Applied Intelligence (2023) 53:18898–18909
https://doi.org/10.1007/s10489-023-04463-1

Semantic segmentation based on double pyramid network with improved global attention mechanism

Xianfeng Ou[1,2] · Hanpu Wang[1,2] · Guoyun Zhang[1,2] · Wujing Li[1,2] · Shuixiang Yu[3]

Accepted: 8 January 2023 / Published online: 14 February 2023

Abstract

Scene semantic segmentation is an important and challenging task, which requires labeling the category of each pixel in the image accurately. The encoder-decoder framework represented by fully convolutional network(FCN) has unique advantages in semantic segmentation. However, it is still hard to segment the small target and object boundary in the FCN framework. So, this paper proposes a global attention double pyramid network(GADPNet) based on an improved global attention mechanism to improve the performance of semantic segmentation. It is composed of deep convolutional neural networks Resnet-101, atrous spatial pyramid pooling(ASPP) module, proposed pyramid decoder structure and improved global attention module. Resnet-101 is the backbone which is used to extract different stages' features. ASPP module is used to capture multi-scale features from a high-level feature branch. Pyramid decoder structure can take advantage of multi-scale features from ASPP module and different stages' low-level multi-scale feature maps guided by improved global attention module. The proposed decoder enhances the ability to capture multi-scale features. GADPNet is an end-to-end network. The experimental results of the value of mIoU on Pascal VOC 2012 test dataset and cityscapes val dataset are 80.5% and 72.9%, which indicate that the proposed GADPNet obtains higher semantic segmentation accuracy compared with the current methods.

Keywords Semantic segmentation · Atrous spatial pyramid pooling · Improved global attention mechanism · Pyramid decoder structure

✉ Wujing Li
liwj@hnist.edu.cn

✉ Shuixiang Yu
missyu107@qq.com

Xianfeng Ou
ouxf@hnist.edu.cn

Hanpu Wang
wanghp568@gmail.com

Guoyun Zhang
gyzhang@hnist.edu.cn

[1] School of Information Science and Engineering, Hunan Institute of Science and Technology, Xiangbei Avenue, Yueyang, 414006, Hunan, China

[2] Machine Vision & Artificial Intelligence Research Center, Hunan Institute of Science and Technology, Xiangbei Avenue, Yueyang, 414006, Hunan, China

[3] Information Center, Hunan Institute of Science and Technology, Xiangbei Avenue, Yueyang, 414006, Hunan, China

1 Introduction

Image semantic segmentation uses the corresponding semantic information and spatial information to classify each pixel in the given image. It is a high-level image task that pave the way for scene understanding. Semantic segmentation has been used in medical image diagnosis [1], indoor scene understanding [2], multimedia content analysis [3] and other fields widely.

Traditional image semantic segmentation algorithms are mainly based on image processing technologies such as threshold segmentation [4], edge segmentation [5], graph theory segmentation [6], or combining these image processing technologies with genetic algorithms and other machine learning technologies for improvement [7]. Due to the need for manual design of parameters, operators, features, etc., the performance of traditional methods cannot meet the complex scene analysis requirements in application scenarios. In the past few years, the way of convolutional neural network [8] has obtained significant results among image classification and target recognition. By replacing the full connection in the original CNN like ResNet with full

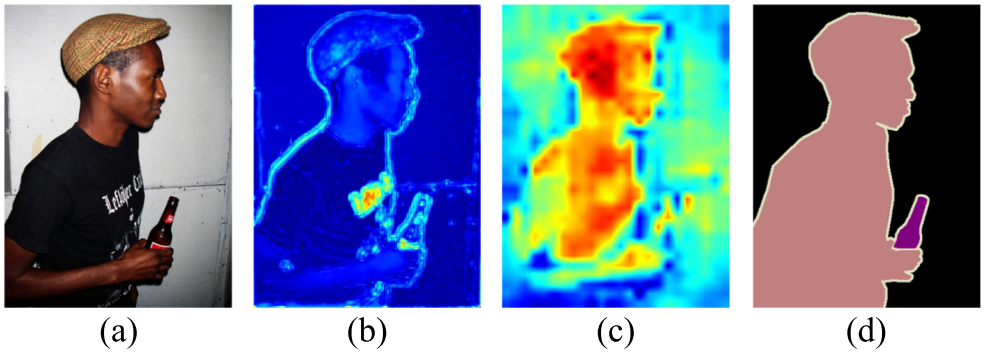

Fig. 1 Visualization results of features at different levels. (a) The RGB images, (b) The visualization map of shallow features, (c) The visualization map of deep features, (d) Ground truth

convolution, the problem of image pixel dense prediction is solved [9]. It achieves end-to-end semantic segmentation, while compared with traditional methods, the result is better.

However, one challenge in semantic segmentation is how to encode and decode efficiently. Deeper Convolutional which has been applied in extracting high-level features from original image, which is conductive to image classification. However, semantic segmentation needs to classify each pixel, so the detailed information and semantic information need to be considered jointly [10, 11]. Taking Fig. 1 for example, Fig. 1(a) is a common RGB image; Fig. 1(b) represents feature map obtained from the low layers and it contains much more details information like boundaries contour; Fig. 1(c) represents the high-level feature map, in which rich semantic information is included representing categories and attributes of object; Fig. 1(d) is the label of RGB image. Semantic segmentation is aimed to match prediction results with the label as much as possible, most current semantic networks are simply to combine different features to improve the performance of semantic segmentation. In global attention pyramid network [12], the decoder provides global context information to guide low-level to locate each pixel where the guidance come from the features of contextual information. This paper designs an effective encoding and decoding structure to enhance the segmentation ability of small targets and object edge details, and it is crucial by taking advantage of features from different levels.

Inspired by the pyramid attention network to recover the details of object [13], among this article, we propose a global attention double pyramid network(GADPNet) based on improved global attention mechanism. In general, the network structure consists of Resnet-101, atrous spatial pyramid pooling, proposed pyramid decoder and improved global attention module. The ASPP module is used to capture the scale characteristics of different receptive field objects and average pooling to obtain the average value of the sub-regions at different stages. In pyramid decoder, we upsample the features of the third and fourth stages to the same resolution of the second stages through the improved global attention module(IGAM), and then the multi-scale features are obtained through the ASPP module, all the features are connected and then are restored to the original resolution through upsampling before making pixel-level predictions finally. Specifically, the IGAM is used to guide the low functions to achieve pixel positioning better and finer details. Compared with the classfic global attention, a new branch consists of 3×3 convolution to captures shallow information over long distances is added. Without using extra training on other datasets like MS COCO, mIoU value of the proposed GADPNet is up to 80.5% above the Pascal VOC 2012 test dataset and 72.9% on the Cityscapes validation dataset.

