Available Online: 14 October 2025 DOI: 10.54254/2977-3903/2025.27681

The application and challenges of deep learning in semantic segmentation of high-resolution remote sensing images

Shijing Hu

College of Environment and Resources, Zhejiang A&F University, Hangzhou, China

1582965112@qq.com

Abstract. Remote sensing images contain a wealth of geospatial information. To accurately identify different geospatial categories and extract relevant data, image semantic segmentation plays a crucial role. In recent years, deep learning technology has brought significant breakthroughs to semantic segmentation of remote sensing images, significantly enhancing its performance. This paper investigates the application of deep learning technologies in remote sensing image semantic segmentation, based on Convolutional Neural Networks (CNN) and Transformer-based semantic segmentation methods. It conducts an in-depth comparison of their structural characteristics and applicable scenarios, summarizes the achievements and shortcomings of existing research, and provides technical references and theoretical support for future studies, thereby contributing to the further development of deep learning technology in the field of remote sensing. Research results indicate that CNN-based semantic segmentation methods still hold advantages in extracting local features and achieving efficient segmentation, whereas Transformers address CNN's limitations in global context modeling and long-range dependency capture. Therefore, the collaborative integration of CNN and Transformers will become an important research direction for enhancing model performance in the future.

Keywords: deep learning, remote sensing image, semantic image segmentation, Transformer, attention mechanism

1. Introduction

Semantic segmentation of remote sensing images refers to the classification of pixels in remote sensing images to form pixel-level semantic labels, thereby accurately classifying each pixel into a specific category. During the early stages of remote sensing technology, semantic segmentation of images primarily relied on traditional methods and could only process grayscale images, resulting in low classification accuracy and high time consumption. However, with the continuous advancement of hardware and software, as well as the emergence of deep learning, methods such as CNN, RNN, and Transformers have gradually been applied to the field of remote sensing image semantic segmentation. The ability of deep learning to automatically extract high-level features has significantly enhanced image processing capabilities and classification accuracy [1-3]. However, existing reviews primarily focus on general image domains, with relatively few specialized reviews on semantic segmentation of remote sensing images. In particular, there are few systematic summaries regarding innovations in deep learning model architectures, high-resolution data processing, and practical application implementation.

This paper investigates the applicability and development trends of diverse deep learning models in remote sensing image processing through systematic review and analysis of mainstream methodologies. The paper examines three key model architectures: first, classic semantic segmentation models based on Convolutional Neural Networks (CNN), including FCN, U-Net, and DeepLabV3+, and mainly analyzes the evolution process of their network structures and their performance in remote sensing image processing [4-6]. Subsequently, the authors analyze CNN-based semantic segmentation enhancement modules incorporating attention mechanisms, with emphasis on SE-Net and CBAM, analyzing how they incorporate channel and spatial attention modules to enhance their ability to focus on key information [7, 8]. Finally, the paper introduces semantic segmentation methods based on the Transformer architecture, with a focus on ViT and Swin Transformer, to analyze the advantages of self-attention mechanisms in modeling long-distance dependencies and extracting global features [9, 10].

2. Semantic segmentation methods based on Convolutional Neural Networks (CNN)

CNNs are widely used in the field of computer vision. Through the extensive interconnection of neurons and convolutional operations, they extract features from multi-dimensional data information. CNN is mainly composed of input layers, convolutional layers, activation layers, pooling layers, fully connected layers, and output layers.

2.1. Evolution of classic architecture

Based on convolutional neural networks, improved networks for semantic segmentation tasks have begun to emerge and gradually improve the efficiency and accuracy of semantic segmentation. This section will specifically introduce the classic architectures FCN, U-Net, and DeeplabV3+, as shown in Table 1.

Table 1. Comparison table of FCN, U-Net, and DeeplabV3+

Architecture and Innovation Points	Advantage	Disadvantage	Usage scenarios
FCN: Modify the fully connected layer to a fully convolutional layer and an upsampling layer.	Accepts inputs of any size; maintains spatial structure.	The boundaries of the prediction results are not fine enough, making it difficult to restore the details.	Basic image segmentation tasks, low-resolution remote sensing image scenes
U-Net: The encoder-decoder structure adopts skip connections to retain the low-level features.	High segmentation accuracy; clear structural symmetry; suitable for tasks with a small number of samples.	Still has difficulty recognizing small targets and complex edges.	Medical image segmentation, high resolution remote sensing image segmentation, satellite image segmentation
DeepLabV3+: Introducing dilated convolution and the ASPP module to achieve multi-scale context modelling	Large receptive field; captures multi-scale information; strong robustness to complex backgrounds.	Complex structure; high dependence on hardware resources	Large scale remote sensing scenes, high precision semantic segmentation, complex scene segmentation

2.1.1. Fully Convolutional Network (FCN)

The model structure of FCN is shown in Figure 1 [4]. Compared with CNN, FCN modifies the traditional fully connected layer into a fully convolutional layer and an upsampling layer. Although CNN's pooling layers expand the receptive field and reduce computational cost, they also decrease the resolution of the feature maps, which is detrimental to image semantic segmentation tasks that rely on fine-grained pixel-level accuracy. Therefore, the fully convolutional layers and upsampling layers of FCN can solve the problems of feature map detail loss caused by CNN pooling layers and the disappearance of spatial position information caused by fully connected layers.

