PSANet: Point-wise Spatial Attention Network for Scene Parsing

Hengshuang Zhao[1⋆], Yi Zhang[2⋆], Shu Liu[1], Jianping Shi[3], Chen Change Loy[4], Dahua Lin[2], and Jiaya Jia[1,5]

[1]The Chinese University of Hong Kong
[2]CUHK-Sensetime Joint Lab, The Chinese University of Hong Kong
[3]SenseTime Research [4]Nanyang Technological University [5]Tencent Youtu Lab
{hszhao,sliu,leojia}@cse.cuhk.edu.hk, {zy217,dhlin}@ie.cuhk.edu.hk,
shijianping@sensetime.com, ccloy@ntu.edu.sg

Abstract. We notice information flow in convolutional neural networks is restricted inside local neighborhood regions due to the physical design of convolutional filters, which limits the overall understanding of complex scenes. In this paper, we propose the *point-wise spatial attention network* (PSANet) to relax the local neighborhood constraint. Each position on the feature map is connected to all the other ones through a self-adaptively learned attention mask. Moreover, information propagation in bi-direction for scene parsing is enabled. Information at other positions can be collected to help the prediction of the current position and vice versa, information at the current position can be distributed to assist the prediction of other ones. Our proposed approach achieves top performance on various competitive scene parsing datasets, including ADE20K, PASCAL VOC 2012 and Cityscapes, demonstrating its effectiveness and generality.

Keywords: Point-wise Spatial Attention, Bi-Direction Information Flow, Adaptive Context Aggregation, Scene Parsing, Semantic Segmentation

1 Introduction

Scene parsing, a.k.a. semantic segmentation, is a fundamental and challenging problem in computer vision, in which each pixel is assigned with a category label. It is a key step towards visual scene understanding, and plays a crucial role in applications such as auto-driving and robot navigation.

The development of powerful deep *convolutional neural networks* (CNNs) has made remarkable progress in scene parsing [26,1,29,4,5,45]. Owing to the design of CNN structures, the receptive field of it is limited to local regions [47,27]. The limited receptive field imposes a great adverse effect on *fully convolutional networks* (FCNs) based scene parsing systems due to insufficient understanding of surrounded contextual information.

⋆ indicates equal contribution.

To address this issue, especially leveraging long-range dependency, several modifications have been made. Contextual information aggregation through dilated convolution is proposed by [4,42]. Dilations are introduced into the classical compact convolution module to expand the receptive field. Contextual information aggregation can also be achieved through pooling operation. Global pooling module in ParseNet [24], different-dilation based *atrous spatial pyramid pooling* (ASPP) module in DeepLab [5] and different-region based *pyramid pooling module* (PPM) in PSPNet [45] can help extract the context information to a certain degree. Different from these extensions, *conditional random field* (CRF) [4,46,2,3] and *Markov random field* (MRF) [25] are also utilized. Besides, *recurrent neural network* (RNN) is introduced in ReSeg [38] for its capability to capture long-range dependencies. However, these dilated-convolution-based [4,42] and pooling-based [24,5,45] extensions utilize homogeneous contextual dependencies for all image regions in a non-adaptive manner, ignoring the difference of local representation and contextual dependencies for different categories. The CRF/MRF-based [4,46,2,3,25] and RNN-based [38] extensions are less efficient than CNN-based frameworks.

In this paper, we propose the *point-wise spatial attention network* (PSANet) to aggregate long-range contextual information in a flexible and adaptive manner. Each position in the feature map is connected with all other ones through self-adaptively predicted attention maps, thus harvesting various information nearby and far away. Furthermore, we design the bi-directional information propagation path for a comprehensive understanding of complex scenes. Each position collects information from all others to help the prediction of itself and vice versa, the information at each position can be distributed globally, assisting the prediction of all other positions. Finally, the bi-directionally aggregated contextual information is fused with local features to form the final representation of complex scenes.

Our proposed PSANet achieves top performance on three most competitive semantic segmentation datasets, *i.e.*, ADE20K [48], PASCAL VOC 2012 [9] and Cityscapes [8]. We believe the proposed point-wise spatial attention module together with the bi-directional information propagation paradigm can also benefit other dense prediction tasks. We give all implementation details, and make the code and trained models publicly available to the community[1]. Our main contribution is three-fold:

- We achieve long-range context aggregation for scene parsing by a learned point-wise position-sensitive context dependency together with a bi-directional information propagation paradigm.
- We propose the *point-wise spatial attention network* (PSANet) to harvest contextual information from all positions in the feature map. Each position is connected with all others through a self-adaptively learned attention map.
- PSANet achieves top performance on various competitive scene parsing datasets, demonstrating its effectiveness and generality.

