J Inf Process Syst, Vol.20, No.5, pp.640~653, October 2024
https://doi.org/10.3745/JIPS.01.0108
ISSN 2092-805X (Electronic)

JOURNAL OF INFORMATION PROCESSING SYSTEMS JIPS

A Semantic Segmentation Method of Remote Sensing Image Based on Feature Fusion and Attention Mechanism

Yiqin Wang[1,*] and Yunyun Dong[2]

Abstract

Current methods for semantic segmentation of remote-sensing images, especially for irregular and small targets, often result in low precision and incomplete feature extraction. To address this issue, an improved semantic segmentation method was developed utilizing DeepLabv3+. First, DeepLabv3+ is combined with the proposed feature fusion module to make full use of the complementary information of low- and high-level features. Second, the channel attention module helps extract effective features while suppressing irrelevant features, thereby enabling the extraction of more meaningful global information from high-level features. Finally, rich spatial information is selected using guided spatial attention, which improves the accuracy of edge segmentation of target objects. The results of the comparison show that the mean F1 score (MF1) and overall accuracy (OA) of the proposed method on the ISPRS Potsdam dataset are 89.81% and 88.45%, respectively. The MF1 of the proposed method is 89.90% and the OA is 89.14% for the UAVid dataset, which are higher than those of the other comparison algorithms. The proposed method exhibits superior semantic segmentation capabilities for remote-sensing images.

Keywords

Channel Attention, DeepLabv3+, Feature Fusion Module, Remote-Sensing Images, Semantic Segmentation

1. Introduction

As unmanned aerial vehicles (UAVs) and remote-sensing technology advance, acquiring high-resolution remote-sensing images has become increasingly important [1–3]. Most remote-sensing image data are processed using manual visual interpretation. However, this manual method cannot efficiently and quickly extract surface feature information from remote-sensing images. Furthermore, it is unsuitable for classifying large volumes of remote-sensing images [4,5]. Currently, remote-sensing technology is employed in land and resource surveys, environmental monitoring, disaster monitoring, urban planning, precision agriculture, military applications, and surveying and mapping. Therefore, effective and accurate automatic extraction of surface feature information (such as buildings, vegetation, rivers, and roads) from remote-sensing images is crucial [6–8]. The use of semantic segmentation to extract terrain information from remote-sensing images could make remote-sensing technology suitable for a wider range of

Manuscript received March 13, 2023; first revision August 22, 2023; accepted September 18, 2023.
* **Corresponding Author:** Yiqin Wang (wangyiqin80@126.com)
[1] School of Information Technology and Engineering, Jinzhong University, Jinzhong, China (wangyiqin80@126.com)
[2] College of Software, Taiyuan University of Technology, Taiyuan, China (dongyunyun312902@126.com)

applications. Although image semantic segmentation has been widely applied and has achieved satisfactory research results, there are several drawbacks inherent to the use of the technology.

The similarity between different types of scenes in remote-sensing images is high, and the internal differences between the same types of scene are large. Traditional image segmentation methods are generally based on edges, thresholds, and regions for image segmentation. Consequently, traditional remote-sensing image classification algorithms fail to fully capture the intricacies of complex remote-sensing image semantic segmentation. Therefore, specific rules and parameters are required to address this drawback [9–12]. Additionally, traditional image segmentation methods are difficult to implement and demonstrate poor accuracy. Computer vision enables computers or machines to perceive the outside world through their "eyes" like people. This technology relies primarily on computers or robots to capture and store external images. Robots use computers to process captured images structurally and convert abstract images into data that can then be analyzed. To obtain relevant information from the outside world, this technology simulates human eyes to capture external information. This core function of computer vision is called a "visual recognition task." It converts images captured by the robot into formatted data and performs the corresponding processing. In recent years, owing to advancements in computer hardware and software technology, deep learning has been applied to the enhancement of remote-sensing image segmentation technology [13–15]. For example, deep learning can effectively be applied to image feature extraction to identify and characterize the land cover category in an image. Deep-learning-based methods rely primarily on datasets to train network models. They can automatically obtain image segmentation results [16–18]. Convolutional neural networks leverage multilevel network features to transform the input colors into features across different layers. They acquire feature representations with distinct meanings from various levels of image features, such as lower-level spatial detail texture features and higher-level semantic concept features. Convolutional neural networks can be applied to the extraction of surface cover information (for example, vegetation, roads, and water bodies) from remote-sensing images [19,20]. Convolutional neural networks can extract both shallow and deep semantic feature information from remote-sensing images, thereby offering a novel approach to the semantic segmentation of such images.

The remainder of this article is structured as follows. Section 2 provides an overview of relevant research in the field of remote-sensing image segmentation. In Section 3, the proposed method for a remote-sensing image segmentation network is described in detail, including the backbone network, feature fusion module, multilocal channel attention mechanism (MCAM) module, and loss function. In Section 4, the experimental specifications and results are presented and analyzed, including the experimental environment, parameter settings, evaluation indicators, training process, and dataset selection. The conclusions of this study are discussed in Section 5, summarizing the contributions of our study and potential directions for future research.

