J Inf Process Syst, Vol.17, No.5, pp.892~904, October 2021
https://doi.org/10.3745/JIPS.02.0164

ISSN 1976-913X (Print)
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

Image Semantic Segmentation Using Improved ENet Network

Chaoxian Dong*

Abstract
An image semantic segmentation model is proposed based on improved ENet network in order to achieve the low accuracy of image semantic segmentation in complex environment. Firstly, this paper performs pruning and convolution optimization operations on the ENet network. That is, the network structure is reasonably adjusted for better results in image segmentation by reducing the convolution operation in the decoder and proposing the bottleneck convolution structure. Squeeze-and-excitation (SE) module is then integrated into the optimized ENet network. Small-scale targets see improvement in segmentation accuracy via automatic learning of the importance of each feature channel. Finally, the experiment was verified on the public dataset. This method outperforms the existing comparison methods in mean pixel accuracy (MPA) and mean intersection over union (MIOU) values. And in a short running time, the accuracy of the segmentation and the efficiency of the operation are guaranteed.

Keywords
Bottleneck, Image Semantic Segmentation, Improved ENet, MIOU, MPA, SE Module

1. Introduction

Image semantic segmentation technology is designed to label each pixel in the image with semantic information, thereby segmenting the image into several regions with different attributes and categories. Image semantic segmentation is a basic research content in the field of computer vision [1]. Image semantic segmentation technology can be applied to many fields such as medical imaging and geographic remote sensing, providing a strong guarantee for intelligent upgrades such as medical auxiliary diagnosis and remote sensing image interpretation [2]. However, the current semantic segmentation technology still needs to overcome the problems such as loss of small-scale objects, discontinuous segmentation, and incorrect segmentation. Therefore, how to enhance the representation ability of spatial detail information is the key research content to improve the segmentation accuracy [3].

In traditional image segmentation, the classic methods range from threshold segmentation, the simplest one, to region growth, edge detection, and graph partition. Among them, normalized cut and GrabCut are two classic segmentation methods based on graph partitioning [4]. Normalized segmentation uses the minimum se gmentation algorithm in graph theory to semantically segment the image. GrabCut is an interactive image segmentation method that uses image texture and boundary information to obtain better

Manuscript received November 19, 2020; first revision May 17, 2021; accepted May 23, 2021.

Corresponding Author: Chaoxian Dong (dong_chaoxian@163.com)
* College of Information Media, Sanmenxia Polytechnic, Sanmenxia, Henan, China (dong_chaoxian@163.com)

foreground and background segmentation results with a small amount of user interaction [5]. Although the computational complexity is not high, the traditional image segmentation algorithm is devoid of no data training stage and it has limited segmentation performance on more difficult segmentation tasks.

With the continuous improvement of classification network performance, there has been growing interest in solving the pixel-level labeling problem with semantic segmentation [6]. Compared with traditional image segmentation methods, the semantic segmentation method based on deep learning automatically learns features from the data, rather than using manually designed features. Using deep neural networks can achieve end-to-end semantic segmentation prediction [7]. Deep learning uses multilayer neural networks to automatically learn high-level features from a large amount of training data. Deep learning has been widely used in a variety of computer vision tasks [8].

However, the technical difficulties of image semantic segmentation in the three aspects of target, category and background still need to be resolved [9,10]. Therefore, an image semantic segmentation using an improved ENet network is proposed. The innovations of the proposed method are:

(1) The proposed model reduces convolution operation in decoder and adopts initialization operation to generate fusion features. In addition, the adaptive bottleneck structure is used to accelerate the segmentation to a great extent.

(2) To improve the accuracy of image semantic segmentation in complex environment, squeeze-and-excitation (SE) module is incorporated into the proposed model. Through learning, the importance of each feature channel is automatically obtained, the weight of useful features is improved, and the features that are not useful for the current segmentation task are suppressed, so as to achieve accurate segmentation of small samples.

The rest of this paper includes: Section 2 summarizes related work, classifies existing image semantic segmentation methods, and analyzes their advantages and disadvantages. Section 3 elaborates on the proposed method, applying the improved ENet model to image semantic segmentation. In Section 4, experiments and discussions are carried out, and the performance of the method in this paper is evaluated. Section 5 is the conclusion.

2. Related Work

The traditional image segmentation algorithm divides the image into different regions based on its color, texture information and spatial structure. The same region has consistent semantic information, and the attributes of different regions are different [11]. At present, the mainstream image semantic segmentation algorithm mainly through feature extraction, restoration, fusion, optimization four processes to obtain the target region of interest in the image to be segmented.

