Received: 7 August 2019 | Revised: 21 October 2020 | Accepted: 25 November 2020

IET Computer Vision

DOI: 10.1049/cvi2.12026

The Institution of Engineering and Technology WILEY

ORIGINAL RESEARCH PAPER

# Enhancing feature fusion with spatial aggregation and channel fusion for semantic segmentation

Jie Hu | Huifang Kong | Lei Fan | Jun Zhou

School of Electrical Engineering and Automation, Hefei University of Technology, Hefei, China

**Correspondence**

Huifang Kong; School of Electrical Engineering and Automation, Hefei University of Technology, Hefei, China.
Email: konghuifang@hfut.edu.cn

**Funding information**

Institute of Intelligent Manufacturing Technology, Hefei University of Technology, Grant/Award Number: IMICZ2017004; The 111 Project BP0719039, Grant/Award Number: BP0719039

**Abstract**

Semantic segmentation is crucial to the autonomous driving, as an accurate recognition and location of the surrounding scenes can be provided for the street scenes understanding task. Many existing segmentation networks usually fuse high-level and low-level features to boost segmentation performance. However, the simple fusion may impose a limited performance improvement because of the gap between high-level and low-level features. To alleviate this limitation, we respectively propose *spatial aggregation* and *channel fusion* to bridge the gap. Our implementation, inspired by the attention mechanism, consists of two steps: (1) Spatial aggregation relies on the proposed pyramid spatial context aggregation module to capture spatial similarities to enhance the spatial representation of high-level features, which is more effective for the latter fusion. (2) Channel fusion relies on the proposed attention-based channel fusion module to weight channel maps on different levels to enhance the fusion. In addition, the complete network with U-shape structure is constructed. A series of ablation experiments are conducted to demonstrate the effectiveness of our designs, and the network achieves mIoU score of 81.4% on Cityscapes test dataset and 84.6% on PASCALVOC 2012 test dataset.

## 1 | INTRODUCTION

Recent years, there is an increasing interest in autonomous driving cars and autonomous driving assistance systems. An essential aspect of autonomous driving is to acquire an accurate understanding of the semantic entities usually found in street scenes, such as the cars, pedestrians, road or sidewalks in the Figure 1. Semantic segmentation, a kind of dense pixel-wise prediction task, can provide abundant information (e.g. object category, location and the shape of each element) for comprehensive scene description. Thanks to the excellent scene recognition performance of semantic segmentation, it has become an important tool for the street scenes understanding task in autonomous driving [1, 2].

Semantic segmentation aims at assigning each pixel in the input image with corresponding semantic labels, which pose strong demand for both semantic and spatial information. The semantic information plays a crucial role in achieving high performance of recognition (i.e. what is visible?) and spatial information is important to boost localization accuracy (i.e. where precisely is something located?). In this study, we pursue the goal of achieving high-quality semantic segmentation by enhancing the correlation between semantic information and spatial information.

Currently, based on the pioneering fully convolutional network (FCN) [3] that takes image as input and outputs a probability map for each class, many state-of-the-art segmentation networks [4–19] have achieved remarkable progress. When performing some form of FCN to boost segmentation performance, the U-shape structure [7, 16, 17, 19] is demonstrated to an effective framework to capture information from both semantic and spatial levels. As shown in Figure 2(a), the U-shape network uses one contracting path to compute high-level features for capturing the semantic information and one symmetric path to fuse low-level features for recovering the spatial information. Low-level and high-level features are complementary by nature, where low-level features are rich in spatial details but lack semantic information and vice versa [18]. Although the networks based on U-shape structure can improve segmentation

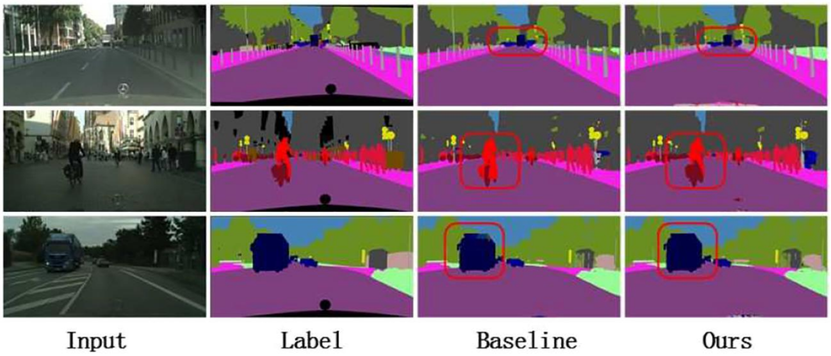


**FIGURE 1** Illustration of the comprehensive semantic entities in the street scenes

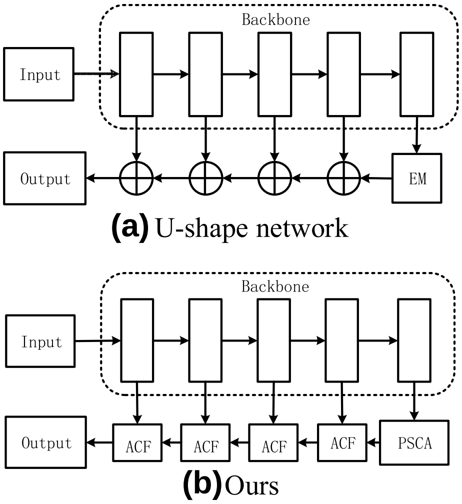


**FIGURE 2** Detailed structure for the U-shape network. EM: encoder module. (a) U-shape network captures context information from both semantic and spatial levels. (b) our network enhances the feature fusion to boost segmentation performance

performance, they fuse the features between low-level and high-level by combining channel maps simply. This simple fusion would result in the limited performance improvement because of the gap between low-level and high-level features. Several methods exist to overcome this challenge and obtain remarkable progress. Zhang et al. [18] proposes a new network, named ExFuse, to bridge the gap between low-level features and high-level features. PAN [15] and DFN [19] adopt high-level features to weight low-level features for selecting precise spatial details.

