Complex & Intelligent Systems (2024) 10:2839–2849
https://doi.org/10.1007/s40747-023-01304-z

ORIGINAL ARTICLE

An improved DeepLabv3+ lightweight network for remote-sensing image semantic segmentation

Hui Chen[1] · Yuanshou Qin[1] · Xinyuan Liu[1] · Haitao Wang[1] · Jinling Zhao[2]

Received: 20 May 2023 / Accepted: 17 November 2023 / Published online: 15 December 2023

Abstract

To improve the accuracy of remote-sensing image semantic segmentation in complex scenario, an improved DeepLabv3+ lightweight neural network is proposed. Specifically, the lightweight network MobileNetv2 is used as the backbone network. In atrous spatial pyramid pooling (ASPP), to alleviate the gridding effect, the Dilated Convolution in original DeepLabv3+ network is replaced with the Hybrid Dilated Convolution (HDC) module. In addition, the traditional spatial mean pooling is replaced by the strip pooling module (SPN) to improve the local segmentation effect. In the decoder, to obtain the rich low-level target edge information, the ResNet50 residual network is added after the low-level feature fusion. To enhance the shallow semantic information, the efficient and lightweight Normalization-based Attention Module (NAM) is added to capture the feature information of small target objects. The results show that, under the INRIA Aerial Image Dataset and same parameter setting, the Mean Pixel Accuracy (MPA) and Mean Intersection over Union (MIoU) are generally best than DeepLabv3+ , U-Net, and PSP-Net, which are respectively improved by 1.22%, − 0.22%, and 2.22% and 2.17%, 1.35%, and 3.42%. Our proposed method has also a good performance on the small object segmentation and multi-object segmentation. What's more, it significantly converges faster with fewer model parameters and stronger computing power while ensuring the segmentation effect. It is proved to be robust and can provide a methodological reference for high-precision remote-sensing image semantic segmentation.

Keywords Remote-sensing image · Semantic segmentation · DeepLabv3+ · Deep learning · Lightweight network

Introduction

In recent years, the rapid development of remote-sensing technology has provided an important means for Earth observation [1]. At the same time, the spatial resolution of remote-sensing images has increased from kilometers, hundreds of meters, and meters to sub-meters, and the time resolution has been shortened from several tens of days or days to hours. Compared with traditional monitoring methods, remote-sensing technology has the advantages of wide data sources, high time and spatial resolution, and low acquisition costs, providing rich data sources for time-series, dynamic, and precise collection and analysis of Earth environment. It has been widely used in scene classification, urban planning, crop classification, climate forecasting, and many other fields [2–5]. Among them, semantic segmentation based on target classification recognition is a key technology in remote-sensing image processing. It identifies and judges the object category to which each pixel in the image belongs as pixels, reasoning from low-level semantics to high-level semantics, and obtaining the final pixel-level segmented image [6].

Traditional semantic segmentation of remote-sensing images mainly involves extracting image features such as areas, linear structures, and points to interpret the required land information. For example, threshold segmentation assigns different object categories to different gray levels based on the gradation of image gray values, and then identifies the target objects accordingly [7]. Edge detection algorithms, on the other hand, use differences in gray values at the edges of the image after filtering to obtain

✉ Haitao Wang
htwang@ahu.edu.cn

✉ Jinling Zhao
zhaojl@ahu.edu.cn

[1] School of Internet, Anhui University, Hefei 230039, China

[2] National Engineering Research Center for Analysis and Application of Agro-Ecological Big Data, Anhui University, Hefei 230601, China

edge images using edge detection operators [8]. In addition, there are also methods such as Structured Regression Forest (SRF) [9], and Support Vector Machine (SVM) [10]. Dai et al. (2020) used a 0.29 cm high-resolution unmanned aerial vehicle optical remote-sensing image to analyze the color features and color indices using the Otsu adaptive threshold method to achieve cotton target recognition and segmentation [11]. Li et al. (2010) enhanced the segmentation accuracy of remote-sensing images using the edge detection algorithm to obtain edge information [12]. Li et al. (2021) selected multi-seasonal Sentinel-1A and Sentinel-2A/B remote-sensing images, used SVM for feature extraction and classification of winter wheat in the fields, and obtained the area of wheat cultivation [13].

It is worth noting that some traditional segmentation methods require manually determining thresholds, which makes the extraction process more complex and sensitive to noise and outliers in the images, resulting in poor segmentation and generalization capabilities. For complex remote-sensing images, conventional image segmentation techniques often struggle to achieve high segmentation accuracy and good classification performance, and cannot meet the high-precision requirements of practical applications. With the rise of deep learning, convolutional neural network models can update the parameters in the network using the loss function of training data, demonstrating strong learning ability and excellent performance [14]. The classic semantic segmentation model DeepLabv3+ adopts an encoder–decoder structure, fully considering shallow and deep semantic information and using depthwise separable convolutions in the spatial pyramid pooling structure, greatly reducing the number of parameters and improving the segmentation performance [15]. To meet the segmentation requirements of different scenes, DeepLabv3+ has been continuously improved. For example, Su et al. (2021) introduced a dual attention mechanism to enhance semantic information at different levels, but there are still issues with holey large object segmentation [16]. Wang et al. (2022) investigated the capabilities of global context information according to the theory of pyramids and improved the atrous spatial pyramid pooling module of DeepLabv3+ network by connecting the atrous convolution with different dilation rates at the receptive field [17].

This article proposes a lightweight method for semantic segmentation of remote-sensing images by improving the DeepLabv3+ model. The method uses the lightweight network MobileNetv2 as the backbone network to reduce the number of model parameters. To enhance the feature extraction network, the standard dilated convolutions in the original network are replaced with the Hybrid Dilated Convolution (HDC) module, effectively solving the grid effects. In addition, the traditional spatial average pooling is replaced with a striped pooling module to enrich local details. In the decoder, a ResNet50 residual module is added after the fusion of low-level features to further obtain rich target edge feature information. Furthermore, a Normalization-based Attention Module (NAM) is added to enhance shallow semantic information and improve the network's ability to capture small target object features.

