Remote Sensing Image Semantic Segmentation Algorithm Based on Improved ENet Network

Yiqin Wang

School of Information Technology and Engineering, Jinzhong University, Jinzhong, Shanxi 030619, China

Correspondence should be addressed to Yiqin Wang; wangyiqin@jzxy.edu.cn

Received 6 September 2021; Revised 22 September 2021; Accepted 23 September 2021; Published 4 October 2021

A remote sensing image semantic segmentation algorithm based on improved ENet network is proposed to improve the accuracy of segmentation. First, dilated convolution and decomposition convolution are introduced in the coding stage. They are used in conjunction with ordinary convolution to increase the receptive field of the model. Each convolution output contains a larger range of image information. Second, in the decoding stage, the image information of different scales is obtained through the upampling operation and then through the compression, excitation, and reweighting operations of the Squeeze and Excitation (SE) module. The weight of each feature channel is recalibrated to improve the accuracy of the network. Finally, the Softmax activation function and the Argmax function are used to obtain the final segmentation result. Experiments show that our algorithm can significantly improve the accuracy of remote sensing image semantic segmentation.

1. Introduction

Image segmentation technology divides the image into different types of uniform areas according to the internal characteristics of the image. It is required that the dividing edge between regions must be accurately delineated; and the internal features of the segmented objects have consistency or similarity. Each area belongs to the same category, and different areas belong to different categories [1]. Its purpose is not only to simplify the information of the image but also to make the image easier to understand and analyze [2–4]. Remote sensing image segmentation technology aims to classify remote sensing images at the pixel level based on actual semantic information. It is divided into a series of areas with landmarks such as roads, farmland, villages, and industrial areas [5]. In recent years, with the advancement of high-resolution earth observation technology, image segmentation has been carried out through data obtained from remote sensing satellites. It is the processing basis for the application research of urban planning, disaster monitoring, target recognition, and so forth [6, 7]. However, the rapid increase in the amount of remote sensing data also brings about many challenges to the segmentation of optical remote sensing images. For example, an increase in spatial resolution brings about higher complexity of ground objects (more shadows and backgrounds). The phenomena of the same substance with different spectrum and foreign substance with the same spectrum are brought about by the drastic change of spectral information. There are huge amounts of data to be processed and features with variable scales to be extracted [8–10].

Deep convolutional neural networks have become the mainstream method to solve image semantic segmentation. The best fitting model is obtained by training the network with a large amount of ground truth (GT) [11]. Existing methods build complex networks by stacking a large number of convolutional layers. Although the recognition accuracy is improved, the amount of calculation is too large [12, 13]. The SE module can increase the influence of useful features in the network. The proportion of useless features is weakened to improve network performance. Through the combination with the ENet network, the accuracy of the remote sensing image semantic segmentation network can be further improved. Therefore, in order to improve the accuracy and speed at the same time, this paper has made improvements based on the ENet network. The specific research content is as follows:
In the encoder, dilated convolution, decomposition convolution, and ordinary convolution are used in an interleaved manner, ensuring the size of the feature map while increasing the receptive field. This makes each convolution output more image information.

(2) In the decoder, the weight of the characteristic channel is adjusted through the SE module. The accuracy of the remote sensing image semantic segmentation network is improved.

(3) The original image size is restored by linear interpolation. Softmax activation function and Argmax function are used to get the final semantic segmentation result of remote sensing image.

2. Related Works

With the development of deep learning, researchers have found that deep neural networks can extract deep features useful for image segmentation [14–16]. At present, the common technical methods of image semantic segmentation can be summarized into two categories. The first category is the traditional image semantic segmentation method. Image segmentation is mainly performed by extracting artificial features as visual information. Some examples are image semantic segmentation method based on threshold, image semantic segmentation method based on edge detection, image semantic segmentation method based on region, image semantic segmentation method based on graph theory, and so forth [17–19]. The features extracted by traditional methods are rough, and the segmentation results lack semantic information. The complexity of the algorithm is low, and the performance of the algorithm is poor in complex scenarios. The second type of method is the image semantic segmentation algorithm based on deep convolutional neural network [20]. In the absence of human operations, the image features are extracted by training a convolutional neural network on the target dataset. Convolutional neural networks (CNN) consist of multiple convolutional layers to form a deep network model. The network obtains the final feature representation through layer-by-layer abstraction of the original input data, so as to realize the deep feature extraction of the original data. At present, most of the segmentation methods based on deep features are extended and developed on the basis of CNN. CNN trains the neural network classifier through layer-by-layer feature learning of optical remote sensing images. Classification is performed at the image pixel level to achieve image segmentation. It is generally composed of convolutional layer, pooling layer, and fully connected layer. Through the combination of these multilayer structures, image features are automatically learned and mapped to a new feature space for representation. As a typical data-driven network, CNN needs a lot of training data to get a good segmentation model. Then the model is used to segment the test image.

