Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090102
ISSN 2371-8838 Vol. 9 Num. 1

A Comparative Study of Deep Learning-Based Semantic Segmentation Methods for High-Resolution Remote Sensing Imagery

Yongsong Jiang*

School of Information Science and Technology, Yunnan Normal University, Yunnan, Kunming, China
junyanshuang@qq.com
**Corresponding author*

Keywords: Deep Learning; Convolutional Neural Networks; Remote Sensing Imagery; Semantic Segmentation

Abstract: Remote sensing image information extraction plays a crucial role in land use planning, environmental monitoring, and natural disaster assessment. However, traditional machine learning-based methods often face challenges such as high computational complexity and limited feature representation ability when processing large-scale remote sensing data, leading to difficulties in meeting both efficiency and accuracy requirements. With the rapid development of deep learning, its application to remote sensing data processing has become a powerful solution. This paper uses the standard Potsdam dataset provided by ISPRS and tests and compares the accuracy of several commonly used deep learning convolutional networks, including SegNet, PspNet, Unet, UNet++, DeepLab V3+, SegFormer, and SegVit, in remote sensing image information extraction. Experimental results show that SegVit performs exceptionally well in accuracy, detail preservation, and edge clarity, achieving higher precision compared to other networks. This finding provides an effective solution for remote sensing image information extraction and offers strong support for research and applications in related fields. It is worth noting that although SegVit excels in accuracy, it may require more computational resources and time during training and inference. Therefore, in practical applications, it is necessary to balance efficiency and accuracy and choose a network model that suits the specific task requirements.

1. Introduction

Remote sensing image segmentation, as a core task in remote sensing data processing, is widely applied in fields such as feature extraction, land use classification, and environmental monitoring. With the rapid advancement of remote sensing technology and the large-scale acquisition of data, image segmentation has become a crucial step in extracting key information from vast datasets. By dividing remote sensing images into different regions or objects, precise feature extraction and

classification can be achieved, providing vital support for surface analysis and decision-making. Traditional image segmentation methods, such as K-means[1], Expectation-Maximization (EM) algorithm[2], decision trees[3], Support Vector Machines (SVM)[4], Maximum Likelihood Estimation[5], and Random Forests (RF)[6], primarily rely on spectral features[7], neglecting the texture and spatial context information in high-resolution remote sensing images, leading to generally lower segmentation accuracy.

With the rise of deep learning, particularly the breakthrough of AlexNet[8] in the ImageNet[9] image classification competition, Convolutional Neural Networks (CNN) have become the mainstream approach in the field of computer vision. Fully Convolutional Networks (FCN)[10] transform semantic segmentation tasks into pixel-level classification by replacing the fully connected layers in image classification networks with convolutional layers, allowing feature extraction networks to use CNNs for more efficient segmentation. This development has led to the rapid emergence of CNN-based semantic segmentation models, which have achieved excellent segmentation results.

Remote sensing image segmentation plays an irreplaceable role in surface information extraction and environmental monitoring. Despite achieving high precision in current remote sensing image semantic segmentation technologies, challenges such as complex object boundary extraction, anomaly data handling, and large-scale data processing still remain. Future research directions should focus on multi-source data fusion, the integration of deep learning with traditional methods, and cross-temporal and spatial scale segmentation technologies. This paper provides an in-depth analysis of existing remote sensing image segmentation methods, their advantages and disadvantages, offering insights for future algorithm improvements and practical applications. To meet the growing demand for high-precision remote sensing data processing, cooperation between governments, academia, and engineering professionals should be strengthened to jointly advance the development of remote sensing image segmentation technology. This paper primarily analyzes the accuracy performance of deep learning methods in remote sensing image semantic segmentation and presents a comparative study of classic network models such as Unet, Unet++, SegNet, PspNet, DeeplabV3+, SegFormer, and SegVit.

2. Common Network Architectures

2.1. Unet

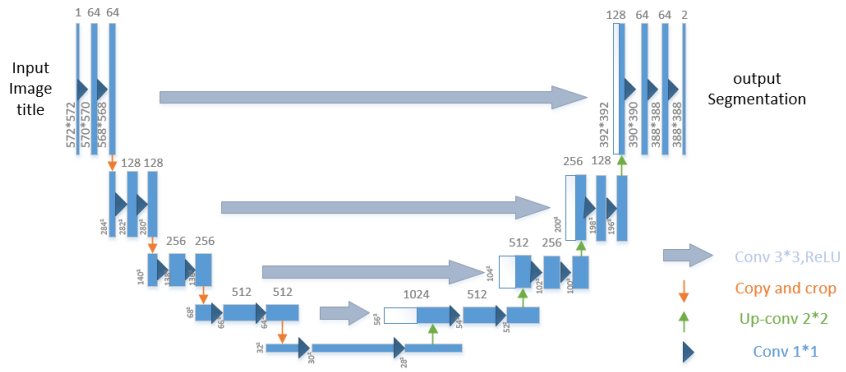

Figure 1: U-Net architecture

U-Net[11] is a classic fully convolutional neural network, named for its shape that resembles the letter "U." Proposed by Olaf Ronneberger et al. in 2015, U-Net achieved significant success in