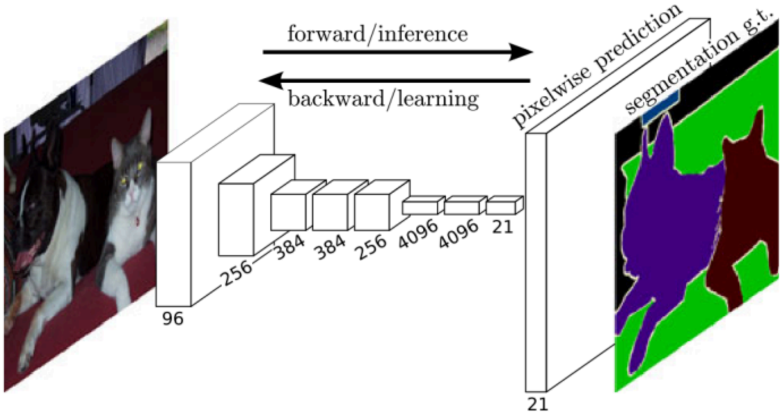

Figure 1. FCN semantic segmentation network model [4]

Although FCN has achieved relatively good results in image semantic segmentation, the prediction results it produces are still not sufficiently refined and do not consider the relationships between pixels [11]. To address the shortcomings of FCN in image semantic segmentation tasks, such as coarse segmentation edges and insufficient recovery of detailed information, Wang et al. used FCN as the base framework, adopted ResNet-50 as the base network, and introduced the ASPP (Atrous Spatial Pyramid Pooling) module to effectively combine multi-scale features [12]. They also replaced traditional transposed convolution with Dense Upsampling Convolution (DUC). Compared with the FCN-8s model, the proposed algorithm achieved a 2.58 percentage-point improvement in mIoU (from 83.604% to 86.185%), significantly enhancing pixel-level classification accuracy. This improvement effectively addressed the original FCN's poor performance in small object and edge segmentation.

2.1.2. U-Net

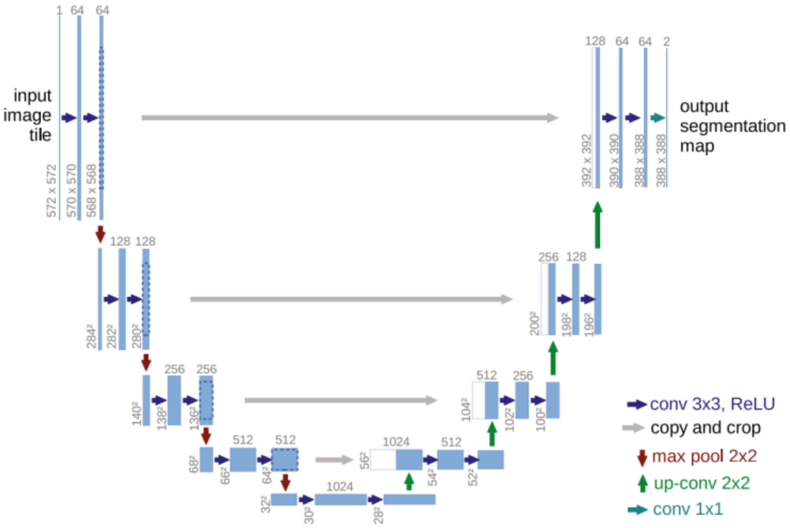

Figure 2. U-Net semantic segmentation network model [5]

U-Net mainly consists of an input layer, an encoder/downsampling stage, a decoder/upsampling stage, and an output layer [5]. As shown in Figure 2, U-Net introduces skip connections, which directly transmit the output features of each layer in the encoder to the corresponding layer in the decoder during the decoding process. This helps the decoder obtain more low-level feature information (such as edges and textures) and improves the accuracy of segmentation.

