electronics

MDPI

Article

LASNet: A Light-Weight Asymmetric Spatial Feature Network for Real-Time Semantic Segmentation

Yu Chen, Weida Zhan *, Yichun Jiang, Depeng Zhu, Renzhong Guo and Xiaoyu Xu

National Demonstration Center for Experimental Electrical, School of Electronic and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China
* Correspondence: zhanweida@cust.edu.cn

Abstract: In recent years, deep learning models have achieved great success in the field of semantic segmentation, which achieve satisfactory performance by introducing a large number of parameters. However, this achievement usually leads to high computational complexity, which seriously limits the deployment of semantic segmented applications on mobile devices with limited computing and storage resources. To address this problem, we propose a lightweight asymmetric spatial feature network (LASNet) for real-time semantic segmentation. We consider the network parameters, inference speed, and performance to design the structure of LASNet, which can make the LASNet applied to embedded devices and mobile devices better. In the encoding part of LASNet, we propose the LAS module, which retains and utilize spatial information. This module uses a combination of asymmetric convolution, group convolution, and dual-stream structure to reduce the number of network parameters and maintain strong feature extraction ability. In the decoding part of LASNet, we propose the multivariate concatenate module to reuse the shallow features, which can improve the segmentation accuracy and maintain a high inference speed. Our network attains precise real-time segmentation results in a wide range of experiments. Without additional processing and pre-training, LASNet achieves 70.99% mIoU and 110.93 FPS inference speed in the CityScapes dataset with only 0.8 M model parameters.

Keywords: asymmetric convolution; real-time semantic segmentation; attention mechanism; residual unit

Citation: Chen, Y.; Zhan, W.; Jiang, Y.; Zhu, D.; Guo, R.; Xu, X. LASNet: A Light-Weight Asymmetric Spatial Feature Network for Real-Time Semantic Segmentation. *Electronics* **2022**, *11*, 3238. https://doi.org/10.3390/electronics11193238

Academic Editors: Muhammad Salman Haleem, Liangxiu Han, Ernesto Iadanza and Baihua Li

Received: 10 September 2022
Accepted: 4 October 2022
Published: 9 October 2022

1. Introduction

In recent years, with the rapid development of deep convolution neural networks (DCNN), semantic segmentation is becoming more and more popular, which has been applied to many computer vision tasks [1–3], and significant progress has been made in semantic segmentation. The low-accuracy and slow-speed problems that are difficult to be solved by previous graph theory and pixel clustering methods have been solved. Semantic segmentation technology plays a crucial role in the medical image [4,5], remote sensing mapping [6,7], automatic driving [8,9], and indoor scene [10]. Moreover, with the rapid development of the GPU industry, more complex DCNN models can be realized and applied, and semantic segmentation technology is constantly improved. However, there are three major problems in the semantic segmentation network based on deep learning:

- The semantic segmentation accuracy is high, but the network model parameters are large.
- The semantic segmentation network is lightweight, but the segmentation accuracy is insufficient.
- The semantic segmentation network cannot fully use context information.

There are three solutions to the above problems: reducing the size of the input feature image, improving the convolution block structure, and using the encoder-decoder architecture.

The first method is to reduce the size of the input feature map, such as ENet [11], SegNet [12], and ERFNet [13], which can improve the inference speed but lose some spatial information. The second method is to strengthen the convolution block structure, such as AGLNet [14], GINet [15], and DSANet [16], which can improve the accuracy of semantic segmentation but reduce the inference speed. The third method uses encoder-decoder architecture, such as LRDNet [17], ERFNet, and FSFNet [18]. After the input image is given, DCNN can learn the feature map of the input image through encoder-decoder architecture. The network can gradually realize the category annotation of each pixel, achieve the end-to-end effect, reduce the amount of calculation, and realize fast inference speed and high-quality segmentation accuracy.

To solve these problems, we propose a lightweight asymmetric spatial feature network(LASNet), which can reduce the loss of spatial details, improve the inference speed, and a better balance between speed and accuracy. Moreover, we design a lightweight asymmetric spatial convolution Module(LAS). We use a residual unit with a skipping connection to prevent network degradation and adopt a channel shuffling operation to enhance the robustness of the network. At the same time, we use the encoder-decoder architecture. We validate LASNet on the CityScapes dataset and achieve satisfactory results. Our LASNet has good semantic segmentation accuracy and fast inference speed compared with state-of-the-art methods, as shown in Figure 1.

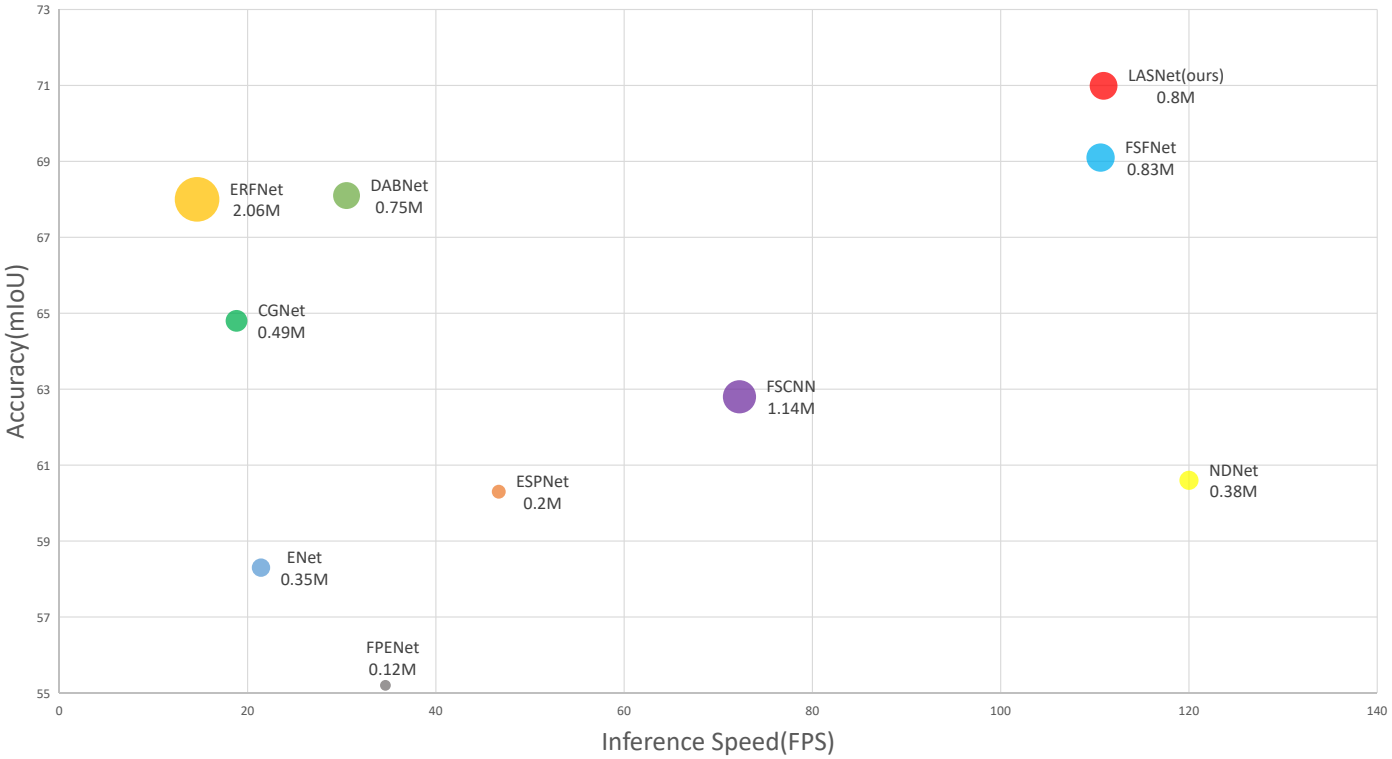


Figure 1. Compare with state-of-the-art network in terms of the available trade-off between accuracy and efficiency. (The circle size represents the parameter).

