Image semantic segmentation method based on GAN network and ERFNet model

Chaoxian Dong

College of Information Media, Sanmenxia Polytechnic, Sanmenxia, Henan, China

Correspondence
Chaoxian Dong, Sanmenxia Polytechnic, No.42 West Xiaoshan Road, Sanmenxia city, Henan province, 472000, China.
Email: dong_chaoxian@163.com

Abstract
This article addresses the problems of traditional methods in image semantic segmentation, such as insufficient segmentation of small-scale targets and weak anti-noise ability. A method of image semantic segmentation using a generative adversarial network (GAN) combined with ERFNet model is proposed. First, the asymmetric residual module (ARM) and weak bottleneck module are used to improve the ERFNet network model. Moreover, dilated convolution is used to reduce information loss and improve the performance of small target image semantic segmentation. Then, a U-shaped network is used to improve the generator of GAN to avoid low-level information sharing. In addition, the residual module is introduced into convolution layer to realize the dynamic adjustment of generator weight. Finally, the improved ERFNet model is used as the input of the generator to output the segmented image. It is input to the discriminator together with the label for judgment, which further improves the performance of image semantic segmentation. The proposed method is demonstrated experimentally based on the PyTorch platform. The results show that mean pixel accuracy and mean intersection over the union of the proposed method on the CamVid and Cityscapes datasets are higher than those of other comparisons. In addition, the execution time is short, and the overall image semantic segmentation performance is relatively ideal.

1 | INTRODUCTION

In recent years, the fields of computer vision, pattern recognition, surveying and remote sensing, and geographic information science have developed rapidly. Semantic segmentation, as a research focus in the above-mentioned fields, has extremely high research value and broad application prospects [1]. Semantic segmentation is a typical computer vision problem. It mainly takes some raw data such as planar images and three-dimensional point clouds as input and transforms it into a mask that highlights features through a series of technical processing [2]. Among them, image semantic segmentation is to segment different types of objects in a picture with semantic information annotation. The goal is to segment the scene image into different image areas related to semantic categories including roads, grasslands, sky and other backgrounds and discrete objects such as people, buildings, and cars [2]. This means that semantic segmentation tasks need to correctly identify different discrete objects and mark semantic information in a complex and changeable background. However, objects are often affected by phenomena such as occlusion, fragmentation and lighting, which increase the difficulty of semantic segmentation. The difficulty of image semantic segmentation mainly comes from three aspects, namely, target, category and background. In terms of targets, the same target under different conditions of illumination, viewing angle, distance, or when it is still and in motion, the images taken will be significantly different, and occlusion may also occur between adjacent targets [3, 4]. In terms of categories, there is a dissimilarity between similar targets and similarity between different targets. In terms of background, a simple background is generally helpful to realise the semantic segmentation of images, but the background in the actual scene is complex [5].

Image semantic segmentation is a key technology in the field of computer vision. It classifies all pixels in an image and extracts semantic information to perceive image content. Traditional image semantic segmentation methods mainly include threshold segmentation method, pixel clustering segmentation...
Due to limited computing power and no data training stage, the method in this period can only process some grayscale images, segmentation by extracting low-level features of the image, and cannot achieve semantic segmentation [11]. With the rapid development of graphics processing unit (GPU), deep learning (DL) technology is widely used in image processing, computer vision, medical imaging, robot control and other fields, bringing new opportunities for semantic segmentation technology [12]. A DL framework for medical magnetic resonance (MR) image segmentation is proposed in [13] through the use of stacked independent subspace analysis network to learn feature datasets in a hierarchical and unsupervised manner, and encode high-level semantic anatomical information. In [14], a new set of genetic operators is proposed for automatic identification of unimportant filters in the whole network. In addition, the pruning operator is designed to eliminate convolution filters from the layers involved in the concatenation of characteristic graphs. The high-order semantics of each pixel is mined by end-to-end training, and the biomedical image classification is realised. But the time cost of training mode is high, and the efficiency that can be achieved needs to be improved.

With the development of computer technology and the continuous improvement of computing power, after the introduction of the convolutional neural network, researchers have designed a series of network models such as Fully convolutional networks (FCN), DeepLab and SegNet for semantic segmentation, constantly refreshing the accuracy of image semantic segmentation [15]. In [16], a dual-path adversarial learning (DAL) was proposed by improving the learning method of the generative adversarial network (GAN) and utilising the image feature extraction characteristics of deep neural network and confrontation learning. Learning region of interest (ROI) function in different complexity in a controlled way, there is still room for improvement in the actual application of complex image segmentation. The authors in [17] proposed a SegNet deep fully convolution neural network architecture for semantic pixel segmentation. Its core consists of an encoder network, a corresponding decoder network and a pixel by pixel classification layer. The architecture of the encoder network is topologically the same as the 13 convolutional layers in the VGG16 network. The function of the decoder network is to map the low-resolution encoder feature map to the full input resolution feature map for pixel classification. But the data interaction between the two-layer network will delay the learning efficiency of the model. In the field of autonomous driving, reference [18] proposed a model called SegFast, which aims to conduct a comprehensive analysis of a large number of driving datasets, and improved the possibility of compact semantic image segmentation, in order to deploy its AI-based extended solution to more vehicles while ensuring the reliability of the model. However, this method has weak anti-interference ability, and its robustness needs to be improved.

