J. Meas. Sci. Instrum., 2025, 16(2): 205-215. http://jmsi.publish.founderss.cn/
DOI: 10.62756/jmsi.1674-8042.2025020 jmsi@nuc.edu.cn

Remote sensing image semantic segmentation algorithm based on improved DeepLabv3+

SONG Xirui[1,2], GE Hongwei[1,2*], LI Ting[1,2]

1. Engineering Research Center of Intelligent Technology for Healthcare, Ministry of Education, Jiangnan University, Wuxi 214122, China;
2. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China

***Corresponding author:** GE Hongwei (ghw8601@163.com)

Received: April 11, 2024 **Revised:** May 10, 2024 **Accepted:** July 19, 2024

Abstract: The convolutional neural network (CNN) method based on DeepLabv3+ has some problems in the semantic segmentation task of high-resolution remote sensing images, such as fixed receiving field size of feature extraction, lack of semantic information, high decoder magnification, and insufficient detail retention ability. A hierarchical feature fusion network (HFFNet) was proposed. Firstly, a combination of transformer and CNN architectures was employed for feature extraction from images of varying resolutions. The extracted features were processed independently. Subsequently, the features from the transformer and CNN were fused under the guidance of features from different sources. This fusion process assisted in restoring information more comprehensively during the decoding stage. Furthermore, a spatial channel attention module was designed in the final stage of decoding to refine features and reduce the semantic gap between shallow CNN features and deep decoder features. The experimental results showed that HFFNet had superior performance on UAVid, LoveDA, Potsdam, and Vaihingen datasets, and its cross-linking index was better than DeepLabv3+ and other competing methods, showing strong generalization ability.

Key words: semantic segmentation; high-resolution remote sensing image; deep learning; transformer model; attention mechanism; feature fusion; encoder; decoder

0 Introduction

With high-resolution remote sensing imagery becoming one of the primary means of acquiring geospatial information[1], it holds immense practical significance in fields such as disaster warning, urban planning, and agricultural planning. Semantic segmentation of remote sensing imagery is a core task in image interpretation, enabling pixel-level classification and obtaining more precise geospatial information[2]. Early remote sensing image segmentation primarily relies on a combination of visual interpretation and computer-assisted methods, including threshold-based segmentation[3], clustering-based segmentation[4], edge-based segmentation[5], graph-based segmentation[6], etc. However, these traditional segmentation methods often only extract low-level features from images, resulting in poor segmentation performance and failing to meet the requirements of high-resolution remote sensing imagery. Therefore, researchers have widely adopted deep learning methods for remote sensing image segmentation.

The DeepLab series of networks are commonly used models for image semantic segmentation. The DeepLabv1 network was first proposed[7], which combined deep convolutional networks, atrous convolution, and conditional random fields to enlarge the receptive field and improve segmentation accuracy. Chen introduced the DeepLabv2 network, which further expanded the receptive field through spatial pyramid pooling and reduced the number of parameters[8]. In DeepLabv3[9], multiple parallel atrous convolutions are used to capture contextual information. Chen proposed the DeepLabv3+ network, which employed an encoder-decoder structure and depthwise separable convolutions to further enhance segmentation accuracy[10].

Although these approaches have improved segmentation effectiveness, DeepLabv3+ utilizes atrous convolutions and pooling operations to enlarge the receptive field, and its decoder structure is relatively simplistic. This can result in the loss of detailed information during the semantic

segmentation process. Moreover, the upsampling and fusion operations in the decoder structure may lead to the loss of spatial details. In remote sensing imagery, detailed information is crucial for accurate semantic segmentation. This issue may result in the appearance of blurry or inaccurate boundaries in the segmentation results, leading to segmentation discontinuities and omissions for object edges and small targets. Therefore, additional mechanisms are required to preserve and restore details.

An improved approach, called hierarchical feature fusion network (HFFNet), was proposed for semantic segmentation tasks of remote sensing images, aiming to improve the performance of DeepLabv3+. Additionally, the potential application of transformers in remote sensing semantic segmentation was explored. The chosen encoders were the Swin transformer and ResNet-50. These encoder selections allowed for feature extraction from different perspectives and scales, enabling the capture of complementary feature information and improving segmentation accuracy. The ResNet-50 encoder is primarily responsible for extracting local detail information and global information, while the Swin transformer complements it by providing semantic information. Subsequently, a hierarchical feature fusion decoder was integrated to preserve more contextual information during the segmentation process. To further enhance the segmentation quality, a spatial channel attention module was introduced. This module refined the features, enhanced the expression of details, and reduced the semantic gap between shallow CNN features and deep decoder features. Comparative experiments and ablation studies demonstrated that HFFNet achieved more competitive results on various remote sensing datasets compared to other competitive methods, indicating its strong generalization capability.

1 Related work

1.1 DeepLabv3+

DeepLabv3+ adopts an encoder-decoder structure, where the atrous spatial pyramid pooling (ASPP) module is employed in the encoder to encode multi-scale contextual information, and the design of the decoder enables the network to restore to the original spatial dimensions.

The key component of the network architecture is the ASPP module. It utilizes atrous separable convolutions for multi-scale feature extraction. Addressing the lack of detail recovery capability in DeepLabv3, DeepLabv3+ improves the decoder structure, as illustrated in Fig. 1. Initially, the features from the encoder undergo a 4× bilinear upsampling, followed by concatenation with the low-level features from the backbone network after dimension reduction. Subsequently, the concatenated features are processed through convolution and bilinear upsampling operations to restore them to the spatial dimensions identical to those of the input image.