Methodology

Architecture of original DeepLabv3+

DeepLabv3+ is based on the encoder-decoder structure, where the encoder is responsible for extracting shallow and high-level semantic information, and the decoder further combines low-level and high-level features to improve the accuracy of segmentation boundaries and classify semantic information of different pixels [18]. In the encoder, the DeepLabv3+ model uses Xception as the backbone network, and extracts shallow and deep features from Xception, with the deep features input into the ASPP module. The ASPP module consists of four convolutional layers with dilation factors of 1, 6, 12, and 18, and a global average pooling operation.

It introduces multiscale information through the pyramid pooling module with dilated convolution, where dilated convolution can expand the receptive field without significantly reducing the image size. This enhances the feature information of the deep features of Xception. Then, these five features of different scales are merged in the channel dimension through Concatenation, and the feature is compressed through a 1×1 convolutional layer to obtain a high-level feature map. In the decoder, 1×1 convolutional layer are used to adjust the number of channels of the bottom features that are compressed twice, and then, they are concatenated with the high-level feature map upsampled four times. After stacking, the features are refined through a 3×3 convolution. Finally, the predicted image with the original image resolution is obtained through $4 \times$ linear interpolation. The specific network framework is shown in Fig. 1.

Architecture of improved DeepLabv3+

This article is based on the classic DeepLabv3+ network model and proposes improvements (Fig. 2).

(1) In the encoding area, DeepLabv3+ uses Xception as the backbone network for feature extraction. However, the issue of parameter volume and training speed that Xception brings still needs optimization. This article adopts a lightweight network, MobileNetv2, based on depth separable convolution as the feature extraction network and improves it to reduce the parameter volume and computational overhead while improving the efficiency of feature extraction, making

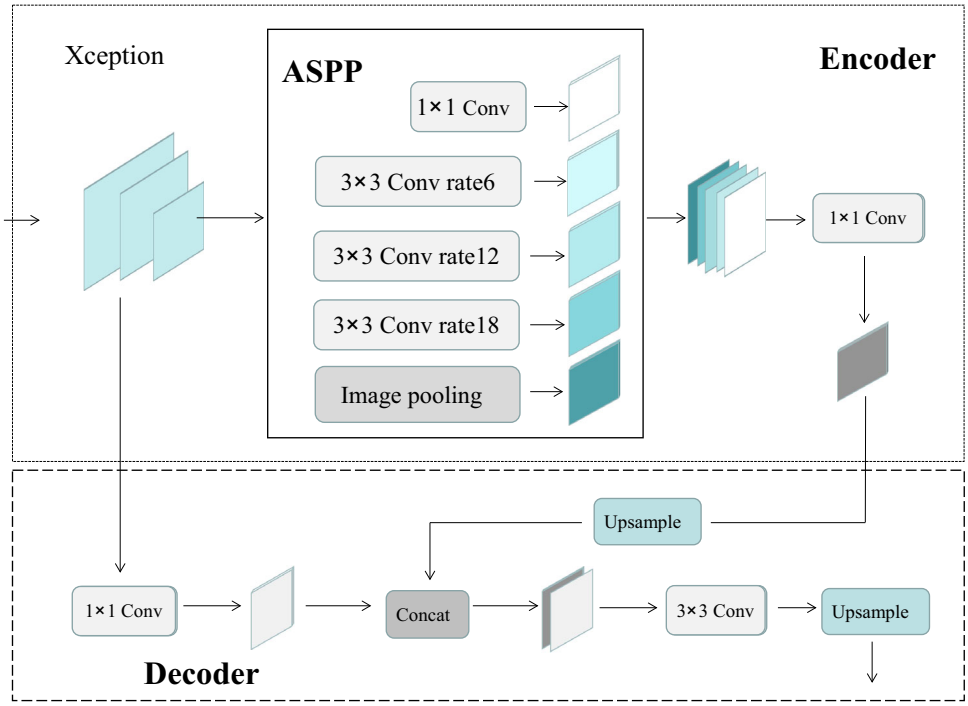

Fig. 1 Traditional network structure of DeepLabv3+

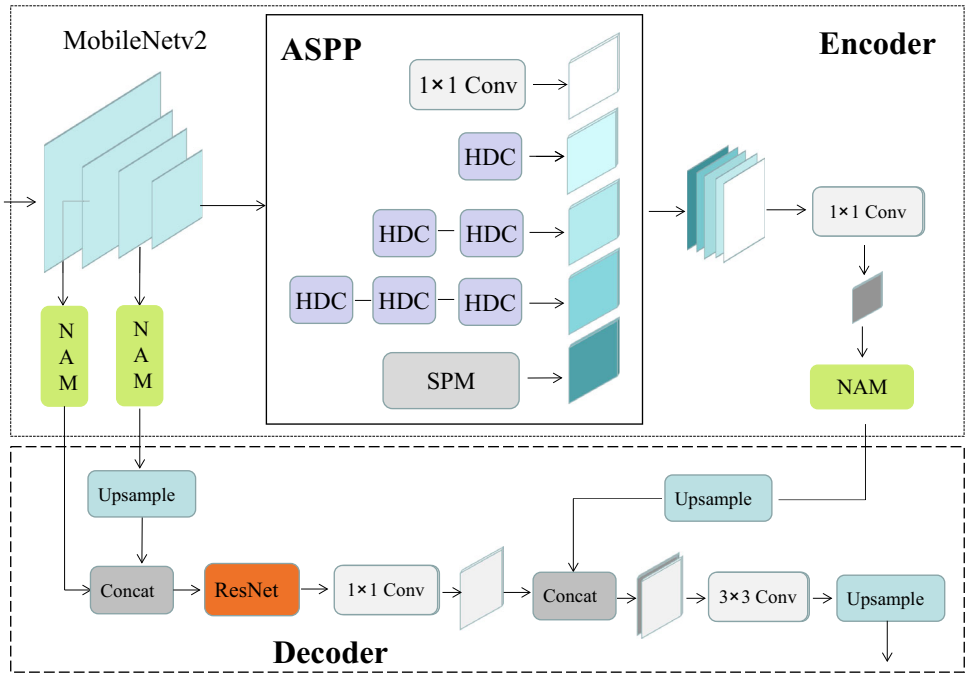

Fig. 2 Network structure of proposed DeepLabv3+

it more suitable for semantic segmentation tasks to extract shallow and deep features. It is worth noting that the decoder part in the classic DeepLabv3+ network model only takes one low-level feature layer, which is too simple. This article extracts two shallow features, the fourth layer and the seventh layer, from the MobileNetv2 network and applies NAM attention mechanism to enhance the semantic information of the lower layers.