In contrast to above methods, we rethink the feature fusion from a more macroscopic point of view. In this way, we regard the feature fusion as a task to explore the correlation between high-level features and low-level features rather than a simple fusion of channel maps. Considering that the high-level features are rich in semantic information details but lack spatial details and the gap between high-level features and low-level features is mainly caused by the imbalance between semantic information and spatial information, we propose to enhance the feature fusion from two aspects: (1) *Spatial aggregation*, aggregating more spatial information into the high-level features for enhancing spatial correlation between the high-level features and low-level features, which are more effective for the latter channel fusion. (2) *Channel fusion*, weighting channel maps on different levels to enhance the feature fusion. Specifically, we construct two attention-based modules for the implementation of spatial aggregation and channel fusion. The attention mechanism [20–24], an important mechanism to select more distinct and informative features, has proven to be an effective method in computer vision tasks. For spatial aggregation, the pyramid spatial context aggregation (PSCA) module is designed, which exploits the attention mechanism to capture the spatial similarities between any two positions. Guided by the spatial similarities, a spatial attention map is generated to update each position of the high-level features. That is to say, spatial representation of high-level features can be enhanced without losing semantic representation. For channel fusion, we build a so-called attention-based channel fusion (ACF) module to generate a weight vector for changing the weights of channel maps on each level, through which low-level and high-level features can be adaptively fused. This effective feature fusion can provide rich feature representations from both semantic and spatial levels for the final prediction. Based on PSCA and ACF modules, we build our segmentation network following the U-shape structure, as shown in Figure 2(b).

In summary, the main contributions of our study include three-fold:

1. We introduce the spatial aggregation and channel fusion to emphasize the spatial correlation and channel interdependence between low-level and high-level features, respectively for enhancing the feature fusion.
2. We propose two attention-based modules (PSCA and ACF modules) for spatial aggregation and channel fusion, and design a network based on U-shape structure.
3. We conduct a series of ablation experiments to demonstrate the effectiveness of our method. And our network achieves comparable performance to some state-of-the-art works.

The remainder of this study is organized as follows. In Section 2, we review some works that have far-reaching effects for semantic segmentation. Section 3 presents our proposed modules and network. In Section 4, we conduct ablation experiments. In Section 5, the conclusion is drawn.

# 2 | RELATED WORKS

## 2.1 | Semantic segmentation

The last years, semantic segmentation has achieved great progress with FCN [3], which was the earliest approach to introduce full convolution into semantic segmentation. Later, many networks based on FCN were proposed to capture multiscale information to alleviate the problem caused by multiscale objects. PSPNet [12] designed a pyramid pooling module to collect the effective contextual prior, containing information of different scales. DenseASPP [14] brought dense connections into atrous spatial pyramid pooling (ASPP) to generate features with various scale. To further enhance feature representation, some works exploited contextual dependencies to generate dense and pixel-wise contextual information. PSANet [26] learnt to aggregate contextual information for each position via a predicted attention map. Liu et al. [27] and Visin et al. [28] utilized RNNs to capture long-range contextual dependency information. In addition, the encoder-decoder framework was widely used in existing works to capture context information from both semantic and spatial levels. For instance, RefineNet [7], U-net [16], SegNet [29] and ParseNet [30] utilized multipath and symmetrical subnetworks (U-shape structure) to fuse the low-level feature for refining prediction results. ENet [31] and ERFNet [32] explored more efficient encoder-decoder framework and multiple optimization methods for semantic segmentation task. Motivated by above methods, our approach investigates the spatial details in high-level features to enhance contextual dependencies and utilizes the multipath and symmetrical subnetworks to adaptively fuse low-level feature and high-level feature.

## 2.2 | Attention mechanism

In computer vision tasks, attention mechanism can be considered as an important mechanism to weight features for selecting more distinct and informative features. SENet [33] enhanced the representational power of the network by modelling channel-wise relationship, which obtained significant improvement on image classification. Woo et al. [34] established a convolutional block attention module to select feature maps along with channel and spatial axes respectively. Wang et al. [35] proposed the nonlocal neural network to calculate the context at one position as a weighted sum of all positions for guiding dense contextual information aggregation. GCNet [36] simplified the nonlocal network and abstracted this simplified version to a global context modelling framework, which can effectively model long-range dependency.

For scene understanding, attention mechanism, as a selector not only was used to select discriminative features, but also was adopted to model spatial dependencies and channel dependencies. PAN [15] and DFN [19] suggested that high-level features can be used as guidance to weight low-level features for selecting precise spatial details. EncNet [37] introduced an encoding layer to capture the semantic information. OCNet [38] proposed an object context pooling (OCP) to represent each pixel by exploiting the set of pixels that belong to the same object category. CCNet [39] proposed the criss-cross attention module on the criss-cross path to aggregate contextual information. DANet [40] explored long-range dependencies over local features across spatial dimensions and channel dimensions simultaneously.

# 3 | METHOD

In this section, we present the details of our study. First, we specifically introduce the two attention-based modules for spatial aggregation and channel fusion. Then, we present the complete architecture of our network, which adopts U-shape structure.

## 3.1 | Spatial aggregation

To encode a wider range of spatial information into high-level features, spatial aggregation is introduced with considering two crucial factors simultaneously: global context dependence and multilevel features. Specifically, the spatial context aggregation (SCA) block is designed to capture global context dependence over the spatial locations, and the PSCA module is further proposed to exploit SCA block to scan the whole image over both positions and pyramid levels. Next, we elaborate the implementation details of SCA block and PSCA module.

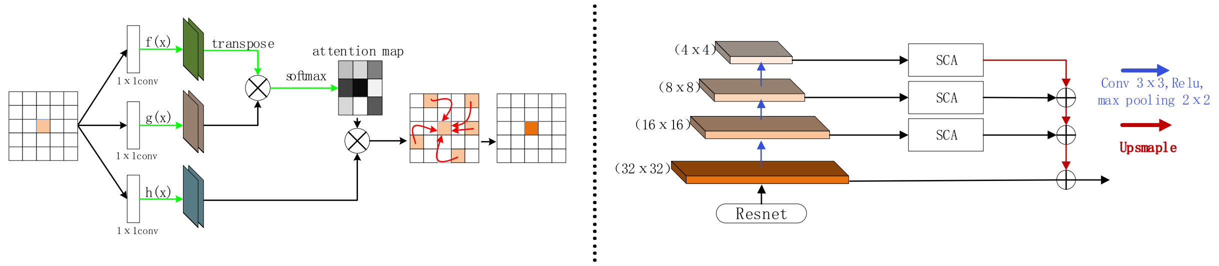
### 3.1.1 | Spatial context aggregation block

The intuition of spatial aggregation is to utilize the spatial similarities to guide global contextual information aggregation, which helps enhance the spatial representation of the high-level features.