In the feature extraction stage, a large number of downsampling and pooling operations lead to the loss of spatial and detailed information, such as fully convolutional networks (FCN). Therefore, a dilated convolution and spatial pyramid pooling module was subsequently proposed to enhance global semantic information [12]. Network models such as DeepLabV1, dense relation network (DRN) and other network models increase the receptive field through serial dilated convolution and obtain richer spatial features. In [13], the authors expands the traditional method by developing a deeper network architecture with a smaller kernel to enhance its discrimination ability. Guo et al. [14] proposed a novel dense-Gram network, which uses clean images and degraded images for training through a pre-processing module based on image restoration, and fine-tunes the pre-trained network. Different from the traditional image semantic segmentation strategy, it can reduce the gap more effectively and realize image degradation segmentation. Reference [15] proposed a prototype image segmentation architecture based on convolution neural network (CNN) to realize automatic laparoscopic control of cholecystectomy. By establishing a recursive network structure that includes multiple use of sub-networks, to alleviate overfitting. The amount of computation, however, is tremendous for this type of method [16]. Network models such as DeepLabV2 and DenseASPP use the spatial pyramid pooling module to extract global semantic information and achieve denser multi-scale feature extraction. The above methods, however, will cause a checkerboard effect, resulting in the loss of local information and the discontinuity of semantic information. Feature restoration restores the resolution of the feature map by upsampling the feature map, which is used for model classification prediction [17]. Methods such as bilinear interpolation and deconvolution have certain limitations in restoring the resolution of feature maps. Zheng et al. [18] extracts high-level semantic features from the network and builds a dense deconvolution network. Finally, super pixel segmentation and spatial smoothness constraints are used to further improve image segmentation recognition results. Although the above methods enhance the expression of features, the segmentation ability of small-scale targets still needs to be improved [19].

Feature fusion obtains richer semantic information through the addition fusion of feature maps, splicing fusion, and cross-layer fusion to improve segmentation accuracy. Addition or splicing is often used to fuse multi-scale features [20]. FCN, U-Net, RefineNet, DeepLabV3+ and other network models adopt the idea of cross layer fusion, which combines the shallow detail features with the deep abstract features. It enhances the representation ability of high-resolution detail information, and opens up a new idea for the research of semantic segmentation. Inspired by the architecture of residual network and deconvolution network, Ozturk and Akdemir [21] proposed an automatic semantic segmentation based on cell type using a new deep CNN (DCNN). Four kinds of semantic information in medical image recognition are proposed and a new DCNN architecture is created. Feature optimization usually uses conditional random fields or Markov random fields to optimize the prediction results of semantic segmentation. By combining low-level image information with pixel-by-pixel classification results, the ability of model capturing fine-grained is improved [22]. In [23], the authors proposed a graph model initialized by a fully convolutional network named Graph-FCN. The graph convolution is used for semantic segmentation and achieves very good results.

With small target segmentation and recognition as the heart of the issue, most of the existing research uses multi-scale fusion enhanced network semantic segmentation algorithm to improve the accuracy of small-scale target segmentation. Zhou et al. [24] constructed the difference merging module in DCNN to extract the edge gradient of the object and obtain better boundary in the segmentation result. Then, the pyramid pooling module and the space-free pyramid pool are combined to extract image global features and contextual structure information by establishing long-distance dependence between pixels. This method stands outs from the traditional methods for it simplifies the original image preprocessing and post-processing steps.

3. Proposed Method

3.1 Solving Steps of Classic Ant Colony Algorithm

ENet network is a lightweight image semantic segmentation network capable of achieving pixel level semantic segmentation. The network features few parameters and fast calculation speed, which meets the real-time and accuracy requirements of image semantic segmentation. At the same time, ENet network also has a certain degree of plasticity. Based on this, the ENet network is pruned and convolution optimized, and integrated into the SE module to automatically learn the importance of each channel. An improved ENet network is proposed to better perform semantic segmentation tasks.

3.2 ENet Network Structure

ENet network adopts the current lightweight encoder-decoder network structure. As network specially designed for low-latency operation tasks, it has huge advantages in model size and parameter amount [25]. The ENet network changes the previous encoder-decoder symmetrical structure, reduces the convolution operation in the decoder, and enhances the processing speed tremendously. The ENet network has an initialization operation for the input image, as shown in Fig. 1. Its main purpose is to generate feature maps, and merge the feature maps generated by pooling and convolution operations.

