Using Features Specifically: An Efficient Network for Scene Segmentation Based on Dedicated Attention Mechanisms

ZHIQIANG XIONG1, ZHICHENG WANG1, JIE LI2, ZHAOHUI YU1, AND XI GU1

1CAD Research Center, College of Electronic and Information Engineering, Tongji University, Shanghai 201800, China
2Shanghai Aerospace Electronic Technology Institute, Shanghai 201109, China

Corresponding author: Zhicheng Wang (zhichengwang@tongji.edu.cn)

ABSTRACT  Semantic segmentation is a challenging task in computer vision, which requires both context information and rich spatial detail. To this end, most methods introduce low-level features for spatial detail. However, low-level features lack global information. Too much low-level features will disturb the segmentation result. In this paper, we extract low-level features guided by abstract semantic features to improve segmentation results. Specifically, we propose a Pixel-wise Attention Module (PAM) to select low-level features adaptively and a Dual Channel-wise Attention Fusion Module (DCAFM) to fuse the context information further. These two modules use the attention mechanism from a more macro perspective, which is not limited to the inter-layer feature adjustments. There are not complicated and redundant processing modules in our architecture. By using features efficiently, the complexity of the network was significantly reduced. We evaluate our approach on Cityscapes, PASCAL VOC 2012, and PASCAL Context datasets, and we achieve 82.3% Mean IoU on PASCAL VOC 2012 test dataset without pre-training on the MS-COCO dataset.

INDEX TERMS  Image processing, deep neural network, semantic segmentation, attention mechanism.

I. INTRODUCTION
Semantic segmentation is an image-classification at the pixel level and segmenting scene into different areas with semantic classes. A few examples are shown in Figure 1. It can be widely applied to the fields of automatic driving, scene understanding, etc.

The semantic segmentation accuracy is affected by the semantic concept for objects and the coordinate alignment between classification labels and pixels, reflected in the consistency within categories and the edge detail of some objects. Many methods based on Fully Convolutional Networks (FCNs) [1] are proposed to address the above problems. On the one hand, to obtain abundant semantic information, context information at different encoder stages is usually fused during decoder [2], [3]. Some works [4]–[6] aggregate multi-scale context information generated by different dilated convolutions or different scale pooling operations, and some works [8] enlarge the kernel size to grab context information. Nevertheless, the context information is not equally important to segmentation, and it should be enhanced selectively with the guidance of global information.

On the other hand, a sizeable spatial detail is lost because of successive convolutions and pooling operations. [2],[3] make up for the spatial details by integrating high-level and mid-level features. Some methods capture spatial details by
increasing the receptive field for high-level features [4]–[6], but there is not much information of detail preserved in high-level features.

High-level features mainly guide most of the segmentation areas, while local areas such as boundaries need rich low-level features to obtain spatial detail. Due to the lack of convolution and global information, the low-level feature cannot accurately predict the category, especially in the intra-object. Excess low-level features will cause incorrect disturbances to high-level features. To address the above problems, we propose a method that uses attention mechanisms in pixel-wise of low-level features and channel-wise of high-level features, respectively. Both attention mechanisms are under the guidance of the information extracted by the encoder.

For channel-wise attention, we designed a Dual Channel-wise Attention Fusion Module (DCAFM) to readjust the proportion of each feature map in the high-level features. Each channel can be regarded as a kind of feature, and different features have different effects on the results. Some channels represent common features of different categories, which have little or even adverse effects on classification, while others are unique to different categories. For the pixel-wise attention module, we use the features fused by DCAFM to filter the low-level features, which can select features to supplement spatial detail for the segmentation result. As a result of successive convolutions and pooling operations, high-level features lack spatial detail and perform poorly in edges of objects.

On the contrary, low-level features retain much spatial detail, making up for the defects of high-level features. However, the context information of low-level features is insufficient to guarantee the intra-consistency of segmentation objects. Therefore, useful spatial detail in low-level features is selected, and unhelpful features are discarded.

A. CONTRIBUTIONS

Our main contributions can be summarized as follows: (a) We explicitly differentiate the encoder’s features and propose a framework with two attention mechanisms from a more macro perspective to process the encoded high-level features and low-level features, respectively, to generate segmentation prediction. (b) A Pixel-wise Attention Module is proposed to select the low-level features with rich spatial detail, and a Dual Channel-wise Attention Fusion Module with dual channel-wise attention to weight different feature maps of high-levels. (c) We proved that our method could soften the fusion of features and avoid the erroneous influence of some low-level features.