The main contributions of this paper are as follows:

- We propose a novel deep convolution neural network called LASNet, which adopts an asymmetric encoder-decoder architecture. Through ablation study, the optimal parameters such as module structure, dilation rate, and dropout rate are obtained, which is helpful to build a high-precision and real-time semantic segmentation network.
- To preserve and utilize spatial information, we propose the LAS module, which adopts asymmetric convolution, group convolution, and dual-stream structure to balance inference speed and segmentation accuracy. However, the LAS module's

computational complexity is much lower. In the encoding part of LASNet, which uses the LAS module to process downsampling features, reduce the number of network parameters, and maintain strong feature extraction ability.

- We propose a multivariate concatenate module, which is used by the decoder of LASNet for upsampling. The module can reuse shallow features of images, which helps to improve the segmentation accuracy and maintain a high inference speed.
- We test LASNet on the CityScapes dataset. The comprehensive experiments show that our network achieves available state-of-the-art results in terms of speed and accuracy. Specifically, LASNet achieves 70.99% mean IoU on the CityScapes dataset, with only 0.8 M model parameters and 110.93 FPS inference speed using NVIDIA Titan XP GPU.

The remainder of this paper is structured as follows. In Section 2, related work on semantic segmentation, convolutional factorization and attention mechanism is introduced. Following that, the detailed architecture of LASNet is introduced in Section 3. Furthermore, the experiments can be found in Section 4. Finally, the concluding remarks and future work are given in Section 5.

2. Related Works

Semantic segmentation is a challenging task in computer vision. Especially in the field of automatic driving, low computational complexity and high segmentation accuracy are needed in practical applications. To meet the above requirements, the design and network architecture of CNNs need to be carefully arranged.For example, ResNet [19], VGG [20], Inception [21–23] and MobileNet [24–26], with using deep learning network frameworks for semantic segmentation, which can predict each information of different semantic categories of the image. So that the automatic driving system can judge the environment around them according to the accuracy of the training model based on the pixel level. For example, roads, cars, pedestrians, sidewalks, and buildings. The fully convolutional network (FCN) [27] transforms the classification network into a network structure for segmentation tasks, which proves end-to-end network training on the segmentation problem.

2.1. Semantic Segmentation

In order to improve the semantic segmentation accuracy of automatic driving or intelligent robot, SegNet was proposed by the University of Cambridge. However, this method had a large calculation and low segmentation accuracy. So it was difficult to be used in the field of real-time semantic segmentation. Deeplab-v3+ [28] used the encoder-decoder structure in semantic segmentation and arbitrarily controlled the resolution of the features extracted by the encoder. At the same time, in order to fuse multi-scale information, Deeplab-v3+ used dilated convolution to expand the receptive field. PSPNet [29] considered the global background of the image to generate the prediction at the local level. It is recognized that ENet proposed the first real-time semantic segmentation network, which adopted encoder-decoder architecture to get good segmentation accuracy and inference speed with few model parameters. ERFNet used residual units and deconvolution to maintain efficiency and improve the accuracy of semantic segmentation, which cannot consume too many resources. The Context Guide block proposed by CGNet [30] can obtain the context information and learn local and global features. AGLNet adopted asymmetric encoder-decoder architecture and used a split-shuffle-non-bottleneck unit to generate downlink sampling characteristics while maintaining strong representation ability, AGLNet made the network scale smaller and improved the segmentation accuracy. LMFFNet [31] extracts sufficient features with fewer parameters and fuses multiscale semantic features to effectively improve the segmentation accuracy. SGCPNet [32] uses the spatial details of shallow layers to guide the propagation of the low-resolution global contexts, in which the lost spatial information can be effectively reconstructed. MAFFNet [33] can effectively extract depth features and combine the complementary information in RGB and depth. Like this work, we use a full resolution of 1024×2048 on the CityScapes dataset.

2.2. Convolutional Factorization

At present, most advanced real-time semantic segmentation networks use convolution factorization, which decomposes the standard convolution into several asymmetric convolutions to reduce the computational complexity and improve the depth of the network.

The calculation cost of the standard convolution layer is usually calculated according to the parameters of "Mult-Adds [24]", which can be shown as:

$$MAC_s = K^2 \times H \times W \times In \times Out \quad (1)$$

Convolution factorization usually decomposes a 2D convolution into two asymmetric convolutions (e.g., decompose n $\times$ n to 1 $\times$ n and n $\times$ 1), such as group convolution [34], depth separable convolution [24], and its extended version [35]. The calculation cost of asymmetric convolution is calculated using the same symbol, which can be shown as:

$$MAC_a = 1 \times K \times H \times W \times In \times Out + K \times 1 \times H \times W \times In \times Out \quad (2)$$

where MAC_s is the calculation cost of standard convolution, MAC_a is the calculation cost of asymmetric convolution, K is the kernel size; in the feature map, H is the spatial height, W is the spatial width; In and Out are the number of input channels and output channels, respectively.

Specifically, the group convolution can reduce training parameters and prevent overfitting. The filter is divided into different groups to reduce training parameters and to avoid overfitting, which has been widely used in many real-time semantic segmentation networks. Different from these efficient networks, our proposed LAS module avoids standard convolution and reduces computational complexity. Compared with ShuffleNet [36,37], which is convoluted only half of the input feature channels, our proposed LASNet makes full use of the input channel with multiple convolution branches. The multi-path structure of the LAS improves the feature extraction ability of the network. In addition, our proposed LAS module enhances the information exchange in the feature channel while maintaining the computational cost similar to the standard convolution.

2.3. Attention Mechanism

In recent years, attention mechanism [38,39] has been widely used in various tasks of computer vision, such as image processing, voice recognition, or natural language processing. SENet [40] is divided into two operations with sequence and exception. The purpose of the squeeze operation is actually to extract the spatial information, and the exception operation is used to fully capture the channel correlation. CBAM [41] is a simple and effective attention module for the convolutional neural network. Given an intermediate feature map, CBAM will infer the attention map along two independent dimensions (channel and space). Then, attention mapping is multiplied by the input feature map for adaptive feature optimization. Triplet attention [42] establishes the dependency relationship between dimensions through rotation operation and residual transformation, which encodes the channel and spatial information with negligible computational overhead. Coordinate attention [43] decomposes channel attention into two 1-dimensional feature coding processes to aggregate features along with two spatial directions. Then, the generated feature map is encoded into a pair of direction aware, and position-sensitive attention maps, which can be complementarily applied to the input feature map to enhance the representation of the object of interest. In this work, triplet attention is used in the Transform Module, and it works well.

$$Z_{pool}(x) = [MaxPool_{0d}(x), AvgPool_{0d}(x)] \quad (3)$$

Equation (3) is one of the core operations of triplet attention, where $0d$ is the 0-th dimension of maximum and average pooling operation, and operator [.] means concatenate operation.

3. LASNet

In order to reduce the computational cost and improve the segmentation accuracy, we propose a lightweight asymmetric spatial feature network called LASNet. Firstly, we propose the LAS module, which is the core component for semantic feature extraction in the network. Thereafter, we designed a new transform module, and multivariate concatenate module. In the transform module, the attention mechanism makes the network pay more attention to the essential features of the feature map. The multivariate concatenate module upsamples the feature image to the size of the input image and completes more complex boundary segmentation. Finally, we introduce the architecture of the LASNet, as shown in Figure 2.