medical image segmentation. The network is divided into an encoder and a decoder. The encoder consists of multiple convolutional blocks, gradually reducing the size of the feature maps while extracting semantic features. The decoder restores the feature map size through upsampling layers, while integrating the semantic features from the encoder with the fine-grained features from the decoder. The skip connection structure allows the network to capture both global and local information simultaneously, thereby improving segmentation accuracy. Its simple yet effective design has made it widely used in various segmentation tasks. Its structure diagram is shown in Figure 1.

2.2. DeeplabV3+

DeepLabv3+[12], proposed by Chen et al. in 2018, is the latest version of the DeepLab series. Its key innovations include the introduction of the Atrous Spatial Pyramid Pooling (ASPP) module and depthwise separable convolutions. The ASPP module processes features with multiple parallel dilated convolution branches to capture semantic information at different scales, effectively addressing varying receptive fields in the image. Depthwise separable convolutions optimize computational efficiency and reduce model complexity. Another significant improvement is the adoption of an encoder-decoder structure. The encoder extracts features and performs multi-scale context feature extraction via the ASPP module, while the decoder upsamples the features and integrates low-level and high-level features, enhancing detail preservation through skip connections. This model has been widely applied in fields such as autonomous driving and medical image segmentation, achieving significant results. Its structure diagram is shown in Figure 2.

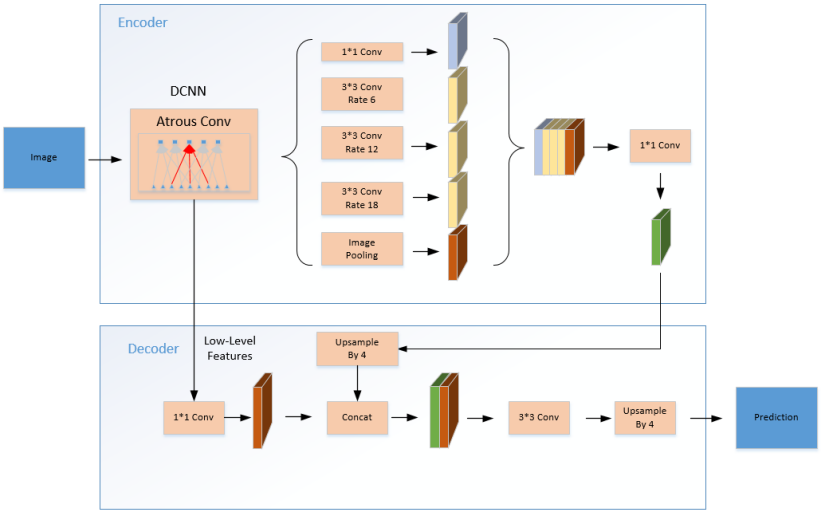

Figure 2: DeepLabV+ architecture

2.3. SegVit

SegVit[13], proposed in 2022, combines the Vision Transformer (ViT) with the U-Net structure, aiming for efficient and accurate image segmentation. The design of SegVit is inspired by ViT, which utilizes a self-attention mechanism to capture relationships between different locations in an image and employs a multi-layer Transformer encoder to extract features. In SegVit, the encoder part of ViT serves as the feature extractor and is combined with a symmetric decoder part, effectively leveraging both global contextual information and local details to enhance semantic segmentation performance. Additionally, SegVit introduces a multi-scale attention mechanism, using multiple attention heads in both the encoder and decoder to focus on features at different

scales, further improving segmentation accuracy and detail preservation. SegVit excels in handling complex scenes, small objects, and fine details, with strong semantic reasoning and global context-awareness capabilities. It has been widely applied in fields like autonomous driving and medical image segmentation, becoming an important segmentation network model. Its structure diagram is shown in Figure 3.