Reference [21] proposed an attentional SegNet network semantic segmentation algorithm combining SegNet and attention mechanism. The network can segment vegetation, buildings, water bodies, and roads in remote sensing images. Reference [22] proposed a semisupervised semantic segmentation method for high-resolution remote sensing images based on generative adversarial networks. It only needs a small number of labeled samples to obtain a better segmentation effect. Reference [23] proposed a semantic segmentation method that combines dense connections and the idea of a fully convolutional network (FCN). In order to automatically provide fine-grained semantic segmentation maps, multiscale filters are used to improve DFCN, which increases the diversity of extracted information. Reference [24] adopts the DeepLabV3 architecture with high computational efficiency, adding spatial pyramid pooling and fully connected fusion path layer. Without reducing the resolution of the feature map, the receptive field of feature points is expanded. The improved algorithm enhances the segmentation effect of high-resolution remote sensing images. Reference [25] used the full convolutional network (FCN) semantic segmentation model to achieve pixel-level image semantic segmentation. It replaces the fully connected layer used for classification mapping in the CNN structure with a convolutional layer. Through the deconvolution operation, the obtained feature heat map is upsampled to the original input image size. At the same time, the intermediate pooling layer information is combined to generate the image prediction segmentation map. Reference [26] combined the spectral recognition index of blue-board houses and bare soil. The index of the remote sensing image is blurred to improve the accuracy and recall rate of segmentation. Reference [27] used Conditional Random Field (CRF), which can obtain global context information, as an optimization method, and it is also a common improvement method. First FCN is used to achieve coarse segmentation, and then CRF is used to refine the segmentation results based on the target’s multiscale information. Reference [28] proposed a deep convolutional encoder-decoder structure SegNet to identify and segment objects in cities such as streets and vehicles. The idea of block-by-block upsampling is used to restore the image to the original input size. The memory space is saved by retaining only the pooling index value of the encoder structure, and the detailed information of the image is retained, thereby effectively improving the segmentation accuracy. In [29], using the spectral and spatial feature information of remote sensing images, SegNet is used to extract the building coverage areas of rural areas in remote sensing images. The above methods can accurately extract network features of remote sensing images. However, due to the influence of network depth, the accuracy of feature extraction is not high. In addition, the remote sensing image network features obtained by deep learning are not smooth, and further improvement is needed. This paper proposes a remote sensing image semantic segmentation algorithm based on SE module and ENet model. In the encoding stage, different convolution methods are introduced to increase the receptive field of the model. Each convolution output contains a larger range of image information.
information. In the decoding stage, the compression, ex-
citation, and reweighting operations of the SE module are
performed. The weight of each feature channel is recal-
ibrated, and the network accuracy is improved.

Segmentation Algorithm Based on Improved
ENet Network

ENet network is a lightweight image semantic segmentation
network, which can achieve target pixel-level semantic
segmentation. This method has the characteristics of few
parameters and fast calculation speed. It meets the real-time
and accuracy requirements of remote sensing image seg-
mentation [30]. Based on this, this paper proposes a remote
sensing image semantic segmentation algorithm based on
improved ENet network.

3.1. ENet Network Structure. The ENet network changes the
previous encoder-decoder symmetry structure. The con-
volution operation is reduced in the decoder, which greatly
speeds up the processing speed. The ENet network has an
initialization operation for the input image, as shown in
Figure 1. Its main purpose is to generate feature maps and to
fuse the feature maps generated by pooling and convolution
operations. The convolution operation has a total of 13 3 × 3
filters with a sliding step of 2, and a total of 13 feature maps
are obtained. Max Pooling is a noncovered 2 × 2 sliding
window, and finally feature fusion is performed.