Aiming at the problems of the above methods that have poor anti-interference ability and are not suitable for the semantic segmentation of small targets in complex environments, an image semantic segmentation method using GAN network combined with ERFNet model is proposed. The innovative points of the proposed method are:

1. Due to the large amount of ERFNet model parameters, which affects the image segmentation effect, the proposed method combines the asymmetric residual module (ARM) and the weak bottleneck module to improve the ERFNet network model. That is, when the amount of feature map calculation is large, ARM is used to increase the running speed. When there are many module channels and large network parameters, the weak bottleneck module is used to reduce the parameter amount and reduce the precision loss so that it is accurately and efficiently suitable for semantic segmentation tasks.

2. In order to avoid the low-level sharing of information between input and output, a U-shaped network is used to improve the network structure of the generator. Then, the residual module is introduced in the convolutional layer for difference learning so as to better adjust the weight of the generator to improve the accuracy of image semantic segmentation in complex environments.

The full text is divided into five sections in total. Section 1 introduces the research significance and technical difficulties of image semantic segmentation, as well as a summary of the innovative points of the proposed method. Section 2 describes the semantic segmentation method based on ERFNet network, including the improvement of ERFNet network, dilated convolution and other operations, and the principle of the GAN network. Section 3 introduces the GAN network in detail. The improved ERFNet model is used as the generator of GAN, and the discriminator is used to improve the accuracy of semantic segmentation.
2 RELATED THEORIES AND METHODS

2.1 Semantic segmentation method based on ERFNet

The proposed image semantic segmentation method is based on the ERFNet model, which uses an end-to-end encoder-decoder structure. Its overall structure is shown in Figure 1.

The image is first input to the encoder. The encoder part is similar to the basic classification network structure and outputs several channels of small resolution feature maps. The decoder is connected behind the encoder to upsample the small-resolution feature map to the initial resolution output. ERFNet network model is improved by using ARM and weak bottleneck modules. When the calculation amount of the feature map is large, the ARB is used to increase the running speed. When the module has many channels and the network parameter is large, the weak bottleneck module is used to reduce the parameter amount and reduce the accuracy loss. This can be accurately and efficiently applied to semantic segmentation tasks.

2.1.1 Encoder-decoder structure of ERFNet network

The improved ERFNet network model is mainly composed of downsampling module, ARM, weak bottleneck module and upsampling module. The encoder part includes two downsampling modules, five ARMs and eight weak bottleneck modules. The decoder part includes two upsampling modules, two ARMs and two weak bottleneck modules.

Among them, the downsampling of images is essential in image semantic segmentation. Not only can it expand the receptive field of the convolutional layer, thereby enriching the context information, it is beneficial to improve the classification accuracy; it can also significantly reduce the size of feature maps, reduce network computing complexity and memory usage, and enhance real-time segmentation [20]. The proposed network adopts the strategy of downsampling as soon as possible and directly performs two downsampling processing in the first two layers of the network to further improve the running speed of the network model. Due to the large size of the initial input image, it contains more redundant information. This leads to high memory usage and computational complexity, and the effect of improving the final output result is not obvious [21].

Due to the information loss and accuracy reduction caused by the downsampling operation, it is solved by fusing the feature information of the first few layers of the network and optimising the output result, but the improvement effect is limited [22]. At present, it is mainly improved by memorising the position index of the selected element in the maximum pooling layer and then combining them in the decoder part to generate an upsampling map. This method also relieves the pressure on memory space and computing power, but the accuracy is relatively low. Therefore, ERFNet uses transposed convolution in the decoder part for upsampling, which can effectively alleviate the loss of spatial information and reduce the loss of image accuracy. Deconvolution can perform image upsampling and convolution operations at the same time rather than through two separate processes, which effectively simplifies the network model [23].

2.1.2 ARM and weak bottleneck modules

The conventional residual and the bottleneck modules are two basic structures, and the proposed network is effectively improved and stacked on this basis. The functions implemented by the two are roughly the same, but they are slightly different in actual effects such as accuracy and efficiency, and each has advantages. The bottleneck module can effectively reduce the computational complexity while increasing the network depth. It is usually used to build very deep networks or networks that focus on efficiency. For the network layer with a large input feature map size, the conventional residual module is often
more computationally efficient than the bottleneck module, which can significantly improve the semantic segmentation effect [24].

Aiming at the problem of low efficiency of the conventional residual module, the asymmetric convolution is used to redesign it to obtain the ARM. The structure comparison is shown in Figure 2.

