China Instrument and Control Society
Instrumentation
INSTRUMENTATION, Vol. 11, No. 2, June 2024 https://doi.org/10.15878/j.instr.202400137

Article

Triple-Branch Asymmetric Network for Real-time Semantic Segmentation of Road Scenes

Yazhi Zhang[1], Xuguang Zhang[1,*], Hui Yu[2]

[1] The Communication Engineering Department, Hangzhou Dianzi University, Zhejiang 310020, China

[2] The School of Creative Technologies, University of Portsmouth, UK

[*] Corresponding author email: zhangxg@hdu.edu.cn

Abstract: As the field of autonomous driving evolves, real-time semantic segmentation has become a crucial part of computer vision tasks. However, most existing methods use lightweight convolution to reduce the computational effort, resulting in lower accuracy. To address this problem, we construct TBANet, a network with an encoder-decoder structure for efficient feature extraction. In the encoder part, the TBA module is designed to extract details and the ETBA module is used to learn semantic representations in a high-dimensional space. In the decoder part, we design a combination of multiple upsampling methods to aggregate features with less computational overhead. We validate the efficiency of TBANet on the Cityscapes dataset. It achieves 75.1% mean Intersection over Union (mIoU) with only 2.07 million parameters and can reach 90.3 Frames Per Second (FPS).

Keywords: encoder-decoder architecture; lightweight convolution; real-time semantic segmentation

Citation: Yazhi Zhang, Xuguang Zhang, and Hui Yu. "Triple-Branch Asymmetric Network for Real-Time Semantic Segmentation of Road Scenes." *Instrumentation* 11, no. 2 (2024). https://doi.org/10.15878/j.instr.202400137.

1 Introduction

Semantic segmentation is defined as one of the tasks in computer vision that labels and classifies each pixel of the input image. Since it is a computationally intensive task for pixel-level classification, it faces challenges such as high computational overhead and high parametric load. Traditional semantic segmentation models, such as VGGNet[4] and SegNet[5], have high accuracy, but the huge number of parameters and slow inference speed make it difficult to meet the requirements for real-time in some fields. In order to solve this problem, the research of lightweight and real-time segmentation networks is essential.

Many real-time semantic segmentation networks have been presented for medical image processing[1], spatial robotics[2], autonomous driving[3] and surveillance environments[4]. The existing mainstream network architectures are mainly divided into two categories: (i) Bilateral structure which usually uses a high-resolution shallow spatial branch and a low-resolution deep semantic branch to extract features, and finally fuse the information descriptors by a feature fusion module, such as the BiseNet series[5-7], DFANet[8], and ContextNet[9]. Although such networks have improved in accuracy over single backbone networks, the computational and parametric quantities have also greatly increased. (ii) Encoder-decoder architecture, which consists of downsampling and convolutional layers in the encoder part to extract the features. In the decoder part, upsampling layers are designed to recover the image resolution and compensate for the lost details, such as the DeepLab series[10-12], LEDNet[13], and FCN[14]. Networks with such architectures tend to have many skip connections, which significantly increases the memory access cost and is not conducive to deploying the network to mobile devices. Obviously, for an encoder-decoder architecture network, improving speed while maintaining high accuracy is a pressing problem.

To this end, we propose the TBANet based on encoder-decoder architecture. Compared with other methods, our network excels in speed, accuracy, and parameters, as shown in Fig.1. The horizontal axis of Fig.1 represents FPS, the vertical axis represents mIoU and the radius of a circle is proportional to the number of parameters. From Fig.1, it can be seen that TBANet

performs well on the Cityscapes dataset. The overall structure of TBANet is shown in Fig.2. Since the existing feature extraction modules are large and cannot meet the demand for real-time semantic segmentation of road scenes, we propose the Triple-Branch Asymmetric (TBA) and ETBA modules. We carry out the TBA module with two branches using dilation convolution and channel split to speed up and the branch using 3×3 convolution to improve segmentation accuracy. The third branch uses ordinary convolution for feature extraction from the input image. Besides, channel shuffle is used to enhance the exchange of information between channels. In addition, we designed and implemented the ETBA module, which improves the level of information fusion between asymmetric convolutions based on the TBA module to meet the needs of gradually growing communication for information between channels in the deep network. The ETBA module adds an additive operation to the asymmetric convolution, enhancing inter-channel communication, resulting in a finer segmentation. In addition, ECA[33] and PSA[34] are applied to improve segmentation accuracy. We apply all the above modules to the encoder part. Furthermore, a Multiple-Methods Aggregation (MMA) module is applied to combine feature maps of the different stages. The module takes full advantage of multiple upsampling methods, so it can restore image resolution accurately.

In summary, the contributions of this paper are as follows:

• A Triple-Branch Asymmetric (TBA) module is proposed to fleetly extract edge and detail information with low computing complexity. It uses convolution with a kernel of 3×3 for fast concatenation, which effectively reduces the gridding artifacts due to dilated convolution.

• An Enhanced Triple-branch Asymmetric (ETBA) module is modified from the TBA module. This module is designed for image feature extraction in the later stages of the network. The module has a powerful semantic feature characterization capability because it obtains an abundance of receptive fields and frequent communication between channels.

• A Multiple-Methods Aggregation (MMA) module using two upsampling methods is employed to aggregate low-level spatial information and high-level semantic information. At the same time, the resolution of inputs is recovered with less computational effort.