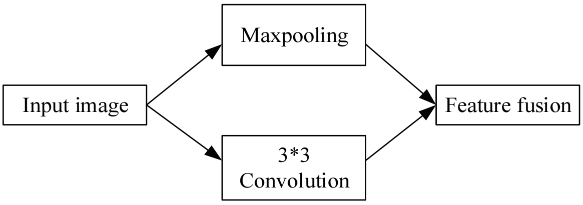

Fig. 1. Initialization operation.

The convolution operation has a total of 14 3×3 filters with a sliding step of 2, and a total of 14 feature maps are obtained. Maxpooling is a non-covered 2×2 sliding window, and four feature maps are obtained. Finally, a total of 18 feature maps are obtained after fusion [26]. In addition, a bottleneck convolution structure is also used in the ENet network. This module runs through the ENet network and is mainly used in encoder-decoder. The specific structure is shown in Fig. 2. Each packaged convolution module contains three convolutional layers [27].

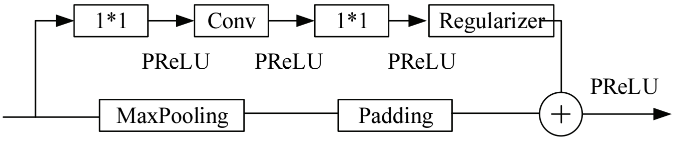

Fig. 2. The structure of bottleneck convolution.

From left to right in Fig. 2 are 1 1×1 projection mapping (used to reduce dimensions), 1 main convolutional layer and 1 1×1 ascending dimension; normalization and PReLU activation operations are performed between convolutional layers. The bottleneck convolution module is not static and will change according to specific operations. If it is a downsampled bottleneck convolution module, the 1×1 projection mapping is replaced by a Maxpooling layer with a kernel size of 2×2 and a step size of 2, and it is filled with 0 to match the size of the feature map. Conv is a 3×3 conventional convolution, expansion convolution or full convolution, and sometimes 1×5 and 5×1 asymmetric convolution are used instead. Regularizer uses Spatial Dropout to solve the problem of model overfitting.

The overall architecture of the ENet network consisting of five parts is between initialization and final full convolution. The first part is 1 downsampling bottleneck convolution module and 4 ordinary convolution bottleneck convolution modules. The second part is the Maxpooling bottleneck convolution module, followed by 8 different bottleneck convolution modules. The third part is 8 different bottleneck convolution modules. The fourth part is 1 upsampled bottleneck convolution and 2 ordinary bottleneck convolutions. The fifth part is an upsampled bottleneck convolution and an ordinary bottleneck convolution module. Finally, full convolution outputs the final result of image semantic segmentation. The fourth and fifth parts do not use the expanded convolution module because the encoding modules of the first three parts have already segmented the image completely, and there is no need to expand the field of view to extract feature information. The decoding structure mainly serves to restore the resolution of the image and improve the efficiency of the network model operation [28,29].

3.3 SE Module

The SE structure is the interdependence of the modeling feature channels displayed on the channel domain, and is used for feature recalibration. The core of the SE module is compression squeeze and excitation. After the convolution operation has obtained the features with multiple channels, the SE module can be used to re-calibrate the weight of each feature channel [30]. The SE module is divided into three steps, namely compression, excitation and reweighting. The schematic diagram is shown in Fig. 3.

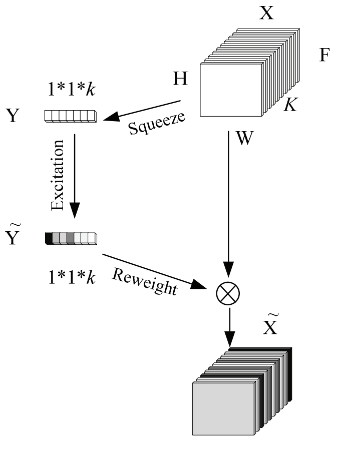

Fig. 3. SE structure.

The given feature map is $X, X \in R^{H \times W \times K}$, where H, W and K refer to the height, width and number of channels of the feature map respectively. After a compression operation (global average pooling), $y \in R^{K \times 1}$ is generated. Where y_m is the y^{th} m element of, and X_m is the m^{th} feature map of X:

$$y_m = F(X_m) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_m(i,j) \tag{1}$$

The excitation operation is realized by using two parameters W_1, W_2, a fully connected layer and two activation functions to generate $\tilde{y} \in R^{K\times 1}$ as shown below:

$$\tilde{y} = F(y,W) = \sigma(g(Y,M)) = \sigma(W_2\phi(W_1 y)) \quad (2)$$

where, σ represents the sigmoid activation function, and ϕ represents the ReLU activation function.

