Complex & Intelligent Systems
https://doi.org/10.1007/s40747-024-01428-w

ORIGINAL ARTICLE

MAFNet: dual-branch fusion network with multiscale atrous pyramid pooling aggregate contextual features for real-time semantic segmentation

Shan Zhao[1] · Yunlei Wang[1] · Xuan Wu[1] · Fukai Zhang[1]

Received: 2 January 2024 / Accepted: 9 March 2024

Abstract
Currently, many real-time semantic segmentation networks aim for heightened accuracy, inevitably leading to increased computational complexity and reduced inference speed. Therefore, striking a balance between accuracy and speed has emerged as a crucial concern in this domain. To address these challenges, this study proposes a dual-branch fusion network with multiscale atrous pyramid pooling aggregate contextual features for real-time semantic segmentation (MAFNet). The first key component, the semantics guide spatial-details module (SGSDM) not only facilitates precise boundary extraction and fine-grained classification, but also provides semantic-based feature representation, thereby enhancing support for spatial analysis and decision boundaries. The second component, the multiscale atrous pyramid pooling module (MSAPPM), is designed by combining dilation convolution with feature pyramid pooling operations at various dilation rates. This design not only expands the receptive field, but also aggregates rich contextual information more effectively. To further improve the fusion of feature information generated by the dual-branch, a bilateral fusion module (BFM) is introduced. This module employs cross-fusion by calculating weights generated by the dual-branch to balance the weight relationship between the dual branches, thereby achieving effective feature information fusion. To validate the effectiveness of the proposed network, experiments are conducted on a single A100 GPU. MAFNet achieves a mean intersection over union (mIoU) of 77.4% at 70.9 FPS on the Cityscapes test dataset and 77.6% mIoU at 192.5 FPS on the CamVid test dataset. The experimental results conclusively demonstrated that MAFNet effectively strikes a balance between accuracy and speed.

Keywords Semantic segmentation · Real time · Multiscale · Pyramid pooling · Autonomous driving

Introduction

Semantic segmentation is an important technique in the field of computer vision, with the objective of assigning each pixel to a distinct semantic category within an image. Currently, the application scenarios of semantic segmentation include, but are not limited to, some practical application scenarios such as autonomous driving [1] and medical imaging analysis [2]. The advent of the convolutional neural network (CNN) marked the inception of early semantic segmentation networks like fully convolutional networks (FCN) [3] and U-Net [4].

In particular, FCN demonstrated remarkable performance in the field of semantic segmentation, marking a significant breakthrough. However, the need for more high-performance semantic segmentation networks became apparent as technology evolved. Additionally, existing networks fall short of meeting the demands of general scenarios, given that achieving higher performance often requires substantial computational resources, especially when relying on complex backbone networks like ResNet101 [5]. ResNet101 is a deep CNN with 101 layers, which addresses challenges related

Shan Zhao, Yunlei Wang, Xuan Wu and Fukai Zhang contributed equally to this work.

✉ Yunlei Wang
212209020083@home.hpu.edu.cn

Shan Zhao
zhaoshan@hpu.edu.cn

Xuan Wu
1248192438@qq.com

Fukai Zhang
zhangfukai@hpu.edu.cn

[1] School of Software, Henan Polytechnic University, 2001 Century Avenue, Jiaozuo 454000, China

Published online: 17 April 2024

to gradient vanishing and exploding during deep network training by introducing a residual learning architecture. Similarly, DeepLabV3+ [6] is a powerful semantic segmentation model that enhances segmentation accuracy by employing techniques such as dilated convolutions and atrous spatial pyramid pooling (ASPP). However, the computational cost of these high-accuracy networks, involving hundreds of giga floating-point operations per second (GFLOs), hinders their suitability for general scenarios like autonomous driving and intelligent transportation. However, some studies have already offered potential insights into semantic segmentation issues in practical application scenarios. Employing the repetitive process control in [7] can enhance batch processing efficiency, potentially contributing practicality to real-time semantic segmentation. Bipartite synchronization in neural networks with event-triggered mechanisms in [8] aims to enhance cooperative operations, with the potential to optimize segmentation strategies. Reference [9] presents a hysteresis-quantified control method for switched systems, offering a potential solution for dynamic scenarios in real-time semantic segmentation.

To address the challenge of high computational costs and meet the real-world demand for network inference speed, the development of real-time semantic segmentation is gradually gaining attention. ENet [10], designed for low-latency operations, offers comparable or superior accuracy to state-of-the-art models. While ICNet [11] introduced a cascaded feature fusion unit for high-quality segmentation results with efficient inference speed. STDCNet [12] presented a short-term dense concatenate module (STDCM), an efficient network structure that progressively reduces the dimension of feature maps. S^2-FPN [13] proposed a scale-aware strip attention module (SSAM) with low computational overhead to collect remote context along the vertical axis by striping operations and reduce computational cost. While these methods achieve commendable segmentation accuracy and speed, striking a balance between accuracy and speed in real-time semantic segmentation remains a challenging task.

To achieve this balance, some models employ lightweight convolution structures to maintain relatively low computational complexity while achieving faster inference speeds. ESPNet [14] introduced a novel convolutional module, decomposing the standard convolution into a spatial pyramid of pointwise and dilated convolutions. EfficientNet [15] systematically explored model scaling, constructing a simple and efficient composite coefficient to regulate the relationship between depth, width, and resolution of the network. Fast-SCNN [16] proposed a shallow learning to downsample module to extract low-level features quickly and efficiently. Although these lightweight convolution structures enhance speed and maintain relatively low computational complexity to some extent, they often fall short of achieving the desired accuracy. Additionally, lightweight structures may lead to the loss of important features in input images, including spatial detail information. Therefore, these issues must be considered when designing and optimizing lightweight semantic segmentation networks. Additionally, the receptive field of the network should be expanded, and richer spatial detail information should be extracted.

Various solutions have been proposed to address the challenges posed by the aforementioned lightweight structures while preserving richer spatial details. Among them, multiscale feature fusion stands out as a common strategy to enhance semantic segmentation performance. This approach integrates features from different hierarchical levels with the goal of improving the accuracy and robustness of segmentation results. For instance, DDRNet [17] introduced the deep aggregation pyramid pooling module (DAPPM) to enhance context aggregation by combining continuous forward transmission flows in Res2Net [18]. LBN-AA [19] proposed the Distinctive Atrous Spatial Pyramid Pooling (DASPP) which was designed according to the receptive field theory of the human visual system [20] using dilated convolution with different rates to enlarge the receptive field for better context information. Although these multiscale feature fusion modules enhance computational efficiency, their simple structure may lead to the loss of fine-grained feature information, potentially causing a drop in segmentation accuracy. PP-LiteSeg [21] addressed this issue by introducing the simple pyramid pooling module (SPPM). These modules use simple feature pyramid pooling to aggregate context information, thereby improving segmentation accuracy with only a slight increase in inference time. However, SPPM eliminates shortcut branches and adopts simple addition operations between elements, making it challenging to preserve the original feature map information.

To enhance segmentation accuracy while ensuring real-time network performance, novel dual-branch network structure such as BiSeNet [22], BiSeNetV2 [23], DDRNet, and RTFormer [24] have been proposed. BiSeNet introduced a detail branch for extracting details and boundary information and a semantic branch for capturing global context information, retaining abundant spatial details. BiSeNetV2 built upon the BiSeNet model, achieving improved performance and real-time inference speed by enhancing the detail branch and introducing a lightweight spatial branch. Additionally, the recent DDRNet, adopting the dual-branch structure, has demonstrated promising results. RTFormer efficiently collected global context information of high-resolution branches through cross-resolution attention propagation, followed by high-level knowledge learned from low-resolution branches. These networks aim to provide multiscale feature fusion and information recovery to enhance segmentation accuracy.

In this study, a dual-branch fusion network with multiscale atrous pyramid pooling aggregate contextual features for real-time semantic segmentation (MAFNet) is proposed.

While ensuring speed, this network improves accuracy by deviating from traditional encoder–decoder structure by adopting a dual-branch network structure. The semantic branch is constructed using the STDC1 [12] backbone network, while the spatial detail branch is constructed using the lightweight RB [17]. For improved spatial-detail extraction, a semantics guide spatial-details module (SGSDM) is designed to enable accurate boundary extraction, fine-grained classification, and semantic-based feature representation, supporting spatial analysis and decision boundaries. Drawing inspiration from LBN-AA [19], a multiscale atrous pyramid pooling module (MSAPPM) is introduced to extract rich context information using different dilated convolution and pooling operations. Finally, to balance the weight relationship between the dual-branch, a bilateral fusion module (BFM) is designed to fuse dual-branch information effectively by calculating feature information to generate weights. The effectiveness of the designed modules is verified in the final ablation experiments.

The main contributions of this paper can be summarized as follows:

1. The SGSDM serves as a pivotal connector between dual branches, seamlessly incorporating ordinary convolution, depth-wise separable convolution, and a sophisticated activation function. This innovative module not only excels in precise boundary extraction and fine-grained classification, but also goes beyond by delivering a distinctive semantic-based feature representation. This feature representation is instrumental in supporting spatial analysis and enhancing decision boundaries with unparalleled accuracy.
2. The MSAPPM is meticulously crafted, leveraging feature pyramid pooling and dilated convolutions with dynamically varying dilation rates. This strategic design not only dramatically expands the network's receptive field but also skillfully aggregates rich context information, resulting in a substantial boost in segmentation accuracy.
3. The revolutionary BFM is introduced with an advanced cross-fusion strategy. This module ingeniously computes feature information from the dual-branch architecture, generating weights that effectively balance the intricate relationship between the dual branches. The result is a seamless aggregation of essential information from both branches, optimizing overall performance.
4. Rigorous experiments are conducted on two prominent datasets to underscore the effectiveness of MAFNet. Operating on a single A100 GPU, MAFNet attains an impressive 77.4% mean Intersection over Union (mIoU) with a swift inference speed of 70.9 FPS on the challenging Cityscapes test dataset. Furthermore, MAFNet achieves a remarkable 77.6% mIoU, coupled with an impressive inference speed of 194.5 FPS on the demanding CamVid test dataset. These experimental results firmly establish MAFNet's exceptional performance and underscore its prowess in real-world applications.

Related work

In this section, some real-time semantic segmentation methods that are relevant to the work presented in this paper is given.

Dual-branch network structure

The dual-branch network structure typically comprises two parallel branches, each dedicated to distinct information extraction and feature learning. Fusing or interacting with these dual-branch features enables comprehensive utilization of feature representation from different levels, enhancing model performance and accuracy. Currently, many dual-branch networks have been proposed. Notable among the existing dual-branch networks is BiSeNetV2 [23], an improved version of the BiSeNet [22] model. Figure 1 shows the backbone architecture of BiSeNetV2.

As depicted in Fig. 1, BiSeNetV2 consists of detail and semantic branches. The detail branch requires a substantial number of channels to encode spatial detail information. This branch adopts a wide channel and shallow structure, avoiding the use of a residual structure to maintain efficiency and speed. Due to its large spatial size and wide channel characteristics for feature representation, the detail branch effectively extracts rich spatial details. On the other hand, the semantic branch operates in parallel and is designed to capture high-level semantics with lower channel capacity. It employs a fast-downsampling strategy and global average pooling to enhance feature representation and expand the network's receptive field. This design choice transforms the semantic branch into a lightweight model capable of meeting the requirements for a large receptive field and global context information extraction.

Efficient backbone network STDCNet

Efficient backbone networks play a crucial role in improving segmentation performance in real-time semantic segmentation models. These networks typically combine optimization strategies for depth and width, considering both computational and storage resource constraints at the same time. Consequently, the accuracy of semantic segmentation and the ability to preserve details are enhanced with the learning of more accurate and richer image feature representations, enabling the model to better understand the semantic information in the image.