II. RELATED WORK

Recently, many methods based on FCNs [1] have made significant progress on different benchmarks of the semantic...
segmentation task. Most of them are designed to fuse the features of adjacent encoder layers for sufficient information.

**A. SPATIAL DETAIL**

FCN-based models obtain high-level semantic information by a convolutional neural network (CNN) [10] with convolution and down-sampling pooling. However, high-level semantic information is not enough for pixel-level semantic segmentation tasks, and the spatial detail is essential for the details of segmentation. Global Convolutional Network (GCN) [8] adopts ‘large kernel’ to increase the receptive field. PSPNet [4] uses multi-scales pooling to preserve the spatial detail of the feature maps, while DUC [11], DeepLab-v2 [5], and DeepLab-v3 [9] use multi-scales dilated convolution.

**B. CONTEXT INFORMATION**

Context information is crucial for distinguishing seemingly similar classes. [4],[9],[12] use global average pooling to supplement global context information. [4],[9],[11],[15] capture and merge different levels of context information by adding the features of the different receptive fields.

**C. ENCODER-DECODER**

The FCNs-based models’ encoder extracted different levels of features, but too much spatial detail is corrupted by the convolution and pooling operations. Some U-shape structure-based methods integrate these features to recover spatial detail and refine the prediction with different decoders. For example, U-net [2] uses the skip connection, while RifineNet [3] utilizes a Multi-Path Refinement structure to optimize prediction results. [41] uses $1 \times 1$ convolution to fuse the class scores output from the feature extractor adaptively, and [42] uses Fusing Attention Interim (FAI) to aggregate the adjacent level’s features. SegNet [13] adds pooling indices in the decoder to retain the detail, and LRR [14] employs the Laplacian Pyramid Reconstruction network. Our method uses two different attention modules to selectively fuse and filter the stage-level features of the ResNet, respectively, and gives full play to each stage-level features’ characteristics.

**D. ATTENTION MECHANISM**

The powerful deep neural network can encode lots of information, and the attention mechanism can act as a leap-forward guide to select the information [16], [17]–[21]. In SENet [16], features were used to learn attention to revise themselves. DFN [17] learns the global context to filter features. [18] uses the attention mechanism on the size of the input images. CBAM [35] uses the attention mechanisms of channel-wise and spatial-wise to refine the feature fusion process of almost every block of the ResNet adaptively. As a comparison, our method utilizes features of different abstraction levels from a perspective of the macro network framework and applies them to adjust the stage-level features of the ResNet. DANet [32] calculates channel similarity and spatial position similarity of high-level features from channel-wise and pixel-wise, respectively, to increase the correlation and differentiation between features. However, after successive convolutions and pooling operations, a large amount of spatial detail has been lost in the high-level features, and calculating the similarity between features requires many calculations, which increases the complexity of the model.

**III. APPROACH**

In this section, we first introduce our method in detail. Then, we elaborate on the design details of the two attention modules. Finally, some other details of our method are supplemented.

For an image to be segmented, the encoder (such as ResNet [22] and VGG [23]) composed of a series of convolution and pooling operations is usually used to capture the information in the image. However, the information in different stages plays different roles in segmentation. To take full advantage of information’s proprietary nature, we processed the information of different stages with different strategies. As illustrated in Figure 2, we employ a pre-trained residual network as the backbone. We firstly divide the features of ResNet into two groups: the feature maps of the layer-1 (low-level) as spatial information and the feature maps of the layer-3 and layer-4 (high-level) as context information. We propose a Dual Channel-wise Attention Fusion Module (DCAF) to fuse the context information. During the fusion of context information, the feature maps of the layer-4 are enlarged to keep consistent with the size of the feature maps of the layer-3. Then, the fused information and spatial information will be fed into the Pixel-wise Attention Module (PAM). In PAM, a pixel-wise attention map was generated adaptively with context information. PAM and DCAF output spatial information weighted by pixel-wise attention map and context information weighted by channel-wise attention vectors, respectively. Finally, we aggregate the two sets of information to obtain the final prediction.

