Hindawi
Mathematical Problems in Engineering
Volume 2023, Article ID 9762309, 1 page
https://doi.org/10.1155/2023/9762309

Retraction

Retracted: DARSegNet: A Real-Time Semantic Segmentation Method Based on Dual Attention Fusion Module and Encoder-Decoder Network

Mathematical Problems in Engineering

Received 19 September 2023; Accepted 19 September 2023; Published 20 September 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] Y. Xing, L. Zhong, and X. Zhong, "DARSegNet: A Real-Time Semantic Segmentation Method Based on Dual Attention Fusion Module and Encoder-Decoder Network," *Mathematical Problems in Engineering*, vol. 2022, Article ID 6195148, 10 pages, 2022.

Hindawi
Mathematical Problems in Engineering
Volume 2022, Article ID 6195148, 10 pages
https://doi.org/10.1155/2022/6195148

Research Article

DARSegNet: A Real-Time Semantic Segmentation Method Based on Dual Attention Fusion Module and Encoder-Decoder Network

Yongfeng Xing,[1,2] **Luo Zhong**,[1] **and Xian Zhong**[1]

[1]*School of Computer Science and Artificial Intelligence, Wuhan University of Technology, Wuhan 430070, China*
[2]*School of Software, Nanyang Institute of Technology, Nanyang 473000, China*

Correspondence should be addressed to Yongfeng Xing; xingyongfeng@whut.edu.cn

Received 27 April 2022; Revised 17 May 2022; Accepted 21 May 2022; Published 6 June 2022

Academic Editor: Hangjun Che

The convolutional neural network achieves excellent semantic segmentation results in artificially annotated datasets with complex scenes. However, semantic segmentation methods still suffer from several problems such as low use rate of the features, high computational complexity, and being far from practical real-time application, which bring about challenges for the image semantic segmentation. Two factors are very critical to semantic segmentation task: global context and multilevel semantics. However, generating these two factors will always lead to high complexity. In order to solve this, we propose a novel structure, dual attention fusion module (DAFM), by eliminating structural redundancy. Unlike most of the existing algorithms, we combine the attention mechanism with the depth pyramid pool module (DPPM) to extract accurate dense features for pixel labeling rather than complex expansion convolution. Specifically, we introduce a DPPM to execute the spatial pyramid structure in output and combine the global pool method. The DAFM is introduced in each decoder layer. Finally, the low-level features and high-level features are fused to obtain semantic segmentation result. The experiments and visualization results on Cityscapes and CamVid datasets show that, in real-time semantic segmentation, we have achieved a satisfactory balance between accuracy and speed, which proves the effectiveness of the proposed algorithm. In particular, on a single 1080ti GPU computer, ResNet-18 produces 75.53% MIoU at 70 FPS on Cityscapes and 73.96% MIoU at 109 FPS on CamVid.

1. Introduction

These years, convolutional neural network is making great progress for semantic image segmentation. Semantic segmentation is a basic topic in the field of computer vision. It is a pixel-level classification and plays an important role in the fields of automatic driving, video surveillance, geographic information system, medical image analysis, and so on [1], [2]. Traditional segmentation methods are limited by feature extraction methods, and the image segmentation effect is poor in complex scenes. The convolutional neural network achieves good segmentation results in artificially annotated datasets with complex scenes [3]. However, recent semantic segmentation methods still suffer from several problems such as low use rate of the features, high computational complexity, and being far from practical application, which bring about challenges for the image semantic segmentation field. Current research mainly focuses on two aspects: applying different network structures to improve the segmentation accuracy and reducing network parameters and computational overheads to meet the real-time requirements with a relatively real-time segmentation accuracy [4].

Real-time segmentation algorithm has attracted more and more attention. Recently, some new real-time semantic segmentation algorithms have been proposed. There are two methods. One is to use GPU efficient backbone, especially ResNet-18, MobileNet, and so forth. Other algorithms have developed complex lightweight coders trained from scratch, and one algorithm BiseNet [5] has reached a new peak in real-time performance. In short, the current mainstream

semantic segmentation framework has some defects, which cannot meet the good balance of high speed and high precision simultaneously. In this paper, we propose a dual attention network with deep high-resolution representation.

Figure 1 shows a comparison of speed and MIoU on the Cityscapes [6] test set. Red color refers to our methods, while green color refers to other methods. We achieve a good speed-accuracy trade-off.

In practical applications such as automatic driving, robotics, and security monitoring, real-time segmentation may be more valuable than accurate segmentation. The lightweight network model aims to reduce the complexity of parameters of the neural network model, while maintaining the accuracy of the model. Lightweight networks not only include in-depth research on network structure but also include the application of model compression technologies such as knowledge extraction and pruning. Together, they promote the application of convolutional neural network technology in mobile terminals and embedded terminals and make due contributions to the development of all walks of life [7–9].

However, the current mainstream semantic segmentation framework has some defects in real-time semantic segmentation field, which cannot meet the good balance between high speed and high precision simultaneously. At present, deep learning has excellent results in various image processing tasks, but a large number of redundant parameter calculations seriously hinder its use in practical projects. It is difficult to comply with real-time requirements in both mobile terminals and embedded devices [2]. For example, the parameters of ResNet-101 are more than 170 MB of storage resources. For instance, images with a 224×224 resolution require more than 7.6 billion floating-point operations, and the parameter memory consumption is 170 MB. This will seriously affect the user's personal experience. Therefore, it is particularly urgent to design a lightweight and efficient neural network.

Compared with the network based on pyramid structure, multibranch network will not increase the output resolution of high-level feature map by changing the reference network. Its operation speed will be faster, but there is a defect that makes it difficult to be applied to real-time semantic segmentation; that is, the contradiction between its spatial branch depth and speed is difficult to coordinate.