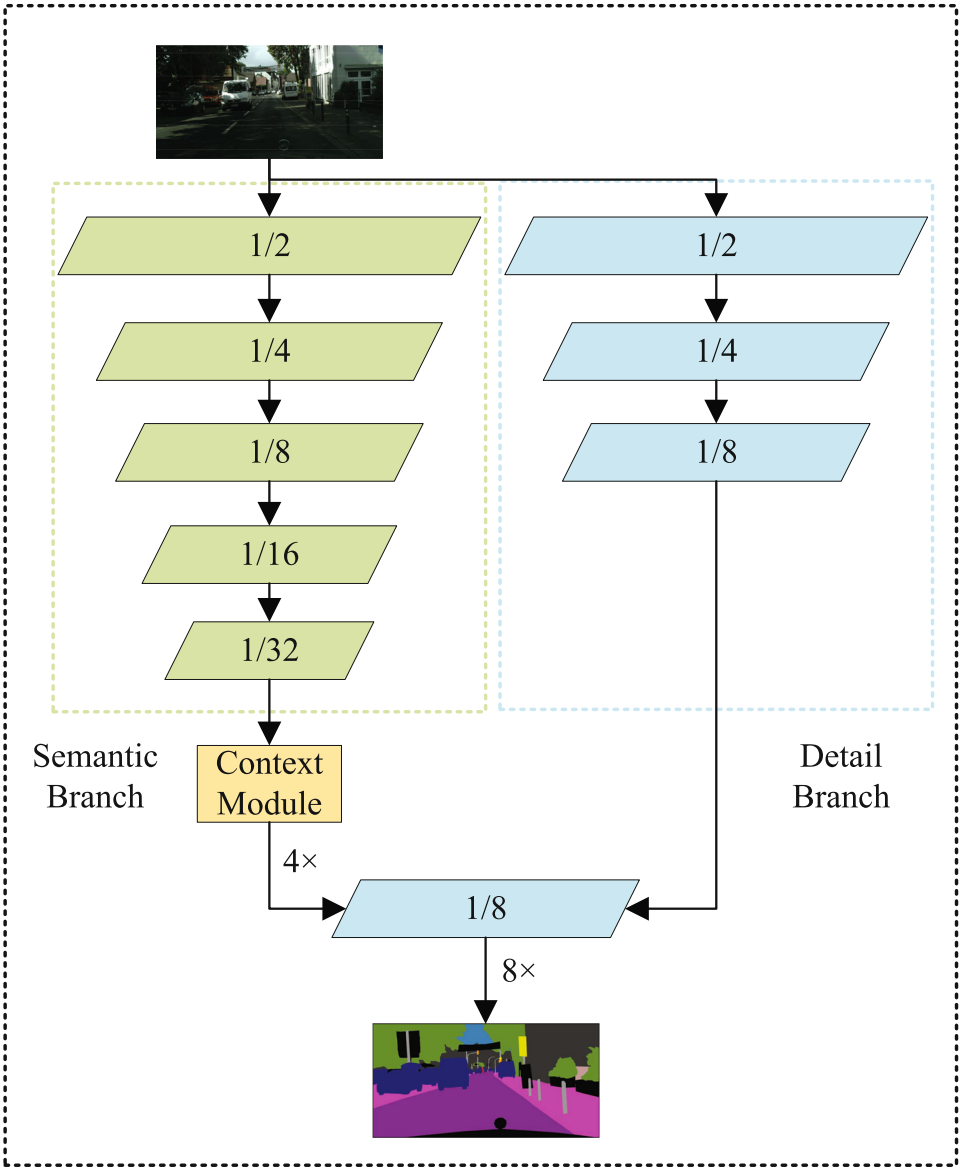


Fig. 1 The backbone architecture of BiSeNetV2

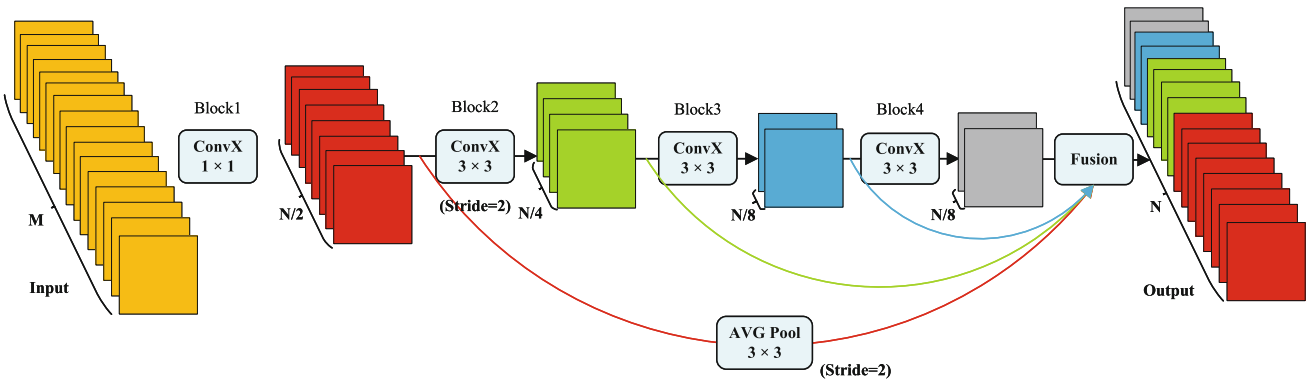


Fig. 2 Overview of STDCM with a step size of 2

STDCNet [12], an efficient backbone network, introduced an STDCM to extract deep feature information with a scalable receptive field and multiscale. As illustrated in Fig. 2, the STDCM structure employs a step size of two, where M is the input channel dimension, and N is the output channel dimension. The kernel sizes of all blocks are set to three, except for the first block, which is set to one. This is achieved by progressively reducing the dimension of the feature map and using the method of feature representation through their aggregation. This not only removes the redundant terms in the structure but also increases the effective receptive field of the network.

Proposed method

In this section, an overview of MAFNet is provided first, followed by the introduction of three modules of MAFNet: SGSDM, MSAPPM, and BFM.

Overall network architecture

MAFNet is designed in this study, aims to overcome common challenges in semantic segmentation. Previous studies such as BiSeNet [22], BiSeNetV2 [23], DDRNet [17] and RTFormer [24], have demonstrated the effectiveness of the dual-branch network structure for real-time semantic segmentation. Therefore, inspired by the dual-branch network structure, MAFNet is constructed with semantic and spatial detail branches, as shown in Fig. 3a. The efficient backbone network STDC1 [12] serves as the semantic branch for extracting semantic information. STDC1 consists of five stages (from stage 1 to 5), with each stage using a downsampling rate of step 2 to halve the input resolution. The spatial detail branch, is built with the lightweight RB [17], has a residual interior structure, as shown in Fig. 3b. If the input image is $F_{\mathrm{i}} \in \mathbb{R}^{H\times W}$, the final output feature map sizes for the semantic and spatial detail branches are $F_{\mathrm{s}} \in \mathbb{R}^{H/32\times W/32}$ and $F_{\mathrm{sd}} \in \mathbb{R}^{H/8\times W/8}$, respectively. The complete implementation process of MAFNet is described in detail below.

The input image undergoes processing through the semantic branch, which extracts overall semantic information. When the semantic branch downsamples to 1/8 of the input image, a spatial details branch is introduced to assist in extracting spatial details. At this stage, the SGSDM is introduced to facilitate better information exchange between the two branches. SGSDM functions to extract semantic information while guiding and preserving crucial spatial details. The innovation of this module lies in its ability to effectively promote interaction between global and local information, providing robust support for subsequent processing stages. Following this, immediately after the semantic branch, the MSAPPM is introduced to compensate for the insufficient contextual information extraction capability of the semantic branch. In this processing flow, MSAPPM combines dilated convolutions with different dilation rates and a feature pyramid pooling structure to expand the network's receptive field and better aggregate contextual information. The introduction of MSAPPM helps the network possess a more powerful contextual awareness when processing semantic information. In the next stage of the network, the BFM is connected to better fuse features generated by the two branches. BFM enhances the interaction between the two branches through a cross-fusion mechanism, improving the network's expressive capability. The inclusion of this module enables the network to more effectively integrate feature information from different branches. Finally, a segmentation head is connected before the network's output, as shown in Fig. 3c, predicting the original image size after 8× upsampling.

In Fig. 3a, the structure of MAFNet reveals a synergistic collaboration among various modules at different stages, forming an efficient information processing workflow. This detailed processing flow ensures each innovative module corresponds to specific processing stages, enhancing the overall performance of the network. Table 1 presents the detailed parameters of the proposed network, where Conv2d represents the Conv–BN–ReLU operation, OPR indicates the operation, and K is the kernel size. S, R, and C denote the stride, repeat times, and output channels, respectively. The basic module of Stages 3, 4, and 5 is the STDC module.

Semantics guide spatial-details module (SGSDM)

Semantic and spatial detail information both play pivotal roles in image understanding. However, a previous method [25] has employed convolution, followed by cross-fusion or direct addition operations to improve performance. While this method encourages complementary information exchange between the semantic and spatial detail branches, the fusion process introduces information conflicts, especially when semantic differences or scale mismatches between features exist. Given that semantic information aids in extracting spatial details, it is crucial to employ a carefully designed fusion approach to enable the model to possess both semantic understanding and detail perception capabilities simultaneously.

SGSDM is introduced for MAFNet to address these challenges, as shown in Fig. 4. This module achieves accurate boundary extraction and fine-grained classification, providing semantic-based feature representation and supporting spatial analysis and decision boundaries. Here, S represents the sum function, wherein the multiplication of two elements is performed first, followed by the sum operation, and UP denotes the upsampling operation. The sum function can be

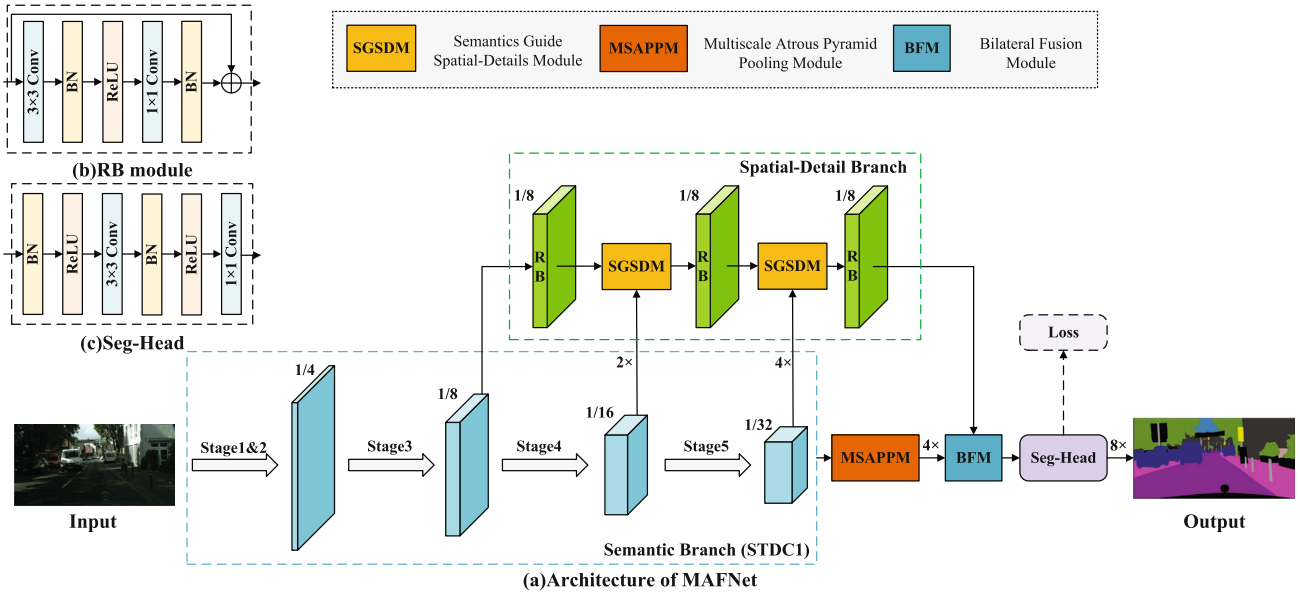


Fig. 3 Overview of the designed MAFNet. **a** Architecture of MAFNet, **b** RB module and **c** Seg-Head