In addition, a Bottleneck convolution structure is also
used in the ENet network. It is mainly used in the encoder-
decoder structure, and the specific structure is shown in
Figure 2. Each Bottleneck convolution module contains 3
convolutional layers. From top to bottom in Figure 2 are a
1 × 1 projection map for reducing dimensionality, a main
convolutional layer, and a 1 × 1 ascending dimension.
Normalization and PReLU activation are performed be-
tween the convolution layers. The Bottleneck module is not
set in stone and will be changed according to specific op-
erations. If it is a downsampled Bottleneck module, the 1 × 1
projection mapping is replaced by the Max Pooling 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
convolutions are used instead. Regularization is used to solve
the problem of model overfitting.

The overall architecture of the ENet network is shown in
Table 1. Between the initialization and the final full con-
volution, there are also 5 parts. In the first part, there is a
downsamled Bottleneck module, followed by four ordinary
convolutional Bottleneck modules. In the second part, the
first is the Maximum Pooling Bottleneck module, followed
by eight different Bottleneck modules in turn: ordinary
Bottleneck module, an expanded Bottleneck module with a
main convolution expansion rate of 2, an asymmetric
Bottleneck module, a Bottleneck module with a main
convolution expansion rate of 4, and four Bottleneck
modules that are repeated in turn except for the maximum
pooling operation. The expansion rate is doubled sequen-
tially; that is, the expansion rate is 2, 4, 8, and 16 in sequence.
In the third part, all operations in the second part except the
maximum pooling Bottleneck operation are repeated. In the
fourth part, there is an upsamled Bottleneck, followed by 2
regular Bottlenecks. The previous operation has extracted
enough feature information. Here we need to restore the
image resolution to its original size and then output it.
Therefore, the fifth part is an upsamled Bottleneck followed
by a regular Bottleneck module. The fourth and fifth parts
did not use the expanded convolution module. This is be-
cause the coding modules of the first three parts have already
segmented the image completely. Therefore, there is no need
to expand the field of view to extract feature information.
The main function of the decoding structure is to restore the
resolution of the image. At the same time, the efficiency of
network model operation is improved.

3.2. SE Module. The core of the SE module is Squeeze and
Excitation. After the convolution operation has obtained
features with multiple channels, the SE module can be used
to recalibrate the weight of each feature channel. The SE module is divided into 3 steps, namely, Squeeze, Excitation, and Reweight. The schematic diagram is shown in Figure 3.

The Squeeze operation uses Global Average Pooling to compress each characteristic channel into a real number. This expands the receptive field to the global scope. The real number is obtained by the following formula:

\[ z_k = \frac{1}{W \times S} \sum_{i=1}^{W} \sum_{j=1}^{S} u_k(i, j), \quad k = 1, 2, \ldots, N, \quad (1) \]

where \( u \) is the feature map obtained after convolution, \( N \) is the number of channels of \( u \), and \( W \times S \) is the spatial dimension of \( u \).

Next, the Excitation operation captures the compressed real number sequence information. Two fully connected (FC) layers are used to increase the nonlinearity of the module. First, the dimension is reduced through the first full connection layer and then activated through ReLU and then through the second fully connected layer to upgrade, and, finally, through the sigmoid activation function. The whole process is as follows:

\[ s = \sigma[W_2 \delta(W_1 z)], \quad (2) \]

where \( \sigma \) is the sigmoid function, \( \delta \) is the nonlinear activation function ReLU, and \( W_1 \) and \( W_2 \) are the parameters of the two FC layers, respectively.

In reweighting, the channel importance coefficient is obtained by multiplying the original feature channel by channel by the excitation operation. The recalibrated features are obtained.

\[ \bar{x}_k = s_k \cdot u_k \quad k = 1, 2, \ldots, N. \quad (3) \]

### 3.3 ENet Network Architecture Integrated with SE Module

The remote sensing image segmentation and extraction process includes three stages: data preprocessing stage, training stage, and testing stage. In the preprocessing stage of remote sensing data, LabelMe is used to manually label and separate remote sensing image features. Training data and test data are generated by cutting the research area. The training data includes the training set and the validation set. K-fold cross-validation is used to realize the automatic division of the training set and the validation set. In the training phase, the preprocessed training samples are input to the improved ENet network with SE structure. The network model architecture is shown in Figure 4. Small batch gradient descent algorithm is used for iterative optimization. When the loss no longer decreases, the iteration task ends. In the test stage, the trained optimal model is applied to the test image for remote sensing image segmentation and extraction.