**A. DUAL CHANNEL-WISE ATTENTION FUSION MODULE**

During the fusion of adjacent stages’ features, the difference between the feature of different stages should be considered, as well as the powerful semantic information of higher stage. Unlike SENet and CBAM, our Dual Attention Fusion Module (DAFM) uses high-level feature maps to change the weight of low-level feature maps and the weight of high-level feature maps from a macro perspective. Specifically, we use the high-level features to learn two weight vectors, which are used to adjust the high-level features and the low-level features at the channel-wise, respectively. In detail, we enter the high-level features into two sets of identical structures, each using global average pooling to capture global information, and the weights are constrained between 0 and 1 by the sigmoid function. Simultaneously, we use $1 \times 1$ convolution to convert high-level features and low-level features to the features with the same number of channels, which is the same as the number of weight vectors’ channels. These two weight
FIGURE 3. A schematic diagram of Dual Channel-wise Attention Fusion Module.

FIGURE 4. A schematic diagram of Pixel-wise Attention Module.

vectors are multiplied separately with the high-level features and low-level features, and then the weighted high-level and low-level features are added.

As shown in Figure 3, given two feature maps, Low and High, where Low, High ∈ R^{C×H×W}. For the Low = X^L, we assume that there are two positions, AI(ia, ja) and BI(ib, jb), which differ by ε^L on channel C_1 and ε^L on channel C_2. (iA, jA), (IB, JB) ∈ D, D is the set of pixel positions.

\[ ε^L_1 = |X^L_{ia,ja,1} - X^L_{ib,jb,1}| \quad (1) \]
\[ ε^L_2 = |X^L_{ia,ja,2} - X^L_{ib,jb,2}| \quad (2) \]

Abstractly, C_1 and C_2 represent two different features, respectively. We assume that A and B need to maintain discrimination on C_1 and remain consistent on C_2, which means that ε^L should be maximized and ε^L should be minimized. To address the problem above, two parameters, α^L ∈ R^{1×1} and α^H ∈ R^{1×1}, are introduced to adjust high-level and low-level features, where \( α = \text{Sigmoid}(X; \omega) \).

\[ ε^L = |α^L X^L_{ia,ja,1} - α^L X^L_{ia,ja,1}| = α^L ε^L_1 \quad (3) \]
\[ ε^L = |α^L X^L_{ia,ja,2} - α^L X^L_{ia,ja,2}| = α^L ε^L_2 \quad (4) \]

where the \( ε^L_1 \) and \( ε^L_2 \) are the weighted differences. The goal above can be achieved by increasing \( α^L_1 \) and decreasing \( α^L_2 \).

For the feature maps obtained by adding \( X^L \) and \( X^H \), \( ε_1 \) and \( ε_2 \) represent the differences between A and B on channel C_1 and C_2, respectively.

\[ ε_1 = ε^L_1 + ε^H \quad (5) \]
\[ ε_2 = ε^L_2 + ε^H \quad (6) \]
\[ ε_1 = α^L_1 ε^L_1 + α^H_1 ε^H \quad (7) \]
\[ ε_2 = α^L_2 ε^L_2 + α^H_2 ε^H \quad (8) \]

However, the features of adjacent stages behave similarly at some positions. When \( ε^L_1 \) equal to \( ε^H_1 \), if only one of \( α^L_1 \) and \( α^H \) is applied, the range of adjustment is roughly halved. Therefore, use \( α^L \) and \( α^H \) at the same time can help adjust the effect of different features at different pixel positions.

B. PIXEL-WISE ATTENTION MODULE

With a large number of convolution and pooling operations, the high-level features obtain enough context information to make the overall judgment of the object more accurate, but the resolution reduction loses a lot of spatial information, resulting in unclear boundaries of the segmentation. In contrast, low-level features preserved high resolution and rich spatial details. However, if low-level features are combined with high-level features directly, the low-level features will disturb the segmentation of high-level features because of the lack of context information. We design a Pixel-wise Attention Module (PAM) to select useful low-level features. As shown in Figure 4, we concatenate low-level features (spatial information) and the features fused by DCAM at first, then the concatenated feature is fed into the network of convolution and activation functions to obtain a confidence map. The low-level features and the confidence map are multiplied to get the filtered spatial information.

\[ X^{SI} = \beta X^{SI} = \begin{bmatrix} \beta_{1,1} X^{SI}_{1,1} & \cdots & \beta_{1,1} X^{SI}_{1,W} \\ \vdots & \ddots & \vdots \\ \beta_{H,1} X^{SI}_{H,1} & \cdots & \beta_{H,1} X^{SI}_{H,W} \end{bmatrix} \quad (9) \]

where \( X^{SI} \in R^{C×H×W} \) is the spatial information (low-level features), \( \beta \in R^{1×H×W} \) is the confidence map. We assume that \( X^{SI}_{ip,jp} \) and \( X^{CI}_{ip,jp} \) are the spatial features and context features at the position \( P(ip, jp) \), respectively. We introduce \( \beta_{ip,jp} \) to adjust the fusion of \( X^{SI}_{ip,jp} \) and \( X^{CI}_{ip,jp} \).