Deep features are enhanced in the ASPP module, but the discrete sampling of dilation convolution makes it easy to ignore the dependence between continuous points with a large dilation rate, resulting in the grid effect, and easily causes the loss of local information and affects the prediction results. This article replaces the dilation convolution with the HDC module, covering the rectangular area of the underlying feature layer with a series of dilation convolutions, and

Table 1 Primary parameters of original MobileNetv2

Layer	Input	Operator	t	c	n	s
1	3	Conv2d	–	32	1	2
2	32	Bottleneck	1	16	1	1
3	16	Bottleneck	6	24	2	2
4	24	Bottleneck	6	32	3	2
5	32	Bottleneck	6	64	4	2
6	64	Bottleneck	6	96	3	1
7	96	Bottleneck	6	160	3	2
8	160	Bottleneck	6	320	1	1
9	320	Conv2d 1 × 1	–	1280	1	1
10	1280	Avgpool 7 × 7	–	–	1	–
11	1280	Conv2d 1 × 1	–	k	–	–

Table 2 Primary parameters of proposed MobileNetv2

Layer	Input	Operator	t	c	n	s	r
1	3	Conv2d	–	32	1	2	1
2	32	Bottleneck	1	16	1	1	1
3	16	Bottleneck	6	24	2	2	1
4	24	Bottleneck	6	32	3	2	1
5	32	Bottleneck	6	64	4	1	1
6	64	Bottleneck	6	96	3	1	1
7	96	Bottleneck	6	160	3	1	4
8	160	Bottleneck	6	320	1	1	1

ensuring that the edges in the rectangular area have no holes or missing parts to improve the problems caused by the grid effect. In addition, this article replaces the global average pooling module used in the original model with the strip pooling module to avoid establishing unnecessary connections at a long distance and collect information from different spatial dimensions through vertical and horizontal pooling to establish the dependence relationship between channels. A lightweight and efficient NAM attention mechanism is also applied to the stacked compressed high-level feature maps to help improve the segmentation accuracy of the image.

(2) In the decoder area, the seventh layer feature with NAM attention is upsampled to the same size as the fourth layer feature after fusion and channel adjustment, and then, the ResNet50 module is added to obtain richer low-level target feature information. Then, the deep features and shallow features are concatenated as in the original model. Finally, after a 3 × 3 convolution and 4 × upsampling, the image is restored to its original size.

Placement of backbone network

MobileNetv2 mainly consists of depthwise separable convolutions. Although its accuracy decreases slightly during training as a feature extraction network, its inverted residual structure greatly improves network performance, reduces parameter count, and enhances network efficiency [19]. The original network structure and parameters of MobileNetv2 are shown in Table 1, where the input represents the number of input channels for each layer, operators include depthwise separable convolution (bottleneck), ordinary convolution (conv2d), and average pooling (avgpool), t represents the ratio of upsampling for the 1 × 1 convolution in the inverted residual structure, c represents the number of output channels, n represents the number of times bottleneck is repeated, and s represents the stride.

The improvements were made to MobileNetv2 to further reduce the model's parameter count and simplify the model. The first 8 layers of MobileNetv2 were used, and the downsampling factor was set to 3. The s value of the fifth and seventh layers was set to 1, and the 3 × 3 ordinary convolution in the seventh layer was replaced with a dilated convolution of dilation rate 4, denoted as r. When $r = 1$, the dilated convolution degenerates into an ordinary convolution. The specific network structure is shown in Table 2.

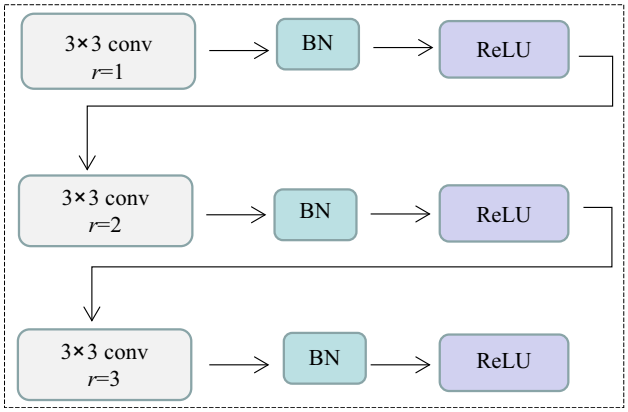

Fig. 3 Structure of the HDC module

Atrous spatial pyramid pooling

The deep-layer feature map output by MobileNetv2 is then passed through ASPP, and dilated convolutions that follow the HDC principle [20] are used to replace the original dilated convolutions. In the HDC principle, the maximum distance formula between two non-zero elements is defined as

$$\begin{aligned} M_i| &= \max[M_{i+1} - 2r_i,\ M_{i+1} - 2(M_{i+1} - r_i),\ r_i] \\ &|= \max[M_{i+1} - 2r_i,\ 2r_i - M_{i+1},\ r_i], \end{aligned} \tag{1}$$

where M_i is the maximum distance between two non-zero values of layer i, and r_i is the dilation rate of layer i. To meet the goal that avoid the final receptive field existing holes which can lead to missing local information, the design maximum distance should better meet the goal that $M_2 \leq K$.