[1] https://github.com/hszhao/PSANet

2 Related Work

Scene Parsing and Semantic Segmentation. Recently, CNN based methods [26,4,5,42,45,6] have achieved remarkable success in scene parsing and semantic segmentation tasks. FCN [26] is the first approach to replace the fully-connected layer in a classification network with convolution layers for semantic segmentation. DeconvNet [29] and SegNet [1] adopted encoder-decoder structures that utilize information in low-level layers to help refine the segmentation mask. Dilated convolution [4,42] applied skip convolution on feature map to enlarge network's receptive field. UNet [33] concatenated output from low-level layers with higher ones for information fusion. DeepLab [4] and CRF-RNN [46] utilized CRF for structure prediction in scene parsing. DPN [25] used MRF for semantic segmentation. LRR [11] and RefineNet [21] adopted step-wise reconstruction and refinement to get parsing results. PSPNet [45] achieved high performance though pyramid pooling strategy. There are also high efficiency frameworks like ENet [30] and ICNet [44] for real-time applications like automatic driving.

Context Information Aggregation. Context information plays a key role for image understanding. Dilated convolution [4,42] inserted dilation inside classical convolution kernels to enlarge the receptive field of CNN. Global pooling was widely adopted in various basic classification backbones [19,35,36,13,14] to harvest context information for global representations. Liu *et al.* proposed ParseNet [24] that utilizes global pooling to aggregate context information for scene parsing. Chen *et al.* developed ASPP [5] module and Zhao *et al.* proposed PPM [45] module to obtain different regions' contextual information. Visin *et al.* presented ReSeg [38] that utilizes RNN to capture long-range contextual dependency information.

Attention Mechanism. Attention mechanism is widely used in neural networks. Mnih *et al.* [28] learned an attention model that adaptively select a sequence of regions or locations for processing. Chen *et al.* [7] learned several attention masks to fuse feature maps or predictions from different branches. Vaswani *et al.* [37] learned a self-attention model for machine translation. Wang *et al.* [40] got attention masks by calculating the correlation matrix between each spatial point in the feature map. Our point-wise attention masks are different from the aforementioned studies. Specifically, masks learned through our PSA module are self-adaptive and sensitive to location and category information. PSA learns to aggregate contextual information for each individual point adaptively and specifically.

3 Framework

In order to capture contextual information, especially in the long range, information aggregation is of great importance for scene parsing [24,5,45,38]. In this

paper, we formulate the information aggregation step as a kind of information flow and propose to adaptively learn a pixel-wise global attention map for each position from two perspectives to aggregate contextual information over the entire feature map.

3.1 Formulation

General feature learning or information aggregation is modeled as

$$\mathbf{z}_i = \frac{1}{N} \sum_{\forall j \in \Omega(i)} F(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij}) \mathbf{x}_j \tag{1}$$

where $\mathbf{z}_i$ is the newly aggregated feature at position i, and $\mathbf{x}_i$ is the feature representation at position i in the input feature map $\mathbf{X}$. $\forall j \in \Omega(i)$ enumerates all positions in the region of interest associated with i, and Δ_{ij} represents the relative location of position i and j. $F(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij})$ can be any function or learned parameters according to the operation and it represents the information flow from j to i. Note that by taking relative location Δ_{ij} into account, $F(\mathbf{x}_i, \mathbf{x}_j, \Delta_{ij})$ is sensitive to different relative locations. Here N is for normalization.

Specifically, we simplify the formulation and design different functions F with respect to different relative locations. Eq. (1) is updated to

$$\mathbf{z}_i = \frac{1}{N} \sum_{\forall j \in \Omega(i)} F_{\Delta_{ij}}(\mathbf{x}_i, \mathbf{x}_j) \mathbf{x}_j \tag{2}$$

where $\{F_{\Delta_{ij}}\}$ is a set of position-specific functions. It models the information flow from position j to position i. Note that the function $F_{\Delta_{ij}}(\cdot, \cdot)$ takes both the source and target information as input. When there are many positions in the feature map, the number of the combination $(\mathbf{x}_i, \mathbf{x}_j)$ is very large. In this paper, we simplify the formulation and make an approximation.

At first, we simplify the function $F_{\Delta_{ij}}(\cdot, \cdot)$ as

$$F_{\Delta_{ij}}(\mathbf{x}_i, \mathbf{x}_j) \approx F_{\Delta_{ij}}(\mathbf{x}_i) \tag{3}$$

In this approximation, the information flow from j to i is only related to the semantic feature at target position i and the relative location of i and j. Based on Eq. (3), we rewrite Eq. (2) as

$$\mathbf{z}_i = \frac{1}{N} \sum_{\forall j \in \Omega(i)} F_{\Delta_{ij}}(\mathbf{x}_i) \mathbf{x}_j \tag{4}$$

Similarly, we simplify the function $F_{\Delta_{ij}}(\cdot, \cdot)$ as

$$F_{\Delta_{ij}}(\mathbf{x}_i, \mathbf{x}_j) \approx F_{\Delta_{ij}}(\mathbf{x}_j) \tag{5}$$

in which the information flow from j to i is only related to the semantic feature at source position j and the relative location of position i and j.

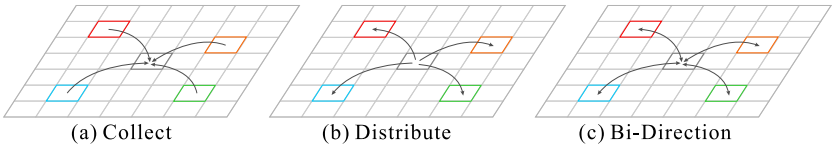


Fig. 1. Illustration of bi-direction information propagation model. Each position both 'collects' and 'distributes' information for more comprehensive information propagation.