\[ X^F = \beta_{ip,jp} X^{SI}_{ip,jp} + X^{CI}_{ip,jp} = \begin{bmatrix} \beta_{ip,jp} X^{SI}_{ip,jp,1} + X^{CI}_{ip,jp,1} \\ \vdots \\ \beta_{ip,jp} X^{SI}_{ip,jp,C} + X^{CI}_{ip,jp,C} \end{bmatrix} \quad (10) \]

where \( \beta_{ip,jp} \) is the confidence at position \( P \). If the effect of \( X^{SI}_{ip,jp} \) on \( X^{CI}_{ip,jp} \) is positive, then \( \beta_{ip,jp} \) increase. Otherwise, to avoid the wrong distribution of \( X^{SI}_{ip,jp} \) affecting the correct distribution of \( X^{CI}_{ip,jp} \), it is necessary to reduce \( \beta_{ip,jp} \). By generating the low-level features’ confidence of each pixel position, the low-level features can be selectively used to take full advantage of the features while reducing the effects of error information.
C. OTHER DETAILS

To apply DCAFM and PAM to the encoder, we adopt a $1 \times 1$ convolution to convert the feature maps to 128 channels. In the final Sum Fusion, we first add the output of DCAFM and PAM, and then restore the feature maps to the size of input with a simple refine block to get the prediction. Moreover, like the deep-supervision [26], we add two auxiliary losses to supervise the output from spatial information and context information. All the loss functions are Cross-Entropy Loss.

$$L(y; w) = \text{CrossEntropyLoss}(y; w)$$

$$loss = L_p(y_p; w) + \lambda_c L_c(y_c; w) + \lambda_s L_s(y_s; w)$$

where the $y_p$, $y_c$, and $y_s$ are the output of spatial information and context information, respectively. The $L_p$ and $L_c$ are the auxiliary loss. The $y_p$ is the prediction of the network, and the $L_p$ is the principal loss function, the $loss$ is the joint loss function. Furthermore, we use the parameters $\lambda_s$, $\lambda_c$ to balance the principal loss and auxiliary loss. The $\lambda_s$, $\lambda_c$ in this paper is equal to 0.1 and 0.4, respectively.

IV. EXPERIMENTS

We evaluate the proposed method on three public datasets: Cityscapes[24], PASCAL VOC 2012[25], and PASCAL Context[36]. We first introduce the datasets and the implementation details, then we investigate the effects of each module of our method. Finally, we report the results of our method on these public datasets.

A. DATASETS

1) CITYSCAPES

The Cityscapes is a large dataset of 5,000 fine annotated images for urban scenes segmentation. The dataset contains 30 classes, 19 classes of which are used for training and evaluation. There are 2,979 images for training, 500 images for validation, and 1,525 images for testing. Each image has $2,048 \times 1,024$ pixels.

2) PASCAL VOC 2012

The PASCAL VOC 2012 is one of the most commonly used semantic segmentation datasets, containing 20 object classes and one background. In the dataset, 1,464 images for training, 1,449 images for validation, and 1,456 images for testing. We augment the original dataset with the Semantic Boundaries Dataset [27], resulting in 10,582 images for training.

3) PASCAL CONTEXT

The PASCAL Context provides detailed semantic labels for scenes, and it contains 4,998 images for training and 5,105 images for testing. Following [3],[33], we evaluate our method on the most frequent 59 classes along with one background category.

B. IMPLEMENTATION DETAILS

In experiments, we apply the ResNet series pre-trained on the ImageNet dataset [28] as the backbone, including ResNet-18, ResNet-50, ResNet-101, and we implement our method based on Pytorch.

1) DATA AUGMENTATION

We adopt random horizontal flip, mean subtraction, and random scale on the input images in the training process. The scales contain {0.75, 1.0, 1.25, 1.5, 1.75, 2.0}. Besides, we randomly crop the input image. The crop size is $1024 \times 1024$ for Cityscapes and $512 \times 512$ for PASCAL VOC 2012, PASCAL Context. Some examples of data augmentation are shown in Figure 5.

