101* represents DANet with BaseNet ResNet-101, *DA* represents data augmentation with random scaling. *Multi-Grid* represents employing multi-grid method, *MS* represents multi-scale inputs during inference.

4.2.3 Visualization of Attention Module

For position attention, the overall self-attention map is in size of $(H \times W) \times (H \times W)$, which means that for each specific point in the image, there is an corresponding sub-attention map whose size is $(H \times W)$. In Figure.6, for each input image, we select two points (marked as #1 and #2) and show their corresponding sub-attention map in columns 2 and 3 respectively. We observe that the position attention module could capture clear semantic similarity and long-range relationships. For example, in the first row, the red point #1 are marked on a building and its attention map (in column 2) highlights most the areas where the buildings lies on. Moreover, in the sub-attention map, the boundaries are very clear even though some of them are far away from the point #1. As for the point #2, its attention map focuses on most positions labeled as "car". In the second row, the same holds for the 'traffic sign' and 'person' in global region, even though the number of corresponding pixels is less. The third row is for the 'vegetation' class and 'person' class. In particular, the point #2 does not respond to the nearby 'rider' class, but it does respond to the 'person' faraway.

For channel attention, it is hard to give comprehensible visualization about the attention map directly. Instead, we show some attended channels to see whether they highlight clear semantic areas. In Figure.6, we display the eleventh and fourth attended channels in column 4 and 5. We find that the response of specific semantic is noticeable after channel attention module enhances. For example, 11th channel map responds to the 'car' class in all three examples, and 4th channel map is for the 'vegetation' class, which benefits for the segmentation of two scene categories. In short, these visualizations further demonstrate the necessity of capturing long-range dependencies for improving feature representation in scene segmentation.