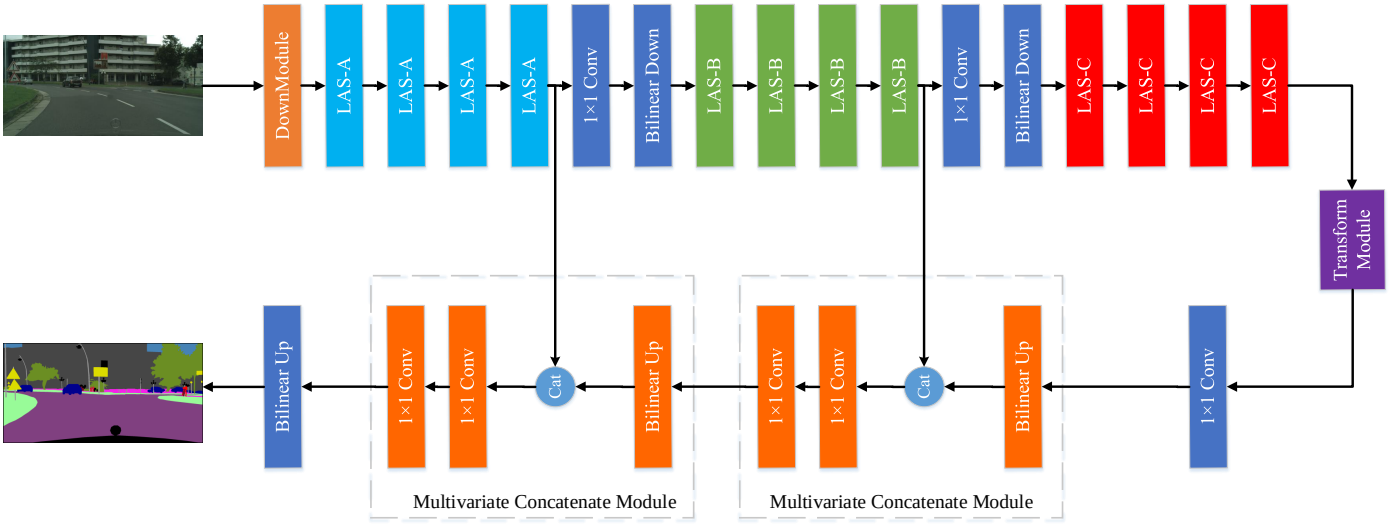
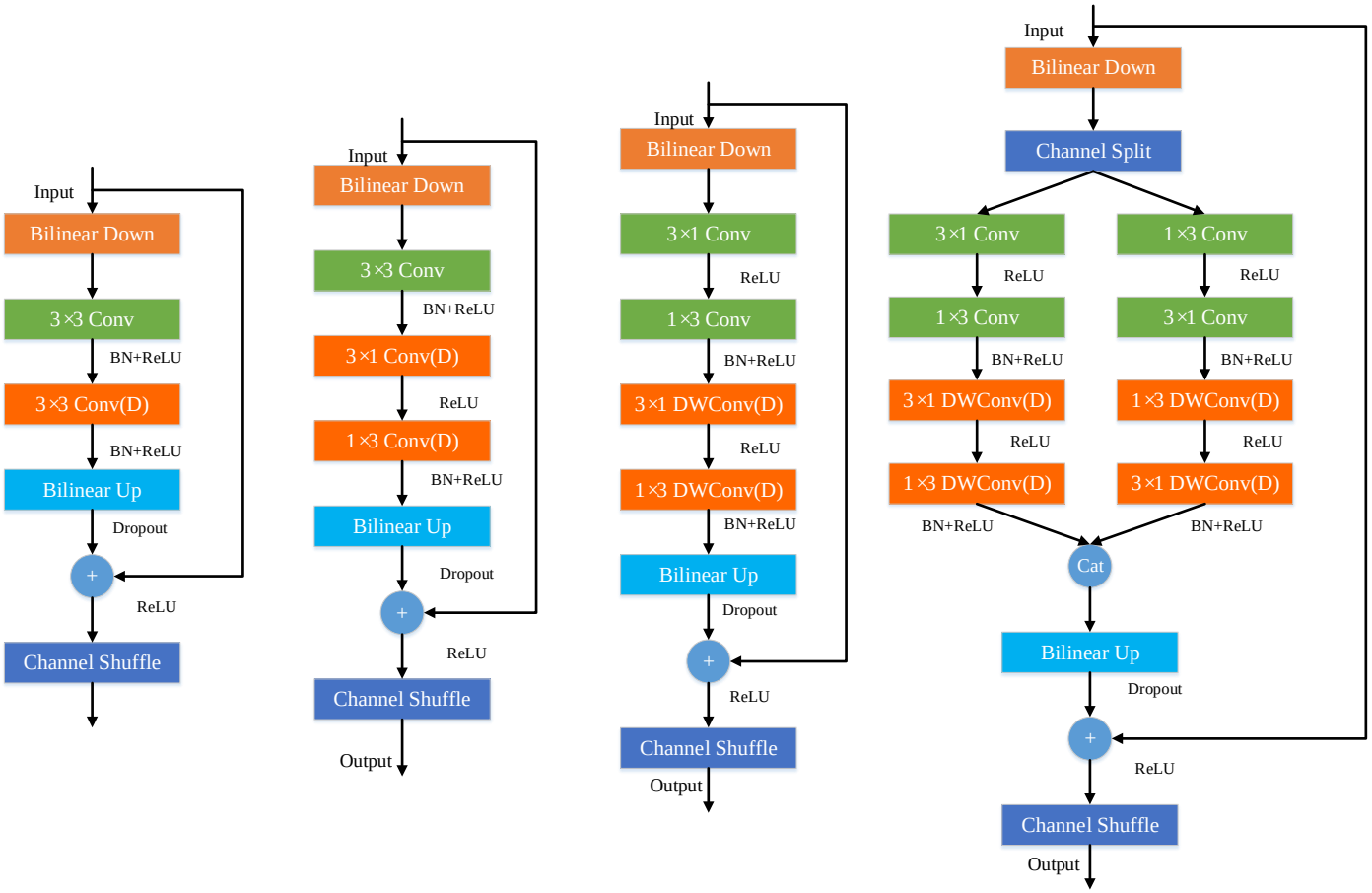


Figure 2. Detailed architecture of LASNet.

3.1. LAS Module

In order to achieve a balance between inference speed and segmentation accuracy, we designed a new module called LAS. Our proposed LAS adopts a residual connection structure, which is to prevent gradient disappearance or gradient explosion by building a high-performance deep network. In order to reduce the computational complexity of convolution, we first adopt downsampling of the feature map in the residual unit to reduce the amount of computation. Then, we use two convolutions to extract the features and upsample the feature map to match the size of the input map. We use bilinear interpolation to upsample and downsample, which is more convenient than convolution. Finally, we use channel shuffle to achieve feature reuse with disrupting channels. We apply asymmetric convolution, group convolution, and dual-stream structure into the residual unit of LAS. There are LAS-A, LAS-B, and LAS-C structures in the LAS module, as shown in Figure 3. Due to this unique design, LASNet has fast inference speed and high accuracy.



(a) LAS-Base (b) LAS-A (c) LAS-B (d) LAS-C

Figure 3. Detailed architecture of LAS Module.

3.1.1. LAS-A Module

In the LAS-A module, the input feature map size and the output feature map size is 128 × 256, and the number of channels is 64. We focus on finding convolution operator combinations with higher accuracy and faster inference speed in this feature map size. We use dilated convolution to integrate the multi-scale context information on the pixel level. Compared with standard convolution, dilated convolution can accommodate a wider receptive field area without increasing the number of parameters. It is very effective to use asymmetric convolution for network structure. 3 × 3 dilated convolution is decomposed into 3 × 1 and 1 × 3 dilated convolution. Although the accuracy decreases slightly, the parameters are reduced by 33%.

3.1.2. LAS-B Module

In the LAS-B module, the input feature map size and the output feature map size is 64 × 128, and the number of channels is 96. In this feature map size, the accuracy of convolution operators with different combinations is not much different. Therefore, we focus on finding convolution operator combinations with faster inference speed. We analyzed various convolution parameters and FLOPs. Furthermore, we found that 3 × 1 and 1 × 3 depth dilated convolution has a fast inference speed and a large receptive field, which is suitable for the structure of the LAS-B module.

3.1.3. LAS-C Module

In the LAS-C module, the input feature map size and the output feature map size is 32 × 64, and the number of channels is 128. We use the dual-stream structure to

extract the features of the feature map and design the LAS-C module. Furthermore, we adopt the split-convolution-cat-shuffle operation, which can reduce computational complexity. At the beginning of the LAS-C module, the number of input channels is evenly divided into two low-dimensional branches by split operation. In order to decrease the computation of standard convolution, we use the asymmetric convolution in the residual unit. The concatenation operation is used to merge the convolution outputs of the two branches so that the number of channels remains the same. Finally, the same channel shuffling operation is used to communicate information between two branches. Channel shuffling can be regarded as feature reuse. With the data flowing to the deepest layer of the network, the network capacity is expanded to a certain extent without significantly increasing the complexity.