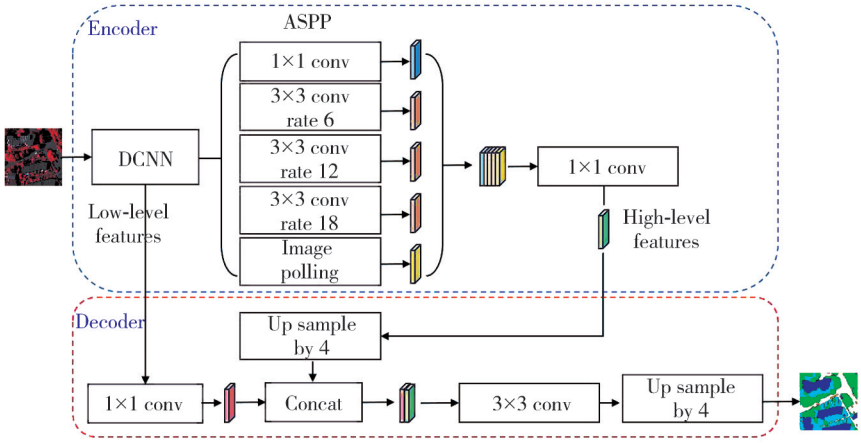

Fig. 1 Schematic diagram of DeepLabv3+ structure

1.2 Transformer

Transformers have made significant contributions to improving semantic segmentation performance in the remote sensing domain. SETR[11], based on vision transformer (ViT)[12], represents a prominent model for semantic segmentation. It replaces the CNN encoder with a purely transformer encoder, driving the development of semantic segmentation in recent years. Liu et al. proposed Swin transformer[13], which employed a shifted window strategy to limit the computation of multi-head self attention (MSA) within non-overlapping windows, forming window-based multi-head self attention (W-MSA). Additionally, to allow

for inter-window information interaction and establish dependency information across windows, they designed shifted window-based multi-head self attention (SW-MSA). Swin transformer exhibits linear computational complexity and achieves outstanding performance in various computer vision tasks.

Many existing works have been inspired by the characteristics and superior performance of Transformers. For instance, BANet[14] addresses challenging urban scene segmentation problems by constructing two feature extraction paths using ResT and convolution, proposing a dual-path cognitive network containing dependency and texture paths. STransFuse[15] combines the strengths of transformers and CNNs, introducing a staged model for extracting coarse-grained and fine-grained feature representations at different semantic scales. Xie et al. presented a framework called the pyramid grafting network (PGNet)[16], which independently extracted features using transformer and CNN encoders and then grafted features from the transformer branch onto the CNN branch. Transformers are widely used in computer vision fields, such as image recognition[17].

1.3 SE

In the realm of deep learning, the core concept of channel attention revolves around dynamically adjusting the weights of each channel based on its importance, allowing the network to learn the varying contributions of each channel. Hu et al. initially introduced the concept of channel attention and a novel neural network unit known as the SE (squeeze-and-excitation) block[18]. The ingenious design of this SE block lies in its ability to explicitly capture dependencies among channels, automatically adjusting the responses of feature channels, thereby enhancing the network's representational capacity.

The workflow of SE module is shown as Fig.2. Firstly, the input undergoes convolutional transformation, followed by squeeze through global average pooling to compute the average value of feature maps on each channel, providing a certain level of global receptive field. Subsequently, an excitation operation is performed to generate an excitation output for each feature channel, representing the weight of that channel's response. Finally, through reweighting, the weights of the excitation output are considered as the contributions of each feature channel, and these weights are multiplied by the original features to achieve reweighting of the original features along the channel dimension. This process aids the network in better understanding the data, highlighting important features, thereby improving the performance and generalization capability of the network.

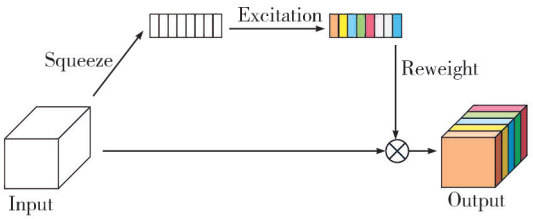

Fig. 2 SE block structure schematic

2 HFFNet

The structure of the HFFNet, an algorithm for semantic segmentation of remote sensing images proposed based on DeepLabv3+, is depicted in Fig.3. The grid structure comprises a CNN encoder, a transformer encoder, a hierarchical feature fusion decoder, and a spatial channel attention module. The CNN encoder utilizes a ResNet-50 pretrained on ImageNet and outputs four different scale feature maps $\{r_1, r_2, r_3, r_4\}$ using ResBlock-1, ResBlock-2, ResBlock-3, and ResBlock-4, respectively. The transformer encoder employs a Swin-T pretrained on ImageNet and outputs four feature maps $\{s_1, s_2, s_3, s_4\}$ at four stages of Swin-T, utilizing a cross-entropy loss function. Feature fusion methods play an important role in feature extraction of remote sensing images[19].

This combination harnesses the distinct advantages of the transformer encoder and CNN encoder to extract both global semantic information and detailed features, leading to more accurate segmentation results. By concurrently processing image inputs at different resolutions, the two encoders can effectively extract and integrate multi-scale features, thereby enhancing the model's performance. Firstly, the features from Swin-T are decoded to extract semantic information from deep layers, embedding this information into the second-stage decoder through concatenation and convolution operations. The second stage involves hierarchical feature fusion, initially fusing features from s_2 and r_4 followed by fusion with the outputs of deep layers of Swin-T, and then employing a series of convolution operations to merge s_1, r_3, and the outputs of decoder. Subsequently, the decoder outputs and shallow CNN features r_2 are passed into the decoder for decoding. In the final decoding stage, the feature information from r_1 is refined through a spatial channel attention module because the shallow layers of CNN retain rich spatial detail information crucial for segmenting small objects and boundaries. Finally, the decoder's output is upsampled to the same spatial resolution as the input, and after passing through convolution layers, the predicted segmentation map is obtained.

