Hindawi
Security and Communication Networks
Volume 2023, Article ID 9836094, 1 page
https://doi.org/10.1155/2023/9836094

Retraction

Retracted: The Segmentation of Road Scenes Based on Improved ESPNet Model

Security and Communication Networks

Received 10 October 2023; Accepted 10 October 2023; Published 11 October 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This article has been retracted by Hindawi following an investigation undertaken by the publisher [1].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] R. Jin, T. Yu, X. Han, and Y. Liu, "The Segmentation of Road Scenes Based on Improved ESPNet Model," *Security and Communication Networks*, vol. 2021, Article ID 1681952, 11 pages, 2021.

Hindawi
Security and Communication Networks
Volume 2021, Article ID 1681952, 11 pages
https://doi.org/10.1155/2021/1681952

Research Article

The Segmentation of Road Scenes Based on Improved ESPNet Model

Ran Jin, Tongrui Yu, Xiaozhen Han, and Yunpeng Liu

College of Big Data and Software Engineering, Zhejiang Wanli University, Ningbo 315100, China

Correspondence should be addressed to Ran Jin; ran.jin@163.com and Xiaozhen Han; 170748822@qq.com

Received 9 April 2021; Revised 17 May 2021; Accepted 21 June 2021; Published 29 June 2021

Academic Editor: Kifayat Ullah

Image segmentation is an important research in image processing and machine vision in which automated driving can be seen the main application scene of image segmentation algorithms. Due to the many constraints of power supply and communication in in-vehicle systems, the vast majority of current image segmentation algorithms are implemented based on the deep learning model. Despite the ultrahigh segmentation accuracy, the problem of mesh artifacts and segmentation being too severe is obvious, and the high cost, computational, and power consumption devices required are difficult to apply in real-world scenarios. It is the focus of this paper to construct a road scene segmentation model with simple structure and no need of large computing power under the premise of certain accuracy. In this paper, the ESPNet (Efficient Spatial Pyramid of Dilated Convolutions for Semantic Segmentation) model is introduced in detail. On this basis, an improved ESPNet model is proposed based on ESPNet. Firstly, the network structure of the ESPNet model is optimized, and then, the model is optimized by using a small amount of weakly labeled and unlabeled scene sample data. Finally, the new model is applied to video image segmentation based on dash cam. It is verified on Cityscape, PASCAL VOC 2012, and other datasets that the algorithm proposed in this paper is faster, and the amount of parameters required is less than 1% of other algorithms, so it is suitable for mobile terminals.

1. Introduction

In recent years, CNN (Convolutional Neural Network) has made great progress in tasks such as image classification and object detection. The most important first step in these tasks is to predict the classification of each pixel in an image, and by segmenting the original image, researchers hopefully achieved accurate identification of what part of the image each pixel belongs to. It is very critical as the first step in computer vision applications. Some traditional methods, such as the Otsu (Maximum Between-Class Variance) method has been used with some success. The FCN (Full Convolutional Network) proposed by Long et al. [1] in 2015 opens up new avenues for image segmentation. The method trains an end-to-end network that uses a convolutional layer instead of an inner layer in a traditional network and can accept image input of any size. Based on the FCN, Chen et al. [2] added conditional random fields to further fine-grained optimization of the FCN model to improve the effect of image segmentation of the boundaries, as the result of achieving 71.6% IOU (Intersection over Union) at the PASCAL VOC 2012 dataset. To address the accuracy problem of image edge information segmentation, Zheng et al. [3] suggested embedding CRF (Conditional Random Fields) as a Recurrent Neural Network (RNN) in FCNs. The average IOU of the PASCAL VOC 2012 dataset increased to 74.7% in the CRF-RNN (Conditional Random Fields-Recurrent Neural Networks) model. To address the problem of overfitting of small samples, the DenseNet (Densely Connected Convolutional Network) model of FCN can achieve the required accuracy without prior training and reduce the number of parameter to 1/10 of the original model, which has a broad application prospect in tasks such as automatic driving, medical images, and satellite images.

The remaining parts of this paper are organized as follows. In Section 2, we first review the ESPNet algorithm and model evaluation criteria. In Section 3, we provide some related work. Section 4 presents an improved ESPNet model.

Then, experiments are conducted in Section 5. Finally, the paper is concluded in Section 6.

2. Preliminaries

2.1. ESPNet. The ESPNet was introduced by Mehta et al. [4] in 2018, where a semantic segmentation network architecture featuring fast calculation and excellent effect of segmentation is presented in detail. ESPNet can be as fast as it achieves processing speeds of 112 frames per second on the GPU and up to 9 frames per second on edge devices. It is faster than the most well-known lightweight networks such as the MobileNet (Efficient Convolutional Neural Networks for Mobile Vision), ENet (a deep neural network architecture for real-time semantic segmentation), and ShuffleNet (an extremely efficient convolutional neural network for mobile) [5, 6], among others. With a loss of only 8% classification accuracy in the control model, ESPNet is only 1/180th as fast as its model parameters and 22 times faster than the best PSPNet (Pyramid Scene Parsing Network) architecture of the time. The design idea of convolutional factor decomposition is used in many deep CNN structures, such as Inception, ResNext (Residual Neural Network), Xception [7, 8], and others. Based on the basic idea of convolutional factor decomposition, the authors introduced a convolutional module called effective space pyramid in ESPNet which makes the network architecture fast, low power, and low latency, making it ideal for deployment in resource-constrained edge devices.

The basic network architecture of ESPNet is shown in Figure 1. In the model, the number of channels is reduced by point convolution and then sent to the convolution pyramid of the cavity. A larger receptive field is obtained by expanding convolution of different proportions, and feature fusion is carried out at the same time. Therefore, the number of parameters is very small. When the number of channels is reduced, the parameters of each expansion convolution are very few. The concatenation strategy is quite different from the ordinary method of feature fusion by expanding convolution. In order to avoid gridding artifacts [9], strategy of adding step by step is adopted. The main architecture of the ESPNet design is shown in Figure 1. The lightweight code-decoding network architecture is shown in Figure 2.