3.2. Multivariate Concatenate Module

We use the multivariable concatenate module (MCM) to filter and fuse feature maps of different scales to achieve better prediction accuracy. MCM uses bilinear interpolation for upsampling to recover the size of the feature map. Then, MCM concatenates the channels of the multivariate information feature map to combine the network with multi-scale context information, which makes the structure can effectively improve the performance of the network. Finally, two 1×1 convolution layers are used to increase the number of channels of the feature map so that it concatenates the number of channels of the multivariate feature map. The structure of MCM is shown in Figure 4. The LAS-A module and LAS-B module output is an input of MCM, respectively, and through concatenate operation, bilinear upsampling, and 1×1 convolution sequential processing.

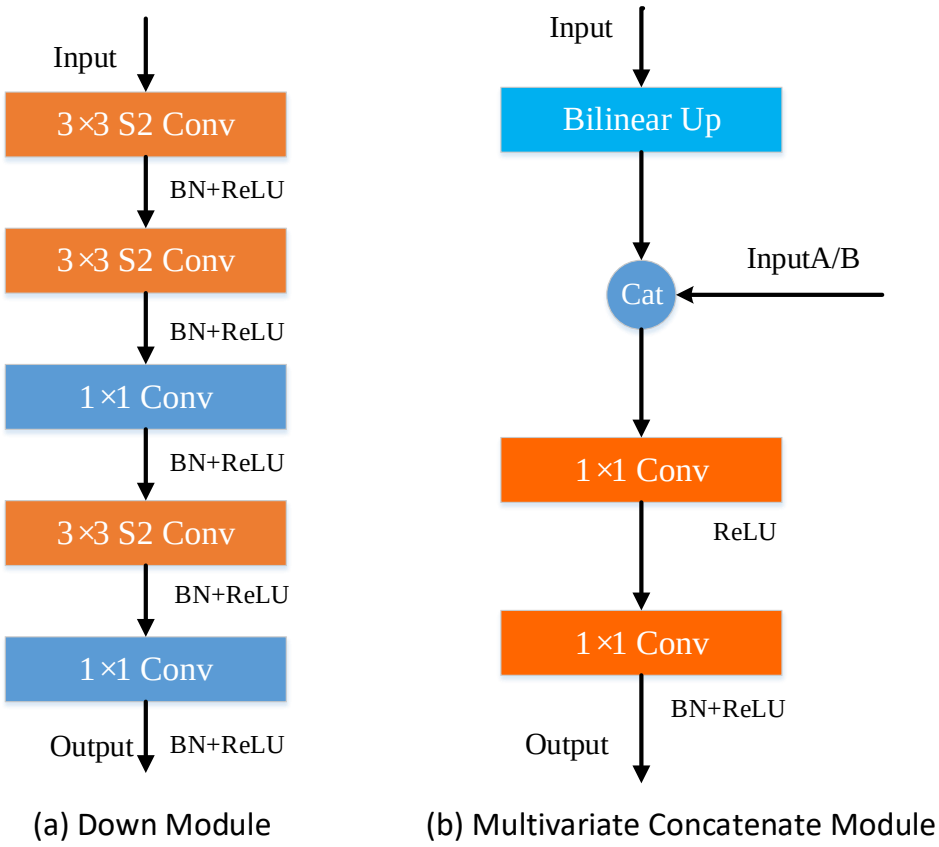


Figure 4. Down sampling module and multivariate concatenate module (S2 conv block denotes a convolution block with a stride of 2).

3.3. Transform Module

The transform module contains three branches. The first branch: channel attention calculation branch; the second branch: channel C and space W dimensional interaction capture branch; the third branch: channel C and space H dimensional interaction capture branch, and finally the output features of the three branches are summed. The structure of the transform module is shown in Figure 5. We use the transform module to extract high-level semantic information from the feature map, which uses the attention mechanism to suppress irrelevant features and focus on the essential features. Meanwhile, the transform module can increase the depth of the network and improve network performance. Subsequently, we perform ablation studies on the transform module to verify the effectiveness of the attention mechanism in the transform module.

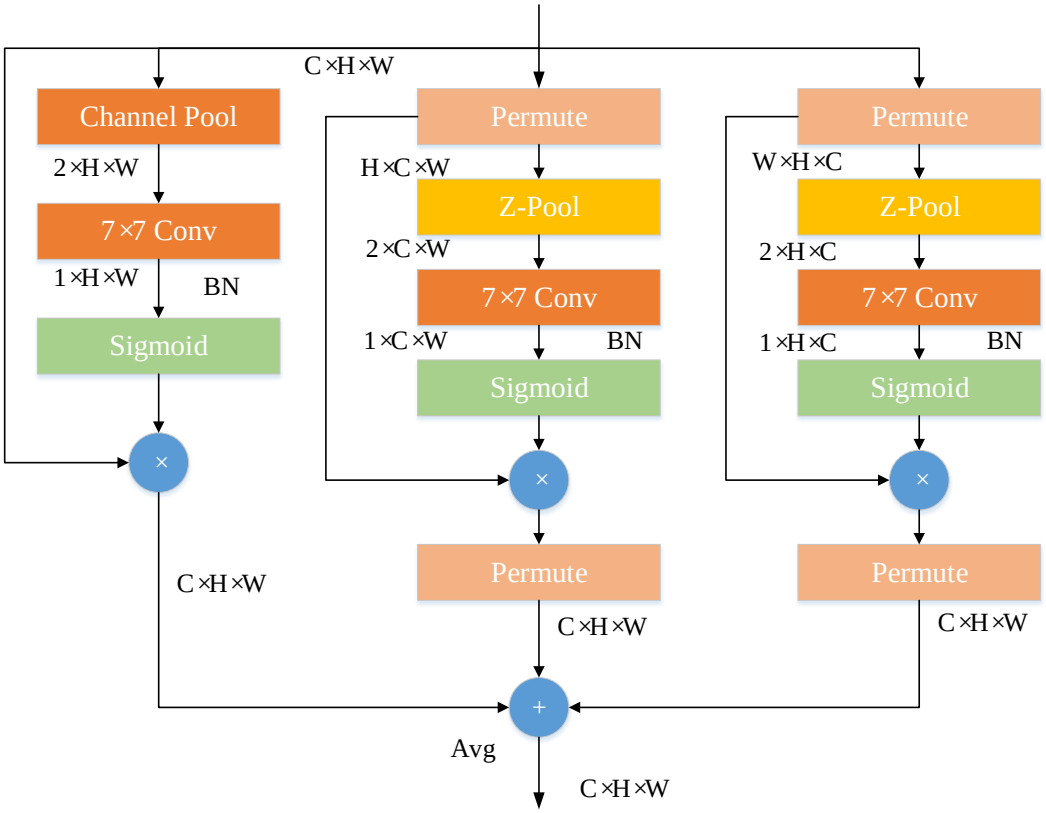


Figure 5. Detailed architecture of transform Module.

3.4. LASNet Architecture

Our LASNet follows a lightweight encoder-decoder architecture. Different from the traditional networks, our LASNet adopts an asymmetric architecture, where an encoder generates downsampled feature maps, and the subsequent decoder upsamples the feature maps to match the input resolution. The detailed structure of our proposed model is shown in Table 1. In our architecture, the first layer is the downsampling module, as shown in Figure 4. Continuous downsampling operations can reduce the size of the feature map in order to extract high-level semantic information. In terms of channel number, the number of channels increases with the downsampling rate. However, we consider the low computational overheads to limit the channel size to 128. We apply LAS-A, LAS-B, and LAS-C modules to feature maps of different scales, and each module is stacked with four. This method can efficiently extract the semantic information of the feature graph according to the size of the feature map.

Table 1. Detailed architecture of proposed LASNet.