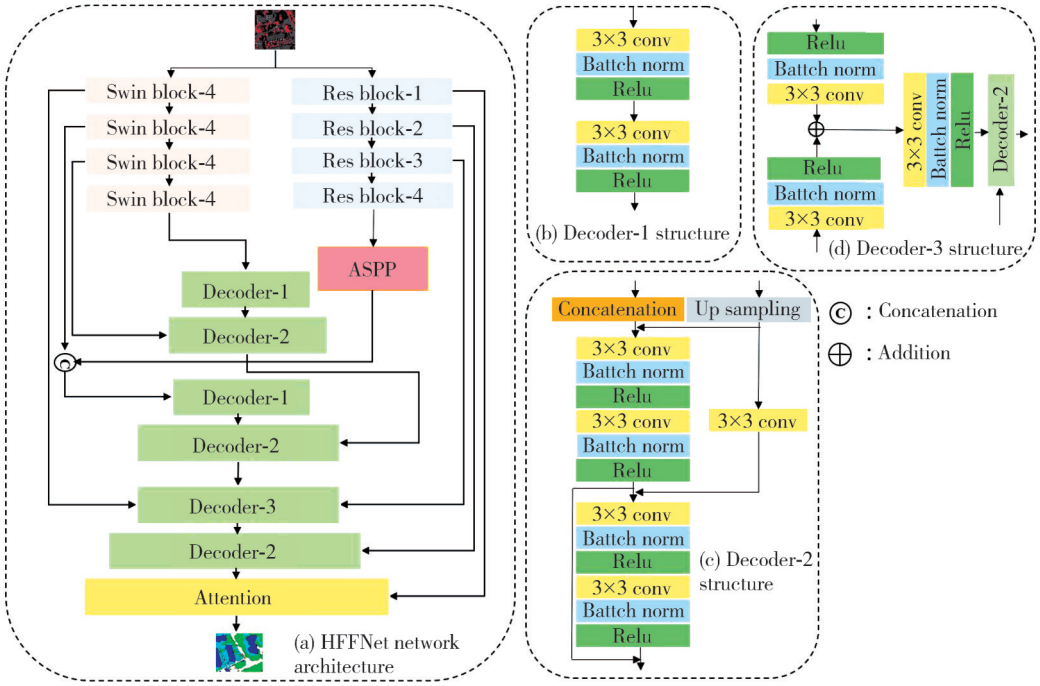

Fig. 3 Schematic diagram of HFFNet structure

2.1 Hierarchical feature fusion decoder

DeepLabv3+ employs numerous dilated convolutions and pooling operations for feature extraction, along with two 4× upsampling operations during image decoding. The use of dilated convolutions and pooling operations to expand the receptive field may lead to the loss of fine-grained details during the encoding and decoding processes. Bilinear interpolation, a local interpolation method, smooths the image to fill in newly generated pixel values. However, it cannot accurately restore the tiny details between original pixels, resulting in a certain degree of information loss. Moreover, high-rate upsampling further blurs the edges of the reconstructed image, adversely affecting boundary detection and segmentation results in remote sensing image segmentation tasks.

To mitigate these issues, a redesign of the decoder structure has been proposed. Firstly, the feature maps extracted at different stages have varying spatial scales, corresponding to different semantic scales. To ensure feature maps contain rich semantic information and retain more feature details, a staged fusion strategy is adopted, leading to the design of a hierarchical feature fusion decoder.

Before inputting features into the Swin encoder, they are downsampled to a fixed size of 224×224. Since the sizes of features s_3 and s_4 are fixed at 14×14 and differ significantly from the spatial dimensions of features output by ResNet-50, direct fusion is not suitable. Thus, s_4 is directly decoded through convolutional operations, and the decoded feature is fused with s_3. During the fusion process, bilinear interpolation and convolution operations are used to adjust the spatial dimensions and channel number to ensure feature dimension alignment. Features r_4 and s_2 have similar spatial dimensions, making them suitable for feature fusion. s_2 is upsampled using bilinear interpolation to match the spatial size of r_4, while r_4 features are directly input into the ASPP module. The concatenated features are then fed into the decoder. Similarly, r_3 and s_1 features, with similar spatial dimensions, are fused using convolutional operations. Additionally, to retain more detail information, skip connections are added between r_1 and r_2, enabling r_2 to be input alongside the fused features into the next decoder. In the final decoding stage, to reduce the semantic gap between shallow CNN features and deep decoder features, r_1 is input alongside decoder features into a spatial channel attention module for final feature refinement before segmentation.

2.2 Spatial channel attention module

The spatial channel attention module comprises two branches, namely spatial attention and channel attention, as depicted in Fig.4. Initially, the fusion of CNN features and decoder features is achieved through convolutional operations, and the resultant fusion is input into the attention module to fully leverage precise spatial detail information and semantic information. The design of the channel pathway is inspired by SE blocks, which generate a channel descriptor for each channel through global average

pooling, 1×1 convolution, and the Sigmoid function. Shallow networks can utilize these descriptors to obtain broader semantic information for better feature representation and learning. Subsequently, the feature maps are reweighted based on these descriptors to produce the output of the channel pathway.