2) TRAINING DETAILS

We use mini-batch stochastic gradient descent (SGD) [10]. Inspired by [5],[9], we employ a ‘poly’ learning rate policy where the current learning rate equals the initial learning rate multiplying $(1 - \frac{iter}{\text{max\_iter}})^{\text{power}}$ with power 0.9. For Cityscapes, we use SGD with batch size 8, initial learning rate 0.01, momentum 0.9, and weight decay $5e^{-4}$. For PASCAL VOC 2012, we use SGD with batch size 16, initial learning rate 0.001, momentum 0.9, and weight decay $1e^{-5}$.

C. ABLATION STUDY

In this subsection, we will decompose the network to verify the effect of each module. We evaluate our method on the validation set of Cityscapes [24] and PASCAL VOC 2012 [24]. Based on ResNet, we add DCAFM and PAM to fuse the features of the first, third, and fourth stages. As a comparison, we replace our module with summation to build the baseline. To verify the effects of our modules, we conduct experiments on PASCAL VOC 2012 validation set with the settings in Table 1 and the Cityscapes validation set with the settings in Table 2.

PASCAL VOC 2012: As shown in Table 1, the modules significantly improved the performance. Compare to the baseline (ResNet-50), the use of the Dual Channel-wise Attention Fusion Module raises the Mean IoU from 68.2% to 73.7%, an improvement of 5.5%. On this basis, the addition of the Pixel-wise Attention Module further improves the Mean IoU to 74.6%. With a deeper backbone, the two modules improve the Mean IoU over baseline (ResNet-101) by 5.0%.
TABLE 1. Ablation study for modules on PASCAL VOC 2012 validation set.

| Backbone  | DCAFM | PAM | Mean IoU% |
|-----------|-------|-----|-----------|
| Resnet50  | ✓     | ✓   | 68.2      |
|           |       |     | 73.7      |
|           |       |     | 74.6      |
| Resnet101 | ✓     | ✓   | 72.9      |
|           |       |     | 76.8      |
|           |       |     | 77.9      |

DCAFM represents the Dual Channel-wise Attention Fusion Module. PAM represents the Pixel-wise Attention Module. In contrast, we sum up features as the baseline.

TABLE 2. Ablation study for modules on Cityscapes validation set.

| Backbone | DCAFM | PAM | Mean IoU% |
|----------|-------|-----|-----------|
| Resnet-18| ✓     | ✓   | 70.3      |
|          |       |     | 72.8      |
|          |       |     | 73.1      |

DCAFM represents the Dual Channel-wise Attention Fusion Module. PAM represents the Pixel-wise Attention Module. In contrast, we sum up features as the baseline.

Cityscapes: As shown in Table 2, the Dual Channel-wise Attention Fusion Module brings a 2.5% improvement over baseline (ResNet-18), and the employ of the Pixel-wise Attention Module further improves the performance to 73.1%.

As shown in Figure 6 and Figure 8, with the DCAFM, some misclassification within the objects was eliminated, such as the ‘car’ in the first row, the ‘sidewalk’ in the second row of Figure 8, and the ‘dog’ in the third row of Figure 6. Furthermore, the PAM filters out the wrong information in the low-level features while preserving the spatial detail that has a positive effect, making the segmentation more accurate and more holistic.

D. VISUALIZATION AND ANALYSIS OF ATTENTION

To illustrate the effect of the attention mechanism in the network explicitly, we visualize the results of the two attention modules.

For the Dual Channel-wise Attention Fusion Module, the features of high-level and low-level to be fused are 128 channels, and feature maps of each channel are in the size of $H \times W$. Therefore, the two vectors used to adjust the high-level and low-level features are in dimensions of 128. To analyze the meaning of the channel weights, we visualized several channel maps of the low-level features to an image with a size of $H \times W$. As shown in Figure 7, the 51$^{st}$ channel map (in the second column) has low weight, while the 11$^{th}$ channel map (in the third column) and the 126$^{th}$ channel map (in the fourth column) have high weights. The features in the 51$^{st}$ channel map are difficult to distinguish since the values of most pixel positions are similar. As a comparison, there are discriminative features in the 11$^{th}$ and 126$^{th}$ channel map. Some of the segmentation objects can be clearly distinguished, such as the car and the boundary of the tree in the 11$^{th}$ channel map; the trees and the road in the 126$^{th}$ channel map. That is to say, the 11$^{th}$ channel and the 126$^{th}$ channel represent some abstract features, which can distinguish the objects with these features from other objects, so the 11$^{th}$ channel and the 126$^{th}$ channel got a high weight, while the 51$^{st}$ channel got a low weight.