As illustrated in Figure 3(a), given a local feature $X$, we denote $x = \{xi\}_{i=1}^{Np}$ as the feature map of one input image, where $Np = H \times W$ is the number of positions in the feature map. We first feed $X$ into a convolution layer with batch normalization and ReLU layers to generate two feature spaces $f$ and $g$. After that we perform a matrix multiplication between the transpose of $f(xi)$ and $g(xj)$, and apply a softmax layer to calculate the spatial attention map$S$:

$$sji = \frac{\exp(\beta ij)}{\sum_{i=1}^{Np} \exp(\beta ij)}, \text{ where } \beta ij = f(xi)^{T} g(xj) \qquad (1)$$

where the spatial attention map $Sji$ is a representation vector for the global position relationship, $\beta ij$ indicates the extent to which the module attends to the $i^{th}$ position when synthesizing the $j^{th}$ position, $f()$ is the query-position transform function and $g()$ is the key-position transform function. Note that the more similar spatial representations of the two positions $i^{th}$ and $j^{th}$, the greater correlation between them. Finally, we construct the spatial aggregation for each key-position by performing a



**(a)** Spatial Context Aggregation block

**(b)** Pyramid Spatial Context Aggregation module

**FIGURE 3** The details of spatial context aggregation (SCA) block and pyramid spatial context aggregation (PSCA) module are illustrated in (a) and (b)

matrix multiplication between the spatial attention map$S$ and the original feature map, as below:

$$yj = \sum_{i=1}^{Np} sji \times h(xi) \qquad (2)$$

where $h()$ is the value transform function following the self-attention. Therefore, the final feature representations achieve mutual compensation.

### 3.1.2 | Pyramid spatial context aggregation module

As illustrated in Figure 3(b), we further implement SCA for the feature maps at multiple levels respectively by introducing a feature pyramid (FP) [41], which has been heavily used in the era of object detection. Then we construct the PSCA module, which involves a bottom-up pathway, lateral connections and a top-down pathway.

The bottom-up pathway is the feed forward computation based on the backbone ResNet-101 [42], which takes a single-scale image as input and computes a pyramidal hierarchy consisting of feature maps at several levels with the pooling operation. For the FP, we denote the outputs of each pyramid level as (P1, P2, P3, P4), which have the resolution of (32×32, 16×16, 8×8, 4×4). We choose the outputs of each pyramid level as our reference set of feature maps, which are integrated to the lateral connections for enriching spatial context information.

The lateral connections utilize the SCA block to compute spatial aggregation for the feature maps at different pyramidal hierarchies, respectively, which generate a multilevel feature representation. This design is natural since the higher levels of FP should have the stronger features. Note that we do not apply SCA block toP1 due to its large memory footprint.

The top-down pathway is an extended path symmetrical to the bottom-up pathway, feature maps from higher pyramid levels and then merges feature maps of the same spatial size from each lateral connection. This process of up-sampling and merging is iterated until the finest resolution maps from P1 are merged. We append a 3×3 convolution on each merged map for reducing the aliasing effect of up-sampling. We fix the feature dimension in all the feature maps, and the final convolutional layer has 512-channel outputs.

Benefit from the simple design, PSCA module can aggregate spatial context information at different semantic levels, which strengthen the spatial representation for the high-level features without losing the semantic representation.

## 3.2 | Channel fusion

Channel maps located in low-level and high-level feature maps can be regarded as the response of spatial and semantic information, respectively, and different responses are associated with each other. Based on the spatial aggregation proposed above, we exploit the interdependent channel maps on different levels to enhance the feature fusion. Specifically, the ACF module is proposed to generate a weight vector for changing the weights of channel maps between high-level and low-level feature maps, through which low-level and high-level features can be adaptively fused.

In general, a common way of feature fusion in the U-shape structure is to formulate as a residual form:

$$yl = \text{Upsample}(yl + 1) + xl \qquad (3)$$

where $y_l$ is the obtained feature by fusing $l$-$th$ level feature, $y_l + 1$ stands for the $l + 1$-$th$ feature generated by the encoder module, which has higher semantic but lower spatial information, and $x_l$ is the feature extracted from the backbone, which has higher spatial resolution but lower semantic level. Equation (3) implicitly indicates that the fusion weights between channels of different levels are equal, which is considered to be simple feature fusion. However, due to the features in different levels have different degrees of discrimination, simple feature fusion results in the semantic and spatial gap between low-level and high-level features. Therefore, we build the ACF module to explicitly model interdependencies between channels for enhancing the channel fusion by changing the weights of the feature maps on each levels, which are inspired by the attention mechanism [15,19].

As illustrated in Figure 4, two input feature maps of ACF module are denoted as high-level features $F_h$ and low-level

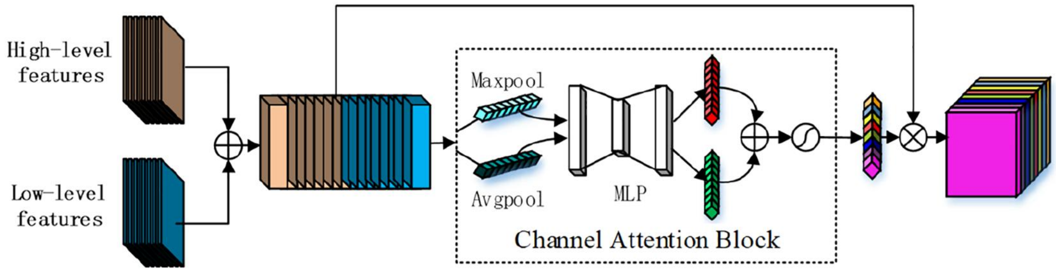


**FIGURE 4** Attention-based channel fusion (ACF) module structure

features $F_l$ respectively, and we directly calculate the feature maps $F_f$ by the simple feature fusion of $F_h$ and $F_l$. Then, the feature maps $F_f$ pass through the channel attention block to generate channel attention vector $M_c$. Specifically, we exploit channel dependencies by applying both average and max pooling operators on channel feature maps $F_f$, producing two different channel description vectors: $V^c_{\max}$ and $V^c_{ave}$. Both vectors are then forwarded to a shared network composed of multilayer perceptron (MLP) with one hidden layer. After the shared network is applied to each descriptor, we merge two output attention vectors using element-wise summation to generate our channel attention vector $M_c$, which is calculated as Equation (4):