ESPNet can achieve an accuracy of 60.3% on the Cityscapes Dataset. Currently, for the application of deep convolutional neural networks in semantic segmentation tasks, the main means of model lightening include convolutional factor decomposition, network compression, low-bit networks, and sparse CNN [10–12]. Convolutional factor decomposition reduces the complexity of convolutional operations by breaking them down into several steps. ESPNet divides the convolutional layer in the network into point convolution and spatial pyramid-based dilated convolution based on the means of convolutional factor decomposition.

The dilated convolution means that holes are injected into the standard convolution operation to increase the size of the receptive field of the convolutional layer. Compared to conventional convolution operations, dilated convolution increases the hyperparameter of the dilation rate. This hyperparameter represents the number of intervals between kernels. The pairing of standard and dilated convolution operations is shown in Figure 3.

The main purpose of using dilated convolution is to solve the problem that the small object information in the image cannot be reconstructed due to the application of a large number of pooling layers in traditional deep CNN, thus affecting the resolution of the semantic segmentation model.

2.2. Model Evaluation Criteria

2.2.1. Accuracy. The accuracy of scene segmentation directly affects the safety performance of driving. The calculations are based on the following four criteria to provide a more comprehensive assurance of accuracy. The main evaluation criteria are shown in equations (1)–(4). These are the various forms of pixel accuracy evaluation. Pixel accuracy (PA) is the most intuitive calculation method for evaluating image segmentation algorithms, and its purpose is to represent the ratio of total pixels in the image of a pixel station with a correct prediction, calculated by the following formula:

$$\mathrm{PA} = \frac{\sum_{i=0}^{k} P_{it}}{\sum_{i=0}^{k} \sum_{j=0}^{k} P_{ij}}. \tag{1}$$

Definition 1: mean pixel accuracy (MPA); calculating the ratio of the total number of correct pixels in each category to the total number of pixels in each category firstly, and then, finding the mean value of each category's PA, which is calculated as follows:

$$\mathrm{MPA} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}. \tag{2}$$

Definition 2: mean intersection over union (MIOU) [13]; calculating the intersection and union ratio to measure the advantages and disadvantages of the algorithm, it is one of the important evaluation indexes in the semantic segmentation model. Here, the intersection and union ratio is the ratio of overlap between the standard labeling of the dataset and the predicted segmentation. It is the calculation of the ratio between TP and TP + FN + FP. The MIOU is first calculated based on each category, and then, its mean value is calculated. The formula is as follows:

$$\mathrm{MIoU} = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - P_{it}}. \tag{3}$$

Definition 3: frequency weighted intersection over union (FWIOU) [14]; assigning different weighting factors for each classification according to the

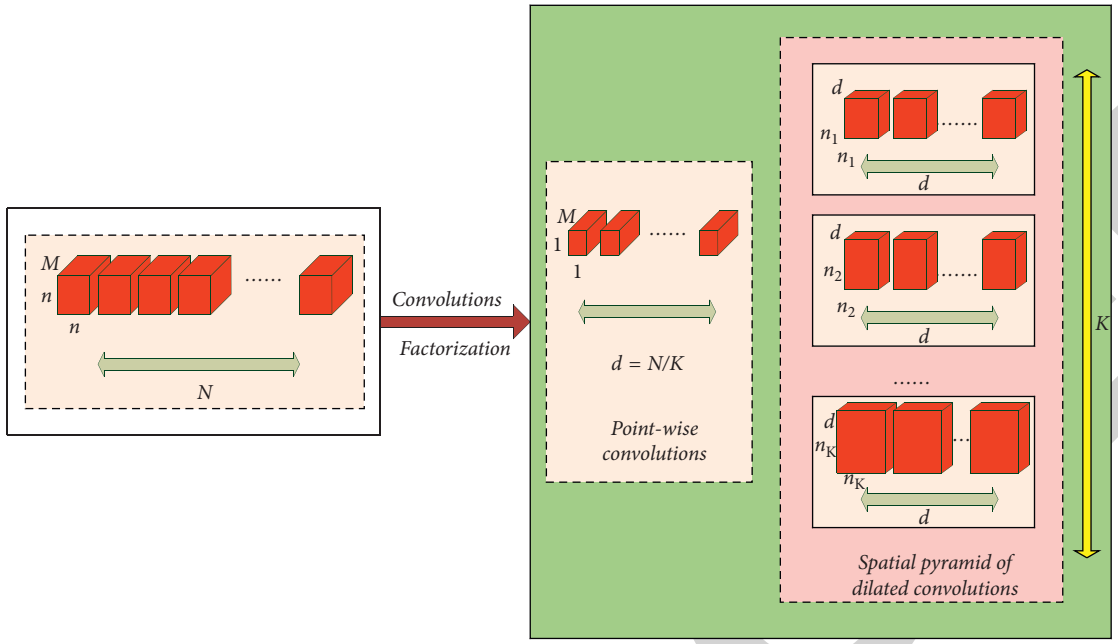

Figure 1: Basic network architecture of ESPNet.

