remote sensing

MDPI

Technical Note

FA-HRNet: A New Fusion Attention Approach for Vegetation Semantic Segmentation and Analysis

Bingnan He, Dongyang Wu, Li Wang and Sheng Xu *

College of Information Science and Technology and Artificial Intelligence, Nanjing Forestry University, Nanjing 210037, China
* Correspondence: xusheng@njfu.edu.cn

Abstract: Semantic segmentation of vegetation in aerial remote sensing images is a critical aspect of vegetation mapping. Accurate vegetation segmentation effectively informs real-world production and construction activities. However, the presence of species heterogeneity, seasonal variations, and feature disparities within remote sensing images poses significant challenges for vision tasks. Traditional machine learning-based methods often struggle to capture deep-level features for the segmentation. This work proposes a novel deep learning network named FA-HRNet that leverages the fusion of attention mechanism and a multi-branch network structure for vegetation detection and segmentation. Quantitative analysis from multiple datasets reveals that our method outperforms existing approaches, with improvements in MIoU and PA by 2.17% and 4.85%, respectively, compared with the baseline network. Our approach exhibits significant advantages over the other methods regarding cross-region and cross-scale capabilities, providing a reliable vegetation coverage ratio for ecological analysis.

Keywords: computer vision; aerial remote sensing images; vision models; semantic segmentation; vegetation coverage; panoptic segmentation

check for updates

Citation: He, B.; Wu, D.; Wang, L.; Xu, S. FA-HRNet: A New Fusion Attention Approach for Vegetation Semantic Segmentation and Analysis. *Remote Sens.* **2024**, *16*, 4194. https://doi.org/10.3390/rs16224194

Academic Editor: Pedro Melo-Pinto

Received: 28 September 2024
Revised: 29 October 2024
Accepted: 5 November 2024
Published: 11 November 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Vegetation encompasses all plant communities within a specific area, indicating significant features of the Earth's surface. It serves as an irreplaceable material resource and is vital for human survival [1]. With the modernization of forestry practices, remote sensing technology has been extensively applied in various domains related to vegetation analysis, including forest resource monitoring, forest fire management [2], biodiversity conservation, forest resource assessment [3], disaster evaluation, and recovery efforts [4].

In the field of vegetation segmentation from aerial remote sensing images, the primary methodologies encompass NDVI (Normalized Difference Vegetation Index) [5], image processing [6], and machine learning approaches [7]. Given the species heterogeneity, seasonal fluctuations, and various characteristics, plants exhibit unique traits and appearances (e.g., forests, grasslands, shrubs). These variations result in differing texture and color features, thereby presenting significant challenges for segmentation tasks. Traditional spectral-based methods are constrained by changes in illumination conditions, which neglect spatial and textural information [8]. Classical machine-learning techniques depend heavily on manual feature extraction. These methods struggle to capture multi-scale information and encounter issues such as over-segmentation and high computation complexity [9]. Nowadays, deep learning methodologies work well in processing high-dimensional data and succeed in capturing intricate patterns. Consequently, this paper presents a semantic segmentation algorithm based on deep learning frameworks, aiming to improve the effectiveness of vegetation detection and semantic segmentation.

In semantic segmentation, three challenges arise: first, aerial remote sensing images are usually collected in extremely high resolution. The existing segmentation network model exhibits a shallow architecture, which limits its ability to extract rich detailed features and spatial characteristics. Consequently, false detection and omission of the target appear. Second, remote sensing images encompass various vegetation types characterized by distinct textures and colors. This diversity can confuse the foreground from the background, as their features may exhibit significant similarities. Finally, images acquired from different regions suffer from substantial differences. For instance, the presence of shadows and the occlusion caused by buildings can significantly impact vegetation segmentation accuracy, requiring the model to address cross-domain issues.

To tackle the above-mentioned challenges, this paper designs an enhanced HRNet network model [10], which incorporates the fusion of attention mechanisms termed FA-HRNet. First, we propose an improved adaptive spatial attention module (ASA) that enables the model to effectively extract and utilize feature maps across different scales. This enhancement facilitates more efficient recognition and emphasis on key features within vegetation. Secondly, we have refined the original upsampling module of HRNet by introducing a multi-level upsampling module. This approach allows for a progressive merging of adjacent branches, thereby minimizing information loss in the feature maps. Finally, after the feature fusion stage, we incorporate a residual channel attention module (RCA) to further extract channel information from the feature maps.

