Dual Attention Network for Scene Segmentation

Jun Fu, Jing Liu, Haijie Tian, Zhiwei Fang, Hanqing Lu
CASIA IVA
{jun.fu, jliu, zhiwei.fang, luhq}@nlpr.ia.ac.cn,{hjtian_bit}@163.com

Abstract

In this paper, we address the scene segmentation task by capturing rich contextual dependencies based on the self-attention mechanism. Unlike previous works that capture contexts by multi-scale features fusion, we propose a Dual Attention Networks (DANet) to adaptively integrate local features with their global dependencies. Specifically, we append two types of attention modules on top of traditional dilated FCN, which model the semantic interdependencies in spatial and channel dimensions respectively. The position attention module selectively aggregates the features at each position by a weighted sum of the features at all positions. Similar features would be related to each other regardless of their distances. Meanwhile, the channel attention module selectively emphasizes interdependent channel maps by integrating associated features among all channel maps. We sum the outputs of the two attention modules to further improve feature representation which contributes to more precise segmentation results. We achieve new state-of-the-art segmentation performance on three challenging scene segmentation datasets, i.e., Cityscapes, PASCAL Context and COCO Stuff dataset. In particular, a Mean IoU score of 81.5% on Cityscapes test set is achieved without using coarse data. *we make the code and trained models publicly available at `https://github.com/junfu1115/DANet`*

Introduction

Scene segmentation is a fundamental and challenging problem, whose goal is to segment and parse a scene image into different image regions associated with semantic categories including stuffs (e.g. sky, road, grass) and discrete objects (e.g. person, car, bicycle). The study of this task can be applied to potential applications, such as automatic driving, robot sensing and image editing. In order to accomplish the task of scene segmentation effectively, we need to distinguish some confusing categories and take into account objects with different appearance. For example, local features for 'field' and 'grass' are indistinguishable in some cases, and 'cars' in different locations of the road may have different scales, occlusion and illumination changing. Therefore, it is necessary to enhance the ability of feature representations for pixel-level recognition.

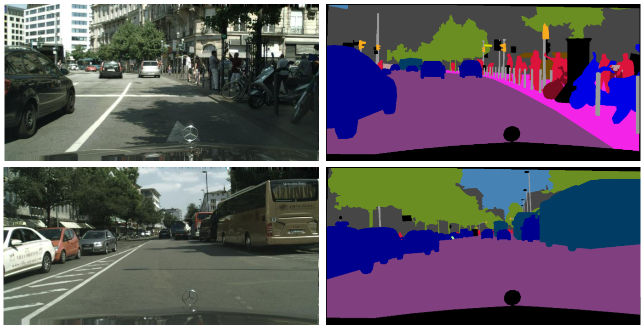


Figure 1: The goal of scene segmentation is to recognize each pixel including stuff, diverse objects. The various scales, occlusion and illumination changing of objects/stuff make it challenging to parsing each pixel.

Recently, state-of-the-art methods based on Fully Convolutional Networks (FCNs) (Long, Shelhamer, and Darrell 2015) have been proposed to address the above issues. One way is to utilize the multi-scale context fusion. For example, some works (Chen et al. 2017; Zhao et al. 2017; Chen et al. 2018) aggregate multi-scale contexts via combining feature maps generated by different dilated convolutions and pooling operations. And some works (Peng et al. 2017; Zhang et al. 2018b) capture richer global context information by enlarging the kernel size with a decomposed structure or introducing an effective encoding layer on top of the network. In addition, the encoder-decoder structures (Lin et al. 2017a; Ronneberger, Fischer, and Brox 2015; Ding et al. 2018) are proposed to fuse mid-level and high-level semantic features. Although the context fusion helps to capture different scales objects, it does not leverage the relationship between objects or stuffs in scene, which is also essential to scene representation.

Another type of methods employs recurrent neural networks to exploit long-range dependencies, thus improving scene segmentation accuracy. The method based on 2D LSTM networks (Byeon et al. 2015) is proposed to capture complex spatial dependencies on labels. The work (Shuai et al. 2018) builds a recurrent neural network with directed acyclic graph to capture the rich contextual dependencies over local features. However, these methods capture the global relationship implicitly with recurrent neural networks, whose effectiveness relies heavily on the learning outcome of the long-term memorization.