The main contribution of the paper is:

- This article introduces the structure of the semantic segmentation encoder built by the classification network Resnet-101 and the atrous spatial pyramid pooling module in detail that captures multi-scale features, which is helpful for readers to understand the encoder structure details of semantic segmentation deeply.
- This article proposes an improved global attention module where a lightweight context branch that captures more local information of low-level features is added to the global attention module.
- This article proposes a pyramid decoder structure which realizes multi-scale features captured by a single high-level feature and fusion of different low-level multi-scale features guided by an improved global attention module. So, the multi-scale feature capture capability of semantic segmentation networks is enhanced.
- The proposed GPDANet network structure achieved mIoU values of 80.5% and 72.9% on the Pascal VOC 2012 test dataset and Cityscapes validation dataset, respectively, and achieved good experimental results.

The other parts of the paper are arranged as follows. The Section 2 presents the related works of semantic segmentation from the perspective of encoder-decoder, image pyramid and attention mechanism briefly. The Section 3 introduces the proposed GADPNet including the improved global attention moudle and the proposed pyramid

decoder. In Section 4, the article demonstrates and analyzes experimental results plenty. Section 5 summaries the article.

2 Related work

2.1 Encoder-decoder

It is important to designed an effective encoder-decoder segmentation structure to predict small target and object edge detail in the given images. For instance, in terms of encoders, the Inception [14], Xception [15], Resent [16] model composed with convolution and the operator of pooling are applied to extract features, which are used for the backbone of network. In terms of the decoders, in the representative FCN [9], the spatial information of the object is maintained through the method of skip connection. In addition, there are many excellent decoder structures. Huang et al. [17] introduce the DenseNet which is like to U-Net, concatenates the encoder with decoder at each level. Shang et al. [18] propose DXNet that uses the Xiphoid Pooling way to integrate more detailed information in the decoder structure. It connects the different features maps from multiple atrous-convolved. Dong et al. [19] design a spatial preservation module and a model to fuse freatures obtained from different branch based on a lightweight baseline. Especially, The Resnet model is very simple and easier to be optimized that is stacked by basic modules, so it is a good tool to extract features from complex images. Besides, the Resnet-101 model can obtain deeper semantic information compared with other Resnet series. So, we choose the Resnet-101 as backbone.

2.2 Image pyramid

It is effective way to obtain multi-scale information which is hidden in high-level feature maps. Chen et al. introduce [20] the atrous spatial pyramid pooling to increase receptive field of different scales. Peng et al. [21] propose a stride spatial pyramid pooling module. The module captures multi-scale semantic information form deeper features maps with faster inference. He et al. [22] propose a dynamic multi-scale network(DMNet) to balance the expansion rate and the range of scale changes. Wu et al. [23] proposed a joint pyramid up sampling module(JPU) instead of dilated convolution to low computational complexity and memory usage without loss of accuracy. We choosed ASPP module to capture multi-scale information of high-level feature maps which is the last layer of the resnet-101 module. On this basis, a pyramid decoder structure is designed that utilizes the multi-scale features obtained by atrous convolution and the joint low-level features of different stages guided by the improved global attention module.

2.3 Attention mechanism

Attention mechanism is another method to improve the performance of image segmentation. In recent years, firstly, the way of attention mechanism is used in natural language processing [24], and then applied in computer vision field like target detection and semantic segmentation. Specifically, it gives a large weight to the target that needs attention, and strengthen the learning of this part of the feature. It mainly contains location and channel attention mechanism. Zhu et al. [25] propose an valid non-local module which is aimed to reduce calculation compared with the of the typical non-local module. Fu et al. [26] introduce the DaNet which is based on the attention mechanism, by inputting the feature map to the position and channel attention module, the specail part of enhanced by those modules are fused to get the final semantic segmentation results. From the perspective of modeling contextual information in images regions, Shen et al. [27] propose a new attention network where all pixels in the feature map are enhanced. This article also studies the application of attention in related fields like video segmentation. A novel network named Co-attention Siamese Network (COSNet) is proposed which pay attention to the inherent correlation among video frames, in the network, they introduce a global co-attention mechanism to improve the performance that focus on learning discriminative foreground representations, the COSNet is used in unsupervised video object segmentation and zero-shot video object segmentation [28, 29] and achieves great results. Wang et al. [30] proposed a new framework to address the task of zero-shot object segmentation. In the framework, the corresponding between two frames are captured by an attention mechanism. Then, based on this work, Lu et al. [31] propose an attentive graph neural network(AGNN) which can examine fine-grained semantic similarities among all location in the data through the differentiable attention. Taking into account that the scale and position force mechanism module require a huge amount of calculation to improve the segmentation performance, so we choose the global attention module which is part of channel attention with low calculation amount relatively as our attention module.

3 Proposed GDAPNet

3.1 Overall architecture

The proposed GADPNet structure is shown in Fig. 2. It consists of encoder structure represented by purple box and decoder structure represented by green box. Firstly, pictures which are sent to ResNet-101 to multiple obtain feature

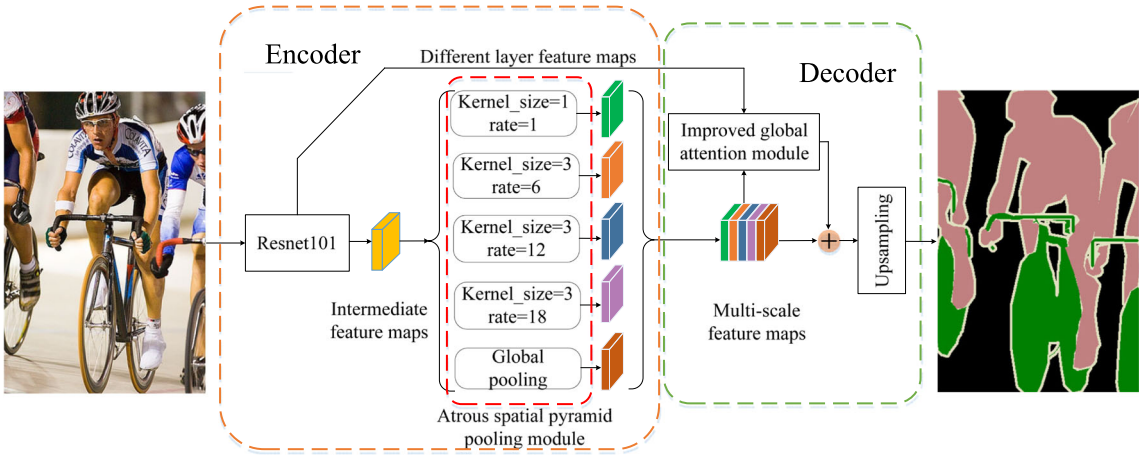

Fig. 2 Double pyramid network structure based on improved global attention module

maps with different level information. Secondly, the third and fourth stage feature map are upsampled by means of bilinear interpolation to the same size as the second stage, and the improved global attention module is used to guide the shallow feature maps with more spatial information. On this baisc, the proposed pyramid decoder architecture is aimed to fuse multi-scale feature obtained by atrous convolution and the joint low-level features of different stages guided by the improved global attention module. Finally, the feature maps are restored to resolution with the size of original image through the process of upsampling and is put into the classifier to perform pixel-segmentation result.