2. Related Work

High-resolution remote-sensing images often have limited band availability and lack spectral information, making classification based solely on spectral features suboptimal. Furthermore, high-resolution remote-sensing images have rich textures, geometries, and semantics that encode detailed information. The drawback of pixel-based methods is that they fail to fully utilize the rich spatial

information present in high-resolution images. In a previous study [21], the advantages of both UNet and deep residual networks were combined to achieve the automatic recognition of roads in remote-sensing images. Although the accuracy of the resultant method is high, the efficiency must be improved. A ResUNet structure was used to extract building information in urban areas, achieving satisfactory extraction results [22]. The residual of the VGGNet of the UNet model was designed and the depth residual UNet was proposed to process remote-sensing images [23]. A ResUNet-a model was designed [24] by combining the idea of a multilevel pooled information aggregation module of the PSPNet and UNet models. A multilevel pooling block was designed at the feature extraction end of UNet, and the results were superior to those of UNet. In one study [25], the DPN model was introduced, which fused and classified multisource remote-sensing data by leveraging the characteristics of DenseNet and a multilevel pooling module. This model improved classification accuracy by fusing LiDAR and remote-sensing images and other data. In another study [26], a mixed attention mechanism model, HMANet, was developed based on the structure of an empty fully convolutional network (FCN). This model had three modules (category attention, channel attention, and spatial attention) and an optimized spatial attention mechanism (nonlocal) module. The DeepLab network, which enhanced the precision and speed of deep-learning semantic segmentation networks, was introduced in [27]. A DeepLabv3+ network model [28] was proposed that utilized porous convolution for remote-sensing image classification. In [29], a segmentation algorithm that combined an attention gate and SegNet was proposed. The segmentation accuracy of this method was high; however, the segmentation of small targets still required improvement and relied heavily on parameter tuning. In another study [30], a generative adversarial network was incorporated into an FCN to capture global image features and achieve the accurate segmentation of complex images.

However, when the above methods are applied to remote-sensing image classification, several problems could arise, such as the inability to recognize small targets, edge segmentation errors, and low classification accuracy. To address these challenges, the main contributions of the proposed algorithm are outlined as follows.

- The DeepLabv3+ network is combined with a feature fusion module. Advanced features are added to the channel attention module to extract effective features. Spatial attention is guided by high-level features, enabling the filtering of rich spatial information from low-level features, thereby enhancing segmentation accuracy.
- A channel attention mechanism is utilized to obtain the weighted advanced features. This is beneficial for extracting the global context and more effective semantic information. Weighted high-level features are employed to guide the extraction of fine low-level features and ensure the retention of more information, such as image edges and textures.

3. Method

3.1 DeepLabv3+

The DeepLabv3+ network is a high-precision architecture for semantic segmentation. The network utilizes a multibranch parallel structure and multiscale feature fusion to optimize object spatial information, thereby enhancing segmentation accuracy. The DeepLabv3+ network model is an improve-

ment on DeepLabv3. This enhancement entailed the integration of the atrous spatial pyramid pooling (ASPP) module and encoder–decoder structure. The boundary information of the image is captured by reconstructing spatial information, and multiscale contextual information is fully exploited. The best segmentation results can be obtained by combining these two features.

The encoder comprises a backbone feature extraction network module and an ASPP module. A backbone network is used for the preliminary extraction of depth features. ASPP is used for better feature extraction. The backbone network uses the ResNet-101 model. Through the residual structure, the problem of gradient disappearance inherent to the deepening of the network can be addressed and accuracy issues can be avoided. There are two types of residual blocks in this model: convolutional residual blocks and identity residual blocks, both of which are bottleneck structures. The difference lies in the input and output dimensions of the convolutional residual blocks. The input and output dimensions of the identical residual blocks are identical. The former can change the network dimensions, whereas the latter can increase its depth. A 101-layer residual network is formed using these two types of residual blocks as the basic units. The ASPP module uses a multibranch parallel structure to extract several kinds of feature information. Serial network architectures only yield a single receptive field, whereas parallel network architectures can capture receptive fields of different sizes, thereby enhancing the recognition of objects across different scales.

The decoder upsamples the high-dimensional feature layer to reconstruct the spatial structure of the input image. The conventional upsampling operation fails to fully restore the spatial information lost during downsampling. DeepLabv3+ enhances semantic segmentation accuracy by integrating shallow and deep features. The size of the shallow feature map in the selected backbone network is 128 × 128 × 256. The deep feature size of the entire encoder is 32 × 32 × 256. To combine the channel and shallow features, a deep feature map must be upsampled. The number of channels in both the DeepLabv3+ shallow and deep feature maps is 256, but the semantic information in the deep feature map is richer. To reduce the weight of the shallow features, a 1 × 1 convolution is used to reduce the number of channels in the shallow feature map. After 3 × 3 convolutions and four upsampling cycles, the feature map is restored to its original image size. The final output image is obtained using a softmax classifier.