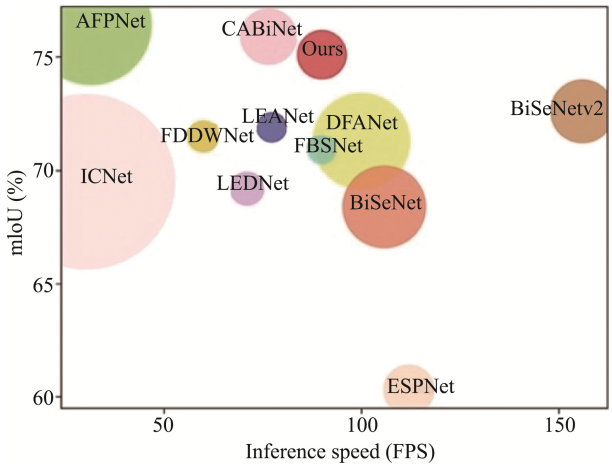

Fig.1 Accuracy, speed, and parameters comparisons on the Cityscapes test set

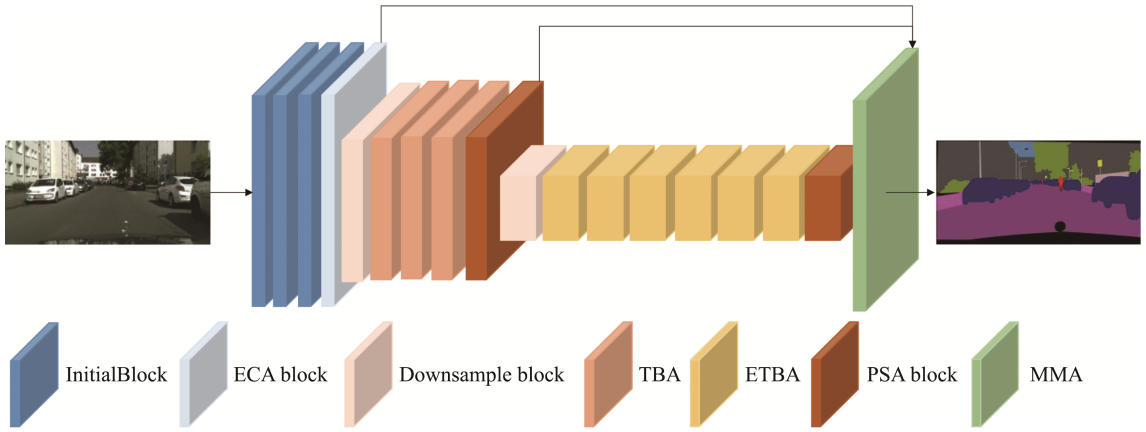

Fig.2 Overview architecture of the proposed TBANet

2 Related work

With the development of mobile terminals, real-time semantic segmentation tasks are gaining more and more attention. This section introduces the three categories most relevant to our work.

2.1 Real-time semantic segmentation

Define Real-time semantic segmentation has received widespread attention due to its advantages such as low device requirements and fast inference, and many network models with excellent performance already exist. FCN[14] proposes to replace the fully connected layer with the convolution, which highly reduces the computation. ENet[17] uses dilated convolution and asymmetric convolution to remove redundancy. ESPNet[18] decouples the convolution to achieve a good result. ICNet[19] processes images of different resolutions separately and finally fuses the acquired feature maps to

have high accuracy. To excel in accuracy and speed, BiSeNetV2[7] designed a bilateral structure for details and semantic information and also proposed an efficient aggregation module to fuse different feature maps. STDC[20] improved the situation that the semantic branching perceptual field of BiSeNetV2[7] is not rich by designing a short-term dense connection module. DDRNet[49] designed a bilateral fusion module to facilitate the fusion of detail and contextual information to improve its accuracy.

2.2 Encoder-decoder architecture

The encoder-decoder architecture is common in computer vision. U-Net[21] designed a symmetric encoder-decoder architecture, which has better accuracy but the model size is too large and not suitable for real-time semantic segmentation tasks. DTT[54] uses an encoder-decoder architecture in video recognition. Most of the subsequent networks adopt asymmetric encoder-decoder architecture to reduce the model size by designing lightweight and efficient decoders, such as ENet[17], RefineNet[22], DABNet[23], etc. Most networks with this structure fuse shallow and deep feature maps by skip connections, which promotes network convergence but also largely increases the inference cost.

2.3 Effective feature extraction method

The selection of the receptive fields is essential because of the variable scales of objects. Many networks are innovative in feature extraction modules, such as EdgeNet[50]. There are some networks[51,52] that use multidimensional features to improve network accuracy. PSPNet[24] designed the Pyramid Pooling Module (PPM) to obtain multi-scale contextual information features. Shufflenet V2[25] uses channel split to improve the model inference speed. ESPNet[26] adopts a pyramid module with a mass of skip connections EADNet[27] that uses asymmetric depth-separable dilated convolution to form a pyramidal pooling module for extracting multi-scale contextual information. However, the existing feature extraction modules are not designed for pre and late-network characteristics. To solve this problem, we propose TBA and ETBA modules.

3 Method

In this section, we focus on the composition and structure of our proposed TBANet. First, we present how the lightweight feature extraction TBA and ETBA modules are constructed. Then, we show the design of the efficient feature fusion MMA module. Finally, we present the overall architecture of TBANet.

3.1 Triple-branch asymmetric module

We propose the TBA module with the aim of extracting image features in the pre-semantic segmentation network. This module can extract features of input images efficiently with limited hardware device resources and time. The structure of the TBA module is shown in Fig.3(c).

To improve the learning ability of the module, the TBA module consists of three main branches. The input of the two branches on the left is generated from the original input by channel split. The channel split operation averages the input segmentation along the channel axis, which can notably reduce the inference time. Meanwhile, many real-time semantic segmentation networks have demonstrated the effectiveness of asymmetric convolution, such as ERFNet[30] (Fig.3(a)), DABNet[23] (Fig.3(b)), EDANet[28], etc. We employ asymmetric depth-wise separable convolution in the two branches on the left, which is an idea that can decouple a two-dimensional convolution into two one-dimensional convolutions. Assuming the input feature map size is and the output map size is $C_{in}\times H\times W$. Meanwhile, C, H and W