It is worth noting that to allow each pixel in the high-level feature map to utilize all the pixels within the receptive field of the low-level feature map, the dilation rate of the first convolution is generally set to 1. Multiple verifications have shown that when the dilation rate of consecutive convolutional layers is jagged and their greatest common divisor is not greater than 1, a completely covered square feature map without any holes can be obtained to effectively alleviate the problem of information loss caused by discrete sampling of dilated convolutions. Therefore, by using three consecutive dilated convolutions with dilation rates of 1, 2, and 3 as the HDC module added to ASPP, the HDC module's specific structure is shown in Fig. 3. To keep the receptive field constant in the network model, one HDC module is used to replace the convolution with a dilation rate of 6, two HDC modules are used to replace the convolution with a dilation rate of 12, and three HDC modules are used to replace the convolution with a dilation rate of 18, effectively avoiding the grid effect and reducing the loss of local information.

Moreover, the standard spatial average pooling which collects context from a fixed square region is used in the original DeepLabv3+ . However, when processing objects with irregular shapes or handle with complex environment, it may build unnecessary connections and inevitably incorporate contaminating information from irrelevant regions.

Here, it is replaced by the strip pooling module [21], which is a band-shape pooling window used to perform pooling along either the horizontal or the vertical dimension. It collects long-range dependencies and meanwhile focus on details which can simultaneously aggregate global and local context (Fig. 4).

In strip pooling, there will be a spatial extent of pooling $H \times 1$ or $1 \times W$. Given the two-dimensional tensor $x \in R^{C \times H \times W}$, where C represents the number of channels, H represents the height and W represents the width of the feature map, and x is the input of two parallel paths which contains a horizontal and vertical strip pooling layer. In the strip pooling module, it averages all feature values in rows or columns, respectively.

In horizontal strip pooling, the pixel values in each row of the feature map are added and then averaged, and the output $y^h \in R^H$ which is a column vector of $H \times 1$ can be written as

$$y_j^h = \frac{1}{W} \sum_{0 \leq i < W} x_{i,\,j}. \tag{2}$$

In vertical strip pooling, the pixel values in each column of the feature map are added and then averaged, and the output $y^v \in R^W$ which is a row vector of $1 \times W$ can be written as

$$y_j^v = \frac{1}{H} \sum_{0 \leq i < H} x_{i,\,j}, \tag{3}$$

where i, j, respectively, represents the ith row and the jth column of the feature map.

To get an output $z \in R^{C \times H \times W}$ that contains more useful global priors, we combine $y^h \in R^{C \times H}$ and $y^v \in R^{C \times W}$, which are composed as follows:

$$y_{c,\,i,\,j} = y_{c,\,i}^h + y_{c,\,j}^v. \tag{4}$$

Then, the output $\mathbf{z}$ can be written as

$$\mathbf{z} = \text{Scale}(\mathbf{x}, \sigma(f(\mathbf{y}))), \tag{5}$$

where Scale() represents element-wise multiplication, σ is the Sigmoid function, and f is the 1×1 convolution.

Compared with global average pooling, strip pooling considers a long and narrow kernel shape focusing on long-range dependencies between regions. Given the horizontal and vertical strip pooling layers, it facilitates the search of global information and performs well in capturing local details.

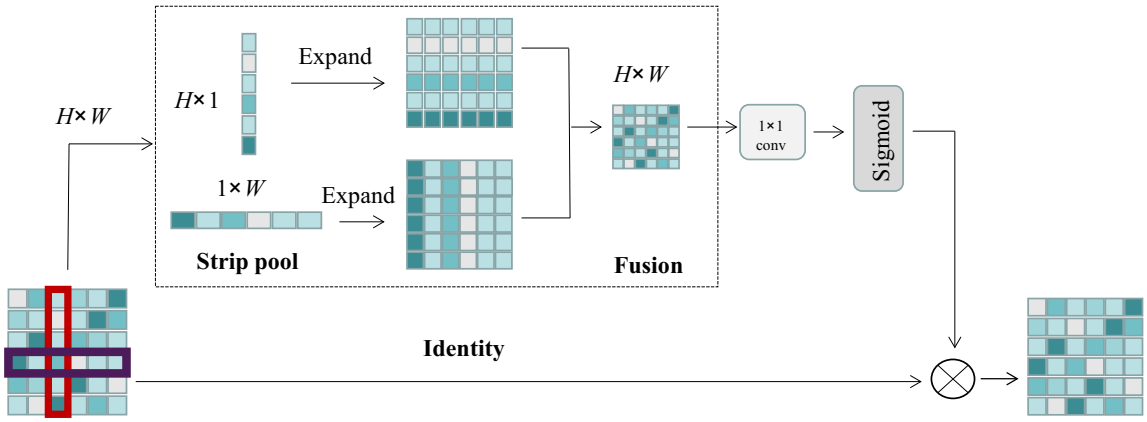

Fig. 4 Schematic illustration of the Strip Pooling (SP) module

Normalization-based attention module

Attention mechanism often captures salient features exploiting the mutual information from different dimensions. Taking account of the contributing factors of the weights, this paper applies the Normalization-based Attention Module (NAM) [22]. NAM is realized by utilizing the variance measurement of the weights which reflects the size of the change in each channel and thus the importance of the channel.

As an efficient and lightweight attention mechanism, NAM uses batch normalized scale factor which expresses the importance of the channel through sparse weight penalty and standard deviation. If the importance is greater, then increase the weight of significant features. The formula is as follows:

$$B_{\text{out}} = BN(B_{\text{in}}) = \gamma \frac{B_{\text{in}} - \mu_B}{\sqrt{\sigma_B^2 + \varepsilon}} + \beta, \tag{6}$$

where μ_B and σ_B, respectively, represents the mean and standard deviation of mini-batch B, γ and β are, respectively, the scale factor and shift, and ε usually is a small number to avoid 0 in the denominator.

The channel attention sub-module is shown in Fig. 5. The information in the channel dimension of the input feature map is normalized, and the final output feature obtained after applying weights is as follows:

$$\mathbf{M}_{\text{C}} = \text{sigmoid}\big(W_\gamma(\text{BN}(\mathbf{F}_1))\big), \tag{7}$$

where γ is the scaling factor of each channel, $\mathbf{M}_{\text{C}}$ represents the output feature, $W\gamma$ represents the weight of this channel, and $\mathbf{F}_1$ represents the input feature map.