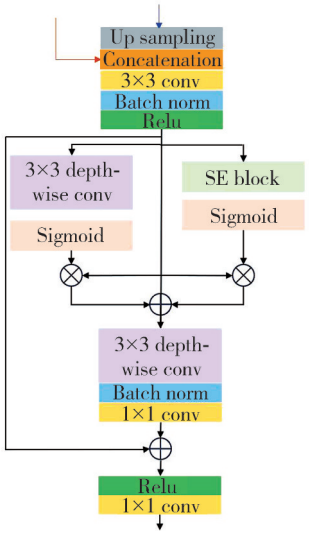

Fig. 4 Schematic diagram of space channel attention module

The spatial pathway generates spatial attention maps $\boldsymbol{S}\in\mathbf{R}^{h\times w\times 1}$ for each channel using deep convolution, where h and w represent the height and width of the feature map, respectively. The attention features generated by both pathways are further fused through summation. This fusion is then processed through deep convolution, normalization, and 1×1 convolution, with a residual connection added to prevent network degradation. Finally, the segmentation map is generated through convolution and the ReLU function. The output formula for spatial channel attention can be represented as

$$F_{\mathrm{c}}(\boldsymbol{X})=\alpha\big(\mathrm{SE}(\boldsymbol{X})\big)\otimes\boldsymbol{X}, \tag{1}$$

$$F_{\mathrm{s}}(\boldsymbol{X})=\alpha\big(\mathrm{Conv}(\boldsymbol{X})\big)\otimes\boldsymbol{X}, \tag{2}$$

$$F_{\mathrm{sc}}(\boldsymbol{X})=$$
$$\mathrm{Conv}_{1\times1}(\mathrm{BN}(\mathrm{Conv}_{3\times3}(F_{\mathrm{c}}(\boldsymbol{X})+F_{\mathrm{s}}(\boldsymbol{X}))))+\boldsymbol{X}, \tag{3}$$

where $\boldsymbol{X}$ represents the input feature; F_{c}, F_{s}, and F_{sc} denote the outputs of the channel pathway, spatial pathway, and spatial channel attention module, respectively. SE(·), Conv(·), and BN(·) represent the SE block, convolution operation, and batch normalization operation, respectively.

2.3 Loss funciton

During the training phase, the cross-entropy loss function is employed as the optimization objective for the model. The formula for calculating the loss is

$$L_{\mathrm{CE}}=-\frac{1}{N}\sum_{n=1}^{N}\sum_{k=1}^{K}y_k^{(n)}\log_{\mathrm{e}}\hat{y}_k^{(n)}, \tag{4}$$

where N and K denote the number of samples and the number of categories, respectively; $y_k^{(n)}$ and $\hat{y}_k^{(n)}$ denote the true label value and the output of the network, $n\in[1,\cdots,N]$; L_{CE} denotes cross-entropy loss.

3 Experiment

3.1 Data sets

The UAVid dataset consists of images captured by drones, focusing on urban street scenes, with two spatial resolutions (3 840×2 160 and 4 096×2 160) and eight categories. Due to its high spatial resolution, significant spatial variations, ambiguous categories, and generally complex scenes, segmentation of UAVid poses challenges. The dataset comprises a total of 420 images, of which, following official recommendations, 200 images are used for training, 70 images for validation, and 150 images for testing.

The LoveDA dataset contains 5 987 finely resolved optical remote sensing images from Nanjing, Changzhou, and Wuhan, with a spatial resolution of 1 024 pixel×1 024 pixel and 0.3 m. It includes urban and rural scenes, presenting significant challenges such as multiscale objects, complex backgrounds, and inconsistent category distributions. As official guidelines, 2 522 images are used for training, 1 669 images for validation, and 1 796 images for testing.

The Vaihingen dataset comprises 33 images with an average size of 2 494 pixel×2 064 pixel and a spatial resolution of 5 cm. It provides near-infrared, red, and green channels, as well as a digital surface model (DSM). Testing is performed using IDs: 2, 4, 6, 8, 10, 12, 14, 16, 20, 22, 24, 27, 29, 31, 33, 35, and 38, validation using ID: 30, and the remaining 15 images are used for training. In experiments, only the red, green, and blue channels and labels with eroded boundaries are utilized.

The Potsdam dataset consists of 38 finely resolved images with a resolution of 6 000 pixel×6 000 pixel and a spatial resolution of 5 cm. It provides near-infrared, red, green, and blue channels, as well as DSM and normalized DSM (NDSM). Testing is conducted using IDs: 2_13, 2_14, 3_13, 3_14, 4_13, 4_14, 4_15, 5_13, 5_14, 5_15, 6_13, 6_14, 6_15, and 7_13, validation using ID: 2_10, and training utilizing 22 images except 7_10 (which contains annotation errors).

3.2 Experimental setup

The experiment was conducted using Python 3.6 and PyTorch 1.10.0 on a single GeForce RTX 2080 graphics card. The AdamW optimizer was employed for training all models. The base learning rate was set to 0.000 6, and a cosine annealing strategy was applied for dynamic learning rate adjustment.

For the UAVid dataset, each image was padded and cropped into eight 1 024×1 024 pixel-sized patches. Random vertical flipping, random horizontal flipping, and random brightness adjustments were used as data augmentation strategies during training. The number of epochs was set to 40, with a batch size of 2. During testing, test-time augmentation (TTA) strategies such as vertical flipping and horizontal flipping were applied.

For the Vaihingen, Potsdam, and LoveDA datasets, images were cropped into 512×512 pixel-sized patches. During training, random scaling, random vertical flipping, random horizontal flipping, and random rotation were utilized as data augmentation strategies. The number of epochs was set to 100, with a batch size of 8. During testing, multi-scale and random flipping data augmentation strategies were applied.