Table 1 Detailed structure of MAFNet

Stage	Output	Semantic branch					Spatial-details branch				
		OPR	K	S	R	C	OPR	K	S	R	C
Input	512 × 1024					3					
S_1	256 × 512	Conv2d	3	2	1	32					
S_2	128 × 256	Conv2d	3	2	1	64					
S_3	64 × 128	Stage3		2	1	256	Conv2d	3	1	2	256
				1	1						
S_4	32 × 64	Stage4		2	1	512	Conv2d	3	1	2	512
				1	1						
S_5	16 × 32	Stage5		2	1	1024	Conv2d	3	1	1	1024
				1	1						

expressed by the following mathematical expression:

$$F_{\text{sum}} = \sum_{i=1}^{C} \mathbb{R}^{B\times C\times H\times W}, \tag{1}$$

where $F_{\text{sum}} \in \mathbb{R}^{B\times H\times W}$, B, C, H and W represent batch size, number of channels, image height, and image width, respectively.

The following steps are taken for the input feature F_{s} of one semantic branch and the input feature F_{sd} of the spatial detail branch. First, F_{s} undergoes a 1 × 1 Conv and batch normalization for channel dimension reduction, followed by the upsampling operation, resulting in features denoted as F_{up1}. To fuse with the spatial detail features, depthwise separable Conv and batch normalization with 3 × 3, and upsampling operations are applied. The resulting feature map is denoted as F_{up2}. For the spatial detail branch, the input feature F_{sd} undergoes a 3 × 3 Conv and batch normalization, and the resulting features are multiplied with F_{up2}. The Sum function is then employed to sum the resulting feature maps, generating a new feature map, F_{sum}. The above process can be expressed as the following equation:

$$\begin{aligned} F_{\text{up1}} &= \text{UP}(C_{1\times1}(F_{\text{s}})), \\ F_{\text{up2}} &= \text{UP}\left(DW_{3\times3}\left(C_{1\times1}(F_{\text{s}})\right)\right), \\ F_{\text{sum}} &= \text{Sum}(C_{1\times1}(F_{\text{sd}}) \otimes F_{\text{up2}}), \end{aligned} \tag{2}$$

where $C_{1\times1}$ is the 1 × 1 Conv and batch normalization, and $DWC_{3\times3}$ is the 3 × 3 depthwise separable convolution and batch normalization. Up indicates the upsampling by bilinear interpolation, and Sum is the operation that multiplies the two elements and then sums them up.

Instructive weights are generated by the Sigmoid function [26] and multiplied and added to the corresponding

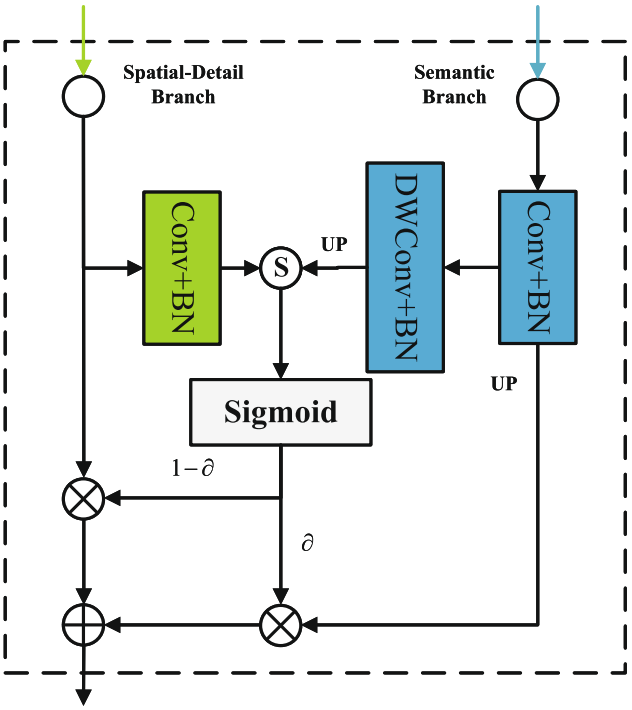


Fig. 4 Overview of SGSDM

branches. The mathematical expression for the Sigmoid activation function is given by the following formula:

$$y = \frac{1}{1 + e^{-x}}. \tag{3}$$

Finally, after a ReLU operation, the output is input to the next stage. Throughout the process, the input to the spatial detail branch remains 1/8 the size of the original image. This complete implementation process not only achieves accurate boundary extraction and fine-grained classification, but also provides semantic-based feature representation to further support spatial analysis and decision boundaries. The output F_{out} of SGSDM can be expressed by the following mathematical expression:

$$\begin{aligned} \partial &= \text{Sigmoid}(F_{\text{sum}}) \\ F_{\text{out}} &= R\left(\left(\partial \otimes F_{\text{up1}}\right) \oplus \left((1-\partial) \otimes F_{\text{sd}}\right)\right), \end{aligned} \tag{4}$$

where Sigmoid denotes the operation by the Sigmoid function, and R represents the ReLU activation function.

Multiscale atrous pyramid pooling module (MSAPPM)

The context aggregation module is commonly used in the field of image processing and computer vision. It provides a broader semantic context, allowing the model to better understand the semantic relationships in the image. Additionally, by expanding the receptive field of the model can be expanded to consider a wider range of contextual information, which is important for processing images with multiscale structure and semantic associations. This expansion of the receptive field enables the model to better understand the context in the image, leading to more accurate pixel classification and segmentation.

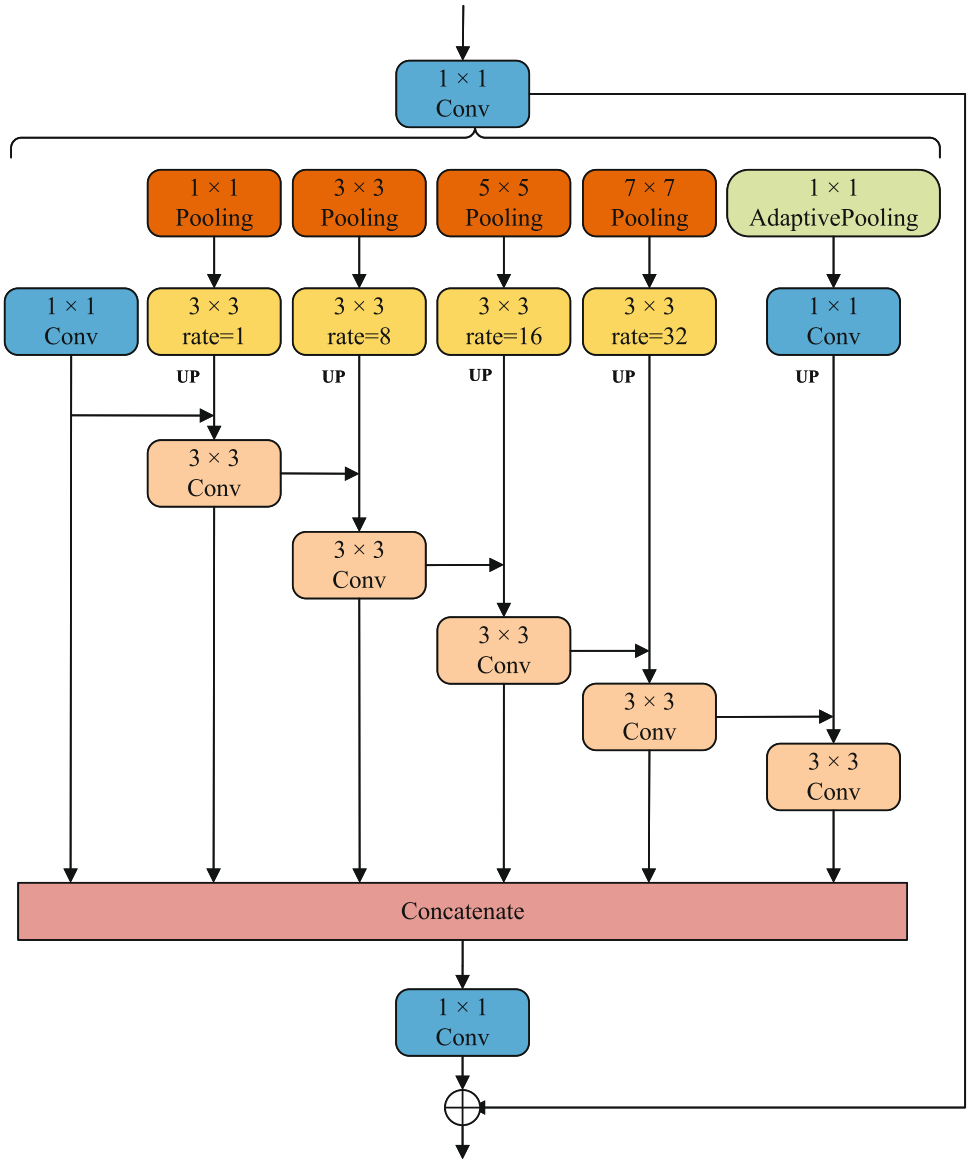
To accurately achieve contextual semantic information in images, this study drew inspiration from the context aggregation proposed in [19] and designed MSAPPM, as shown in Fig. 5. In this module, Pooling, UP, and Concatenate represents the pooling, upsampling, and splicing operation, respectively. Specifically, for an input feature map F_{m}, the dimensionality is first reduced by a 1×1 Conv and batch normalization. Subsequently, a 1×1 Conv and batch normalization is applied from the left side, and the resulting feature maps are used for subsequent information passing. In parallel, a set of $\{1 \times 1, 3 \times 3, 5 \times 5, 7 \times 7\}$ pooling operations and dilation convolutions with dilation rates $r = \{1, 8, 16, 32\}$ are followed by upsampling to restore the original size for subsequent concatenation operations. Additionally, global average pooling, 1×1 Conv, and batch normalization are concatenated in parallel to better capture global context information. After passing through the parallel connections, starting from the left side, the result is added to the next branch to form a continuous forward flow of information. Simultaneously, a 3×3 Conv and batch normalization is added after the addition to construct a hierarchical residual connection for better integration of multiscale global context information, and the output of each branch is denoted as F_{i}. Therefore, for an input feature F_x of the MSAPPM parallel branch, the output feature F_{i} of each branch can be expressed by the following mathematical expression:

$$F_{\text{i}} = \begin{cases} C_{1\times1}\left(F_x\right); i = 1 \\ C_{3\times3}\left(\text{UP}\left(DC_{3\times3,r=2^{0,3,4,5}}\left(H_{\text{pooling},j}\left(F_x\right)\right)\right)\right. \\ \left.\oplus F_{i-1}\right); 1 < i < n,\ j = 1, 3, 5, 7 \\ C_{3\times3}\left(\text{UP}\left(C_{1\times1}\left(H_{\text{gpooling}}\left(F_x\right)\right)\right) \oplus F_{i-1}\right); i = n, \end{cases} \tag{5}$$

where $C_{3\times3}$ represents the 3×3 Conv and batch normalization. $DC_{3\times3,r=2^{0,3,4,5}}$ is the 3×3 dilated convolution and batch normalization with dilation rate r. $H_{\text{pooling},j}$ represents the size of the convolution kernel and is used to represent the Pooling operation with $k \times k$ convolution kernel size, and H_{gpooling} indicates a global average pooling operation of size 1×1.