3.2 Feature Fusion Module

The function of the encoder in a deep network is to extract feature information continuously from the input image across multiple network layers. The deeper the network layer, the richer the semantic information contained in the extracted advanced features, which is more conducive to pixel classification. Low-level features contain rich texture, boundary, and other details but also contain background noise. Advanced features have plentiful semantic information that is conducive to target location and noise suppression. Therefore, many researchers have combined the low- and high-level features. However, simple splicing cannot effectively utilize the complementary information of the two types of features. Some irrelevant information and background noise information in the underlying features lead to the deterioration of the performance of the entire network. Therefore, the feature fusion module of the DeepLabv3+ decoder is introduced. By weighting lower-level features with higher-level features, useful lower-level features are selected and stacked with higher-level features to improve their effectiveness. Prior to fusing high- and low-level features, a channel attention module is added to the high-level features. Because each channel has different weights, the importance of each channel is reflected by the channel

attention map generated using high-level features. This provides rich global contextual information. The weighted high-level features guide the extraction of low-level features and enhance details, such as edges.

To improve the segmentation effect of the network and compensate for the absence of spatial detail information in high-level features, the DeepLabv3+ decoder incorporates low-level feature fusion. The implementation process is as follows. First, the feature map output by Block1 in ResNet is passed through a 3 × 3 convolution. The rectified linear unit (ReLU) activation function is added to extract low-level feature information to obtain a low-level feature map, as described in Eq. (1). Simultaneously, the resolution of the high-level features produced by the encoder is matched with that of the low-level feature map through bilinear interpolation, resulting in an upsampled high-level feature map.

$$F_L' = \delta\left(\mathrm{Conv}_{3\times3}\left(F_L\right)\right), \tag{1}$$

$$F_H' = F_{up}(F_H), \tag{2}$$

where F_L is the low-level feature map, δ is the ReLU activation function, F_H is the high-level feature map, and F_{up} is a bilinear interpolation operation. The high-level feature map is then entered into the channel attention module.

3.3 Multilocal Channel Attention Mechanism Module

The attention mechanism is derived from the study of human visual observation. When faced with a complex scene, humans can quickly identify key areas and process them effectively. With this insight, researchers have applied this mechanism to deep learning, which assigns large coefficients of attention to the parts that must be focused on according to the critical degree of the features. The important feature information can be extracted selectively using a deep-learning model to improve the precision of the algorithm. Attention mechanisms are an important research trend in deep neural networks. They enable the feature information extracted by adaptive calibration to be trained effectively and are used in computer vision tasks. A simple attention mechanism can assign a weight of 1 to the features being focused on and 0 to the others. This is not applicable to image super-resolution reconstruction tasks. Another attention mechanism belongs to the [0,1] continuous distribution problem, which assigns all feature values between [0,1] to clarify their importance. Because each channel of the feature map corresponds to a specific semantic response, the semantic responses are interrelated. Squeeze-and-excitation (SE) blocks help learn effective channel attention using channel context dependencies to improve the representation of specific semantic features to avoid dimension reduction. An efficient channel attention (ECA) block, based on one-dimensional convolution, is proposed after the average pooling of two-dimensional spatial feature graphs into one-dimensional features. Both the SE block and ECA block enhance the representation capability of the model by capturing the global channel interaction information. The channel attention module is shown in Fig. 1.

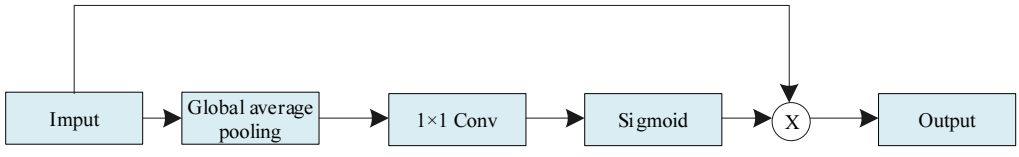


Fig. 1. Channel attention module.