Stage	Layer	Type	Out_Channel	Out_Size
Encoder	0	Input Image	3	1024 × 2048
	1	Down Module	64	128 × 256
	2	LAS_A	64	128 × 256
	3	LAS_A	64	128 × 256
	4	LAS_A	64	128 × 256
	5	LAS_A	64	128 × 256
	6	Bilinear Down	64	64 × 128
	7	Convolution 1 × 1	96	64 × 128
	8	LAS_B	96	64 × 128
	9	LAS_B	96	64 × 128
	10	LAS_B	96	64 × 128
	11	LAS_B	96	64 × 128
	12	Bilinear Down	96	32 × 64
	13	Convolution 1 ×1	128	32 × 64
	14	LAS_C	128	32 × 64
	15	LAS_C	128	32 × 64
	16	LAS_C	128	32 × 64
	17	LAS_C	128	32 × 64
Transform Module	18	Triplet Attention	128	32 × 64
Decoder	19	Convolution 1 × 1	32	32 × 64
	20	Bilinear Up(×2)	32	64 × 128
	21	Concat	128	64 × 128
	22	Convolution 1 × 1	32	64 × 128
	23	Convolution 1 × 1	48	64 × 128
	24	Bilinear Up(×2)	48	128 × 256
	25	Concat	112	128 × 256
	26	Convolution 1 × 1	32	128 × 256
	27	Convolution 1 × 1	19	128 × 256
	28	Bilinear Up(×8)	19	1024 × 2048

In addition, the use of dilated convolutions allows our structure to expand the receptive field without losing resolution, and obtain multi-scale context information, which further improves the accuracy. Compared with the larger kernel sizes, this technology can reduce the amount of calculation without introducing additional parameters. We also added dropout to the LAS module to achieve regularization and a slightly improved dropout rate in LAS-B and LAS-C to enhance the regularization effect, which will bring better benefits. We will prove this later in the experiment. Each bilinear interpolation layer in the encoder will pass through a 1 × 1 convolution, which can adjust the number of channels without significantly increasing the number of parameters. In the transform module, the attention mechanism enables the network to suppress irrelevant characteristics and focus on essential features. The multivariable concatenate module of the decoder completes more complex boundary segmentation, which gradually recovers the spatial information lost by multiplexing the shallow features. Because the number of channels in the deep semantic feature map is too large, we use 1 × 1 convolution for dimensionality reduction and feature fusion. Finally, bilinear upsampling is used to recover the resolution step-by-step.

4. Experiment

In this section, we conducted semantic segmentation experiments on the challenging dataset CityScapes [44] to demonstrate the high segmentation accuracy and inference speed of our proposed LASNet. In order to better understand the potential behavior of semantic segmentation networks in machine vision, we also carried out some ablation studies.

4.1. Implement Details

We tested LASNet on the CityScapes dataset, which is a common benchmark for real-time semantic segmentation. The CityScapes dataset has 5000 images from driving scenes in 50 urban environments, which is including of 2975 training images, 500 validation images, and 1525 test images, with the image size of 1024×2048. It has 19 categories of dense pixel annotations. For a fair comparison, we use the original image size 1024×2048 as the input resolution of the CityScapes dataset.

LASNet is trained end-to-end using Adam optimizer for CityScapes dataset. For CityScapes dataset, we prefer a large batchsize (set to 8) to use GPU memory fully. The initial learning rate is set to 1e-3. During our training process, we adopted the "poly" learning rate strategy [29], in which the power of the learning rate is 0.9, and the weight attenuation is set to 5e-4. Furthermore, for the CityScapes dataset, the maximum number of the training epoch is set to 350.

$$lr = lr_{base} \times \left(1 - \frac{epoch}{epoch_{\max}}\right)^{power} \tag{4}$$

In the training process of data enhancement, we use random horizontal flipping and random scaling from 0.5 to 2 for the input image. Finally, we randomly cut the image into a fixed size for training. All images of the CityScapes dataset were normalized to zero mean and unit variance.

4.2. Comparative Experiments

In order to demonstrate the advantages of our network, we selected 11 most advanced lightweight models as comparison networks, including SegNet, ENet, ICNet [45], ESPNet [46], CGNet, ERFNet, DABNet [47], FSCNN [48], FPENet [49], FSFNet, NDNet [50]. The experimental results of some network models are generated using the default parameter settings given by the author, while others are directly reproduced from the published literature. All comparison networks are evaluated and measured by the mean Intersection over Union (mIoU) class score, which is commonly used in the evaluation of semantic segmentation model indicators. mIoU represents the ratio of the intersection and union of the real value to the predicted value. Each class calculates its own IoU as follows:

$$IoU = TP/(FN + FP + TP) \tag{5}$$

TP, FN, and FP, respectively, represent true positive, false negative, and false positive. After calculating the average value of IoU of each class, the main evaluation indicator mIoU of semantic segmentation is obtained, and the expression is as follows:

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{6}$$

where i represents the real value, and $k+1$ represents the number of categories (including empty categories).

4.3. Analysis of CityScapes Evaluation Results

For the fairness of experimental data, all comparison networks will use the same hardware platform and NVIDIA Titan XP GPU for training. Table 2 compares our LASNet with selected state-of-the-art networks. Experimental data shows that LASNet is superior to these networks, with high classification accuracy and high inference speed. In these

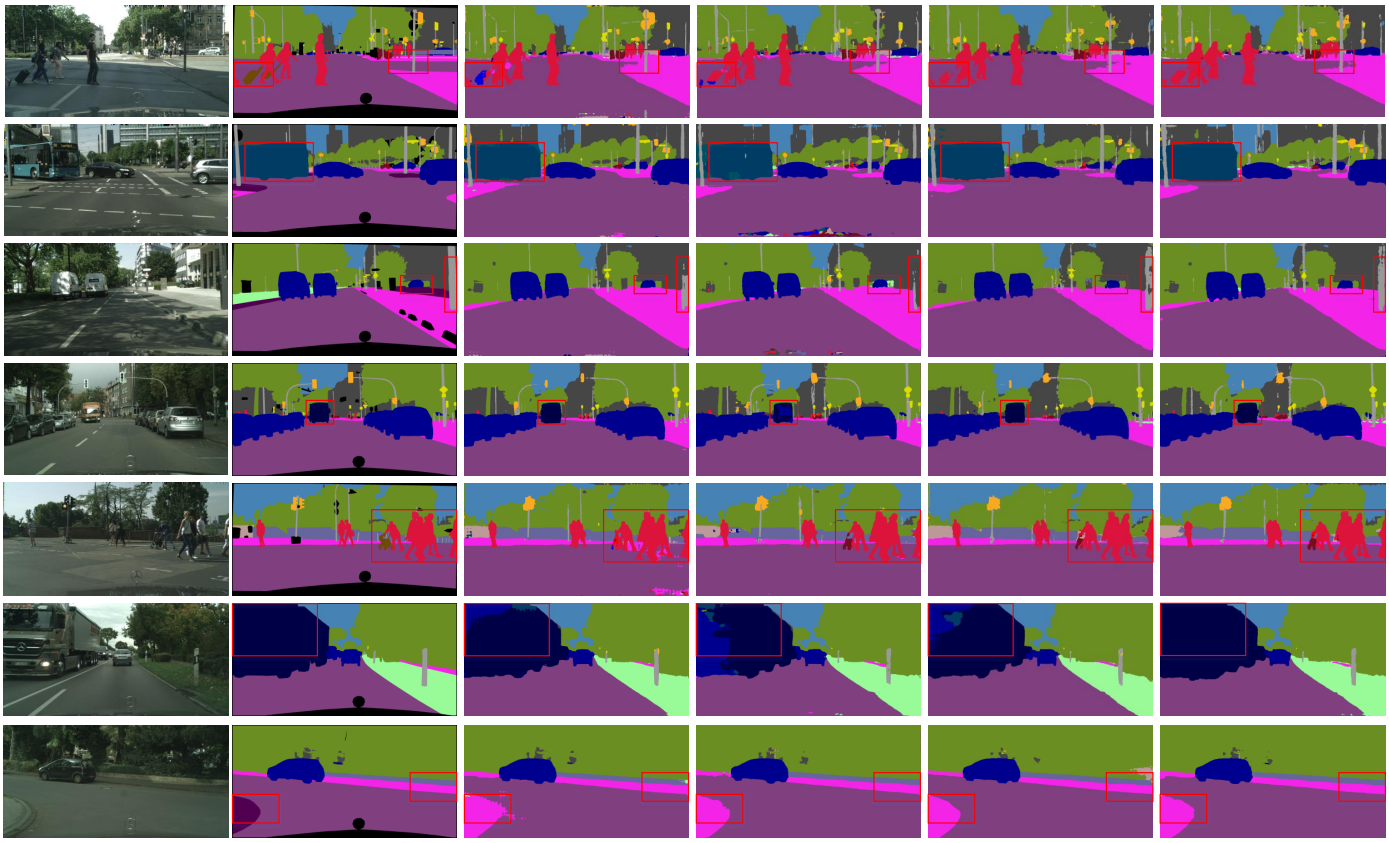
methods, our proposed LASNet has only 0.8 M network parameters without pre-training in ImageNet, which has 110.93FPS inference speed and 70.99% mIoU. As can be seen from the experimental data in Table 2, LASNet still has 110.93FPS in terms of inference speed when the size of the input feature map is 1024 × 2048. Compared with FSFNet, the segmentation accuracy of LASNet is higher than 1.8%. The inference speed of other lightweight networks is similar to our LASNet. However, the segmentation accuracy is low. For example, the network model parameters of FPENet are only 0.13 M, and the inference speed is 110 FPS, but the segmentation accuracy is 15% lower than that of our LASNet. We also compare with some relatively large networks and give Table 2, the detailed IoU of each class is shown in Table 3. The results show that compared with ERFNet and ICNet, our LASNet has similar segmentation accuracy, but the inference speed is lower than 70–90 FPS. Figure 6 shows the results of the CityScapes dataset after these comparative network segmentation. The experimental results show that compared with these networks, our proposed LASNet has higher accuracy and faster inference speed for different scales of target segmentation, which proves the advanced level of our network.