Furthermore, we visualized the high-level and low-level features’ weights of the first two examples in Figure 7. Figure 9 (a) is the weights of low-level and high-level features of the first example of Figure 7, while Figure 9 (b) is the weights of low-level features of the first two examples of Figure 7. From Figure 9, we can learn that: 1) the weights of channels are different in different images, which indicates that the convolution operation cannot meet the requirements of the feature weights’ adjustment, so it is useful to introduce the channel attention mechanism; 2) as shown in Figure 9 (b), the weights’ distribution of the low-level features of different images are roughly the same, which implies that each channel represents a specific abstract feature, and each abstract feature is of different importance; 3) as shown in Figure 9 (a), the weight of high-level features is generally higher than the weight of the low-level features, and the distribution of them is very different, the weight of high-level is dominant.

For the Pixel-wise Attention Module, we visualize the confidence map of spatial information. As shown in the fifth columns of Figure 7, the brighter the pixel position represents more useful and credible spatial details. It can be clearly seen that the highlights are almost at the boundaries of the segmentation objects. It shows that the spatial information lacking in the high-level features is preserved by confidence map, while the context information in low-level features is not enough, so the part lacking the semantic information is filtered.

E. RESULTS ON CITYSCAPES

In evaluation, following [17], [29], we adopt the multi-scale input with scales $\{0.5, 0.75, 1.0, 1.25, 1.5, 1.75, 2.0\}$ and left-right flip on the image, and we train our method with only fine data of Cityscapes. As shown in Table 3, the multi-scale input improved the Mean IoU by 1.5% to
FIGURE 7. The attention modules’ visualization results on Cityscapes validation set. For each row, we show an input image, three channel maps (51st, 126th, 11th) of the low-level features, and the confidence map of the Pixel-wise Attention Module. Finally, corresponding ground truths are provided.

FIGURE 8. Examples of ablation study on Cityscapes dataset. Column 1: input image. Column 2: results of baseline. Column 3: results of the baseline + DCAFM. Column 4: results of baseline + DCAFM + PAM. Column 5: ground truth. Compare the area in the red box.

| Method          | MS     | Flip | Mean IoU% |
|-----------------|--------|------|-----------|
| Ours (Res18)    |        |      | 73.1      |
| Ours (Res18)    | √      |      | 74.6      |
| Ours (Res18)    | √      | √    | 74.9      |

MS represent Multi-Scale. Flip represent left-right flip.

TABLE 3. Ablation study for improvement strategies on the Cityscapes validation set.

74.6%, and the left-right flip brings an improvement by 0.3%. We train our network with the best setting and experiment on the Cityscapes test dataset [24].

Furthermore, we compare our approach with other representative methods in various aspects, including Params, GFLOPs, and Mean IoU. As shown in Table 4, our method yields Mean IoU 73.6% on the Cityscapes test set. Compared with other methods, our method greatly reduces the parameters and the GFLOPs. Our method is 4.8% less accurate than PSPNet [4] while the learning parameters is 6.5× fewer, and GFLOPs is 22.3× fewer. Compared to RefineNet [3] with the same Mean IoU, our learning parameters isP 10× fewer, and GFLOPs is 64× fewer. RefineNet uses multiple identical refine blocks to fuse multi-scale features, which complicates the network, while our approach adopts two dedicated

| Method          | Backbone | mIoU% | Params(M) | GFLOPs |
|-----------------|----------|-------|-----------|--------|
| FCN-8s-heavy [1]| VGG-16   | 65.3  | 134.5     | 1319.1 |
| Dilation10 [15] | VGG-16   | 67.1  | 134.4     | 997.2  |
| DeepLab [5]     | ResNet-101 | 70.4 | 43.9      | 1481.4 |
| RefineNet [3]   | ResNet-101 | 73.6 | 118.0     | 6748.2 |
| DUC [11]        | ResNet-101 | 77.6 | 163.7     | 1135.7 |
| BiSeNet[37]     | ResNet-18 | 77.7 | 12.9      | 104.1  |
| PSPNet [4]      | ResNet-101 | 78.4 | 75.0      | 2350.9 |
| Auto-DeepLab[38] | Searched-F20-ASPP | 79.9 | -         | 333.3  |
| Dynamic[39]     | Layer33-PSP | 80.7 | -         | 270.0  |
| DANet[32]       | ResNet-101 | 81.5 | 66.5      | 376.4  |
| Ours            | ResNet-18 | 73.6 | 11.57     | 105.2  |

GFLOPs are estimated for input of $3 \times 1024 \times 2048$. 