At present, the real-time algorithms based on multibranch networks use relatively simple high-resolution branches. Although they run fast, their segmentation accuracy is low. Some branch information will be extracted in different network contexts. The deep branches of the network use separable convolution and other lightweight operations to obtain semantic context information, and the shallow branches use convolution to retain effective spatial details. The network model with this structure is lighter and promotes the real-time application of semantic segmentation, but it is difficult to extract effective semantic context information. In addition, there is a large gap between the two pieces of feature information, so the fusion cannot produce good results.

Although real-time semantic segmentation has made good progress, there are still three main problems [4]. Firstly, the image may contain similar objects with different scales, such as cars and houses. How to capture and integrate different proportions of image features is very important for semantic segmentation. In the mainstream semantic segmentation framework, image classification network is usually used to extract features, while pyramid feature fusion is used to extract multiscale feature information, such as spatial pyramid pooling module. In this case, a lot of computing resources are generally required. Secondly, the multiscale context extraction module of this pyramid method is not flexible enough and needs to manually set the kernel size, so it can only extract a limited feature scale range, which is not conducive to the learning of network semantic features. Finally, the deep convolution neural network has a hierarchical structure, and the characteristics of different levels are different. The high level has rich semantics but lacks accurate location information, while the low one contains spatial detail information but lacks discriminative semantic features. Because semantic segmentation involves object positioning, there are different levels of feature fusion. If the information flow in the model is not well controlled, some redundant features, including background noise in low level and rough boundary in high level, will be introduced into subsequent features and may lead to network performance degradation.

2. Related Work

The segmentation accuracy largely depends on the choice of backbone network. Generally speaking, the more accurate the segmentation model is, the better the relative effect of semantic segmentation is. There are three very important indicators: accuracy, speed, and memory. The performance of these indicators depends on the CNN you choose and any modifications you make to it. Networks have different trade-offs on these indicators. In addition, these network structures can be modified, such as by reducing some layers and adding some layers. Usually, adding more layers will improve accuracy, while sacrificing some speed and memory. However, researchers have realized that this trade-off is subject to marginal effects; that is, the more layers are added, the less accuracy improvement will be brought by adding each layer.

The segmentation accuracy and speed of some classification network models are shown in Table 1.

Generally speaking, using a larger convolution kernel will always lead to the highest accuracy, but it will lose both speed and memory. However, this is not always the case, because it has been found many times that using a large convolution kernel will make the network difficult to diverge. Using smaller cores, such as 3×3 convolution, the effect will be better. ResNet [10] and VGGNet [11] both fully explain this fact, as shown in the papers related to these two models.

In this paper, ResNet-18, the lightweight form of ResNet, is taken as the backbone network of semantic segmentation. Compared with ResNet-50, ResNet-101, and ResNet-152,

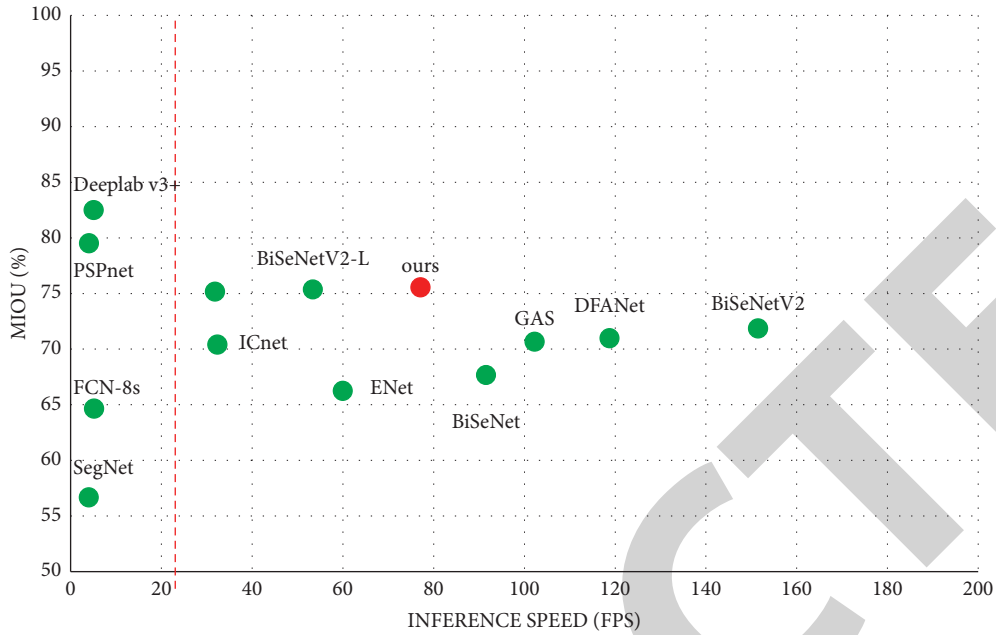

Figure 1: Comparison of speed and MIoU on the Cityscapes dataset.

ResNet-18 has fewer network layers and output channels. Although the information classification accuracy of the network is sacrificed, it can significantly improve the running speed of it. ResNet-18 is also composed of convolution layer, pool layer, and four blocks, but each block of ResNet-18 is not composed of bottleneck units but of more lightweight residual units. It can be seen that each block of ResNet-18 contains two residual units, a number which is far less than the number of units of each block in the backbone network ResNet-101, making it run faster. According to the data provided by ResNet [10], its detection speed is 31.54 ms and 156.44 ms, respectively.