The final step is to determine the weighting operation. By multiplying the weight obtained by the excitation operation with the previous feature channel by channel, the recalibration of the feature on the channel domain is completed. Generate the rescaled feature map cluster $\tilde{X} \in R^{H\times W\times K}$, each feature map $\tilde{X}_m \in R^{H\times W\times 1}$ is as follows:

$$\tilde{X}_n = F_{scale}(X_m, y_m) = X_m \cdot y_m \quad (3)$$

where, $F_{scale}(X_m, y_m)$ is a channel by channel multiplication, $\tilde{X}_n$ is the n^{th} characteristic graph of $\tilde{X}$.

3.4 ENet Network Architecture Integrated with SE Module

Image semantic segmentation involves three stages: data preprocessing stage, training stage and testing stage. Labelme is used for manual labeling.in the data preprocessing stage. Training data and test data are generated by cutting the research area, where the training data includes the training set and the validation set. K-fold cross-validation is used to realize the automatic division of training set and validation set [31]. In the training phase, the pre-processed training samples are put into the improved ENet network fused with the SE module. The model architecture of the network is shown in Fig. 4.

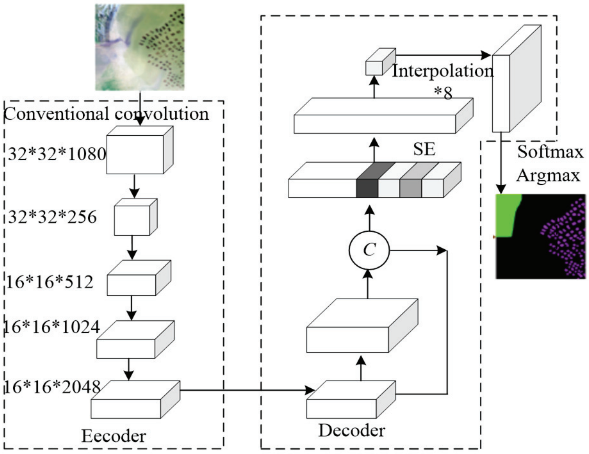

Fig. 4. Improved ENet network architecture with SE module.

The improved ENet network adopts an encoder-decoder structure. The encoder uses conventional convolution and a residual structure with dilated convolution to extract high-level semantic features, followed by batch normalization and PReLU activation functions after each layer of convolution. The decoder reduces the convolution operation, but incorporates the Bottleneck convolution structure. The SE module automatically learns the importance of each channel to better perform semantic segmentation tasks. The feature map is then restored to the original image size by linear interpolation, and the softmax activation function and the argmax function are used to obtain the final segmentation result, thereby achieving the end-to-end classification task.

4. Experimental Results and Analysis

In the experiment, the TensorFlow deep learning framework released by Google was used to construct an improved ENet network. The proposed model is experimentally demonstrated based on the PyThon simulation platform. The GPU model is RTX 2080Ti, the operating system is Ubuntu 16.04, the CPU model is i7-8700k, and the memory is 12G.

4.1 Network Parameter Setting

When training the network, the input image undergoes local response normalization before the first layer of convolution, $\alpha = 0.0001$, $\beta = 0.75$. The learning rate is set to 0.001, the weight attenuation is set to 0.0001, and the number of iterations is set to 60,000. The data set is randomly shuffled during the training process, and the batch size is set to 5. Use the cross entropy loss function and add L2 regularization to the network to prevent overfitting.

4.2 Evaluation Index

In order to evaluate the performance of the network model in this paper, we used the following evaluation indicators.

(1) Running time: Including training time and test time. In some cases, it is difficult to determine the exact running time of the model, because it depends to a large extent on the hardware device and the background implementation. However, providing information about the hardware and running time of the model can help evaluate its effectiveness.

(2) Accuracy: Pixel accuracy (PA) refers to the ratio of correctly classified pixels to the total pixels. In case unbalanced categories arise in the test set, the pixel accuracy rate cannot serve as a reliable indicator of the model's performance. Two evaluation indicators, therefore, are defined here: mean pixel accuracy (MPA) and mean intersection-over-union (MIOU).