To address above problems, in this paper, we propose a novel framework, called as Dual Attention Network (DANet), for natural scene image segmentation, which is illustrated in Figure. 2. It introduces a self-attention mechanism to capture visual features dependencies in the spatial and channel dimensions respectively. Specifically, we append two parallel attention modules on top of traditional dilated FCN. One is a *position attention module* and the other is a *channel attention module*. For the position attention module, we introduce the self-attention mechanism to capture the spatial dependencies between any two positions of the feature maps. For feature at a certain position, it is updated via aggregating features at all positions with weighted summation, where the weights are decided by the feature similarities between the corresponding two positions. That is, any two positions with similar features can contribute mutual improvement regardless of their distance in spatial dimension. For the channel attention module, we use the similar self-attention mechanism to capture the channel dependencies between any two channel maps, and update each channel map with a weighted sum of all channel maps. Finally, the outputs of these two attention modules are fused to further enhance the feature representations.

It should be noted that our method is more effective and flexible than previous methods (Chen et al. 2017; Zhao et al. 2017) when dealing with complex and diverse scenes. Take the street scene in Figure. 1 as an example. First, some 'person' and 'traffic light' in the first row are inconspicuous or incomplete objects due to lighting and view. If simple contextual embedding is explored, the context from dominated salient objects (e.g. car, building) would not assist labeling inconspicuous objects. By contrast, in our attention model, we selectively aggregate the similar semantic features of inconspicuous objects to enhance their feature representations and avoid the influence of salient objects. Second, the scales of the 'car' and 'person' are diverse, and recognizing such diverse objects requires contextual information at different scales. That is, the features at different scale should not be treated equally to represent different semantics. Our model with attention mechanism just aims to adaptively integrate similar features at any scales from a global view, and this can solve the above problem to some extent. Third, we explicitly take spatial relationships and channel relationships into consideration, so that scene understanding could benefit from long-range dependencies.

Our main contributions can be summarized as follows:

- We propose a Dual Attention Network (DANet) to capture the global feature dependencies in the spatial and channel dimensions for the task of scene understanding.
- A position attention module is proposed to learn the spatial interdependencies of features and a channel attention module is designed to model channel interdependencies. It significantly improves the segmentation results by modeling rich contextual dependencies over local features.
- We achieve new state-of-the-arts results on three popular benchmarks including Cityscapes dataset (Cordts et al. 2016), PASCAL Context dataset (Mottaghi et al. 2014) and COCO Stuff dataset (Caesar, Uijlings, and Ferrari 2016).

Related Work

Semantic segmentation. Fully Convolutional Networks (FCNs) based methods have made great progress in image semantic segmentation. There are several model variants proposed to enhance multi-scale contextual aggregation. For example, Deeplabv2 (Chen et al. 2018) proposes atrous spatial pyramid pooling (ASPP) to embed contextual information, which consists of parallel dilated convolutions with different dilated rates. Deeplabv3 (Chen et al. 2017) extends the ASPP module with image-level features to further capture global contexts. PSPNet (Zhao et al. 2017) designs a pyramid pooling module to collect the effective contextual prior, containing information of different scales.

In addition, some works exploit contextual dependencies over local features. The work (Shuai et al. 2018) proposes a Directed Acyclic Graph-Recurrent Neural Network to capture long-range contexts and embed them into local features to enhance their representative capability. The work (Liu et al. 2015) adopts additional convolution layers to approximate the mean field algorithm (MF) for capturing high-order relations. The work (Liang et al. 2016) proposes a Graph LSTM to deal with general graph-structured data in semantic object parsing tasks.

Different from previous works, in this paper, we propose two types of attention modules to model long-range dependencies, thus adaptively integrating interdependent semantic features for better scene understanding.

Attention modules. Attention modules can model long-range dependencies and have recently been widely applied in the Natural Language Processing (NLP) field. In particular, the work (Vaswani et al. 2017) is the first to propose the self-attention mechanism to directly draw global dependencies of inputs and apply it in the task of machine translation. Such a mechanism is dispensed with recurrence and convolutions entirely, thus more parallelizable and more efficient in model training. This attention mechanism has been extended in many NLP applications, such as natural language inference (Shen et al. 2018), text representation (Lin et al. 2016), sentence embedding (Lin et al. 2017b) and so on.

Meanwhile, the attention modules are also increasingly applied in the image vision flied. For example, the work (Hu et al. 2017) proposes an object relation module to model the relationships among a set of objects, which improves object recognition. The work (Zhang et al. 2018a) introduces self-attention modules to efficiently find global dependencies within internal representations for better image generation.

Our approach is motivated by the success of attention modules in the above works. We explore long-range dependencies over local features across spatial dimensions and channel dimensions simultaneously. To the best of our knowledge, this is the first to introduce self-attention mechanism to model dependency relationship of visual features in scene segmentation.