The structure of U-Net makes it particularly suitable for high precision positioning scenarios. Through skip connections, it can be used for satellite image segmentation, cell segmentation, and other applications. In recent years, improvements to U-Net have mainly focused on enhancing feature extraction capabilities and introducing attention mechanisms. Xiang and Huang [13] built upon the U-Net architecture by adding random data augmentation modules, Dual Encoder Fusion U-Net (DEFU-net) modules, and random voting prediction modules, enhancing the model across three dimensions: sample generation, model training, and prediction. The DEFU-net module enlarges the receptive field of convolutional layers and mitigates the adverse effects of small batch sizes on normalization, thereby enhancing model robustness. The random voting prediction module makes the model more robust and improves the accuracy of the results. After training and validation, the improved method achieved a stable accuracy of 0.9508 and a converged loss of approximately 0.0813 on the building extraction task, outperforming the original U-Net. The mIoU in the validation set improved from 0.7398 to 0.8128 [13]. These enhancements strengthen local feature representation while suppressing irrelevant cues, rendering U-Net more robust for segmenting complex structures in remote sensing imagery.

2.1.3. Deeplabv3+

Like U-Net, DeepLabV3+ also adopts an encoder-decoder structure, but unlike U-Net, DeepLabV3+ introduces Dilated Convolution and ASPP (Atrous Spatial Pyramid Pooling) modules [6].

DeepLabV3+ primarily consists of an input layer, encoder stages, decoder stages, and an output layer. The encoder stage comprises convolutional layers, activation layers, and pooling layers, as well as dilated convolutions and ASPP modules. Figure 3 shows the model structure of DeepLabV3+. In the encoder stage, as shown in Figure 4, DeepLabV3+ enlarges the receptive field by inserting holes between the elements of the convolution kernel using dilated convolution and utilizes the ASPP module to capture contextual information at different receptive field scales through parallel multi-branch dilated convolution, helping the model to capture feature information of different sizes and scales in the image.

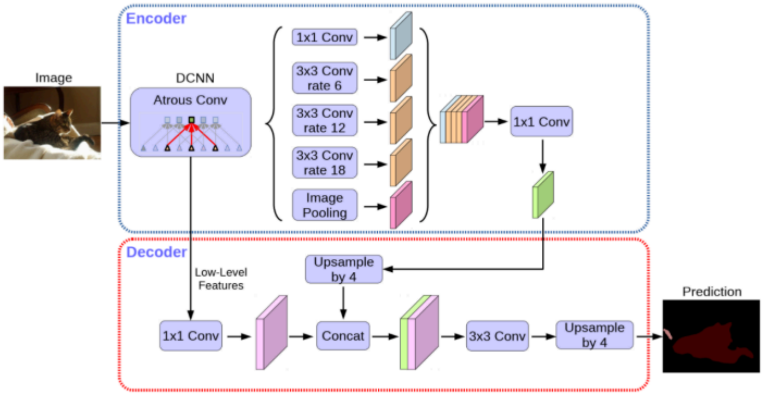

Figure 3. DeepLabV3+ semantic segmentation network model [6]

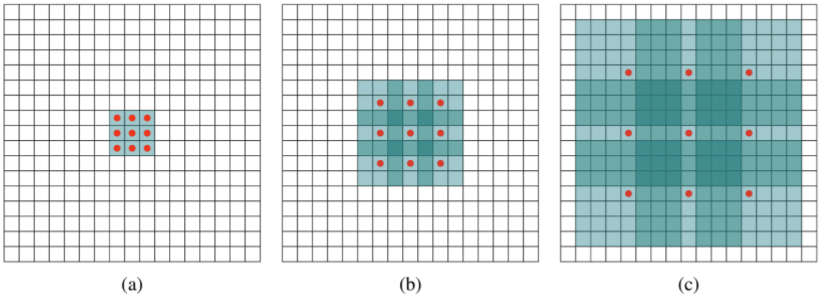

Figure 4. Schematic diagram of dilated convolution [14]

Therefore, DeepLabV3+ is suitable for high precision semantic segmentation and semantic segmentation in complex scenes, such as background blurring in video conferences and other types of portrait segmentation. The optimization of DeepLabV3+ focuses on combining attention mechanisms with improved feature fusion methods. Bai and Tang proposed the SMANet network model, which introduces a Strip Pooling Module (SPM) and the Multi-Parallel Atrous Spatial Pyramid Pooling (MASPP) Module to enhance DeepLabV3+'s multi-scale object segmentation capabilities in remote sensing scenarios [15, 16]. Experimental results on the Wuhan Dense Labeling Dataset (WHDLD) demonstrate that SMANet elevates DeepLabV3+'s mIoU from 61.30% to 64.18%.