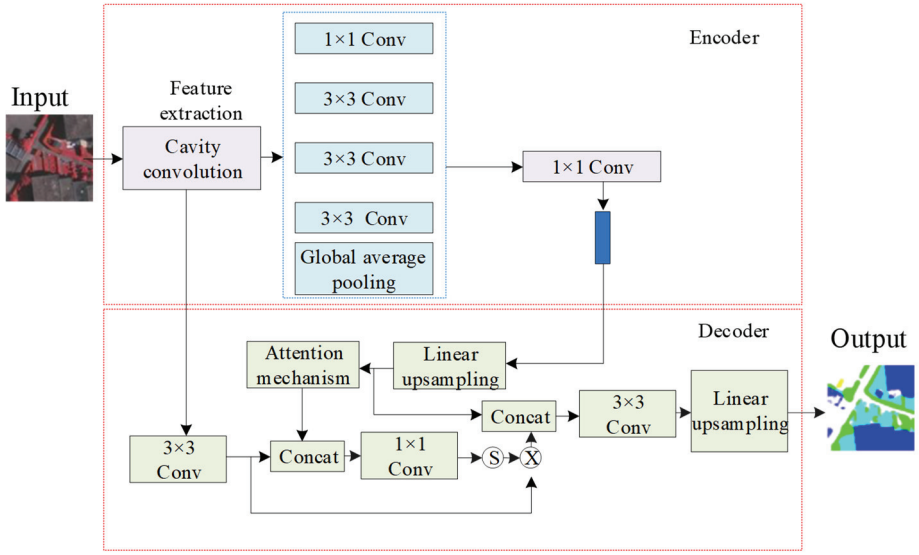


Fig. 2. Improved DeepLabv3+ network structure.

Owing to the fusion of information from different levels of characteristics, it is necessary not only to establish a certain connection with the feature information of the distant channel but also to minimize the calculation of the correlation of the near-distance channel information. Inspired by ASPP, an MCAM module was developed to learn the semantic dependencies between image feature channels and readjust the feature mapping. The MCAM module operates with four one-dimensional hole convolution layers that are arranged in parallel on the ECA block to perform sparse convolution operations on the averaged pooling features. The convolution kernel size is 9, the void contents are 1, 2, 4, and 8, and the receptive fields are 9, 17, 33, and 65, respectively. The network structure of the proposed method is shown in Fig. 2, where "S" represents the sigmoid and "X" represents element multiplication.

3.4 Loss Function

Image segmentation can be viewed as a pixel classification task, and cross-entropy is commonly employed as the loss function. For a fair comparison, the cross-entropy loss function was adopted in this study. Cross entropy, an important concept in information theory, quantifies the difference between two probability distributions. Minimizing cross entropy can yield an approximation of the target probability distribution. Cross entropy serves as a metric for the distance between two probability distributions. It is primarily utilized to measure the difference in information between two probability distributions. The formula for cross-entropy is

$$L = -\frac{1}{N}\sum_{i}\sum_{c=1}^{M} y_{ic}\log(p_{ic}), \tag{3}$$

where M represents the number of sample data categories in the dataset used and y_{ic} represents a symbolic function.

4. Experimental Results and Analysis

4.1 Experimental Environment and Parameter Setting

The experimental platform was the CentOS7 operating system, the central processing unit was the Intel Xeon Gold5215, the graphics processing unit (GPU) was the Tesla V100, and the memory was 32 GB. The basic code and data-processing-related code of the experimental model were written in Python. Python was selected as the training framework for this experiment because it supports the creation of dynamic graphs owing to the availability of Python interfaces. The software environment was set up using Anaconda to complete the Python framework.

Table 1 lists the configuration information used to train the two datasets. The Adam solver was used for parameter optimization, and the initial learning rate was 0.0003. The minimum learning rate was 1×10^{-7}, beta1 was 0.9, beta2 was 0.999, and the weight attenuation was set to 0.02. The learning strategy was poly. The number of times that a batch of data was trained was measured in iterations; when a batch of data had been trained once, it was considered to have undergone one iteration of the training process.

Table 1. Training configuration of each dataset

Item	UAVid	ISPRS Potsdam
Input size	1024 × 1024	512 × 512
Training size	768 × 768	512 × 512
Test size	1024 × 1024	512 × 512
Iterations	80,000	80,000
Batch size	8	16
Learning strategy	Poly	Poly

4.2 Evaluation Index

The performance indicators of the method were evaluated using precision, recall, and F1 values. The precision represents the number of correctly classified samples. However, when positive and negative datasets are unbalanced, additional indicators must be considered. Recall and precision influence each other; thus, the F1 value is introduced. F1 represents the harmony between recall and precision. F1, the mean F1 score (MF1), and the overall accuracy (OA) are expressed by Eqs. (4), (6), and (7), respectively.

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}, \tag{4}$$

$$\text{precision} = \frac{tp}{tp+fp}, \text{recall} = \frac{tp}{tp+fn}, \tag{5}$$

$$MF1 = \left(\sum_{1}^{n} F1_n\right)/n, \tag{6}$$

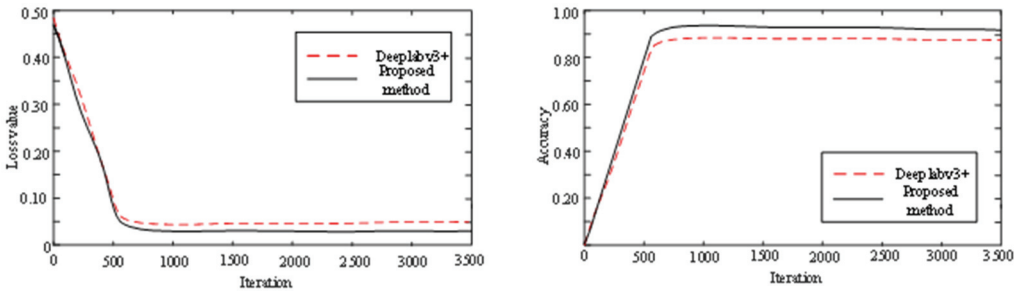
$$OA = \frac{tp + tn}{tp + fp + tn + fn}, \tag{7}$$

where tp denotes the number of correctly identified positive samples. The fp represents the number

of negative FP samples for false positives. tn represents the number of correctly identified negative samples. fn represents the number of false positive samples.