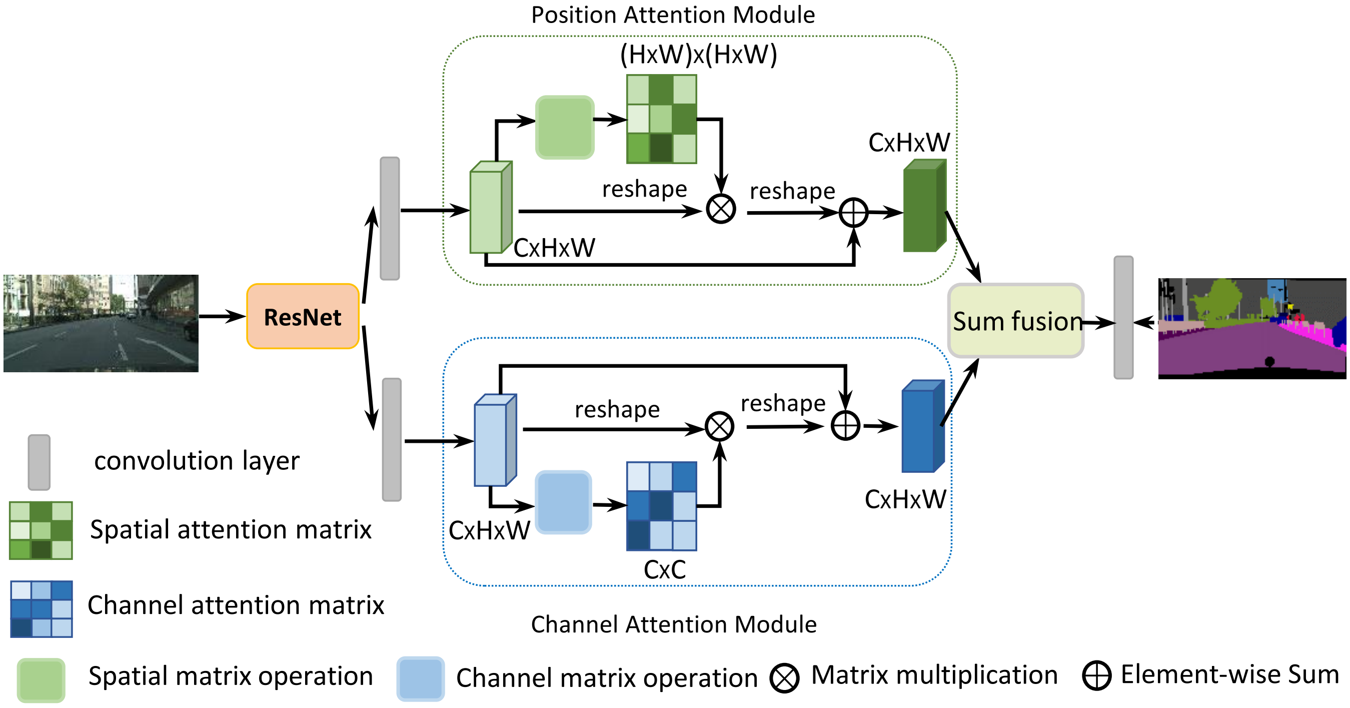


Figure 2: An overview of the Dual Attention Network. (Best viewed in color)

Dual Attention Network

In this section, we first present a general framework of our network and then introduce in detail the two attention modules which capture long-range contextual information in spatial dimension and channel dimension respectively. Finally we describe how to aggregate them together for further refinement.

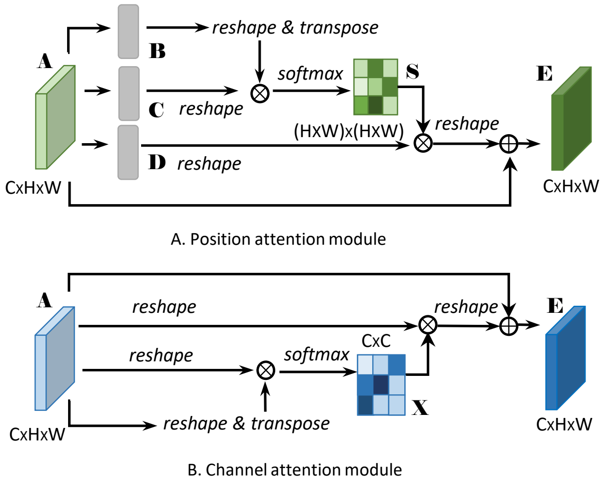
Overview

Given a picture of scene segmentation, stuff or objects, are diverse on scales, lighting, and views. Since convolution operations have a local receptive field, the features corresponding to the pixels with the same label do have some differences. These differences affect the recognition accuracy of traditional FCNs. To address this issue, we design DANet to capture global dependencies by building associations among features with the attention mechanism. Our method could adaptively aggregate long-range contextual information, thus improving feature representation for scene segmentation.