Suppose there are a total of $c + 1$ categories. p_{ij} is the number of points for predicting i type as j type; p_{ii} represents the number of points whose true value is i and predicted value is i; p_{ij} represents the number of points whose true value is i and predicted value is j; p_{ji} represents the number of points whose true value is j and predicted value is i. Then MIOU is calculated as follows:

$$mIOU = \frac{1}{c+1}\sum\nolimits_{i=0}^{c} \frac{p_{ii}}{\sum\nolimits_{j=0}^{c} p_{ij} + \sum\nolimits_{j=0}^{c} p_{ji} - p_{ii}} \tag{4}$$

MPA is the average of pixel accuracy of each category, and is calculated as follows:

$$MPA = \frac{1}{c+1}\sum\nolimits_{i=0}^{c} \frac{p_{ii}}{\sum\nolimits_{j=0}^{c} p_{ij}} \tag{5}$$

4.3 CamVid Dataset

CamVid is the earliest semantic segmentation data set used in the field of autonomous driving. At first, five video sequences with a resolution of 960×720 pixels were shot on the car dashboard, with the shooting angle basically the same as that of the driver. Using image annotation software, 700 images were continuously annotated in the video sequence, including 32 categories such as buildings, trees, sky, roads, cars, and buses.

In order to more intuitively reflect the improvement of pixel category consistency by the improved ENet network, compare it with the segmentation results of the traditional ENet network, as shown in Fig. 5. To represent more intuitively the improvement of pixel category consistency with the improved ENet network, we compare it with the segmentation results of the traditional ENet network, as shown in Fig. 5.

As can be seen from Fig. 5, different from the traditional ENet model, this method adopts initialization operation to generate fusion features. The adaptive bottleneck convolution structure is used to replace the traditional convolution layer, and the fusion of the SE module can significantly improve the category consistency between adjacent pixels. And the misdetection of pixel categories contained in the same target is greatly reduced.

In order to further demonstrate the segmentation performance of the proposed model, we compare it with the models in [13,18,24]. The results of MPA and MIOU of each model on the CamVid data set are shown in Table 1.

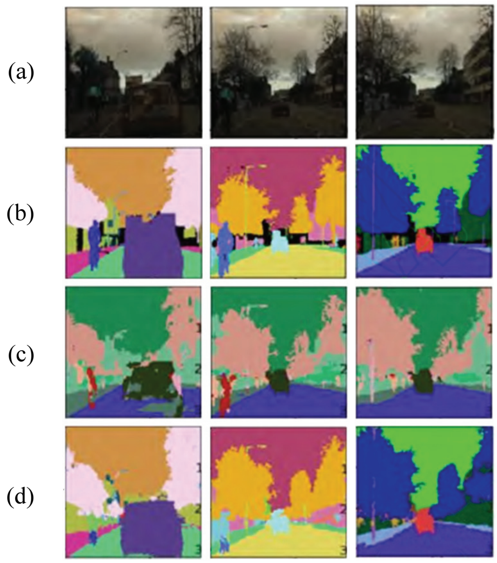

Fig. 5. Segmentation results of different models on CamVid dataset: (a) original image, (b) manual marking, (c) ENet, and (d) improved ENet.

Table 1. Comparison of the results of different methods on the CamVid dataset

Methods	MPA	MIOU
Yuan and Lo [13]	0.7690	0.6593
Zheng et al. [18]	0.7983	0.6810
Zhou et al. [24]	0.8047	0.7293
Proposed method	0.8385	0.7562

As can be seen from Table 1, the improved ENet network used in the proposed model outperforms other comparison methods in MPA and MIOU. In [13], the authors enhanced its discrimination ability by employing a smaller kernel and a deeper network architecture to achieve high-precision image segmentation. However, the model shows a low overall performance since it extraction accuracy is not desirable in complex environments. Zheng et al. [18] extracts high-level semantic features in deep convolutional networks, and introduces short connections in the deconvolution stage. Superpixel segmentation and spatial smoothness constraints are used to further improve the image segmentation recognition results. However, the accuracy of this method in the segmentation of small-scale targets still needs to be improved. Zhou et al. [24] constructed a difference merging module in DCNN, and established a long-distance dependency relationship between pixels through the combination of a pyramid pooling module and a space-free pyramid pool to extract image global features and contextual structure information, and achieved good results effect. Since the proposed model adopts an improved ENet network, the network is adaptively adjusted through the bottleneck convolution structure to better adapt to complex images, and the SE module is used to increase the weight of useful information. Therefore, the MAP and MIOU values have been further improved, reaching 0.8385 and 0.7562, respectively.

The running time is used as an evaluation index of the image segmentation model. Results of the running time comprised with the proposed model and different three methods [13,18,24] are shown in Table 2.