In the module, \( h \) is the number of feature map channels. Theoretically, it can be deduced that by stacking a \( 1 \times n \) convolution and a \( n \times 1 \) convolution, the same receptive field as a \( n \times n \) convolution can be achieved. But its network parameters are greatly reduced, and the model complexity is also significantly reduced. With the increase of \( n \), the reduction in the number of parameters and computational complexity will be more obvious. In the proposed network, a \( 1 \times 5 \) and a \( 5 \times 1 \) convolutional layers are stacked to replace a \( 5 \times 5 \) convolutional layer, thereby further decomposing the small-size convolution kernel. In the course of the experiment, it was found that all ARM networks could not achieve the desired effect. Therefore, ARM in the module with a large amount of feature map calculation is only used, which has the effect of reducing the number of network parameters and calculation complexity and has a small impact on the model segmentation accuracy.

As the depth of the network model increases, bottleneck modules are usually used to reduce the number of model parameters and loss of accuracy. But the bottleneck structure is often affected by degradation problems [25, 26]. Considering that each convolutional layer can be decomposed by a combination of filters, the resulting low-dimensional decomposition layer has a simple structure and can reduce computational cost. For this reason, a filter is used to redesign the bottleneck module, and the improved residual layer structure is called a weak bottleneck module as shown in Figure 3. Compared with the bottleneck module, this module has stronger computing power and fewer network parameters while maintaining the same learning ability and accuracy as the non-bottleneck module.

The proposed network designs some ARMs as weak bottleneck modules. In order to make the module calculation more sufficient, the number of channels of the module convolution layer is increased, and the space and depth information are decoupled by using grouped convolution, which reduces the number of parameters and the possibility of model overfitting. In order to make full use of context information, dilated convolution is used in the convolution layer of some weakly bottleneck modules. It can not only maintain the image spatial information but also increase the local receptive field of the network, thereby improving the accuracy of semantic segmentation [27].

### 2.2 Batch normalisation (BN)

The BN method ensures the consistency of data transmission in the network. Especially in the deep neural network model, the training process is prone to gradient explosion and gradient disappearance phenomena, which makes it difficult for the training results to converge. BN has shown good performance in some network models, speeds up the training of the model, and improves the generalisation ability of the model [28].

### 2.3 Dilated convolution

Dilated convolution can be regarded as a special convolution operation [29]. In the field of image semantic segmentation, the output label size needs to be consistent with the input image. The deep convolutional network structure contains multiple downsampling operations, and the semantic information becomes more abstract as the network deepens. However, in this process, a large amount of detailed information is inevitably lost and cannot be recovered. The dilated convolution alleviates this problem to a certain extent, expands the receptive field without reducing the resolution of the feature map, and does not increase the parameters of the convolution kernel.

In the one-dimensional case, the output \( y[i] \) of the hole convolution can be expressed as

\[
y[i] = \sum_{z=1}^{n} x[i+z] \cdot \varepsilon \cdot \varphi[n]
\]

where the one-dimensional convolution kernel is \( \varphi[n] \), \( n \) is the length of the convolution kernel, \( \varepsilon \) is the input sampling step size. When its value is 1, it is standard convolution. The two-dimensional situation is shown in Figure 4.

Figure 4(a) corresponds to a convolution kernel with a size of \( 3 \times 3 \) and a step size of 1, which is the standard convolution.

**FIGURE 2** Comparison between the conventional residual and asymmetric residual modules

**FIGURE 3** Structure of weak bottleneck and bottleneck modules
The receptive field of dilated convolution

Figure 4 corresponds to a convolution kernel with a size of $3 \times 3$ and a step size of 2. At this time, the size of the convolution kernel is still essentially $3 \times 3$, but the hole is 1, that is, only nine points are involved in the operation for the $7 \times 7$ area, and the other points can be regarded as the corresponding convolution kernel weights with the value as 0. Therefore, when the step size is 2, the convolution kernel with a size of $3 \times 3$ extends the receptive field to $7 \times 7$.

2.4 Deconvolution

Deconvolution is mainly used for upsampling in the field of image semantic segmentation. The deconvolution operation is the inverse process of the convolution operation, that is, the forward propagation process of the deconvolution operation is the backpropagation process of the convolution operation. Therefore, the multiplication matrix of the deconvolution can also be regarded as the permutation matrix of the convolution, so it is often called the permutation convolution [30]. The deconvolution operation can increase the resolution of the feature map and expand the receptive field [31].

