Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230874

Semantic Segmentation of Remote Sensing Images Based on Swin Transformer

Yinghao LIN[a,b] Shihao ZHAO[a], Yuye WANG[a], Yi XIE[a], Baojun QIAO[a,1]
[a]*Henan Key Laboratory of Big Data Analysis and Processing, School of Computer and Information Engineering, Henan University, Kaifeng, China*
[b] *Shenzhen Research Institute of Henan University, Shenzhen, China*

Abstract. With the evolution of Earth observation technology and remote sensing technologies, the amount of data available for high-resolution remote sensing images has exploded, and high-precision image segmentation has become a current research hotspot. Semantic segmentation technology is becoming increasingly important in fields such as urban planning, land use management, and autonomous driving. Large disparities within intraclass and modest differences intraclass are hallmarks of high resolution remote sensing pictures. Traditional image semantic segmentation methods rely on human-computer interaction and have poor generalization ability. When facing remote sensing images with rich types of ground objects and significant differences in target scales, the segmentation accuracy is not high. In this paper, we suggest an UperSwin decoder structure. The decoder includes several Swin transformer blocks and a fusion upsampling module, where the multi head contextual attention module in the Swin transformer block simultaneously uses multi-scale features and upsampling output features from the backbone network. In addition, the fusion upsampling module concatenates the backbone network features with the output features of the Swin transformer block, and then performs upsampling operations, preserving more detailed information. This article evaluates the accuracy and intersection ratio indicators on the Potsdam and Vaihingen datasets, verifying the feasibility and effectiveness of the model.

Keywords. Remote sensing images, semantic segmentation, attention mechanism, Transformer

1. Introduction

The semantic segmentation of remote sensing images is one of the key issues in remote sensing research. Its major objective is to assign relevant category labels to each pixel in the image and categorize each one [1]. Globally, more and more high-resolution remote sensing images are being caputured as a result of the development of both remote sensing and Earth observation technologies. These high-resolution remote sensing images provide a wealth of spatial features as well as possible Semantic information, and have important application value for urban planning [2], road extraction [3], natural disaster monitoring [4] and other fields. However, remote sensing images contain significant differences in the scale of land objects, with the same type of land objects having different shapes and sizes, while different types of land objects may have similar features. Models that rely solely on spectral information cannot effectively distinguish various

[1] Corresponding Author, Baojun QIAO, Henan Key Laboratory of Big Data Analysis and Processing, School of Computer and Information Engineering, Henan University; e-mail: qbj@henu.edu.cn

ground objects, and require the assistance of multi-scale background information to improve the segmentation's accuracy [5]. In semantic segmentation tasks, if only local information is modeled, the classification results are usually fuzzy. Therefore, effectively obtaining Images' global contextual information and enhancing the correlation between features has always been a highly concerned and explored issue.

Convolutional neural networks have been a popular way for addressing semantic segmentation tasks in recent years, thanks to the continual advancement of deep learning technology. Deep learning for image semantic segmentation has been developed and applied owing to the fully convolutional neural network FCN [6]. However, due to the limitation of the Receptive field, more contextual information cannot be effectively collected. In order to deal with the restriction of Receptive field, the pyramid scene analysis network (PSPNet) was suggested by Zhao et al [7]. Which proposed a spatial pyramid module to extract information at different locations through pooling layers of different scales, increasing the Receptive field and enhancing the ability of feature representation. The Atrous Space Pyramid Pooling Module (ASPP) was suggested by Chen et al [8]. It captures multi-scale contextual information by parallelizing atrous convolutions with various expansion rates. UNet [9] adopts an encoder-decoder structure to obtain different levels of feature map information by skipping links, enhancing the expression of feature map information. In addition, the design of self attention mechanism can also be used to make up for the small Receptive field. The Position Attention Module (PAM) and Channel Attention Module (CAM) were added by Fu et al [10]. into the semantic segmentation paradigm, and calculated the position attention map and channel attention map respectively through weight redistribution, capturing remote contextual information and global dependencies. However, during the convolution process, the image resolution will be reduced due to convolution and pooling operations, which inevitably leads to spatial information loss during feature extraction. The attention mechanism is not entirely focused on features, but rather adds CNN fragments based on prior knowledge. And although integrating multi-scale information can help identify targets of different scales, it does not perform well in understanding the relationships between features [11]. Unlike CNN, Transformer utilizes serialization to process images and employs positional embeddings to capture spatial relationships [12], instead of using convolutions or pooling for a more comprehensive spatial information acquisition.

Based on the concept of Swin transformer, this paper designs a decoder structure based on Swin transformer for remote sensing image semantic segmentation. The decoder integrates multi-scale features from the encoder, achieving effective fusion of multi-scale features and improving segmentation accuracy.