Table 2. Compared with the state-of-the-art approaches on the CityScapes. The red color font indicates the optimal result.

Method	Input Size	Params (M)	FPS	FLOPs (G)	mIoU (%)
SegNet	512 × 1024	29.45	4.50	326.26	57.00
ENet	1024 × 2048	0.35	21.42	21.76	58.30
ICNet	1024 × 2048	26.72	24.61	87.83	68.50
ESPNet	1024 × 2048	0.20	46.68	16.65	60.30
CGNet	1024 × 2048	0.49	18.82	27.73	64.80
ERFNet	1024 × 2048	2.06	14.63	120.22	68.00
DABNet	1024 × 2048	0.75	30.51	41.50	68.10
FSCNN	1024 × 2048	1.14	72.25	6.94	62.80
FPENet	1024 × 2048	0.12	34.64	6.17	55.20
FSFNet	1024 × 2048	0.83	110.61	13.47	69.10
NDNet	1024 × 2048	0.38	120.00	2.01	60.60
LASNet(Ours)	1024 × 2048	0.80	110.93	11.90	70.99

Table 3. Compared with other approaches on the CityScapes in terms of per-class results. The red color font indicates the optimal result.

Method	Roa	Sid	Bui	Wal	Fen	Pol	Lig	Sig	Veg	Ter
ENet	96.30	74.20	85.00	32.10	33.20	43.40	34.10	44.00	88.60	61.40
ERFNet	97.71	81.00	89.80	42.50	48.00	56.20	59.80	65.31	91.40	68.20
CGNet	95.90	73.90	89.90	43.90	46.00	52.90	55.90	63.80	91.70	68.30
ESPNet	95.70	73.30	86.60	32.80	36.40	47.00	46.90	55.40	89.80	66.00
FSFNet	97.70	81.10	90.20	41.70	47.00	54.10	61.10	65.30	91.80	69.40
NDNet	96.60	75.20	87.20	44.20	46.10	29.60	40.40	53.30	87.40	57.90
LASNet	97.18	80.34	89.15	64.59	58.89	48.62	48.54	62.60	89.95	62.05
Method	**Sky**	**Per**	**Rid**	**Car**	**Tru**	**Bus**	**Tra**	**Mot**	**Bic**	**mIoU**
ENet	90.60	65.50	38.40	90.60	36.90	50.50	48.10	38.80	55.40	58.30
ERFNet	94.21	76.80	57.10	92.82	50.80	60.10	51.80	47.30	61.70	68.00
CGNet	94.10	76.70	54.20	91.30	41.30	55.90	32.80	41.10	60.90	64.80
ESPNet	92.50	68.50	45.90	89.90	40.00	47.70	40.70	36.40	54.90	60.30
FSFNet	94.20	77.80	57.80	92.80	47.30	64.40	59.40	53.10	66.20	69.10
NDNet	90.20	62.60	41.60	88.50	57.80	67.30	35.10	31.90	59.40	60.60
LASNet	91.84	70.83	51.38	91.10	77.39	81.72	69.22	48.02	65.84	70.99



(a)input images (b)ground truth (c)CGNet (d)FSFNet (e)NDNet (f)LASNet

Figure 6. Visual Detail results on CityScapes validation dataset. From left to right are input images, ground truth, segmentation outputs from CGNet, FSFNet, NDNet, and our LASNet.

4.4. Ablation Study

4.4.1. LAS Module Structure

In order to prove the effectiveness of our proposed LAS module, we used the CityScapes dataset to ablation the LAS module and combined LAS-base, LAS-A, LAS-B, and LAS-C into our network. Table 4 analyzes the contribution of each combination to LASNet performance. It can be observed that the introduction of different LAS modules can improve the segmentation accuracy. Compared with the basic module, the LAS module concatenates the semantics of high-level features and the spatial details of low-level features to improve performance. The mIoU results using the combination of LAS-A and LAS-B reached 71.39%, which is 1.74% higher than the baseline. On the other hand, the combination of LAS-A, LAS-B, and LAS-C was 1.34% higher than the baseline, and the segmentation accuracy was 70.99%. Considering FPS, parameters, and FLOPs, we finally chose the combination of LAS-A, LAS-B, and LAS-C, which has the smallest parameters and FLOPs, faster inference speed, and higher segmentation accuracy.

Table 4. Ablation studies of different module structures. The red color font indicates the optimal result.

Module Structure	FPS	Params (M)	FLOPs (G)	mIoU (%)
Structure1:base·base·base	79.06	2.27	13.36	69.65
Structure2:base·LAS-B·LAS-C	111.58	0.85	12.15	69.86
Structure3:LAS-A·base·LAS-C	95.03	1.24	12.65	67.51
Structure4:LAS-A·LAS-B·base	111.54	1.78	12.36	71.39
Structure5:LAS-A·LAS-B·LAS-C	110.93	0.80	11.90	70.99

4.4.2. LAS Module Number

Our LASNet design architecture uses different numbers of LAS modules, and we verified the impact of different numbers of each LAS module on the LASNet performance. Table 5 analyzes the contribution of the number of each LAS module on the performance of LASNet. It can be seen that the introduction of different numbers of LAS modules can improve the accuracy of segmentation. The increased number of different LASmodules can improve the performance compared to the basic module. However, when the number of LAS Modules is much or little, it has a negative effect. When the number of LAS Modules is 2, the inference speed is fast but the quality is too low, Furthermore, when the number of LAS Modules is 6, the accuracy is reduced. Finally, considering the FPS, parameters and FLOPs, we finally choose the combination of LAS-A, LAS-B and LAS-C with the number of all 4.

Table 5. Ablation studies of different module numbers. The red color font indicates the optimal result.

Module Numbers	FPS	Params (M)	FLOPs (G)	mIoU (%)
LAS-A:2·LAS-B:2·LAS-C:2	123.44	0.33	9.27	66.32
LAS-A:3·LAS-B:3·LAS-C:3	120.89	0.57	10.36	68.45
LAS-A:5·LAS-B:5·LAS-C:5	101.78	2.25	13.25	71.86
LAS-A:6·LAS-B:6·LAS-C:6	90.22	2.85	14.50	71.46
LAS-A:3·LAS-B:4·LAS-C:5	98.13	1.31	12.56	69.31
LAS-A:5·LAS-B:4·LAS-C:3	101.24	1.48	12.33	70.32
LAS-A:4·LAS-B:4·LAS-C:4	110.93	0.80	11.90	70.99

4.4.3. Dilation Rate

We use dilated convolution [51] to expand the receptive field and aggregate semantic information to realize the flexible fusion of multi-scale context information in the LAS module. There are four expansion convolution layers in LAS-A, LAS-B, and LAS-C blocks. We performed ablation experiments on the dilation rate, which are {1, 1, 1, 1}, {1, 2, 3, 4}, {1, 2, 5, 9} and {1, 2, 4, 8}. We compare the network segmentation accuracy in these four cases. The experimental results are given in Table 6. We can find that according to the use of dilated convolution, the segmentation accuracy of CityScapes dataset can differ by up to 3%. Therefore, in the structure of LASNet, we use the dilation rate of 1, 2, 4, 8 to achieve the best segmentation accuracy.