2.5 Generative adversarial networks

GAN is composed of two parts: Generator and discriminator [32]. The discriminator is a simple convolutional neural network model, which takes real and fake images constructed by the generator as input. Through a series of convolution, excitation, normalization, and pooling layers, feature extraction of input data is performed, and the probability value of the interval $[0,1]$ is finally output. The generator is an inverse convolutional neural network model, which uses a series of deconvolution layers for upsampling and combines the excitation layer to convert low-dimensional vectors into vector outputs with the same dimensions as the real image. The input of the generator is randomly generated Gaussian white noise, which is decoded by the generator network and finally outputs a vector with the same size as the real image. Then, the difference between the value predicted by the discriminator and the label is calculated, and this error is directly used as the backpropagation error to update the parameters and the initial input vector [33]. The cross-entropy loss function is used to optimise the parameters of GAN, and the process can be expressed as

$$\min_G \min_D V(D, G) = E_{x \sim \mathcal{P}_{\text{data}}(x)} \{ \xi [D(x)] \} + E_{z \sim \mathcal{P}_{\text{G}(z)}} \{ \xi [1 - D(G(z))] \}$$

where $V$ is the loss function, $E$ represents the mean, $D$ is the discriminator model, $G$ is the generator model, $P$ is the vector distribution type, $x$ is the real image, $z$ is the input random noise vector, and $\xi$ is the model parameter. The discriminator needs to distinguish whether the input is a real image or an image generated by the generator. When the input is a real image, the value of $D(x)$ approaches 1; when the input is an image generated by the generator, the value of $D(x)$ approaches 0. The generator needs to adjust the distribution of $z$ as much as possible to minimise the difference between the distribution of $G(z)$ and $x$, and make $D(G(z))$ approach 1. Since GAN has two sub-network models, a stepwise crossover training method is adopted. The loss function when training the discriminator is

$$\min_D V(D, G) = E_{x \sim \mathcal{P}_{\text{data}}(x)} \{ \xi [D(x)] \} + E_{z \sim \mathcal{P}_{\text{G}(z)}} \{ \xi [1 - D(G(z))] \}$$

At this point, the default generator performance has reached the optimal, and only the classification performance of the discriminator needs to be trained. When training the loss function of the generator, it is considered that the discriminator has the best discriminative ability. That is, part $E_{x \sim \mathcal{P}_{\text{data}}(x)} \{ \xi [D(x)] \}$ is always a fixed constant, and training is not required. Therefore,
the loss function when training the generator is
\[
\min_G \mathcal{L}(D, G) = E_{x \sim \mathcal{D}_{G(x)}} \{ \mathbb{E} [1 - D(G(x))] \} \tag{4}
\]

3 | PROPOSED METHOD

3.1 | Generator network structure

The generator uses \( x \in \mathbb{R}^{W \times H \times C} \) as the input image, \( W = H = 512, C = 3 \). The structure of the ERFNet network is encoder-decoder. This network structure is continuously down-sampling until it reaches a bottleneck layer. It turns the extracted information into a one-dimensional vector, performs deconvolution at this point, gradually upsamples, and finally restores it to an image. Such a network structure requires all information to pass through all network layers, including the bottleneck layer. In many image segmentation problems, input and output share a large amount of low-level information that could directly pass through the network layer \([34, 35]\). In order to enable the generator to avoid this situation, the proposed method uses the idea of a U-shaped network structure. For the \( n \) layer network structure, the output of the \((n - i)\)th layer network is connected with the output of the \((n - i + 1)\)th layer node. The network structure is shown in Figure 5. Among them, A is the residual convolution unit, and B is the forward propagation convolution unit.

Leaky-ReLU (Rectified Linear Unit) is used as a non-linear activation function in the encoding layer, and BN is used in each layer. By normalising the current layer input (\( \mu = 0, \delta = 1 \)), it is helpful to accelerate the convergence of the entire network and improve the independence between layers \([36]\). The decoding layer uses ReLU as a non-linear function, and the Tanh activation function is used to generate images in the last layer.

In order to improve the accuracy of model segmentation, the residual network structure is added to the generator, the structure is shown in module A in Figure 6. The residual network structure can alleviate the gradient reduction, solve the problem of gradient disappearance, and improve the sensitivity of the network structure to weight changes so that the generator can fully learn the distribution of different types of images, thereby improving the effect of segmentation \([37]\). In order to reduce the complexity of the model and reduce the amount of calculation, and the number of training parameters, a \( 1 \times 1 \) convolution kernel is added as the bottleneck layer to reduce the dimensionality of the input feature layer before each \( 3 \times 3 \) convolution kernel is used. The number of channels is reduced to half of the original number and then restored to the original number of channels through a \( 3 \times 3 \) convolution kernel. The specific structure is shown in module B in Figure 6.

3.2 | Discriminator network structure

The network structure of the discriminator is shown in Figure 6, with a total of seven layers. Using \( \{x_i, y_i\}_{i=1}^N \) or \( \{x_i, G(x_i)\}_{i=1}^N \) as input, \( x \in \mathbb{R}^{512 \times 512 \times 3}, y \in \mathbb{R}^{512 \times 512 \times 1} \). When the input is \( \{x, y\} \), the correct output of the discriminator is 1, and when \( \{x, G(x)\} \) is input, the correct output of the discriminator is 0. In each convolutional layer in the discriminator, Leaky-ReLU is used as the non-linear activation function, and strided convolutions is used instead of pooling layer \([38]\).