We finally decompose and simplify the function as a bi-direction information propagation path. Combining Eq. (3) and Eq. (5), we get

$$F_{\Delta_{ij}}(\mathbf{x}_i, \mathbf{x}_j) \approx F_{\Delta_{ij}}(\mathbf{x}_i) + F_{\Delta_{ij}}(\mathbf{x}_j) \tag{6}$$

Formally, we model this bi-direction information propagation as

$$\mathbf{z}_i = \frac{1}{N} \sum_{\forall j \in \Omega(i)} F_{\Delta_{ij}}(\mathbf{x}_i)\mathbf{x}_j + \frac{1}{N} \sum_{\forall j \in \Omega(i)} F_{\Delta_{ij}}(\mathbf{x}_j)\mathbf{x}_j. \tag{7}$$

For the first term, $F_{\Delta_{ij}}(\mathbf{x}_i)$ encodes to what extent the features at other positions can help prediction. Each position 'collects' information from other positions. For the second term, the importance of the feature at one position to features at other positions is predicted by $F_{\Delta_{ij}}(\mathbf{x}_j)$. Each position 'distributes' information to others. This bi-directional information propagation path, shown in Fig. 1, enables the network to learn more comprehensive representations, evidenced in our experimental section.

Specifically, our PSA module, aiming to adaptively predict the information flow over the entire feature map, takes all the positions in feature map as $\Omega(i)$ and utilizes the convolutional layer as the operation of $F_{\Delta_{ij}}(\mathbf{x}_i)$ and $F_{\Delta_{ij}}(\mathbf{x}_j)$. Both $F_{\Delta_{ij}}(\mathbf{x}_i)$ and $F_{\Delta_{ij}}(\mathbf{x}_j)$ can then be regarded as predicted attention values to aggregate feature $\mathbf{x}_j$. We further rewrite Eq. (7) as

$$\mathbf{z}_i = \frac{1}{N} \sum_{\forall j} \mathbf{a}^c_{i,j}\mathbf{x}_j + \frac{1}{N} \sum_{\forall j} \mathbf{a}^d_{i,j}\mathbf{x}_j, \tag{8}$$

where $\mathbf{a}^c_{i,j}$ and $\mathbf{a}^d_{i,j}$ denote the predicted attention values in the point-wise attention maps $\mathbf{A}^c$ and $\mathbf{A}^d$ from 'collect' and 'distribute' branches, respectively.

3.2 Overview

We show the framework of the PSA module in Fig. 2. The PSA module takes a spatial feature map $\mathbf{X}$ as input. We denote the spatial size of $\mathbf{X}$ as $H \times W$. Through the two branches as illustrated, we generate pixel-wise global attention maps for each position in feature map $\mathbf{X}$ through several convolutional layers.

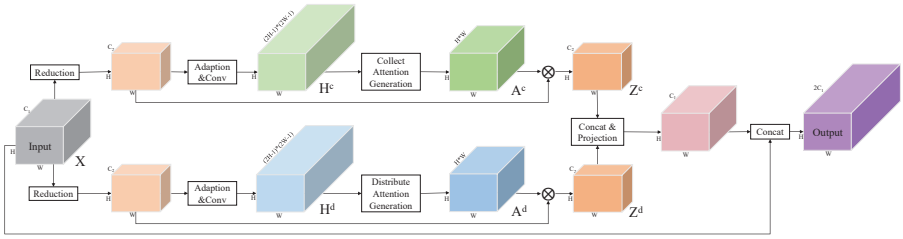


Fig. 2. Architecture of the proposed PSA module.

We aggregate input feature map based on attention maps following Eq. (8) to generate new feature representations with the long-range contextual information incorporated, *i.e.*, $\mathbf{Z}^c$ from the 'collect' branch and $\mathbf{Z}^d$ from the 'distribute' branch.

We concatenate the new representations $\mathbf{Z}^c$ and $\mathbf{Z}^d$ and apply a convolutional layer with batch normalization and activation layers for dimension reduction and feature fusion. Then we concatenate the new global contextual feature with the local representation feature $\mathbf{X}$. It is followed by one or several convolutional layers with batch normalization and activation layers to generate the final feature map for following subnetworks.

We note that all operations in our proposed PSA module are differentiable, and can be jointly trained with other parts of the network in an end-to-end manner. It can be flexibly attached to any feature maps in the network. By predicting contextual dependencies for each position, it adaptively aggregates suitable contextual information. In the following subsections, we detail the process of generating the two attention maps, *i.e.*, $\mathbf{A}^c$ and $\mathbf{A}^d$.