Now the mainstream image semantic segmentation methods mainly focus on the improvement of network performance. Although the segmentation method based on deep convolution neural network has significantly improved the performance of image segmentation, it still faces a lot of computational overhead. However, for the real-time semantic segmentation method, the biggest concern is how to construct a real-time system with low delay. In order to solve this problem, many researchers have studied the lightweight image semantic segmentation model and summarized some experience, which is also reflected in our paper. For example, there are some achievements in the lightweight convolution structure, such as 1×1 convolution, decomposition convolution, grouping convolution, and depth separation convolution.

3. Real-Time Semantic Segmentation Network Architecture

3.1. Dual Channel Attention Mechanism Network. Therefore, reducing the communication between high-level and low-level feature maps is the most effective way to improve their fusion efficiency. Inspired by the GAU attention module in PANet, we strengthen its performance and use it as part of the attention mechanism we proposed. The attention part of the upper channel is shown in Figure 2. GDAM is a structure that can be used to enhance acquisition ability of lower feature map. For high-level feature map, GDAM first uses average pooling and max pooling to reduce its resolution to 1×1. There is one more maximum pooling operation than the original GAU module. In paper [12], the authors proved that global average pooling is not the optimal choice for channel attention, and global maximum pooling will also extract some unique features of objects, which can infer more meaningful feature information. Then, combining BN and sigmoid function uses 1×1 convolution to generate channel attention mask. For the low-level feature map, GDAM uses 3×3 combining BN and ReLU convolution layer to optimize it; here 3×3 convolution is combined with 1×3 and 3×1 convolution.

Table 1: Classification performance of common network models.

Network	Layers	Top-1 error	Top-5 error	Speed (ms)
VGG-16	16	27.00	8.80	128.62
ResNet-18	18	30.43	10.76	31.54
ResNet-101	101	22.44	6.21	156.44

3.2. DARSegNet Semantic Segmentation Model. With the asymmetric encoder-decoder and the dual attention mechanism, the DARSegNet (deep asymmetric real-time semantic segmentation network model) is illustrated in Figure 3.

In this section, we first describe in detail our proposed segmented network DARSegNet. In addition, we also explain the effectiveness of these two paths accordingly. Finally, we show how to combine the features of these two

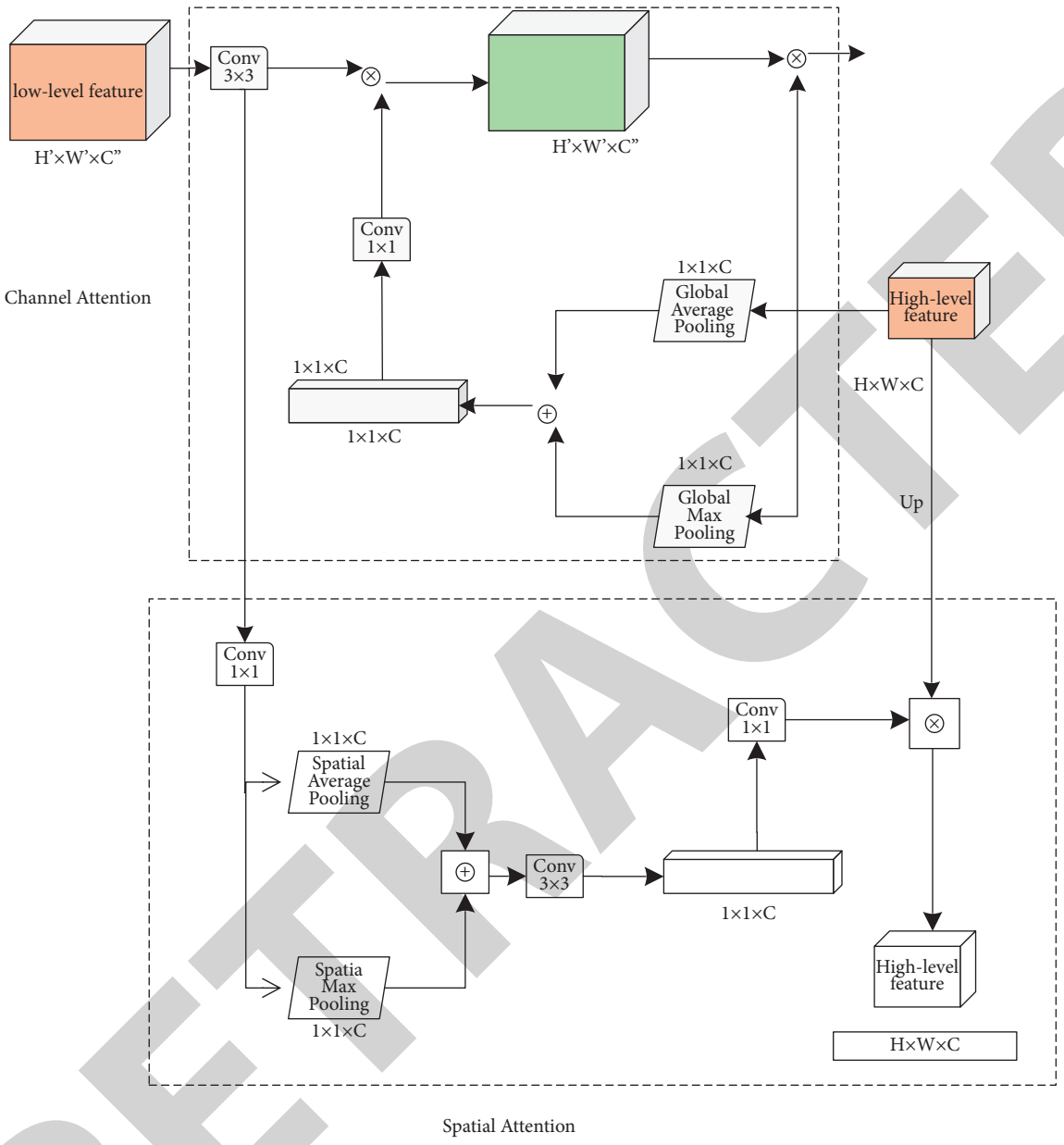

Figure 2: Dual channel attention mechanism network.

paths with the feature fusion module and the overall architecture of our DARSegNet.