Table 6. Ablation studies of different dilation rates. The red color font indicates the optimal result.

Dilation Rates	mIoU (%)
Dilation1:LAS-A:(1,1,1,1) LAS-B:(1,1,1,1) LAS-C:(1,1,1,1)	67.92
Dilation2:LAS-A:(1,2,3,4) LAS-B:(1,2,3,4) LAS-C:(1,2,3,4)	70.59
Dilation3:LAS-A:(1,2,5,9) LAS-B:(1,2,5,9) LAS-C:(1,2,5,9)	70.54
Dilation4:LAS-A:(1,2,4,8) LAS-B:(1,2,4,8) LAS-C:(1,2,4,8)	70.99

4.4.4. Dropout Rate

In this section, we will show to select the appropriate dropout rate to improve the segmentation accuracy of LASNet. In Table 7, we analyze the impact of dropout on performance in the LAS module and modify the dropout rate in each LAS block. The experimental results show that the use of a dropout rate is effective, and the segmentation accuracy of the CityScapes dataset can differ by 0.49%. In particular, the dropout rate of the LAS module gradually increases from 0.01, which shows the best segmentation performance. Because dropout in the LAS module can simplify the model, improve the regularization effect, and model generalization force, which can avoid overfitting. Therefore, the dropout rate increased from small to large is suitable for our architecture and shows good performance.

Table 7. Ablation studies of different dropout rates. The red color font indicates the optimal result.

Dropout Rates	mIoU (%)
Dropout1:LAS-A:(0.00,0.00,0.00,0.00) LAS-B:(0.00,0.00,0.00,0.00) LAS-C:(0.00,0.00,0.00,0.00)	70.85
Dropout2:LAS-A:(0.01,0.01,0.01,0.01) LAS-B:(0.01,0.01,0.01,0.01) LAS-C:(0.01,0.01,0.01,0.01)	70.91
Dropout3:LAS-A:(0.01,0.02,0.03,0.04) LAS-B:(0.01,0.02,0.03,0.04) LAS-C:(0.01,0.02,0.03,0.04)	70.50
Dropout4:LAS-A:(0.01,0.02,0.03,0.04) LAS-B:(0.05,0.06,0.07,0.08) LAS-C:(0.05,0.06,0.07,0.08)	70.99

4.4.5. Transform Module

We use the attention mechanism in the transform module to alleviate the contradiction between model complexity and expression ability. With the help of the way the human brain processes information overload, the spatial attention is used as the critical part with greater weight so that the model's attention can be more focused on this part. In Table 8, we compare four cases in the transform module of LASNet: no-transform module, CBAM, triplet attention, and coordinate attention. Extensive experimental results show that selecting the appropriate attention mechanism is effective. The segmentation accuracy of CityScapes dataset can differ by 0.36%. In particular, the use of triplet attention in the transform module shows the best segmentation performance. This is because the dependency between dimensions is established through rotation operation and residual transformation, which can encode the channel and spatial information. Therefore, triplet attention is more suitable for our architecture. Figure 7 shows the IoU of the CityScapes dataset after segmentation by all ablation studies.

Table 8. Ablation studies of different transform modules. The red color font indicates the optimal result.

Transform	mIoU (%)
Not use	70.84
CBAM	70.63
Coordinate	70.47
Triplet	70.99

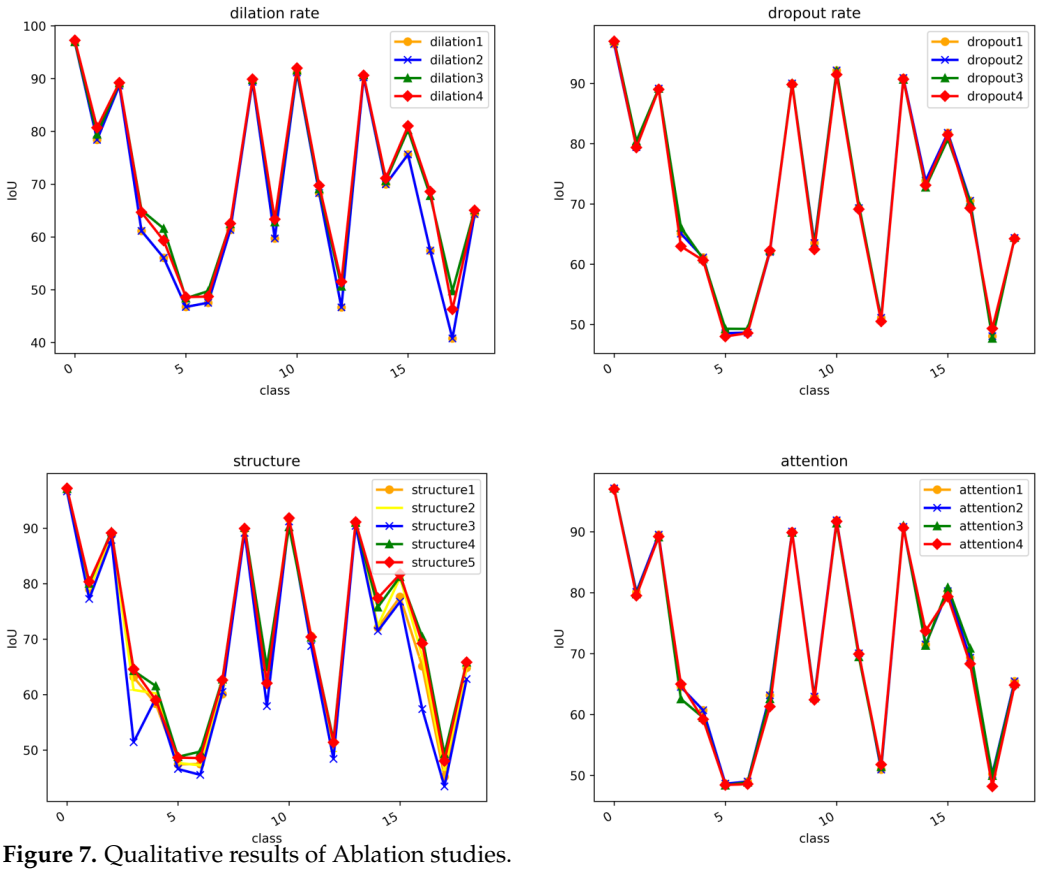


Figure 7. Qualitative results of Ablation studies.

5. Conclusions

This paper describes a lightweight asymmetric spatial feature network (LASNet), which is an encoder-decoder network for real-time semantic segmentation of automatic driving. The encoder adopts channel splitting and shuffling operations in the residual unit, which strengthens information exchange in the way of feature reuse. LAS-A, LAS-B, and LAS-C modules quickly extract the semantic information of the feature map according to the size of the feature map. Then, the attention mechanism in the transform module makes the network pay more attention to the semantic features of the feature map. Finally, the multivariable concatenate module of the decoder completes more complex boundary segmentation and gradually recovers the spatial information lost by the encoder due to the reduction of the size of the feature map. The entire network proves end-to-end network training. To evaluate our network, we conducted experiments on popular datasets. The experimental results show that our LASNet is better than the comparative SOTA network in the segmentation accuracy and efficiency of the urban landscape dataset. In the future, we will strive to quantify the model parameters and deploy them in embedded devices.

Author Contributions: Conceptualization, Y.C.; Data curation, Y.J. and D.Z.; Formal analysis, D.Z.; Funding acquisition, Y.J.; Investigation, D.Z., R.G. and X.X.; Methodology, Y.C.; Project administration, Y.J.; Resources, W.Z.; Software, X.X.; Supervision, W.Z. and R.G.; Validation, Y.C.; Visualization, Y.C.; Writing—original draft, Y.C.; Writing—review & editing, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by Jilin Province Development and Reform Commission (Grant No. FG2021236JK).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Conflicts of Interest: We declare that we have no financial and personal relationships that can influence our work.