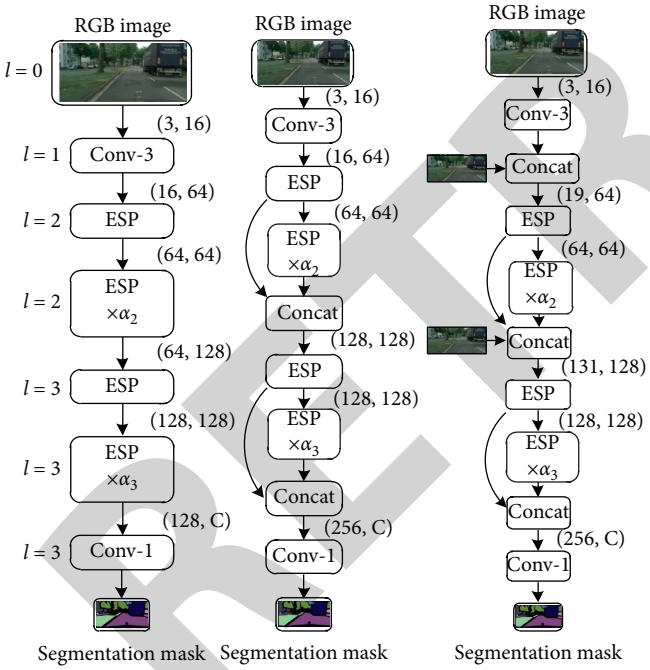

Figure 2: Main architecture of ESPNet.

frequency of each classification, it is an improved version of MIOU. The formula is as follows:

$$\mathrm{FWIoU} = \frac{1}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \sum_{i=0}^{k} \frac{\sum_{j=0}^{k} p_{ij} p_{ii}}{\sum_{j=0}^{k} p_{ji} - P_{ii}}, \tag{4}$$

where P_{ij} is the total number of pixels that belong to class i but are predicted to be class j and k indicates the total number of categories.

2.2.2. Latency. Latency represents the time of a CNN processing a single image and is usually evaluated by the number of frames processed per second. The latency rate is an important reference index for intelligent driving. Therefore, this paper adopts a distributed computing method to deal with the real-time image recognition analysis of roads, with the advantage of being able to label massive data to perform an optimal solution. It is calculated as follows:

$$\min_{\theta \in \vartheta} \sum_{1}^{S} \frac{1}{N} \sum_{1}^{N} L\left(f_\theta^s(x_i^s), y_i^s\right) + \lambda \mathrm{Reg}\left(f_\theta^s\right), \tag{5}$$

where S is the number of replicas of the model to be optimized, i.e., the number of replicas of the model improved by the ESPNet model in this paper.

Network parameters represent the number of parameters to be learned in the neural network. The network size indicates the amount of storage space required to store the network. Sensitivity to GPU frequency is important for evaluating the computational power of the model. This is usually expressed as the ratio of the rate of change in execution time and the rate of change in GPU frequency. The higher this ratio, the better the ability of that deep learning application to utilize the GPU. Resource utilization refers to the ability to use a combination of CPU and GPU resources when running on an edge device [15]. In fact, edge computing devices such as the Jetson TX2 and the CPU and GPU share storage space.

3. Related Works

Many scholars have carried out useful research on image segmentation [16–27]. Zhao et al. [16] introduced horizontal crossover search (HCS) and vertical crossover search (VCS)

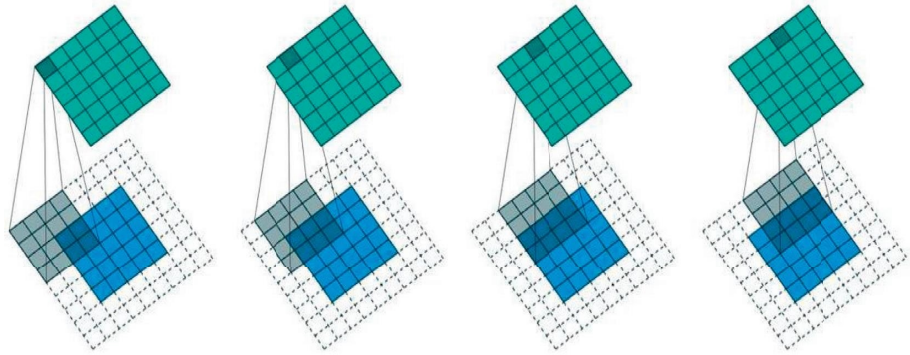

Figure 3: Conventional convolution on the left and dilated convolution on the right.

into the ACOR and improved the selection mechanism of the original ACOR to form an improved algorithm (CCACO) for the first time. Liu et al. [17] proposed a novel structure to fuse image and LiDAR point cloud in an end-to-end semantic segmentation network, in which the fusion is performed at the decoder stage instead of at, more commonly, the encoder stage. Ji et al. [18] proposed a new architecture of feature aggregation, which is designed to deal with the problem that the information of each convolutional layer cannot be used reasonably and the shallow layer information is lost in the process of transmission. Reis et al. [19] took advantage of the learned model in a deep architecture, by extracting side outputs at different layers of the network for the task of image segmentation. Parajuli et al. [20] performed pixel-wise segmentation to classify each pixel as road or nonroad based on color and depth features in a larger neighborhood context and described a cost-effective, modular, deep convolution network design. In order to improve the effect of image segmentation, directly at the deficiency of single seed point and fixed threshold of traditional region growing algorithm, a seed selection method based on the gray level of two-dimensional histogram and local variance is proposed, and the dynamic threshold is used to change the region growing rule [21]. Badrinarayanan et al. [22] presented a novel and practical deep fully convolutional neural network architecture for semantic pixel-wise segmentation termed SegNet. Akagic et al. [23] proposed an efficient unsupervised vision-based method for pothole detection without the process of training and filtering. Zhang et al. [24] adapted the multidimensional Haar-like features as well as the AdaBoost algorithm, to implement training of the cascade classifier, which will achieve the reliable vehicle detection.

In the past two years, some scholars have carried out useful research based on ESPNet [13–15, 28]. Kim and Heo [13] proposed ESCNet based on ESPNet architecture which is one of the state-of-the-art real-time semantic segmentation network that can be easily deployed on edge devices. Nuechterlein and Sachin [14] extended ESPNet, a fast and efficient network designed for vanilla 2D semantic segmentation, to challenging 3D data in the medical imaging domain.

4. Improvements Based on the ESPNet Model

In this section, we improve ESPNet, describe the core module that builds it, and compare the improved ESP module with similar CNN modules, such as Inception, ResNet, MobileNet, and ShuffleNet.