2.2. CNN semantic segmentation enhancement module based on attention mechanism

In image semantic segmentation tasks, while the aforementioned methods can effectively extract spatial features from images, they often encounter issues such as excessive calculation time and low segmentation efficiency when handling complex scenes and detailed information, due to a lack of focus on important regions. Therefore, the attention mechanism has been introduced into image semantic segmentation methods. Based on the different dimensions of the information being focused on, attention mechanisms can be categorized into channel attention mechanisms, spatial attention mechanisms, temporal attention mechanisms, and self-attention mechanisms, among others. This section will focus on SE-Net and CBAM.

2.2.1. Squeeze-and-Exitation Networks (SE-Net)

SE-Net, a canonical channel attention mechanism, characterizes channel-wise dependencies via sequential squeeze and excitation operations.

Figure 5 shows the workflow of SE-Net. In the Squeeze stage, the SE module compresses each channel from H×W×C to 1×1×C through global average pooling, forming a channel descriptor. This descriptor encapsulates the global statistics of channel activations, facilitating bottom-up propagation of global contextual information [7]. In the excitation stage, this descriptor is used to obtain channel attention through two fully connected layers and activation functions that weight the channels.

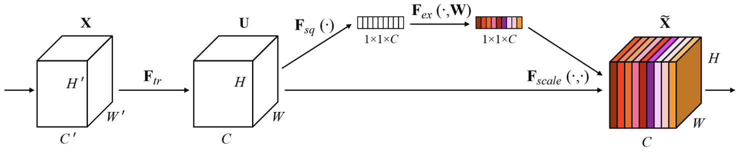

Figure 5. SE-Net workflow diagram [7]

Owing to its lightweight architecture and computational efficiency, which requires low computational resources, the SE module can be inserted at any position in the semantic segmentation network to improve model performance. However, SE-Net also has drawbacks, such as the difficulty of capturing complex global information through global average pooling operations and the increase in model parameters and computational overhead caused by subsequent fully connected operations [17]. Wang introduced a channel attention mechanism similar to SE-Net in U-Net, generating channel weights through global average pooling and fully connected layers to focus the model on building features, achieving an overall accuracy of 98.38% [18].

2.2.2. CBAM (Convolutional Block Attention Module)

CBAM is composed of a Channel Attention Module (CAM) and a Spatial Attention Module (SAM), and its workflow is shown in Figure 6. CBAM enables the network to concurrently emphasize informative channels and salient spatial locations [8].

Figure 7 shows a schematic diagram of CAM and SAM. The CAM module aggregates spatial information while preserving the channel dimension. CAM first performs global max pooling and global average pooling on the input feature map, compressing the feature map from H×W×C to 1×1×C. After passing through the activation functions (ReLU and Sigmoid) of the Multi-Layer Perceptron (MLP) module, it outputs the attention weights for each channel. The SAM module compresses the channel dimension while keeping the spatial dimension unchanged. SAM processes the CAM-refined features by concatenating max- and average-pooled maps of size H×W×1, then employs a 7×7 convolution followed by a sigmoid activation to produce the spatial attention map.

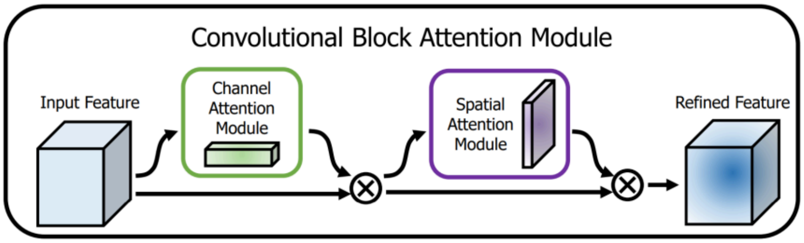

Figure 6. CBAM workflow diagram [8]

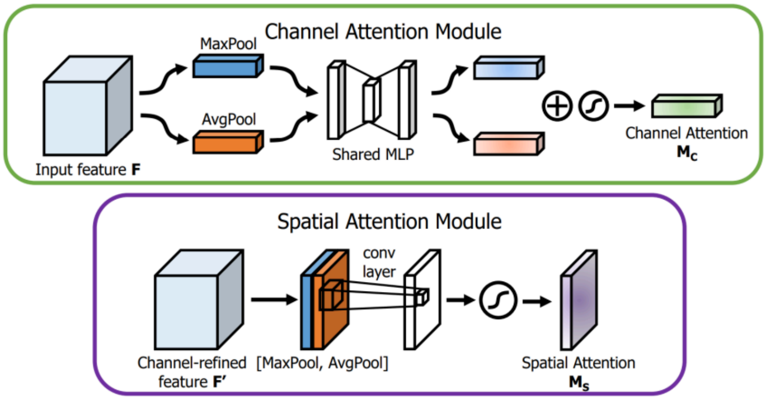

Figure 7. Schematic diagram of CAM and SAM [8]