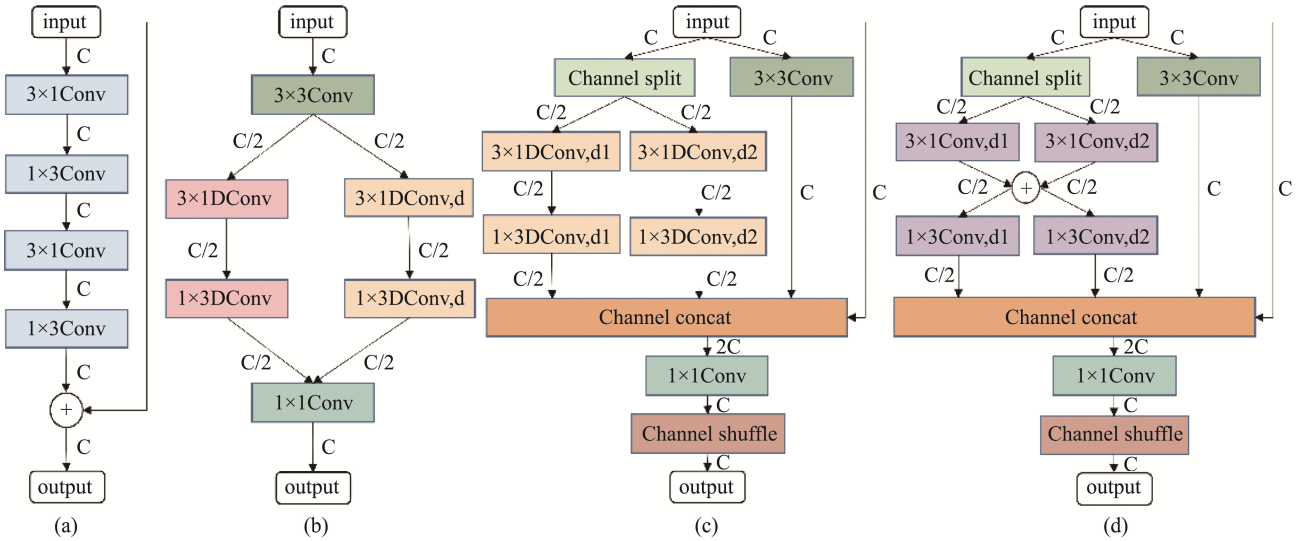

Fig.3 The different structures of feature extraction modules. (a)Non-bottleneck-1D module. (b)DAB module. (c)TBA module. (d)ETBA module. "Conv" denotes convolution, "Dconv" denotes depth-wise separable convolution, and "d" denotes different dilation rates

represent channels, height and width respectively. P is parameters and Com is computational effort. The standard 3×3 convoltion is following formula:

$$P_{3\times3}=3\times3\times C_{in}\times C_{out}\# \quad (1)$$

$$Com_{3\times3}=3\times3\times H\times W\times C_{in}\times C_{out}\# \quad (2)$$

The 3×3 asymmetric depth-separable dilation convolution is following formula:

$$P_{\mathrm{ADD}}=2\times3\times C_{in}+C_{in}\times C_{out}\# \quad (3)$$

$$Com_{\mathrm{ADD}}=2\times3\times H\times W\times C_{in}+C_{in}\times C_{out}\times H\times W\ \# \quad (4)$$

The diversity of target object sizes for semantic segmentation is such that the network needs to have size-rich receptive fields to capture features efficiently. To address this need, we use different dilation rates on the two branches to vary the size of the receptive field with a small computational overhead. In the pre-network stage, the image details are more complete, we want the TBA module to focus on small objects and edge segmentation. So, we set a small dilation rate. Considering that the dilation convolution loses some pixel information, we use a 3×3 convolution operation on the third branch to compensate for the loss. Also, it can capture the features of short-range objects. Inspired by ResNet[31], to solve the gradient disappearance problem, we used the residual concatenation method, which performs a channel concat operation on the input and the output from the three branches. Then the number of channels is reduced using 1×1 convolution to decrease the computational burden. Finally, to enable information communication between channels, we use the channel shuffle strategy.

3.2 Enhance triple-branch asymmetric module

With the deepening of the network and the increasing number of image channels, communication between channels is especially important. Based on this, we design the ETBA module to better extract the high-level semantic information of the input. The structure of the ETBA module is shown in Fig.3(d).

In the late stage of the semantic segmentation network, the number of channels of the feature image is generally high and contains rich semantic information. Different from the TBA module, the two left branches of ETBA use asymmetric dilation convolution with a 3×3 convolution kernel. This modification avoids the damage of deeply separable convolution on the information flow between channels and ensures the accuracy of the network. It is worth noting that we also add the feature map addition operation inside the asymmetric dilation module significantly improves the module's ability to extract semantic information with less computational overhead. This part is expressed as formulas:

$$x_1=f_{3\times1,\,d_1}(x_{1,1})+f_{3\times1,\,d_2}(x_{1,2})\# \quad (5)$$

$$y_1=f_{1\times3,\,d_1}(x_1)\# \quad (6)$$

$$y_2=f_{1\times3,\,d_2}(x_1)\# \quad (7)$$

where $x_{1,1}$ and $x_{1,2}$ mean the two outputs of the channel split operation. y_1 and y_2 represent the output of the left and middle branch. $f_{3\times1}$ and $f_{1\times3}$ represent the two steps of asymmetric convolution. d_1 and d_2 denotes different dilation rates. In order to have a richer receptive field, the dilation rates of the ETBA module are larger than those of the TBA module. The ETBA module is more suitable for feature extraction in the later stages of the network.

3.3 Multiple-methods aggregation module

Common semantic segmentation networks usually use a more complex decoder structure to reach a high-precision segmentation target, and this approach is not suitable for real-time semantic segmentation tasks. To address this problem, we propose a multiple-methods aggregation (MMA) module, as shown in Fig.4.

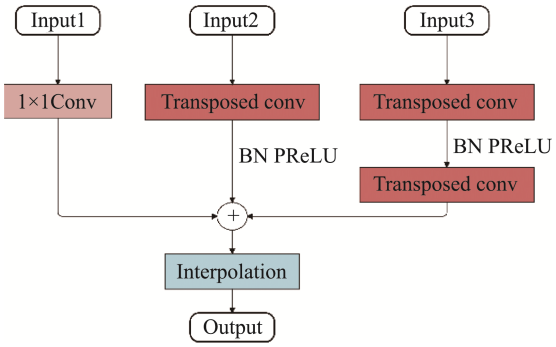

Fig.4 The structure of the multiple-methods aggregation module. "Conv" denotes convolution, and "Interpolation" denotes bilinear interpolation

The previous BiAttnNet[3] uses bilinear interpolation for upsampling, which is a low computational cost upsampling method. However, if this method is used singly, a large amount of high-dimensional information may be lost, which affects the network segmentation accuracy. Inspired by the operation of FCN[14] using transposed convolution to recover image information, we use transposed convolution to recover the main feature image elements. This is a learnability upsampling method that is more in line with the characteristics of deep learning networks. So the MMA module combines the above two upsampling methods to recover the image resolution efficiently. In addition, the decoder we designed incorporates three different stages of feature maps, and the feature reuse rate reaches a high level. We select the output of the ECA block containing rich edge details as the x_1 and make it pass through a 1×1 convolution layer to normalize the number of channels. x_2 and x_3 are processed using transposed convolution to normalize their dimensions. Because the late image resolution is extensive, using transposed convolution causes expensive computational costs and severely delays the network inference, so the bilinear interpolation method is used to process the feature maps after the fusion of the three branches. The expression for MMA is as follows:

$$y=F\{Con(x_1)+T(x_2)+T(x_3)\}\# \quad (8)$$

Here, F means bilinear interpolation, Con is convolution computations, and T denotes transposed convolution. x_1 is the output of the ECA module. The two PSA block outputs are used as x_2 and x_3 in turn.