$$Mc = \delta\left(\mathrm{MLP}\left(\frac{1}{s^2}\sum_{i,m\in[i,i+s]} \mathrm{P}(Ff)ave\right) \oplus \mathrm{MLP}\left(\max_{i,m\in[i,i+s]} \mathrm{P}(Ff)\max\right)\right) \tag{4}$$

where $\delta$ is the sigmoid activation function, $\oplus$ is element-wise addition, $P$ denotes the pooling function and $s$ is the total number of pixels in feature maps $F_f$. Finally, $M_c$ is used to weight the channel maps of high level and low level to obtain final output $F_o$, as given in Equation (5):

$$Fo = Ff \times Mc \tag{5}$$

With this design, the ACF module is used to emphasize interdependent features stage-wise to enhance channel fusion, which is vital to bridge the gap between low-level and high-level features.

## 3.3 | Network architecture

With the PSCA and ACF modules, we design the complete network, which is given in Figure 5. Our network based on the U-shape structure consists of two paths, one contracting path to compute the high-level features, and the other symmetric expanding path to fuse low-level features of adjacent stages.

First, in the contraction path, we use pretrained ResNet-101 [42] as our backbone, similar to the PAN [15]. The ResNet-101 is widely used to extract the low-level features. Our approach feeds the input image to the ResNet-101, outputting a feature map X of size W×H×C. In detail, we divide the ResNet-101 into five blocks Res2 to Res5. In order to retain more details, we remove the last two down-sampling operations and use atrous convolutions [48] in the Res4 and Res5 blocks, thus enlarging the feature map size of the Res5 to one-eighth of the input image. Next, the feature maps from Res5 are fed into our proposed PSCA module. The PSCA module aggregates spatial context information, which spans level from low to high. Through this path, more spatial information is embedded into the high-level features without losing semantic information, which is better for the latter channel fusion. Then, in the symmetric expanding path, the proposed ACF module is used to adaptively fuse high-level and low-level features from Res2 to Res4 layer by layer. In this process, attention mechanism is used to change the weights of channel maps on each level for enhancing the channel fusion. Finally, through the U-shape structure, we get our final prediction map. With the support of this network, the feature fusion of the low-level and high-level features is enhanced to boost the segmentation performance.

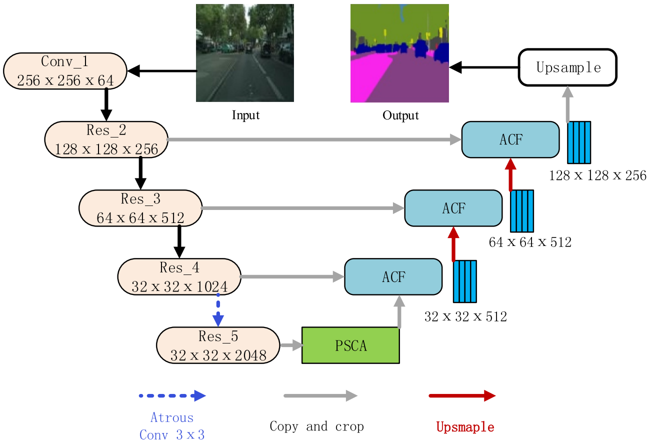


**FIGURE 5** Overview of the proposed network for semantic segmentation

## 4 | EXPERIMENT

In this section, we carried out comprehensive experiments on Cityscapes dataset [25] and PASCAL VOC 2012 [50] to evaluate the proposed method. First, we introduced the dataset and implementation details, then the contributions of each module were investigated in ablation experiments. Finally, we performed comprehensive experiments on Cityscapes and PASCAL VOC 2012 benchmark and the test results are submitted to a dedicated evaluation server for comparison to other methods. For quantitative evaluation, mean of class-wise Intersection over Union (mIoU) are used.

## 4.1 | Experimental settings

### 4.1.1 | Cityscapes dataset

The Cityscapes dataset [25] is composed of a large, diverse set of high resolution (2048 × 1024) images recorded in street scenes, where 5000 of these images have high-quality pixel-level labels of 19 classes and results $9.43 \times 10^9$ labelled pixels in total. High quality images include 2975 images for training, 500 images for validation, as well as 1525 images for testing.

### 4.1.2 | PASCAL VOC 2012 dataset

This PASCAL VOC 2012 [50] benchmark contains 20 foreground object classes and one background class. The original dataset contains 1464 (training), 1449 (validation) and 1456 (testing) pixel-level annotated images.

### 4.1.3 | Implementation details

We implemented our experiments on the public platform PyTorch [43] with four NVIDIA GTX1080ti cards.

For data augmentation, common data augmentations were used as preprocessing, including mean subtraction, random flipping horizontally and random scaling in the range of [0.5, 2].

For training, following the previous studies [30, 44], we used the 'poly' learning rate strategy where the learning rate is shown in Equation (6):

$$lr = lrbase * \left(1 - \frac{epoch}{\mathrm{max}epoch}\right)^{power} \quad (6)$$

where $lr_{base}$ denoted the base learning rate, $power$ denoted decayed index and max_$epoch$ denoted the number of total epochs. Specifically, we empirically set $lr_{base}$ to 0.01, $power$ to 0.9 and weight decay to 0.0001. The training was performed by using minibatch stochastic gradient descent [45] with batch size eight and momentum 0.99. We used 60 K training iterations.

