Advances in Computer, Signals and Systems (2024)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2024.080209
ISSN 2371-8838 Vol. 8 Num. 2

Study on Vegetation Extraction from Riparian Zone Images Based on Cswin Transformer

Yuanjie Ma[1,a], Yaping Zhang[1,b,*]

[1]School of Information Science and Technology, Yunnan Normal University, Kunming, Yunnan, 650500, China
[a]1945984520@qq.com, [b]zhangyp@ynnu.edu.cn
[*]Corresponding author

Keywords: Vegetation extraction; deep learning; semantic segmentation; Transformer

Abstract: In the field of ecological conservation, accurately extracting vegetation areas in UAV images is a critical task. This study aims to accurately identify vegetation from high-resolution riverine zone UAV images. Facing the challenges of complex factors such as light variations and water ripples, a deep learning technique, which combines Convolutional Neural Networks and Vision Transformer, is used in this study, which proposes a semantic segmentation network structure based on an encoder-decoder. We innovatively introduce the Explicit Visual Center mechanism (EVC) and CSWin Transformer structure to optimize image feature capture, especially in dealing with the classification challenges caused by the similarity between vegetation and water ripples. The experimental results show that the proposed network has the best results compared with the classical network models such as U-Net, PSP-Net, DeepLabv3+, etc., and the mIOU phase of U-Net, which is the highest among the three networks, is 1.3 percentage points higher. In this paper, an effective scheme is proposed for vegetation extraction from UAV images in the riparian zone.

1. Introduction

The vegetation coverage condition of the riparian zone is a key indicator for assessing the health of ecosystems, and accurately extracting vegetation information from drone aerial images is crucial for ecological protection. However, the interlaced water and land environment, along with factors such as light refraction, make it difficult for traditional methods to distinguish between vegetation and water bodies. Therefore, exploring an effective method for extracting vegetation from riparian zone images is of great importance. Deep learning has shown significant effectiveness in the field of image semantic segmentation, especially in extracting vegetation information from complex backgrounds. This is due to its excellent ability to learn features, allowing deep learning models to adapt to different interference factors, effectively recognizing and segmenting vegetation.

CSWin Transformer [1], designed for visual tasks, is an advancement over the SWin Transformer[2]. It addresses the computational efficiency issues of vision Transformer[3] in processing high-resolution images. By introducing an innovative Cross-Shaped Window self-attention mechanism, it reduces computational costs while expanding the model's receptive field, effectively enhancing the processing capability for pixel-level tasks.

Compared to Swin Transformer, Patch Merging is replaced with a convolution of 3 × 3 steps of 2. This strategy is similar to the design of downsampling in traditional CNN architectures, which further improves the efficiency and performance of the model.The design of the CSWinBlock allows the model to not only emphasize the local location information when dealing with features within a small window, but also to flexibly adapt to inputs of different sizes and shapes. This layered design strategy enables CSWin Transformer to efficiently extract features from shallow to deep layers and provide rich semantic information for a variety of visual tasks such as image classification, target detection and semantic segmentation.

This study aims to explore a network structure suitable for extracting vegetation from high-resolution riparian zone drone images using deep learning semantic segmentation network methods. The goal is to address the issue of difficult-to-recognize water ripples in riparian zone images, thereby improving the accuracy of vegetation segmentation.

2. Materials & Methods

In the CSWin Transformer, images are first downsampled through a convolutional layer and mapped to a higher dimensional space, similar to the initial processing method of the Swin Transformer. Subsequently, the model processes through several stages, each including multiple CSWin Transformer Blocks. These stages utilize Local Enhanced Positional Encoding (LePE) to highlight the importance of local positional information, while Multilayer Perceptrons (MLP) [4]are used for feature extraction and transformation.The encoder part consists of four stages, each of which does extraction and refinement of features. In the initial stage, the input image is downsampled through a convolutional layer and then transformed into a series of non-overlapping image blocks (patches) which are further mapped to a high-dimensional feature space through a linear layer.