The spatial attention sub-module is shown in Fig. 6. The same normalization method is used for each pixel in the input feature map, and the final output feature is as follows:

$$\mathbf{M}s = \text{sigmoid}(W_\lambda(BN_S(\mathbf{F}_2))), \tag{8}$$

where λ is the scaling factor of each channel, $\mathbf{M}_s$ represents the output feature, W_λ represents the weight of this channel, and $\mathbf{F}_2$ represents the input feature map.

ResNet50 module

In the decoder, to further enrich the detailed information in the low-level features, this paper adds the ResNet50 module [23] after the connection of the fourth and seventh layers of MobileNetv2. The ResNet50 module has a stack of 3 layers consists of 1 × 1, 3 × 3, and 1 × 1 convolutions. The 1 × 1 layers are responsible for reducing and then increasing dimensions, leaving the 3 × 3 layer a bottleneck with smaller input/output dimensions to further refine the semantic information. What's more, the 3 × 3 convolution in ResNet50 is replaced by the dilated convolution with the dilation rate of 4 to expand the receptive field and improve the segmentation effect (Fig. 7).

Datasets and evaluation metrics

Dataset and experimental environment

The algorithm is implemented based on the PyTorch deep learning framework, and the improved algorithm is compared with the original DeepLabv3+ , U-Net [24], and PSPNet [25]. The hardware configuration is chosen as 16-core CPU, RTX 3090, 43 GB RAM, and 500 GB hard disk. For training, the image size was compressed to 512 × 512, batch size was set to 8, the learning rate was 0.005, and 50 iterations were performed.

The INRIA Aerial Image Labeling Dataset (https://project.inria.fr/aerialimagelabeling/), released by INRIA (the French National Institute for Research in Computer Science and Automation) in 2017 [26], was used to perform the experiments. The INRIA dataset is a remote-sensing image

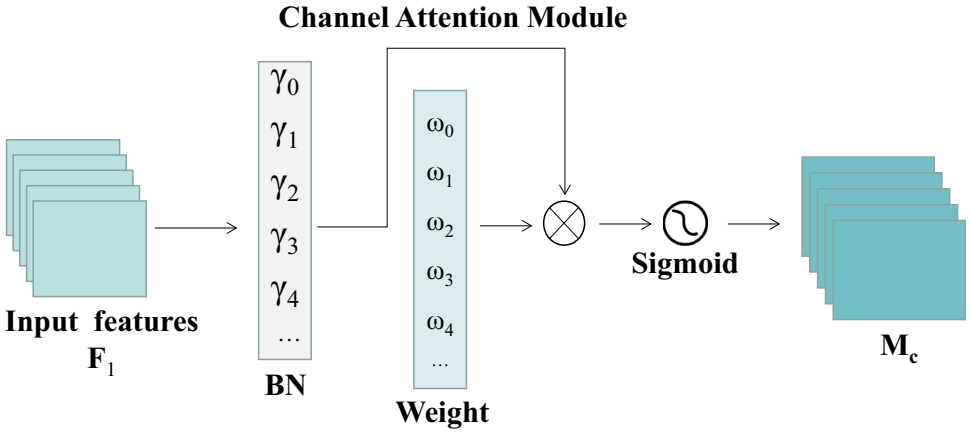

Fig. 5 Channel attention module

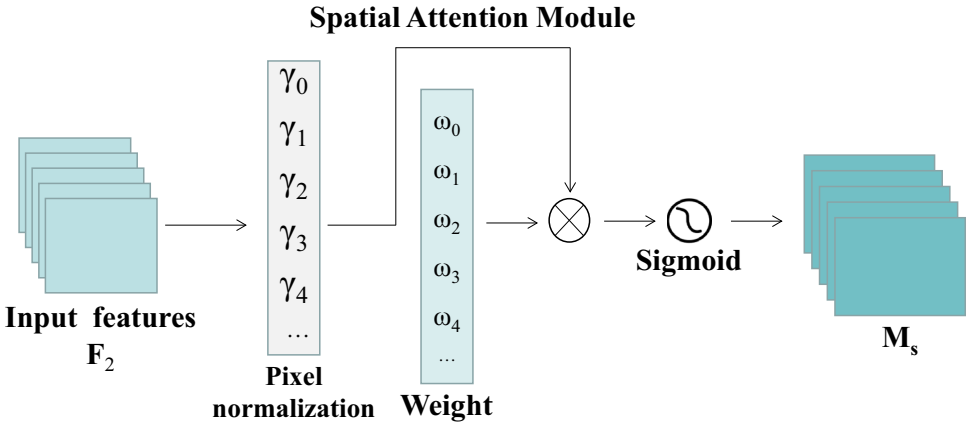

Fig. 6 Spatial attention module

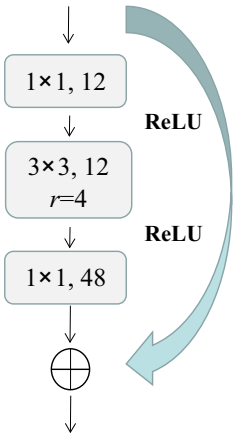

Fig. 7 Improved ResNet50 module

dataset used for the automatic pixelwise labeling of aerial imagery. The annotations are divided into two categories: buildings and non-buildings, primarily for semantic segmentation. The spatial resolution of the images is 0.3 m, ranging from densely populated areas like the financial district of San Francisco to small mountain towns like Lienz in Tyrol, Austria. The dataset consists of 360 Aerial orthorectified color imagery with a size of 5000 × 5000 pixels covers, including 5 cities: Austin, Chicago, Kitsap, Tyrol-w, and Vienna, with 72 images per city, and encompassing a ground area of 810 km.

In our study, the entire dataset is divided into two parts, the training set and the test set, each containing 180 images. Additionally, for validation purposes, this study selects the first five images from each city in the training set. During the training and testing process, each image with a resolution of 5000 × 5000 is first segmented into images with a resolution of 500 × 500.