CBAM combines channel attention and spatial attention through a cascading approach, first performing channel attention and then spatial attention, thereby achieving dual attention enhancement of feature maps. This cascaded design significantly enhances feature discriminability and segmentation accuracy. Sun et al. integrated the CBAM module into Res-UNet, where the model backbone adopts a U-Net model with an encoder-decoder structure, embedding residual structures into the encoder to effectively avoid model degradation issues [19]. Integrating CBAM enables the network to extract inter-object discriminative cues, sharpen object boundaries, and boost semantic-segmentation accuracy on remote-sensing imagery [19]. Compared to models without CBAM, the model's mIoU improved from 0.9481 to 0.9573, and the model also showed improvements in accuracy, pixel accuracy, and other metrics. These applications demonstrate that CBAM can effectively enhance the model's ability to focus on target features in remote sensing images by dynamically adjusting feature weights.

3. Transformer-based semantic segmentation method

To overcome the limitations of convolutional neural networks in capturing long-range pixel dependencies and to enhance the global modeling capacity of semantic segmentation models, the Transformer architecture has been introduced into computer vision. It achieves global modelling of features through the self-attention mechanism, offering stronger expressive capabilities and parallel processing advantages. This section will first introduce the principles of the self-attention mechanism, followed by an overview of two classic Transformer architectures: Vision Transformer (ViT) and Swin Transformer.

3.1. Transformer

Transformer mainly consists of encoder and decoder structures, with its core being the Self-Attention mechanism and Multi-head Attention mechanism. Its model structure is shown in Figure 8 [3].

The Transformer's encoder first uses an Input Embedding to represent the input as a sequence of embedding vectors of fixed dimension, and then stacks multiple identical layers on top of each other. Each layer includes the following steps:

(1) Multi-Head Attention: First, the input vector sequence X is linearly transformed into Key, Value, and Query vectors, and then the attention weights for each position are calculated according to formula (1):

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{1}$$

This operation calculates the attention weight for each position by calculating the relevance of each token to all other tokens, and then concatenates the parallel calculation results of multiple heads, as shown in Formula (2) and (3):

$$MultiHead(Q,K,V) = Concat(head_1,\ldots,head_h)W^O \tag{2}$$

$$wherehead_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

(2) Residual connection and layer normalization (Add&Norm): Add the output of Attention to the original input and normalize it.

(3) Feedforward neural network (Feed Forward): The network typically consists of fully connected layers and non-linear activation functions.

(4) Residual connection and layer normalization again.

Repeat these steps L times, then output the final feature representation.

The decoder of the Transformer first takes the decoder's output from the previous time step as input via the Output Embedding layers. Its structure is composed of multiple identical layers stacked together, with the following sequence: Masked Multi-Head Attention, Add&Norm, Multi-Head Attention, Add&Norm, Feed Forward, and Add&Norm. The key distinction lies in the use of Masked Multi-Head Self-Attention, which—via a causal mask—prevents the decoder from accessing future positions during training. Subsequently, the encoder-decoder Multi-Head Attention layer receives Queries from the masked self-attention output and Keys/Values from the encoder, aligning the target sequence with the source representation; Additionally, the input to the decoder's Multi-Head Attention is composed of the Query output from the Masked Multi-Head Attention and the Value and Key outputs from the encoder, thereby establishing a connection between the input and output. The overall structure is similar to that of the encoder.