In this article, the deep convolutional neural network Resnet-101 is the backbone acted as the feature extraction tool which is modified from classification network and the ASPP module is used to extracted multi-scale feature with different atrous rates effectively. Below, we present the two parts detaily.

ResNet-101 is composed of multiple residual blocks. The degradation problem is addressed by presenting typical deep residual learning network. In detail, x represents input, and $F(x)$ is the output, which can be calculated by (1):

$$F(x) = W_2\sigma(W_1x + b_1) + b_2 \tag{1}$$

Where W_1 and W_2 represent the weights of two layers, b_1 and b_2 represent the bias of two layers. σ is activation function. $\sigma(F(x) + x)$ is the output.

In addition to the activation layer and pooling layer, the Resnet-101 network contains 101 convolution layers. Assuming the input size is 320×480, the network structure parameters are shown in the following Table 1.

As can be seen from Table 1, the fourth layer of the modified ResNet-101 uses atrous convolution to extract contextual information further, but the resolution of image features is unchanged. In the actual training, the full connection layer in Table 1 is discarded, and the results are input into the spatial pyramid atrous pooling module for feature extraction further.

Atrous spatial pyramid pooling module is composed of one 1×1 convolution, three 3×3 with $rates = (6, 12, 18)$ when $output_{stride} = 16$ which all have 256 filters and the operator of batch normalization, a global average pooling layer. Broadly, the atrous convolution solves a problem of loss of details with respect to the reduction of image resolution during down sampling. Besides, Compared with VGG network, which uses small convolution kernel superposition to increase receptive field linearly, atrous convolution can increase receptive field exponentially. The global average pooling module is aimed to achieve global context only. Finally, the input feature maps obtained from

Table 1 Resnet-101 structural parameters

Name of Layer	Size of output	Special parameter
Conv1	160×240	7×7, 64
Layer1	80×120	3×3 Max pooling
		$\begin{bmatrix} 1\times1 & 64 \\ 3\times3 & 64 \\ 1\times1 & 128 \end{bmatrix} \times 3$
Layer2	40×60	$\begin{bmatrix} 1\times1 & 128 \\ 3\times3 & 128 \\ 1\times1 & 512 \end{bmatrix} \times 4$
Layer3	20×30	$\begin{bmatrix} 1\times1 & 256 \\ 3\times3 & 256 \\ 1\times1 & 1024 \end{bmatrix} \times 23$
Layer4	20×30	$\begin{bmatrix} 1\times1 & 512 \\ 3\times3 & 512 \\ 1\times1 & 2048 \end{bmatrix} \times 3$
FC	1×1	Average Pool, FC, activate function

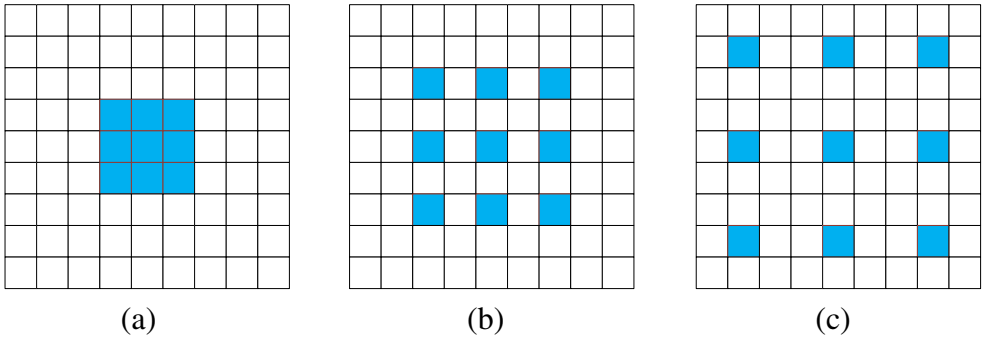

Fig. 3 The size of convolution kernel in three picture is 3 × 3 (a) The atrous rate is 1; (b) The atrous rate is 2; (c) The atrous rate is 3

Resnet-101 are processed by ASPP module, and multi-scale information is extracted without reducing the spatial resolution of the image.

To be more specific, the difference between the atrous convolution and the conventional convolution is that it has a factor, the expansion rate, which enables the filter receptive field to be expanded. The atrous convolution in the multi-scale feature extraction module is shown in (2):

$$y[i] = \sum_{k=1}^{K} x[i + s \times k]w[k] \tag{2}$$

where $y[i]$ is the output signal at the i-th position on the output feature map, and x is the start input, w is the convolution filter, k is the convolution kernel size, s is the expansion rate.

Taking the 3×3 convolution kernel as an example. When the expansion rate is equal to 1, its function is similar to conventional convolution, as shown in Fig. 3(a). However, when the expansion rate is set to 2, it has the function of expanding the expanding the convolution kernel. Theoretically, it works like this. Firstly, it expands the convolution kernel according to the expansion rate; then, it fills the empty space with zeros to create a sparse similar convolution kernel; finally, it uses the expanded convolution kernel for regular convolution.

Therefore, the k size of 3×3 and the s rate of 2 can cover a 5×5 area, as shown in Fig. 3(b). In a similar way, a conventional 3×3 convolution with an expansion ratio of 3 can obtain the corresponding 7×7 area signal, as shown in Fig. 3(c). This effect permits us to control the resolution for calculating feature response. In addition, the atrous convolution obtains a larger range of semantic information. Besides, the number of parameters during calculating is not increase.