VOLUME 8, 2020
attention mechanisms to adjust features of different phases for complementary fusion, making the network lightweight.

As shown in Table 5, we test the speed of our method on a single RTX 2080Ti GPU, the model is based on ResNet-18, and the image size is $W \times H$. The results are the average value of 200 images.

### F. RESULTS ON PASCAL VOC 2012 AND PASCAL CONTEXT

As the same as the experiment on the Cityscapes test set, we employ multi-scale input and left-right flip in evaluation on PASCAL VOC 2012 test set [24], as shown in Table 6. Furthermore, since the PASCAL VOC 2012 dataset provides a higher quality of annotation than the augmented datasets [27], we fine-tune our model on PASCAL VOC 2012 train-val set for evaluation on the test set.

As shown in Table 7, our method achieves 82.3%\(^1\) Mean IoU on PASCAL VOC 2012 test set without pre-training on the MS-COCO dataset [34], and the detailed results are listed in Table 10. Our method is 0.7% less Mean IoU than DUC, which trained their model with extra MS-COCO dataset and employed a deeper backbone (ResNet152). With the same backbone, our method is 0.2% less Mean IoU than PSPNet [4], but our GFLOPs are much less than PSPNet (about 1/7 of PSPNet).

As shown in Table 8 and Table 9, we evaluate our method on the PASCAL Context dataset. Our method aims at proving that the different weighted selection and adjustment of high-level features and low-level features can soften the fusion of features and avoid the erroneous influence of some low-level features. Although we use a very simple decoder and do not pre-train the network on additional datasets, such as MS-COCO, our method achieves good performance with lower complexity.

### V. CONCLUSION

In this paper, we propose an efficient network that adopts attention mechanisms from a macro perspective to fully utilize the different features and reduce the complexity.

---

\(^1\)The result link to the VOC evaluation server: http://host.robots.ox.ac.uk:8080/anonymous/O5UYFF.html
We evaluate our method on the Cityscapes, PASCAL VOC 2012, and PASCAL Context datasets. Specifically, two attention modules are introduced to optimize the high-level and low-level features from the feature extraction network, respectively, and the optimized features are used as the source of context information and spatial details. We demonstrated that these two modules could improve the segmentation performance remarkably by ablation experiments. Besides, we visually analyzed the intermediate features, which verified the effects of the two attention mechanisms in the architecture, and the soft fusion with a weighted selection of features could optimize the information for segmentation. In particular, our method, which has no complicated and repetitive structure for feature fusion, aims to differentiation of the features. Our method reduces the complexity of the model significantly while ensuring access to sufficient information.

### REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3431–3440.

[2] T. Ronnberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Medical Image Computing and Computer-Assisted Intervention. Springer, 2015, pp. 234–241.

[3] G. Lin, A. Milan, C. Shen, and I. Reid, “RefineNet: Multi-path refinement networks for high-resolution semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5168–5177.

[4] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 6230–6239.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.

[6] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, “Psanet: Point-wise spatial attention network for scene parsing,” in Proc. Eur. Conf. Comput. Vis., Sep. 2018, pp. 267–283.

[7] H. Ding, X. Jiang, B. Shuai, A. Q. Liu, and G. Wang, “Context contrasted feature and gated multi-scale aggregation for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 2393–2402.

[8] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, “Large kernel matters—Improve semantic segmentation by global convolutional network,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1743–1751.

[9] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, arXiv:1706.05587.

[10] A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in Proc. NIPS, 2012, pp. 1–9.

[11] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, “Understanding convolution for semantic segmentation,” in Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV), Mar. 2018, pp. 1451–1460.

[12] W. Liu, A. Rabinovich, and A. C. Berg, “Parsenet: Learning to see better,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2016.

[13] V. Badrinarayanan, A. Kendall, and R. Cipolla, “SegNet: A deep convolutional encoder-decoder architecture for image segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[14] G. Ghiasi and C. Fowlkes, “Laplacian pyramid reconstruction and refinement for semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2016, pp. 519–534.

[15] F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” in Proc. Int. Conf. Learn. Represent. (ICLR), 2016.