ESPNet is a decomposed form of convolution based on the Efficient Spatial Pyramid (ESP) module. It decomposes the standard convolution into spatial pyramids of point and unfolded convolution. The point convolution in the ESP module uses $1 * 1$ convolution to map high-dimensional features to low-dimensional space. The spatial pyramid of extended convolution resamples these low-dimensional feature maps simultaneously using K and $N * N$ extended convolution kernels. The expansion rate of each convolutional core is $2K - 1$ ($K = F1$). Based on this decomposition, the number of parameters and memory required for the ESP module is greatly reduced, while preserving a large effective receive domain $(n - 1)\ 2K - 1$. The pyramidal convolution operation is called a spatial expansion convolution pyramid.

Designed for fast semantic segmentation of high-resolution images with limited resource, ESPNet is efficient in terms of computational memory and power consumption and is 22 times faster than PSPNet [28] on the GPU, with 180 times smaller files and only 8% accuracy loss. ESPNet is validated on the Cityscapes, PASCAL VOC 2012, and other datasets and outperformed all current efficient CNN networks such as MobileNet, ShuffleNet, and ENet in both standard metrics and newly introduced performance metrics (measuring the efficiency of edge devices) under the same memory and compute conditions. ESPNet is fast, small, low power, and low latency to ensure a network with segmentation accuracy. The improved ESPNet model follows the principle based on convolution factor decomposition, as shown in Figure 4, and can be easily adopted to resource-constrained end devices based on the ESP module.

On the basis of extended convolution, the ASPP (Atrous Spatial Pyramid Pooling) module is introduced to realize multiscale information collection, and image level feature information is integrated in the existing ASPP module. ASPP uses four different expansion rates of extended convolution to capture multiscale information in parallel on the

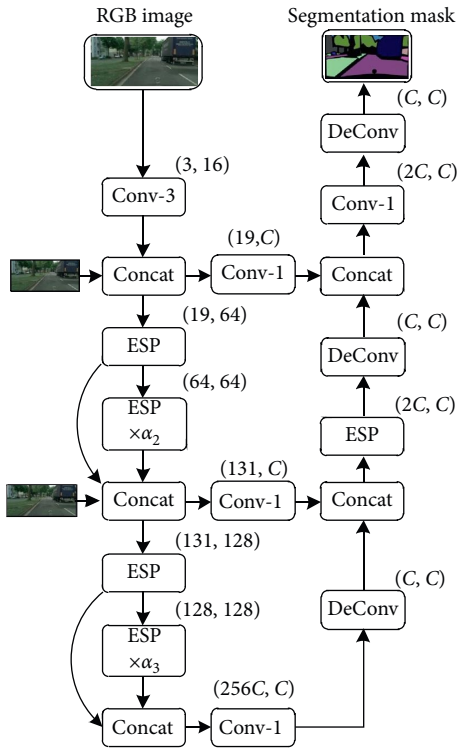

Figure 4: Basic structure of the improved model.

top-level feature responses of the backbone. The improved ASPP module gives the neurons a larger receptive field. The Pyramid Pooling Module (PPM) is introduced into the proposed ESPNet in DeepLab-v3 (Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs). Thus, in aggregating the contextual semantic information of different regions, a better segmentation is obtained. The basic structure and working principle of the improved ESPNet model are shown in Figures 4 and 5.

It is well known that irrational use of void convolution can lead to mesh artifacts, and ESPNet's use of stacked convolutional structures with large void ratios is also easy to form artifacts [29]. This paper uses HFF (Hierarchical Feature Fusion) to enrich the use of void convolution and effectively reduce the formation of artifacts. The minimum $(n_1 * n_1)$ feature map of the hole kernel is directly output, and the hole kernel $(n_2 * n_2)$ feature map is output as a residual with the previous output. The summation is used as the output. Subsequent feature maps are similar to this operation to obtain fused features with different void rates, which are then stitched together and later form residuals with the original input. In summary, HFF ensures the quality of the output through the restriction of residuals in a way that stitches together different layers of feature maps, preserving local details and global semantic features. The HFF structure allows the use of large void-rate convolution kernels, speeding up the extraction of semantic features. The decoder is similar to UNet (Convolutional Networks for Biomedical Image Segmentation), which uses layer-by-layer up-sampling, hopping connection, and restoring detailed information. Because of the fusion method of residual calculation, we use the PreLU activation function and finally connect softmax for network training.

4.1. Activate Function Module. In this paper, scene segmentation experiments are conducted using a variety of different activation functions based on ESPNet to investigate which type of activation function can lead to better network performance improvement for CNN. In this section, pairs of functions of different forms of Maxout, Tanh, ReLU, ELU, and PreLU are used as activation functions for neural networks [30, 31], respectively, while extensive comparative experiments are conducted. The experiments on activation function selection are performed on the PASCAL VOC 2012 dataset. Based on the variation in segmentation accuracy observed in this paper, the advantages and disadvantages of different activation functions throughout the training process were deeply studied.

An intuitive idea is to apply softmax to each weight [32], such that all weights are normalized to a probability with values ranging from 0 to 1, indicating the importance of each input. However, as shown in our previous studies, the additional softmax results in a significant slowdown in the GPU hardware. In order to minimize the cost of the additional delay, we further propose a fast fusion method. The formula is shown in the following equation:

$$O = \sum_i \frac{e^{w_i}}{\sum_j e^{w_j}} . I_i. \tag{6}$$

4.2. Pooling Module. Figure 6 explains the principle of the pooling approach with the aim of analyzing the impact of different pooling approaches on the performance of this paper's scene segmentation network. In this paper, three approaches, average pooling, max pooling, and random sampling pooling [33], are investigated in depth, and comparative experiments are conducted.