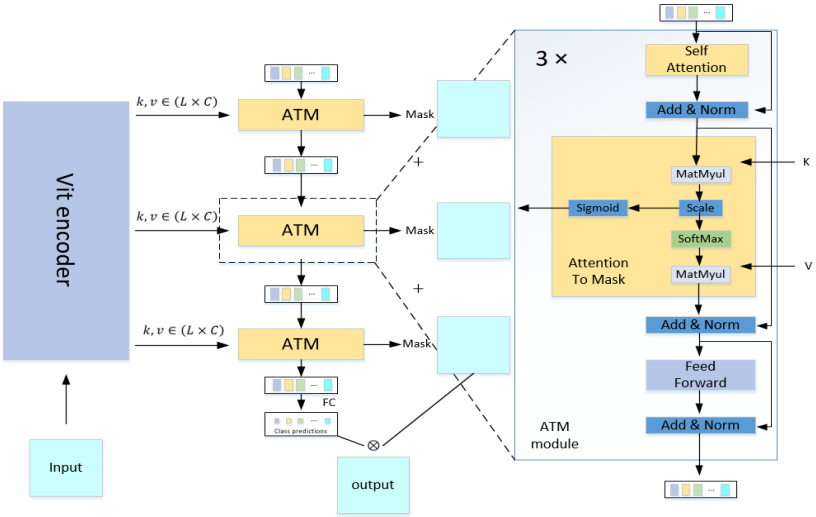

Figure 3: SegVit architecture

3. Experiment and Analysis

3.1. Dataset Introduction

In this experiment, the Potsdam dataset is used for remote sensing image segmentation tasks. This dataset was created by the Remote Sensing Research Group of the University of Potsdam in Germany, containing 24 high-resolution multispectral images from four regions of Potsdam city and corresponding ground truth label images, covering an area of approximately 38.4 square kilometers. Each image has a resolution of 6000x6000 pixels and includes 16 spectral bands, including red, green, blue, and near-infrared bands. The label images are manually annotated and consist of six categories: buildings, low vegetation, trees, roads, and background. The resolution of the label images is consistent with the multispectral images, making them suitable for segmentation tasks. In the experiment, images numbered 6_7, 6_8, 6_9, 7_7, 7_8, and 7_9 are selected as the test set, while the remaining images are used for training. The original images and labels are shown in Figure 4.

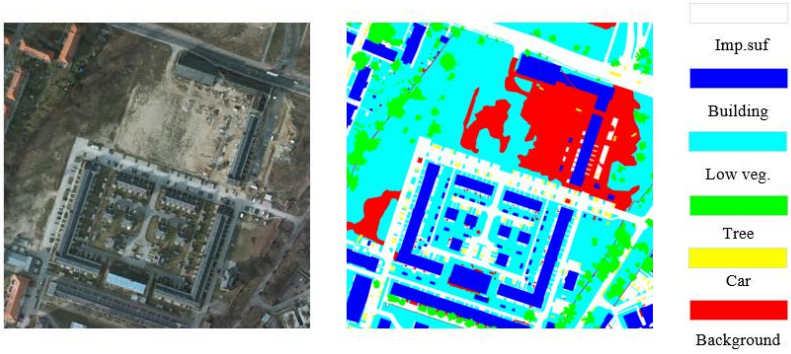

Figure 4: Potsdam DataSet

3.2. Model training

During the training process, we will use batch processing to train the data, dividing the dataset into training and testing sets with an 80:20 ratio. Throughout the training process, we will perform 200 iterations, updating the model parameters and optimizing the loss function at each iteration. This setup allows the model to gradually learn the features and patterns in the dataset, enhancing its generalization ability and performance. Through the above training settings and hyperparameter choices, we can train high-performance models for different network architectures to meet the specific task requirements. This training process will provide us with reliable models and lay the foundation for future research and applications.

3.3. Analysis

In this study, we perform a comparative analysis of seven networks, including Unet, Unet++, SegNet, PspNet, DeeplabV3+, SegFormer, and SegVit, under the same experimental conditions. The experimental results are shown in Table 1.

Table 1: Model Accuracy Table

Model	Low. veg	background.	Tree	buildings	CAR	Road	MIOU	OA
	IOU	IOU	IOU	IOU	IOU	IOU		
Unet	73.49	67.82	73.59	70.37	76.66	70.71	72.10	87.05
Unet++	72.28	69.53	69.84	73.54	76.09	72.09	72.36	88.23
SegNet	70.89	53.22	72.26	88.92	81.24	79.79	74.39	87.17
PspNet	72.24	47.77	70.97	88.59	57.85	76.51	68.99	86.14
DeeplabV3+	76.08	55.72	73.97	91.94	81.26	81.81	76.8	88.53
Segformer	75.86	56.99	74.07	91.85	80.4	82.63	76.97	89.2
SegVit	76.45	57.62	73.56	90.56	82.25	84.52	77.49	89.36