2. Related Theories and Techniques

2.1. Semantic Segmentation Method Based on Convolutional Neural Networks

The first CNN structure to comprehensively address semantic segmentation issues is the fully convolutional neural network (FCN). However, due to the simplicity of the decoder, the resolution is low, which limits the improvement of segmentation accuracy. In order to address this issue, UNet uses the encoder decoder structure to extract recovery features. The encoder extracts feature information of different spatial resolutions by gradually Downsampling the image, and gradually recovers spatial resolutions in the decoder stage to learn more context information. Later, the encoder decoder structure also became the

mainstream structure for semantic segmentation of remote sensing images. SegNet [13] records pooling indices during the encoding process and utilizes them to supervise the decoding process, urging the adoption of a more uniform decoding process. Numerous optimization modules are introduced by Refinenet to improve the fusion outcomes of feature maps [14], thereby improving the information capturing capability of the fused feature maps. Zhou et al [15]. designed a unique skip connection to capture more contextual information. Chen et al. proposed that DeepLabv3+ [16] retains the advantages of DeepLabv3 in the encoder phase. The decoder portion enhances the model segmentation performance by the efficient merging of low-level semantic information with high-level semantic information. Although the encoder decoder structure has achieved good results in the field of image segmentation, CNN can only focus on local features, providing insufficient global correlation information between local pixels, and lacking the ability to model the global relationships between target objects in the image. In addition, the limited Receptive field can not provide enough contextual feature information, which has a great impact on the segmentation accuracy.

2.2. Semantic Segmentation Method Based on Transformer

Unlike traditional CNN structures, Transformer converts two-dimensional images into one-dimensional sequences for computation. Due to its sequence to sequence modeling ability, Transformer exhibits excellent feature extraction in capturing global context compared to the aforementioned models. Therefore, it has achieved the most advanced results in basic visual tasks such as object detection and semantic segmentation. VIT [17] is the first pure Transformer structure model applied in the realm of computer vision. As a result, numerous remote sensing researchers have implemented Transformer to remote sensing image sceneries. Xie et al. [18] introduced SegFormer, a semantic segmentation framework that is simple, efficient, and powerful. SegFormer employs a hierarchical feature representation approach by combining Transformer with lightweight multilayer perceptron (MLP) modules. The Swin transformer proposed by Liu [19] et al. introduces moving windows to perform attention computation within non overlapping local windows, while allowing cross window computation, which compensates for the lack of connectivity between windows caused by window partitioning in VIT and improves image segmentation accuracy.

3. The Proposed Method

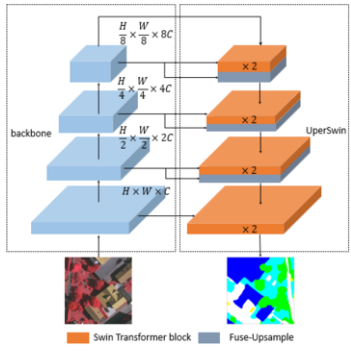

Figure 1. Overall network architecture

The UperSwin decoder consists of several Swin Transformer blocks and Fuse upsample, as shown in Figure 1, In this experiment, we used two different backbone networks, ResNet50 and Swin-T, to verify the performance of the module. The specific components will be described in the following text.

3.1. Multi Head Contextual Attention Module

We utilized the idea of the second multi head attention mechanism in the Decoder of the Transformer model to improve the W-MSAand the SW-MSA in the Swin Transformer. The improved Swin transformer block is shown in Figure 2. This module uses the output M from the previous stage as the Key-Value, and uses the multi-scale feature F from the backbone network as the Query. At this point, the self attention calculation can be represented by Eq. (1).

$$Attn(Q_F, K_M, V_M) = softmax\left(\frac{Q_F(K_M)^T}{\sqrt{d}} + B\right)V_M \tag{1}$$

Among them, Q_F as a query calculated by W-MCA, is a linear transformation of feature F, and K_M, V_M as the key value pair of the multihead contextual attention mechanism, is a linear transformation of feature M, where d is the vector dimension of Q_F、K_M and B is the relative position offset.Similarly, the output obtained from W-MCA after MLP calculation is used as a Query for SW-MCA self attention calculation, and the output M from the previous stage is used as a Key-Value. This calculation can be represented by Eq. (2).

$$Attn(Q_{out}, K_M, V_M) = softmax\left(\frac{Q_{out}(K_M)^T}{\sqrt{d}} + B\right)V_M \tag{2}$$

Q_{W-MCA} represents the result of W-MCA after MLP calculation, K_M,V_M as a key value pair of multi head contextual attention mechanism, is a linear transformation of M.