## 4.2 | Ablation studies for attention-based modules

We used the ResNet-101 as our baseline, and directly upsampled the output. First, we evaluated the performance of

TABLE 1 Performance of ablation studies on Cityscapes val set

| Method | Setting | mIoU (%) | Road | Sidewalk | Building | Wall | Fence | Pole | Traffic light | Traffic sign | Vegetation | Terrain | Sky | Person | Rider | Car | Truck | Bus | Train | Motorcycle | Bicycle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Baseline | | 71.05 | 97.9 | 82.4 | 90.0 | 52.5 | 51.2 | 54.2 | 58.7 | 70.2 | 90.3 | 65.1 | 92.1 | 77.5 | 48.9 | 92.2 | 63.2 | 73.3 | 65.2 | 55.1 | 70.1 |
| PSCA | SCA | 74.89 | 98.3 | 83.5 | 91.4 | 48.4 | 53.5 | 55.2 | 66..2 | 74.4 | 92.6 | 70.9 | 94.5 | 82.3 | 57.9 | 94.3 | 68.3 | 77.6 | 71.3 | 61.1 | 71.0 |
| | FP | 73.23 | 98.1 | 83.7 | 92.3 | 50.1 | 53.2 | 57.2 | 59.2 | 69.1 | 92.4 | 70.2 | 94.2 | 77.2 | 54.1 | 93.7 | 67.9 | 79.2 | 67.2 | 59.1 | 70.7 |
| | FP&SCA | 76.65 | 98.5 | 85.4 | 92.7 | 49.1 | 52.9 | 59.1 | 71.2 | 76.3 | 93.1 | 71.2 | 95.3 | 84.3 | 67.2 | 95.4 | 67.4 | 80.9 | 68.4 | 63.1 | 72.2 |
| ACF | NOR | 77.01 | 98.6 | 86.1 | 92.8 | 57.2 | 54.1 | 63.2 | 70.8 | 76.8 | 93.4 | 72.2 | 95.4 | 84.9 | 67.8 | 95.7 | 68.6 | 77.5 | 69.4 | 65.3 | 74.5 |
| | AVE | 78.23 | 98.6 | 86.6 | 93.0 | 58.9 | 56.9 | 67.2 | 74.3 | 78.9 | 93.6 | 73.1 | 95.3 | 86.2 | 69.7 | 96.1 | 65.8 | 82.1 | 76.2 | 66.7 | 75.7 |
| | MAX | 78.72 | 98.5 | 86.4 | 92.8 | 57.3 | 59.7 | 63.8 | 73.2 | 77.2 | 93.2 | 72.7 | 95.5 | 85.3 | 70.1 | 95.7 | 75.4 | 84.1 | 75.1 | 67.2 | 74.2 |
| | AVE + MAX | 79.91 | 98.5 | 86.6 | 93.1 | 59.1 | 61.1 | 67.2 | 76.5 | 77.3 | 93.3 | 72.3 | 95.2 | 86.0 | 70.5 | 96.5 | 74.1 | 90.6 | 77.1 | 67.4 | 75.7 |
| | (AVE + MAX) &MLP | 81.16 | 98.6 | 86.8 | 93.4 | 61.9 | 60.7 | 68.5 | 75.8 | 78.9 | 93.7 | 73.5 | 95.6 | 86.9 | 71.6 | 97.9 | 78.9 | 92.1 | 77.2 | 68.7 | 75.5 |
| Network (PSCA&ACF) | | 81.31 | 98.7 | 87.5 | 93.5 | 61.6 | 62.6 | 66.7 | 77.5 | 78.6 | 93.7 | 73.7 | 95.4 | 86.2 | 72.5 | 97.8 | 81.2 | 92.8 | 78.6 | 70.8 | 75.3 |

the baseline on the validation set, as shown in Table 1. Then we extended baseline to U-shape structure mentioned in Section 1 with our proposed PSCA and ACF modules, and conducted a set of ablation experiments for the two modules. Finally, we experimented the complete network.

### 4.2.1 | Ablation studies for PSCA module

Based on the observation in Section 3.2, we performed different settings for PSCA module to verify the validity of each component. As shown the second row block in Table 1, using SCA or FP alone yielded a result of 74.89% and 73.23% in mIoU, respectively. When we integrated SCA and FP together by embedding SCA into each level of the FP, the performance further was improved to 76.65%, which was an obvious improvement comparing with the baseline. Noted that, FP&SCA significantly outperformed the baseline in recognizing the objects with diverse scales, for example the ‘rider’ and ‘person’.

### 4.2.2 | Ablation studies for ACF module

As shown the third row block in Table 1, we set five experiments (i.e. NOR, AVE, MAX, AVE + MAX and (AVE + MAX)&MLP) to explore the best design for ACF module. Noted that the five experiments denoted different designs used to generate the channel attention vector.

1. NOR: the simple fusion, which obtained 77.01% mIoU.
2. AVE: using global average pooling to generate channel attention vector, which obtained 78.23% mIoU.
3. MAX: using global max pooling to generate channel attention vector, which obtained 78.72% mIoU.
4. AVE + MAX: combing global max and average pooling to generate channel attention vector, which obtained 79.91% mIoU.
5. (AVE + MAX)&MLP: combing global max and average pooling to generate channel attention vector firstly and using MLP to generate the final attention vector, which obtained 81.16% mIoU.

Comparing with NOR, four other designs all obtained performance improvement. Because the global average and max pooling can utilize global information to guide the channel fusion. We experimentally verified that utilizing MLP helps integrate the attention vector from AVE and MAX, which can bring superior performance improvement for the ACF module. So we adopted (AVE + MAX)&MLP in our final module.

### 4.2.3 | Complete network

Combined our best setting on PSCA and ACF modules, we experimented the complete network. In evaluation, we applied the multiscale inputs (with scales {0.5, 0.75, 1.0, 1.5, 1.75, 2.0}) and also left-right flipped the images. Finally, our network