DARSegNet combines the sequential and parallel structure of the general network model and can provide multiscale visualization. Because DARSegNet can provide multiscale processing of image features, it can effectively improve the accuracy of image segmentation. Compared with other segmentation methods, the hierarchical convolution module used in DARSegNet model can effectively reduce the depth of the network model and improve the coupling of the model under the condition of providing receptive fields of the same size. The global pooling layer in the model can effectively reduce the computational complexity and prevent overfitting in the training process. The module uses the attention mechanism, combined with the high-level and low-level features of the network and the introduced supervision strategy to correct the wrong details in the features, which can obtain more accurate results in the image segmentation task.

The image segmentation model DARSegNet combines the sequential and parallel structure of the general network model and can provide multiscale visualization. Because DARSegNet segmentation model can provide multiscale processing of image features, it can effectively improve the accuracy of image segmentation. Compared with other segmentation methods, the hierarchical convolution module used in DARSegNet model can effectively reduce the depth of the network model and improve the coupling of the model under the condition of providing receptive fields of the same size. The global pooling layer in the model can effectively reduce the computational complexity and prevent overfitting in the training process. The module uses the attention

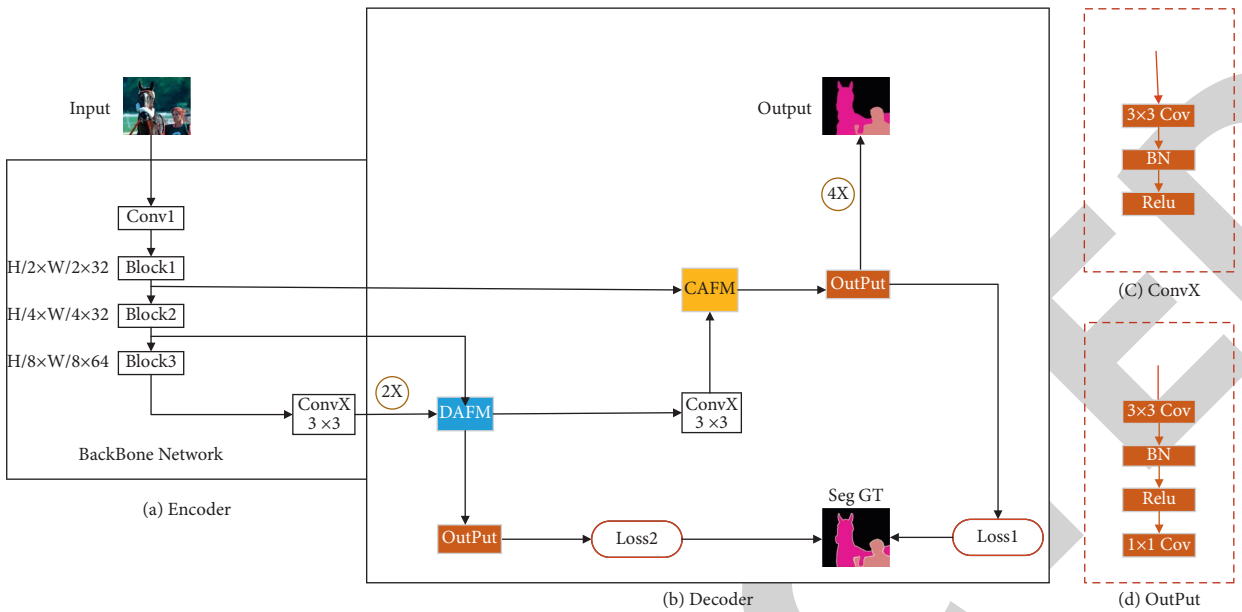

Figure 3: The architecture of the asymmetric encoder-decoder and attention mechanism of the semantic segmentation model.

mechanism, combined with the high-level and low-level features of the network and the introduced supervision strategy to correct the wrong details in the features, which can obtain more accurate results in the image segmentation task.

The encoder-decoder model is improved, the basic network structure of the encoder module is redesigned, and the feature extraction ability of the encoder is improved. On the one hand, an asymmetric convolution block (ACB) is connected behind each convolution of the backbone network. On the other hand, combining several atrous convolutions [13] with different expansion rates, according to the idea of dense connection, a dense atrous spatial pyramid pooling (DASPP) module is proposed; that is, ACB and DASPP together constitute the encoder for feature extraction. In order to improve the efficiency of the decoder to fuse high-level feature maps, a dual channel attention decoder (DCAM) is proposed, which can significantly reduce the information gap between high-level and low-level feature maps by using the attention mechanism and provide guarantee for the accurate fusion of high-level and low-level feature maps. Experimental results show that the proposed network and module can significantly improve the segmentation accuracy of the network [14].

After summarizing the above, combined with our proposed feature processing module, this paper proposes a new asymmetric encoder network model DARSegNet. The model in this chapter combines the general sequential and parallel structure and can provide multiple scales of visual domain. Therefore, the method in this chapter provides a feature of multiscale processing, which effectively improves the accuracy of the algorithm. In addition, compared with other methods, the hierarchical convolution module of the segmentation model in this chapter effectively reduces the depth of the network model when providing the same size of receptive field. The global pooling layer included in the model effectively reduces the computational overhead and prevents the training from overfitting. The module uses the hop connection strategy, combined with the characteristics of the lower layer of the network and the strategy of intermediate supervision to correct the wrong details in the features, so as to get more precise and accurate results in image segmentation.