Evaluation metrics and training procedure

(1) Evaluation metrics for algorithms. Here, the mean Intersection over Union (mIoU) and mean Pixel Accuracy (mPA) are used as evaluation metrics. Here, mIoU refers to the fraction of predicted pixels that are correct in the union of predicted and true pixels. mPA represents the proportion of correctly labeled pixels over the total pixels. The formulas of mIoU and mPA are as follows:

Table 3 Results of repeated experiments

Number of tests	mIoU/%	mPA/%
1	83.51	91.54
2	83.55	91.55
3	83.45	91.50
4	83.58	91.44
5	83.66	91.74
6	83.44	91.51
7	83.45	91.5
8	83.53	91.54
9	83.58	91.48
Mean	83.53	91.54
Std	0.00484	0.00638

$$\text{mIoU} = \frac{\text{TP}}{\text{FP} + \text{FN} + \text{TP}} \tag{9}$$

$$\text{mPA} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{10}$$

where *TP* represents the true example, that is, the model predicts a positive example, but the actual example is also positive. *FP* is for false positives, where the model predicts a positive example but the actual example is negative. *FN* stands for false negative example, which is a positive example when the model predicts a negative example.

(2) Network training process. The cross-entropy loss is chosen as the loss of the algorithm

$$\text{Loss} = \text{Ent}_{\text{loss}} = -\sum_{i=0}^{n} [y_i \times \log(\hat{y}_i) + (1 - \hat{y}_i) \times \log(1 - \hat{y}_i)], \tag{11}$$

where y_i is the truth value of a pixel (truth value 0 or 1 in binary classification tasks), $\hat{y}_i$ is the predicted value of a pixel, and n is the sample size for each calculation of loss.

Results and discussion

Repeated experiments

To ensure the statistical significance of the experiment with the improved method, nine repeated experiments were conducted. These experiments were performed under the same conditions, with 50 iterations, a batch size of 8, and a learning rate of 0.005. As shown in Table 3, considering the smaller variance, the mean value was selected as the final precision parameter for the improved model.

Table 4 Results of ablation studies

Method	mIoU/%	mPA/%
DeepLabv3+ (MobileNet)	82.31	89.37
ASPP	83.51	91.31
ASPP + ResNet50	83.49	91.43
ASPP + NAM	83.51	91.51
Improved DeepLabv3+	83.53	91.54

Ablation studies

Ablation studies in deep learning image segmentation is to analyze the importance and impact of different components within the model, such as network structures, loss functions, data augmentation methods, optimizers, etc. [27]. To analyze the importance and impact of added and improved components within the improved DeepLabv3+ , we conducted ablation studies to test the mIoU and mPA in three cases: improving only the ASPP module, improving ASPP and adding ResNet50, and improving ASPP and adding NAM. It is important to note that, to avoid the influence of the backbone network, MobileNet was chosen for all these experiments. As shown in Table 4, it can be observed that all the added modules are effective, and the comprehensively improved model exhibits the best prediction performance. Compared to the original DeepLabv3+ , the mIoU and mPA are increased by 1.22% and 2.17%, respectively, which represents a significant improvement. Furthermore, we can find that there are few differences for the values of mIoU and mPA within the three cases, showing that the improved ASPP module demonstrates the most noticeable effect in extracting high-level semantic information. Through the ablation studies, it can be also observed that not all module improvements and additional components can significantly enhance the segmentation accuracy of the network. It is essential to thoroughly investigate the original network structure to identify and improve the components that can substantially improve accuracy.

Comparison of different algorithms

To verify the segmentation accuracy of this method, DeepLabv3+ , U-Net and PSP-Net are selected as comparison models. These three comparison models and the improved DeepLabv3+ model are trained on the INRIA Aerial Image Dataset. The trained network model is used to predict the data in the validation set.

The comparison results are shown in Table 5. It can be seen that the improved DeepLabv3+ model in this paper achieves the best detection effect. In terms of accuracy, our method achieves an average pixel accuracy (MPA) of 83.58% and

Table 5 Accuracy comparison of different segmentation methods

Methods	mIoU/%	mPA/%
Original DeepLabv3+	82.31	89.37
U-Net	83.75	90.19
PSP-Net	81.31	88.1
Improved DeepLabv3+	83.53	91.54

Table 6 Efficiency of different segmentation methods

Method	Parameters	GFLOPs
Original DeepLabv3+	54,708,674	1882.83
U-Net	24,891,202	1806.69
PSP-Net	2,375,874	76.659
Improved DeepLabv3+	9,170,018	425.44

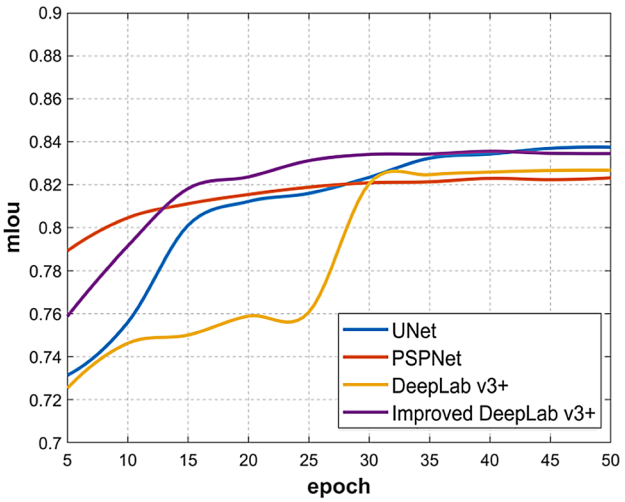

Fig. 8 Comparison of mIoU values on the INRIA Aerial Image Dataset

a mean intersection over union (MIoU) of 91.48%, which is an improvement of 0.22%, − 0.22%, 2.22% compared to the original DeepLabv3+ , U-Net, and PSP-Net, respectively. However, during model training, our improved method demonstrates significantly faster convergence compared to the original DeepLabv3+ , U-Net, and PSP-Net.