Different pooling layers in ESPNet are used to compare and analyze the average pooling (Mean), max pooling (Max), and random pooling (Stoh) on the network performance with three different pooling layer structures. Figure 6 shows the change of accuracy during the iteration of ESPNet with different pooling layers. The x-axis edge is the number of iterations, and the y-axis represents the accuracy [34, 35]. To more clearly compare the training of the network under the three pooling layers, Figure 7 shows only the change in accuracy during the beginning rounds of iteration. In this paper, various pooling methods are further analyzed, and local pixel maxima are often extracted as feature points in traditional image features such as textures and gradients because these local extreme points are better able to describe the edge information of the image.

The function of the maximization operation is that whenever a feature is extracted in any quadrant, it is retained in the maximized pooled output. So what the maximization operation actually does is that if a feature is extracted in the

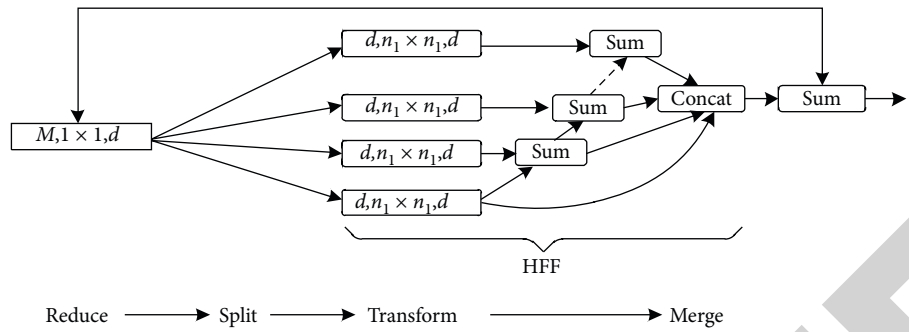

Figure 5: Schematic diagram of the improved model.

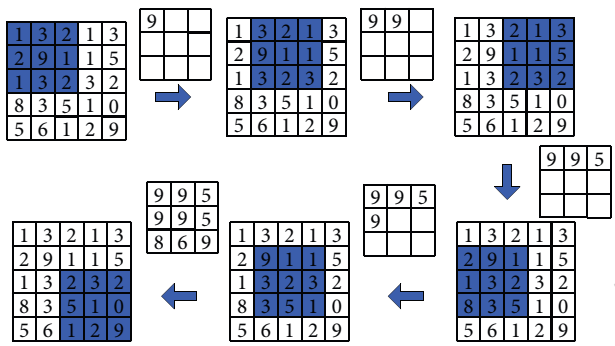

Figure 6: Pooling layer basic mode diagram.

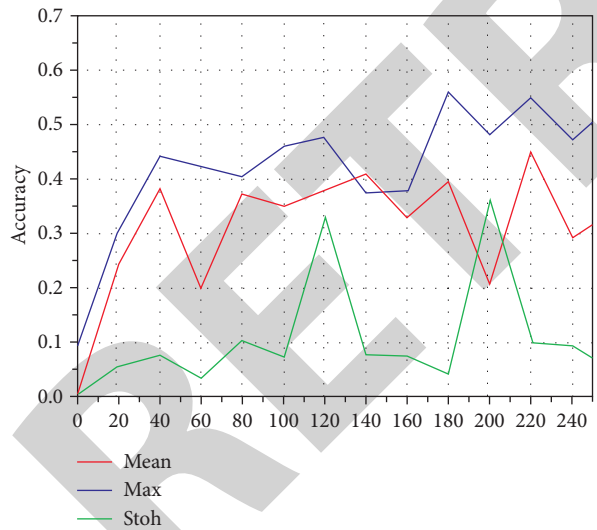

Figure 7: Comparison of basic approaches to pooling layers.

filter, then the maximum value is preserved. If this feature is not extracted, it may not exist in the upper right quadrant, and the maximum value of it is still small, which is an intuitive understanding of max pooling.

It can be seen that when there are several examples of superparameters and the input is a 5 ∗ 5 matrix, we use max pooling. The filter parameter is 3 ∗ 3, i.e., f is 3, the step is 1, i.e., s is 1, and the output matrix is 3 ∗ 3. The same formula for calculating the output size of the convolutional layer described earlier also applies to the max pooling, as shown in the following equation:

$$S = n + 2p - fs + 1. \tag{7}$$

In Section 5, all experiments are conducted on the PASCAL VOC 2012 dataset. To search for a more suitable semantic segmentation for urban road scenarios, we studied the accuracy of the proposed model for image segmentation and the influence of various pooling methods on the model.

5. Experimental Classification Results and Analysis

5.1. Common Dataset.

PASCAL VOC 2012 Dataset. This set contains 20 object categories, and the training sample contains 11,530 images. This includes 27,450 regions of interest for labeled object types and 6,929 semantic segmentation regions.

Cityscapes Dataset. This set is derived from a large number of video sequences recorded from streets in different cities, covering the 50 cities in spring, summer, and autumn, mainly in Germany and neighbouring countries. By using camera systems and postprocessing that represent the current state-of-the-art in the automotive field, a total of 5,000 images with high quality pixel-level labeled fine images and 20,000 additional images with coarse labels were obtained.

5.2. Experimental Environment. Considering that the scene objects are mainly partitioned within the city, their driving speed and equipment costs are limited, so the experimental equipment in this paper hardware and software are matched to ensure low latency and high efficiency while avoiding the use of costly equipment. Therefore, the experimental environment setup is shown in Table 1.

5.3. Experimental Results on the Cityscapes Dataset. Figure 8 describes the experimental results on the Cityscapes Dataset.

Table 1: Experimental environment configuration table.