The encoder-decoder model is improved, the basic network structure of the encoder module is redesigned, and the feature extraction ability of the encoder is improved. On the one hand, asymmetric convolution blocks (ACBs) are connected behind each convolution in the backbone. On the other hand, combined with several ATOS convolutions with different expansion rates, according to the idea of dense connection, a dense atrous (dilated) spatial pyramid pool (DASPP) module is proposed; that is, ACB SegNet and DASPP together constitute an encoder for feature extraction. In order to improve the efficiency of the decoder in fusing high-level feature maps, a dual channel attention decoder (DCAM) is proposed. The decoder uses the attention mechanism to significantly narrow the information gap between high-level and low-level feature maps, which provides a guarantee for the accurate fusion of high-level and low-level feature maps. The experimental results show that the proposed network and module can significantly improve the segmentation accuracy of the network.

3.3. Loss Function. We also use the auxiliary loss function to supervise the training of DARSegNet. In order to make the semantic segmentation model converge effectively, similar to PSPNet [15], the model proposed adding supervision information in the backbone network; that is, additional auxiliary loss function is introduced to supervise and learn the initial segmentation results generated by the model. The auxiliary loss function and the main loss function of the final segmentation result use loss function, as shown in formula

(4). Softmax function is shown in formula (1), pred is the prediction segmentation diagram, Y_t is the truth segmentation diagram, Cost $(*)$ represents the multivariate cross entropy loss function, and its definition is shown in formula (3), where n is the number of samples.

Loss function is an important part of convolution network. It is used to calculate the difference between the network prediction result and the true value, so as to update the network parameters through the back-propagation algorithm. The most widely used loss function in deep learning semantic segmentation is softmax cross entropy.

$$\text{soft max}(Z_i) = \frac{e^{z_i}}{\sum e^{z_j}}, \tag{1}$$

$$\text{Loss}(\text{pred}, Y_t) = \text{Cost}(\text{soft max}(\text{pred}), Y_t), \tag{2}$$

$$\text{Cost} = -\frac{1}{N}\sum_i ((1 - Y_t) \times \log(1 - \text{soft max}(\text{pred})) + Y_t \times \log(\text{soft max}(\text{pred}))). \tag{3}$$

In general, additional supervision in the model training stage can optimize the deep convolution neural network.

$$\text{Loss}_f = \alpha_1 \text{Loss}_1 + \alpha_2 \text{Loss}_2. \tag{4}$$

Here, Loss_f, Loss_1, and Loss_2, respectively, represent the final loss, main loss, and auxiliary loss; α_1 and α_2 represent the balance parameters of main loss and auxiliary loss. According to a large number of experiments [5], when $\alpha_1 = 1$ and $\alpha_2 = 0.4$, Loss_f is the joint loss function.

4. Experiments

This section is the experimental part. We will introduce the effect of semantic segmentation, configuration environment, network structure, and experimental results.

4.1. Experiment Environment. Facebook developed the PyTorch framework based on Python and used the Python version of the torch library in image processing. The advantage is that it provides dynamic calculation diagrams, which means that images are generated at runtime and are easier to run on GPU. However, due to the short development time and lack of reference materials, it is still to be developed.

This paper selects the advanced PyTorch platform. The specific configuration is shown in Table 2.

This section makes relevant experiments on the Cityscapes [6] and CamVid [16] datasets and compares the performance with those of other advanced models. The software and hardware configuration of the experimental platform is shown in Table 2.

Table 2: Real-time image semantic segmentation model environment.

Item	Configuration
OS	Ubuntu 16.04
CPU	Intel Core i7 4790K
GPU	GeForce 1080TI 11 GB
RAM	32 G
Framework	PyTorch 1.6
Programming language	Python 3.6
GPU acceleration	CUDA 10.2//cuDNN 7.6.5

4.2. Datasets

Cityscapes. It contains 2975 images for training, 500 images for verification, and 1525 images for testing. It has 19 dense pixel annotations. Cityscapes is a new large-scale dataset containing street scenes from 50 different cities. In addition to 20000 weak annotation frames, it also contains 5000 high-quality pixel level annotation frames [6].

CamVid. CamVid (the Cambridge driving labeled video database) dataset was released by the Engineering Department of Cambridge University in 2008. It is the first video set with target category semantic tags. It is the first video dataset containing semantic labels of object classes. It is selected from driving videos taken during the day and dusk. It contains 701 color images and notes of 11 semantic classes. The dataset consists of four video clips, each of which contains an average of 5000 frames with a resolution of 720×960 pixels, about 40 K frames [16].

4.3. Parameter Setting

Cityscapes Setting. Following [5], the SGD optimizer with initial learning rate of 0.01, momentum of 0.9, and weight attenuation of 0.0001 is used in this paper. The learning strategy with power of 0.9 is adopted to reduce the learning rate, and data enhancement methods including random clipping image, random scaling from 0.5 to 2.0, and random horizontal flip are used. The image is randomly cropped to 1024×1024 for training following [5]. We use the linear warmup strategy, from $0.1 \times \text{lr}$ (learning rate) to $1 \times 1\text{r}$ which only works in the previous 5000 iterations.

CamVid Setting. The initial learning rate is 0.001, and models are trained in 968 stages. The image is randomly cropped to 960×720 pixels for subsequent training stages. Other settings followed Cityscapes.

4.4. Analysis of Network Training Process. This section mainly analyzes the network training process of DARSegNet model and introduces and analyzes the loss rate, MIoU, and PA in the network training process of DARSegNet in the Cityscapes dataset.

4.4.1. Loss Rate Analysis of the Network Model. During the training on the Cityscapes dataset, the change of loss rate of DARSegNet model can be seen in Figure 4, in which the blue one indicates the change of loss rate.