3.4 Network architecture

The proposed triple-dilation asymmetric network for real-time semantic segmentation has two main architectures: encoder and decoder, as shown in Fig.2. more detailed architectural information on TBANet is shown in Table 1.

Table 1 The architectural details of TBANet

Stage	Type	Mode	Channel	Output
	Initial Block	Stride=2	32	256×512
	ECA	-	35	256×512
	Downsample	Stride=2	64	128×256
	TBA×3	Dilated=(2,5)	64	128×256
Encoder	PSA	-	128	128×256
	Downsample	Stride=2	128	64×128
	ETBA-1×3	Dilated=(3,7)	128	64×128
	ETBA-2×3	Dilated=(9,11)	128	64×128
	PSA	-	256	64×128
Decoder	MMA	-	19	512×1024

TBANet designed the model with fewer layers to reduce the risk of overfitting. The activation function of PReLU and batch normalization layer are also added to the model, which enhances the expressive ability of the model to handle the distribution of different data better.

a) Downsampling: The input image is fed to TBANet first by an Initial Block consisting of three 3×3 convolutional layers cascaded, where the first 3×3 convolution has a step size of 2, thus completing the downsampling of the image. It is worth mentioning that in TBANet, we only perform three downsampling operations on the images, which ensures the accuracy and real-time performance of our network. For the other two downsampling tasks, we use the Initial Block of ENet[17] as the Downsample Block, which reduces the image resolution with less detailed information lost.]

b) Mechanism of attention: Networks such as SENet[43] and DANet[44] have demonstrated the importance of using attention in the network. There is a wealth of high-level semantic information contained in the channels, which is vital for the segmentation task. To boost the information shared between channels, we design to add ECA[33] behind Initial Block and use PSA[34] behind TBA and ETBA which can fuse different scales of contextual information and is more suitable for application in deeper network layers.The output of PSA[34]is as follows:

$$\mathrm{PSA(X)}=A^{ch}(X)\Theta^{ch}X+A^{sp}(X)\Theta^{sp}X\# \quad (9)$$

Where A^{ch} and A^{sp} are the output of channel-only branch and spatial-only branch, respectively.The Θ^{ch} is a channel-wise multiplication operator and Θ^{sp} is a spatial-wise multiplication operator.

4 Experiments

In this section, two datasets (Cityscapes and CamVid) are used to test the performance of our network. We present the specific details and relevant parameters in our experiments. We perform ablation and comparison experiments on the designed modules to verify their effectiveness of the modules. Finally, we compare the experimental results of TBANet and state-of-art networks in various aspects.

4.1 Datasets

Cityscapes[15] is a large dataset for the semantic understanding urban street scenes with high pixel resolution. It contains 5000 images of city street scenes from 50 cities. In addition, it also has 19998 images with rough annotations, which we did not use in our experiments. Each image in the dataset has a resolution of 1024×2048, and there are 19 classes for semantic classification.

Table 2 Ablation study results about TBA and ETBA on cityscapes validation set

Feature Extraction Module		mIoU (%)	FPS	Params (M)
TBA	TBA	71.4	94.5	1.75
ETBA	ETBA	74.7	73.1	2.10
Non-bottleneck-1D	Non-bottleneck-1D	68.5	88.0	1.40
DAB module	DAB module	70.8	100.6	1.00
TBA-r = [2,4]	ETBA-r = [4,8],[8,16]	74.2	89.3	2.07
TBA-r = [2,5]	ETBA-r = [5,9],[9,13]	75.1	89.8	2.07
TBANet		75.7	90.3	2.07

CamVid[16] is an autonomous driving scene understanding dataset with 11 semantic categories. CamVid is a smaller dataset that contains only 701 images (367 of them as a training set, 101 as a validation set, and 233 for testing) with an image resolution of 360×480.

4.2 Implementation details

We conducted experiments on Pytorch 1.7.0 with an NVIDIA RTX2080Ti GPU, using stochastic gradient descent to train the network. A "poly" learning rate strategy with momentum 0.9, and weight decay $2e^{-4}$ was used during training, and the initial learning rate was set to $4.5e^{-2}$, the learning rate was calculated as $lr_{initial}\times\left(1-\frac{inter}{inter_{max}}\right)^{0.9}$. Regarding the initial loss function, we use the cross-entropy loss: $\mathrm{L}=\frac{1}{N}\sum_{i}-\sum_{C=1}^{M}y_{ic}\log(p_{ic})$. We reduce the resolution of the input image, which greatly reduces the memory footprint. We crop the input image to 512×1024. Unlike the Cityscapes dataset, we

adopt the weight decay $1e^{-3}$ on the CamVid dataset. Regarding the data enhancement strategy, we used random cropping, random horizontal flipping, and random scaling to enhance the data. Regarding the conclusion section, we evaluate the network performance regarding mIoU, FPS, and parameters.