Subsequently, the feature map enters the core of the CSWin Transformer, which consists of multiple CSWin Transformer blocks, each of which contains the Cross Shape Window's self-attention (CSWin self-attention) mechanism. In this mechanism, self-attention is computed within a window of bars in the vertical and horizontal directions, respectively, which not only reduces the complexity of self-attention, but is allowed to capture details and contextual information in both directions. As the network hierarchy deepens, these bar windows are progressively merged and narrowed, allowing information to flow over a larger area, resulting in a hierarchical feature representation.Inside the CSWin Transformer block, the detailed features are further processed and refined to improve the sensitivity to the location information without increasing the computational cost through the Local Enhancement of Position Awareness (LePE) attentional mechanism. In addition, a Merge Block is used at the end of each stage, which downsamples and dimensionally increases the feature map through a convolution operation to prepare for the next stage of feature extraction.

The entire encoder structure is able to gradually construct feature representations from coarse to fine by the above staged processing, which provides rich semantic information for the subsequent decoder.CSWin Transformer, while maintaining high efficiency, improves the model's ability to adapt to features at different scales through its cross shape window design, which is crucial for accurate pixel-level prediction tasks.

Facing the challenges of complex environmental factors in high-resolution drone images of riparian zones, especially the issue of misclassification caused by the similarity between water ripples and vegetation textures, this network design optimizes through the collaborative work of the EVC[5] module and the decoder. The global and local feature fusion capability of the EVC module helps the model capture more detailed vegetation features. Meanwhile, the decoder ensures the flow of information from the encoder to the decoder through skip connections and gradual upsampling, preserving key vegetation information and effectively reconstructing the spatial distribution of vegetation. In summary, the lightweight MLP structure achieves effective capture of global information while maintaining high efficiency by combining the properties of deep convolution and MLP, while LVC effectively encodes and processes the input features to provide richer information and context for subsequent model processing and inference.

This paper presents a network structure based on CSWin Transformer and EVC, aimed at achieving high-precision vegetation segmentation. The CSWin Transformer structure optimizes the capture of vegetation image features by combining an encoder-decoder structure with self-attention mechanisms, while the EVC module aggregates global information and local details to enhance the model's ability to recognize complex vegetation areas, making it more accurate in processing high-resolution riparian zone images. The decoder part draws on the architecture of UNet[6] to achieve gradual upsampling of feature maps, starting with bilinear upsampling to enlarge the feature maps, which are then concatenated with feature maps output from skip connections of the encoder. Subsequently, the concatenated feature maps are processed through two convolutional layers. This process of upsampling and feature fusion is carried out step by step, compensating for the loss of vegetation detail and edge features caused by downsampling and enhancing feature richness. Finally, the network uses a convolutional layer to convert the final feature map into prediction results. The output channels of this convolutional layer equal the number of target classes for classification, thereby achieving classification prediction for each pixel. The network structure is shown in Figure 1.

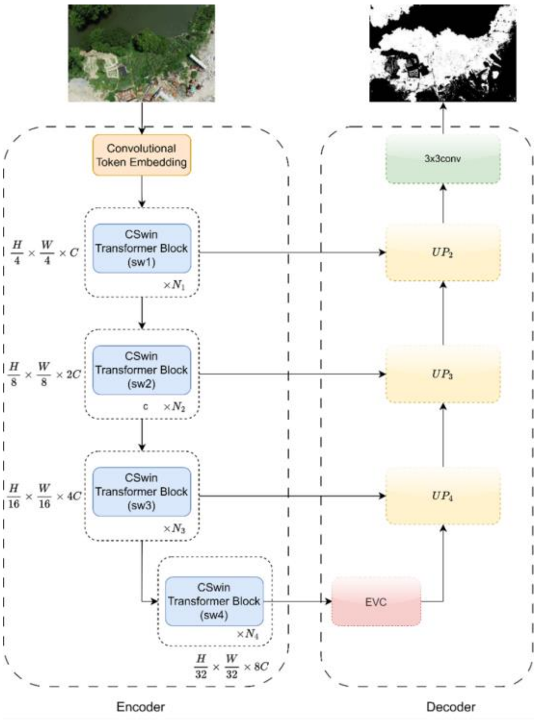
Convolutional Token Embedding
$\frac{H}{4}\times\frac{W}{4}\times C$
CSwin Transformer Block (sw1)
$\times N_1$
$\frac{H}{8}\times\frac{W}{8}\times 2C$
CSwin Transformer Block (sw2)
$\times N_2$
$\frac{H}{16}\times\frac{W}{16}\times 4C$
CSwin Transformer Block (sw3)
$\times N_3$
CSwin Transformer Block (sw4)
$\times N_4$
$\frac{H}{32}\times\frac{W}{32}\times 8C$
3x3conv
UP_2
UP_3
UP_4
EVC
Encoder
Decoder