Finally, the features generated above are concatenated, denoted as F_{cat}, for further downstream propagation. The concatenated feature map is connected to a 1×1 Conv and

Fig. 5 Overview of MSAPPM



batch normalization, and a residual connection is performed with the feature map F_x that enters the parallel branch. After these operations, the output feature of MSAPPM denoted as F_{out}, is obtained. This process can be expressed through the following mathematical expression:

$$\begin{aligned} F_{\text{cat}} &= \text{Cat}\left(F_{\text{i}}\right) \\ F_{\text{out}} &= C_{1\times1}\left(F_{\text{cat}}\right) \oplus F_x, \end{aligned} \tag{6}$$

where Cat is the concatenation of the output features of each branch F_{i} ground operation. The complete implementation process of MSAPPM described above effectively expands the network's receptive field, and rich contextual information can be aggregated.

Bilateral fusion module (BFM)

Semantic and spatial detail information provide different types of information at different levels. Semantic information mainly studies the semantic meaning and relationship of objects, while spatial detail information provides detailed information such as the location, shape, and size of objects. Direct summation of these two types of information may lead to information mixing or neglect because semantic and spatial detail information differ in importance. Therefore, assigning accurate weights for the direct summation is challenging, resulting in the excessive prominence or neglect of certain information.

To effectively merge these two types of information, the BFM is designed to fuse the feature information learned by the dual-branch. The feature information of the dual-branch

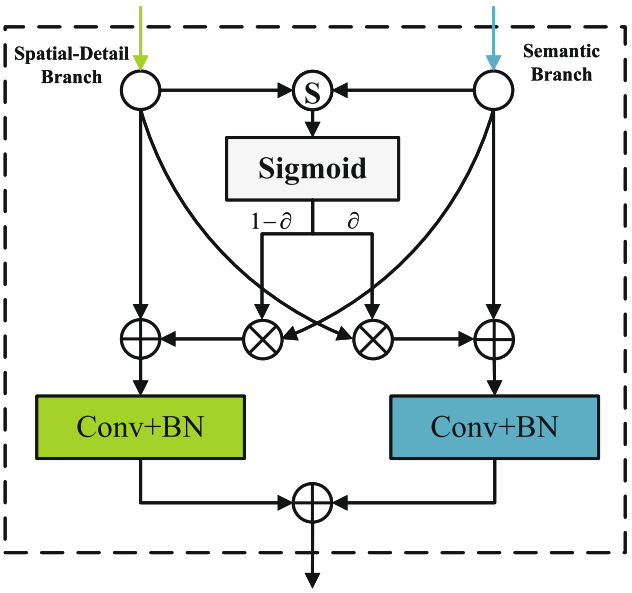


Fig. 6 Overview of BFM

is calculated to generate weights to control the weight relationship of the different branches. Thus, the importance of semantic and spatial detail information in fusion can be flexibly adjusted. Figure 6 shows an overview of BFM. For the input feature F_s of one semantic branch and the input feature F_{sd} of the spatial detail branch, the following steps are taken. First, F_s and F_{sd} are multiplied, and the resulting feature map is then summed using the sum function, denoted as F_{sum}. Subsequently, F_{sum} undergoes a Sigmoid operation to generate weights for the corresponding branches. These weights are multiplied by the features of the respective branches and added to the other branch. This method eliminates some redundant operations found in other aggregation modules and, through careful design, preserves the cross-fusion approach. Meanwhile, a 3×3 Conv and batch normalization are connected after each branch. Finally, the feature maps generated by the two branches are added to obtain the output of BFM, referred to as F_{out}. Therefore, the output F_{sd} of BFM can be expressed by the following mathematical expression:

$$\begin{aligned} F_{sum} &= \mathrm{Sum}\left(F_s \otimes F_{sd}\right) \\ \partial &= \mathrm{Sigmoid}\left(F_{sum}\right) \\ F_{out} &= C_{1\times1}\left(\left((1-\partial)\otimes F_s\right)\oplus F_{sd}\right) \\ &\quad \oplus C_{1\times1}\left(\left(\partial \otimes F_{sd}\right)\oplus F_s\right), \end{aligned} \tag{7}$$

where Sigmoid denotes the Sigmoid function operation. The complete implementation process of BFM described above effectively enhances the interaction of information between the two branches, thereby improving the network's expressive capability.

Experiments

To validate the performance of the designed method, extensive training and ablation experiments are conducted in this section. First, the dataset used in this study is introduced, and the experimental setup is described in detail. Next, ablation experiments performed to evaluate the performance impact of the key components of the method designed are discussed. Finally, MAFNet is compared with state-of-the-art methods to demonstrate its generalization ability through qualitative analysis.

Datasets

Cityscapes [1] This dataset is widely used for evaluating the performance of semantic segmentation algorithms because it provides rich annotated information and high-quality images. It is composed of keyframes extracted from complete video sequences of various cities [27]. It contains about 5000 high-resolution street view images from 50 cities, each sized at 1024×2048, covering a variety of different urban scenes such as downtown, residential areas, industrial areas, and parks. The images are all manually annotated with 19 common semantic segmentation categories, including vehicles, buildings, roads, and pedestrians. Additionally, the training, validation, and test sets consist of 2975, 500, and 1525 images, respectively.

CamVid [28] This dataset captures images from cars driving on roads in Cambridge, UK, making it highly representative for studying urban traffic scene segmentation. Despite its smaller size, CamVid is a popular choice for testing real-time semantic segmentation models. The dataset contains about 500 images with a resolution of 720×960, each manually annotated with 11 common categories, including roads, buildings, cars, and pedestrians. The training, validation, and test sets comprise 367, 101, and 233 images, respectively.

Implementation details and evaluation

Training settings In the experiment, the training settings closely align with [29]. On Cityscapes, stochastic gradient descent (SGD) with a momentum of 0.9, an initial learning rate of 0.05, and a weight decay of 0.0001 is utilized. This configuration is chosen to facilitate rapid convergence in the early stages of training while mitigating oscillations throughout the training process. The "poly" learning rate scheduler is implemented, along with a linear warmup [30] from 0.1 LR to LR for the first 3000 iterations. During training, random horizontal flipping, random scaling within the range [400, 1600], and random cropping of 512×1024 are employed, as a specific image cropping method for segmentation [31] yields better results. A reduced set of RandAug [32] opera-

tions are used, including auto contrast, equalize, rotate, color, contrast, brightness, and sharpness. Due to the critical importance of batch size selection for the overall performance of the model [33], therefore, the cross-entropy loss is employed with a batch size of 8. In MAFNet, the backbone network is trained using weights pre-trained on the ImageNet [34] dataset, while other components are trained using the default initialization weights provided by PyTorch [35]. The training lasted for 500 epochs on a single A100 GPU, employing mixed precision training for accelerated convergence without sacrificing accuracy. For the test server submission, training is conducted on the trainval set, and online hard example mining (OHEM) [36] loss is used additionally.

On the CamVid, the training setup resembles Cityscapes, but Cityscapes pre-trained model is employed due to the smaller dataset size. Random horizontal flipping, random scaling within the range [288, 1152], and random cropping of 720 × 960 with a batch size of 12 are applied. RandAug and class uniform sampling are excluded, and the training span 200 epochs.

Inference settings During the inference stage, the model is applied to maintain the original size of Cityscapes (1024 × 2048) and CamVid (720 × 960). Inference speed is measured on a single A100 GPU, utilizing PyTorch 1.10, CUDA 11.3, cuDNN 8.0, TenserRT 8.4.0, and the Anaconda environment. FP32 data accuracy is employed, batch size is set to 1, and the inference speed protocol proposed in [37] is used for measurement.

Evaluation The study utilizes a set of standard performance evaluation indicators tailored to the semantic segmentation task and specific research goals. The chosen metrics include IoU, mIoU, along with considerations for computational efficiency such as GFLOPs and model parameters (Params). The specific calculation method of IoU is as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{8}$$

and the specific calculation method of mIoU is as follows:

$$\mathrm{mIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}, \tag{9}$$

where FN represents the number of negative categories with incorrect prediction results, FP denotes the number of positive categories with incorrect prediction results, and TP is the number of positive categories with incorrect prediction results [38].

The rationale behind the selection of IoU and mIoU is rooted in their effectiveness for quantifying pixel-wise segmentation accuracy. A higher IoU and mIoU value signify superior segmentation performance, indicating the model's ability to accurately delineate semantic boundaries and capture intricate details. Additionally, the inclusion of GFLOPs and model parameters serves to assess the computational complexity of the proposed model. A higher value in GFLOPs and Params suggests increased model complexity, indicating potential resource requirements during deployment.

Table 2 Ablation study of backbone

Backbone	Params (M)	GFLOPs	mIoU (%)	FPS
ResNet18	14.32	31.75	77.77	60.9
STDC1	12.00	22.30	78.54	70.9
STDC2	16.03	28.91	78.85	64.9

Ablation study

In this subsection, ablation experiments on the Cityscapes dataset are conducted to explore the effectiveness of the designed module.

Ablation of backbone

To validate the effectiveness of the STDC1 [12] backbone network, ResNet18 [5] and STDC2 [12] methods are chosen as comparative benchmarks, and the results are presented in Table 2. To ensure a fair comparison, the same training settings and complexity calculation methods are employed in constructing the three backbone network models. Employing STDC1 as the backbone network an mIoU of 78.54%, whereas replacing it with ResNet18 results in an mIoU of 77.77%. In comparison, the proposed method demonstrates a 0.77% increase in mIoU, accompanied by a 1.4-fold reduction in computational complexity and an improvement in inference speed. Notably, even when using STDC2 to construct larger models, an mIoU of 78.85% is achieved. However, its computational complexity increases by 1.3 times compared to STDC1. Hence, after conducting a thorough analysis of backbone ablation studies and taking into account factors such as segmentation accuracy, computational complexity, and inference speed, the STDC1 backbone network is chosen for building the designed MAFNet, demonstrating its superior overall performance.