4.3 Ablation Studies

In this chapter, to demonstrate the effectiveness of the modules in TBANet, we perform ablation experiments on the Cityscapes dataset.

a) Ablation Study for TBA and ETBA: To verify that the TBA module is suitable for shallow network feature extraction and the ETBA module is suitable for the deep, we designed comparison experiments. As shown in Table 2, when using TBA alone as the feature extraction module in the decoder, the accuracy is substantially reduced by 4.3%, although there is a slight improvement in FPS and parameters. The TBA module possesses a small receptive field and cannot meet the growing demand for cross-channel information exchange in the later stages of the network, so the network segmentation accuracy is poor when the TBA module is used alone. When using ETBA alone, the performance in terms of mIoU, FPS and parameters is poor. This is due to the fact that at the beginning of the network, the size of the channel dimension of the feature map is small and the feature extraction ability of the network is weak, so the ETBA module enables the channel dimension to exchange information frequently may increase the impact of the wrong prediction information, which impairs the segmentation effect of the network. In order to verify the efficiency of TBA and ETBA modules, we use the non-bottleneck 1D module and DAB module respectively in Fig.3(a) and Fig.3(b) instead of TBA and ETBA. It can be learned from Table 2 that the accuracy drops from 75.7% to 68.5% and the speed slows down when using the non-bottleneck 1D module as the feature extraction module. When the DAB module is used as the feature extraction module, it improves 10.3 FPS, but the accuracy is destroyed substantially. It can be seen that the TBA and ETBA modules are complementary and efficient.

b) Ablation Study for dilation rates: The list of dilation rates used in TBA and ETBA in TBANet is {[2,5], [3,7], [9,11]}, and to prove the effectiveness of this dilation rate, we design comparison experiments for different dilation rates cases as shown in Table 2. EFRNet[41] argued that large multiplying dilation rates could obtain better results; however, applied in TBANet, only 74.2% mIoU was obtained. We visualized the output using multiplying dilation rates and found that the picture produces a gridding effect. After we set the dilation rate according to the coprime rules, the gridding effect of the segmented image of TBANet becomes smaller. We also experimented with large coprime dilation rates with 0.6% lower accuracy than TBANet.

c) Ablation Study for MMA: In the decoder section, we designed the MMA module to recover the image resolution. To verify the performance of this module, we use bilinear interpolation instead of MMA. As shown in Table 3, the results prove that MMA is more suitable to TBANet than the single use of bilinear interpolation, with a higher 4.7% mIoU at a low time consumption. It can be seen that the MMA module we designed is beneficial to TBANet.

Table 3 Ablation study results about MMA and attention on cityscapes validation set

ECA	PSA-1	PSA-2	MMA	mIoU(%)	FPS
√		√	√	73.2	95.4
√	√		√	71.7	94.2
	√	√	√	74.2	91.8
√	√	√		71.0	88.9
√	√	√	√	75.7	90.3

d) Ablation Study for attention: In the encoder part, we used ECA[33] in the shallow phase of the network and applied PSA[34] after TBA and ETBA. PSA-1 and PSA-2 represent PSA modules added after the TBA and ETBA modules, respectively. Experiments are implemented for each attention module separately. As shown in Table 3, each attention module has different degrees of influence on network accuracy, among which PSA-2 has the largest influence on mIoU up to 4%. Although removing the attention module resulted in a small increase in the network's speed, it sacrificed the segmentation accuracy of the network, which is not what we expected. And a series of comparative experiments have been done on the location of the PSA module, the experimental results are shown in Fig.5. The horizontal coordinates of the image represent the location where the PSA module was placed, and the vertical coordinates represent mIoU. After fixing the PSA-2 position constant and placing PSA-1 in the first TBA and second TBA modules, respectively, the network segmentation accuracy in these cases decreased slightly. When the PSA-1 position is constant and the PSA-2 module position is changed, the further the PSA-2 is placed, the higher the network accuracy. The experiments proved that the PSA module has the best effect when it is put behind the last TBA module and ETBA module. It shows that the attention mechanism in TBANet benefits the accuracy improvement.

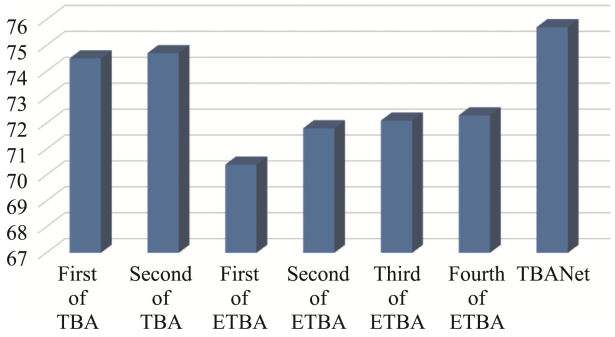

Fig.5 Ablation study for location of PSA module.

e) Visualizing Ablation Results: In order to demonstrate more intuitively the impact of our proposed module on the segmentation results, we visualize the results of the relevant ablation experiments, as shown in Fig.6. It can be seen that when the TBA module is removed, the network is unable to accurately segment small-sized objects in the distant view, and the segmentation results for objects in the pole class are also poor. After ablating the ETBA module, the network cannot recognize large-sized objects because of the restricted receptive field. When the MMA module is not used as the up-sampling module, the network cannot recover the resolution accurately, resulting in a rougher quality of the final predicted image. From the above results, it can be proved that our proposed module is beneficial for the model to segment the input image accurately.

4.4 Evaluation results on cityscapes

In this phase, we test the accuracy and inference rate of the proposed TBANet on the widely used test set of Cityscapes and compare it with other advanced real-time semantic segmentation models. For a fair comparison, we do not use testing techniques such as multi-crop and multi-scale testing during the testing period, and data for other networks are obtained from the relevant literature.

The results in Table 4 show that TBANet achieves 75.1% mIoU on the Cityscapes test set, already surpassing