Figure 4 shows the improved ENet model structure with SE structure proposed in this paper. The model uses an encoder-decoder structure; and the encoder uses conventional convolution and a residual structure with dilated convolution to extract high-level semantic features. After that each layer of convolution is followed by BN and ReLU activation functions. In the decoding stage, image information of different scales is obtained through upsampling operation. Then the weight of each feature channel is recalibrated through the SE module to improve the accuracy of the network. Then the feature map is restored to the original image size through linear interpolation. Finally, the Softmax activation function and the Argmax function are used to obtain the final prediction result, so as to realize the end-to-end remote sensing image segmentation task.

### Table 1: Framework of ENet network.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Output size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial</td>
<td></td>
<td>16 × 256 × 256</td>
</tr>
<tr>
<td>Bottleneck</td>
<td>Downsampling</td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>4 × Bottleneck 1.x</td>
<td></td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>Bottleneck 2.0</td>
<td>Downsampling</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.1</td>
<td></td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.2</td>
<td>Dilated 2</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.3</td>
<td>Asymmetric 5</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.4</td>
<td>Dilated 2</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.5</td>
<td></td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.6</td>
<td>Dilated 2</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.7</td>
<td>Asymmetric 5</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 2.8</td>
<td>Dilated 16</td>
<td>128 × 64 × 64</td>
</tr>
<tr>
<td>Bottleneck 4.0</td>
<td>Upsampling</td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>Bottleneck 4.1</td>
<td></td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>Bottleneck 4.2</td>
<td></td>
<td>64 × 128 × 128</td>
</tr>
<tr>
<td>Bottleneck 5.0</td>
<td>Upsampling</td>
<td>16 × 256 × 256</td>
</tr>
<tr>
<td>Bottleneck 5.1</td>
<td>Upsampling</td>
<td>16 × 256 × 256</td>
</tr>
<tr>
<td>Fullconv</td>
<td></td>
<td>C × 512 × 512</td>
</tr>
</tbody>
</table>
4. Experiment

4.1. Experimental Environment and Network Parameters. The computer used in this paper is configured with 16 G memory, i7-8700K CPU, and GTX 1080ti graphics card. The experiment was carried out on TensorFlow. TensorFlow can run on one or more CPU/GPU.

When training the network, in order to enhance the generalization ability of the network, the input image undergoes local response normalization before the first layer of convolution; $\alpha = 0.0001$ and $\beta = 0.75$. The learning rate is set to 0.001 and iterated until the loss function converges. In the training process, the training dataset is randomly shuffled, and then every 5 images are treated as a batch. The objective function of the network uses the cross-entropy loss function; and L2 regularization is added to the last layer of the network to prevent overfitting.

4.2. Evaluation Index. F1 score is a measure of classification problems. It is the harmonic mean of precision and recall. The formula is as follows:
where \( tp \) is true positive, \( tn \) is true negative, \( fp \) is false positive, and \( fn \) is false negative.

MF1 score (mean \( F_1 \) score) is the average of \( F_1 \) scores, used to evaluate the global standard:

\[
MF1 = \frac{\sum_n F_1_n}{n},
\]

where \( n \) is the number of feature types.

Overall accuracy (OA) is the ratio between the correct prediction of the model on all test sets and the total number. It is used to evaluate the global standard:

\[
OA = \frac{tp + tn}{tp + fp + tn + fn},
\]

4.3. Vaihingen Dataset. The Vaihingen dataset was taken in Vaihingen, Germany. The dataset contains 33 remote sensing images of different sizes. Each image is extracted from a larger top-level orthophoto image. It contains 3 channels of IRRG (Infrared, Red, and Green) images, DSM (Digital Surface Model) images, and NDSM (Normalized Digital Surface Model) images. The average size of the images is \( 2494 \times 2064 \). The experiment only used IRRG images, not DSM images or NDSM images.