As can be seen from Table 2, of the three methods, the model in [13] has the shortest running time, 0.0537 seconds. The model is less time-consuming since it uses a learning network with a smaller kernel and has a simple structure. Methods in [18] and [24] have a long running time. Although these methods have good segmentation effects, their computational efficiency is sacrificed due to their complex structure. The improved ENet network in the proposed model reduces the convolution operation and integrates the SE module to speed up the extraction of useful features, thus ensuring segmentation accuracy and operating efficiency as well.

Table 2. Running time comparison of the CamVid dataset

Methods	Running time (s)
Yuan and Lo [13]	0.0537
Zheng et al. [18]	0.0726
Zhou et al. [24]	0.0873
Proposed method	0.0692

4.4 Cityscapes Dataset

The Cityscapes dataset contains 5,000 image scenes. The training set, validation set, and test set of the urban landscape dataset consist of 2,975,500 and 1,525 images, respectively, including 19 categories such as ground, building, sky, people, and vehicles.

Similarly, we compare it with the segmentation results of the traditional ENet network, as shown in Fig. 6.

As can be seen from the first column of Fig. 6, the proposed algorithm uses the SE module to improve the ENet network, which can effectively segment the pedestrians lost in the traditional ENet network, and improves the segmentation ability of small-scale targets. In the second column of segmentation results, the traditional ENet network mistakenly identified the bus rearview mirror as a pedestrian, while the improved network uses a weak bottleneck module to avoid segmentation errors for small targets. The segmentation results in the third column also prove that the improved network is better at segmenting and predicting small-scale targets than the traditional ENet network.

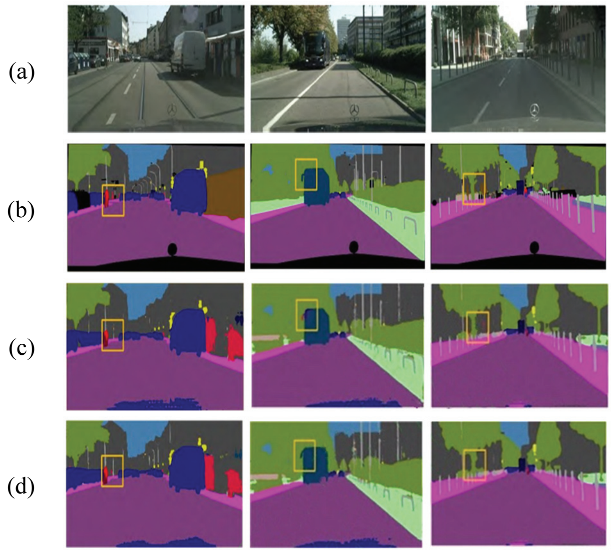

Fig. 6. Segmentation results of different models on Cityscapes dataset: (a) original image, (b) manual marking, (c) ENet, and (d) improved ENet.

In addition, the segmentation performance of the proposed model is compared with that of [13,18,24]. The results of MPA and MIOU of each model on the Cityscapes data set are shown in Table 3.

As can be seen from Table 3, the proposed model achieves the best segmentation effect with both MPA (0.9056) and MIOU (0.8465) higher than those gained with other models. Compared with the simple learning model in [13], the MPA and MIOU of the proposed model increased by 15.08% and 13.01%, respectively, and by 4.96% and 6.91%, respectively compared with [18]. Similarly, Zhou et al. [24] can obtain image global feature and context feature information, and achieve better segmentation effect. However, the proposed model integrates the SE module and the bottleneck convolution structure, and its performance is higher than that of [24] by 1.62% and 2.77%, respectively.

Table 3. Comparison of the results of different methods on the Cityscapes dataset

Methods	MPA	MIOU
Yuan and Lo [13]	0.7869	0.7491
Zheng et al. [18]	0.8628	0.7918
Zhou et al. [24]	0.8912	0.8237
Proposed method	0.9056	0.8465

In addition, Table 4 shows the running time comparison between the proposed model and the comparison models (e.g., [13,18,24]) on the Cityscapes dataset.

As can be seen from Table 4, the structure of [13] is simple, easy to train, and has the shortest running time. Methods in [18] and [24], both improve the learning network, and improve the segmentation accuracy through the optimized network model, but the running time is longer, with 0.0519 seconds and 0.0668 seconds, respectively. The improved ENet network in the proposed model reduces the convolution operation, and integrates the SE module to speed up the extraction of useful features. Therefore, while ensuring the accuracy of segmentation, the operating efficiency is ensured, and the running time is 0.0437 seconds. The proposed model, therefore, can ensure both segmentation accuracy and operating efficiency with a running time of 0.0437 seconds.