4.3 Training Process

Fig. 3 shows the changes in the loss and accuracy of the original DeepLabv3+ method and the improved DeepLabv3+ method during the training process. Fig. 3(a) shows that the loss values of the improved DeepLabv3+ and DeepLabv3+ methods gradually decreased and finally tended to converge. The loss value of the improved method was better than that of the comparison method. This proves that the introduction of the feature fusion module and attention module in DeepLabv3+ does not increase the difficulty of the model training process. Fig. 3(b) shows that the precision increase speed of the improved DeepLabv3+ is better than that of the comparison method. The improved DeepLabv3+ algorithm achieved a maximum accuracy of 92.36%.



(a) (b)

Fig. 3. Model training process: (a) loss change curves and (b) accuracy change curves.

4.4 ISPRS Potsdam Dataset

The ISPRS Potsdam dataset was obtained from the German International Photogrammetry and Remote Sensing Association. The dataset consists of 38 6000 × 6000 high-resolution images, including images for which the ground truth is available and the digital surface model map. The ground resolution of each image is 5 cm. The ground images provided by this dataset are composed of four bands. The number of ground category labels and red–green–blue (RGB) values of the dataset are consistent with those of the Vaihingen dataset. The Potsdam dataset training and test sets contain 24 and 14 images, respectively. The dataset is not divided into verification sets; therefore, the 24 images specified in the dataset for training and 14 images for testing were used. The experiment used only red, green, and blue bands to synthesize RGB images.

To evaluate the effectiveness of the improved DeepLabv3+ method, the ISPRS Potsdam dataset was used to compare previous results [29,30], original DeepLabv3+, and PSPNet with the proposed method. To be fair, all models were trained for 200 epochs on two Ge Force RTX 2080Ti GPUs, and all models used by other researchers were trained from scratch using open-source code. The comparison results for the ISPRS Potsdam dataset are presented in Fig. 4 and Table 2. As shown in the first row, the proposed method mitigated the issue of dividing the background into cars. The second row highlights the advantages of the proposed method over the other methods for accurately defining boundaries and categories. The MF1 of the proposed method was 89.81%, and the OA was 88.45%. The MF1 and OA

values of [29] were 87.87% and 87.36%, respectively. The MF1 and OA values of [30] were 88.91% and 87.96%, respectively. The MF1 and OA of the DeepLabv3+ method were 86.80% and 86.85%, respectively. The MF1 and OA values of the PSPNet method were 86.36% and 86.53%, respectively. This is because the proposed method combines the DeepLabv3+ network with the proposed feature fusion module. It leverages the complementary information of low- and high-level features to enhance segmentation accuracy.

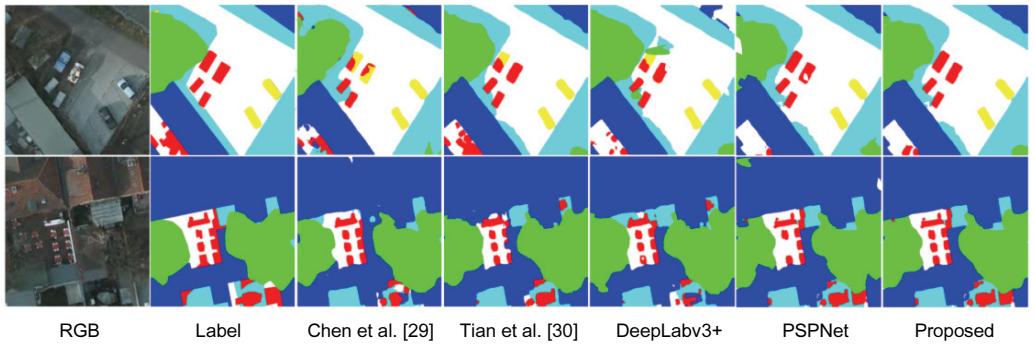


Fig. 4. Comparison of results of different methods on the Potsdam dataset.