[16] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7132–7141.

[17] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Learning a discriminative feature network for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1857–1866.

[18] L.-C. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, “Attention to scale: Scale-aware semantic image segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3640–3649.

[19] V. Mnih, V. Mnih, N. Heess, and A. Graves, “Recurrent models of visual attention,” in Proc. Neural Inf. Process. Syst., 2014, pp. 2204–2212.

[20] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang, “Residual attention network for image classification,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 3156–3164.

[21] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua, “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 5659–5667.

[22] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 770–778.

[23] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. ICLR, 2015, pp. 1–14.

[24] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2016, pp. 3213–3223.
[25] M. R. Everingham and J. Winn. (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012). [Online]. Available: http://pascalin.ecs.soton.ac.uk/challenges/VOC/voc2012/index.html

[26] S. Xie and Z. Tu, “Holistically-nested edge detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1395–1403.

[27] B. Hariharan, P. Arbelaez, L. Bourdev, S. Maji, and J. Malik, “Semantic contours from inverse detectors,” in Proc. Int. Conf. Comput. Vis., Nov. 2011, pp. 991–998.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet large scale visual recognition challenge,” Int. J. Comput. Vis., vol. 115, no. 3, pp. 211–252, Dec. 2015.

[29] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, and C. Lawrence Zitnick, “Microsoft coco: Common objects in context,” in Proc. Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 991–998.

[30] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2018, pp. 7151–7160.

[31] S. Xie and Z. Tu, “Semantic contours from inverse detectors,” in Proc. Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3194–3203.

[32] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3146–3154.

[33] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7151–7160.

[34] G. Lin, C. Shen, A. van den Hengel, and I. Reid, “Efficient piecewise training of deep structured models for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3376–3385.

[35] ZHIQIANG XIONG received the B.S. degree from Shanghai University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the M.Sc. degree with Tongji University, Shanghai, China. His research interests include deep learning and its applications in scene parsing.

[36] ZHICHENG WANG received the Ph.D. degree from the Institute of Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, China, in 2006. He is currently a Full Professor and the Director of the CAD Research Center, Tongji University. He has published approximately 40 papers in major international journals and conferences. His research interests include pattern recognition, computer vision, and machine learning.

[37] XI GU is currently pursuing the master’s degree with Tongji University, China. Her current research interests include object detection and their applications in remote sensing.

[38] M. Mostajabi, P. Yadollahpour, and G. Shakhnarovich, “Feedforward semantic segmentation with zoom-out features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2015, pp. 3376–3385.

[39] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 3–19.

[40] R. Mottaghi, X. Chen, X. Liu, N.-G. Cho, S.-W. Lee, S. Maji, and J. Malik, “The role of context for object detection and semantic segmentation in the wild,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 891–898.

[41] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “BiseNet: Bilateral segmentation network for real-time semantic segmentation,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[42] S. Xie and Z. Tu, “Semantic contours from inverse detectors,” in Proc. Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 3194–3203.

[43] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, “Dual attention network for scene segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2019, pp. 3146–3154.

[44] H. Zhang, K. Dana, J. Shi, Z. Zhang, X. Wang, A. Tyagi, and A. Agrawal, “Context encoding for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7151–7160.

[45] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, and C. Lawrence Zitnick, “Microsoft coco: Common objects in context,” in Proc. Eur. Conf. Comput. Vis., 2014, pp. 740–755.

[46] Z. Xiong received the Ph.D. degree in electronic engineering from Wuhan University of Science and Technology, China, in 2017. He is currently an Engineer with the Shanghai Aerospace Electronic Technology Institute, China. His research interests include deep learning and its applications in object detection.

[47] ZHAOHUI YU graduated from Shanghai University. He is currently pursuing the master’s degree with Tongji University. His current research interests include objection detection, instance segmentation, model compression, and he has considerable experience in these fields.

[48] XI GU received the Ph.D. degree in electronic engineering from Tongji University, China. Her current research interests include object detection and their applications in remote sensing.

[49] JIE LI received the Ph.D. degree in electronic engineering from Wuhan University of Science and Technology, China, in 2017. He is currently an Engineer with the Shanghai Aerospace Electronic Technology Institute, China. His research interests include pattern recognition, computer vision, and machine learning.

[50] ZHIQIANG XIONG received the B.S. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the M.Sc. degree with Tongji University, Shanghai, China. His research interests include deep learning and its applications in scene parsing.