CPU	Intel core i5-6200K
RAM	16G
Operating system	Ubuntu 16.04
Display card (computer)	GTX 1080Ti
PyTorch	1.2.0
Python	3.7.1

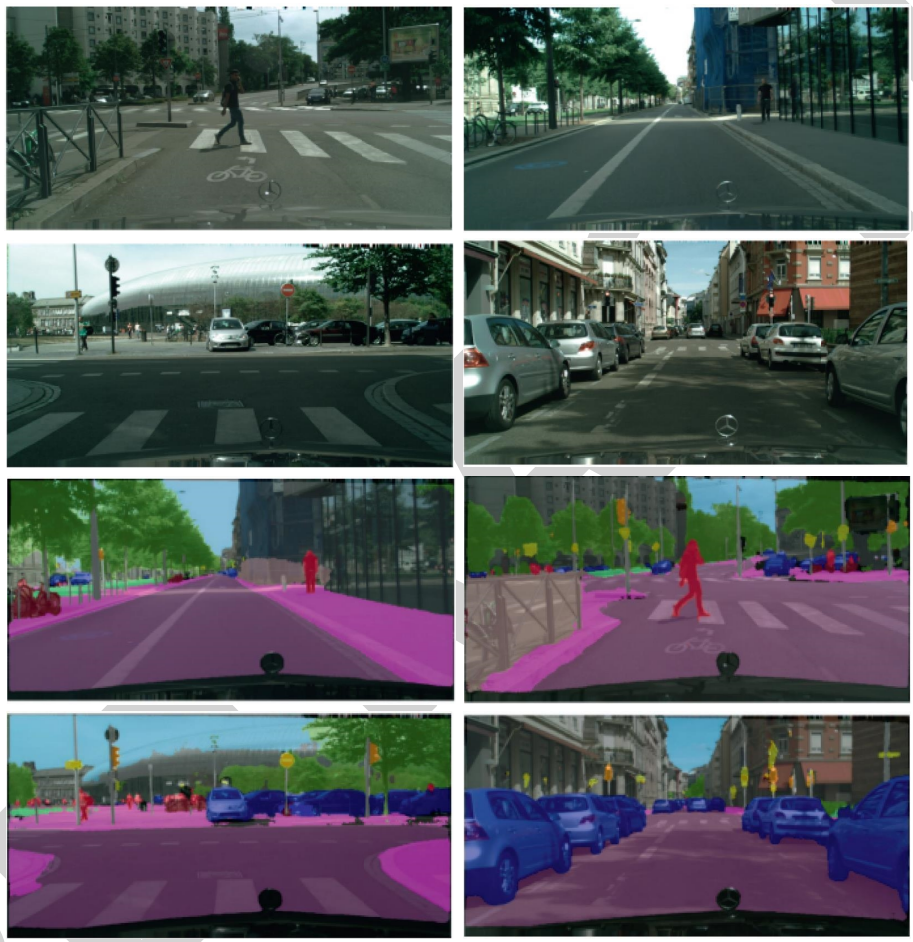

Figure 8: Plot of Cityscapes experimental segmentation results.

5.4. Experimental Results on the PASCAL VOC 2012 Dataset. Since most data of the PASCAL VOC 2012 dataset is not obtained from the camera loaded on the vehicle, it is moderately effective when using the proposed model in this paper for identification. The data and results of this experiment are shown in Table 2 and Figure 9, respectively. The upper part of the results is the original image, and the lower part is the segmentation result. Although the data segmentation results are missing compared to the Cityscapes Dataset, the overall object segmentation is basically correct.

For the characteristics of more pedestrians and vehicles on the road, this paper adjusts the type distribution of PASCAL VOC 2012 dataset to ensure the maximum number of pedestrian and vehicle data and appropriately reduces other type of data, so as to obtain more targeted experimental results to illustrate the segmentation effect of the model on the urban road scene.

5.5. Experimental Results on Self-Selected Data. This section shows the segmentation results of the road scene segmentation model designed in this paper in a continuous video. Two frames with an interval of about 1 second are extracted for illustration. The results are shown in Figure 10. It can be seen that the network designed in this paper has high accuracy and generalization ability in road segmentation and can identify the existence of obstacles on the road ahead in good prospects for application on road obstacle prediction for driving assistance.

Table 2: Experiment data table of PASCAL VOC 2012.

	Training set		Test set		Training set 1		Test set 1	
	Images	Object	Images	Object	Images	Object	Images	Object
Aeroplane	112	151	126	155	238	306	204	285
Bicycle	116	176	127	177	243	353	239	337
Bus	97	115	89	114	186	229	174	213
Car	376	625	337	625	713	1250	721	1201
Horse	139	182	148	180	287	362	274	348
Motorbike	120	167	125	172	245	339	222	325
Person	1025	2358	983	2332	2008	4690	2007	4528
Total	1985	3774	1935	3755	3920	7529	3841	7237