Table 2. Comparison of different methods on the Potsdam dataset (unit: %)

Method	F1					MF1	OA
	Road	Architecture	Vegetation	Trees	Cars		
Chen et al. [29]	88.63	93.65	82.14	84.56	90.36	87.87	87.36
Tian et al. [30]	89.74	94.74	83.25	85.74	91.08	88.91	87.96
DeepLabv3+	87.59	92.78	80.76	84.01	88.87	86.80	86.85
PSPNet	87.42	92.14	80.36	83.28	88.63	86.36	86.53
Proposed method	91.28	95.08	84.12	86.36	92.22	89.81	88.45

4.5 UAVid Dataset

The UAVid dataset was obtained from high-resolution, low-altitude UAV remote-sensing image data provided by Tunt University in the Netherlands. The dataset offers 420 high-resolution images with an average size of 3840 × 2160, along with the corresponding ground truth. There are eight categories in total. The categories and corresponding RGB values are background clutter (0, 0, 0), building (128, 0, 0), road (128, 64, 128), tree (0, 128, 0), low vegetation (128, 128, 0), moving car (64, 0, 128), static car (192, 0, 192), and humans (64, 64, 0). The labels are RGB images, so the Python language was used for the format conversion of the tag images. The label category for semantic segmentation is generally encoded from 0. For example, the label value of the background in the UAVid dataset is (0, 0, 0). At this time, it is necessary to convert all pixels with the same gray value into a single-channel image with a pixel value of 0. The other categories are converted in turn. The UAVid dataset comprises 200 images for the training set, 100 images for the verification set, and 120 images for the test set.

To assess the effectiveness of the improved DeepLabv3+method, the UAVid dataset was used to compare the previous results [29,30], the original DeepLabv3+, and PSPNet with the proposed method. All networks were trained and tested under identical hardware and software conditions to ensure fair comparison and data reliability. This experiment employed the same learning rate strategy, with a total

of 200 iterations. The model that achieved the best segmentation performance for the validation set was saved. The results for the UAVid dataset are summarized in Fig. 5 and Table 3. The segmentation effect of the proposed method for people and vehicles was better than those of the other methods, and the people in the image were correctly segmented. The other methods ignored some humans. Because it is difficult to determine whether the target moves semantically for a single image, all the methods in Fig. 5 have some errors in distinguishing between static and dynamic vehicles. The results indicate that the proposed method excelled in various object segmentation scenarios. The MF1 of the proposed method was 89.90%, and the OA was 89.14%. The MF1 and OA of [29] were 87.61% and 87.74%, respectively. The MF1 and OA of [30] were 89.16% and 89.85%, respectively. The MF1 and OA of the DeepLabv3+ method were 86.59% and 86.77%, respectively. The MF1 and OA of the PSPNet method were 86.24% and 86.35%, respectively. This is because the proposed method uses a feature fusion module combined with channel attention to replace the decoder in the DeepLabv3+ network. To select useful spatial details adaptively from low-level features guided by high-level features and filter the relevant interference information, the channel attention mechanism helps in obtaining weighted high-level features, facilitating the extraction of global context and more meaningful semantic information. This leads to the retention of additional image details, ultimately enhancing the segmentation outcome.

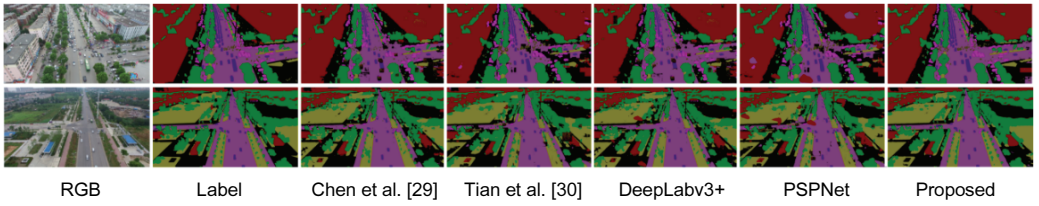


Fig. 5. Comparison of results of different methods on the UAVid dataset.

Table 3. Comparison of different methods on the UAVid dataset (unit: %)

Method	F1					MF1	OA
	Road	Architecture	Vegetation	Trees	Cars		
Chen et al. [29]	88.90	93.36	82.56	84.47	88.74	87.61	87.74
Tian et al. [30]	90.06	94.76	84.08	85.21	91.69	89.16	89.85
DeepLabv3+	88.27	92.45	81.74	83.21	87.29	86.59	86.77
PSPNet	87.95	92.03	81.38	82.74	87.11	86.24	86.35
Proposed method	91.36	95.04	85.65	85.69	91.78	89.90	89.14

4.6 Ablation Experiment

Ablation experiments were conducted using the UAVid dataset to demonstrate the effectiveness of the proposed module. The segmentation performance was evaluated quantitatively using the MF1 and OA evaluation indices. DeepLabv3+ was used as the base network, and three ablation comparison experiments were conducted. The results of the ablation experiments are listed in Table 4. The results show that, when the feature fusion module was used, the MF1 and OA metrics of the network on the UAVid dataset improved by 1.42% and 1.27%, respectively. This suggests that the feature fusion module exhibits superior classification capabilities. Subsequently, an MCAM was added to the network, and the MF1 and OA indicators increased by 3.31% and 2.37%, respectively, demonstrating the effectiveness of

the MCAM. The experimental outcomes indicate that each key module introduced in this study enhanced the accuracy of the semantic segmentation of remote-sensing images.