3.2. Fuse-Unsample Module

The Fuse-Unsample module supplements the upsampling process with additional feature maps to increase the amount of output information and enhance detail information. By fusing multi-scale feature information from the backbone network, the spatial details lost during the downsampling process can be effectively restored. The structure of the fusion upsampling module is shown in Figure 3.

The specific process is as follows: first, concatenate the output feature maps of the multi head context attention module with the feature maps of the backbone network with equal resolution. At this point, the number of channels increases from C to 2C, and then use a size of 1 × 1. The convolutional kernel with a step size of 1 reduces the feature channel to 1/4 of its original size, while establishing a connection between the two feature maps. Then, the bilinear interpolation method is used to enlarge the feature map with the resolution of $H \times W$ to $2H \times 2W$, and finally the feature map with the size of $2H \times 2W \times \frac{C}{2}$ is obtained.

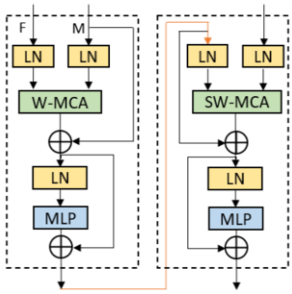

Figure 2. Improved Swin transformer block

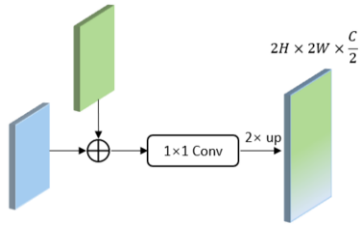

Figure 3. Fuse-Unsample module

4. Experiment

4.1. Datasets and Evaluation Metrics

Potsdam dataset contains 38 sheets of 6000×6000 pixel orthophoto image, each covering approximately 3.42 km^2 of ground area, with a spatial resolution of 5cm. This dataset consists of six categories: buildings, trees, cars, low vegetation, impermeable surfaces, and backgrounds. We selected 24 out of 38 images as the network training set and 14 images as the test set. Vaihingen is a smaller village. This dataset contains 33 remote sensing images of different sizes, each with a size of approximately 2500 × 2000 pixels, with a spatial resolution of 9cm, each remote sensing image covers approximately 1.38km^2 of ground area. This dataset, like the Potsdam dataset, is divided into 6 categories. We selected 16 images as the training set and 17 images as the test set. This paper uses mean pixel accuracy (mPA), and mean intersection over union (MIoU) to evaluate experimental results. The calculation of the four evaluation indicators can be based on the confusion matrix, and the formula is defined as follows. The specific meanings of its symbols are as follows: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).

$$mPA = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP+TN}{TP+TN+FP+FN} \tag{3}$$

$$MIoU = \frac{1}{k+1}\sum_{i=0}^{k}\frac{TP}{FN+FP+TP} \tag{4}$$

4.2. Implementation Details

The GPU we use is NVIDIA GeForce GTX3060, with CUDA version 11.1. The deep learning framework is the open source Pytorch framework, and the Python version is 3.8. The training of all datasets in this article adopts a multiple learning rate decay strategy, with an energy factor of 1 and an initial learning rate of 1e-6. The learning rate is multiplied by left $\left(1-\frac{iter}{total\_iter}\right)$ for each round. We use the AdamW algorithm for the

optimizer, with betas set to 0.9 and 0.999, respectively. The optimizer momentum is set to 0.9, and the initial learning rate is set to 1e-4, Batch_ Set the size to 2.

4.3. Result Display

The comparative experimental results on two datasets are shown in Table 1 Specifically, our proposed Swin-T+UperSwin method has a MIoU value of 79.06% and mPA value of 86.24% on the Potsdam dataset, and a MIou value of 74.25% and mPA value of 81.69% on the Vaihingen dataset. The results are superior to most resnet based methods. Thanks to the multi head contextual attention mechanism, our model is able to fully utilize multi-scale information while taking into account pixel connections between objects. Compared with popular networks such as DeepLabv3+and PSPNet, the accuracy has been improved. Figure.4 shows the visualization results of each model, it can be seen that Swin-T+UperSwin provides clearer segmentation of target edges and more accurate recognition of low vegetation and trees, thanks to global contextual relationships and multi-scale structures.