The modules as mentioned above make FA-HRNet focus on critical areas such as vegetation while also improving its ability to handle complex scenarios involving occlusions and shadows. These modules empower the network to extract more comprehensive features and achieve fine segmentation results. Our main contributions are summarized as follows:

1. An improved Spatial Attention Module and Channel Attention Module are developed and fused for comprehensive extraction of vegetation features from remote sensing images.
2. An enhanced multi-level upsampling module is proposed, which is better suited for segmenting and achieving semantic information from high-resolution vegetation images.
3. In high-resolution remote sensing imagery, the method achieves an accuracy of 73.81%, surpassing the original model by 2.17%, in terms of vegetation segmentation.

2. Related Works

2.1. Unsupervised Learning-Based Segmentation Methods

Unsupervised learning-based methods analyze input data without any label information [11,12]. Researchers do not need to pre-label remote sensing images or apply supervised training processing [13]. The advantages of unsupervised learning include the ability to perform segmentation in the absence of labeled data and the fact that it does not require extensive prior knowledge. Those methods minimize human errors during the segmentation process [14].

Commonly employed unsupervised learning techniques include K-means [15], Principal Component Analysis (PCA) [16], and Gaussian Mixture Model (GMM) [17]. For example, the K-means has been successfully applied in remote sensing images for a long time [18]. Zhiyong Lv et al. [19] proposed an enhanced K-means-AMV, which incorporates contextual information to construct adaptive regions around pixels. By leveraging context information, this algorithm effectively determines labels for each pixel.

Despite their adaptability and capacity for large-scale processing, unsupervised learning methods are limited by the lack of class definitions. Therefore, results accuracy depends on the parameter selection heavily.

2.2. Traditional Supervised Learning-Based Segmentation Methods

Supervised learning methods train models using a substantial amount of labeled information [20]. The model learns from the training set and is subsequently tested on the validation set to optimize their performance [21]. These models effectively learn the intricate characteristics of vegetation in remote sensing images, thereby achieving high segmentation accuracy and robustness.

Traditional supervised learning techniques include Support Vector Machine (SVM) [22], decision trees [23], and Random Forest (RF) [24]. For example, the Random Forest (RF) classifier is an ensemble method that employs randomly selected training samples to generate multiple decision trees [25]. RF has gained widespread application in remote sensing image segmentation [26]. Recently, researchers have constructed an optimal feature space by comprehensively considering factors such as spectral information, texture, and terrain and established an object-oriented random forest approach to complete vegetation extraction [27].

Traditional machine learning methods rely on hand-crafted feature extraction and selection. Therefore, they may not fully capture deep features or complex relationships present in sensing images characterized by intricate landforms.

2.3. Deep-Learning-Based Segmentation Methods

In recent years, the rapid advancement of deep learning has led to its widespread application across various fields [28–30]. Key research areas include the forest resource monitoring [31], forest fire detection [32], crop classification [33], and vegetation classification [34]. The segmentation techniques employed to detect vegetation can be broadly categorized into decoding networks methods (such as U-Net and SegNet) [35], spatial pyramid networks method (such as DeepLab and PSPNet) [36], and multi-branch network architectures method (such as HRNet and BiSeNet) [37]. For example, Mahendra et al. [38] conducted convolutional neural network (CNN) models to precisely calculate urban vegetation coverage. Liu et al. [39] utilized DeepLabV3+ for the extraction and classification of wetland vegetation, determining the optimal spatial resolution for images.

The learning network significantly influences the final segmentation results when high-resolution remote sensing images are used [40]. Decoding networks and spatial pyramid networks often reduce spatial resolution through sampling and pooling operations. They exhibit certain limitations regarding precision in image segmentation [41]. In contrast, multi-branch network structures process features at varying scales and levels of abstraction via parallel pathways. Those approaches compensate for the detail loss by integrating high-resolution and low-resolution features. Meanwhile, the original multi-branch network structure models, such as HRNet, still exhibit several shortcomings. For instance, when processing low-resolution feature maps, issues related to insufficient feature extraction may arise. To address these challenges, Che et al. [42] improved the existing multi-branch architecture by proposing a pyramid feature attention module aimed at achieving precise segmentation of buildings. Similarly, Li et al. [43] opted for an attention mechanism module that incorporates compression and excitation mechanisms to tackle this issue.

This paper aims to focus on the above-mentioned problems encountered in vegetation extraction by integrating various attention mechanisms and enhancing the original HRNet architecture, resulting in achieving accurate vegetation segmentation.