Compared to the original DeepLabv3+ , our method shows significant improvements in both segmentation effectiveness and speed. Compared to PSP-Net, our method demonstrates a notable improvement of 2.22% in average pixel accuracy (MPA) and 3.42% in mean intersection over union (MIoU), indicating a significant enhancement. Compared to U-Net, our improved method achieves a substantial reduction in model parameter quantity while maintaining excellent segmentation effectiveness and stronger computational capabilities.

In addition, we made further experiments to compare the efficiency of the training. The model parameters and floating-point operand of the proposed method are compared with the other methods when the same 500 × 500, three-channel data set image is input for prediction. As can be seen from the data in Table 6, the model parameters of the improved method are greatly reduced compared with the original DeepLab v3+ and U-Net which means the prediction will be faster. At the same time, the number of floating-point operations is smaller than the original DeepLab v3+ and U-Net requiring less computing power. What′s more, compared with the lightweight network PSP-Net, it does not have a significant advantage in segmentation speed, but in the above accuracy experiment, the improved method in this paper has shown its absolute advantage.

The improved method not only improves the accuracy, but also converges faster (Fig. 8). On the whole, the semantic segmentation effect of the proposed method on the INRIA Aerial dataset is better than that of the comparison methods.

In the experimental results, three representative result images are selected, the first two are different areas of the city Austin, the third is part of the city Chicago (Fig. 9). To better display the segmentation effect, we choose the houses which are scattered in first two images and houses which are clustered in the third image. The comparison shows that the improved method in this paper can better improve the hole problem in the original algorithm, the edge segmentation is more regular and clearer, and has good adaptability to the recognition and segmentation of small targets.

Comparison of different datasets

To verify the robustness of the model, the improved DeepLabv3+ network and the original DeepLabv3+ network are compared in the prediction results of INRIA Aerial Image Dataset, PASCAL VOC 2012 and SBD dataset, and BH-pools dataset, respectively. The prediction results on the same validation set are shown in Table 7.

From the experimental results, the improved DeepLabv3+ model has a significant improvement over the original DeepLabv3+ model on the three datasets tested. Moreover, on PASCAL VOC 2012 and SBD datasets, the improved DeepLabv3+ network in this paper has the largest improvement in mIoU compared with the original network, with an improvement value of 2.99%. This shows that the improved model in this paper has better effect on multi-category image segmentation tasks.

In general, this paper focuses on improving the original DeepLabv3+ in terms of segmentation speed and accuracy, resulting in a better overall performance compared to current

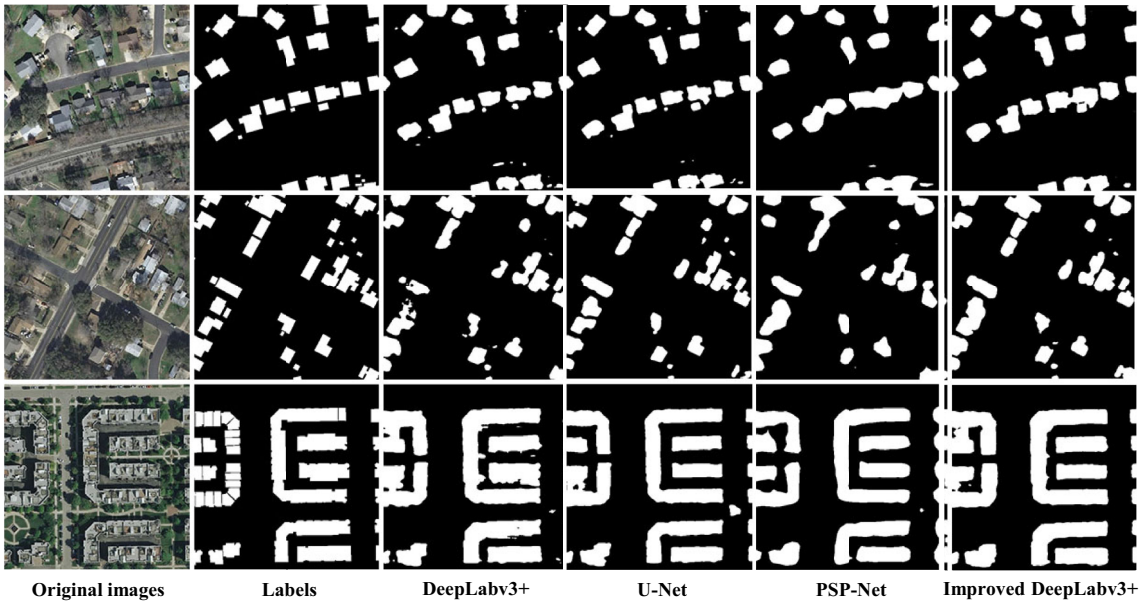

Fig. 9 Comparison of segmentation results among different methods

Table 7 Comparison of accuracies between original and improved DeepLabv3+

Method	Metrics	INRIA Aerial	BH-pools	PASCAL VOC 2012 and SBD
Original DeepLabv3+	mIoU/%	82.31	66.53	58.57
	mPA/%	89.37	99.21	88.91
Improved DeepLabv3+	mIoU/%	83.58	68.53	61.56
	mPA/%	91.48	99.37	89.98

mainstream algorithms. However, it is important to acknowledge that the current model still has some limitations. While the improved model exhibits faster segmentation speed compared to U-Net, the improvement in mIoU accuracy is not significant. On the other hand, compared to PSP-Net, the segmentation effect is notably enhanced, but the segmentation speed is still lacking. In the future, building upon the existing improvement ideas, we will continue to work toward further enhancing the segmentation effect while maintaining excellent segmentation speed. Our goal is to establish a network model with a superior overall performance.

Conclusion

This paper proposes an improved DeepLabv3+ model. To adapt to the real-time and dynamic data analysis requirements of remote-sensing images, it is necessary to reduce parameters while taking into account both efficiency and segmentation effect. In the improved network model, the backbone network is replaced by the lightweight network MoibleNetv2. In ASPP, the dilated convolution following the HDC principle can preserve the local information as much as possible without changing the size of the receptive field. What's more, we use the strip pooling module capture the local details in each direction from the horizontal and vertical dimensions. In the decoder, the network residual module of ResNet50 is added after low-level feature fusion to further obtain rich target edge feature information at low level. In addition, to enhance the ability of the network to capture the feature information of small target objects, the NAM attention mechanism is applied at multiple places to adapt to the segmentation requirements of complex environments. Experimental results show that compared with other classical semantic segmentation methods, the proposed method has faster convergence, higher accuracy, better effect, and better robustness on both remote-sensing image small object segmentation and multi-object classification datasets.

