Noted that we do not employ convolution layers to embed features before computing relationshoips of two channels, since it can maintain relationship between different channel maps. In addition, different from recent works [27] which explores channel relationships by a global pooling or encoding layer, we exploit spatial information at all corresponding positions to model channel correlations.

3.4. Attention Module Embedding with Networks

In order to take full advantage of long-range contextual information, we aggregate the features from these two attention modules. Specifically, we transform the outputs of two attention modules by a convolution layer and perform an element-wise sum to accomplish feature fusion. At last a convolution layer is followed to generate the final prediction map. We do not adopt cascading operation because it needs more GPU memory. Noted that our attention modules are simple and can be directly inserted in the existing FCN pipeline. They do not increase too many parameters yet strengthen feature representations effectively.

4. Experiments

To evaluate the proposed method, we carry out comprehensive experiments on Cityscapes dataset [5], PASCAL VOC2012 [7], PASCAL Context dataset [14] and COCO Stuff dataset [2]. Experimental results demonstrate that DANet achieves state-of-the-art performance on three datasets. In the next subsections, we first introduce the datasets and implementation details, then we perform a series of ablation experiments on Cityscapes dataset. Finally, we report our results on PASCAL VOC 2012, PASCAL Context and COCO Stuff.

4.1. Datasets and Implementation Details

Cityscapes The dataset has 5,000 images captured from 50 different cities. Each image has 2048×1024 pixels, which have high quality pixel-level labels of 19 semantic classes. There are 2,979 images in training set, 500 images in validation set and 1,525 images in test set. We do not use coarse data in our experiments.

PASCAL VOC 2012 The dataset has 10,582 images for training, 1,449 images for validation and 1,456 images for testing, which involves 20 foreground object classes and one background class.

PASCAL Context The dataset provides detailed semantic labels for whole scenes, which contains 4,998 images for training and 5,105 images for testing. Following [10, 27], we evaluate the method on the most frequent 59 classes along with one background category (60 classes in total).

COCO Stuff The dataset contains 9,000 images for training and 1,000 images for testing. Following [6, 10], we report our results on 171 categories including 80 objects and 91 stuff annotated to each pixel.

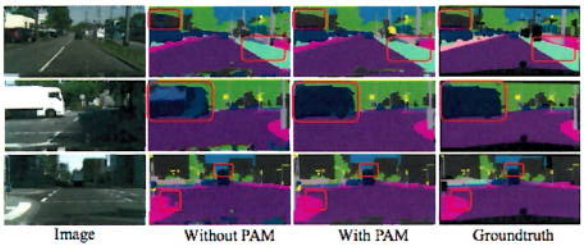

Figure 4: Visualization results of position attention module on Cityscapes val set.

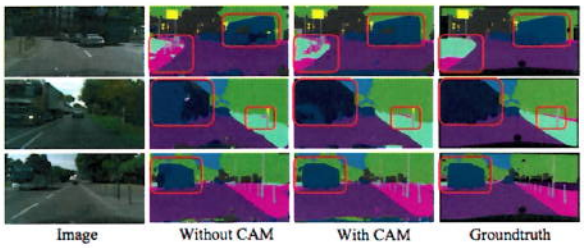

Figure 5: Visualization results of channel attention module on Cityscapes val set.

4.1.1 Implementation Details

We implement our method based on Pytorch. Following [4,27], we employ a poly learning rate policy where the initial learning rate is multiplied by $(1-\frac{iter}{total\_iter})^{0.9}$ after each iteration. The base learning rate is set to 0.01 for Cityscapes dataset. Momentum and weight decay coefficients are set to 0.9 and 0.0001 respectively. We train our model with Synchronized BN [27]. Batchsize are set to 8 for Cityscapes and 16 for other datasets.When adopting multi-scale augmentation, we set training time to 180 epochs for COCO Stuff and 240 epochs for other datasets. Following [3], we adopt multi-loss on the end of the network when both two attention modules are used. For data augmentation, we apply random cropping (cropsize 768) and random left-right flipping during training in the ablation study for Cityscapes datasets.

4.2. Results on Cityscapes Dataset

4.2.1 Ablation Study for Attention Modules

We employ the dual attention modules on top of the dilation network to capture long-range dependencies for better scene understanding. To verify the performance of attention modules, we conduct experiments with different settings in Table 1.

As shown in Table 1, the attention modules improve the performance remarkably. Compared with the baseline FCN (ResNet-50), employing position attention mod-