3.3 Evaluation metric

The performance of HFTNet on the four datasets is evaluated using overall accuracy (OA), mean intersection over union (mIoU), and $F1$ score.

$$OA = \frac{\sum_{k=1}^{N} TP_k + TN_k}{\sum_{k=1}^{N} TP_k + FP_k + TN_k + FN_k}, \tag{5}$$

$$mIoU = \frac{1}{N}\sum_{k=1}^{N} \frac{TP_k}{TP_k + FP_k + FN_k}, \tag{6}$$

$$F1 = 2 \times \frac{precision \times recall}{precision + recall}, \tag{7}$$

where TP_k, FP_k, TN_k, and FN_k represent true positives, false positives, true negatives, and false negatives, respectively, for objects of class k; *precision* represents precision; and *recall* represents recall.

3.4 Comparision experiments

To evaluate the segmentation performance of the proposed model, HFFNet, comparisons were made with other widely used remote sensing image semantic segmentation models that also incorporated feature fusion, including ABCNet[19], BANet[14], MANet[20], UNetFormer[21], etc. To ensure fairness, all methods were evaluated using the same testing code, and the results are highlighted in bold font in the table.

A detailed comparison with other competitive networks is conducted on the official UAVid test set, as shown in Table 1. The proposed method achieves the highest mIoU (68.1%). Specifically, HFFNet not only outperforms the CNN-based ABCNet by 4.3% in terms of mIoU but also surpasses the recently introduced hybrid network UNetFormer based on transformer and CNN by 0.3%. Additionally, it outperforms the CNN and transformer based feature fusion network BANet by 3.5%.

Table 1 Performance comparison with other five methods on UAVid dataset

Model	IoU/%								mIoU/%
	background	Building	Road	Tree	Vegetation	Static car	Moving car	Human	
DeepLabv3+	68.7	**88.5**	81.1	**80.3**	63.7	53.9	61.1	22.9	65.0
BANet	66.6	85.4	80.7	78.9	62.1	52.8	69.3	21.0	64.6
ABCNet	67.4	86.4	81.2	79.9	63.1	48.4	69.8	13.9	63.8
UnetFormer	68.4	87.4	81.5	80.2	63.5	56.4	**73.6**	**31.0**	67.8
MANet	67.4	88.0	79.5	79.4	63.0	**64.6**	67.4	21.7	66.3
Swin SegFormer[22]	59.3	88.7	76.0	55.5	**73.3**	24.7	64.5	70.9	64.2
HFFNet	**69.2**	88.3	**82.3**	80.2	63.9	62.9	72.8	24.8	**68.1**

Particularly in the segmentation of static car, human, building, and road, HFFNet demonstrates significant advantages, with IoU values higher than the BANet method by 10.1%, 3.8%, 2.9%, and 1.6%, respectively. Static car and human are very small objects and are severely affected by lighting conditions in many scenarios, making them challenging to handle. Segmentation of building and road requires more semantic information and global context. Furthermore, the visualization results on the UAVid test set (Fig. 5) also demonstrate the effectiveness of HFFNet. Due to the lack of contextual and detailed information, DeepLabv3+ cannot accurately discern the surrounding scenes of segmented objects, resulting in blurred object edges and difficulty in identifying extremely small objects, while HFFNet can accurately segment them.

Comparative experiments were conducted on the LoveDA dataset to further evaluate the performance of HFFNet. As shown in Table 2, due to the guidance from Swin features during the decoding process, richer

semantic information was obtained, while the skip connections and attention mechanism strengthened the network's ability to preserve detailed features. The highest mIoU (53.4%) is achieved by HFFNet. Segmentation advantages are also maintained across multiple categories, with outstanding performance demonstrated in road and water surface segmentation results, surpassing other methods by at least 1.7% and 1% in IoU, respectively. The visualization results of the LoveDA test set are depicted in Fig.6.