Table 4. Running time comparison of the Cityscapes dataset

Methods	Running time (s)
Yuan and Lo [13]	0.0326
Zheng et al. [18]	0.0519
Zhou et al. [24]	0.0668
Proposed method	0.0437

5. Conclusion

In recent years, the persistent development in automatic driving and security monitoring has placed higher requirements for model size, calculation cost, and segmentation accuracy in image semantic segmentation. For this reason, an image semantic segmentation model using improved ENet network is proposed. The ENet network is improved by using the initialization operation and the bottleneck convolution structure, and the SE module is integrated, and the importance of each feature channel is automatically acquired through learning. The improved ENet network integrated into the SE module is used for image segmentation of small targets in the complex environments. Finally, the proposed model is experimentally demonstrated based on the CamVid and Cityscapes datasets. Results show that the MPA and MIOU values of the three datasets of the proposed model are higher than other comparison methods, and the running time is shorter. For the Cityscapes dataset, its MPA, MIOU, and running time are 0.9056, 0.8465, and 0.0437 seconds, respectively. For the CamVid dataset, its MPA, MIOU, and running time are 0.8385, 0.7562, and 0.0692 seconds, respectively. The proposed model in this paper ensures both operating efficiency and segmentation accuracy.

In our future work, we will endeavor to improve the accuracy of target boundary segmentation and the ability to successfully segment small targets, and overcome the problem of discontinuous target segmentation, which will further improve the performance of semantic segmentation model.

References

[1] R. Moreno, M. Grana, D. M. Ramik, and K. Madani, "Image segmentation on spherical coordinate representation of RGB colour space," *IET Image Processing*, vol. 6, no. 9, pp. 1275-1283, 2012.

[2] Z. C. Jing, J. Ye, and G. L. Xu, "A geometric flow approach for region-based image segmentation-theoretical analysis," *Acta Mathematicae Applicatae Sinica, English Series*, vol. 34, no. 1, pp. 65-76, 2018.

[3] 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, 2017.

[4] T. Ling and W. Wu, “An image segmentation algorithm using visual saliency and graph cut,” *Paper Asia*, vol. 2, no. 1, pp. 114-118, 2019.

[5] A. V. Anjikar, K. Ramteke, and S. Chauvan, “Color image segmentation using region growth and merge improved technique,” *International Journal of Computer Sciences and Engineering*, vol. 7, no. 3, pp. 1070-1072, 2019.

[6] D. Stosic, D. Stosic, T. B. Ludermir, and T. I. Ren, “Natural image segmentation with non-extensive mixture models,” *Journal of Visual Communication and Image Representation*, vol. 63, article no. 102598, 2019. https://doi.org/10.1016/j.jvcir.2019.102598

[7] U. Anitha, S. Malarkkan, G. A. Jebaselvi, and R. Narmadha, “Sonar image segmentation and quality assessment using prominent image processing techniques,” *Applied Acoustics*, vol. 148, pp. 300-307, 2019.

[8] M. Aamir, Y. F. Pu, W. A. Abro, H. Naeem, and Z. Rahman, “A hybrid approach for object proposal generation,” in *The Proceedings of the International Conference on Sensing and Imaging*. Cham, Switzerland: Springer, 2017, pp. 251-259.

[9] Y. Wang, Q. Qi, Y. Liu, L. Jiang, and J. Wang, “Unsupervised segmentation parameter selection using the local spatial statistics for remote sensing image segmentation,” *International Journal of Applied Earth Observation and Geoinformation*, vol. 81, pp. 98-109, 2019.

[10] X. Wang, W. Li, C. Zhang, W. Lou, and R. Song, “An adaptable active contour model for medical image segmentation based on region and edge information,” *Multimedia Tools and Applications*, vol. 78, no. 23, pp. 33921-33937, 2019.

[11] T. Arora and R. Dhir, “A variable region scalable fitting energy approach for human Metaspread chromosome image segmentation,” *Multimedia Tools and Applications*, vol. 78, no. 7, pp. 9383-9404, 2019.

[12] G. Qin and Q. Li, “Pavement image segmentation based on fast FCM clustering with spatial information in Internet of Things,” *Multimedia Tools and Applications*, vol. 78, no. 5, pp. 5181-5191, 2019.

[13] Y. Yuan and Y. C. Lo, “Improving dermoscopic image segmentation with enhanced convolutional-deconvolutional networks,” *IEEE Journal of Biomedical and Health Informatics*, vol. 23, no. 2, pp. 519-526, 2017.