The network convergence process is stable. In the first 10,000 iterations, the loss rate of the DARSegNet network model decreases rapidly and steadily, and the network model

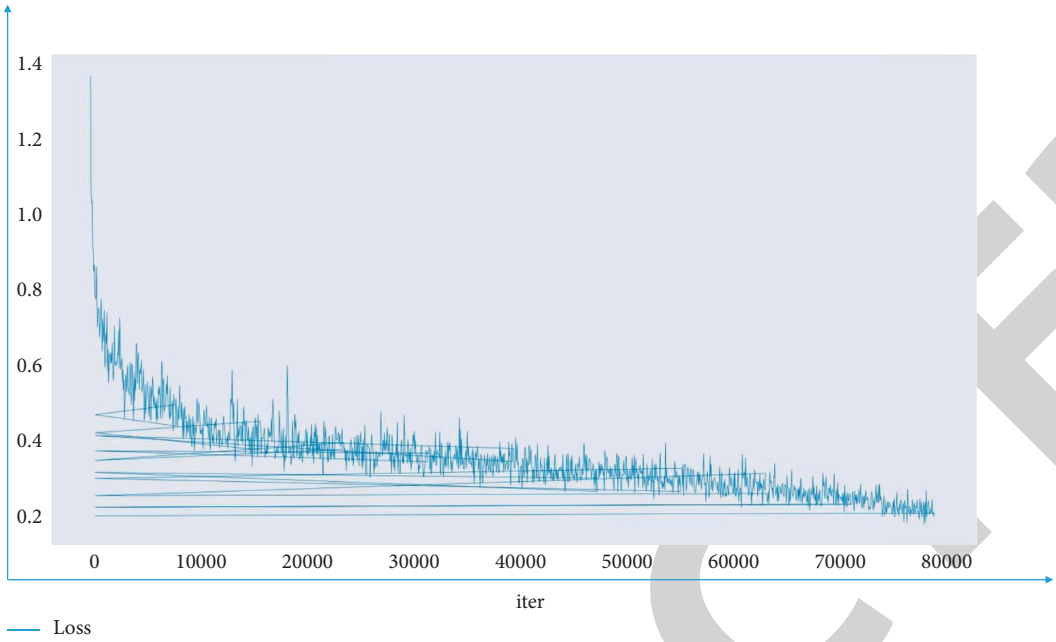

Figure 4: Decline curve of loss rate during network training.

converges quickly. During 10,000 to 18,000 iterations, the loss rate drop of the network model slows down rapidly. In the process of 18,000 to 20,000 iterations, the loss rate of the model fluctuated greatly. After 20,000 iterations of the model, the loss rate slowly decreased in the oscillation. At 70,000 epochs, the loss rate of the training model was about 0.20, and then the loss rate dropped to about 0.10. After 80,000 iterations, the loss value stabilized, and the loss rate hardly decreased.

4.4.2. Network Model MIoU Analysis. The network model has carried out 100,000 iterations in total. Around 25,000 iterations, MIoU fluctuates slightly, which is consistent with the time of loss value oscillation of loss rate in Figure 5. After that, MIoU increased rapidly and the effect was good. From 59,000 iterations to 60,000 iterations, MIoU rose very slowly in the shock, and, after the shock, MIoU began to stabilize. Near 80,000 iterations, MIoU increases to about 72%. In the process of 90,000 to 100,000 iterations, MIoU has only very small changes and tends to be stable as a whole. After 100,000 iterations, the MIoU value tends to be stable and hardly decreases, and the MIoU value is about 75%. It can be seen from the trend of the curve in Figure 5 that the MIoU of the network model rises rapidly in the early stage, the convergence speed of the model is good and tends to be stable in the later stage, there is no obvious large fluctuation, and finally it stabilizes at about 75%.

4.4.3. Pixel Accuracy Analysis of the Network Model. On the Cityscapes dataset, the changes of the pixel accuracy in the training are shown in Figure 6.

The network model has carried out 100,000 iterations in total. Around 25,000 iterations, MIoU fluctuates slightly, which is consistent with the time of loss value oscillation of loss rate in Figure 6. After that, MIoU increased rapidly and the effect was good. From 59,000 iterations to 60,000 iterations, MIoU rose very slowly in the shock, and, after the shock, MIoU began to stabilize. Near 80,000 iterations, MIoU increases to about 72%. In the process of 90,000 to 100,000 iterations, MIoU has only very small changes and tends to be stable as a whole. After 100,000 iterations, the MIoU value tends to be stable and hardly decreases, and the MIoU value is about 75%. It can be seen from the trend of the curve in Figure 6 that the MIoU of the network rises rapidly in the early stage, the convergence speed of the model is good and tends to be stable in the later stage, there is no obvious large fluctuation, and it finally stabilizes at about 75%.

The pixel accuracy of the network model increases rapidly. Although there is a small fluctuation in the middle, the pixel accuracy is stable around 0.948 in the end.

On the Cityscapes dataset, the comparisons of accuracy, speed, and parameters of some lightweight segmentation model are shown in Table 3.

Using a single GTX 1080Ti GPU card, with 32 G memory, DARSegNet achieves 75.53% MIoU and carries out image segmentation test at the speed of 70 FPS. In the Cityscapes test set, its performance is better than that of the current SOTA BiseNet V2, with an increase of 0.8%, respectively, while the number of parameters is reduced by about 23%, and FPS is nearly twice that. The visualization of image segmentation results is shown in Figure 7.

4.4.4. Comparative Analysis of Experiments on the CamVid Dataset. It can be seen that the method proposed in this paper also achieves competitive accuracy and speed on CamVid dataset and realizes a good balance between accuracy and speed. As shown in Table 4, DARSegNet achieves 73.96% MIoU and 109 FPS test speed.