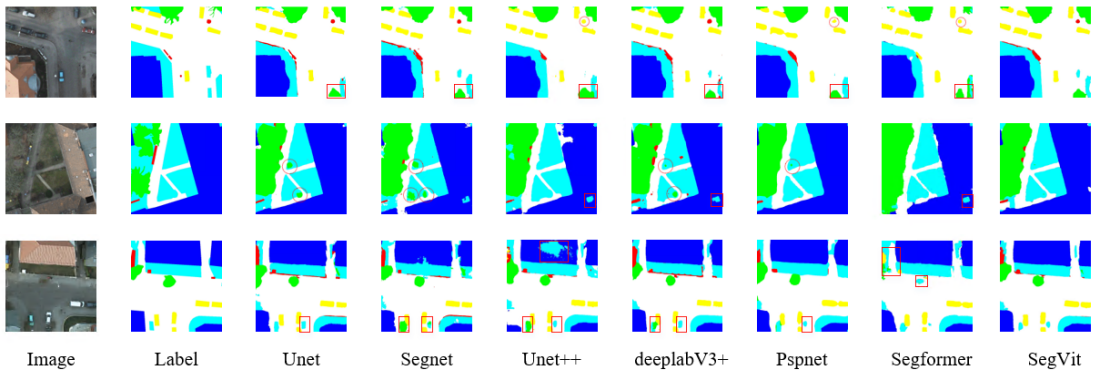

Figure 5: Experimental Results

From Table 1, it can be seen that SegVit achieves the highest overall accuracy among the other networks. It has the highest IoU in the segmentation of low vegetation, vehicles, and roads compared to the other networks. Particularly in small object detection, such as vehicles, SegVit outperforms PspNet by 24.4% in IoU. However, apart from the performance on small object detection, PspNet's segmentation results for other categories are quite similar to those of the other networks. The second-best network is SegFormer, with its IoU and OA being only 0.52% and 0.16% lower than SegVit's, respectively.The specific visual results are shown in Figure 5. In the first image, it can be observed that Unet++, PspNet, and SegVit misclassify the background as vehicles, and all networks misidentify the low vegetation in the lower right corner as trees. This

indicates that these models tend to misclassify objects at isolated edges. In the second image, it is noticeable that for objects with detailed contours, the segmentation results from all models lose fine details. In the third image, where multiple and complex objects appear, the segmentation results from all networks are suboptimal, with instances of segmentation confusion.

4. Summary

In this study, we applied deep learning convolutional networks for semantic segmentation of remote sensing images, aiming to accurately classify pixels into different categories. Through comparative experiments, we evaluated the performance of several mainstream networks and found that SegVit achieved the best segmentation accuracy. However, we also observed that DeepLabV3+ encountered some challenges in handling edge segmentation of small objects. This may be attributed to the network architecture and parameter settings. To further improve the model's performance, we can explore modifications to the network architecture and optimization of parameter settings.In conclusion, although SegVit has shown promising results in remote sensing image semantic segmentation, there is still room for improvement. Through further research and refinement, we can expect better edge segmentation of small objects, thereby enhancing the overall segmentation accuracy.

References

[1] Likas A, Vlassis N, Verbeek J J. The global k-means clustering algorithm[J]. Pattern recognition, 2003, 36(2): 451-461.

[2] Carson C, Belongie S, Greenspan H, et al. Blobworld: Image segmentation using expectation-maximization and its application to image querying [J]. IEEE Transactions on pattern analysis and machine intelligence, 2002, 24(8): 1026-1038.

[3] Shuang L, Sheng-yan D, Le-xiang Q. The decision tree classification and its application research in land cover[J]. Remote sensing technology and application, 2011, 17(1): 6-11.

[4] Song M, Civco D. Road extraction using SVM and image segmentation[J]. Photogrammetric Engineering & Remote Sensing, 2004, 70(12): 1365-1371.

[5] Guindon S, Gascuel O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood [J]. Systematic biology, 2003, 52(5): 696-704.

[6] Jordan M I, Mitchell T M. Machine learning: Trends, perspectives, and prospects[J]. Science, 2015, 349(6245): 255-260.

[7]Gu Xiaotian, Gao Xiaohong, Ma Huijuan, et al. Comparison of machine learning methods for land use/land cover classification in the complicated terrain regions [J]. Remote Sensing Technology and Application, 2019, 34(1): 59-69.

[8] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Advances in neural information processing systems, 2012, 25.

[9] Deng J, Dong W, Socher R, et al. Imagenet: A large-scale hierarchical image database[C]//2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009: 248-255.

[10] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

[11] Wang Y, Tan L, Wang G, et al. Study on the impact of spatial resolution on fractional vegetation cover extraction with single-scene and time-series remote sensing data[J]. Remote Sensing, 2022, 14(17): 4165.

[12] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[13] Zhang B, Tian Z, Tang Q, et al. Segvit: Semantic segmentation with plain vision transformers[J]. Advances in Neural Information Processing Systems, 2022, 35: 4971-4982.