3.3 Point-wise Spatial Attention

Network Structure. PSA module firstly produces two point-wise spatial attention maps, *i.e.*, $\mathbf{A}^c$ and $\mathbf{A}^d$ by two parallel branches. Although they represent different information propagation directions, network structures are just the same. As shown in Fig. 2, in each branch, we firstly apply a convolutional layer with 1×1 filters to reduce the number of channels of input feature map $\mathbf{X}$ to reduce computational overhead (*i.e.*, $C_2 < C_1$ in Fig. 2). Then another convolutional layer with 1×1 filters is applied for feature adaption. These layers are accompanied with batch normalization and activation layers. Finally, one convolutional layer is responsible for generating the global attention map for each position.

Instead of predicting a map with size $H \times W$ for each position i, we predict an over-completed map $\mathbf{h}_i$, *i.e.*, with size $(2H-1) \times (2W-1)$, covering the input feature map. As a result, for feature map $\mathbf{X}$, we get a temporary representation map $\mathbf{H}$ with spatial size $H \times W$ and $(2H-1) \times (2W-1)$ channels. As illustrated

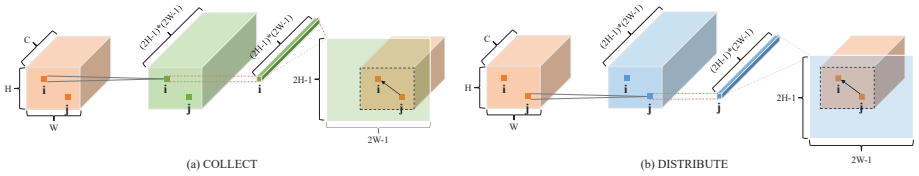


Fig. 3. Illustration of Point-wise Spatial Attention.

by Fig. 3, for each position i, $\mathbf{h}_i$ can be reshaped to a spatial map with $2H-1$ rows and $2W-1$ columns and centers on position i, of which only $H \times W$ values are useful for feature aggregation. The valid region is highlighted as the dashed bounding box in Fig. 3.

With our instantiation, the set of filters used to predict the attention maps at different positions are not the same. This enables the network to be sensitive to the relative positions by adapting weights. Another instantiation to achieve this goal is to utilize a fully-connected layer to connect the input feature map and the predicted pixel-wise attention map. But this will lead to an enormous number of parameters.

Attention Map Generation. Based on the predicted over-completed map $\mathbf{H}^c$ from the 'collect' branch and $\mathbf{H}^d$ from the 'distribute' branch, we further generate attention maps $\mathbf{A}^c$ and $\mathbf{A}^d$, respectively.

In the 'collect' branch, at each position i, with k_{th} row and l_{th} column, we predict how current position is related to other positions based on feature at position i. As a result, $\mathbf{a}_i^c$ corresponds to the region in $\mathbf{h}_i^c$ with H rows and W columns starting from $(H-k)_{th}$ row and $(W-l)_{th}$ column.

Specifically, element at s_{th} row and t_{th} column in attention mask $\mathbf{a}_i^c$, *i.e.*, $\mathbf{a}_{[k,l]}^c$ is

$$\mathbf{a}_{[k,l],[s,t]}^c = \mathbf{h}_{[k,l],[H-k+s,W-l+t]}^c, \quad \forall s \in [0,H), \quad t \in [0,W) \tag{9}$$

where $[\cdot,\cdot]$ indexes position in rows and columns. This attention map helps collect informative in other positions to benefit the prediction at current position.

On the other hand, we distribute information at the current position to other positions. At each position, we predict how important the information at the current position to other positions is. The generation of $\mathbf{a}_i^d$ is similar to $\mathbf{a}_i^c$. This attention map helps to distribute information for better prediction.

These two maps encode the context dependency between different position pairs in a complementary way, leading to improved information propagation and enhanced utilization of long-range context. The benefits of utilizing those two different attentions are manifested in experiments.

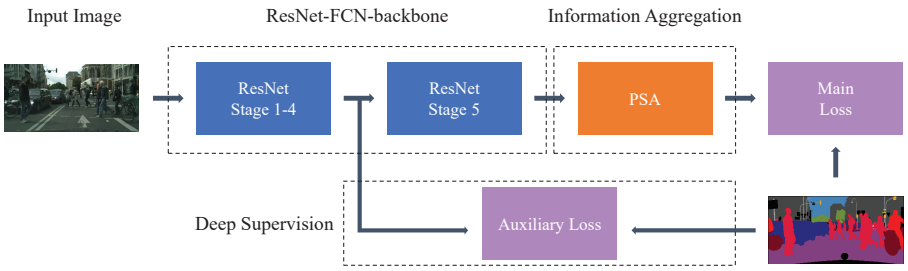


Fig. 4. Network structure of ResNet-FCN-backbone with PSA module incorporated. Deep supervision is also adopted for better performance.

3.4 PSA Module with FCN

Our PSA module is scalable and can be attached to any stage in the FCN structure. We show our instantiation in Fig. 4.

Given an input image **I**, we acquire its local representation through FCN as feature map **X**, which is the input of the PSA module. Same as that of [45], we take ResNet [13] as the FCN backbone. Our proposed PSA module is then used to aggregate long-range contextual information from the local representation. It follows stage-5 in ResNet, which is the final stage of the FCN backbone. Features in stage-5 are semantically stronger. Aggregating them together leads to a more comprehensive representation of long-range context. Moreover, the spatial size of the feature map at stage-5 is smaller and can reduce computation overhead and memory consumption. Referring to [45], we also utilize the same deep supervision technique. An auxiliary loss branch is applied apart from the main loss as illustrated in Fig. 4.