Ablation of SGSDM

The impact of spatial detail branches on the overall performance of the network is explored, as shown in the first part of Table 3. The first row, displaying the segmentation accuracy and computational complexity obtains solely with the main backbone network, achieves an mIoU of 72.06%. The

Table 3 Ablation study of SGSDM

Semantic	Spatial-detail	Add	Concat	SGSDM	OHEM	mIoU (%)	GFLOPs
√						72.06	9.25
√	√					74.68	21.06
√	√	√				78.00	22.14
√	√		√			78.11	23.96
√	√			√		78.54	22.30
√	√			√	√	78.73	22.30

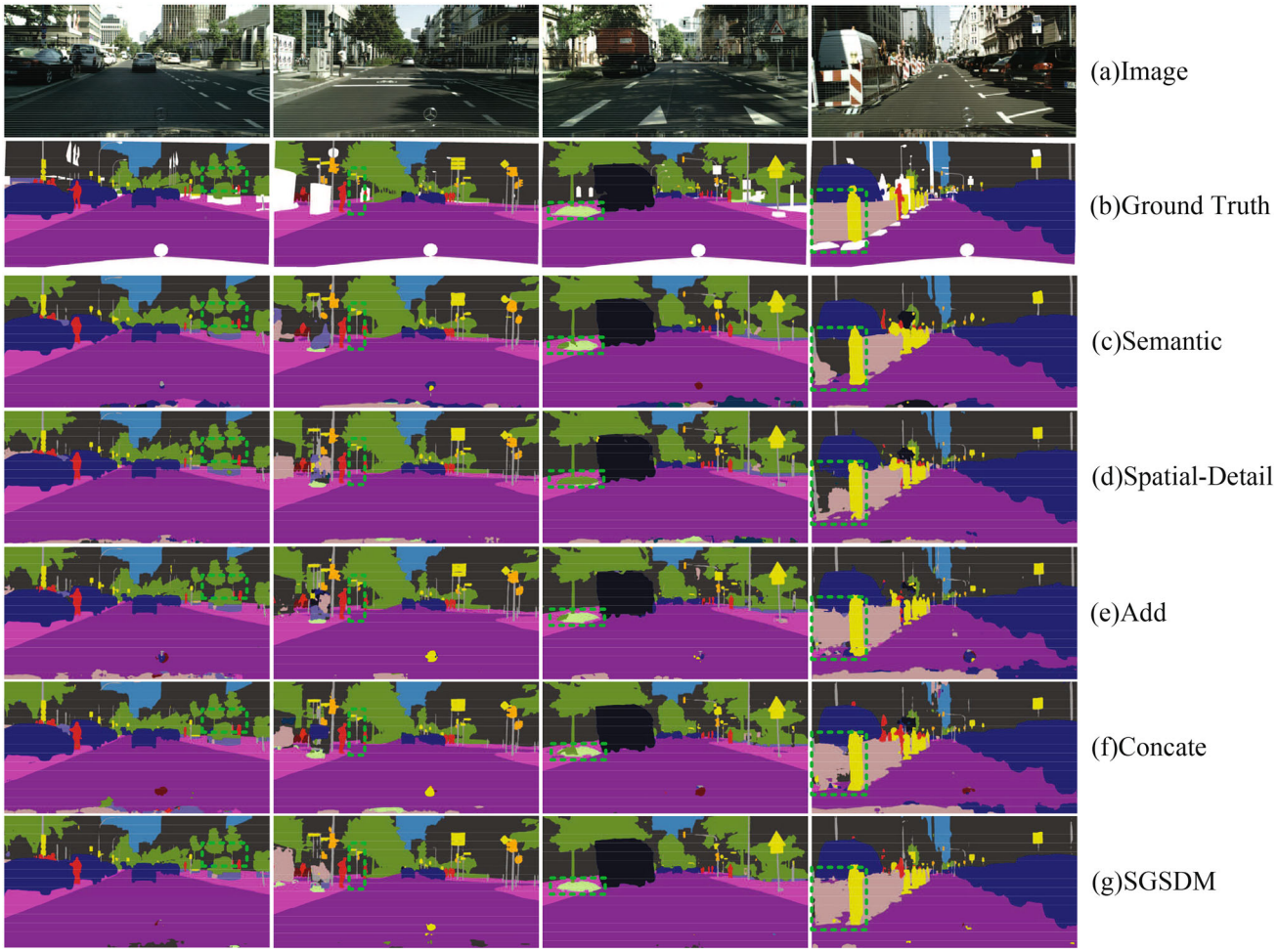


Fig. 7 Comparison of SGSDM visualization on the Cityscapes dataset

introduction of the spatial detail branch elevates the segmentation accuracy to 74.68% mIoU. Since the semantic branch focuses solely on high-level semantic information and lacks the ability to extract low-level spatial detail information, the addition of the spatial detail branch results in a significant improvement of 2.62% mIoU in segmentation accuracy.

Following that, the effectiveness of SGSDM is validated. Initially, the network is modified by removing the SGSDM to be verified and replacing it with a simple addition operation, yielding an mIoU of 78.00%. Furthermore, when replaced with a concatenation operation, the method achieves an mIoU of 78.11%. Finally, using the designed SGSDM method in this study results in an mIoU of 78.54%. The results indicate that employing SGSDM better establishes a bridge for information transfer between dual branches, improving segmentation accuracy by 0.54% mIoU compared to the simple addition method, with only a marginal increase in computational complexity. Compared to the concatenation method, there is a 0.43% mIoU improvement and a reduction of 1.66 GFLOs in computational complexity. Figure 7 provides a visual comparison of SGSDM on the Cityscapes dataset. In Fig. 7a, the original RGB image is displayed, Fig. 7b shows

Table 4 Performance of MSAPPM with different dilation rate strategies

Strategy	Dilation rates	mIoU (%)	FPS
s0	(1,1,1,1)	77.69	72.1
s1	(1,1,2,2)	78.38	70.0
s2	(1,1,2,3)	77.81	68.3
s3	(1,1,3,3)	77.77	70.1
s4	(1,2,3,4)	78.05	70.3
s5	(1,2,4,8)	77.95	69.9
s6	(1,3,5,7)	78.14	69.7
s7	(1,4,8,16)	78.23	70.3
s8	(1,6,10,14)	77.95	70.2
s9	(1,8,16,32)	78.54	70.9
s10	(1,16,32,64)	78.37	70.3

the true labels of the original image, and Fig. 7c–g present the predictions of semantic, spatial-detail, add, concatenate, and the SGSDM method proposed in this study, respectively. In Fig. 7, the third column highlights SGSDM's superior accuracy in predicting the terrain category. The experiments and visualized results unequivocally affirm SGSDM's effectiveness, demonstrating that leveraging advanced semantic information significantly enhances the comprehension of low-level spatial details. Furthermore, SGSDM contributes a semantic-based feature representation to the spatial detail branch, reinforcing spatial analysis and decision boundaries.

Ablation of MSAPPM with different dilation rates

To determine a set of optimal dilation rate parameters for the MSAPPM, MSAPPMs with different dilation rates are constructed, and the experimental results are presented in Table 4.

A total of 11 networks with different dilation rates are constructed, divided into two intervals. Here, s0 is the baseline network, where dilated convolution is replaced with ordinary convolution to build MSAPPM. s1–s6 are MSAPPM with a smaller dilation rate, while s7–s10 are MSAPPM with a larger dilation rate. The experimental results show that s1 obtains the highest mIoU when using a smaller dilation rate, reaching 78.38%, which proves that using different dilation rates impacts the overall network. Subsequently, MSAPPM with a large dilation rate are evaluated, and s9 obtains the highest mIoU with 78.54%. These results show that the accuracy of the entire network may be improved as the dilation rate increases, but the increase in the dilation rate is not infinite. For example, when the dilation rate is set to s10, the mIoU is reduced by 0.17% compared to s9. Importantly, the dilation rate cannot be set too large due to the special structure of dilated convolutions. If the dilation rate is too large, some input image features will not be fully utilized. Therefore, while ensuring the performance of semantic segmentation, the dilated convolution with an appropriate dilation rate should be selected, and it is necessary to avoid setting too large dilation rates to prevent information loss. In this study, s9 is finally selected as the dilation rate parameter of MSAPPM.

Ablation and efficiency analysis of MSAPPM

After determining the internal structure of MSAPPM, an ablation study on the entire MSAPPM is conducted, and comparative experiments are performed using advanced context aggregation methods to demonstrate the effectiveness of MSAPPM in aggregating rich contextual information. The experimental results are shown in Table 5.

(1) Ablation analysis: Initially, the baseline network is designed without any context aggregation module, where the output of the last stage of the backbone network directly fuses with the output of the spatial detail branch. From Table 5, it is observed that the addition of the designed MSAPPM improves the mIoU from 75.93 to 78.54% for the baseline network, resulting in a 2.61% increase in mIoU while only slightly elevating computational complexity and maintaining inference speed.
(2) Efficiency analysis: MSAPPM is also compared with similar methods, such as PPM [40], PAPPM [41], and DAPPM [17]. Compared to the baseline network, the mIoU increases by 0.76%, 0.88%, and 1.13%, respectively. Further analysis of results from the ablation experiments reveals that MSAPPM outperforms these advanced context extraction modules, enhancing segmentation accuracy while maintaining high computational efficiency. Figure 8 illustrates the effect of MSAPPM on aggregating context information, with the first, second, and third columns focusing on the motorcycle, truck, and bus classes, respectively. Figure 8c illustrates that PPM has the least satisfactory prediction performance, while

Table 5 Ablation study of MSAPPM

Baseline	PPM	PAPPM	DAPPM	MSAPPM	mIoU (%)	GFLOPs
√					75.93	21.49
√	√				76.69	21.93
√		√			76.81	21.46
√			√		77.06	21.93
√				√	78.54	22.30

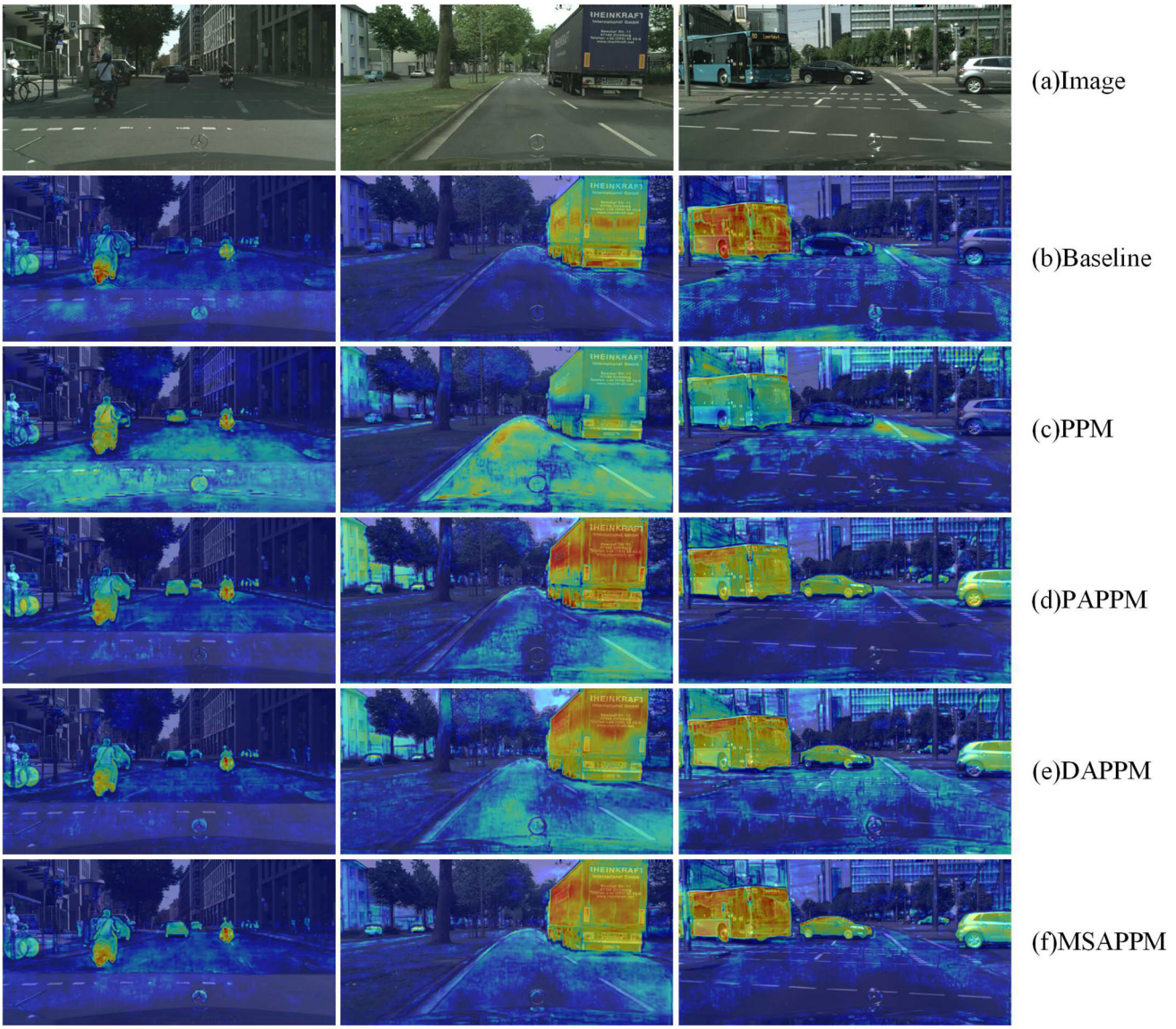


Fig. 8 A visual interpretation of the different context extraction modules is shown using the Grad-CAM [39] method. Here, **a** original image, **b** baseline, **c** PPM, **d** PAPPM, **e** DAPPM, and **f** MSAPPM

Table 6 Ablation study of BFM

Method	mIoU (%)	GFLOPs
Baseline	77.83	22.02
+Concat	78.02	23.23
+FFM	78.26	22.30
+BFM	78.54	22.30

MSAPPM demonstrates a more balanced focus on the regions of interest for the three mentioned categories. This affirms the module's enhanced ability to perceive global context effectively. The experiments above indicate that the designed MSAPPM in this study is an effective method for context aggregation, allowing for better exploration of latent contextual information and thereby improving the model's performance.