In order to study the effect of the time coverage of the dicing and the crop size on the accuracy of the segmentation, in this paper, experiments with different dicing strategies are carried out on the Vaihingen dataset. Firstly, the crop size is fixed to \( 512 \times 512 \). Experiments were performed according to the interblock coverage rates of 0\%, 25\%, 50\%, and 75\% (step sizes: 512, 384, 256, and 128, respectively). The experimental results are shown in Figure 5. It can be seen that OA and MF1 score the highest when the coverage rate is 75%.

Table 2 shows the comparison between the proposed algorithm and Semi-GAN [22], DFCNN [23], and DeepLabV3+ [24] on the Vaihingen dataset. Judging from the final score, the algorithm proposed in this article is 2.0\% higher than the DNCC algorithm on MF1 and 2.3\% higher on OA. From the perspective of different classifications, according to the \( F_1 \) scoring standard, the algorithm in this paper has achieved the first place in the classification of roads, buildings, vegetation, trees, and vehicles. Figure 6 shows the qualitative comparison between this algorithm and the Semi-GAN, DFCNN, and DeepLabV3+ on the Vaihingen dataset. From the challenging high-density car scene in the second row, we can see that the algorithm in this paper divides each car very finely. Other networks even divide the gray cars in the shadows into buildings. From the scenes with more trees and vegetation in the first and third rows, it can be seen that the classification error of the algorithm in this paper is less. Other algorithms often confuse similar features when classifying.

4.4. Potsdam Dataset. The Potsdam dataset was taken in Potsdam, Germany, and it contains IRRG images, IRGB images, DSM images, and NDSM images. The image size is \( 6000 \times 6000 \), and the total number of images is 38, of which 14 are used as the test set. In this experiment, 17 images are used as the training set, 5 images are used as the verification set, and 14 images are used as the test set. In the experiment, only IRRG images were used, and two incorrectly labeled images in the training set were deleted.

Table 3 shows the comparison between the proposed algorithm and the Semi-GAN, DFCNN, and DeepLabV3+ on the Potsdam dataset. From the final score, the algorithm proposed in this paper is 1.5\% higher than the second DFCNN algorithm on MF1 and 1.1\% on OA. From the perspective of different categories, according to the \( F_1 \) scoring standard, the algorithm proposed in this article has
achieved the first place in all categories. Figure 7 shows the qualitative comparison between the proposed algorithm and other general networks on the Potsdam dataset. From the scene where there are many cars in the second line and a single tree, it can be seen that the division of trees and cars is fine. However, other networks have serious misclassification. It can be seen from the third row that the forest area segmentation is more accurate, and there is less confusion between vegetation and trees. Other networks often misclassify similar features.

### 4.5. Massachusetts Buildings Dataset.

In order to verify the generality of the model, experiments are carried out on the Massachusetts building dataset. The Massachusetts buildings dataset is a large dataset used for building segmentation. The dataset consists of 151 groups of aerial images and corresponding single-channel label images. The size of the images in the data set is 1500 pixels × 1500 pixels. Among them, there are 137 training images, 10 test images, and 4 verification images. The annotated image only contains buildings. Random cropping and data enhancement are used to

### Table 2: Comparison of F1, OA, and MF1 in different algorithms of Vaihingen dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Road</th>
<th>Architecture</th>
<th>Vegetation</th>
<th>Tree</th>
<th>Vehicle</th>
<th>OA (%)</th>
<th>MF1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-GAN</td>
<td>87.8</td>
<td>90.1</td>
<td>75.9</td>
<td>86.5</td>
<td>74.2</td>
<td>84.5</td>
<td>82.9</td>
</tr>
<tr>
<td>DFCNN</td>
<td>89.1</td>
<td>90.5</td>
<td>75.2</td>
<td>85.6</td>
<td>75.8</td>
<td>86.1</td>
<td>83.2</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>85.2</td>
<td>89.4</td>
<td>75.8</td>
<td>82.6</td>
<td>69.2</td>
<td>82.9</td>
<td>80.4</td>
</tr>
<tr>
<td>SE-ENet</td>
<td>89.8</td>
<td>91.5</td>
<td>79.3</td>
<td>86.7</td>
<td>78.9</td>
<td>88.4</td>
<td>85.2</td>
</tr>
</tbody>
</table>