3.5 Discussion

There has been research making use of context information for scene parsing. However, the widely used dilated convolution [4,42] utilized a fixed sparse grid to operate the feature map, losing the ability to utilize information of the entire image. While pooling strategies [24,5,45] captures global context with fixed weight at each position, they can not adapt to the input data and are less flexible. Recently proposed non-local method [40] encodes global context by calculating the correlation of semantic features between each pair of positions on the input feature map, ignoring the relative location between these two positions.

Different from these solutions, our PSA module adaptively predicts global attention maps for each position on the input feature map by convolutional layers, taking the relative location into account. Moreover, the attention maps can be predicted from two perspectives, aiming at capturing different types of information flow between positions. The two attention maps actually build the bi-direction information propagation path as illustrated in Fig. 1. They collect

and distribute information over the entire feature map. The global pooling technique is just a special case of our PSA module in this regard. As a result, our PSA module can effectively capture long-range context information, adapt to input data and utilize diverse attention information, leading to more accurate prediction.

4 Experimental Evaluation

The proposed PSANet is effective on scene parsing and semantic segmentation tasks. We evaluate our method on three challenging datasets, including complex scene understanding dataset ADE20K [48], object segmentation dataset PASCAL VOC 2012 [9] and urban-scene understanding dataset Cityscapes [8]. In the following, we first show the implementation details related to training strategy and hyper-parameters, then we show results on corresponding datasets and visualize the learned masks generated by the PSA module.

4.1 Implementation Details

We conduct our experiments based on Caffe [15]. During training, we set the mini-batch size as 16 with synchronized batch normalization and base learning rate as 0.01. Following prior works [5,45], we adopt 'poly' learning rate policy and the power is set to 0.9. We set maximum iteration number to 150K for experiments on the ADE20K dataset, 30K for VOC 2012 and 90K for Cityscapes. Momentum and weight decay are set to 0.9 and 0.0001 respectively. For data augmentation, we adopt random mirror and random resize between 0.5 and 2 for all datasets. We further add extra random rotation between -10 and 10 degrees, and random Gaussian blur for ADE20K and VOC 2012 datasets.

4.2 ADE20K

The scene parsing dataset ADE20K [48] is challenging for up to 150 classes and diverse complex scenes up to 1,038 image-level categories. It is divided into 20K/2K/3K for training, validation and testing, respectively. Both objects and stuffs need to be parsed for the dataset. For evaluation metrics, both *mean of class-wise intersection over union* (Mean IoU) and *pixel-wise accuracy* (Pixel Acc.) are adopted.

Comparison of Information Aggregation Approaches. We compare the performance of several different information aggregation approaches on the validation set of ADE20K with two network backbones, *i.e.*, ResNet with 50 and 101 layers. The experimental results are listed in Table 1. Our baseline network is ResNet-based FCN with dilated convolution module incorporated at stage 4 and 5, *i.e.*, dilations are set to 2 and 4 for these two stages respectively.

Table 1. Contextual information aggregation with different approaches. Results are reported on *validation* set of ADE20K dataset. 'SS' stands for single-scale testing and 'MS' means multi-scale testing strategy is utilized.

Method	Mean IoU(%) / Pixle Acc.(%)	
	SS	MS
ResNet50-Baseline	37.23/78.01	38.48/78.92
ResNet50+DenseCRF[a] [18]	37.97/78.51	38.86/79.32
ResNet50+GlobalPooling [24]	40.07/79.52	41.22/80.35
ResNet50+ASPP [5]	40.39/79.71	42.18/80.73
ResNet50+NonLocal [40]	40.93/79.97	41.94/80.71
ResNet50+PSP [45]	41.68/80.04	42.78/80.76
ResNet50+COLLECT(Compact)	41.07/79.61	41.99/80.32
ResNet50+COLLECT	41.27/79.74	42.56/80.56
ResNet50+DISTRIBUTE	41.46/80.12	42.63/80.90
ResNet50+COLLECT+DISTRIBUTE	41.92/80.17	42.97/80.92
ResNet101-Baseline	39.66/79.44	40.71/80.17
ResNet101+COLLECT	42.70/80.53	43.68/81.24
ResNet101+DISTRIBUTE	42.11/80.01	43.38/81.12
ResNet101+COLLECT+DISTRIBUTE	42.75/80.71	43.77/81.51

[a] CRF parameters: bi_w=3.5, bi_xy_std=55, bi_rgb_std=3, pos_w=2, pos_xy_std=1.

Based on the feature map extracted by FCN, DenseCRF[18] only brings slight improvement. Global pooling [24] is a simple and intuitive attempt to harvest long-range contextual information, but it treats each position on the feature map equally. Pyramid structures [5,45] with several branches can capture contextual information at different scales. Another option is to use an attention mask for each position in the feature map. A non-local method was adopted in [40], in which attention mask for each position is generated by calculating the feature correlation between each paired positions. In our PSA module, apart from the uniqueness of the attention mask for each point, our point-wise masks are self-adaptively learned with convolutional operations instead of simply matrix multiplication adopted by non-local method [40]. Compared with these information aggregation methods, our method performs better, which shows that the PSA module is a better choice in terms of capturing long-range contextual information.