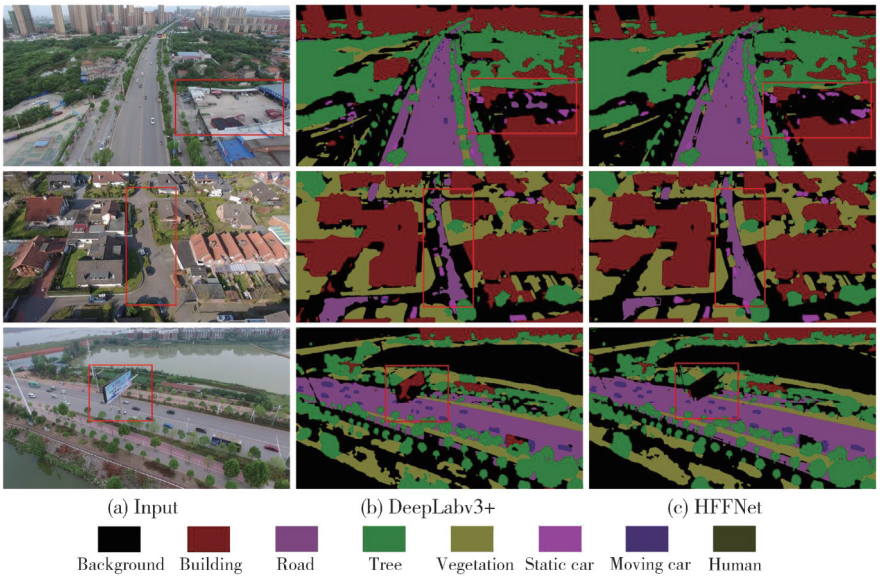

Fig. 5 Visualization results for UAVid test set

Table 2 Performance comparison with other six methods on LoveDA dataset

Model	IoU/%							mIoU/%
	Background	Building	Road	Water	Barren	Forest	Agriculture	
DeepLabv3+	45.5	56.7	57.6	80.2	20.9	41.5	62.8	52.1
BANet	43.7	51.5	51.1	76.9	16.6	44.9	62.5	49.6
SwinUperNet[13]	43.3	54.3	54.3	78.7	14.9	45.3	59.6	50.0
UnetFormer	44.7	58.8	54.9	79.6	**20.1**	46.0	62.5	52.4
MANet	46.5	57.5	53.2	79.3	18.3	47.0	**67.2**	52.7
LSKNet-T[23]	45.6	59.3	59.0	80.3	18.7	47.0	62.4	53.2
HFFNet	**46.8**	**59.0**	**59.3**	**81.2**	18.5	**47.7**	61.8	**53.4**

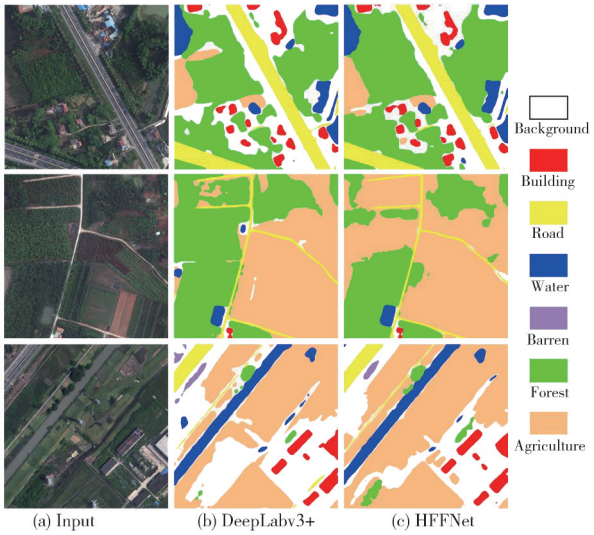

Fig. 6 Visualization results of LoveDA test set

To further validate the effectiveness of the proposed HFFNet, a comparison was conducted between HFFNet and other advanced approaches on the Vaihingen and Potsdam datasets. As shown in Table 3, the best average $F1$ (90.7%), OA (91.5%), and mIoU (83.2%) on the Vaihingen test set are achieved by HFFNet. In addition, segmentation advantages are maintained in multiple categories, outperforming other CNNS and transformer-based feature fusion networks, and mIoU outperforms other networks by at least 0.5%. Notably, for the low vegetation category, an $F1$ score of 85.9% is achieved by HFFNet, significantly surpassing other networks.

In the experiments on the Potsdam dataset, as presented in Table 4, HFFNet also outperforms recent hybrid networks based on transformer and CNN, such as UNetFormer and BANet, as well as CNN-based ABCNet and MANet. The segmentation results for each category demonstrate strong competitiveness compared to mainstream segmentation networks. On the Potsdam dataset, an average $F1$ score of 93.1%, OA of 91.6%, and mIoU of 87.2% are achieved by HFFNet.

Table 5 presents a comparison of parameter count and

computational complexity between HFFNet and others with an image input size of 1 024×1 024. It is found that HFFNet significantly improves segmentation accuracy compared to lightweight networks, with only a slight increase in parameter count and computational complexity.

Additionally, HFFNet maintains a certain advantage compared to other CNN and transformer fusion networks. Furthermore, in comparison with DeepLabv3+, there was only a marginal increase in parameter count and computational complexity, yet segmentation accuracy improved significantly.

Table 3 Performance comparison with other six methods on Vaihingen dataset

Model	*F*1/%					Mean *F*1/%	OA/%	mIoU/%
	Impervious surface	Building	Low vegetation	Tree	Car			
DeepLabv3+	91.6	94.1	82.5	88.0	77.7	86.7	89.1	77.1
BANet	92.2	95.2	83.8	89.9	86.8	89.6	90.5	81.4
ABCNet	92.7	95.2	84.5	89.7	85.3	89.5	90.7	81.3
UnetFormer	92.7	95.3	84.9	90.6	88.5	90.4	91.0	82.7
MANet	93.0	95.4	84.6	90.0	**88.9**	90.4	91.0	82.7
Hi-ResNet[24]	92.3	95.1	84.9	88.5	83.5	89.1	90.7	79.8
HFFNet	**93.4**	**95.8**	**85.9**	**90.7**	87.7	90.7	91.5	83.2

Table 4 Performance comparison with other six methods on Potsdam dataset

Model	*F*1/%					Mean *F*1/%	OA/%	mIoU/%
	Impervious surface	Building	Low vegetation	Tree	Car			
DeepLabv3+	92.1	95.3	85.6	86.5	94.8	90.9	89.2	84.2
BANet	93.3	96.7	87.4	89.1	96.0	92.5	91.1	86.3
ABCNet	93.5	96.9	87.9	89.1	95.8	92.7	91.3	86.5
UnetFormer	**93.6**	**97.2**	87.7	88.9	**96.5**	92.8	91.3	86.8
MANet	93.4	97.0	88.3	89.4	**96.5**	92.9	91.3	87.0
Hi-ResNet	93.2	96.5	87.9	88.6	96.1	92.4	91.1	86.1
HFFNet	93.5	97.1	**88.4**	**89.8**	96.4	93.1	91.6	87.2

Table 5 Comparison of parameters and calculations

Model	Number of parameters/$\times 10^6$	Number of FLOPs/$\times 10^9$
DeepLabv3+	40.4	123.3
BANet	12.8	53.2
ABCNet	14.0	62.9
SwinUperNet	60.0	349.1
UnetFormer	11.7	46.9
MANet	35.9	311.6
Swin SegFormer	29.5	277.2
LSKNet-T	4.5	27.4
Hi-ResNet	54.3	49.7
HFFNet	53.5	119.7

3.5 Ablation experiments

To validate the enhancement of network performance by the backbone network, hierarchical feature fusion decoder, and spatial channel attention module, ablation experiments were conducted on the UAVid dataset. Hierarchical feature fusion decoder is denoted as ①, and spatial channel attention module is denoted as ②.