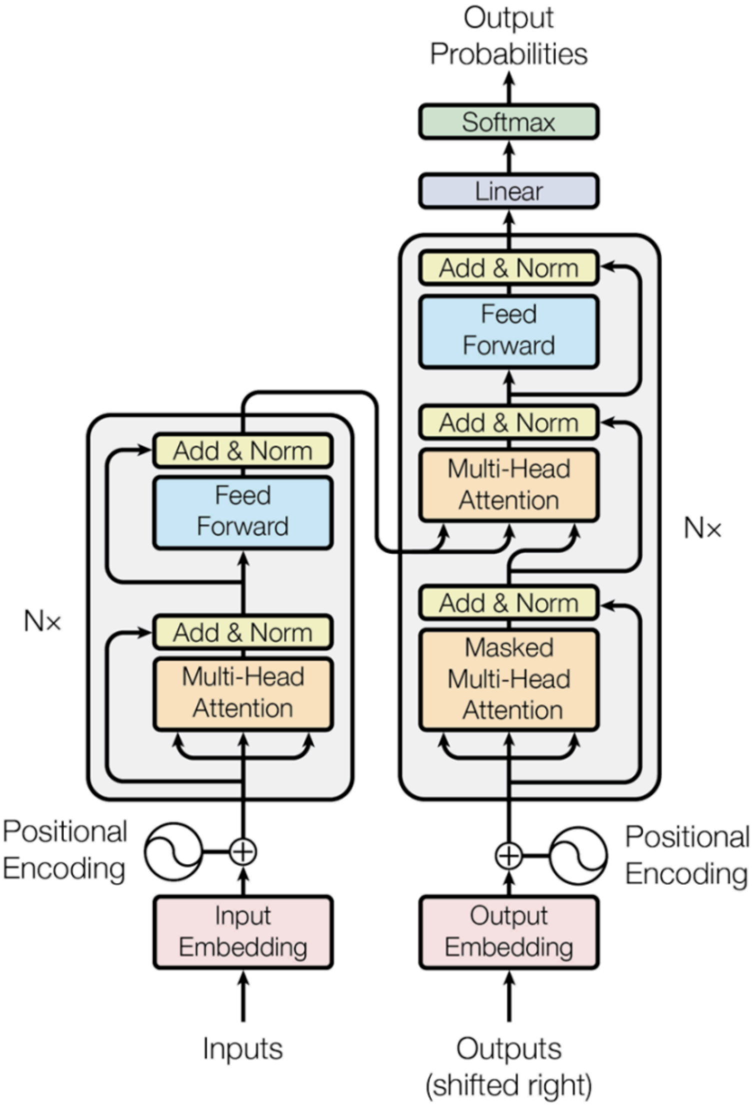

Figure 8. Transformer model architecture [3]

3.2. ViT (Vision Transformer)

Although Transformers have become standard in natural language processing, their application in vision is still limited. Vision Transformer (ViT), as the visual version of Transformer, is the first model to successfully apply the Transformer architecture to image classification tasks. Figure 9 shows the model architecture of ViT [9].

The structure of the Vit model is primarily divided into the following steps: First, the input image is divided into fixed-size non-overlapping patches, such as dividing a 224×224×3 image into 16×16 patches. Each patch is flattened into a 768-dimensional vector and linearly projected into a fixed-length embedding, yielding the initial patch-token sequence. However, since the Transformer itself cannot capture positional information, ViT introduces learnable Positional Embedding to preserve spatial structural information by adding it to the patch embedding. Additionally, to extract global features from the entire image, ViT adds a special character CLS at the beginning of the input sequence. This CLS can learn useful information from other embeddings to serve as a global representation of the entire image. This enriched sequence is forwarded through a stack of Transformer encoder layers—each composed of multi-head self-attention and feed-forward sub-networks—to generate final features for downstream tasks.

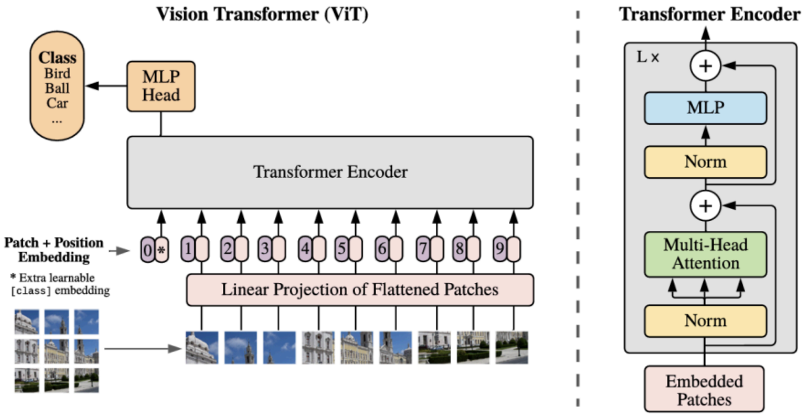

Figure 9. ViT model architecture [9]

The advantage of ViT lies in its strong global modelling capabilities. However, dividing images into small patches inevitably results in information loss (such as the loss of local continuity and adjacency relationships), which affects the final accuracy. Therefore, it is difficult to directly apply ViT to high resolution remote sensing images. Therefore, they proposed a feature-guided image semantic segmentation method based on ViT called ConTransNet. Through experimental evaluation on the Cityscapes dataset, the model achieved an mIoU value of 78.1%, significantly outperforming the base ViT [20].

3.3. Swin Transformer

Swin Transformer is an improvement on the ViT structure. It fully integrates the advantages of CNN in local modelling and hierarchical feature extraction while retaining the global modelling capabilities of Transformer [10]. This model introduces a sliding window attention mechanism to enable information exchange between different windows, thereby achieving efficient processing of high-resolution images.