We further explore the two branches in our PSA module. Taking ResNet50 as an example with information flow in 'collect' mode (denoted as '+COLLECT') in Table 1, our single scale testing results get 41.27/79.74 in terms of Mean IoU and Pixel Acc. (%)., exceeding the baseline by 4.04/1.73. This significant improvement demonstrates the effectiveness of our proposed PSA module, even with only uni-directional information flow in a simplified version. With our bi-direction information flow model (denoted as '+COLLECT +DISTRIBUTE'), the performance further increases to 41.92/80.17, outperforming the baseline model by 4.69/2.16 in terms of absolute improvement and 12.60/2.77 in terms

Table 2. Methods comparison with results reported on ADE20K *validation* set.

Method	Mean IoU(%)	Pixel Acc.(%)
FCN-8s [26]	29.39	71.32
SegNet [1]	21.64	71.00
DilatedNet [42]	32.31	73.55
CascadeNet [48]	34.90	74.52
RefineNet101 [21]	40.20	-
RefineNet152 [21]	40.70	-
PSPNet50 [45]	42.78	80.76
PSPNet101 [45]	43.29	81.39
WiderNet [41]	43.73	81.17
PSANet50	42.97	80.92
PSANet101	43.77	81.51

Table 3. Methods comparison with results reported on VOC 2012 *test* set.

Method	mIoU(%)
LRR [11]	79.3
DeepLabv2 [5]	79.7
G-CRF [3]	80.4
SegModel [34]	81.8
LC [20]	82.7
DUC_HDC [39]	83.1
Large_Kernel_Matters [31]	83.6
RefineNet [21]	84.2
ResNet-38 [41]	84.9
PSPNet [45]	85.4
DeepLabv3 [6]	85.7
PSANet	85.7

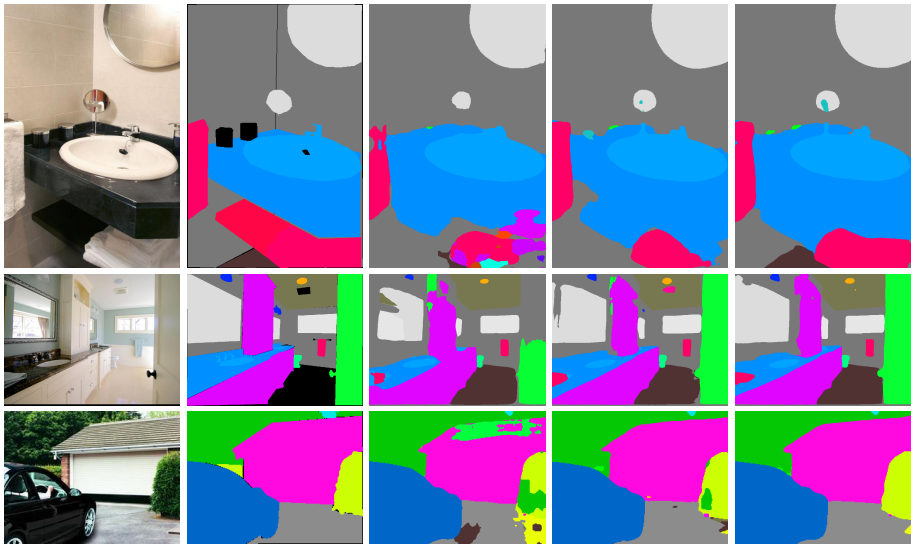
of relative improvement. The improvement is just general to backbone networks. This manifests that both of the two information propagation paths are effective and complementary to each other. Also note that our location-sensitive mask generation strategy plays a key role for our high performance. Method denoted as '(compact)' means compact masks are generated with size $H \times W$ instead of the over-completed ones with doubled size, ignoring the relative location information. The performance is higher if the relative location is taken into account. However, the 'compact' method outperforms the 'non-local' method, which also indicates that the long-range dependency adaptively learned from the feature map as we propose is better than that calculated from the feature correlation.

Method Comparison. We show the comparison between our method and others in Table 2. With the same network backbone, PSANet gets higher performance than those of RefineNet [21] and PSPNet [45]. PSANet50 even outperforms RefineNet with much deeper ResNet-152 as the backbone. It is slightly better than WiderNet [41] that uses a powerful backbone called Wider ResNet.

Visual Improvements. We show the visual comparison of the parsing results in Fig. 5. PSANet much improves the segmentation quality, where more accurate and detailed predictions are generated compared to the one without the PSA module. We include more visual comparisons between PSANet and other approaches in the supplementary material.

4.3 PASCAL VOC 2012

PASCAL VOC 2012 segmentation dataset [9] is for object-centric segmentation and contains 20 object classes and one background. Following prior works [4,5,45], we utilize the augmented annotations from [12] resulting 10,582, 1,449 and 1,456 images for training, validation and testing. Our introduced PSA module is also very effective for object segmentation as shown in Table 4. It boosts the performance greatly, exceeding the baseline by a large margin.