Ablation of BFM

In this subsection, an ablation study focusing on validating BFM is conducted, emphasizing the crucial role of effectively balancing semantic and spatial detail information generated by the dual-branch network. Initially, features generated by the two branches are fused using a simple addition operation, serving as a baseline network, and are compared to BFM. To enhance the credibility of BFM, splicing operations are also chosen for comparison. The experimental results are presented in Table 6, where baseline denotes an addition operation, Concat involves a concatenation operation, and FFM [12] represents the feature fusion module.

Table 6 reveals segmentation accuracies of 77.83%, 78.02%, 78.26%, and 78.54% mIoU for baseline, Concat, FFM, and BFM methods, respectively. The proposed BFM method has a 0.71% mIoU improvement compared to the baseline. The BFM has a 0.52% mIoU improvement compared to the concate method, with computational complexity largely comparable between these two methods. Lastly, compared to the FFM method, BFM demonstrates a 0.28% mIoU improvement while maintaining identical computational complexity. Figure 9 illustrates the visual comparison results of BFM on the Cityscapes dataset: Fig. 9a shows the original RGB image; Fig. 9b shows the true labels of the original image; Fig. 9c–f present the predictions of baseline, Concat, FFM, and the proposed BFM method, respectively. From the second column in Fig. 9, BFM effectively captures details within the truck category and accurately predicts spatial positional information. The experiments above indicate that the cross-fusion approach, achieved by separately calculating weights generated by the dual-branch, not only better balances the weight relationship but also efficiently integrates information from both branches. This demonstrates the effectiveness of the BFM module in enhancing detail capture and spatial prediction.

Comparison with state-of-the-art methods

In this section, MAFNet is compared with state-of-the-art methods. The results on the Cityscapes and CamVid datasets are given as follows.

Comparison on Cityscapes dataset

First, MAFNet is compared with nonreal-time segmentation algorithms such as DeepLab [42] and PSPNet [40], and real-time segmentation algorithms such as ICNet [11], ERFNet [43], DFANet-A [44], BiSeNet1 [22], BiSeNet2 [22], TD^4-Bise18 [45], LBN-AA [19], BiSeNetV2-L [23], SwiftNetRN-18 [46], HyperSeg-M [47], S^2-FPN18 [13], and STDC2-Seg75 [12]. Table 7 presents the segmentation accuracy and inference speed of the designed and state-of-the-art methods on the validation and test sets on Cityscapes.

It is evident from Table 7 that MAFNet exhibits strong overall performance, achieving a 78.5% mIoU on the validation set and 77.4% mIoU on the test set at an inference speed of 70.9 FPS. The proposed method outperforms the earlier DeepLab in terms of segmentation accuracy, parameter quantity, and computational complexity. Although PSPNet achieves an mIoU of 81.2% on the test set, it loses real-time performance, limiting its practical application.

When compared to real-time approaches, MAFNet outperforms STDC2-Seg75 by 1.7% mIoU on the validation set and 0.6% mIoU on the test set in terms of segmentation accuracy. Against S^2-FPN18, MAFNet achieves a 1.2% mIoU improvement on the test set. Compared to HyperSeg-M, MAFNet demonstrates superior performance with a 2.5% mIoU increase on the validation set, a 1.6% mIoU increase on the test set and a 34 FPS improvement in inference speed. Against SwiftNetRN-18, MAFNet achieves a 1.9% mIoU improvement on the test set and a 31.6 FPS increase in inference speed. Compared to BiSeNetV2-L, MAFNet surpasses by 2.9% mIoU on the validation set and 2.1% mIoU on the test set. Against LBN-AA, MAFNet outperforms by 3.8% mIoU on the test set. Furthermore, MAFNet demonstrates competitive performance against other real-time methods, boasting the highest segmentation accuracy on both the test and validation sets, coupled with reasonable parameter size, computational complexity, and inference speed. Figure 10 presents a more visual comparison of the results, where the red dashed line represents the real-time boundary, blue dots represent state-of-the-art segmentation algorithms, and red dots represent the proposed MAFNet.

Finally, Fig. 11 presents a color-annotated map of the segmentation on the Cityscapes dataset: Fig. 11a shows the original RGB image; Fig. 11b shows the Ground Truth of

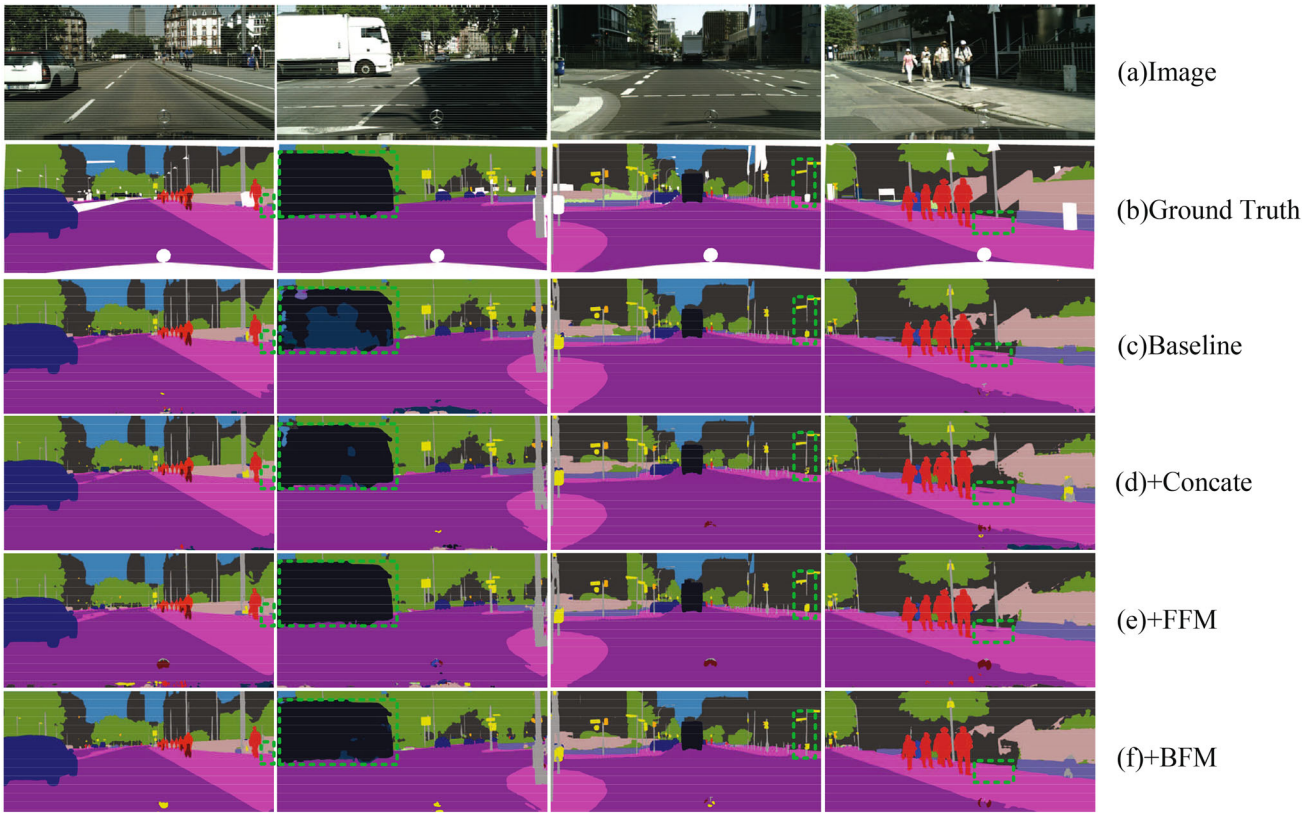


Fig. 9 Comparison of BFM visualization on the Cityscapes dataset

Table 7 Comparison with state-of-the-art methods on the Cityscapes dataset

Method	Backbone	Parameters (M)	GFLOPs	mIoU (%)		FPS
				Val	Test	
DeepLab [42]	VGG16	262.1	457.8	–	63.1	0.3
PSPNet [40]	ResNet101	250.8	412.2	–	81.2	0.8
ICNet [11]	PSPNet50	26.5	28.3	–	69.5	30.3
ERFNet [43]	No	2.1	27.7	–	69.7	41.7
DFANet-A [44]	Xception A	7.8	3.4	–	71.3	100.0
BiSeNet1 [22]	Xception39	5.8	14.8	69.0	68.4	105.8
BiSeNet2 [22]	ResNet18	49.0	54.0	74.5	74.7	62.1
TD^4-Bise18 [45]	BiseNet18	–	–	75.0	74.9	47.6
LBN-AA [19]	MobileNetV2	6.2	49.5	–	73.6	51.0
BiSeNetV2-L [23]	No	5.2	118.5	75.8	75.3	47.3
SwiftNetRN-18 [46]	ResNet18	11.8	104.0	75.4	75.5	39.3
HyperSeg-M [47]	EfficientNet-B1	10.1	7.5	76.2	75.8	36.9
S^2-FPN18 [13]	ResNet18	17.8	29.1	76.6	76.2	67.6
STDC2-Seg75 [12]	STDC2	22.2	54.9	77.0	76.8	97.0
MAFNet	STDC1	12.0	22.23	78.7	77.4	70.9