4 | EXPERIMENTAL RESULTS AND ANALYSIS

In the experiment, the PyTorch1.1.0 DL framework is used to implement the proposed method. The system environment is set up under Ubuntu 16.04 and NVIDIA GTX 1070 graphics card for GPU acceleration is used. The CPU is Intel Core(TM) i7-8750H, and the memory is 64 G.

The batch selection of the training process consists of eight images. The optimiser uses the Adam optimiser with adaptive capabilities, and the initial value of the learning rate is set to 0.0002. An exponential decay strategy is used for the learning rate, and the momentum coefficient is 0.9.

4.1 | Evaluation index

In order to measure the performance of the semantic segmentation model, it is necessary to use general objective evaluation indicators to ensure the fairness of the evaluation. Running time, video memory occupancy and accuracy are three commonly used evaluation indicators. The algorithm application environment and test purpose are different, and the algorithm evaluation standards are also different. For example, for a real-time semantic segmentation model, the accuracy can be lost within a certain range to increase the calculation speed. For general algorithms, it is necessary to improve all measurement performance. The proposed method focuses on real-time image semantic segmentation, so the running time and accuracy are selected as its evaluation indexes.

1. Running time: The running time includes the training and testing time of the network model. Most models need to predict the segmentation results in real time. In some cases, it may be difficult to provide the exact running time of the model because the running time is very dependent on hardware devices and background implementation. However, providing information and running time of the model running hardware is helpful to evaluate its effectiveness and test the fastest execution method under the same environment.

2. Accuracy: Pixel accuracy (PA) refers to the ratio of correctly classified pixels to the total pixels. When the test set has unbalanced categories, the PA rate cannot objectively reflect the performance of the model. Therefore, two evaluation indicators are defined: Mean PA (MPA) and mean intersection-over-union (MIOU). Among them, MIOU is
FIGURE 5  Generator network structure
the most important performance evaluation index, which can better reflect the accuracy of the model.

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 the predicted value is \( i \); \( p_{ij} \) represents the number of points whose true value is \( i \) and the predicted value is \( j \); \( p_{ji} \) represents the number of points whose true value is \( j \) and the predicted value is \( i \). Then, MIOU is calculated as follows:

\[
mIOU = \frac{1}{c + 1} \sum_{i=0}^{c} \frac{p_{ii}}{\sum_{j=0}^{c} p_{ij} + \sum_{j=0}^{c} p_{ji} - p_{ii}}
\]

(5)

MPA is the average of PA of each category, calculated as follows:

\[
MPA = \frac{1}{c + 1} \sum_{i=0}^{c} \frac{p_{ii}}{\sum_{j=0}^{c} p_{ij}}
\]

(6)

4.2 Training process

When training the network, in order to enhance the generalisation ability of the network, the input image undergoes local response normalisation before the first layer of convolution, \( \alpha = 0.0001, \beta = 0.75 \). The objective loss function is optimised by using the Adam algorithm with a learning rate of 0.0002 and iterated until the loss function converges; the weight decays to 0.0001, and the number of iterations is set to 60,000. In the training process, the training dataset is randomly shuffled, and then every eight training images are treated as a batch, that is, batch size is set to eight. The objective function of the network uses the cross-entropy loss function, and L2 regularisation is added to the last layer of the network to prevent overfitting. Due to the large deviation of the number of pixels in each category in the dataset, the median frequency equalisation method is used to balance between classes. In order to prevent overfitting in the process of training the network, the data is enhanced by cropping and left-right flipping.

Because the improved ERFNet network structure is more complex, the network level is usually deeper. The training process consumes a lot of time, and the final result of the model may appear near the optimal solution in actual experiments. Therefore, in the process of training the network model, the idea of model retraining is adopted. First, the parameters of each layer are trained with less training samples, then the network model is initialised with these parameters, and finally, the final model is formally trained. In the training process, Adam stochastic gradient descent method is used to update the network parameters. According to the distribution of various objects, the values of various loss functions are weighted accordingly to solve the problem of uneven distribution of objects in the dataset. The convergence curves of the total loss function corresponding to the model retraining and the unused model retraining methods are shown in Figure 7.

It can be seen from Figure 8 that the convergence speed of using the model retraining method is significantly faster than
that of not using the model retraining method. As the iteration progresses, the convergence speed of model training tends to be flat and finally reaches a steady state.

4.3 CamVid dataset

CamVid is the earliest semantic segmentation dataset used in the field of autonomous driving. At first, five video sequences with a resolution of 960 × 720 pixels were shot on the dashboard of the car, and the shooting angle of view was 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, traffic lights, sky, roads, pedestrians, motorcycles, cars, and buses.