**TABLE 2** Category-wise comparison with state-of-the-art methods on the Cityscapes test set

| Method | mIoU (%) | Road | Sidewalk | Building | Wall | Fence | Pole | Traffic light | Traffic sign | Vegetation | Terrain | Sky | Person | Rider | Car | Truck | Bus | Train | Motorcycle | Bicycle |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FRRN [2] (Pohlen et al. 2017) | 71.8 | 98.2 | 83.3 | 91.6 | 45.8 | 51.1 | 62.2 | 69.4 | 72.4 | 92.6 | 70 | 94.9 | 81.6 | 62.7 | 94.6 | 49.1 | 67.1 | 55.3 | 53.5 | 69.5 |
| RefineNet [7] (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 | 70.0 |
| PEARL [47] (Jin *et al.* 2017) | 75.4 | 98.4 | 84.5 | 92.1 | 54.1 | 56.6 | 60.4 | 69 | 74 | 92.9 | 70.9 | 95.2 | 83.5 | 65.7 | 95 | 61.8 | 72.2 | 69.6 | 64.8 | 72.8 |
| PSPNet [12] (Zhao et al. 2017) | 78.4 | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - | - |
| DepthSeg [49] (Kong et al. 2018) | 78.2 | 98.5 | 85.4 | 92.5 | 54.4 | 60.9 | 60.1 | 72.3 | 76.8 | 93.1 | 71.6 | 94.9 | 85.2 | 69.0 | 95.7 | 70.1 | 86.5 | 75.5 | 68.3 | 75.5 |
| DUC [48] (Wang et al. 2018) | 77.6 | 98.5 | 85.5 | 92.8 | 58.6 | 55.5 | 65.0 | 73.5 | 77.9 | 93.3 | 72.0 | 95.2 | 84.8 | 68.5 | 95.4 | 70.9 | 78.8 | 68.7 | 65.9 | 73.8 |
| DenseASPP [14] (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 |
| DFN[19] (Yu et al. 2018) | 80.3 | 98.6 | 85.9 | 93.2 | 59.6 | 61.0 | 66.6 | 73.2 | 78.2 | 93.5 | 71.6 | 95.5 | 86.5 | 70.5 | 96.1 | 77.1 | 90.0 | 84.7 | 68.2 | 76.5 |
| Ours | 81.4 | 98.5 | 87.2 | 94.3 | 61.8 | 63.2 | 69.2 | 76.5 | 78.3 | 93.7 | 72.1 | 95.3 | 87.0 | 71.9 | 97.5 | 81.2 | 92.7 | 79.3 | 70.1 | 76.5 |

achieved performance of 81.3% in mIoU without MS-COCO [46]. More performance details were listed in the fourth row block of Table 1.

## 4.3 | Comparison with state-of-the-art methods

In this section, we evaluate our complete network on Cityscapes and PASCALVOC 2012 datasets and compare our results with several state-of-the-art methods.

### 4.3.1 | Results on Cityscapes dataset

We trained our network with only fine annotated data, and submitted the test results to the official evaluation server, which achieved mIoU of 81.4% on the test set. We compared our network with existing methods on the Cityscapes test set and more performance details were shown in Table 2 and Figure 6. Our network outperformed many existing methods with dominant advantage. In particular, our model outperformed the PSANet [14] by a large margin with the same backbone ResNet-101. In addition, we observed that it also surpasses DenseASPP [26], which used more powerful pretrained models than ours.

In order to show the superior performance of our network more intuitively, we visualized some examples of the Cityscapes dataset, as shown in Figure 7. Without extra measures to capture context information, the baseline (in column 3 of Figure 6) cannot correctly distinguish 'terrain' and 'fence', and misclassified the category of 'bus' and 'truck'. For PSCA, due to the SCA and multiscale transformation, it can provide accurate feature representation for the confusion categories and