### Table 3: Comparison of F1, OA, and MF1 in different algorithms of Potsdam dataset.

<table>
<thead>
<tr>
<th>Method</th>
<th>Road</th>
<th>Architecture</th>
<th>Vegetation</th>
<th>Tree</th>
<th>Vehicle</th>
<th>OA (%)</th>
<th>MF1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-GAN</td>
<td>86.7</td>
<td>90.3</td>
<td>76.1</td>
<td>82.5</td>
<td>75.8</td>
<td>83.5</td>
<td>82.3</td>
</tr>
<tr>
<td>DFCNN</td>
<td>87.5</td>
<td>90.8</td>
<td>76.5</td>
<td>84.6</td>
<td>76.1</td>
<td>86.2</td>
<td>83.1</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>84.6</td>
<td>87.4</td>
<td>75.7</td>
<td>83.9</td>
<td>65.5</td>
<td>91.3</td>
<td>79.4</td>
</tr>
<tr>
<td>SE-ENet</td>
<td>88.8</td>
<td>91.4</td>
<td>78.4</td>
<td>84.9</td>
<td>79.7</td>
<td>87.3</td>
<td>84.6</td>
</tr>
</tbody>
</table>

Figure 6: Comparison of different algorithms on Vaihingen dataset.
Remote sensing image | Label | Semi-GAN | DFCNN | DeepLabV3+ | SE-ENet |
|----------------------|-------|----------|--------|------------|---------|

- Street
- Architecture
- Vegetation
- Tree
- Vehicle
- Background

**Figure 7:** Comparison of different algorithms on Potsdam dataset.

Remote sensing image | Label | Semi-GAN | DFCNN | DeepLabV3+ | SE-ENet |
|----------------------|-------|----------|--------|------------|---------|

- Street
- Architecture
- Vegetation
- Tree
- Vehicle
- Background

**Figure 8:** Comparison of different algorithms on Massachusetts building dataset.
generate experimental images with a size of 256 pixels × 256 pixels. The red pixels represent the building area, and the black pixels are the background. The details of the building segmentation experiment are similar to the remote sensing satellite image segmentation experiment. During the training process of segmentation experiment, the number of training images in batch size is 4. The number of training epochs is 10 rounds, and 10,000 batches are trained in each round. There are 3000 test images with a size of 256 pixels × 256 pixels.

The experiment chooses Semi-GAN, DFCNN, and DeepLabV3+ for comparison. Each network is trained 10 times, and the best model is selected for testing. The segmentation result of the network is shown in Figure 8. It can be found from Figure 8 that there is little difference between the networks for the segmentation of small buildings (villas). For the recognition of large buildings (factories, shopping malls), the segmentation results of this algorithm are more complete and refined. In this paper, a segmentation experiment is carried out on 3000 experimental datasets of 256 pixels × 256 pixels. The experimental results are shown in Table 4. It can be seen from Table 4 that SE-ENet achieved an F1 value of 89.7%, which is the highest among the comparison algorithms.

5. Conclusion

Through experiments on the TensorFlow platform, this paper proves that the ENet network integrated with the SE module has higher accuracy. It can be better applied to remote sensing image segmentation. Compared with other semantic segmentation algorithms, the image features generated by the proposed method are more complete. Through the introduction of dilated convolution and de-composition convolution and the cross-use of ordinary convolution, the receptive field of the model is increased. Each convolution output contains a larger range of image information. The SE module recalibrates the weight of each characteristic channel to improve the accuracy of the network. Experimental results prove that the proposed algorithm has higher segmentation accuracy. In the following research, we will focus on the segmentation of small and medium targets in remote sensing images to further improve the accuracy and segmentation speed of the model.

Data Availability

The data included in this paper are available without any restriction.

Conflicts of Interest

The author declares that there are no conflicts of interest regarding the publication of this study.

Acknowledgments

This work was supported by the special project of “Internet + Education” in the 13th Five-Year Plan of Shanxi Province in 2020 (no. HLW-20111) and “1331 Project” Maker Team Project of Jinzhong University (no. jzyxckt2019039).

References


Table 4: Massachusetts building image segmentation results.

<table>
<thead>
<tr>
<th>Method</th>
<th>F1 (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semi-GAN</td>
<td>85.4</td>
</tr>
<tr>
<td>DFCNN</td>
<td>88.3</td>
</tr>
<tr>
<td>DeepLabV3+</td>
<td>88.5</td>
</tr>
<tr>
<td>SE-ENet</td>
<td>89.7</td>
</tr>
</tbody>
</table>