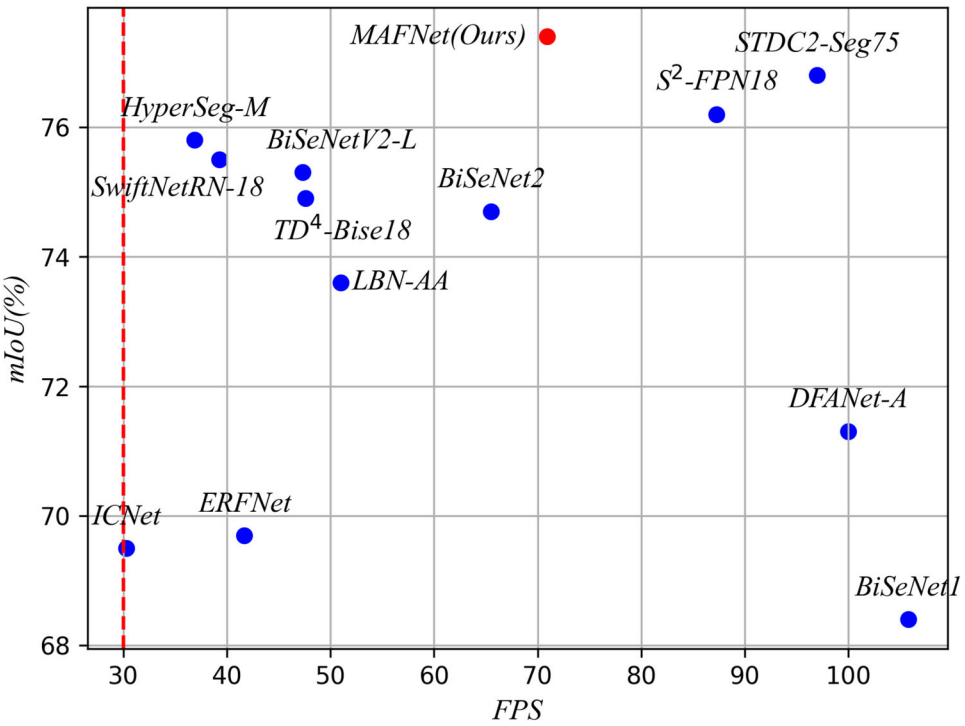


Fig. 10 Comparison of MAFNet with some state-of-the-art methods on the Cityscapes test dataset

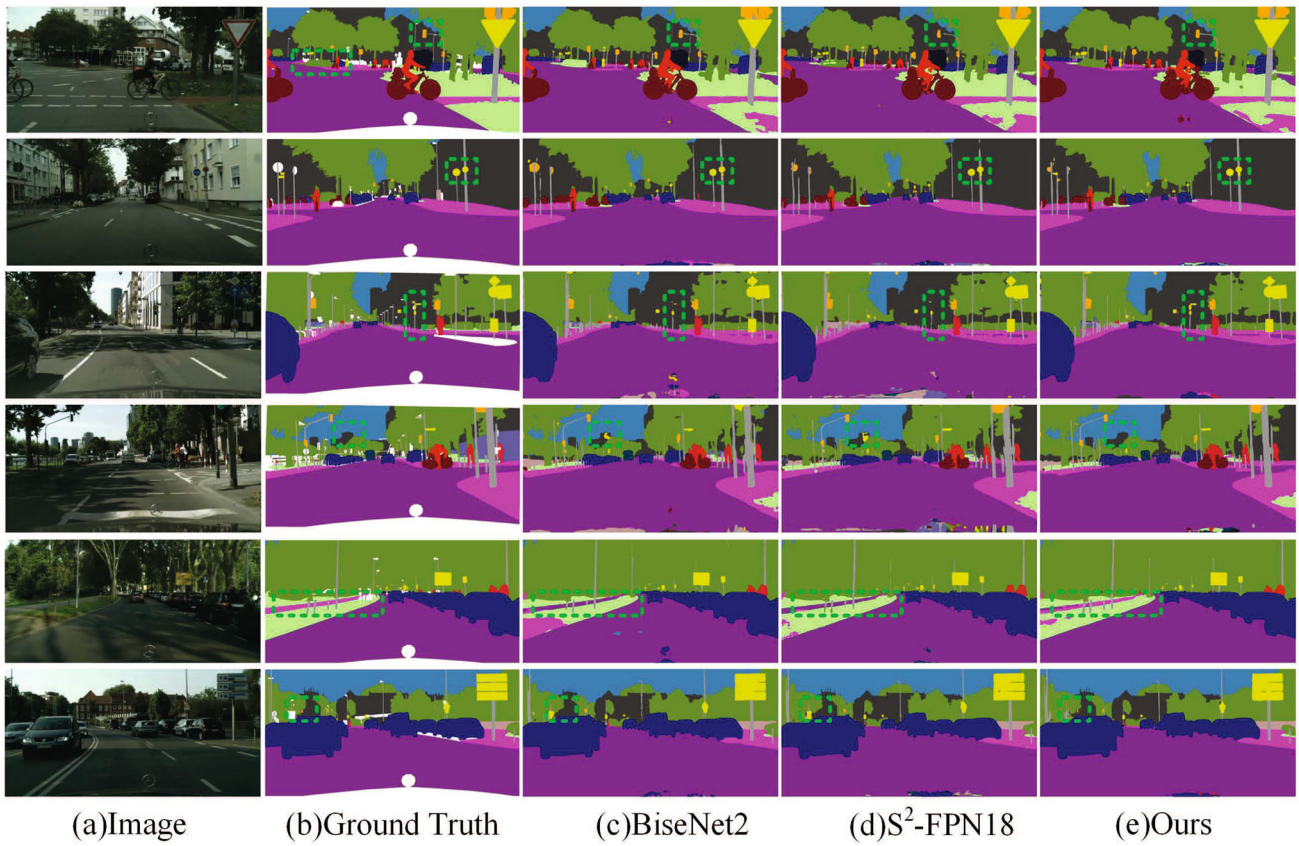


Fig. 11 Visualization results with different methods on the Cityscapes dataset

Table 8 Comparison with state-of-the-art methods on the CamVid dataset

Method	Extra data	mIoU (%)	FPS
LBN-AA [19]	–	68.0	39.3
BiSeNet2 [22]	IM	68.3	116.3
TD^4-Bise18 [45]	IM	72.6	25.0
SFNet [48]	IM	73.8	36
STDC2-Seg75 [12]	IM	73.9	97.0
S^2-FPN18 [13]	IM	68.7	124.2
S^2-FPN34 [13]	IM	71.0	107.2
S^2-FPN34M [13]	IM	73.9	55.5
VideoGCRF [49]	C	75.2	–
MSFNet [50]	IM	75.4	91.0
BiSeNetV2 [23]	C	76.7	124.5
MAFNet	C	77.6	192.5

the original image; Fig. 11c–e shows the prediction results of BiSeNet2, S^2-FPN18, and MAFNet, respectively. Green rectangles are used to highlight some regions with good segmentation. It is evident from Fig. 11 that MAFNet is better at recognizing traffic lights, traffic signs, trees, and sidewalks. In contrast, BiSeNet2 and S^2-FPN18 struggle to recognize the rough outline and even completely fail to recognize the above categories.

In conclusion, MAFNet exhibits excellent performance on the Cityscapes dataset, showcasing its superiority in key metrics such as real-time performance and segmentation accuracy compared to various real-time and non-real-time segmentation algorithms. Notably, MSAPPM plays a crucial role in extracting contextual information, contributing to MAFNet's outstanding performance. BFM effectively achieves the fusion of high-level semantic and low-level spatial details, further enhancing the network's overall segmentation performance. Lastly, SGSDM excels in precise boundary extraction and fine-grained classification, playing a supportive role in spatial analysis and decision boundary determination for the network.

Comparison on CamVid dataset

To further validate MAFNet, a comparison with state-of-the-art methods on the CamVid dataset is conducted, including LBN-AA [19], BiSeNet2 [22], TD^4-Bise18 [45], SFNet [48], STDC2-Seg75 [12], S^2-FPN18 [13], S^2-FPN34 [13], S^2-FPN34M [13], VideoGCRF [49], MSFNet [50], and BiSeNetV2 [23]. Table 8 presents the segmentation accuracy on the CamVid test dataset and the inference speed, where IM represents the ImageNet dataset and C represents the Cityscapes dataset. The original image size of 720×960 is maintained as input to the model during the training of the CamVid dataset. Additionally, the pre-trained model weights on Cityscapes are used.

Table 8 presents a clear overview of MAFNet's performance on the CamVid dataset, achieving a 77.6% mIoU at an inference speed of 192.5 FPS. In contrast, various state-of-the-art methods, such as LBN-AA, BiSeNet2, TD^4-Bise18, SFNet, STDC2-Seg75, S^2-FPN18, S^2-FPN34, S^2-FPN34M, VideoGCRF, MSFNet, and BiSeNetV2, achieved 68%, 68.3%, 72.6%, 73.8%, 73.9%, 68.7%, 71.0%, 73.9%, 75.2%, 75.4% and 76.7% mIoU, respectively. These results show

Table 9 IoU (%) for 11 classes on Camvid test set

Method	Bui	Tree	Sky	Car	Sig	Roa	Ped	Fen	Pol	Side	Bic	mIoU(%)
BiSeNet1 [22]	82.2	74.4	91.9	80.8	42.8	93.3	53.8	49.7	25.4	77.3	50.0	65.6
BiSeNet2 [22]	83.0	75.8	92.0	83.7	46.5	94.6	58.8	53.6	31.9	81.4	54.0	68.7
LBN-AA [19]	83.2	70.5	92.5	81.7	51.6	93.0	55.6	53.2	36.3	82.1	47.9	68.0
S^2-FPN18 [13]	83.0	77.2	91.8	88.9	48.2	95.7	56.4	43.4	32.4	84.8	62.5	69.6
S^2-FPN34 [13]	85.3	77.4	91.7	91.2	49.6	95.7	59.1	46.8	33.2	85.4	66.5	71.0
S^2-FPN34M [13]	86.0	78.8	92.6	**92.2**	56.2	96.0	67.1	47.3	42.1	86.8	70.7	74.2
MAFNet	**89.2**	**81.5**	**92.7**	89.2	**60.6**	**96.1**	**72.3**	**69.1**	**43.3**	**87.7**	**71.6**	**77.5**

The bolded values in columns 2–12 indicate the highest values of the class IoU (%) metric for each column, while the bolded value in the last column indicates the highest mIoU (%) metric among the compared methods. This approach highlights the superiority of our proposed method