As illustrated in Figure. 2, we design two types of attention modules to draw global dependencies over local features generated by a dilated residual network, thus obtaining better feature representations for pixel-level prediction. We employ a pretrained residual network with the dilated strategy (Chen et al. 2018) as the backbone. Note that we remove the downsampling operations and employ dilation convolutions in the last two ResNet blocks, thus enlarging the size of the final feature map size to 1/8 of the input image. This retains more details without adding extra parameters. Then the features from the dilated residual network would be fed into two parallel attention modules. Take the spatial attention modules in the upper part of the Figure. 2 as an example, we first apply a convolution layer to obtain the features of dimension reduction. Then we feed the features into the position attention module and generate new features of spatial long-range dependencies through the following three. The first step is to generate a spatial attention matrix which models the spatial relationship between any two pixels of the features. Next, we perform a matrix multiplication between the attention matrix and the original features. Third, we perform an element-wise sum operation on the above multiplied resulting matrix and original features to obtain the final representations reflecting long-range contexts. Meanwhile, channel long-range dependencies are captured by a channel attention module. The process to capture the channel relationship is similar to the position attention module except for the first step, in which channel attention matrix is calculated in the channel dimension. Finally we aggregate the output features from the two attention modules to obtain better feature representations for pixel-level prediction.

Position Attention Module

Context relationship is essential for scene understanding, which aims at capturing global dependencies regardless of locations. However, many works (Zhao et al. 2017; Peng et al. 2017) suggest that local feature representations generated by traditional FCNs could lead to misclassification of objects and stuff. In order to model rich contextual de-



A. Position attention module

B. Channel attention module

Figure 3: The details of Position Attention Module and Channel Attention Module are illustrated in (A) and (B). (Best viewed in color)

pendencies over local feature representations, we introduce a position attention module. The position attention module encodes a wider range of contextual information into local features, thus enhancing their representative capability. Next, we elaborate the process to adaptively aggregate spatial contexts.

As illustrated in Figure. 3(A), given a local feature $\mathbf{A} \in \mathbb{R}^{C \times H \times W}$, we first feed it into a convolution layers with batch normalization and ReLU layers to generate two new feature maps $\mathbf{B}$ and $\mathbf{C}$, respectively, where $\{\mathbf{B}, \mathbf{C}\} \in \mathbb{R}^{C \times H \times W}$. Then we reshape them to $\mathbb{R}^{C \times N}$, where $N = H \times W$ is the number of features. After that we perform a matrix multiplication between the transpose of $\mathbf{C}$ and $\mathbf{B}$, and apply a softmax layer to calculate the spatial attention map $\mathbf{S} \in \mathbb{R}^{N \times N}$:

$$s_{ji} = \frac{exp(B_i \cdot C_j)}{\sum_{i=1}^{N} exp(B_i \cdot C_j)} \tag{1}$$

where s_{ji} measures the i^{th} position's impact on j^{th} position. Note that the more similar feature representations of the two position contributes to greater correlation between them.

Meanwhile, we feed feature $\mathbf{A}$ into a convolution layer with batch normalization and ReLU layers to generate a new feature map $\mathbf{D} \in \mathbb{R}^{C \times H \times W}$ and reshape it to $\mathbb{R}^{C \times N}$. Then we perform a matrix multiplication between $\mathbf{D}$ and the transpose of $\mathbf{S}$ and reshape the result to $\mathbb{R}^{C \times H \times W}$. Finally, we multiply it by a scale parameter α and perform a element-wise sum operation with the features $\mathbf{A}$ to obtain the final output $\mathbf{E} \in \mathbb{R}^{C \times H \times W}$ as follows:

$$E_j = \alpha \sum_{i=1}^{N} (s_{ji} D_i) + A_j \tag{2}$$

where α is initialized as 0 and gradually learn to assign more weight (Zhang et al. 2018a). It can be inferred from Equation 2 that the resulting feature $\mathbf{E}$ at each position is a weighted sum of the features at all positions and original features. Therefore, it has a global contextual view and selectively aggregates contexts according to the spatial attention map. This feature representations achieve mutual gains and are more robust for scene segmentation.

Channel Attention Module

Each channel map of high level features can be regarded as a class-specific response, and different semantic responses are associated with each other. By exploiting the interdependencies between channel maps, we could emphasize interdependent feature maps and improve the feature representation of specific semantics. Therefore, we build a channel attention module to explicitly model interdependencies between channels.