References

1. Jiang, Y.; Liu, Y.; Zhan, W.; Zhu, D. Lightweight Dual-Stream Residual Network for Single Image Super-Resolution. *IEEE Access* **2021**, *9*, 129890–129901. [CrossRef]
2. Zhu, D.; Zhan, W.; Jiang, Y.; Xu, X.; Guo, R. MIFFuse: A Multi-Level Feature Fusion Network for Infrared and Visible Images. *IEEE Access* **2021**, *9*, 130778–130792. [CrossRef]
3. Zhu, D.; Zhan, W.; Jiang, Y.; Xu, X.; Guo, R. IPLF: A Novel Image Pair Learning Fusion Network for Infrared and Visible Image. *IEEE Sens. J.* **2022**, *22*, 8808–8817. [CrossRef]
4. Luo, X.; Wang, G.; Song, T.; Zhang, J.; Aertsen, M.; Deprest, J.; Ourselin, S.; Vercauteren, T.; Zhang, S. MIDeepSeg: Minimally interactive segmentation of unseen objects from medical images using deep learning. *Med. Image Anal.* **2021**, *72*, 102102. [CrossRef]
5. Feng, R.; Zheng, X.; Gao, T.; Chen, J.; Wang, W.; Chen, D.Z.; Wu, J. Interactive Few-shot Learning: Limited Supervision, Better Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2021**, *40*, 2575–2588. [CrossRef]
6. Cui, W.; He, X.; Yao, M.; Wang, Z.; Hao, Y.; Li, J.; Wu, W.; Zhao, H.; Xia, C.; Li, J.; et al. Knowledge and Spatial Pyramid Distance-Based Gated Graph Attention Network for Remote Sensing Semantic Segmentation. *Remote Sens.* **2021**, *13*, 1312. [CrossRef]
7. Li, Y.; Shi, T.; Zhang, Y.; Chen, W.; Wang, Z.; Li, H. Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation. *ISPRS J. Photogramm. Remote Sens.* **2021**, *175*, 20–33. [CrossRef]
8. Lv, Q.; Sun, X.; Chen, C.; Dong, J.; Zhou, H. Parallel complement network for real-time semantic segmentation of road scenes. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 4432–4444. [CrossRef]
9. Dong, G.; Yan, Y.; Shen, C.; Wang, H. Real-time high-performance semantic image segmentation of urban street scenes. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 3258–3274. [CrossRef]
10. Chen, X.T.; Li, Y.; Fan, J.H.; Wang, R. RGAM: A novel network architecture for 3D point cloud semantic segmentation in indoor scenes. *Inf. Sci.* **2021**, *571*, 87–103. [CrossRef]
11. Paszke, A.; Chaurasia, A.; Kim, S.; Culurciello, E. Enet: A deep neural network architecture for real-time semantic segmentation. *arXiv* **2016**, arXiv:1606.02147.
12. Badrinarayanan, V.; Kendall, A.; Cipolla, R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
13. Romera, E.; Alvarez, J.M.; Bergasa, L.M.; Arroyo, R. Erfnet: Efficient residual factorized convnet for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* **2017**, *19*, 263–272. [CrossRef]
14. Zhou, Q.; Wang, Y.; Fan, Y.; Wu, X.; Zhang, S.; Kang, B.; Latecki, L.J. AGLNet: Towards real-time semantic segmentation of self-driving images via attention-guided lightweight network. *Appl. Soft Comput.* **2020**, *96*, 106682. [CrossRef]
15. Wu, T.; Lu, Y.; Zhu, Y.; Zhang, C.; Wu, M.; Ma, Z.; Guo, G. GINet: Graph interaction network for scene parsing. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 34–51.
16. Elhassan, M.A.; Huang, C.; Yang, C.; Munea, T.L. DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes. *Expert Syst. Appl.* **2021**, *183*, 115090. [CrossRef]
17. Zhuang, M.; Zhong, X.; Gu, D.; Feng, L.; Zhong, X.; Hu, H. LRDNet: A lightweight and efficient network with refined dual attention decorder for real-time semantic segmentation. *Neurocomputing* **2021**, *459*, 349–360. [CrossRef]
18. Kim, M.; Park, B.; Chi, S. Accelerator-aware fast spatial feature network for real-time semantic segmentation. *IEEE Access* **2020**, *8*, 226524–226537. [CrossRef]
19. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
20. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
21. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1–9.
22. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
23. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A.A. Inception-v4, inception-resnet and the impact of residual connections on learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017.
24. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.

25. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
26. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
27. Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3431–3440.
28. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 801–818.
29. Zhao, H.; Shi, J.; Qi, X.; Wang, X.; Jia, J. Pyramid scene parsing network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2881–2890.
30. Wu, T.; Tang, S.; Zhang, R.; Cao, J.; Zhang, Y. Cgnet: A light-weight context guided network for semantic segmentation. *IEEE Trans. Image Process.* **2020**, *30*, 1169–1179. [CrossRef]
31. Shi, M.; Shen, J.; Yi, Q.; Weng, J.; Huang, Z.; Luo, A.; Zhou, Y. LMFFNet: A Well-Balanced Lightweight Network for Fast and Accurate Semantic Segmentation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**, 1–5. [CrossRef]
32. Hao, S.; Zhou, Y.; Guo, Y.; Hong, R.; Cheng, J.; Wang, M. Real-Time Semantic Segmentation via Spatial-Detail Guided Context Propagation. *IEEE Trans. Neural Netw. Learn. Syst.* **2022**. [CrossRef]
33. Lv, T.; Zhang, Y.; Luo, L.; Gao, X. MAFFNet: Real-time multi-level attention feature fusion network with RGB-D semantic segmentation for autonomous driving. *Appl. Opt.* **2022**, *61*, 2219–2229. [CrossRef]
34. Huang, G.; Liu, S.; Van der Maaten, L.; Weinberger, K.Q. Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2752–2761.
35. Mehta, S.; Rastegari, M.; Shapiro, L.; Hajishirzi, H. Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seoul, Korea, 27 October–2 November 2019; pp. 9190–9200.
36. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
37. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
38. Zhou, Z.; Zhou, Y.; Wang, D.; Mu, J.; Zhou, H. Self-attention feature fusion network for semantic segmentation. *Neurocomputing* **2021**, *453*, 50–59. [CrossRef]
39. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 603–612.
40. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
41. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
42. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Nashville, TN, USA, 20–25 June 2021; pp. 3139–3148.
43. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
44. Cordts, M.; Omran, M.; Ramos, S.; Rehfeld, T.; Enzweiler, M.; Benenson, R.; Franke, U.; Roth, S.; Schiele, B. The cityscapes dataset for semantic urban scene understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 3213–3223.
45. Zhao, H.; Qi, X.; Shen, X.; Shi, J.; Jia, J. Icnet for real-time semantic segmentation on high-resolution images. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 405–420.
46. Mehta, S.; Rastegari, M.; Caspi, A.; Shapiro, L.; Hajishirzi, H. Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 552–568.
47. Li, G.; Yun, I.; Kim, J.; Kim, J. Dabnet: Depth-wise asymmetric bottleneck for real-time semantic segmentation. *arXiv* **2019**, arXiv:1907.11357.
48. Poudel, R.P.; Liwicki, S.; Cipolla, R. Fast-scnn: Fast semantic segmentation network. *arXiv* **2019**, arXiv:1902.04502.
49. Liu, M.; Yin, H. Feature pyramid encoding network for real-time semantic segmentation. *arXiv* **2019**, arXiv:1909.08599.

50. Yang, Z.; Yu, H.; Fu, Q.; Sun, W.; Jia, W.; Sun, M.; Mao, Z.H. Ndnet: Narrow while deep network for real-time semantic segmentation. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 5508–5519. [CrossRef]
51. Chen, L.C.; Papandreou, G.; Schroff, F.; Adam, H. Rethinking atrous convolution for semantic image segmentation. *arXiv* **2017**, arXiv:1706.05587.