3.2 Improved global attention module

Previous methods have represented that contextual is important to enhance the ability of semantic segmentation. It can be seen from the overall framework that the feature maps obtained in ResNet-101 as the backbone stage are sent to the improved global attention module, and then the feature maps with the corresponding resolution guided by high-level features are obtained. Compared with the classical global attention module, the proposed module uses convolution block composed with two 1×1 and one 3×3 to obtain more information while learning non-linear characteristics better by 1×1 convolution. The improved global attention module is represented in Fig. 4. Different from the SE block in the Sequeeze-and-Excitation, after extracting the global semantic information, to match the number of channels of low-level feature maps for the task of semantic segmentation, we perform a dimensionality reduction operation while performing feature map reweighting. Besides, when the global attention map is multiplied with the original feature map, a lightweight context branch that captures more local information of low-level features is added to the global

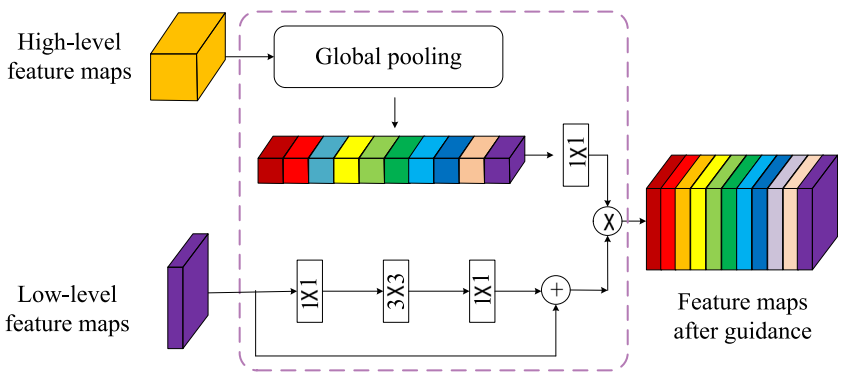

Fig. 4 Improved global attention module

attention module. The article will introduce the improved global attention module in the following part from three aspects.

In the first step, the high-level feature maps are pooled by global average to extract global context, therefore, the size of feature map achieved is resized to 1×1. For the purpose of matching the number of low-level feature channels numbers, the feature map' dimension we have obtained needs to be reduced by 1×1 convolution with the operator of Batch Normalization(BN) and ReLU. The specific calculation is shown in (3):

$$m_i = ReLu\left(BN\left(Conv\left(\frac{\sum_{h=0,x=0}^{H,W} x_i(h,w)}{H \times W}\right)\right) + b_i\right) \tag{3}$$

where H and W represent the size of deep feature maps, i shows off the corresponding channel number, and b_i represents the offset of the element and m_i shows off the weight value of the corresponding channel.

In the second step, the low-level feature maps reduce amount of calculation through 1×1 convolution which shows the channel dimension reduced, and then through a given convolution to capture more spatial information which will be modeled by global semantic information, the obtained feature maps are restored to the number of original channels by another 1×1 convolution marked as y_i'. Combined with the original low-level feature maps, the new feature map contains more spatial features. The processing in shown in (4):

$$n_i = y_i + y_i' \tag{4}$$

In the third step, we multiply the feature maps obtained by (4) by the feature maps containing multiple stage features pixel by pixel. So, the low-level features element-wise multiplication with the global attention maps generate a weighted shallow feature maps. The processing process is shown in (5):

$$v_i = m_i \otimes n_i \tag{5}$$

where $\otimes$ represent multiplication pixel-by-pixel, and v_i shows off the i-th channel value among low-level feature maps guided by high-level feature maps.

3.3 Pyramid decoder structure

The function of the decoder structure is not only to restore the boundary information of object, but also repairs the pixel-category position. In general, the decoder structure is to unsample the last layer of feature maps directly or combine intermediate feature maps simply to restore detailed object information and repair category pixel positioning. However, it is not beneficial to take advantage of the semantic which implies in high-level features or detail information of low-level features. As we have known, high-level feature maps with enough context information which can be used to rechange the weights of low-level feature maps. Due to this reason, pixels are classified better with more precise details.

Inspired by the pyramid encoder structure where multi-scale features can be achieved, in this article, we build the pyramid decoder structure where the multi low-level feature maps guided by high-level feature maps through

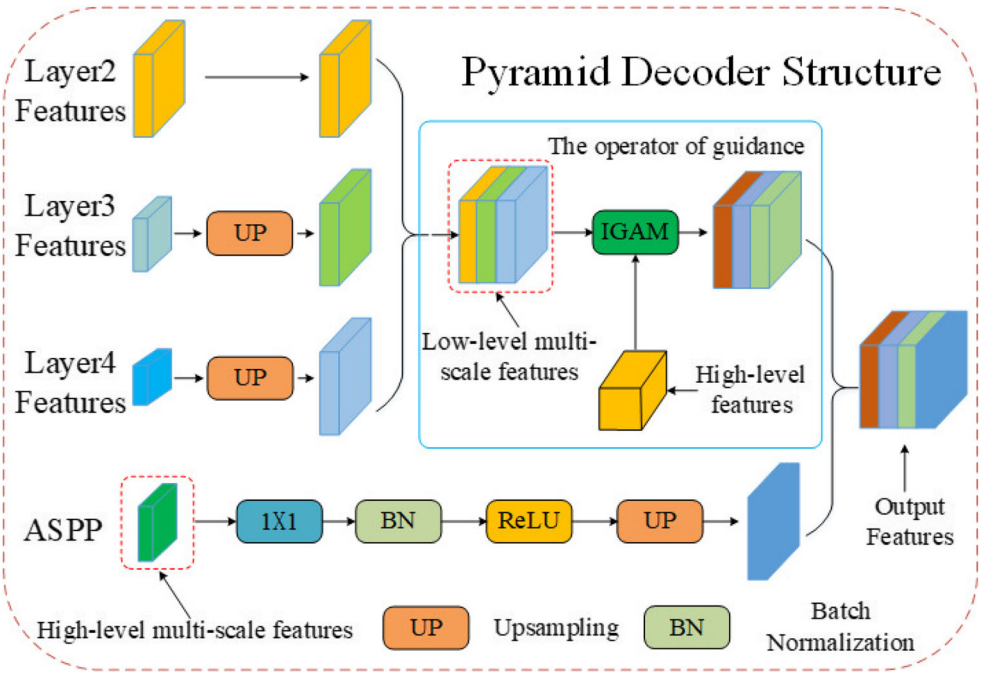

Fig. 5 The detail of pyramid decoder

Table 2 Results from global attention module guides low-level feature maps at different layers on the PASCAL VOC 2012 validation

Method	Layer2	Layer3	Layer4	mIoU (%)	PA (%)
ResNet-101+ASPP				75	93.3
ResNet-101+ASPP+GAM	✓			75.9	93.6
		✓		75.92	93.7
			✓	75.1	93.3
	✓	✓	✓	75.49	93.5

improved global attention module directly, then the high-level and the low-level feature maps are connected together. Specifically, Since the number of channels of the low-level feature maps of different layers is inconsistent, the number of channels of the low-level feature maps of each layer is reduced to 48 after guidance, and the number of channels of the feature maps captured by the ASPP module is reduced to 256 through 1×1 convolution. The number of channels which are connected is 400, and pixel-by-pixel prediction is performed from the operator of dimensionality reduction and upsampling gradually. As far as we know, the features extracted by deeper network contain enough category and context information, which can be used to guide the low-level feature maps, so pixel category and positioning information can be obtained better.