Table 6 presents the results of the ablation experiments on the UAVid dataset. Compared to DeepLabv3+, when Swin-T is used as the backbone network alone, the segmentation accuracy sharply decreases due to its difficulty in extracting detailed features, especially for segmenting small objects, where the IoU for the human category is only 0.3%. When the hierarchical feature fusion decoder is used, the mIoU is significantly improved by 2.2%, attributed to the sufficient semantic information and rich contextual features, leading to more precise discrimination of small objects. This improvement respectively increases the IoU for static car, moving car, and human by 4.8%, 9.6%, and 1.6%. After incorporating the spatial channel attention module, the features are further refined, enhancing the network's ability to retain detailed information, resulting in a 1.1% increase in mIoU and improvements in segmentation accuracy for all categories to varying degrees. When all modules are utilized, leveraging both the semantic information of Swin and reinforcing the network's feature details, the mIoU increases by 3.1%. Notably, there is a significant improvement in the segmentation of distant and small objects, with the IoU for road, static car, moving car, and human increasing by 1.2%, 9.0%, 11.7%, and 1.9%, respectively.

In addition, to investigate the impact of different types of Swin transformers on feature fusion, ablation experiments targeting Swin transformer types were conducted on the UAVid dataset. The experimental results are shown in Table 7.

Table 6 Ablation experiments on UAVid dataset

Model	Backbone	IoU/%								mIoU/%	OA/%
		Clutter	Building	Road	Tree	Vegetation	Static car	Moving car	Human		
Base	Res50	68.7	88.5	81.1	80.3	63.7	53.9	61.1	22.9	65.0	86.8
Base	Swin-T	58.7	81.5	75.2	73.8	57.0	30.3	48.5	0.3	53.2	81.6
Base+①	Res50+Swin-T	69.2	88.8	81.0	80.5	64.0	58.7	70.7	24.5	67.2	87.1
Base+②	Res50	68.9	88.5	81.8	80.6	64.2	54.4	63.3	27.3	66.1	87.0
HFFNet	Res50+Swin-T	69.2	88.3	82.3	80.2	63.9	62.9	72.8	24.8	68.1	87.1

Table 7 Ablation experiments of different types of Swin transformer on UAVid dataset

Model	Backbone	mIoU/%
Base+①	Res-50+ Swin-B	65.6
Base+①	Res-50+ Swin-S	66.3
Base+①	Res-50+ Swin-T	67.2

The results indicate that Swin-T, with the fewest parameters, performs the best. While Swin-B, with the most parameters, exhibits the poorest performance. Swin-B, being deeper and larger compared to Swin-T, possesses more layers and parameters, giving it an advantage in modeling complex semantic information and handling larger-scale contexts. However, if the dataset is small or training resources are limited, Swin-B may require more data and longer training time to fully leverage its advantages. Therefore, in scenarios with small datasets or limited resources, Swin-B is more prone to overfitting during the training process.

Conversely, Swin-T is a more lightweight model with fewer parameters and shallower layers. This makes it easier to train and optimize, particularly in situations with smaller datasets, and it is less susceptible to overfitting, exhibiting stronger generalization capabilities.

4 Conclusions

A semantic segmentation algorithm HHFNet based on DeepLabv3+ for remote sensing images was proposed. To address the issue of insufficient semantic feature extraction by CNN networks and the limited capacity of DeepLabv3+ decoder to retain detailed information, a hierarchical feature fusion decoder was introduced. Swin-T was incorporated as the encoder, adopting a hierarchical feature fusion strategy guided by Swin-T features to provide a more comprehensive feature representation, enhancing object discrimination capability, and incorporating more detailed information during the decoding process. To tackle the coarse spatial resolution of deep decoder features, a spatial-channel attention module was designed. Through skip connections, high-resolution and detail-rich shallow CNN features were fused with decoder features to restore lost details. The spatial-channel attention module enabled the model to focus on important spatial positions and channel information, thereby enhancing the finesse of the segmentation map. Extensive experimental results demonstrated that the proposed algorithm exhibited superior feature representation capability and segmentation performance. Compared to other algorithms, the proposed method showed significant improvements in various metrics and performed well in different scenarios, especially for small objects and object boundaries. However, the network has a large number of parameters, and its computational complexity was higher compared to lightweight networks. Future work will focus on improving the network structure and optimizing the encoder and decoder further to enhance algorithm performance.

Acknowledgement

This work was supported by National Natural Science Foundation of China (No. 52374155), and Anhui Provincial Natural Science Foundation (No. 2308085MF218).

Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

References

[1] LI D R, WANG M, JIANG J. China's high-resolution optical remote sensing satellites and their mapping applications. Geospatial Information Science, 2021, 24(1): 85-94.

[2] TIAN X, WANG L, DING Q. Review of image semantic segmentation based on deep learning. Journal of Software, 2019, 30(2): 440-468.