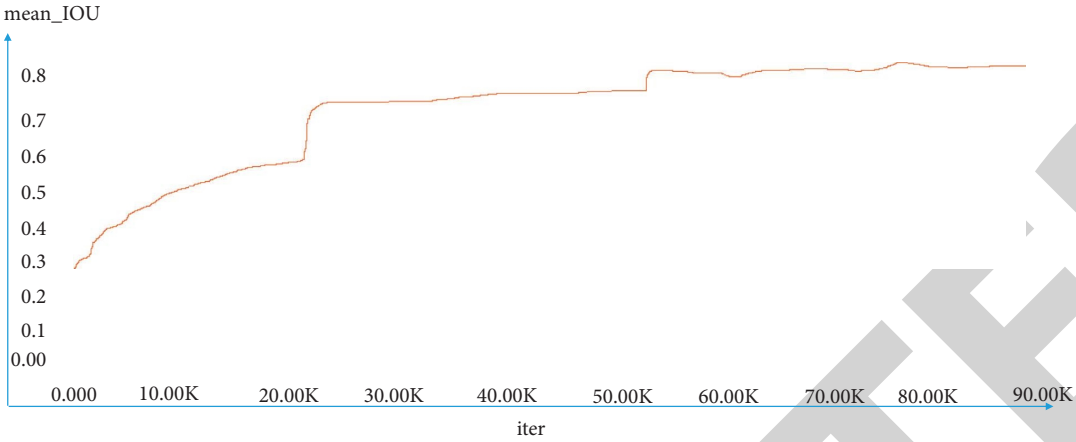

Figure 5: Changes of MIoU during training.

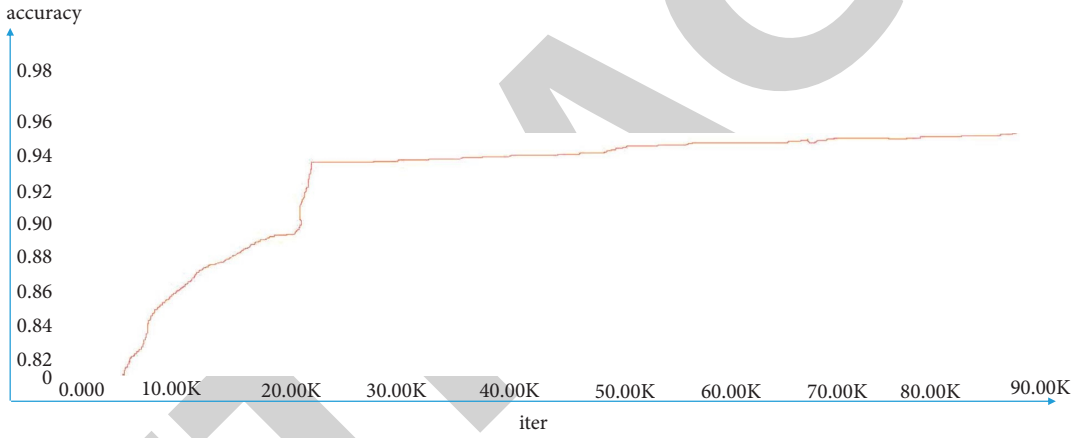

Figure 6: Changes in pixel accuracy during training.

Table 3: Comparison on cityscapes test set.

Method	Pretraining	GPU	MIoU	Speed (FPS)	Params
SegNet	ImageNet	TitanX	56.1	15	29.5 M
ENet	No	TitanX	58.3	31	0.36 M
ESPNet	No	TitanX	60.3	113	0.36 M
CGNet	No	1080Ti	64.8	17	0.50 M
ContextNet	ImageNet	1080Ti	66.1	18	0.85 M
EDANet	No	1080Ti	67.3	81	0.68 M
ERFNet	No	1080Ti	68.0	42	2.10 M
BiseNet	ImageNet	1080Ti	68.4	106	5.80 M
ICNet	ImageNet	TitanX M	69.5	30	7.80 M
DABNet	No	1080Ti	70.1	28	0.80 M
BiSeNet V2	No	1080Ti	73.4	156	49M
BiSeNet V2-L	No	1080Ti	**75.8**	47.3	NULL
DARSegNet (ours)	No	1080Ti	75.49	69.3	4.15 M
DARSegNet (ours)	ImageNet	1080Ti	75.53	69.5	4.15 M