As shown in Fig. 5, Layer2 stands for the second layer operator while the yellow block represents feature maps obtained from the Layer2. In the experiment, we find that the global attention feature map obtained by the high-level feature through the improved global attention module guide the three-stage features directly, and is fused with the multi-scale feature map. However, the effect is not good for a single third-stage guidance and fusion. The study find that the details of object are lost after the guided feature map is unsampled. Therefore, in Fig. 5, feature maps from the third and fourth stage features in Resnet-101 are upsampled firstly, and the resolution is restored to the second stage's size, then three feature maps represented by the color blocks contact together as one input to the improved global pooling module to guide better pixel classification of shallow feature maps. The guided shallow multi-scale feature maps and the multi-scale feature maps achieved from ASPP module unsampled after contacting, which is input to the classifier to perform pixel-by-pixel prediction to recover detailed information better.

4 Experiments

This article evaluates the proposed GADPNet on the PASCAL VOC 2012 test dataset [32] and the Cityscapes val dataset [33]. To ensure the correctness of our experiment, we have done enough ablation experiments on the PASCAL val dataset. Then, we submit segmentation results image achieved on the test dataset to the official website to evaluate to compare with others exellent methods. Among the Cityscapes dataset, we evaluate our method on the validation dataset only.

4.1 Implementation details

The experiment in this paper depends on the pytorch framework, like the ANN and OCRNet structure, this article employs the typical "poly" strategy, the learning rate(lr) is calculated by (6):

$$lr = \left(1 - \frac{current\_iter}{\max\_iter}\right)^{power} \tag{6}$$

where $current\_iter$ and $\max\_iter$ represnet the current and max number of iterations respectively , the value of power is set to 0.9. In this paper, In the PASCAL VOC 2012 dataset and Cityscapes data set, the final number of iterations is 40K, 120K, the value of initial learning rate is 0.007, 0.01. The training strategy is stochastic gradient

Table 3 Different inference methods on the PASCAL VOC 2012 val dataset beyond global attention module

Method	Decoder way	$OS = 16$	MS	Flip	$mIoU$ (%)
ResNet-101 +ASPP+GAM	basic decoder	✓			75.49
	pyramid decoder	✓			76.49
		✓	✓		76.74
		✓	✓	✓	77.15

Table 4 Results on the PASCAL val dataset beyond improved global attention module

Methods	Decoder	Params	Run times	mIoU
ResNet-101+ASPP+GAM	Pyramid decoder	68.02M	7.93h	77.15%
ResNet-101+ASPP+IGAM		68.43M	8.43h	77.31%

descent(SGD), besides the loss function in network training is the cross-entropy function.

In this paper, the Resnet-101 classification network is modified to become the backbone network for feature extraction. The fourth layer of the modified ResNet-101 uses atrous convolution to further extract contextual information while keeping the resolution of image features unchanged. The specific parameters of the atrous convolution in the Resnet-101 network are as follows: $kernel = 3, stride = 1, padding = 2, rate = 2$. We run the model on four 2080Ti graphics cards. During the operation, we use the CUDNN engine to improve the speed of training processing.

In order to evaluate the semantic segmentation effect of the GADPNet in this paper, the experimental results are analyzed from both the subjective and objective evaluation aspects. Subjective evaluation mainly visualizes smaller targets, edge segmentation and part of object from the results of semantic segmentation. As for objective evaluation, this article uses mean Intersection over Union(mIoU) and Pixel Accuracy(PA) as evaluation indicators. PA represents the ratio of the special number of pixels classified correctly among all pixels. MIoU is defined as the ratio of the intersection and union of ground truth and predicted segmentation. The calculation formulas are as follows:

$$PA = \frac{\sum_{i=0}^{K} p_{ii}}{\sum_{i=0}^{K} \sum_{j=0}^{K} P_{ij}} \tag{7}$$

$$mIoU = \frac{1}{K+1} \sum_{i-0}^{K} \frac{p_{ii}}{\sum_{j=0}^{K} p_{ij} + \sum_{j=0}^{K} p_{ji} - p_{ii}} \tag{8}$$

Where K represents foreground categories numbers, p_i represents the true value of i, which is predicted to be the number of j.

4.2 Experiments on PASCAL VOC 2012 dataset

4.2.1 Ablation study

This article conducts ablation experiments on the Pascal val dataset to confirm our ideas from the following tables. It is a very popular data set that provides 20 common different

Table 5 IoU of each class on the Pascal VOC 2012 val dataset

Methods	FCN [9]	ESPNetv2 [34]	PTIA-Net [35]	ANN [36]	OCRNet [37]	DSM [38]	SFFN [39]	DeepLabv2 [40]	Proposed Way
aero	76.8	87.5	89.4	88.1	89.7	90.6	90.5	93.2	94.5
bike	34.2	36.9	42.7	56.4	54.9	37.6	67.3	66.8	68.6
bird	68.9	75.9	86	84.6	87.9	80	88.8	91.6	92.9
boat	49.4	64	66.3	72	72.7	67.8	74.5	64.9	68.3
bottle	60.3	63.8	78.8	74.3	72.2	74.4	72.3	79.6	80.2
bus	75.3	87.2	93.1	88.4	88.7	92	85.5	94.1	94.4
car	74.7	73.7	83.6	84.3	86.3	85.2	85.7	87.2	89.3
cat	77.6	76.5	89.9	86.9	86.7	86.2	88.8	90.7	92
chair	21.4	26.7	36.4	29.6	32.9	39.1	29.6	36.5	36.9
cow	64.5	70.3	79.5	78.3	85	81.2	88.1	85.3	86.7
table	46.8	57.5	55.8	53.7	60.7	58.9	61	63.3	63.8
dog	71.8	68.9	83.1	78.8	80.3	83.8	84.1	84.5	85.9
horse	63.9	70.6	77.7	69.9	77.4	83.9	80.1	85.4	86.6
mbike	76.5	82.9	81.4	79	84.3	84.3	83.9	86.0	88.3
person	73.9	78.9	85.3	83.6	83.2	84.8	85.3	86.9	89.4
plant	45.2	48.1	57.6	54.4	55.4	62.1	61.5	65.4	67.3
sheep	72.4	76.4	83.6	78.5	77.6	83.2	81.4	91.8	92.8
sofa	37.4	46.9	45.8	44.5	45.2	58.2	45.3	56.8	57.8
train	70.9	77.7	84.8	80.8	80.2	80.8	82.3	81.8	84.4
tv	55.1	64.1	71.5	69.7	74.5	72.3	75.6	71.1	73.8
mIoU	62.2%	68.0%	74.6%	74.5%	74.7%	75.3%	76.4%	79.0%	80.5%