Figure 1: Network infrastructure

3. Dataset & Evaluation metrics

The publicly available datasets show significant differences in environment and resolution compared to the drone aerial visible light images of riparian zones used in this study. Therefore, for the collected riparian drone images, vegetation distribution images for training labels were obtained using the watershed algorithm [7] combined with the vegetation index[8] extraction method. However, this approach does not capture labels for images with high-intensity water ripples. Therefore, the Labelme tool was used to precisely annotate the water bodies in these images, generating binary images of water and non-water as supplementary labels. After organization, we obtained a total of 1330 data samples, which were randomly divided into a training dataset and a validation dataset at a ratio of 8:2.

The performance evaluation metrics used in the experiments include mean intersection over union (mIoU), mean pixel accuracy (mPA), and overall accuracy (Accuracy). mIoU, as a standard evaluation metric in the field of semantic segmentation, is selected as the main evaluation metric in this paper.

4. Results

To demonstrate the effectiveness of the network proposed in this chapter, a comparison of riparian zone vegetation extraction was conducted between several existing semantic segmentation networks (including DeepLabV3+, PSPNet, UNet) and our network. The specific results are shown in Table 1.

According to Table 1, several different network models worked well in the extraction of the vegetation dataset we constructed for the riparian zone, with accuracies above 0.9. However, in general, our network model outperformed several other networks in terms of mIoU, mPA and Accuracy.

Table 1: Comparison with other network models

network	# params	mIoU /%	mPA /%	Accuracy /%
Ours(CSwin-T)	31M	97.33	98.65	98.64
DeepLabV3+	59M	95.47	97.69	97.69
UNet	25M	96.08	98.01	98.00
PSPNet	24M	94.61	97.24	97.23

As can be seen in Table 1, the two metrics mPA and Accuracy are very close to each other, which is mainly attributed to the dataset characteristics. In our vegetation extraction task, the images in the dataset were cropped from the high-resolution vegetation extraction result images to a size of 512 × 512, resulting in a very uniform spatial distribution of most of the samples, i.e., either exclusively vegetated pixels or exclusively non-vegetated pixels. This consistency in spatial distribution, coupled with the roughly balanced distribution of vegetation and non-vegetation samples across the entire dataset, motivates the model to exhibit consistent predictive performance for both categories. As a result, the values of mPA and Accuracy even appear to be identical after retaining to two decimal places. This also reflects the model's ability to predict both vegetation and non-vegetation categories uniformly and effectively when dealing with the relatively simple task of binary classification of vegetation images.

To validate the effectiveness of the CSWin Transformer as the encoder backbone network, we conducted ablation experiments, comparing it with structures such as Resnet50[9], Biformer_tiny[10], Davit_tiny[11], and Efficientformer[12]. Additionally, the CSWin Transformer is divided into different variants according to scale parameters, with CSwin-T, CSwin-S, CSwin-B, and CSwin-L representing the tiny, small, standard, and large versions, respectively. Ablation experiments were also performed for these four variants, with results shown in Table 2.