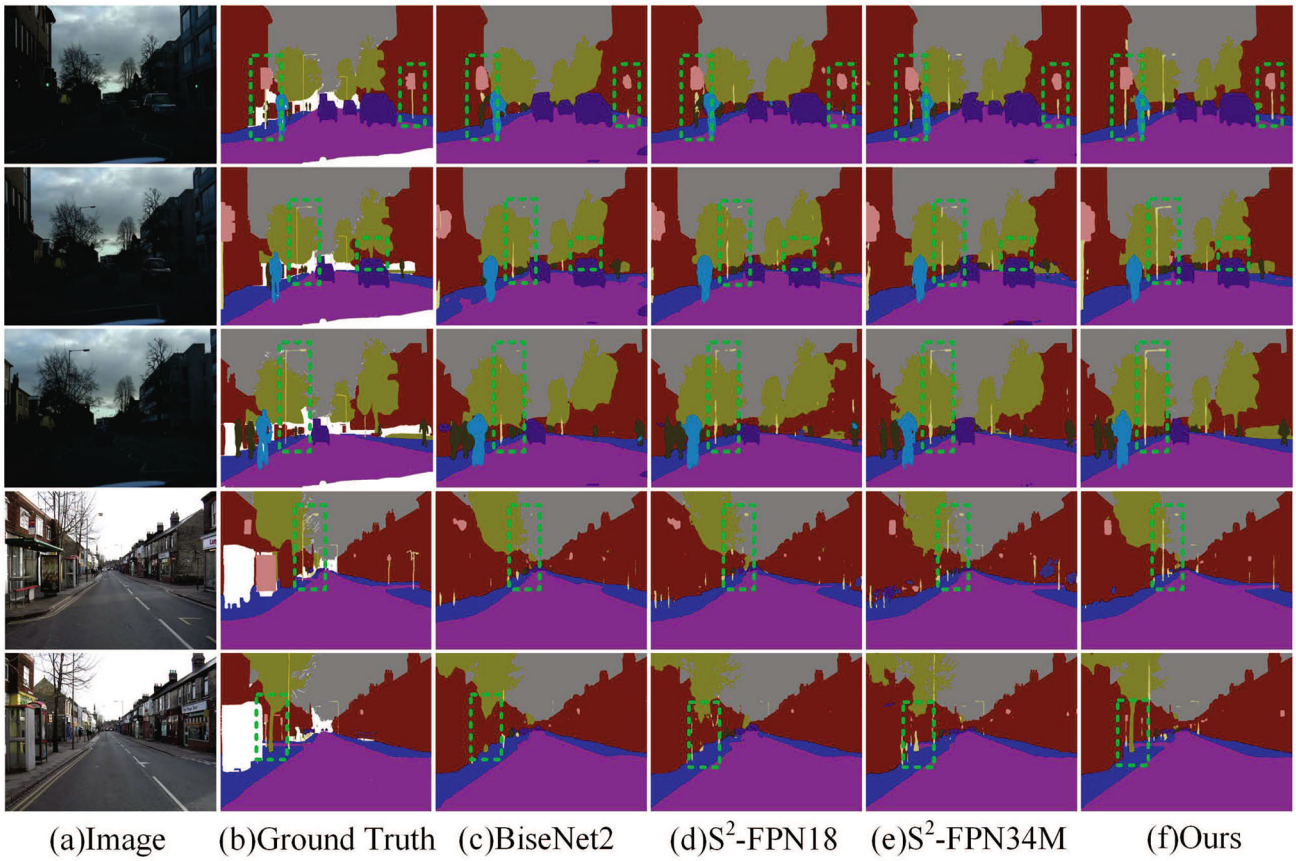


Fig. 12 Visualization results with different methods on the CamVid dataset

that MAFNet outperforms some state-of-the-art methods by 0.9–9.6% in terms of segmentation accuracy, maintaining competitive inference speed. In the Table 9, the superior accuracy of the proposed network is thoroughly validated by the fact that the IoU for 10 out of the 11 classes achieved the highest precision. This outcome underscores the effectiveness of the network in accurately capturing and delineating semantic classes, affirming its excellence in semantic segmentation tasks on the Camvid dataset.

Figure 12 presents the segmented results of MAFNet and three other methods (BiSeNet2, S^2-FPN18, and S^2-FPN34M) on the CamVid dataset. Eleven classes of objects are color-labeled for clarity, and a green dashed box highlights regions with superior segmentation. Only MAFNet can accurately recognize small objects such as poles, trees, pedestrians, cyclists, and traffic signs. BiSeNet2 shows the weakest segmentation effect, particularly struggling with the recognition of poles and trees. S^2-FPN18 performs better than BiSeNet2, identifying some object contours. S^2-FPN34M, a larger model, exhibits more distinct object features but struggles to clearly identify overall shape and position information. MAFNet stands out by accurately identifying object shapes and positions, highlighting the spatial detail branch's ability to retain location information. The comprehensive comparison of experimental results on the CamVid dataset confirms that MAFNet excels in balancing segmentation accuracy and inference speed, showcasing the superiority of the key modules designed in this study.

Conclusions

In this study, a real-time semantic segmentation network, MAFNet, is designed based on a dual-branch network structure to optimize the trade-off between accuracy and speed in real-time semantic segmentation. The design includes an SGSDM that not only enables accurate boundary extraction and fine-grained classification but also provides semantic-based feature representation to support spatial analysis and decision boundaries. Additionally, to address the limited receptive field problem of the proposed network, an MSAPPM is designed using an effective combination of dilated convolution and a feature pyramid pooling structure to aggregate rich contextual information. The BFM is designed to balance weight relationships between the dual branches, using feature information of the dual branch to generate weights for cross-fusion. Through extensive experiments and qualitative analysis, the effectiveness of the MAFNet method

is proved. Future work will focus on exploring more effective ways of information interaction between dual-branch networks, potentially incorporating attention mechanisms, and a more in-depth discussion on balancing accuracy and speed in real-time semantic segmentation networks.

Acknowledgements This work was supported by the National Natural Science Foundation of China [grant number 61602157], Henan Science and Technology Planning Program [grant number 202102210167]. The authors thank TopEdit (https://www.topeditsci.com) for its linguistic assistance during the preparation of this manuscript.

Funding The National Natural Science Foundation of China [Grant No: 61602157], Henan Science and Technology Planning Program [Grant No: 202102210167].

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code Availability The code are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Consent to participate All authors agreed to participate in this paper.

Consent for publication Not applicable.

Ethics approval Not applicable

References

1. Cordts M, Omran M, Ramos S, Rehfeld T, Enzweiler M, Benenson R, Franke U, Roth S, Schiele B (2016) The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3213–3223
2. Zhou Z, Rahman Siddiquee MM, Tajbakhsh N, Liang J (2018) Unet++: a nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support: 4th international workshop, DLMIA 2018, and 8th international workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, September 20, 2018, Proceedings 4, Springer, pp 3–11
3. 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
4. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer, pp 234–241
5. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 770–778
6. Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 801–818
7. Tao H, Zheng J, Wei J, Paszke W, Rogers E, Stojanovic V (2023) Repetitive process based indirect-type iterative learning control for batch processes with model uncertainty and input delay. J Process Control 132:103112
8. Song X, Wu N, Song S, Zhang Y, Stojanovic V (2023) Bipartite synchronization for cooperative-competitive neural networks with reaction-diffusion terms via dual event-triggered mechanism. Neurocomputing 550:126498
9. Peng Z, Song X, Song S, Stojanovic V (2023) Hysteresis quantified control for switched reaction-diffusion systems and its application. Complex Intell Syst 9(6):7451–7460
10. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: a deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147
11. Zhao H, Qi X, Shen X, Shi J, Jia J (2018) Icnet for real-time semantic segmentation on high-resolution images. In: Proceedings of the European conference on computer vision (ECCV), pp 405–420
12. Fan M, Lai S, Huang J, Wei X, Chai Z, Luo J, Wei X (2021) Rethinking bisenet for real-time semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9716–9725
13. Elhassan MA, Yang C, Huang C, Legesse Munea T, Hong X (2022) s^2-fpn: scale-ware strip attention guided feature pyramid network for real-time semantic segmentation
14. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H (2018) Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 552–568
15. Tan M, Le Q (2019) Efficientnet: rethinking model scaling for convolutional neural networks. In: International conference on machine learning, pp 6105–6114
16. Poudel RP, Liwicki S, Cipolla R (2019) Fast-scnn: fast semantic segmentation network. arXiv preprint arXiv:1902.04502
17. Hong Y, Pan H, Sun W, Jia Y (2021) Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv:2101.06085
18. Gao S-H, Cheng M-M, Zhao K, Zhang X-Y, Yang M-H, Torr P (2019) Res2net: a new multi-scale backbone architecture. IEEE Trans Pattern Anal Mach Intell 43(2):652–662
19. Dong G, Yan Y, Shen C, Wang H (2020) Real-time high-performance semantic image segmentation of urban street scenes. IEEE Trans Intell Transport Syst 22(6):3258–3274
20. Liu S, Huang D et al (2018) Receptive field block net for accurate and fast object detection. In: Proceedings of the European conference on computer vision (ECCV), pp 385–400
21. Peng J, Liu Y, Tang S, Hao Y, Chu L, Chen G, Wu Z, Chen Z, Yu Z, Du Y et al (2022) Pp-liteseg: a superior real-time semantic segmentation model. arXiv preprint arXiv:2204.02681

22. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N (2018) Bisenet: bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV), pp 325–341
23. Yu C, Gao C, Wang J, Yu G, Shen C, Sang N (2021) Bisenet v2: bilateral network with guided aggregation for real-time semantic segmentation. Int J Comput Vis 129:3051–3068
24. Wang J, Gou C, Wu Q, Feng H, Han J, Ding E, Wang J (2022) Rtformer: efficient design for real-time semantic segmentation with transformer. Adv Neural Inf Process Syst 35:7423–7436
25. Sun K, Zhao Y, Jiang B, Cheng T, Xiao B, Liu D, Mu Y, Wang X, Liu W, Wang J (2019) High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514
26. Thukral R, Aggarwal AK, Arora AS, Dora T, Sancheti S (2023) Artificial intelligence-based prediction of oral mucositis in patients with head-and-neck cancer: a prospective observational study utilizing a thermographic approach. Cancer Res Stat Treat 6(2):181–190
27. Maini D, Aggarwal AK (2018) Camera position estimation using 2d image dataset. Int J Innov Eng Technol 10:199–203
28. Brostow GJ, Shotton J, Fauqueur J, Cipolla R (2008) Segmentation and recognition using structure from motion point clouds. In: Computer vision–ECCV 2008: 10th European conference on computer vision, Marseille, France, October 12-18, 2008, Proceedings, Part I 10. Springer, pp 44–57
29. Roland G (2021) Rethink dilated convolution for real-time semantic segmentation. arXiv:2111.09957
30. Goyal P, Dollár P, Girshick R, Noordhuis P, Wesolowski L, Kyrola A, Tulloch A, Jia Y, He K (2017) Accurate, large minibatch sgd: training imagenet in 1 hour. arXiv preprint arXiv:1706.02677
31. Aggarwal AK, Jaidka P (2022) Segmentation of crop images for crop yield prediction. Int J Biol Biomed 7:40–44
32. Cubuk ED, Zoph B, Shlens J, Le QV (2020) Randaugment: practical automated data augmentation with a reduced search space. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pp 702–703
33. Brar DS, Aggarwal AK, Nanda V, Kaur S, Saxena S, Gautam S (2024) Detection of sugar syrup adulteration in unifloral honey using deep learning framework: an effective quality analysis technique. Food Hum 2:100190
34. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M et al (2015) Imagenet large scale visual recognition challenge. Int J Comput Vis 115:211–252
35. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32
36. Shrivastava A, Gupta A, Girshick R (2016) Training region-based object detectors with online hard example mining. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 761–769
37. Chen W, Gong X, Liu X, Zhang Q, Li Y, Wang Z (2019) Fasterseg: searching for faster real-time semantic segmentation. arXiv preprint arXiv:1912.10917
38. Brar DS, Aggarwal AK, Nanda V, Saxena S, Gautam S (2024) Ai and cv based 2d-cnn algorithm: botanical authentication of Indian honey. Sustain Food Technol
39. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
40. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2881–2890
41. Xu J, Xiong Z, Bhattacharyya SP (2023) Pidnet: a real-time semantic segmentation network inspired by pid controllers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 19529–19539
42. Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
43. Romera E, Alvarez JM, Bergasa LM, Arroyo R (2017) Erfnet: efficient residual factorized convnet for real-time semantic segmentation. IEEE Trans Intell Transport Syst 19(1):263–272
44. Li H, Xiong P, Fan H, Sun J (2019) Dfanet: deep feature aggregation for real-time semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 9522–9531
45. Hu P, Caba F, Wang O, Lin Z, Sclaroff S, Perazzi F (2020) Temporally distributed networks for fast video semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 8818–8827
46. Orsic M, Kreso I, Bevandic P, Segvic S (2019) In defense of pre-trained imagenet architectures for real-time semantic segmentation of road-driving images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 12607–12616
47. Nirkin Y, Wolf L, Hassner T (2021) Hyperseg: Patch-wise hypernetwork for real-time semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 4061–4070
48. Li X, You A, Zhu Z, Zhao H, Yang M, Yang K, Tan S, Tong Y (2020) Semantic flow for fast and accurate scene parsing. In: Computer Vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. Springer, pp 775–793
49. Chandra S, Couprie C, Kokkinos I (2018) Deep spatio-temporal random fields for efficient video segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 8915–8924
50. Si H, Zhang Z, Lv F, Yu G, Lu F (2019) Real-time semantic segmentation via multiply spatial fusion network. arXiv preprint arXiv:1911.07217