Since the CamVid dataset is relatively small, in order to prevent overfitting in the process of training the network, the data is enhanced by cropping and left-right flipping. In the experiment, the semantic categories of the dataset are divided into 10 categories. After 50 epochs, the training is stopped, and the changes of MIOU, MPA and loss of the validation set during the training process are shown in Figure 8.

It can be seen from Figure 8 that the MIOU reached 60% when the network was trained to the 13th epoch, and it reached 70% after 27 epochs. In the subsequent training, although MIOU has a very slow improvement, the effect is not obvious. The MPA reached 80% when the network iterated to 28 epochs. The Loss is reduced to about 15% when the network is iterated to 32 epochs. In the following training process, the model basically begins to converge. After 50 epoch iterations, the MIOU of the model reached 75%, the MPA basically remained at around 82%, and the Loss dropped below 12%.

When the accuracy of the network is no longer significantly improved during the training process, the test set is used to test the network model at this time. The visualisation results of the proposed method on the CamVid dataset are shown in Figure 9.

It can be seen from Figure 9 that compared to a single ERFNet model, the proposed method using GAN network combined with ERFNet model can significantly improve the category consistency between adjacent pixels. The misdetection of pixel categories contained in the same target is greatly reduced.

At the same time, the segmentation performance of the proposed method is compared and analysed quantitatively with [9, 16, 18]. The MPA and MIOU results of each method on the CamVid dataset are shown in Figure 10.

It can be seen from Figure 10 that the values of the proposed method MPA and MIOU are 0.85 and 0.78, respectively, which are higher than other comparative methods. Because the proposed method uses the improved ERFNet model as the generator, the result of image semantic segmentation is obtained through the discrimination processing of GAN, and the segmented image is closer to the result of manual annotation. The study in [9] realises image semantic segmentation based on the copy-movement forgery detection method based on superpixel segmentation and cluster analysis. This method mainly extracts low-level features of the image, the segmentation accuracy is generally poor, and its MIOU is only 0.55. In [16], the authors improved GAN network by using deep neural network and proposed a DAL model for image segmentation, but in practical application, the accuracy of complex image segmentation needs to be improved. A SegFast model is proposed in [18] for compact semantic image segmentation. Although the accuracy has been improved, its anti-interference ability is weak. Compared with the proposed method, the MPA and MIOU are reduced by 0.08 and 0.09, respectively.

In addition, the running time is used as an evaluation index of the image segmentation model. The running time of the proposed method and [9, 16, 18] are shown in Table 1.

| Methods | [9] | [16] | [18] | The proposed method |
|---------|-----|-----|-----|---------------------|
| Running time (ms) | 63.72 | 71.68 | 82.51 | 67.87 |

It can be seen from Table 1 that the execution time of image segmentation in [9] is the shortest, which is 63.72 ms. Since there is no data training stage, the computational complexity is not high, so it takes a short time. In [16, 18], the execution takes a long time to execute. Because of their complex structure, although the segmentation effect is better, the efficiency is low. The proposed method uses an improved ERFNet model, in which the ARM and weak bottleneck modules reduce the number of network parameters and the computational complexity, thus, ensuring the accuracy of segmentation while ensuring the efficiency of execution.

4.4 Cityscapes dataset

Cityscapes is a large-scale dataset of 5000 high-quality images collected from street scenes in 50 different cities. The training,
validation and test sets of the Cityscapes dataset consist of 2975, 500 and 1525 images, respectively, including 19 categories such as ground, building, road signs, nature, sky, people and vehicles. The dataset also provides 20,000 coarsely segmented images for training the performance of the classification network based on weakly supervised learning.

In the experiment, the proposed method is trained on the Cityscapes dataset for 50 epochs iteratively. The changes of MIOU, MPA and loss of the validation set during the training process are shown in Figure 11.

As can be seen from Figure 11, the MIOU reached 80% at the 25th epoch of the network training and 85% after the iteration to 37 epochs. In the subsequent training, the value of MIOU stabilised. The MPA reached 92% when the network iteration reached 30 epochs. The Loss value is reduced to about 6% when the network iteration reaches 30 epochs, and the model basically converges in the subsequent training process. After 50 epoch iterations, the model MIOU reached 86%, the MPA basically remained at about 92%, and the Loss dropped below 5%.

Similarly, in order to intuitively reflect the improvement of the image semantic segmentation performance of the proposed method on the Cityscapes dataset, it is compared with the segmentation result of a single ERFNet model. The comparison result is shown in Figure 12.

It can be seen from the first row of Figure 12 that the proposed algorithm can effectively segment the pedestrians lost in the ERFNet model and improve the segmentation ability of small-scale targets. In the second row of segmentation results, the ERFNet model incorrectly identified the rearview mirror of the bus as a pedestrian, and the improved model uses a weak bottleneck module to avoid segmentation errors for small targets.
In addition, the segmentation accuracy of the proposed method is compared with [9, 16, 18]. The MPA and MIOU results of each method on the Cityscapes dataset are shown in Figure 13.