The structure of channel attention module is illustrated in Figure. 3(B). Different from the position attention module, we directly calculate the channel attention map $\mathbf{X} \in \mathbb{R}^{C \times C}$ from the original features $\mathbf{A} \in \mathbb{R}^{C \times H \times W}$. Specifically, we reshape $\mathbf{A}$ to $\mathbb{R}^{C \times N}$. We then perform a matrix multiplication between $\mathbf{A}$ and the transpose of $\mathbf{A}$. Finally, we apply a softmax layer to obtain the channel attention map $\mathbf{X} \in \mathbb{R}^{C \times C}$:

$$x_{ji} = \frac{exp(A_i \cdot A_j)}{\sum_{i=1}^{C} exp(A_i \cdot A_j)} \tag{3}$$

where x_{ji} measures the i^{th} channel's impact on the j^{th} channel. In addition, we perform a matrix multiplication between the transpose of $\mathbf{X}$ and $\mathbf{A}$ and reshape their result to $\mathbb{R}^{C \times H \times W}$. Then we multiply the result by a scale parameter β and perform an element-wise sum operation with $\mathbf{A}$ to obtain the final output $\mathbf{E} \in \mathbb{R}^{C \times H \times W}$:

$$E_j = \beta \sum_{i=1}^{C} (x_{ji} A_i) + A_j \tag{4}$$

where β gradually learn a weight from 0. The Equation 4 shows that the final feature of each channel is a weighted sum of the features of all channels and original features, which models the long-range semantic dependencies between feature maps. It emphasizes class-dependent feature maps and helps to boost feature discriminability. Different from recent work (Zhang et al. 2018b) which explores channel relationships by a global pooling or encoding layer, we exploit spatial information at all corresponding positions to model channel correlations.

Attention Module Embedding with Networks

In order to take full advantage of long-range dependencies, 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. 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.

Experiments

To evaluate the proposed method, we carry out comprehensive experiments on Cityscapes dataset (Cordts et al. 2016), PASCAL Context dataset (Mottaghi et al. 2014) and COCO Stuff dataset (Caesar, Uijlings, and Ferrari 2016). Experimental results demonstrate that DANet achieves state-of-the-art performance on these 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 Context and COCO Stuff.

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 Context The dataset provides detailed semantic labels for whole scenes, which contains 4,998 images for training and 5,105 images for testing. Following (Lin et al. 2017a; Zhang et al. 2018b), 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 (Lin et al. 2017a; Ding et al. 2018), we report our results on 171 categories including 80 objects and 91 stuff annotated to each pixel.

Implementation Details We implement our method based on Pytorch. Following prior works (Chen et al. 2017; Zhang et al. 2018b), 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 and 0.001 for others. Momentum and weight decay coefficients are set to 0.9 and 0.0001 respectively. For data augmentation, we apply random cropping and random left-right flipping during training in the ablation study for Cityscapes datasets.

Following the common procedure of (Chen et al. 2018; Zhang et al. 2018b; Yang et al. 2018), we apply Mean IoU (Percentage of correctly labeled pixels in a class over the union set of pixels predicted to this class and groundtruth, and then averaged over all classes) to evaluate our approach.

Results on Cityscapes Dataset

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 module yields a result of 75.74% in Mean IoU , which brings 5.71% improvement. Meanwhile, employing channel contextual module individually outperforms the baseline by 4.28% gain. 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.

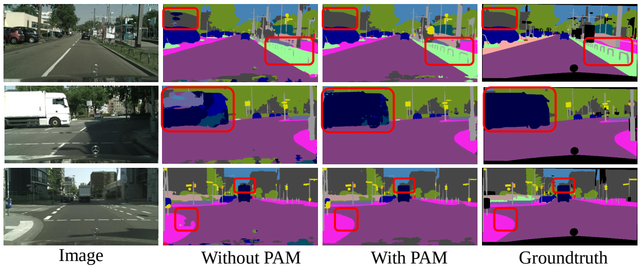


Figure 4: Visualization results of position attention module on Cityscapes val set

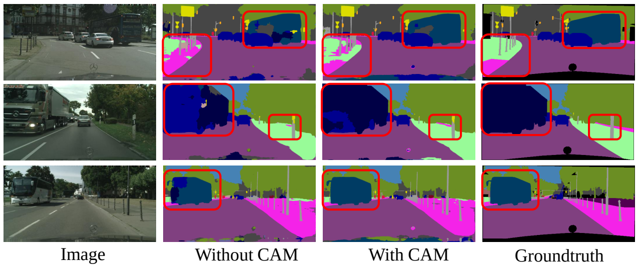


Figure 5: Visualization results of channel attention module on Cityscapes val set

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.