(a) Image (b) Ground Truth (c) Baseline (d) PSA-COL (e) PSA-COL-DIS

Fig. 5. Visual improvement on *validation* set of ADE20K. The proposed PSANet gets more accurate and detailed parsing results. 'PSA-COL' denotes PSANet with 'COLLECT' branch and 'PSA-COL-DIS' stands for bi-direction information flow mode, which further enhances the prediction.

Table 4. Improvements introduced by PSA module. Results are reported with models trained on *train_aug* set and evaluated on *val* set of VOC 2012.

Method	Mean IoU(%) / Pixle Acc.(%) SS	MS
Res50-Baseline	67.12/92.83	67.57/92.98
+COL	76.96/94.79	78.00/95.01
+COL+DIS	77.24/94.88	78.14/95.12
Res101-Baseline	70.64/93.82	71.22/93.95
+COL	77.90/95.02	79.07/95.32
+COL+DIS	78.51/95.18	79.77/95.43

Table 5. Improvements introduced by PSA module. Results are reported with models trained on *fine_train* set and evaluated on *fine_val* set of Cityscapes.

Method	Mean IoU(%) / Pixle Acc.(%) SS	MS
Res50-Baseline	71.93/95.53	72.99/95.76
+COL	76.51/95.95	77.50/96.15
+COL+DIS	76.65/95.99	77.79/96.24
Res101-Baseline	74.83/96.03	75.89/96.23
+COL	77.06/96.18	78.05/96.39
+COL+DIS	77.94/96.10	79.05/96.30

Following methods of [4,5,45,6], we also pre-train on the MS-COCO [23] dataset and then finely tune the system on the VOC dataset. Table 3 lists the performance of different frameworks on VOC 2012 test set – PSANet achieves top performance. Visual improvement is clear as shown in the supplementary material. Similarly, better prediction is yielded with PSA module incorporated.

4.4 Cityscapes

Cityscapes dataset [8] is collected for urban scene understanding. It contains 5,000 finely annotated images divided into 2,975, 500, and 1,525 images for

Table 6. Methods comparison with results reported on Cityscapes *test* set. Methods trained using both *fine* and *coarse* data are marked with †.

Method	mIoU(%)
DeepLabv2 [5]	70.4
LC [20]	71.1
Adelaide [22]	71.6
FRRN [32]	71.8
RefineNet [21]	73.6
PEARL [16]	75.4
DUC_HDC [39]	77.6
SAC [43]	78.1
PSPNet[a] [45]	78.4
ResNet-38 [41]	78.5
SegModel [34]	78.5
Multitask Learning [17]	78.5
PSANet[a]	78.6
PSANet[b]	80.1

Method	mIoU(%)
LRR-4x† [11]	71.8
SegModel† [34]	79.2
DUC_HDC† [39]	80.1
Netwarp† [10]	80.5
ResNet-38† [41]	80.6
PSPNet† [45]	81.2
DeepLabv3† [6]	81.3
PSANet†	81.4

[a] Trained with *fine_train* set only
[b] Trained with *fine_train* + *fine_val* set

training, validation and testing. 30 common classes of road, person, car, etc. are annotated and 19 of them are used for semantic segmentation evaluation. Besides, another 20,000 coarsely annotated images are also provided.

We first show the improvement brought by our PSA module based on the baseline method in Table 5 and then list the comparison between different methods on test set in Table 6 with two settings, *i.e.*, training with only *fine* data and training with *coarse+fine* data. PSANet achieves the best performance under both settings. Several visual predictions are included in the supplementary material.

4.5 Mask visualization

To get a deeper understanding of our PSA module, we visualize the learned attention masks as shown in Fig. 6. The images are from the validation set of ADE20k. For each input image, we show masks at two points (red and blue ones), denoted as the red and blue ones. For each point, we show the mask generated by both 'COLLECT' and 'DISTRIBUTE' branches. We find that attention masks pay low attention at the current position. This is reasonable because the aggregated feature representation is concatenated with the original local feature, which already contains local information.

We find that our attention mask effectively focuses on related regions for better performance. For example in the first row, the mask for the red point, which locates on the beach, assigned a larger weight to the sea and beach which is beneficial to the prediction of red point. While the attention mask for the blue point in the sky assigns a higher weight to other sky regions. A similar trend is also spotted in other images.

(a) Input Image (b) RED-COL (c) RED-DIS (d) BLUE-COL (e) BLUE-DIS

Fig. 6. Visualization of learned masks by PSANet. Masks are sensitive to location and category information that harvest different contextual information.

The visualized masks confirm the design intuition of our module, in which each position gather informative contextual information from regions both nearby and far away for better prediction.

5 Concluding Remarks

We have presented the PSA module for scene parsing. It adaptively predicts two global attention maps for each position in the feature map by convolutional layers. Position-specific bi-directional information propagation is enabled for better performance. By aggregating information with the global attention maps, long-range contextual information is effectively captured. Extensive experiments with top ranking scene parsing performance on three challenging datasets demonstrate the effectiveness and generality of the proposed approach. We believe the proposed module can advance related techniques in the community.