After comparison, we find that the evaluation index of the model improves accordingly as the number of parameters increases from CSwin-T to CSwin-B variants. However, despite having the largest number of parameters, the CSwin-L variant does not optimize its performance, but rather performs worse than CSwin-T. This may be caused by overfitting, i.e., a large model that overperforms on the training set and underperforms on the validation set, especially more pronounced when the amount of data is small. In addition, if the task or data complexity does not require a large model, the performance improvement of the larger variant is limited. However, the evaluation metrics of the CSwin structure have significant advantages over other non-CSwin encoder structures.

Table 2: Comparison of evaluation metrics for encoder backbone networks

network	#params	mIoU /%	mPA /%	Accuracy /%
Resnet50	44M	97.07	98.52	98.50
Biformer_tiny	22M	97.13	98.55	98.53
Davit_tiny	39M	96.69	98.32	98.31
Efficientformer	19M	96.55	98.26	98.24
Ours(CSwin-T)	31M	97.33	98.65	98.64
Ours(CSwin-S)	43M	97.42	98.70	98.70
Ours(CSwin-B)	88M	97.48	98.73	98.72
Ours(CSwin-L)	186M	97.14	98.55	98.54

5. Conclusion

To address the difficulty of distinguishing between vegetation and water bodies in drone aerial images of riparian zones, we constructed a semantic segmentation neural network based on an encoder-decoder structure. In the encoder, the CSWin Transformer structure was selected as the main feature extraction network, and the EVC module was introduced in the decoder for feature fusion. Additionally, we conducted model testing and ablation experiments to verify that our method achieves reliable accuracy in extracting vegetation from riparian zones.

Future research could focus on introducing and developing more advanced image preprocessing and enhancement techniques, especially when dealing with complex environmental factors such as light variations, shadows and reflections. This will help to improve the performance of the algorithm for vegetation extraction under various environmental conditions. Consider combining different types of remote sensing data, such as hyperspectral, infrared, and radar data, to enhance the generalization capability of the model. Multi-source data fusion can provide more comprehensive information and help improve the accuracy of vegetation classification and identification. Through in-depth research and exploration in these directions, future work will not only improve the accuracy and efficiency of vegetation distribution detection in the riparian zone, but also further expand the prospects for its application in a wider range of fields such as ecological monitoring and environmental protection.

References

[1] Dong X, Bao J, Chen D, et al. Cswin transformer: A general vision transformer backbone with cross-shaped windows[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 12124-12134.

[2] 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.

[3] Arnab A, Dehghani M, Heigold G, et al. Vivit: A video vision transformer[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 6836-6846.

[4] Tolstikhin I O, Houlsby N, Kolesnikov A, et al. Mlp-mixer: An all-mlp architecture for vision [J]. Advances in neural information processing systems, 2021, 34: 24261-24272.

[5] Quan Y, Zhang D, Zhang L, et al. Centralized feature pyramid for object detection[J]. IEEE Transactions on Image

Processing, 2023, 32:4341-4354

[6] 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.

[7] Bieniek A, Moga A. An efficient watershed algorithm based on connected components [J]. Pattern recognition, 2000, 33(6): 907-916.

[8] Xiaoqin W, Miaomiao W, Shaoqiang W, et al. Extraction of vegetation information from visible unmanned aerial vehicle images[J]. Nongye Gongcheng Xuebao/Transactions of the Chinese Society of Agricultural Engineering, 2015, 31(5): 152-159.

[9] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[10] Zhu L, Wang X, Ke Z, et al. Biformer: Vision transformer with bi-level routing attention[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 10323-10333.

[11] Ding M, Xiao B, Codella N, et al. Davit: Dual attention vision transformers[C]//European conference on computer vision. Cham: Springer Nature Switzerland, 2022: 74-92.

[12] Li Y, Yuan G, Wen Y, et al. Efficientformer: Vision transformers at mobilenet speed[J]. Advances in Neural Information Processing Systems, 2022, 35: 12934-12949.