Acknowledgements This work was supported in part by the College Students' Innovative Entrepreneurial Training Plan Program (0510), and in part by the Provincial Quality Engineering Project of Anhui Provincial Department of Education (2021jyxm0060), in part by the Science and Technology Major Project of Anhui Province (202003a06020016), and in part by the Natural Science Foundation of Anhui Province (2008085MF184).

Data availability The dataset generated or analyzed during this study is available in the Inria Aerial Image Labeling, https://project.inria.fr/aerialimagelabeling/.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Huadong G, Changlin W (2005) Building up national Earth observing system in China. Int J Appl Earth Obs Geoinf 6(3–4):167–176
2. Song H, Yang W (2022) GSCCTL: a general semi-supervised scene classification method for remote-sensing images based on clustering and transfer learning. Int J Remote Sens 43(15–16):5976–6000
3. Huang C, Xiao C, Rong L (2022) Integrating Point-of-Interest density and spatial heterogeneity to identify urban functional areas. Remote Sens 14(17):4201
4. Adrian J, Sagan V, Maimaitijiang M (2021) Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine. ISPRS J Photogramm Remote Sens 175:215–235
5. Chen Y, Huo J, Li X, Bi K, Ma N, Jing Y, Ma X (2022) Classification and characteristic analysis of the clouds and dust in a dust-carrying precipitation process based on multi-source remote-sensing observations. Atmos Pollut Res 13(1):101267
6. Zhao T, Xu J, Chen R, Ma X (2021) Remote-sensing image segmentation based on the fuzzy deep convolutional neural network. Int J Remote Sens 42(16):6264–6283
7. Aouat S, Ait-hammi I, Hamouchene I (2021) A new approach for texture segmentation based on the Gray Level Co-occurrence Matrix. Multimed Tools Appl 80:24027–24052
8. Tian X, Chen L, Zhang X (2021) Classifying tree species in the plantations of southern China based on wavelet analysis and mathematical morphology. Comput Geosci 151:104757
9. Wang X, Zhai S, Niu Y (2019) Automatic vertebrae localization and identification by combining deep SSAE contextual features and structured regression forest. J Digit Imaging 32:336–348
10. Rao CS, Karunakara K (2022) Efficient detection and classification of brain tumor using kernel based SVM for MRI. Multimed Tools Appl 81(5):7393–7417
11. Dai J, Xue J, Zhao Q, Wang Q, Chen B, Zhang G, Jiang N (2020) Extraction of cotton seedling growth information using UAV visible light remote-sensing image. Trans Chin Soc Agric Eng 36(4):63–71
12. Li D, Zhang G, Wu Z, Yi L (2010) An edge embedded marker-based watershed algorithm for high spatial resolution remote-sensing image segmentation. IEEE Trans Image Process 19(10):2781–2787
13. Li C, Chen W, Wang Y, Ma C, Wang Y, Li Y (2021) Extraction of winter wheat planting area in county based on multi-sensor Sentinel data. Trans Chin Soc Agric Machinery 52(12):207–215
14. Wang Z, Wang J, Yang K, Wang L, Su F, Chen X (2022) Semantic segmentation of high-resolution remote-sensing images based on a class feature attention mechanism fused with Deeplabv3+. Compu Geosci 158:104969
15. Du S, Du S, Liu B, Zhang X (2021) Incorporating DeepLabv3+ and object-based image analysis for semantic segmentation of very high-resolution remote-sensing images. Int J Digital Earth 14(3):357–378
16. Su H, Peng Y, Xu C, Feng A, Liu T (2021) Using improved DeepLabv3+ network integrated with normalized difference water index to extract water bodies in Sentinel-2A urban remote-sensing images. J Appl Remote Sens 15(1):018504
17. Wang Z, Zhang H, Huang Z, Lin Z, Wu H (2022) Multi-scale dense and attention mechanism for image semantic segmentation based on improved DeepLabv3+. J Electron Imaging 31(5):053006
18. Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H (2018) Encoder-decoder with atrous separable convolution for semantic image segmentation, In: Proceedings of the European conference on computer vision (ECCV). 801–818
19. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T et al. (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861
20. Wang P, Chen P, Yuan Y, Liu D, Huang Z, Hou X, Cottrell G (2018) Understanding convolution for semantic segmentation. In: IEEE Winter Conference on Applications of Computer Vision (WACV). 1451–1460
21. Hou Q, Zhang L, Cheng MM, Feng J (2020) Strip pooling: Rethinking spatial pooling for scene parsing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 4003–4012
22. Liu Y, Shao Z, Teng Y, Hoffmann N (2021) NAM: Normalization-based attention module. arXiv preprint arXiv:2111.12419
23. Sun L, Cheng S, Zheng Y, Wu Z, Zhang J (2022) SPANet: successive pooling attention network for semantic segmentation of remote-sensing images. IEEE J Sel Topics Appl Earth Observ Remote Sens 15:4045–4057
24. Ronneberger O, Fischer P, Brox T (2015) U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III 18, Springer International Publishing. 234–241
25. Chen C, Papandreou G, Kokkinos I, Murphy K, Yuille AL (2017) Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans Pattern Anal Mach Intell 40(4):834–848
26. Maggiori E, Tarabalka Y, Charpiat G, Alliez P (2017) Can semantic labeling methods generalize to any city? The INRIA aerial image labeling benchmark. In: IEEE International Geoscience and Remote-sensing Symposium (IGARSS). 3226–3229
27. Meyes R, Lu M, de Puiseau CW, Meisen T (2019) Ablation studies in artificial neural networks. arXiv preprint arXiv:1901.08644