Table 1. Experimental results on the Potsdam and Vaihingen datasets

dataset	model	backbone	MIoU	mPA
Potsdam	FCN	ResNet50	76.19	84.15
	UNet	-	73.76	81.67
	DeepLabv3+	ResNet50	77.33	84.54
	PSPNet	ResNet50	77.58	84.57
	DANet	ResNet50	77.10	84.30
	UperNet	ResNet50	77.56	84.89
	UperSwin	ResNet50	77.10	84.49
	UperNet	Swin-T	78.71	85.70
	UperSwin	Swin-T	**79.06**	**86.24**
Vaihingen	FCN	ResNet50	71.56	78.45
	UNet	-	66.85	76.54
	DeepLabv3+	ResNet50	72.97	80.81
	PSPNet	ResNet50	72.53	79.61
	DANet	ResNet50	71.25	79.03
	UperNet	ResNet50	72.76	79.68
	UperSwin	ResNet50	73.57	81.20
	UperNet	Swin-T	73.11	80.28
	UperSwin	Swin-T	**74.25**	**81.69**

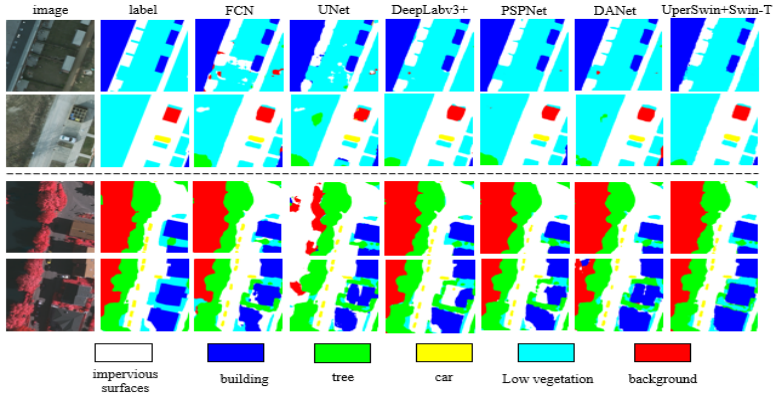

Figure 4. Visualization results on the Potsdam and Vaihingen datasets

5. Conclusion

Remote sensing image features are rich in variety. In order to enhance the global Semantic information and strengthen the semantic association between features, this paper designs a remote sensing image semantic segmentation model based on Swin transformer (UperSwin). This article guarantees the superiority of the model from two aspects. Firstly, attention calculation is performed using features from different scales of the backbone network and fused upsampled features. The unique contextual attention mechanism can effectively capture the correlation between features. Secondly, during the upsampling process, the backbone network features were also integrated to enhance detail information and establish global semantic associations. While achieving excellent segmentation results, the number of model parameters is also relatively large. If the model is better simplified, further research is needed.

References

[1] Tapasvi B, Udaya Kumar N, Gnanamanoharan E. A Survey on Semantic Segmentation using Deep Learning Techniques[J]. Int. J. Eng. Res. Technol, 2021, 9: 50-56.
[2] Shi Y, Qi Z, Liu X, et al. Urban land use and land cover classification using multisource remote sensing images and social media data[J]. Remote Sensing, 2019, 11(22): 2719.
[3] Kestur R, Farooq S, Abdal R, et al. UFCN: A fully convolutional neural network for road extraction in RGB imagery acquired by remote sensing from an unmanned aerial vehicle[J]. Journal of Applied Remote Sensing, 2018, 12(1): 016020-016020.
[4] Goswami S, Chakraborty S, Ghosh S, et al. A review on application of data mining techniques to combat natural disasters[J]. Ain Shams Engineering Journal, 2018, 9(3): 365-378.
[5] Guo Y, Liao J, Shen G. A deep learning model with capsules embedded for high-resolution image classification[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 214-223.
[6] 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.
[7] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network[C].Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
[8] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[J]. arXiv preprint arXiv:1706.05587, 2017.
[9] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.
[10] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.
[11] Marcos D, Volpi M, Kellenberger B, et al. Land cover mapping at very high resolution with rotation equivariant CNNs: Towards small yet accurate models[J]. ISPRS journal of photogrammetry and remote sensing, 2018, 145: 96-107.
[12] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.
[13] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
[14] Lin G, Milan A, Shen C, et al. Refinenet: Multi-path refinement networks for high-resolution semantic segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 1925-1934.
[15] Zhou Z, Rahman Siddiquee M M, Tajbakhsh N, et al. Unet++: A nested u-net architecture for medical image segmentation[C]//Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: 4th International Workshop, DLMIA 2018, and 8th International Workshop, ML-CDS 2018, Held in Conjunction with MICCAI 2018, Granada, Spain, September 20, 2018, Proceedings 4. Springer International Publishing, 2018: 3-11.

[16] 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.
[17] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
[18] Xie E, Wang W, Yu Z, et al. SegFormer: Simple and efficient design for semantic segmentation with transformers[J]. Advances in Neural Information Processing Systems, 2021, 34: 12077-12090.
[19] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]. Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.