Acknowledgments

This work is partially supported by The Early Career Scheme (ECS) of Hong Kong (No.24204215). We thank Sensetime Research for providing computing resources.

References

1. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. TPAMI (2017)
2. Chandra, S., Kokkinos, I.: Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In: ECCV (2016)
3. Chandra, S., Usunier, N., Kokkinos, I.: Dense and low-rank gaussian crfs using deep embeddings. In: ICCV (2017)
4. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. ICLR (2015)
5. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. TPAMI (2018)
6. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv:1706.05587 (2017)
7. Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR (2016)
8. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: CVPR (2016)
9. Everingham, M., Gool, L.J.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes VOC challenge. IJCV (2010)
10. Gadde, R., Jampani, V., Gehler, P.V.: Semantic video cnns through representation warping. In: ICCV (2017)
11. Ghiasi, G., Fowlkes, C.C.: Laplacian pyramid reconstruction and refinement for semantic segmentation. In: ECCV (2016)
12. Hariharan, B., Arbelaez, P., Bourdev, L.D., Maji, S., Malik, J.: Semantic contours from inverse detectors. In: ICCV (2011)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
14. Huang, G., Liu, Z., Weinberger, K.Q., van der Maaten, L.: Densely connected convolutional networks. In: CVPR (2017)
15. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM MM (2014)
16. Jin, X., Li, X., Xiao, H., Shen, X., Lin, Z., Yang, J., Chen, Y., Dong, J., Liu, L., Jie, Z., et al.: Video scene parsing with predictive feature learning. In: ICCV (2017)
17. Kendall, A., Gal, Y., Cipolla, R.: Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In: CVPR (2018)
18. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: NIPS (2011)
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS (2012)
20. Li, X., Liu, Z., Luo, P., Loy, C.C., Tang, X.: Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade. In: CVPR (2017)
21. Lin, G., Milan, A., Shen, C., Reid, I.D.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: CVPR (2017)
22. Lin, G., Shen, C., Reid, I.D., van den Hengel, A.: Efficient piecewise training of deep structured models for semantic segmentation. In: CVPR (2016)

23. Lin, T., Maire, M., Belongie, S.J., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: ECCV (2014)
24. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. arXiv:1506.04579 (2015)
25. Liu, Z., Li, X., Luo, P., Loy, C.C., Tang, X.: Semantic image segmentation via deep parsing network. In: ICCV (2015)
26. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: CVPR (2015)
27. Luo, W., Li, Y., Urtasun, R., Zemel, R.: Understanding the effective receptive field in deep convolutional neural networks. In: NIPS (2016)
28. Mnih, V., Heess, N., Graves, A., et al.: Recurrent models of visual attention. In: NIPS (2014)
29. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: ICCV (2015)
30. Paszke, A., Chaurasia, A., Kim, S., Culurciello, E.: Enet: A deep neural network architecture for real-time semantic segmentation. arXiv:1606.02147 (2016)
31. Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J.: Large kernel matters–improve semantic segmentation by global convolutional network. In: CVPR (2017)
32. Pohlen, T., Hermans, A., Mathias, M., Leibe, B.: Full-resolution residual networks for semantic segmentation in street scenes. In: CVPR (2017)
33. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
34. Shen, F., Gan, R., Yan, S., Zeng, G.: Semantic segmentation via structured patch prediction, context crf and guidance crf. In: CVPR (2017)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR (2015)
36. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S.E., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: CVPR (2015)
37. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: NIPS (2017)
38. Visin, F., Romero, A., Cho, K., Matteucci, M., Ciccone, M., Kastner, K., Bengio, Y., Courville, A.: Reseg: A recurrent neural network-based model for semantic segmentation. In: CVPR Workshop (2016)
39. Wang, P., Chen, P., Yuan, Y., Liu, D., Huang, Z., Hou, X., Cottrell, G.W.: Understanding convolution for semantic segmentation. In: WACV (2018)
40. Wang, X., Girshick, R., Gupta, A., He, K.: Non-local neural networks. In: CVPR (2018)
41. Wu, Z., Shen, C., van den Hengel, A.: Wider or deeper: Revisiting the resnet model for visual recognition. arXiv:1611.10080 (2016)
42. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. ICLR (2016)
43. Zhang, R., Tang, S., Zhang, Y., Li, J., Yan, S.: Scale-adaptive convolutions for scene parsing. In: ICCV (2017)
44. Zhao, H., Qi, X., Shen, X., Shi, J., Jia, J.: ICNet for real-time semantic segmentation on high-resolution images. In: ECCV (2018)
45. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR (2017)
46. Zheng, S., Jayasumana, S., Romera-Paredes, B., Vineet, V., Su, Z., Du, D., Huang, C., Torr, P.H.S.: Conditional random fields as recurrent neural networks. In: ICCV (2015)

47. Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., Torralba, A.: Object detectors emerge in deep scene cnns. ICLR (2015)
48. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20K dataset. In: CVPR (2017)