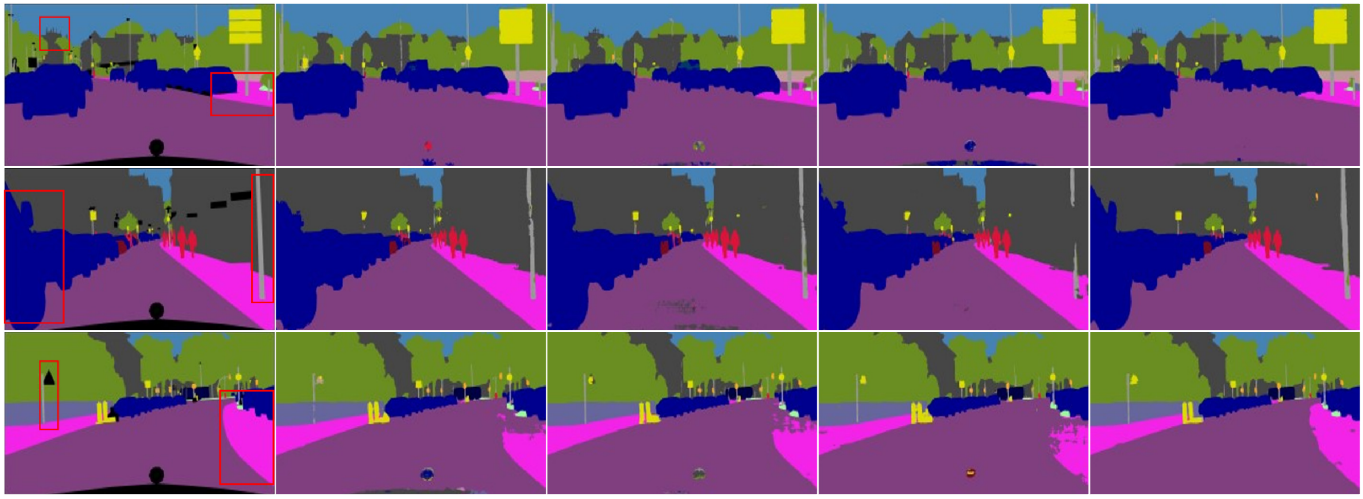

Fig.6 Visualizing ablation results. From left to right: Ground-truth, ablation of TBA, ETBA, MMA and TBANet.

Table 4 Evaluation results between TBANet and other state-of-art networks on the cityscapes test set

Method	Input Size	Pretrain	GPU	mIoU(%)	FPS	Params(M)
ENet[17]	512×1024	No	Titan	58.3	76.9	0.36
ESPNet[26]	512×1024	No	Titan	60.3	11	2.10
ICNet[19]	512×1024	ImageNet	Titan	69.5	30.3	26.50
BiseNet[6]	768×1536	ImageNet	TitanXp	68.4	105.8	5.80
DABNet[23]	512×1024	No	1080Ti	70.1	104.0	0.76
LEDNet[13]	512×1024	No	1080Ti	69.2	71	0.94
DFANet[8]	1024×1024	ImageNet	TitanX	71.3	100	7.80
MSCFNet[45]	512×1024	No	TitanXp	71.9	50	1.15
BiSeNetv2[7]	512×1024	No	1080Ti	72.6	156	3.4
CFPNet[36]	1024×2048	No	2080Ti	70.1	30	0.55
AFPNet[35]	1024×2048	No	TitanXp	76.4	31.5	12.2
FDDWNet[37]	512×1024	No	2080Ti	71.5	60	0.80
FBSNet[38]	512×1024	No	2080Ti	70.9	90	0.62
CABiNet[39]	1024×2048	No	2080Ti	75.9	76.5	0.64
LEANet[39]	512×1024	No	1080Ti	71.9	77.3	0.74
LAANet[29]	512×1024	No	1080Ti	73.6	95.8	0.67
BiAttnNet[32]	512×1024	No	2080Ti	74.7	89.2	2.20
EANet[46]	1024×2048	No	1080Ti	74.6	35.4	12.6
PCNet[47]	1024×2048	Scratch	2080Ti	72.9	79.1	1.49
TBANet(ours)	512×1024	No	2080Ti	75.1	90.3	2.07

Many real-time semantic segmentation networks. AFPNet[35] is slightly more accurate than TBANet, but is almost three times slower and has six times more parameters. TBANet has almost the same speed as BiAttnNet[32] and is inferior to our network regarding parameters and accuracy. In conclusion, our network has excellent speed and performs well in terms of accuracy.

To show the superior segmentation capability of our model in more detail, Fig.7 shows the images segmented on the Cityscapes dataset. In the first set of comparison figures, it can be seen that the TBANet is more accurate in segmenting fences and poles. Good segmentation results of our network for street lights and traffic signs are shown in the second and third sets of figures. The last set of figures, it is demonstrated that the TBANet can segment trucks more accurately. As can be seen from the segmentation results, TBANet not only has excellent segmentation ability for small objects, but also can accurately categorize large objects.

4.5 Evaluation results on CamVid

To further demonstrate the generalization ability and efficiency of the proposed TBANet, we conduct experiments on another frequently used CamVid dataset. A comparison with segmentation networks that have performed well in recent years is shown in Table 5. TBANet has achieved good experimental results on CamVid both in terms of accuracy and speed, better balancing speed and accuracy. DSANet[42], although slightly more accurate than TBANet, is much slower than our network. Although the mIoU of MLFNet[48] is higher than that of TBANet on the CamVid dataset, the network does not segment as well as TBANet on the Cityscapes dataset. It can be seen that our model accomplishes the two goals of real-time semantic segmentation: fast speed and high accuracy.

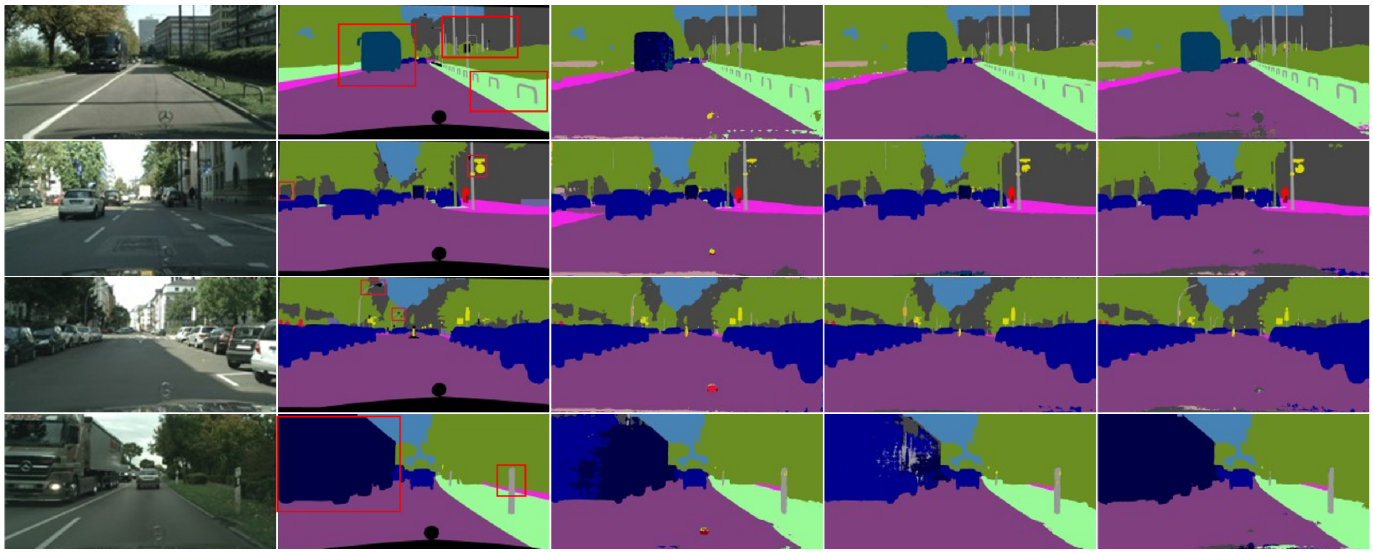

Fig.7 Visual comparison on Cityscapes validation set. From left to right: Input image, ground-truth, prediction of DABNet, FDDWNet, and TBANet.