Figure 10 shows the model architecture of the Swin Transformer. The Swin Transformer first divides the input image into 4×4 patches and flattens each patch before mapping it to a C-dimensional vector via linear projection, resulting in an initial feature map of size $\frac{H}{4} \times \frac{H}{4} \times C$. The model then extracts multi-scale features from the image through four stages. The first stage consists of a Linear Embedding layer and a Swin Transformer Block, where the Linear Embedding layer resizes the vector dimensions to a predefined value; the other three stages consist of Patch Merging and Swin Transformer Block. After each stage, Patch Merging merges adjacent small patches into a larger patch, thereby performing downsampling to reduce the spatial dimension of the feature map and the expansion of the number of channels.

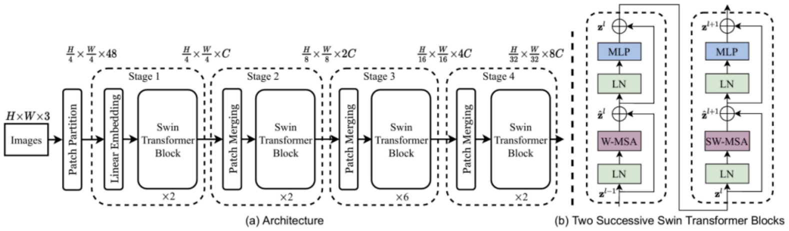

Figure 10. Swin Transformer model architecture [10]

Consequently, Swin Transformer outperforms ViT in modeling high-resolution imagery, making it particularly well-suited for remote-sensing semantic segmentation tasks.

Wang, Hu, Wu, Yan and Wang proposed an architecture that combines Swin Transformer and ResNet18 in parallel, with the former modelling global semantic relationships and the latter collecting rich spatial information [21]. On the Potsdam dataset, the mIoU score reached 86.1%, and on the Vaihingen dataset, the mIoU score reached 82.4%, showing significant improvements compared to other models such as FCN and DeepLabV3.

4. Discussion

4.1. Existing technical bottlenecks

Despite substantial advances, deep-learning-based remote-sensing semantic segmentation remains hindered by several practical challenges:

(1) Challenges in data acquisition and annotation:

High-quality, large-scale datasets are indispensable for training deep-learning models, yet pixel-level annotation demands extensive manual effort from domain experts. This pixel-level annotation work requires professionals to spend a significant amount of time and effort using specialized software. Scarcity and high annotation costs frequently result in overfitting and poor generalization.

(2) Limited model generalization ability:

Model generalization capability refers to the ability of a trained model to maintain good performance in new environments and regions. Due to the variability of remote sensing data influenced by different regions, weather conditions, and time, there is often a significant discrepancy between training data and actual application scenarios. Additionally, the high cost of labelling and the limited number of datasets can easily lead to model overfitting. Therefore, model generalization capability remains a challenge in remote sensing imagery. Although the CNN-based methods introduced in this paper, such as FCN, U-Net, and DeepLabV3+, perform well on specific datasets, they still suffer from issues like overfitting and weak transferability. Modules like SE-Net and CBAM have improved the ability to focus on key areas, but they still cannot fully address the shortcomings of models in cross-scenario and cross-region applications; Transformer models also face issues such as large parameter counts and strong dependence on training data, making it difficult to fully ensure their generalization capabilities.

4.2. Solution direction

In response to the current bottlenecks in semantic segmentation of remote sensing images in terms of data and application generalization, future research can be improved in the following areas:

(1) Building a high-quality, diverse remote sensing image dataset

Future efforts should prioritize curating and openly sharing diverse, high-quality remote-sensing datasets to enhance model efficacy and generalizability. Augmentation strategies and unsupervised learning should be jointly leveraged to mitigate annotation costs and data scarcity, boosting performance in low-resource settings.

(2) Combining the advantages of multiple deep learning architectures to improve model adaptability

Relying on a single architecture rarely achieves an optimal trade-off between local detail and global context. Hybrid designs that synergize CNNs' local inductive biases with Transformers' global receptive fields are therefore a promising direction.

(3) Enhance model generalization capabilities across regions and scenarios

The transferability and robustness of models are key to the practical application of remote sensing segmentation. Future research can use cross-modal learning and other methods to improve the adaptability of models to images obtained from different regions and sensors, thereby enhancing their performance stability in different scenarios.