objects, including cars, bicycles, people, etc. The data contains 1464 training pictures, 1449 verification pictures, and 1445 test pictures. Besides, the Pascal dataset also includes 10582(trainaug) training images which can be used for training. In order to prevent over-fitting, we use random mirror images and pictures with a size between 0.5 and 2 to enhance the data set. We also use -15 to 15° random rotation to enlarge the semantic segmentation dataset. We choose 513×513 as the input crop size.

As shown in Table 2, symbol "✓" of Layer 3 represents the shallow feature maps got from the third layer of ResNet-101 is guided by semantic information achieved by gloabl attention meanchinm. At this moment, the obtained value of mIoU is 75.90% and the PA is 93.7%. However, the mIoU value is only 75.49% when the global pooling guidance for all the three stages'feature, the effect is not improved because the details are lost after the feature map is sampled after the guidance, which is not conducive to pixel-by-pixel prediction.

In order to confirm our ideas, as shown in Table 3, output stride(OS) represents the ratio of the size of feature maps get from ResNet-101 with the size of the original picture. multiscale(MS) represents multi scale prediction and Flip represents flip prediction. multiscale and flip represent the different way to get the final prediction results. The low-level feature maps are upsampled before being guided by high-level features while the rest remains unchanged. As a result, the mIoU value is 76.49%, which means the segmentation performance improved effective. In an effort to further enhance the ability of semantic segmentation, the mIoU value is increased to 76.74% when we use multi-scale prediction. Finally, we combine multi-scale with random flip prediction, the final mIoU value is up to 77.15%.

We introduce the improved global attention module to enhance the ability of segmentation further. As show in Table 4, the original model based on the methods of GAM, the parammeters is 68.02M and the corresponding mIoU is 77.15%. When the IGAM module is added, the amount of parameters grows to 68.43M, the mIoU value increased by 0.16% to 77.31%. Besides, in order to minimize the impact of multiple GPU training data exchange on the model time, we train on a single 3090Ti server. When the number of iterations is 40K, after adding the IGAM module, the training time of the model is 8.43h.

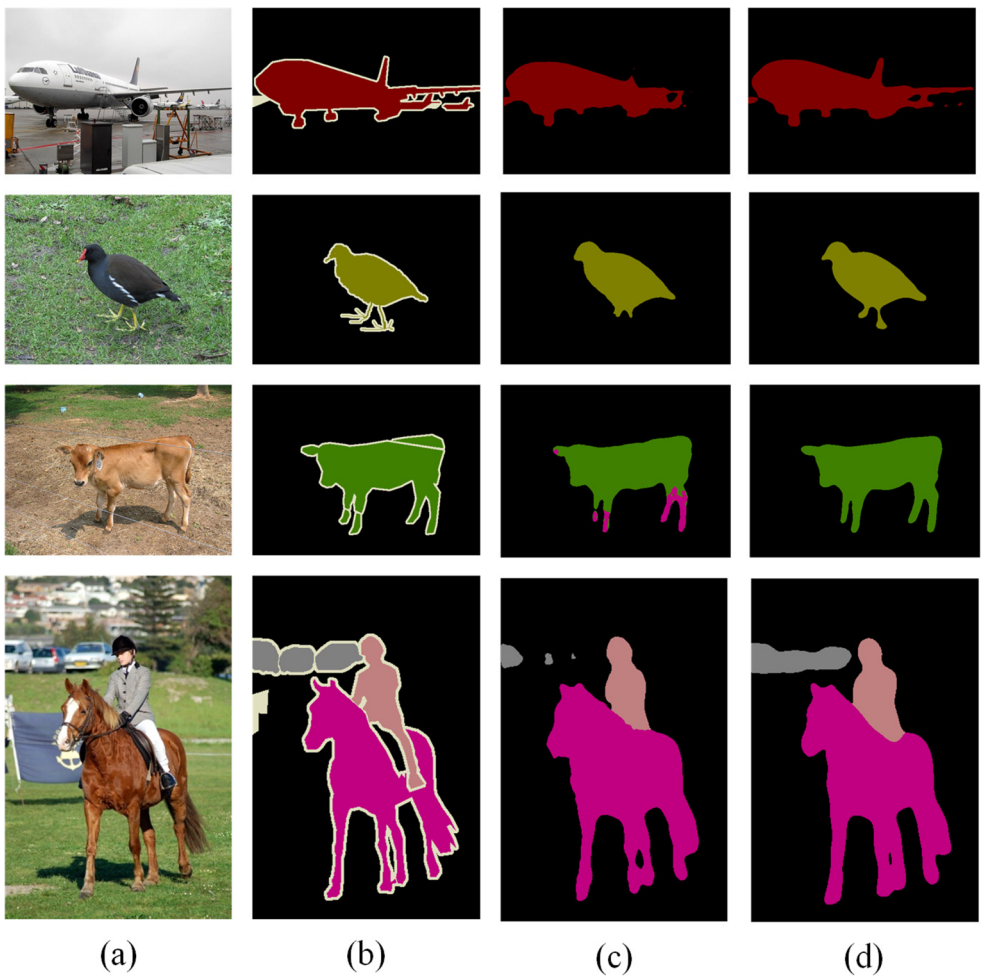

Fig. 6 Visualization results of PASCAL VOC 2012 test. (a) RGB Images, (b) Ground truth, (c) DeepLabv2, (d) Proposed GADPNet

4.2.2 Comparative experiments

Under the circumstances of no post-processing like CRF, or pretraining model on large-scale MS-COCO dataset is used, this article compare with other advanced semantic segmentation ways. We only train our model with the learning rate = 0.0001 on trainval dataset for it provides higher annotations. Specifically, the batch normalization parameters are frozen when training another model. To prove the effect of the proposed GADPNet, this article does some contrast experiments on dataset with advanced methods, such as FCN [9], ESPNetv2 [34], PTIA-Net [35], ANN [36], OCRNet [37], DSM [38], SFFN [39], DeepLavb2 [40]. The ways of DeepLab and DeepLabv2 only combine high and low level simply. As can be seen from Table 5, the final mIoU on test dataset is 80.5%, and it shows best performance. Besides, when the proposed decoder and imrpoved global attention module is attached to the network of DeepLabv3, the better mIoU can be got. The mIoU value on the Pascal VOC 2012 test dataset is 81.6% finally while the mIoU value obtained by the classic DeepLabv3 method is 80.3% (Fig. 6).