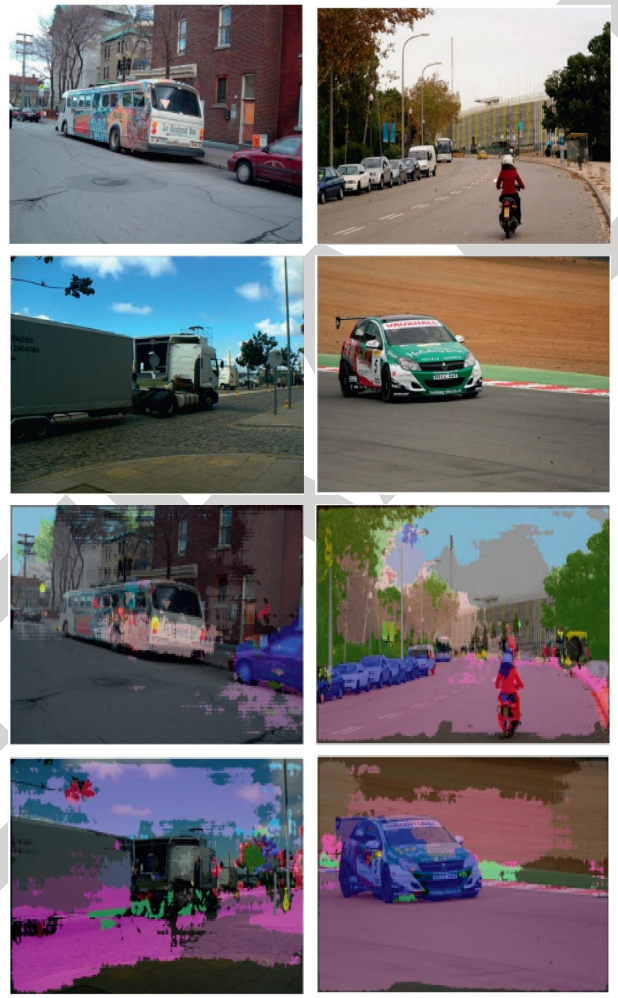

Figure 9: Experimental segmentation result graph of PASCAL VOC 2012.

5.6. Test of Different Activation Function. In this paper, different activation functions are used in ESPNet to test the effect of different activation functions on the training results. The activation functions tested are Maxout, Tanh, ReLU, ELU, and PReLU. In this paper, we document the changes in the training sample segmentation accuracy metric during the iterative process, and the experimental results are shown in Figure 11 with the x-axis of the figure indicating the number of iterations and the y-axis indicating the accuracy. In terms of final accuracy, better experimental results were obtained using ESPNet with ELU and PReLU. However, throughout the iterations, the ELU model fluctuates sharply several times during the iterations, and the accuracy of the ascension process is slow. In contrast, the accuracy of the PReLU model improves rapidly to near the peak during iteration, and there are some fluctuations after that; the whole iterative process always maintains the highest accuracy level in the same period.

Figure 10: Comparison of two frames of self-selected data.

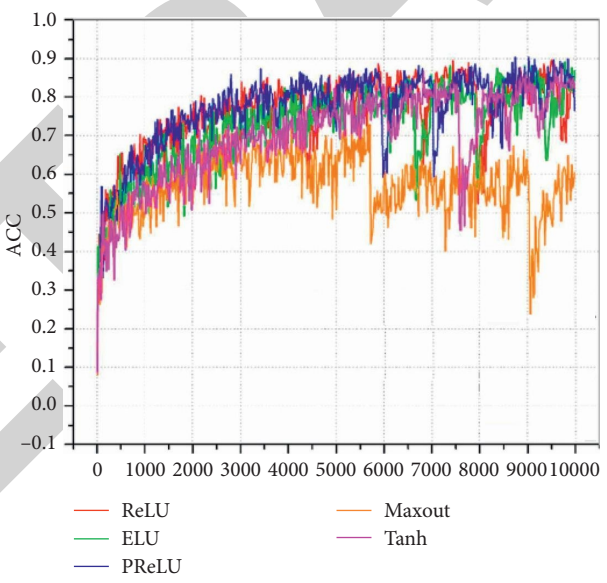

Figure 11: Comparison of experimental results for activation functions.

5.7. Comparison of Approaches to Pooling Layers. Edge segmentation is important for semantic segmentation, especially in semantic segmentation oriented towards road scene understanding. This paper argues that the reason why max pooling can achieve better segmentation is related to its ability to preserve better boundary information. The experimental results are shown in Figure 7, indicating that the model with the max pooling layer (MAX) exhibits better segmentation results.

5.8. Comparison of the Proposed Model and Common Models. Provided with the same memory and calculation condition, performance of the proposed model is superior to some efficient convolutional neural networks under the standard metrics and introduced performance metrics, with the test results given in Table 3.

Referring to Table 3, the amount of parameters involved in the paper is very small, and the recognition and segmentation are fast.

Table 3: Comparison of experimental results.

	This paper	SegNet	RefineNet	DeepLab	PSPNet	LRR	Dilation-8	FCN-8s
Params	0.364	29.5	42.6	44.04	65.7	48	141.13	134.5
MIOU	80.01	59.10	82.40	79.70	85.40	79.30	75.30	67.20

Note: parameter unit: millions; velocity unit: frames per second (fps).

6. Conclusions

Image segmentation consists of creating partitions within an image into meaningful areas and objects. It can be used in scene understanding and recognition, in fields such as biology, medicine, robotics, and satellite imaging, amongst others. This paper focuses on ESPNet as the underlying network structure and proposed an improved ESPNet model based on ESPNet to optimize the segmental results of road scenes. The proposed model in this paper is verified on Cityscape, PASCAL VOC 2012, and other datasets. Under the same memory and computing conditions, its performance is better than some efficient convolutional neural networks in the standard metrics and the newly introduced performance metrics. In the processing of high-resolution images, it has the characteristics of fast, small size, low power consumption, and low delay and ensures the segmentation accuracy. Although the proposed method in this paper has achieved good experimental results, there is also the problem of high accuracy that cannot achieve nonreal-time performance. The main reason may be the fuzziness and unclear boundary semantics caused by less parameters in the aspect of pooling layer and HFF feature fusion.

Data Availability

The experimental datasets used in this work are publicly available, and the bundled data and code of this work are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant nos. 61472348 and 61672455, Humanities and Social Science Fund of the Ministry of Education of China under Grant no. 17YJCZH076, Zhejiang Science and Technology Project under Grant nos. LGF18F020001 and LGF21F020022, and Ningbo Natural Science Foundation under Grant no. 202003N4324.

References

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," 2014, https://arxiv.org/abs/1411.4038.

[2] L.-C. Chen, J. Barron, G. Papandreou et al., "Semantic image segmentation with task-specific edge detection using CNNs and a discriminatively trained domain transform," 2015, https://arxiv.org/abs/1511.03328.

[3] S. Zheng, S. Jayasumana, B. Romera-Paredes et al., "Conditional random fields as recurrent neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, Santiago, Chile, December 2015.