5. Conclusion

With the advancement of deep learning, semantic segmentation techniques for remote sensing imagery have gradually shifted from traditional CNN-based semantic segmentation methods to more flexible Transformer-based semantic segmentation frameworks. CNN-based approaches—exemplified by FCN, U-Net, and DeepLabV3+—have continually pushed the boundaries of segmentation accuracy and computational efficiency. They have also introduced structural innovations such as skip connections, dilated convolutions, and attention mechanisms to adapt to the diversity and complexity of remote sensing images. For example, attention mechanisms represented by SE-Net and CBAM can significantly enhance the model's focus on key regions and important features, thereby improving segmentation accuracy while maintaining computational efficiency. However, due to the limited receptive field of convolutional operations, it is challenging to model relationships between distant pixels. Therefore, Transformers have been introduced into image semantic segmentation tasks. Transformers break the local constraints of convolutional operations by introducing self-attention mechanisms, demonstrating strong global modelling capabilities. For example, ViT and Swin Transformer have shown promising application potential in high-resolution processing tasks for remote sensing images. In the future, research on remote sensing-based semantic segmentation will further develop in the directions of structural model integration, application scenario generalization, and model lightweighting. By combining the advantages of CNN and Transformer, it is expected to improve segmentation accuracy while addressing model generalizability and practicality issues, thereby enabling the efficient application of remote sensing images in fields such as smart cities and disaster monitoring.

References

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (2002). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https: //doi.org/10.1109/5.726791

[2] Jordan, M. I. (1997). Serial order: A parallel distributed processing approach. In J. W. Donahoe & V. P. Dorsel (Eds.), *Advances in psychology* (Vol. 121, pp. 471–495). Elsevier Science. https: //doi.org/10.1016/S0166-4115(97)80111-2

[3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*. https: //proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[4] Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3431–3440). https: //doi.org/10.1109/CVPR.2015.7298965

[5] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention* (pp. 234–241). Springer. https: //doi.org/10.1007/978-3-319-24574-4_28

[6] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., & Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. *Proceedings of the European Conference on Computer Vision* (ECCV) (pp. 801–818). https: //doi.org/10.1007/978-3-030-01234-2_49

[7] Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7132–7141). https: //doi.org/10.1109/CVPR.2018.00745

[8] Woo, S., Park, J., Lee, J.-Y., & Kweon, I. S. (2018). Cbam: Convolutional block attention module. *Proceedings of the European Conference on Computer Vision* (ECCV) (pp. 3–19). https: //doi.org/10.1007/978-3-030-01234-2_1

[9] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2020). *An image is worth 16x16 words: Transformers for image recognition at scale*. arXiv. https: //arxiv.org/abs/2010.11929

[10] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 10012–10022). https: //doi.org/10.1109/ICCV48922.2021.00986

[11] He, C., Li, S., Xiong, D., & Fang, L. (2020). Remote sensing image semantic segmentation based on edge information guidance. *Remote Sensing, 12*(9), 1501. https: //doi.org/10.3390/rs12091501

[12] Wang, H., Qu, Y., Liu, Z., Wang, J., & Liu, J. (2022). Research on Image Semantic Segmentation Algorithm based on FCN. *Journal of Chengdu Technological University, 25*(1), 36–41.

[13] Xiang, Y., & Huang, Z. (2024). A Building Segmentation Method for Remote Sensing Imaged Based on Improved Unet Network. Urban Geotechnical Investigation & Surveying, (1), 109–113.

[14] Yu, F., & Koltun, V. (2015). *Multi-scale context aggregation by dilated convolutions*. arXiv. https: //arxiv.org/abs/1511.07122

[15] Bai, J., & Tang, B. (2023). Research on Semantic Segmentation of Remote Sensing Images Based on Deep Learning. *Practical Electronics, 31*(14), 79–82.

[16] Hou, Q., Zhang, L., Cheng, M.-M., Feng, J., & Yan, S. (2020). Strip pooling: Rethinking spatial pooling for scene parsing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4003–4012). https: //doi.org/10.1109/CVPR42600.2020.00406

[17] Song, K. (2023). *Research on Image Semantic Segmentation Depth Model Based on Attention Mechanism* [Master's thesis, Inner Mongolia University].

[18] Wang, Z. (2024). *Research on building semantic segmentation method in high-resolution remote sensing images based on deep learning* [Master's thesis, Anhui University of Technology]. https: //doi.org/10.26918/d.cnki.ghngc.2024.000360

[19] Sun, L., Zhao, L., Li, C., & Ma, Y. (2024). Res-UNet High-resolution Remote Sensing Image Semantic Segmentation Model Integrated into CBAM. *Geospatial Information, 22*(2), 68–70.

[20] Wang, L. (2024). *Research on Image Semantic Segmentation Method Based on Vision Transformer* [Master's thesis, Qufu Normal University]. https: //doi.org/10.27267/d.cnki.gqfsu.2024.000920

[21] Wang, Y., Hu, Y., Wu, X., Liu, Y., & Zhang, Y. (2024). *Semantic Segmentation Method for Remote Sensing Images Based on Improved Swin Transformer* [Master's thesis, Lanzhou Jiaotong University]. https: //doi.org/10.27205/d.cnki.gltec.2024.001078