Table 4. Results of ablation experiment

Base	Feature fusion module	MCAM	MF1 (%)	OA (%)
√			86.59	86.77
√	√		88.01	88.04
√	√	√	89.90	89.14

5. Conclusion

To address the challenges of low segmentation accuracy and insufficient feature extraction for small and irregular objects in existing semantic segmentation methods for remote-sensing images, an improved DeepLabv3+-based segmentation method was proposed. The DeepLabv3+ network was combined with the proposed feature fusion module, and high-level features were added to the channel attention module to obtain more effective global context information. The MF1 and OA values of the proposed method on the ISPRS Potsdam dataset were 89.81% and 88.45%, respectively. For the UAVid dataset, the MF1 of the proposed method was 89.90% and the OA was 89.14%, which were higher than those of the other comparison algorithms. The experimental findings demonstrate that this method outperforms other techniques and achieves superior results that can serve as a valuable reference for remote-sensing image segmentation methods.

In this study, only the spatial attention and channel attention modules were considered to determine the influence of the attention mechanism on semantic segmentation. Other types of attention mechanisms have been proposed recently and have achieved good results, and they can be combined with various attention mechanisms for future research. In addition, in a planned subsequent study, improving the image feature extraction branch network, selectively fusing useful features, and ignoring useless features will be considered. Designing a lighter network while maintaining the same segmentation accuracy should also be a future research focus.

Conflict of Interest

The authors declare that they have no competing interests.

Funding

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), "1331 Project" Maker Team Project of Jinzhong University (No. jzxycktd2019039).

References

[1] B. Cui, X. Chen, and Y. Lu, "Semantic segmentation of remote sensing images using transfer learning and deep convolutional neural network with dense connection," *IEEE Access*, vol. 8, pp. 116744-116755, 2020. https://doi.org/10.1109/ACCESS.2020.3003914

[2] W. Sun and R. Wang, "Fully convolutional networks for semantic segmentation of very high resolution remotely sensed images combined with DSM," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 3, pp. 474-478, 2018. https://doi.org/10.1109/LGRS.2018.2795531

[3] J. Jiang, C. Lyu, S. Liu, Y. He, and X. Hao, "RWSNet: a semantic segmentation network based on SegNet combined with random walk for remote sensing," *International Journal of Remote Sensing*, vol. 41, no. 2, pp. 487-505, 2020. https://doi.org/10.1080/01431161.2019.1643937

[4] L. Ding, H. Tang, and L. Bruzzone, "LANet: Local attention embedding to improve the semantic segmentation of remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 1, pp. 426-435, 2021. https://doi.org/10.1109/TGRS.2020.2994150

[5] J. Zhang, S. Lin, L. Ding, and L. Bruzzone, "Multi-scale context aggregation for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 12, no. 4, article no. 701, 2020. https://doi.org/10.3390/rs12040701

[6] M. Kampffmeyer, A. B. Salberg, and R. Jenssen, "Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Las Vegas, NV, USA, 2016, pp. 1-9. https://doi.org/10.1109/CVPRW.2016.90

[7] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Systems with Applications*, vol. 169, article no. 114417, 2021. https://doi.org/10.1016/j.eswa.2020.114417

[8] R. Dong, X. Pan, and F. Li, "DenseU-net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65347-65356, 2019. https://doi.org/10.1109/ACCESS.2019.2917952

[9] R. Shang, J. Zhang, L. Jiao, Y. Li, N. Marturi, and R. Stolkin, "Multi-scale adaptive feature fusion network for semantic segmentation in remote sensing images," *Remote Sensing*, vol. 12, no. 5, article no. 872, 2020. https://doi.org/10.3390/rs12050872

[10] L. Mi and Z. Chen, "Superpixel-enhanced deep neural forest for remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 159, pp. 140-152, 2020. https://doi.org/10.1016/j.isprsjprs.2019.11.006

[11] Y. Hua, D. Marcos, L. Mou, X. X. Zhu, and D. Tuia, "Semantic segmentation of remote sensing images with sparse annotations," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, article no. 8006305, 2021. https://doi.org/10.1109/LGRS.2021.3051053

[12] Z. Xu, W. Zhang, T. Zhang, and J. Li, "HRCNet: High-resolution context extraction network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 13, no. 1, article no. 71, 2020. https://doi.org/10.3390/rs13010071

[13] H. Wei, X. Xu, N. Ou, X. Zhang, and Y. Dai, "DEANet: dual encoder with attention network for semantic segmentation of remote sensing imagery," *Remote Sensing*, vol. 13, no. 19, article no. 3900, 2021. https://doi.org/10.3390/rs13193900