[4] S. Mehta, M. Rastegari, A. Caspi et al., "ESPNet: efficient spatial pyramid of dilated convolutions for semantic segmentation," 2018, https://arxiv.org/abs/1803.06815.

[5] M. Siam, M. Gamal, M. Abdel-Razek et al., "Rtseg: real-time semantic segmentation comparative study," in *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*, Athens, Greece, October 2018.

[6] X. Zhang, X. Zhou, M. Lin et al., "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, June 2018.

[7] F. Chollet, "Xception: deep learning with depthwise separable convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[8] A. Rosebrock, *ImageNet: VGGNet, ResNet, Inception, and Xception with Keras*, Springer, Berlin, Germany, 2017.

[9] L. Gómez-Chova, R. Zurita-Milla, and L. Alonso, "Gridding artifacts on medium-resolution satellite image time series: MERIS case study," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 49, no. 7, pp. 2601–2611, 2011.

[10] K. Ullrich, E. Meeds, and M. Welling, "Soft weight-sharing for neural network compression," in *Proceedings of the International Conference on Representation Learning (ICLR)*, Toulon, France, April 2017.

[11] R. Gong, X. Liu, S. Jiang et al., "Differentiable soft quantization: bridging full-precision and low-bit neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, Seoul, South Korea, 2019.

[12] H. Zhou, J. Alvarez, and F. Porikli, "Less is more: towards compact cnns," in *Proceedings of the European Conference on Computer Vision*, Amsterdam, Netherlands, October 2016.

[13] J. Kim and Y. S. Heo, "Efficient semantic segmentation using spatio-channel dilated convolutions," *IEEE Access*, vol. 7, pp. 154239–154252, 2018.

[14] N. Nuechterlein and M. Sachin, "3D-ESPNet with pyramidal refinement for volumetric brain tumor image segmentation," *Lecture Notes in Computer Science*, vol. 11384, pp. 245–253, 2018.

[15] D. Franklin, "Nvidia jetson TX2 delivers twice the intelligence to the edge," 2017, https://developer.nvidia.com/blog/jetson-tx2-delivers-twice-intelligence-edge/.

[16] D. Zhao, L. Liu, F. Yu et al., "Ant colony optimization with horizontal and vertical crossover search: fundamental visions for multi-threshold image segmentation," *Expert Systems with Applications*, vol. 167, pp. 1–45, 2021.

[17] H. Liu, Y. Yao, Z. Sun et al., "Road segmentation with image-LiDAR data fusion in deep neural network," *Multimedia Tools and Applications*, vol. 79, no. 47-48, pp. 35503–35518, 2020.

[18] J. Ji, S. Li, J. Xiong et al., "Semantic image segmentation with propagating deep aggregation," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 12, pp. 9732–9742, 2020.

[19] F. Reis, R. Almeida, E. Kijak et al., "Combining convolutional side-outputs for road image segmentation," in *Proceedings of the International Joint Conference on Neural Networks*, Budapest, Hungary, July 2019.

[20] B. Parajuli, P. Kumar, T. Mukherjee et al., "Fusion of aerial lidar and images for road segmentation with deep CNN," in *Proceedings of the 26th ACM-SIGSPATIAL International Conference on Advances in Geographic Information Systems*, Seattle, WA, USA, November 2018.

[21] D. Yang, J. Gan, and Y. Luo, "Urban road image segmentation algorithm based on statistical information," in *Proceedings of the 26th International Conference on Geoinformatics*, June 2018.

[22] V. Badrinarayanan, A. Kendall, and R. Cipolla, "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.

[23] A. Akagic, E. Buza, and S. Omanovic, "Pothole detection: an efficient vision based method using RGB color space image segmentation," in *Proceedings of the 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, May 2017.

[24] Y. Zhang, P. Sun, J. Li et al., "Real-time vehicle detection in highway based on improved adaboost and image segmentation," in *Proceedings of the IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER)*, Shenyang, China, June 2015.

[25] J. Ran, C. Gang, H. Anthony, K. Tung et al., "An optimized iterative semantic compression algorithm and parallel processing for large scale data," *KSII Transactions on Internet and Information Systems*, vol. 12, no. 6, pp. 2761–2781, 2018.

[26] J. Ran, C. Gang, H. Anthony, K. Tung et al., "DIM A distributed spatial query processing based on mapreduce for spatial query in road networks," *EURASIP Journal on Wireless Communications and Networking*, vol. 208, pp. 1–15, 2018.

[27] J. Ran, K. Chunhai, L. Ruijuan, and G. Tao, "A common framework of partition-based clustering for large scale dataset using sampling and its mapreduce implementation," *Tehnicki vjesnik-Technical Gazette*, vol. 23, no. 1, pp. 25–33, 2016.

[28] H. Zhao, J. Shi, X. Qi et al., "Pyramid scene parsing network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Honolulu, HI, USA, July 2017.

[29] S. Mehta, M. Rastegari, L. Shapiro et al., "ESPNetv2: a light-weight, power efficient, and general purpose convolutional neural network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, USA, June 2019.

[30] I. Goodfellow, D. Warde-Farley, M. Mirza et al., "Maxout networks," in *Proceedings of the International Conference on Machine Learning*, Atlanta, GA, USA, June 2013.

[31] D. Clevert, T. Unterthiner, and S. Hochreiter, *Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUS)*, ICLR, San Juan, Puerto Rico, 2016.

[32] W. Liu, Y. Wen, Z. Yu et al., *Large-Margin Softmax Loss for Convolutional Neural Networks*, ICML, New York, NY, USA, 2016.

[33] Y.-L. Boureau, J. Ponce, and Y. Lecun, "A theoretical analysis of feature pooling in visual recognition," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, Haifa, Israel, June 2010.

[34] M. Everingham, L. Van Gool, C. Williams et al., "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.

[35] A. Geiger, P. Lenz, C. Stiller et al., "Vision meets robotics: the kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.