Table 5 Evaluation results on the CamVid test set

Method	GPU	mIoU	FPS
ENet [17]	TitanX	51.3	98.8
ESPNet [26]	TitanX	58.2	112
ICNet [19]	TitanX	67.1	27.8
BiseNet [5]	TitanXp	65.6	175
DABNet [23]	1080Ti	66.2	124.4
DFANet [8]	TitanX	64.7	120
LEANet [40]	1080Ti	67.5	98.6
FDDWNet [37]	2080Ti	66.9	79
DSANet [42]	1080Ti	69.9	75.3
LAANet [29]	1080Ti	67.9	112.5
MLFNet-Res34 [48]	2080Ti	69.0	57.2
TBANet(ours)	2080Ti	68.4	112.9

5 Conclusion

In this paper, we propose a TBANet applied to real-time road scene segmentation and experimentally demonstrate that it shows better capabilities in both speed and accuracy. In the encoder part, our proposed TBA and ETBA can extract image features efficiently under the constraints of small computational costs. In the decoder part, the MMA is designed based on the characteristics of standard upsampling methods that can aggregate multi-stage feature information. In conclusion, TBANet is a real-time semantic segmentation network with outstanding performance in parametric number, accuracy and speed.

Author Contributions:

Yazhi Zhang is responsible for the main network

design, experimental research and writing original draft. Xuguang Zhang and Yu Hui are responsible for checking and reviewing the draft, data curation and formal analysis.

Funding Information:

This research received no external funding.

Data Availability:

The authors declare that the main data supporting the findings of this study are available within the paper and its Supplementary Information files.

Conflict of Interest:

The authors declare no competing interests.

Dates:

Received 15 February 2024; Accepted 2 April 2024; Published online 30 June 2024

References

[1] K. Yang, X. Hu, L. M. Bergasa, E. Romera, and K. Wang. (2020). Pass: Panoramic annular semantic segmentation. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 10, pp. 4171-4185.

[2] V.-C. Miclea and S. Nedevschi. (2020). Real-time semantic segmentation-based stereo reconstruction. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems, vol. 21, no. 3,4 pp. 1514-1524.

[3] Y. Kang, K. Yamaguchi, T. Naito, and Y. Ninomiya. (2011). Multiband image segmentation and object recognition for understanding road scenes. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems, vol. 12, no. 4, pp. 1423-1433.

[4] Simonyan K, Zisserman A. (2015). Very deep convolutional networks for large-scale image recognition. In Proceedings of the International Conference on Learning Representation.

[5] V. Badrinarayanan, A. Kendall, and R. Cipolla. (2017). SegNet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE transactions on pattern analysis and machine intelligence , vol. 39, no. 12, pp. 2481-2495.

[6] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang. (2018). BISENet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the European conference on computer vision (ECCV), pp. 325-341.

[7] C. Yu, C. Gao, J.Wang, G. Yu, C. Shen, and N. Sang. (2021). BiseNet v2:Bilateralnetwork with guided aggregation for real-time semantic segmentation. In Proceedings of the International Journal of Computer Vision, vol. 129, no. 11, pp. 3051-3068.

[8] H. Li, P. Xiong, H. Fan, and J. Sun. (2019). DFANet: Deep feature aggregation for real-time semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 9522-9531.

[9] R. P. Poudel, U. Bonde, S. Liwicki, and C. Zach. (2018). CONTEXTNet: Exploring context and detail for semantic segmentation in real-time. In Proceedings of the British Machine Vision Conference(BMVC), pp. 146-146.

[10] Chen L C, L. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2018). Deeplab: Semantic image segmentation with deep convolutional nets,atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Anal- ysis and Machine Intelligence , vol. 40, no. 4, pp. 834-848, Apr. 2018.

[11] Chen L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam. (2017). Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587.

[12] 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.833-851.

[13] Y. Wang, Q. Zhou, J. Liu, J. Xiong, G. Gao, X. Wu, and L. J. Latecki. (2019). LEDNet: A lightweight encoder-decoder network for real-time semantic segmentation. In Proceedings of the IEEE International Conference on Image Processing, pp. 1860-1864.

[14] E. Shelhamer, J. Long, and T. Darrell. (2017). Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Anal- ysis and Machine Intelligence, vol. 39, no. 4, pp. 640-651.

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele. (2020). The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 3213-3223.

[16] G. J. Brostow, J. Shotton, J. Fauqueur, and R. Cipolla. (2008). Segmentation and recognition using structure from motion point clouds. In Proceedings of the European conference on computer vision (ECCV), pp. 44-57.

[17] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello .(2016). ENET: A deep neural network architecture for real-time semantic segmentation. arXiv preprint arXiv:1606.02147.

[18] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi. (2019). Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), pp. 9190-9200.

[19] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia. (2018). ICNet for real-time semantic segmentation on high-resolution images. In Proceedings of the European conference on computer vision (ECCV), pp. 405-420.

[20] M. Fan et al. (2021, June). Rethinking BiseNet for real-time semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9716-9725.

[21] O. Ronneberger, P. Fischer, and T. Brox. (2015). U-net: Convolutional networks nfor biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234-241.

[22] G. Lin, A. Milan, C. Shen and I. Reid. (2017). RefineNet: Multi-path inement Networks for High-Resolution Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5168-5177.

[23] G. Li, I. Yun, J. Kim, and J. Kim. (2019). DABNet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. In Proceedings of the British Machine Vision Conference(BMVC), pp. 1-12.