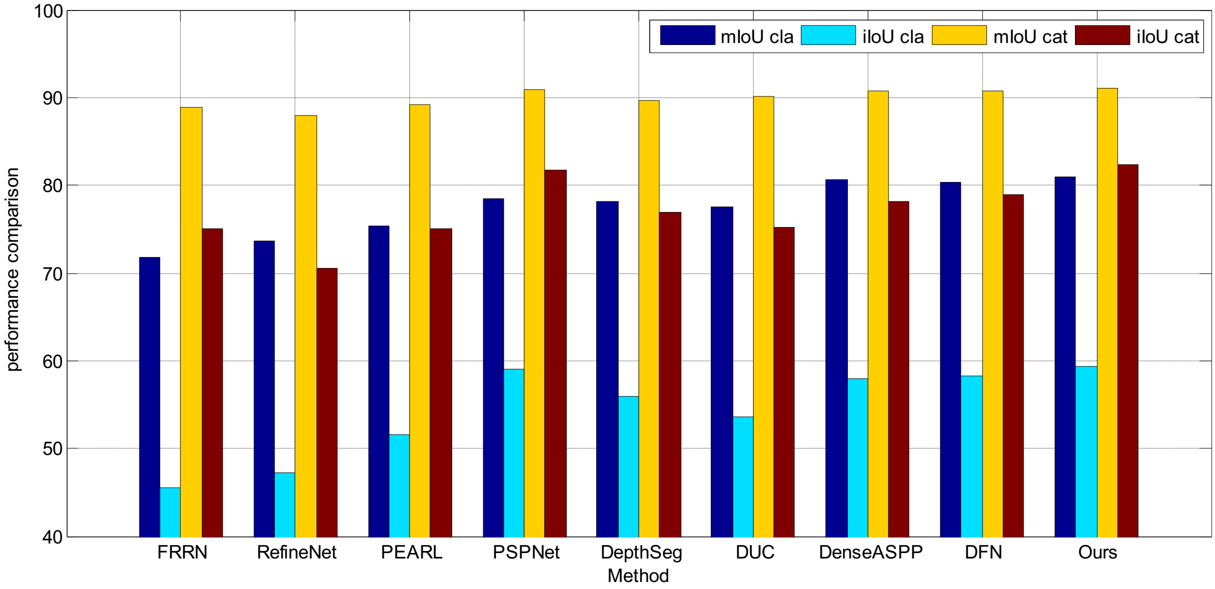


**FIGURE 6** Performance comparison on Cityscapes test set

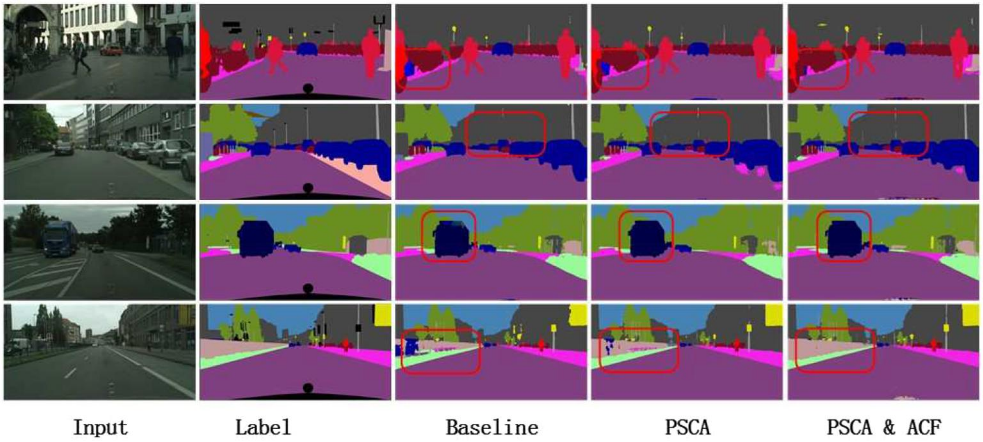


**FIGURE 7** Visualization results of attention modules on Cityscapes val set

TABLE 3 Per-class results on PASCAL VOC 2012 testing set

| Method | mIoU(%) |
|---|---|
| FCN 8s [3] | 62.2 |
| DeepLabv2 [44] | 71.6 |
| Piecewise [10] | 75.3 |
| ResNet38 [8] | 82.5 |
| EncNet [37] | 82.6 |
| DFN [19] | 82.7 |
| DUC [48] | 83.1 |
| **Ours** | **84.6** |

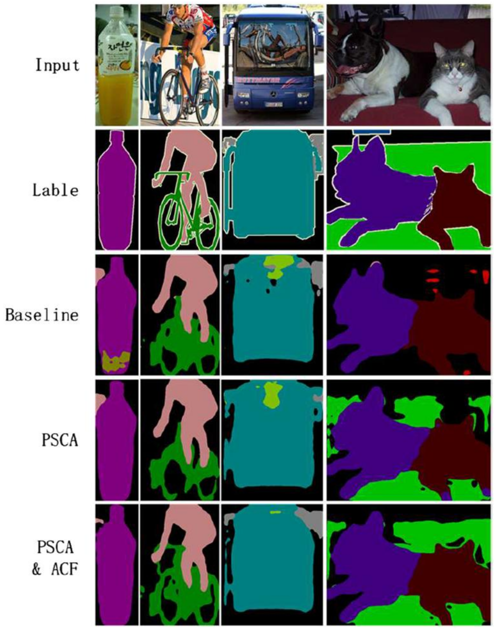


FIGURE 8 Visualization results of attention modules on PASCAL VOC 2012 val set

diverse scales objects. For instance, in the column 4 of Figure 6, the PSCA module can correctly distinguish the big 'bus' and 'truck', comparing with the baseline. Meanwhile, we observed that some details and object boundaries were clearer with the ACF module, such as the 'pole' and the 'sidewalk' in the column 5. Furthermore, combined PSCA and ACF modules, our complete network showed excellent performance for both inconspicuous and incomplete objects.

### 4.3.2 | Results on PASCAL VOC 2012 dataset

In order to further demonstrate the generalization of our method, we test our proposed method on PASCAL VOC 2012. Comparisons with previous state-of-the-art methods are reported in Table 3. Results show that our model achieves 84.6% in mIoU, which outperforms these methods by a large margin. Different from these previous methods, we introduce the attention modules to capture global dependencies explicitly, and the proposed method can achieve better performance. We also offer some examples of visualization to illustrate the effects of our method, as shown in Figure 8.

## 5 | CONCLUSION

We have proposed to enhance feature fusion between high-level features and low-level features to capture context information from both semantic and spatial levels. Spatial aggregation and channel fusion were introduced to emphasize the spatial correlation and channel interdependence between low-level features and high-level features, respectively for bridging the semantic and spatial gap, which was effective for the feature fusion. For the implementation, we designed two attention-based modules, that is PSCA module and ACF module. Theoretical analysis, visualization and quantitative experimental results on Cityscapes and PASCAL VOC 2012 datasets were presented to demonstrate the effectiveness of our designs.

## ACKNOWLEDGEMENTS

This work is supported by the project no. IMICZ2017004 and the 111 Project BP0719039, and the computational resources used in this study are provided by the Automotive Electronics and Control Research Centre, Hefei University of Technology.

## ORCID

*Jun Zhou* https://orcid.org/0000-0001-7569-5589

## REFERENCES

1. Xu, H., et al.: End-to-end learning of driving models from large-scale video datasets. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 2174–2182. CVPR, Honolulu (2017)
2. Pohlen, T., et al.: Full-resolution residual networks for semantic segmentation in street scenes. In: Proceedings IEEE Computer Society conference on computer vision and pattern recognition, pp. 4151–4160. CVPR, Honolulu (2017)
3. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 3431–3440. CVPR, Boston (2015)
4. Liu, S., et al.: Learning affinity via spatial propagation networks. In: Proceedings of advances in neural information processing systems, pp. 1520–1530. NIPS, Long Beach (2017)
5. Chen, L. C., et al.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of European conference On computer vision, pp. 833–821. ECCV, Munich (2018)
6. Sun, K., et al.: High-resolution representations for labelling pixels and Regions. (2019). arXivpreprint arXiv:1904.04514. https://arxiv.org/abs/1904.04514
7. Lin, G., et al.: Refinenet: Multi-path refinement networks for high-resolution semantic segmentation. In: Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition, pp. 5168–5177. (CVPR), Honolulu (2017)
8. Wu, Z, Shen, C, van den Hengel, A: Wider or Deeper: Revisiting the ResNet Model for Visual Recognition (2016). arXivpreprint arXiv. 1611.10080. https://arxiv.org/abs/1611.10080
9. Li, X., Hu, X., Yang, J.: Spatial group-wise enhance: improving semantic feature learning in convolutional networks (2019). arXivpreprint arXiv. 1905.09646. https://arxiv.org/abs/1905.09646