[3] AL-AMRI S S, KALYANKAR N V, KHAMITKAR S D, et al. Image segmentation by using threshold techniques. 2010: 1005.4020. https://arxiv.org/abs/1005.4020v1.

[4] COATES A, NG A Y. Learning feature representations with k-means. Neural Networks: Tricks of the Trade: Second Edition. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 561-580.

[5] AL-AMRI M S S, KALYANKAR D N, DR

KHAMITKAR S D. Image segmentation by using edge detection. International Journal on Computer Science and Engineering, 2010, 2(3): 804.

[6] PENG B, ZHANG L, ZHANG D. A survey of graph theoretical approaches to image segmentation. Pattern Recognition, 2013, 46(3): 1020-1038.

[7] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. Semantic image segmentation with deep convolutional nets and fully connected CRFs. 2014: 1412.7062. https://arxiv.org/abs/1412.7062v4.

[8] CHEN L C, PAPANDREOU G, KOKKINOS I, et al. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4): 834-848.

[9] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation. 2017: 1706.05587. https://arxiv.org/abs/1706.05587v3.

[10] CHEN L C, ZHU Y K, PAPANDREOU G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation//Computer Vision-ECCV 2018, September 8-14, 2018, Munich, Germany, Cham: Springer International Publishing, 2018: 833-851.

[11] ZHENG S X, LU J C, ZHAO H S, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 20-25, 2021, Nashville, TN, USA. New York: IEEE, 2021: 6877-6886.

[12] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale. https://arxiv.org/abs/2010.11929.

[13] LIU Z, LIN Y T, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows//2021 IEEE/CVF International Conference on Computer Vision, October 10-17, 2021, Montreal, QC, Canada. New York: IEEE, 2021: 9992-10002.

[14] WANG L B, LI R, WANG D Z, et al. Transformer meets convolution: a bilateral awareness network for semantic segmentation of very fine resolution urban scene images. Remote Sensing, 2021, 13(16): 3065.

[15] GAO L, LIU H, YANG M H, et al. STransFuse: fusing swin transformer and convolutional neural network for remote sensing image semantic segmentation. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 10990-11003.

[16] XIE C X, XIA C Q, MA M C, et al. Pyramid grafting network for one-stage high resolution saliency detection//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-24, 2022, New Orleans, LA, USA. New York: IEEE, 2022: 11707-11716.

[17] YANG J, JIN Y X, LIU Y B, et al. Research on fabric material recognition method based on improved transformer. Journal of North University of China (Natural Science Edition), 2023, 44(2): 138-145.

[18] HU J, SHEN L, SUN G. Squeeze-and-excitation networks//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 7132-7141.

[19] LI R, ZHENG S Y, ZHANG C, et al. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of fine-resolution remotely sensed imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2021, 181: 84-98.

[20] LI R, ZHENG S Y, ZHANG C, et al. Multiattention network for semantic segmentation of fine-resolution remote sensing images. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 5607713.

[21] WANG L B, LI R, ZHANG C, et al. UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 190: 196-214.

[22] HE J N, LI W J, QU J T. Shifted window SegFormer for remote sensing image segmentation//2023 8th International Conference on Intelligent Informatics and Biomedical Sciences, November 23-25, 2023, Okinawa, Japan. New York: IEEE, 2023: 117-121.

[23] LI Y X, HOU Q B, ZHENG Z H, et al. Large selective kernel network for remote sensing object detection//2023 IEEE/CVF International Conference on Computer Vision, October 1-6, 2023, Paris, France. New York: IEEE, 2023: 16748-16759.

[24] CHEN Y, FANG P, YU J, et al. Hi-ResNet: A High-Resolution Remote Sensing Network for Semantic Segmentation. https://arxiv.org/abs/2305.12691.

基于改进DeepLabv3+的遥感图像语义分割算法

宋熙睿[1,2], 葛洪伟[1,2*], 李　婷[1,2]

1. 江南大学 康养智能化技术教育部工程研究中心, 江苏 无锡 214122;
2. 江南大学 人工智能与计算机学院, 江苏 无锡 214122

摘　要： 针对DeepLabv3+在高分辨率遥感影像的语义分割任务中, 卷积神经网络(Convolutional neural network, CNN)方法提取特征存在着感受野大小固定导致缺乏语义信息、解码器的高倍率上采样操作和简单的结构导致细节信息保留能力欠缺的问题, 本文提出了一种分层特征融合网络HFFNet(Hierarchical feature fusion network)。首先, 采用Transformer模型和CNN网络架构来提取不同分辨率图像的特征。然后, 独立地进行处理, 将来自Transformer的特征与CNN特征融合, 受不同来源特征的引导, 帮助网络在解码阶段更全面地还原信息。之后, 在解码的最后阶段, 设计了一个空间通道注意力模块, 细化特征, 缩小CNN浅层特征和解码器的深层特征之间的语义差距。实验结果表明, HFFNet在UAVid、LoveDA、Potsdam和Vaihingen数据集上表现出色, 交并比指标优于DeepLabv3+和其他有竞争性的先进方法, 展现出较强的泛化能力。

关键词： 语义分割; 高分辨率遥感图像; 深度学习; Transformer模型; 注意力机制; 特征融合; 编码器; 解码器

引用格式： SONG Xirui, GE Hongwei, LI Ting. Remote sensing image semantic segmentation algorithm based on improved DeepLabv3+. Journal of Measurement Science and Instrumentation, 2025, 16(2): 205-215. DOI: 10.62756/jmsi.1674-8042.2025020