[14] D. Guo, Y. Pei, K. Zheng, H. Yu, Y. Lu, and S. Wang, “Degraded image semantic segmentation with dense-gram networks,” *IEEE Transactions on Image Processing*, vol. 29, pp. 782-795, 2019.

[15] T. Manabe, K. Tomonaga, K. Fujita, Y. Shibata, T. Kosaka, and T. Adachi, “CNN architecture for surgical image segmentation with recursive structure and flip-based upsampling,” *International Journal of Networking and Computing*, vol. 10, no. 2, pp. 259-276, 2020.

[16] R. Ratnakumar and S. J. Nanda, “A low complexity hardware architecture of K-means algorithm for real-time satellite image segmentation,” *Multimedia Tools and Applications*, vol. 78, no. 9, pp. 11949-11981, 2019.

[17] A. W. Rosyadi and N. Suciati, “Image segmentation using transition region and k-means clustering,” *IAENG International Journal of Computer Science*, vol. 47, no. 1, pp. 47-55, 2020.

[18] Y. Zheng, X. Zhang, F. Wang, T. Cao, M. Sun, and X. Wang, “Detection of people with camouflage pattern via dense deconvolution network,” *IEEE Signal Processing Letters*, vol. 26, no. 1, pp. 29-33, 2019.

[19] R. Jin and G. Weng, “Active contour model based on fuzzy c-means for image segmentation,” *Electronics Letters*, vol. 55, no. 2, pp. 84-86, 2019.

[20] M. Aamir, Y. F. Pu, Z. Rahman, W. A. Abro, H. Naeem, F. Ullah, and A. M. Badr, “A hybrid proposed framework for object detection and classification,” *Journal of Information Processing Systems*, vol. 14, no. 5, pp. 1176-1194, 2018.

[21] S. Ozturk and B. Akdemir, "Cell-type based semantic segmentation of histopathological images using deep convolutional neural networks," *International Journal of Imaging Systems and Technology*, vol. 29, no. 3, pp. 234-246, 2019.

[22] Y. Wu and S. Misra, "Intelligent image segmentation for organic-rich shales using random forest, wavelet transform, and hessian matrix," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 7, pp. 1144-1147, 2020.

[23] Y. Lu, Y. Chen, D. Zhao, and J. Chen, "Graph-FCN for image semantic segmentation," in *Advances in Neural Networks – ISSN 2019*. Cham, Switzerland: Springer, 2019, pp. 97-105.

[24] H. Zhou, A. Han, H. Yang, and J. Zhang, "Edge gradient feature and long distance dependency for image semantic segmentation," *IET Computer Vision*, vol. 13, no. 1, pp. 53-60, 2019.

[25] A. Abu and R. Diamant, "Enhanced fuzzy-based local information algorithm for sonar image segmentation," *IEEE Transactions on Image Processing*, vol. 29, pp. 445-460, 2019.

[26] Z. Huang, G. Huang, and L. Cheng, "Medical image segmentation of blood vessels based on Clifford algebra and Voronoi diagram," *Journal of Software*, vol. 13, no. 6, pp. 360-373, 2018.

[27] H. Fakhi, O. Bouattane, M. Youssfi, and H. Ouajji, "Distributed GPU-based k-means algorithm for data-intensive applications: large-sized image segmentation case," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 12, pp. 171-178, 2017.

[28] A. R. Subhamathi, "Ultrasound image segmentation based on information diffusion model," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 3, pp. 205-210, 2018.

[29] T. C. Zhang, J. Zhang, J. P. Zhang, M. L. Smith, and E. R. Hancock, "A novel model and method based on Nash equilibrium for medical image segmentation," *Journal of Medical Imaging and Health Informatics*, vol. 8, no. 5, pp. 872-880, 2018.

[30] S. Ren and F. Liu, "The optimal thresholding technique for image segmentation using fuzzy Otsu method," *Advances in Computational Sciences and Technology*, vol. 11, no. 6, pp. 445-454, 2018.

[31] A. K. M. Khairuzzaman and S. Chaudhury, "Masi entropy based multilevel thresholding for image segmentation," *Multimedia Tools and Applications*, vol. 78, no. 23, pp. 33573-33591, 2019.

Chaoxian Dong **https://orcid.org/0000-0003-0828-5410**

He was born in 1981. He has got his master's degree in computer science and technology. He graduated from Henan University of Science and Technology in 2010. He is an associate professor in Sanmenxia Polytechnic. His research interests include computer network and algorithm design.