Ablation Study for Improvement Strategies Following (Chen et al. 2017), 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 in the last ResNet block. (3) MS: We average the segmentation probability maps from 6 image scales{0.75 1 1.25 1.5 1.75 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 enrich-

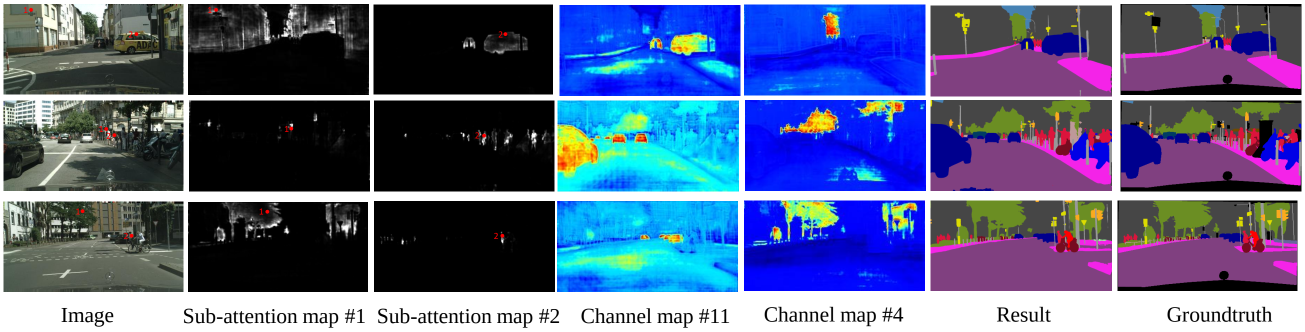


Figure 6: Visualization results of attention modules on Cityscapes val set. For each row, we show an input image, two sub-attention maps $(H \times W)$ corresponding to the ponits marked in the input image. Meanwhile, we give two channel maps from the outputs of channel attention module, where the maps are from 4^{th} and 11^{th} channels, respectively. Finally, corresponding result and groundtruth are provided.

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.

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.

ing scale diversity of training data. We adopt Multi-Grid to obtain better feature representations of pretrained network, which further achieves 1.11% improvement. Finally, segmentation map fusion further improves the performance to 81.50%, which outperforms well-known method Deeplabv3 (Chen et al. 2017) (79.30% on Cityscape val set) by 2.20%.

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 point (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. Whats more, 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, 11^{th} channel map responds to the 'car' class in all three examples, and 4^{th} 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.

Comparing with State-of-the-art We further compare our method with existing methods on the Cityscapes test set. Specifically, we train our DANet-101 with only fine annotated data and submit our test results to the official evaluation server. Results are shown in Table 3. DANet outperforms existing approaches with dominantly advantage. In particular, our model outperforms the PSANet by a large margin with the same backbone ResNet-101. Moreover, it also surpasses DenseASPP, which use more powerful pretrained models than ours.

Noted that we mainly discuss the effectiveness of our method in a general dialted FCN network, and our modules can be used directly in some methods of multi-scale context fusion (Chen et al. 2017; Zhao et al. 2017; Lin et al. 2017a). In addition, some others improvement strategy, such as on-

Methods	Mean IoU	road	sidewalk	building	wall	fence	pole	traffic light	traffic sign	vegetation	terrain	sky	person	rider	car	truck	bus	train	motorcycle	bicycle
DeepLab-v2(Chen et al. 2018)	70.4	97.9	81.3	90.3	48.8	47.4	49.6	57.9	67.3	91.9	69.4	94.2	79.8	59.8	93.7	56.5	67.5	57.5	57.7	68.8
RefineNet (Lin et al. 2017a)	73.6	98.2	83.3	91.3	47.8	50.4	56.1	66.9	71.3	92.3	70.3	94.8	80.9	63.3	94.5	64.6	76.1	64.3	62.2	70
GCN (Peng et al. 2017)	76.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DUC (Wang et al. 2018)	77.6	98.5	85.5	92.8	58.6	55.5	65	73.5	77.9	93.3	72	95.2	84.8	68.5	95.4	70.9	78.8	68.7	65.9	73.8
ResNet-38 (Wu, Shen, and Hengel 2016)	78.4	98.5	85.7	93.1	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	76.7
PSPNet (Zhao et al. 2017)	78.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
BiSeNet (Yu et al. 2018)	78.9	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PSANet (Zhao et al. 2018)	80.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
DenseASPP (Yang et al. 2018)	80.6	**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
DANet	**81.5**	98.6	86.1	**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**

Table 3: Per-class results on Cityscapes testing set. DANet outperforms existing approaches and achieves 81.5% in Mean IoU.