As can be seen from Figure 13, the overall MPA and MIOU values of this dataset are higher than those of the CamVid dataset. Since the CamVid dataset has many types and finer classifications, the requirements for image semantic segmentation are higher. Compared with other comparison methods, the MPA and MIOU values of the proposed method are the highest, reaching 0.93 and 0.86 on the Cityscapes dataset because the ERFNet model in the proposed method uses the ARM and the weak bottleneck module to strengthen the image segmentation of small targets. As the generator of GAN, the performance of image semantic segmentation is further improved by the discriminator. Compared with the reference [9], the MPA and MIOU of the proposed method are increased by 12.15% and 11.98%, respectively; compared with [16], increased by 5.73% and 6.09%, respectively; compared with [18], increased by 2.45% and 2.36%, respectively.

In addition, the execution time comparison between the proposed method and the methods in [9, 16, 18] on the Cityscapes dataset is shown in Table 2.

It can be seen from Table 2 that the structure of [9] is simple and the execution time is the shortest. It can be seen from Figure 13 that the segmentation accuracy is the lowest, so the overall segmentation effect of this method is not good. In [16], using deep neural network to improve GAN network, the segmentation accuracy is guaranteed, but the execution time is increased to 51.93 ms. Similarly, in [18], SegFast model is used to realize image segmentation. The model is complex and the execution time is long, which is 66.82 ms. The proposed method uses an improved ERFNet model in which ARM and weak bottleneck modules reduce the number of network parameters and reduce the computational complexity. Therefore, the execution time is only higher than [9], which is 43.75 ms.

5 | CONCLUSION

In order to improve the accuracy of small target image segmentation, an image semantic segmentation method using GAN network combined with ERFNet model is proposed. ARM and weak bottleneck modules are used to improve the ERFNet network model, and dilated convolution is used to reduce information loss. At the same time, the U-shaped network is used to improve the GAN generator, and the residual module is introduced in the convolution layer to realize the dynamic adjustment of generator weight. Combined with the characteristics of the two, the improved ERFNet model is used as the generator to output the segmented image. The algorithm and the artificial standard image are input into the discriminator to further improve the performance of image semantic segmentation. The results show that the MPA and MIOU of the proposed
method on the CamVid dataset are 0.85 and 0.78, respectively. The MPA and MIOU on the Cityscapes dataset are 0.93 and 0.86, respectively. The values are higher than other comparison methods, and the running time is shorter. The running time on the two datasets is 67.87 ms and 43.75 ms, respectively. The overall image of semantic segmentation performance is the best.

In the future, the proposed method should not be limited to a small number of datasets but should train a semantic segmentation model with very good generalisation ability and use multiple image sets to evaluate it.

ACKNOWLEDGEMENTS
This work is not supported by any fund projects.