[24] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia. (2017). Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2881-2890.

[25] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun. (2018). Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European conference on computer vision (ECCV), pp. 116-131.

[26] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi. (2018). Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the European conference on computer vision (ECCV), pp. 552-568.

[27] Q. Yang, T. Chen, J. Fan, Y. Lu, C. Zuo and Q. Chi. (2021, June). EADNet: Efficient Asymmetric Dilated Network For Semantic Segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2315-2319.

[28] S.-Y. Lo, H.-M. Hang, S.-W. Chan, and J.-J. Lin. (2019). Efficient dense modules of asymmetric convolution for real-time semantic segmentation. In Proceedings of the ACM Multimedia Asia, pp. 1-6.

[29] X. Zhang, B. Du, Z. Wu, and T. Wan. (2022). LAANet: Lightweight attention-guided asymmetric network for real-time semantic segmentation. Neural Computing & Applications, vol. 34, pp. 1-15.

[30] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo. (2018). ERFNet:Efficient residual factorized convnet for real-time semantic segmentation. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems, vol. 19, no. 1, pp. 263-272.

[31] K. He, X. Zhang, S. Ren, and J. Sun. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770-778.

[32] G. Li, L. Li and J. Zhang. (2022). BiAttnNet: Bilateral Attention for Improving Real-Time Semantic Segmentation. IEEE Signal Processing Magazine, vol. 29, pp. 46-50.

[33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu. (2020). ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11534-11542.

[34] H. Liu, F. Liu, X. Fan, D. Huang. (2021). Polarized self-attention: Towards high-quality pixel-wise regression. arXiv preprint arXiv:2107.00782.

[35] J. Hyun, H. Seong, S. Kim and E. Kim. (2022). Adjacent Feature Propagation Network (AFPNet) for Real-Time Semantic Segmentation.IEEE Transactions on Systems, Man, and Cybernetics, vol. 52, no. 9, pp. 5877-5888.

[36] A. Lou and M. Loew. (2021). CFPNET: Channel-Wise Feature Pyramid For Real-Time Semantic Segmentation. IEEE International Conference on Image Processing, pp. 1894-1898.

[37] J. Liu, Q. Zhou, Y. Qiang, B. Kang, X. Wu and B. Zheng. (2020). FDDWNet: A Lightweight Convolutional Neural Network for Real-Time Semantic Segmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2373-2377.

[38] G. Gao, G. Xu, J. Li, Y. Yu, H. Lu and J. Yang. (2022). FBSNet: A Fast Bilateral Symmetrical Network for Real-Time Semantic Segmentation. In Proceedings of the IEEE Transactions on Multimedia.

[39] S. Kumaar, Y. Lyu, F. Nex and M. Y. Yang. (2021). CABiNet: Efficient Context Aggregation Network for Low-Latency Semantic Segmentation. In Proc. IEEE Int. Conf. Robotics and Automation, pp. 13517-13524.

[40] X.-L. Zhang, B.-C. Du, Z.-C. Luo, and K. Ma. (2022). Lightweight and efficient asymmetric network design for real-time semantic seg- mentation. International Journal of Speech Technology, vol. 52, no. 1, pp. 564-579.

[41] Y Li, M Li, Z. Li, C. Xiao, and H. Li. (2022). EFRNet: Efficient Feature Reuse Network for Real-time Semantic Segmentation. In Proceedings of the Neural Processing Letters, pp: 1-13.

[42] M. A. Elhassan, C. Huang, C. Yang, and T. L. Munea. (2021). DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes. In Proceedings of the Expert Systems with Applications, vol. 183, p. 115090.

[43] J. Hu, L. Shen, and G. Sun. (2018). Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132-7141.

[44] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu. (2019). Dual attention network for scene segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3146-3154.

[45] G. Gao, G. Xu, Y. Yu, J. Xie, J. Yang and D. Yue. (2021). MSCFNet: A Lightweight Network With Multi-Scale Context Fusion for Real-Time Semantic Segmentation. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems], pp. 25489-25499.

[46] J. Dong, J. Guo, H. Yue and H. Gao. (2022). EANET: Efficient Attention-Augmented Network for Real-Time

Semantic Segmentation. IEEE International Conference on Image Processing, pp. 3968-3972.

[47] Q. Lv, X. Sun, C. Chen, J. Dong and H. Zhou. (2021). Parallel Complement Network for Real-Time Semantic Segmentation of Road Scenes. In Proceedings of the IEEE Transactions on Intelligent Transportation Systems, pp. 4432-4444.

[48] J. Fan, F. Wang, H. Chu, X. Hu, Y. Cheng and B. Gao. (2023), MLFNet: Multi-Level Fusion Network for Real-Time Semantic Segmentation of Autonomous Driving. IEEE Transactions on Intelligent Vehicles, pp. 756-767.

[49] Y. Hong, H. Pan, W. Sun, Y. Jia. (2021). Deep dual-resolution networks for real-time and accurate semantic segmentation of road scenes. arXiv preprint arXiv: 2101.06085.

[50] H.-Y. Han, Y.-C. Chen, P.-Y. Hsiao, and L.-C. Fu. (2020). Using channel- wise attention for deep cnn based real-time semantic segmentation with class-aware edge information. IEEE Transactions on Intelligent Vehicles, vol. 22, no. 2, pp. 1041-1051.

[51] L. Ding, J. Terwilliger, R. Sherony, B. Reimer, and L. Fridman. (2022). Value of temporal dynamics information in driving scene segmentation. IEEE Transactions on Intelligent Vehicles, vol. 7, no. 1, pp. 113-122.

[52] Y. Zhu, Z. Li, F. Wang and L. Li. (2023). Control Sequences Generation for Testing Vehicle Extreme Operating Conditions Based on Latent Feature Space Sampling. IEEE Transactions on Intelligent Vehicles, pp. 1-11.

[53] P. W. Patil, K. M. Biradar, A. Dudhane, and S. Murala. (2021, June). An end-to-end edge aggregation network for moving object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8149-8158.

[54] P. W. Patil, A. Dudhane, A. Kulkarni, S. Murala, A. B. Gonde, and S. Gupta. (2021). An unified recurrent video object segmentation framework for various surveillance environments. IEEE Transactions on Image Processing, vol. 30, pp. 7889-7902.