Method	**mIoU%**
FCN-8s (Long, Shelhamer, and Darrell 2015)	37.8
HO_CRF (Arnab et al. 2016)	41.3
Piecewise (Lin et al. 2016)	43.3
DeepLab-v2 (Res101-COCO) (Chen et al. 2018)	45.7
RefineNet (Res152) (Lin et al. 2017a)	47.3
PSPNet (Res101) (Zhao et al. 2017)	47.8
Ding et al.(Res101) (Ding et al. 2018)	51.6
EncNet (Res101) (Zhang et al. 2018b)	51.7
Dilated FCN(Res50)	44.3
DANet (Res50)	50.1
DANet (Res101)	**52.6**

Table 4: Segmentation results on PASCAL Context testing set.

line bootstrapping of difficult pixels and pictures, also can be used in training phase. Moreover, Our method aims at capturing global dependencies in the spatial and channel dimensions respectively, which could be explored further in some other visual tasks, such as instance segmentation (He et al. 2017), pose estimation (Newell, Yang, and Deng 2016), image caption (Xu et al. 2015), and so on.

Results on PASCAL Context Dataset

In this subsection, we carry out experiments on the PASCAL Context dataset to further evaluate the effectiveness of our method. Quantitative results of PASCAL Context are shown in Table. 4. The baseline (Dilated FCN-50) yields Mean IoU 44.3%. DANet-50 boosts the performance to 50.1%. Furthermore, with a deep pretrained network ResNet101, our model results achieve Mean IoU 52.6%, which outperforms previous methods by a large margin. Among previous works, Deeplab-v2 and RefineNet adopt multi-scale feature fusion using features from different atrous convolution or different stage of encoder. In addition, they trained their model with extra COCO data or adopt a deeper model (ResNet152) to improve their segmentation results. Different from previous methods, we introduce attention modules to capture global dependencies explicitly, and the proposed method can achieve better performance.

Method	**mIoU%**
FCN-8s (Long, Shelhamer, and Darrell 2015)	22.7
DeepLab-v2(Res101-COCO) (Chen et al. 2018)	26.9
DAG-RNN (Shuai et al. 2018)	31.2
RefineNet (Res101) (Lin et al. 2017a)	33.6
Ding et al.(Res101) (Ding et al. 2018)	35.7
Dilated FCN (Res50)	31.9
DANet (Res50)	37.2
DANet (Res101)	**39.7**

Table 5: Segmentation results on COCO Stuff testing set.

Results on COCO Stuff Dataset

We also conduct experiments on the COCO Stuff dataset to verify the generalization of our proposed network. Comparisons with previous state-of-the-art methods are reported in Table. 5. Results show that our model achieves 39.7% in Mean IoU, which outperforms these methods by a large margin. Among the compared methods, DAG-RNN (Shuai et al. 2018) utilizes chain-RNNs for 2D images to model rich spatial dependencies, and Ding et al. (Ding et al. 2018) adopts a gating mechanism in the decoder stage for improving inconspicuous objects and background stuff segmentation. our method could capture long-range contextual information more effectively and learn better feature representation in scene segmentation.

Conclusion

In this paper, we have presented a Dual Attention Network (DANet) for scene segmentation, which adaptively integrates local semantic features using the self-attention mechanism. Specifically, we introduce a position attention module and a channel attention module to capture global dependencies in the spatial and channel dimensions respectively. The ablation experiments show that dual attention modules capture long-range contextual information effectively and give more precise segmentation results. Our attention network achieves outstanding performance consistently on three scene segmentation datasets, i.e. Cityscapes, Pascal Context, and COCO Stuff.

Acknowledgment

We would like to acknowledge Yong Li, Yongjun Bao and Weipeng Yan (JD.com) for their valuable support.

References

[Arnab et al. 2016] Arnab, A.; Jayasumana, S.; Zheng, S.; and Torr, P. H. S. 2016. Higher order conditional random fields in deep neural networks. In *the European Conference on Computer Vision,*, 524–540.

[Byeon et al. 2015] Byeon, W.; Breuel, T. M.; Raue, F.; and Liwicki, M. 2015. Scene labeling with LSTM recurrent neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 3547–3555.

[Caesar, Uijlings, and Ferrari 2016] Caesar, H.; Uijlings, J. R. R.; and Ferrari, V. 2016. Coco-stuff: Thing and stuff classes in context. *CoRR* abs/1612.03716.

[Chen et al. 2017] Chen, L.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *CoRR* abs/1706.05587.

[Chen et al. 2018] Chen, L.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2018. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 40(4):834–848.

[Cordts et al. 2016] Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; and Schiele, B. 2016. The cityscapes dataset for semantic urban scene understanding. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223.