[14] Y. Liu, Q. Ren, J. Geng, M. Ding, and J. Li, "Efficient patch-wise semantic segmentation for large-scale remote sensing images," *Sensors*, vol. 18, no. 10, article no. 3232, 2018. https://doi.org/10.3390/s18103232

[15] M. Alam, J. F. Wang, C. Guangpei, L. V. Yunrong, and Y. Chen, "Convolutional neural network for the semantic segmentation of remote sensing images," *Mobile Networks and Applications*, vol. 26, pp. 200-215, 2021. https://doi.org/10.1007/s11036-020-01703-3

[16] C. He, S. Li, D. Xiong, P. Fang, and M. Liao, "Remote sensing image semantic segmentation based on edge information guidance," *Remote Sensing*, vol. 12, no. 9, article no. 1501, 2020. https://doi.org/10.3390/rs12091501

[17] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 145, pp. 60-77, 2018. https://doi.org/10.1016/j.isprsjprs.2018.04.014

[18] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, and H. Li, "Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 175, pp. 20-33, 2021. https://doi.org/10.1016/j.isprsjprs.2021.02.009

[19] R. Li, S. Zheng, C. Zhang, C. Duan, J. Su, L. Wang, and P. M. Atkinson, "Multiattention network for semantic segmentation of fine-resolution remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, article no. 5607713, 2021. https://doi.org/10.1109/TGRS.2021.3093977

[20] S. Ouyang and Y. Li, "Combining deep semantic segmentation network and graph convolutional neural network for semantic segmentation of remote sensing imagery," *Remote Sensing*, vol. 13, no. 1, article no. 119, 2020. https://doi.org/10.3390/rs13010119

[21] Z. Zhang, Q. Liu, and Y. Wang, "Road extraction by deep residual U-Net," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 5, pp. 749-753, 2018. https://doi.org/10.1109/LGRS.2018.2802944

[22] H. Yang, P. Wu, X. Yao, Y. Wu, B. Wang, and Y. Xu, "Building extraction in very high resolution imagery by dense-attention networks," *Remote Sensing*, vol. 10, no. 11, article no. 1768, 2018. https://doi.org/10.3390/rs10111768

[23] Y. Yi, Z. Zhang, W. Zhang, C. Zhang, W. Li, and T. Zhao, "Semantic segmentation of urban buildings from VHR remote sensing imagery using a deep convolutional neural network," *Remote Sensing*, vol. 11, no. 15, article no. 1774, 2019. https://doi.org/10.3390/rs11151774

[24] F. I. Diakogiannis, F. Waldner, P. Caccetta, and C. Wu, "ResUNet-a: a deep learning framework for semantic segmentation of remotely sensed data," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 162, pp. 94-114, 2020. https://doi.org/10.1016/j.isprsjprs.2020.01.013

[25] X. Pan, L. Gao, B. Zhang, F. Yang, and W. Liao, "High-resolution aerial imagery semantic labeling with dense pyramid network," *Sensors*, vol. 18, no. 11, article no. 3774, 2018. https://doi.org/10.3390/s18113774

[26] R. Niu, X. Sun, Y. Tian, W. Diao, K. Chen, and K. Fu, "Hybrid multiple attention network for semantic segmentation in aerial images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, article no. 5603018, 2021. https://doi.org/10.1109/TGRS.2021.3065112

[27] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. 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. https://doi.org/10.1109/TPAMI.2017.2699184

[28] L. Yuan, J. Yuan, and D. Zhang, "Remote Sensing Image Classification Based on DeepLab-v3+," *Laser & Optoelectronics Progress*, vol. 56, no. 15, article no. 152801, 2019. http://dx.doi.org/10.3788/LOP56.152801

[29] B. Chen, J. Zhang, J. Zhou, Z. Chen, J. Yang, and Y. Zhang, "Semantic image segmentation network based on deep learning," in *Proceedings of SPIE 11429: MIPPR 2019: Automatic Target Recognition and Navigation.* Bellingham WA: International Society for Optics and Photonics, 2020, pp. 77-81. https://doi.org/10.1117/12.2538067

[30] L. Tian, X. Zhong, and M. Chen, "Semantic segmentation of remote sensing image based on GAN and FCN network model," *Scientific Programming*, vol. 2021, article no. 949137, 2021. https://doi.org/10.1155/2021/9491376

Yiqin Wang **https://orcid.org/0000-0003-4970-9224**

She is an associate professor and graduated from Tianjin Normal University in 2007. She received a master's degree in computer science and now works in Jinzhong University. Her research interests include deep learning and image processing.

Yunyun Dong **https://orcid.org/0009-0005-5584-9778**

She is a doctor of computer science and graduated from Taiyuan University of Technology in 2021. She is now a lecturer at Taiyuan University of Technology. Her research interests include intelligent information processing and computer simulation.