10. Lin, G, et al.: Efficient piecewise training of deep structured models for semantic segmentation. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 3194–3203. CVPR, Las Vegas (2016)
11. Chen, L.C., et al.: Searching for efficient multi-scale architectures for dense image prediction. In: Proceedings of advances in neural information processing systems, pp. 8699–8710. NIPS, Montreal (2018)
12. Zhao, H., et al.: Pyramid scene parsing network. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 2881–2890. CVPR, Honolulu (2017)
13. Chaurasia, A., Culurciello, E.: Linknet: Exploiting encoder representations for efficient semantic segmentation. In: 2017 IEEE visual communications and image processing, pp. 1–4. (VCIP) (2017)
14. Yang, M., et al.: Denseaspp for semantic segmentation in street scenes. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, pp. 3684–3692. CVPR, Salt Lake City (2018)
15. Li, H., et al.: Pyramid attention network for semantic segmentation. (2018). arXivpreprint arXiv:1805.10180. https://arxiv.org/abs/1805.10180
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: Proceedings of international conference on medical image computing and computer-assisted intervention, pp. 234–241. MICCAI, Munich (2015). https://doi.org/10.1007/978-3-319-24574-4_28
17. Peng, C., et al.: Large kernel matters-improve semantic segmentation by global convolutional network. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR), pp. 4353–4361. CVPR, Honolulu (2017)
18. Zhang, Z., et al.: ExFuse: Enhancing feature fusion for semantic segmentation. In: Proceedings of the European conference on computer vision, pp. 269–284. CVPR, Munich (2018)
19. Yu, C., et al.: Learning a discriminative feature network for semantic segmentation. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, pp. 1857–1866. CVPR, Salt Lake City (2018)
20. Chen, L., et al.: Sca-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, pp. 5659–5667. CVPR, Honolulu (2017)
21. Chen, L.C., et al.: Attention to scale: Scale-aware semantic image segmentation. In: Proceedings CVPR IEEE computer society conference on computer vision and pattern recognition (CVPR), pp. 3640–3649. Las Vegas NV, USA (2016)
22. Cheng, J., Dong, L., Lapata, M.: Long short-term memory-networks for machine reading (2016). arXivpreprint arXiv:1601.06733
23. Vaswani, A., et al.: Attention is all you need. In: Proceedings of the advances in neural information processing systems, pp. 5998–6008. NIPS, Long Beach (2017)
24. Zhang, Y., et al.: Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision, pp. 286–301. ECCV, Munich (2018)
25. Cordts, M., et al.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, pp. 3213–3223. CVPR, Seattle (2016)
26. Zhao, H., et al.: PSANet: Point-wise spatial attention network for scene parsing. In: Proceedings of the European conference on computer vision, pp. 267–283. ECCV, Munich (2018)
27. Liu, S., et al.: Learning affinity via spatial propagation networks. In: Proceedings of the advances in neural information processing systems, pp. 1520–1530. NIPS, Long Beach (2017)
28. Visin, F., et al.: Reseg: A recurrent neural network-based model for semantic segmentation. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition, pp. 41–48. CVPR, Las Vegas (2016)
29. Badrinarayanan, V., Kendall, A., Cipolla, R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495. (2017)
30. Liu, W., Rabinovich, A., Berg, A.C.: Parsenet: Looking wider to see better. (2015). arXivpreprint arXiv:1506.04579. https://arxiv.org/abs/1506.04579
31. Paszke, A., et al.: Enet: A deep neural network architecture for real-time semantic segmentation. (2016). arXivpreprint arXiv:1606.02147. https://arxiv.org/abs/1606.02147
32. Romera, E., et al.: Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. 19(1), 263–272 (2017)
33. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. Proceedings CVPR IEEE Computer Society conference on computer vision and pattern recognition (CVPR), 7132–7141 (2018)
34. Woo, S., et al.: CBAM: convolutional block attention module. In: Proceedings of the European conference on computer vision, pp. 3–19. ECCV, Munich Munich (2018)
35. Wang, X., et al.: Non-local neural networks. In: Proceedings IEEE Computer Society conference on computer vision and pattern recognition, pp. 7794–7803. CVPR, Salt Lake City (2018)
36. Cao, Y., et al.: GCNet: Non-local networks meet squeeze-excitation networks and beyond. (2019). arXivpreprint arXiv:1904.11492. https://arxiv.org/abs/1904.11492
37. Zhang, H., et al.: Context encoding for semantic segmentation. In: Proceedings IEEE computer society conference on computer vision and pattern recognition, pp. 7151–7160. CVPR, Salt Lake City (2018)
38. Yuan, Y., Wang, J.: OCNet: Object context network for scene parsing. (2018). arXivpreprint arXiv:1809.00916. https://arxiv.org/abs/1809.00916
39. Huang, Z., et al.: CCNet: Criss-cross attention for semantic segmentation. (2018). arXivpreprint arXiv:1811.11721. https://arxiv.org/abs/1811.11721
40. Fu, J., et al.: Dual attention network for scene segmentation. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 3146–3154. CVPR, Long Beach (2019)
41. Lin, T. Y., et al.: Feature pyramid networks for object detection. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 2117–2125. CVPR, Honolulu (2017)
42. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition (CVPR), pp. 770–778. CVPR, Las Vegas (2016)
43. Paszke, A., et al.: Automatic differentiation in pytorch. In: Proceedings of the Autodiff workshop. The future of gradient-based machine learning software and techniques. NIPSW, Long Beach (2017)
44. Chen, L. C., et al.: Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFS. IEEE Trans. Pattern Anal. Mach. Intell. 40(4), 834–848 (2017)
45. Bottou, L.: Large-scale machine learning with stochastic gradient descent. In: Proceedings of International Conference on Computational Statistics. vol. 2010, pp. 177–186. COMPSTAT, Paris (2010)
46. Lin, T.Y., et al.: Microsoft coco: Common objects in context. In: Proceedings of the European conference on computer vision, pp. 740–755. ECCV, Zurich (2014)
47. Jin, X., et al.: Video scene parsing with predictive feature learning. In: Proceedings of IEEE computer society conference on computer vision and pattern recognition (CVPR), pp. 5580–5588 CVPR, Venice (2017)
48. Wang, P., et al.: Understanding convolution for semantic segmentation. In: Proceedings of IEEE Winter conference on applications of computer vision, pp. 1451–1460. WACV, NV (2018)
49. Kong, S., Fowlkes, C.C.: Recurrent scene parsing with perspective understanding in the loop. In: Proceedings of IEEE Computer Society conference on computer vision and pattern recognition, pp. 956–965. CVPR, Salt Lake City (2018)
50. Everingham, M., et al.: The Pascal visual object classes VOC challenge. Int. J. Comput. Vision. 88(2), 303–33 (2010)

**How to cite this article:** Hu, J., et al.: Enhancing feature fusion with spatial aggregation and channel fusion for semantic segmentation. IET Comput. Vis. 15 (6), 418–427 (2021). https://doi.org/10.1049/cvi2.12026