[Ding et al. 2018] Ding, H.; Jiang, X.; Shuai, B.; Liu, A. Q.; and Wang, G. 2018. Context contrasted feature and gated multi-scale aggregation for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2393–2402.

[He et al. 2017] He, K.; Gkioxari, G.; Dollár, P.; and Girshick, R. 2017. Mask r-cnn. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, 2980–2988. IEEE.

[Hu et al. 2017] Hu, H.; Gu, J.; Zhang, Z.; Dai, J.; and Wei, Y. 2017. Relation networks for object detection. *CoRR* abs/1711.11575.

[Liang et al. 2016] Liang, X.; Shen, X.; Feng, J.; Lin, L.; and Yan, S. 2016. Semantic object parsing with graph LSTM. In *Computer Vision - ECCV 2016 - 14th European Conference*, 125–143.

[Lin et al. 2016] Lin, G.; Shen, C.; van den Hengel, A.; and Reid, I. D. 2016. Efficient piecewise training of deep structured models for semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 3194–3203.

[Lin et al. 2017a] Lin, G.; Milan, A.; Shen, C.; and Reid, I. D. 2017a. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 5168–5177.

[Lin et al. 2017b] Lin, Z.; Feng, M.; Santos, C. N. d.; Yu, M.; Xiang, B.; Zhou, B.; and Bengio, Y. 2017b. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130.*

[Liu et al. 2015] Liu, Z.; Li, X.; Luo, P.; Loy, C. C.; and Tang, X. 2015. Semantic image segmentation via deep parsing network. In *2015 IEEE International Conference on Computer Vision*, 1377–1385.

[Long, Shelhamer, and Darrell 2015] Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3431–3440.

[Mottaghi et al. 2014] Mottaghi, R.; Chen, X.; Liu, X.; Cho, N.; Lee, S.; Fidler, S.; Urtasun, R.; and Yuille, A. L. 2014. The role of context for object detection and semantic segmentation in the wild. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 891–898.

[Newell, Yang, and Deng 2016] Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *European Conference on Computer Vision*, 483–499. Springer.

[Peng et al. 2017] Peng, C.; Zhang, X.; Yu, G.; Luo, G.; and Sun, J. 2017. Large kernel matters - improve semantic segmentation by global convolutional network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 1743–1751.

[Ronneberger, Fischer, and Brox 2015] Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 234–241.

[Shen et al. 2018] Shen, T.; Zhou, T.; Long, G.; Jiang, J.; Pan, S.; and Zhang, C. 2018. Disan: Directional self-attention network for rnn/cnn-free language understanding. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

[Shuai et al. 2018] Shuai, B.; Zuo, Z.; Wang, B.; and Wang, G. 2018. Scene segmentation with dag-recurrent neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 1480–1493.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, 6000–6010.

[Wang et al. 2018] Wang, P.; Chen, P.; Yuan, Y.; Liu, D.; Huang, Z.; Hou, X.; and Cottrell, G. W. 2018. Understanding convolution for semantic segmentation. In *IEEE Winter Conference on Applications of Computer Vision*, 1451–1460.

[Wu, Shen, and Hengel 2016] Wu, Z.; Shen, C.; and Hengel, A. v. d. 2016. Wider or deeper: Revisiting the resnet model for visual recognition. *arXiv preprint arXiv:1611.10080.*

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, 2048–2057.

[Yang et al. 2018] Yang, M.; Yu, K.; Zhang, C.; Li, Z.; and Yang, K. 2018. Denseaspp for semantic segmentation in street scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3684–3692.

[Yu et al. 2018] Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; and Sang, N. 2018. Bisenet: Bilateral segmentation network for real-time semantic segmentation. *arXiv preprint arXiv:1808.00897*.

[Zhang et al. 2018a] Zhang, H.; Goodfellow, I. J.; Metaxas, D. N.; and Odena, A. 2018a. Self-attention generative adversarial networks. *CoRR* abs/1805.08318.

[Zhang et al. 2018b] Zhang, H.; Dana, K.; Shi, J.; Zhang, Z.; Wang, X.; Tyagi, A.; and Agrawal, A. 2018b. Context encoding for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Zhao et al. 2017] Zhao, H.; Shi, J.; Qi, X.; Wang, X.; and Jia, J. 2017. Pyramid scene parsing network. In *IEEE Conference on Computer Vision and Pattern Recognition*, 6230–6239.

[Zhao et al. 2018] Zhao, H.; Zhang, Y.; Liu, S.; Shi, J.; Loy, C. C.; Lin, D.; and Jia, J. 2018. Psanet: Point-wise spatial attention network for scene parsing. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 267–283.