As shown in Fig. 7, the proposed GADPNet in this article is compared with the DeepLabv2 method from subjective aspect. As shown by the airplane in the first row, the DeepLabv2 method cannot segment the tail part very well, while our GADPNet can better identify the tiny part of the tail; in the third row, DeepLabv2 can identify the main body as a cow. But for some parts like horse legs, the cow legs are misidentified. In the GADPNet of this paper, the cow can be completely segmented using the global attention mechanism; for the last row, compared with DeepLabv2, our proposed GADPNet reduces partial division. It can be found that our proposed GADPNet helps to fine-tune the details of objects such as airplanes and birds, and reduce the misclassification of some objects.

4.3 Results on cityscapes datasets

This article evaluates the proposed GADPNet on the Cityscapes data set. The dataset provides 19 different common objects for semantic segmentation tasks. Two different types of images are annotated, 5000 images are annotated finely while 19,998 images are annotated coarsely. Among them, images which are annotated finely are divided into 2975, 500 and 1525 for training, testing and verification.

In the verification phase, we use multi-scale and random flip prediction. To prove the effect of the proposed GADPNet, this article does some contrast experiments on dataset with state-of-the-art methods, such as FCN [9], GANet [12], SqueezeNAS [41], FSFNet [42], DeepLabv2 [40], Degenet [43]. In result, we only train our model on images which are annoted finely. The validation performance results are reported in Table 6. The visualization results of the method proposed and two comparative experiments in this paper on the Cityscapes verification data set are shown in Fig. 7. From left to right are RGB Images, GANet, DeepLabv2, Proposed Ways. In general, as can be seen from the Fig. 7, as indicated by the blue color box, the method proposed in this paper has a higher ability to capture the edges of objects such as buildings, trees, and vehicles than GANet which is marked by the red color box. It is also marked that the method

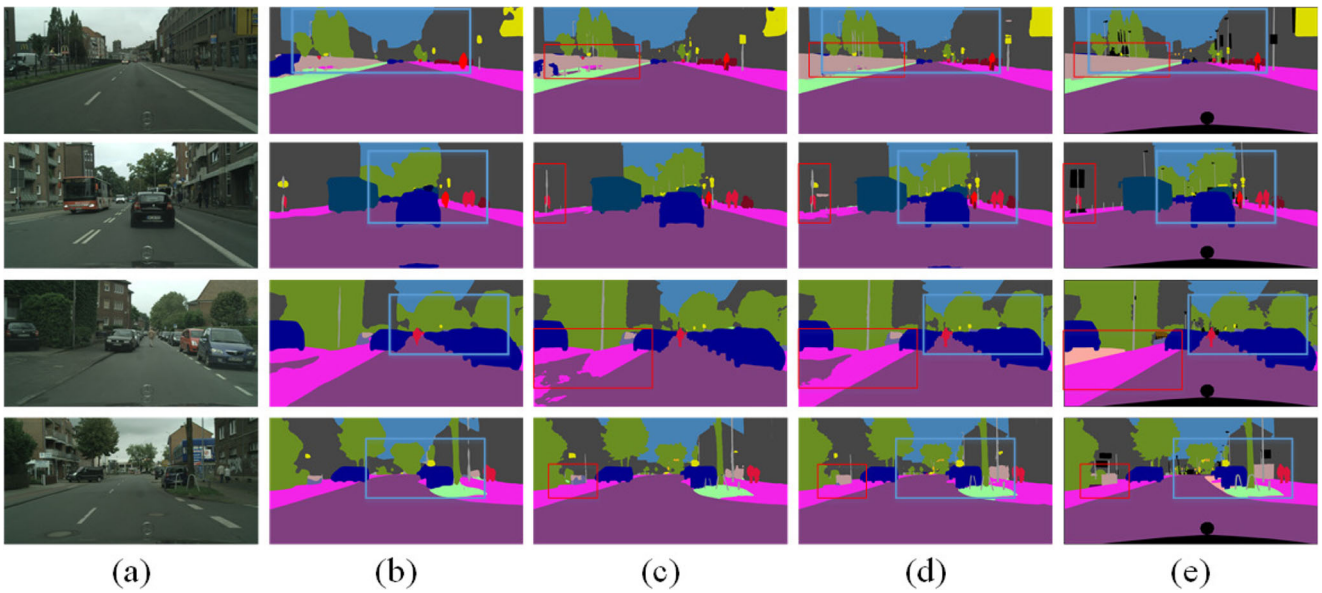

Fig. 7 Visualization results of Cityscapes validation dataset. (a) RGB images, (b) GANet, (c) DeepLabv2, (d) Proposed Way, (d) Ground Truth

Table 6 The mIoU value on the Cityscapes dataset

Methods	mIoU
FCN [9]	65.30%
GANet [12]	67.82%
SqueezeNAS [41]	68.00%
FSFNet [42]	69.70%
DeepLabv2 [40]	70.40%
Ddgenet [43]	71.00%
Proposed Way	72.90%

proposed in this paper enhances the ability to capture global information after introducing the improved global attention module, and the segmentation accuracy is enhanced for large objects like roads and small objects like people.

5 Conclusion

In this article, based on the FCN network structure, we proposed a global attention double pyramid network(GADPNet) structure based on the improved global attention mechanism. We used the IGAM to generate a global contextual attention map, which was used to guide the low-level feature maps to located better and classified correctly. The pyramid decoder was designed to enhance the ability to capture of multi-scale featuers by combining the different multi-features got from many layers. The results showed that our proposed GADPNet achieved better results in PASCAL and Cityscapes datasets without pre-training on larger datasets or post-processing. Future work of this study is as follows:

1) In addition to do experiments on these two datasets like Pascal and Cityscapes, We will apply the proposed GADPNet to some other challenging semantic segmentation such as ADE20K dataset [40] and PASCAL-Context dataset, the deeper features should be captured in the improved global attention module. [41].
2) We will study the relationship about channels and global attention before decoder in semantic segmentation to further improve the performance of semantic segmentation.