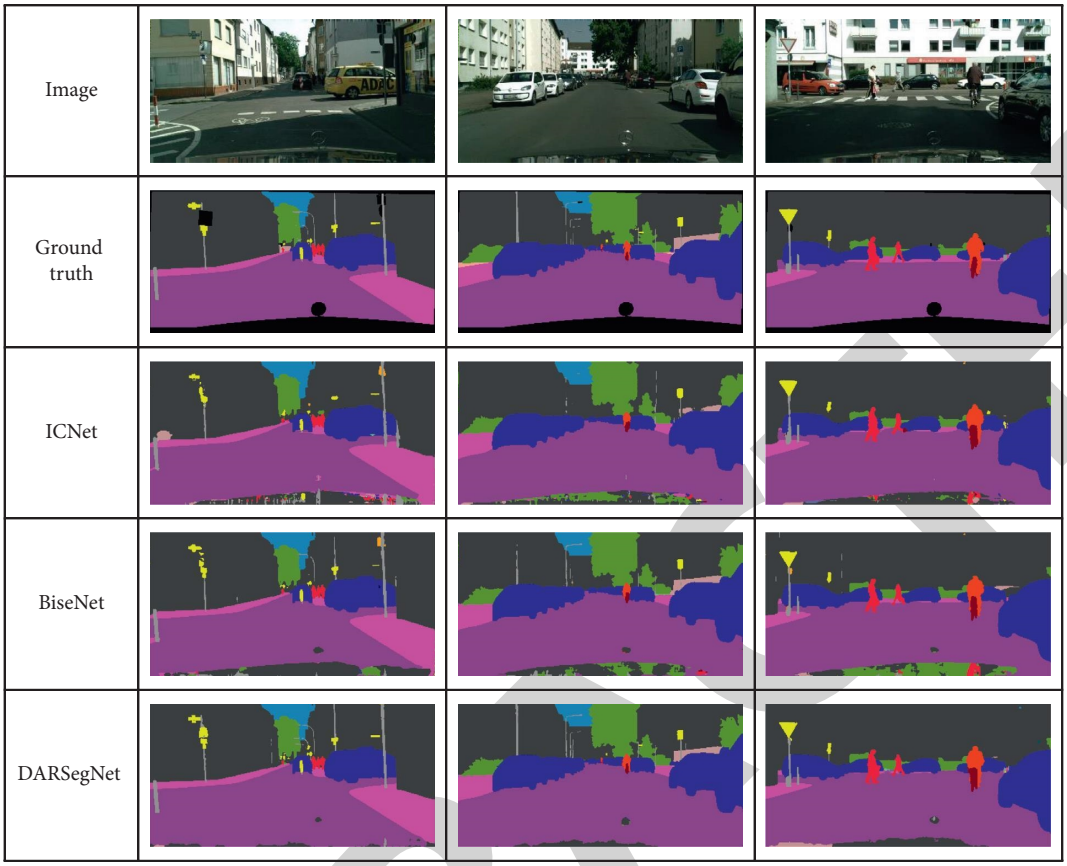

Figure 7: Visualization results on the Cityscapes test set.

Table 4: Comparison of accuracy, speed, and parameters of the lightweight segmentation model on the CamVid test set.

Method	Pretraining	GPU	MIoU (%)	Speed (FPS)
SegNet	ImageNet	TitanX	55.6	15
ENet	No	TitanX	51.3	61.2
BiseNet	ImageNet	1080Ti	65.6	175
ICNet	ImageNet	TitanX	67.1	34.5
SFNet (DF2)	No	1080Ti	70.4	134
CAS	No	Titan Xp	71.2	169
BiSeNet V2	No	1080Ti	68.7	116
BiSeNet V2	No	1080Ti	73.2	32.7
SwiftNet	No	1080Ti	72.6	75.9
DARSegNet (ours)	No	1080Ti	73.27	109
DARSegNet (ours)	ImageNet	1080Ti	73.96	109

The best results in each class are shown in bold.

According to the setting of [5], the split network model is trained with a combined dataset consisting of training set and validation set. The MIoU is measured using the test set. To test the running speed of the network, an image with a size of $720 \times 960 \times 3$ resolution was input to each network, and the FPS of each network was measured.

5. Conclusions

In this paper, we propose an effective real-time semantic segmentation network model for asymmetric encoder and decoder. This segmentation result is efficient and in real time and has been recognized. We improve the performance by using asymmetric encoder structure and dual attention module, increasing the receptive field of backbone network. We introduce a novel attention mechanism to increase the accuracy of semantic segmentation without losing local details and balance the speed. The real-time segmentation performance is verified by several experiments in the Cityscapes and CamVid datasets. We also propose a lightweight decoder module, which has better performance than the ordinary decoder. The next research will focus on improving the generalization performance.

Data Availability

The simulation experiment data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant no. 61303029).

References

[1] D. Jiang, G. Li, Y. Sun, J. Kong, and B. Tao, "Gesture recognition based on skeletonization algorithm and CNN with ASL database," *Multimedia Tools and Applications*, vol. 78, no. 21, pp. 29953–29970, 2019.

[2] G. Li, D. Jiang, Y. Zhou, G. Jiang, J. Kong, and G. Manogaran, "Human lesion detection method based on image information and brain signal," *IEEE Access*, vol. 7, pp. 11533–11542, 2019.

[3] A. Liu, Y. Yang, Q. Sun, and Q. Xu, "A deep fully convolution neural network for semantic segmentation based on adaptive feature fusion," in *Proceedings of the 2018 5th International Conference on Information Science and Control Engineering*, pp. 16–20, (ICISCE), Zhengzhou, China, July 2018.

[4] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[5] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European conference on computer vision*, pp. 325–341, (ECCV), Munich, Germany, September 2018.

[6] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, and R. Benenson, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3213–3223, Las Vegas, NV, USA, June 2016.

[7] J. Yang, W. Zhang, J. Liu, J. Wu, and J. Yang, "Generating de-identification facial images based on the attention models and adversarial examples," *Alexandria Engineering Journal*, vol. 61, no. 11, pp. 8417–8429, 2022.

[8] W. Wei, S. Liu, W. Li, and D. Du, "Fractal intelligent privacy protection in online social network using AttributeBased encryption schemes," *IEEE Transactions on Computational Social Systems*, vol. 5, no. 3, pp. 736–747, 2018.

[9] K. Cai, H. Chen, W. Ai, X. Miao, Q. Lin, and Q. Feng, "Feedback convolutional network for intelligent data fusion based on near infrared collaborative IoT technology," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 2, pp. 1200–1209, 2022.

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, USA, July 2016.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[12] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "A simple and light-weight Attention module for convolutional neural networks," *International Journal of Computer Vision*, vol. 128, no. 4, pp. 783–798, 2020.

[13] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[14] C. Tian, Y. Xu, W. Zuo, C. W. Lin, and D. Zhang, "Asymmetric CNN for image superresolution," *IEEE Transactions on Systems, Man, a Cybernetics: Systems*, vol. 99, pp. 1–13, 2021.

[15] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6230–6239, CVPR), Honolulu, HI, USA, July 2017.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.