REFERENCES

1. Wang, W., et al.: Semantic segmentation of remote sensing ship image via a convolutional neural networks model. IET Image Proc. 13(6), 1016–1022 (2019)
2. Zhang, L., et al.: Image object detection and semantic segmentation based on convolutional neural network. Neural Comput. Appl. 32(7), 1949–1958 (2020)
3. Khelifi, I., Mignotte, M.: MC-SSM: Nonparametric semantic image segmentation with the ICM algorithm. IEEE Trans. Multimedia 21(8), 1946–1959 (2019)
4. Anthimopoulos, M., et al.: Semantic segmentation of pathological lung tissue with dilated fully convolutional networks. IEEE J. Biomed. Health Inf. 23(2), 714–722 (2019)
5. Zhou, H., et al.: Edge gradient feature and long distance dependency for image semantic segmentation. IET Comput. Vision 13(1), 53–60 (2019)
6. Guo, D., et al.: Degraded image semantic segmentation with dense-gram networks. IEEE Trans. Image Process. 29(99), 782–795 (2019)
7. Li, J., et al.: Multilevel thresholding selection based on variational mode decomposition for image segmentation. Signal Process. 147(06), 80–91 (2018)
8. Breve, F.: Interactive image segmentation using label propagation through complex networks. Expert Syst. Appl. 123(06), 18–33 (2019)
9. Liu, Y., et al.: A passive forensic scheme for copy-move forgery based on superpixel segmentation and K-means clustering. Multimedia Tools Appl. 79(1/2), 477-500 (2020)
10. Li, X., Ma, H., Luo, X.: Weaker supervised semantic segmentation with only one image level annotation per category. IEEE Trans. Image Process. 29(1), 128-141 (2019)
11. Yan, B., et al.: Semantic segmentation guided pixel fusion for image retargeting. IEEE Trans. Multimedia 22(3), 676–687 (2020)
12. Niu, X., et al.: Effective image restoration for semantic segmentation. Neurocomputing 374(01), 100–108 (2020)
13. Liao, S., et al.: Representation learning: A unified deep learning framework for automatic prostate MR segmentation. Med. Image Comput. Comput. Assisted Intervention 16(2), 254–261 (2013)
14. Zhou, Y., Yin G, G., Yi, Z.: Evolutionary compression of deep neural networks for biomedical image segmentation. IEEE Trans. Neural Networks Learn. Syst. 31(8), 2916–2929 (2020)
15. Li, Y., et al.: Joint CRF and locality-consistent dictionary learning for semantic segmentation. IEEE Trans. Multimedia 21(4), 875–886 (2019)
16. Bi, L., Feng, D., Kim, J.: Dual-path adversarial learning for fully convolutional network (FCN)-based medical image segmentation. Visual Comput. 34(6-8), 1043–1052 (2018)
17. Badrinarayanan, V., Kendall, A., Cipolla, R.: SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. 39(12), 2481–2495 (2019)
18. Ghosh, S., et al.: SegFast-V2: Semantic image segmentation with less parameters in deep learning for autonomous driving. Int. J. Mach. Learn. Cybern. 10(11), 3145–3154 (2019)
19. Romera, F., et al.: ERFNet: Efficient residual factorized ConvNet for real-time semantic segmentation. IEEE Trans. Intell. Transp. Syst. (1), 1–10 (2017)
20. Ates, H.F., Sunetci, S.: Multi-hypothesis contextual modeling for semantic segmentation. Pattern Recognit. Lett. 117, 104–110 (2019)
21. Peng, C., et al.: Densely based multi-scale and multi-modal fully convolutional networks for high-resolution remote-sensing image semantic segmentation. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12(8), 2621–2626 (2019)
22. Shang, R., et al.: Semantic segmentation for SAR image based on texture complexity analysis and key superpixels. Remote Sens. 12(13), 2141 (2020)
23. Ozturk, S., Akdemir, B.: Cell-type based semantic segmentation of histopathological images using deep convolutional neural networks. Int. J. Imaging Syst. Technol. 29(3), 234–246 (2019)
24. Igonin, I.A., Tramuttey, Y.V.: Comparative efficiency analysis for various neural architectures for semantic segmentation of images in remote sensing applications. Ope. Mem. Neural Networks 28(4), 306–320 (2019)
25. Zheng, C., Wang, L., Chen, X.: A hybrid Markov random field model with multi-granularity information for semantic segmentation of remote sensing imagery. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 12(8), 2728–2740 (2019)
26. Lin, D., Huang, H.: Zig-zag network for semantic segmentation of RGB-D images. IEEE Trans. Pattern Anal. Mach. Intell. 42(10), 2642–2655 (2020)
27. Zhao, Z., et al.: Lidar mapping optimization based on lightweight semantic segmentation. IEEE Trans. Intell. Veh. 4(3), 353–362 (2019)
28. Zhu, X., et al.: A novel framework for semantic segmentation with generative adversarial network. J. Visual Commun. Image Represent. 58, 532–543 (2019)
29. Bike, C., et al.: Importance-aware semantic segmentation for autonomous vehicles. IEEE Trans. Intell. Transp. Syst. 20(1), 137–148 (2019)
30. Jiang, J., et al.: RWSNet: A semantic segmentation network based on SegNet combined with random walk for remote sensing. Int. J. Remote Sens. 41(2), 487–505 (2020)
31. Pan, B., et al.: Cross-view semantic segmentation for sensing surroundings. IEEE Rob. Autom. Lett. 5(3), 4867–4873 (2020)
32. Goodfellow, I., et al.: Generative adversarial nets. Adv. Neural Inf. Process. Syst., 3(11), 2672–2680 (2014)
33. Jamshidi, M.B., et al.: Artificial intelligence and COVID-19: Deep learning approaches for diagnosis and treatment. IEEE Access, PP(99): 109581–109595 (2020)
34. Deng, L., et al.: Semantic segmentation-based lane-level localization using around view monitoring system. IEEE Sens. J. 19(21), 10077–10100 (2019)
35. Wen, S., et al.: Widening residual refine edge reserved neural network for semi-automated segmentation. Multimedia Tools Appl. 78(13), 18229–18247 (2019)
36. Xia, K., et al.: Liver semantic segmentation algorithm based on improved deep adversarial networks in combination of weighted loss function on abdominal CT images. IEEE Access 7(99), 96349–96358 (2019)
37. Mei, J., et al.: Semantic segmentation of 3D LiDAR data in dynamic scene understanding. IEEE Trans. Intell. Syst., 3(11), 2672–2680 (2014)
38. Pai, M.M.M., et al. Improved semantic segmentation of water bodies and land in SAR images using generative adversarial networks. Int. J. Semant. Comput. 14(01), 55–69 (2020)

How to cite this article: Dong C. Image semantic segmentation method based on GAN network and ERFNet model. J. Eng. 2021;2021:189–200. https://doi.org/10.1049/tje2.12025