Acknowledgments This work was supported by Hunan Provincial Natural Science Foundation (2020JJ5218), the Scientific Research Fund of Education Department of Hunan Province (22A0417), and the Hunan Provincial Innovation Foundation for Postgraduate (CX20201114), General project of Hunan Water Resources Department (XSKJ2021000-13), the Open Fund of Education Department of Hunan Province (20K062), Hunan University Students Innovation and Entrepreneurship Training Project (2021-20-3151).

References

1. Jiang F, Grigorev A, Rho S, Tian Z, Fu Y, Jifara W, Adil K, Liu S (2018) Medical image semantic segmentation based on deep learning. Neural Comput Applic 29(5):1257–1265
2. Wang Y, Chen Q, Chen S, Wu J (2020) Multi-scale convolutional features network for semantic segmentation in indoor scenes. IEEE Access 8:89575–89583
3. Wang M, Li H, Tao D, Wu X (2012) Multimodal graph-based reranking for web image search. IEEE Trans Image Process 21(11):4649–4661
4. Bhargavi K, Jyothi S (2014) A survey on threshold based segmentation technique in image processing. Int J Innov Res Develop 3(12):234–239
5. Zhang Y, Li X, Gao X, Zhang C (2016) A simple algorithm of superpixel segmentation with boundary constraint. IEEE Trans Circuits Syst Video Technol 27(7):1502–1514
6. Bhargavi K, Jyothi S (2013) A survey of graph theoretical approaches to image segmentation. Pattern Recognit 46(3):1020–1038
7. Kang W, Yang QQ, Liang RP (2009) The comparative research on image segmentation algorithms. In: 2009 first international workshop on education technology and computer science, vol 2, pp 703–707
8. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556
9. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3431–3440
10. Matthew D, Fergus R (2014) Visualizing and understanding convolutional networks
11. Yosinski J, Clune J, Nguyen A, Fuchs T, Lipson H (2019) Pyramid context contrast for semantic segmentation. IEEE Access 7:173679–173693
12. Zhang N, Li J, Li Y, Du Y (2019) Global attention pyramid network for semantic segmentation. In: 2019 chinese control conference (CCC), pp 8728–8732
13. Sang H, Zhou Q, Zhao Y (2020) Pcanet: pyramid convolutional attention network for semantic segmentation. Image Vis Comput 103:103997
14. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Scott D, Dragomir E, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. Proc IEEE Conf Comput Vision Pattern Recognit:1–9
15. Chollet F (2017) Xception: deep learning with depthwise separable convolutions. Proc IEEE Conf Comput Vision Pattern Recognit:1251–1258
16. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Conf Comput Vision Pattern Recognit:770–778
17. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. Proc IEEE Conf Comput Vision Pattern Recognit:4700–4708
18. Shang Y, Zhong S, Gong S, Zhou L, Ying W (2019) DXNEt: an encoder-decoder architecture with XSPP for semantic image segmentation in street scenes. Int Conf Neural Inf Process:550–557
19. Dong G, Yan Y, Shen C, Wang H (2021) Real-time High-performance semantic image segmentation of urban street scenes. Trans Intell Trans Syst 22(6):3258–3274
20. Chen LC, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: European conference on computer vision (ECCV), pp 801–818

21. Peng C, Ma J (2020) Semantic segmentation using stride spatial pyramid pooling and dual attention decoder. Pattern Recognit 107:107498
22. He J, Zhang Z, Qiao Y (2019) Dynamic multi-scale filters for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 3562–3572
23. Wu H, Zhang J, Huang K, Liang K, Yu Y (2019) Rethinking dilated convolution in the backbone for semantic segmentation. arXiv:1903.11816
24. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A (2017) I, Polosukhin, attention is all you need. Adv Neural Inf Process Syst:5998–6008
25. Zhu Z, Xu M, Bai S, Huang T, Bai X (2019) Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 593–602
26. Fu J, Liu J, Tian H, Li Y, Fang YBZ, Lu H (2019) Dual attention network for scene segmentation. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3141–3149
27. Shen D, Ji Y, Li P, Wang Y, Lin D (2020) Ranet: region attention network for semantic segmentation. Adv Neural Inf Process Syst 33:13927–13938
28. Lu X, Wang W, Shen J, Ma C, Shen J, Shao L, Porikli F (2019) See more, know more: unsupervised video object segmentation with co-attention siamese networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 3623–3632
29. Lu X, Wang W, Shen J, Ma C, Shen J, Shao L, Porikli F (2020) Zero-shot video object segmentation with co-attention siamese networks. IEEE Trans Pattern Anal Mach Intell
30. Wang W, Lu X, Shen J, Crandall D, Shao L (2019) Zero-shot video object segmentation via attentive graph neural networks. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 9236–9245
31. Wang W, Lu X, Shen J, Crandall D, Shao L (2021) Segmenting objects from relational visual data. IEEE Trans Pattern Anal Mach Intell
32. Everingham M, Van Gool L, Williams CK, Winn J (2010) The pascal visual object classes (voc) challenge. Int J Comput Vision 88(2):303–338
33. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223
34. Mehta S, Rastegari M, Shapiro L, Hajishirzi H (2019) Espnetv2: a light-weight, power efficient, and general purpose convolutional neural network. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9190–9200
35. Zhu H, Zhang M, Zhang X, Zhang L (2021) Two-branch encoding and iterative attention decoding network for semantic segmentation. Neural Comput Applic 33(10):5151–5166
36. Zhu Z, Xu M, Bai S, Huang T, Bai X (2019) Asymmetric non-local neural networks for semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 593–602
37. Yuan Y, Chen X, Wang J (2020) Object-contextual representations for semantic segmentation. European Conf Comput Vision:173–190
38. Lin G, Shen C, van den Hengel A, Reid I (2018) Exploring context with deep structured models for semantic segmentation. IEEE Trans Pattern Anal Mach Intell 40(6):1352–1366
39. Zhou Z, Zhou Y, Wang D, Mu J, Zhou H (2021) Self-attention feature fusion network for semantic segmentation. Neurocomputing 453:50–59
40. Chen LC, Papandreou G, Kokkinos I (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Patt Anal Mach Intell 40(4):834–848
41. Shaw A, Hunter D, Landolar F, Sidhu S (2019) Squeezenas: fast neural architecture search for faster semantic segmentation. In: Proceedings of the IEEE/CVF international conference on computer vision workshops, pp 1–11
42. Kim M, Park B, Chi S (2020) Accelerator-aware fast spatial feature network for real-time semantic segmentation. IEEE Access 8:226524–226537
43. Han HY, Chen YC, Hsiao PY, Fu LC (2020) Using channel-wise attention for deep CNN based real-time semantic segmentation with class-aware edge information. IEEE Trans Intell Transp Syst 22(2):1041–1051