3. The Method

3.1. Subject Network

The network architecture of the model presented is illustrated in Figure 1. The FA-HRNet network represents an enhanced HRNet architecture, incorporating a variety of attention mechanisms and advanced upsampling modules. This model effectively capitalizes on the strengths of HRNet's multi-branch structure to facilitate detailed feature extraction from input drone remote sensing images. The overall architecture consists of four stages, which produces feature maps with varying resolutions. First, stage 1 processes the original image to generate an initial high-resolution feature map. In stage 2, multiple branches are introduced, including both high-resolution and low-resolution branches. These branches are fused with the low-resolution branch features through upsampling techniques. In stage 3, additional branches with varying resolutions are incorporated, allowing for further fusion of features from each branch. Finally, in stage 4, the features from different branches undergo enhancement via an adaptive spatial attention module.

The resulting features are then paired and fused using multi-level upsampling modules to produce a smooth and refined high-resolution feature map. After processing through these four stages, the features are enhanced by a residual channel attention module. This approach improves the generalization capability for new data while mitigating the risk of overfitting.

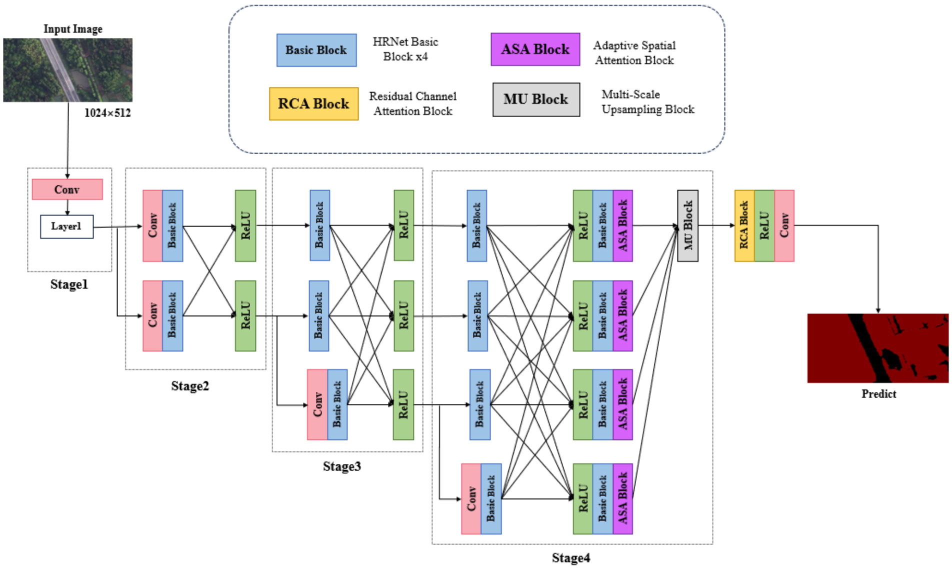


Figure 1. The structure of the proposed FA-HRNet.

3.2. ASA Module

The existing HRNet model is renowned for its high-resolution feature extraction capabilities, which enable it to comprehensively capture the details and semantic information across various scales [10]. However, the fusion of feature maps from different scales may lead to a loss of important features. To address this issue, we enhance the spatial attention mechanism and integrates it into the feature fusion stage [44,45]. This improvement allows for more effective integration of feature maps at varying scales, thereby enabling the model to more accurately delineate object boundaries, texture details, and other complex semantic information. Consequently, this enhancement improves both the accuracy and precision of semantic segmentation. The structure of the ASA mechanism is illustrated in Figure 2.

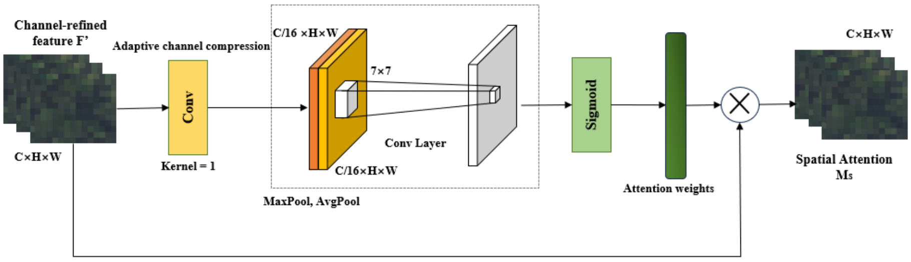


Figure 2. Module structure of ASA mechanism.

Firstly, the obtained feature map is processed through a compressed convolution layer. Maintaining the original length and width of the feature map, then an adaptive channel compression convolution module is employed to reduce the number of channels to decrease computational complexity and minimize the number of parameters. Secondly, it passes through both a maximum pooling layer and an average pooling layer independently, with their results concatenated thereafter. Finally, this concatenated feature map undergoes further processing via a convolutional layer followed by an activation function. This stage maps the processed results into a range between 0 and 1, thereby facilitating the extraction of weights for each feature point. The specific processing process is shown in Equation (1):

$$\begin{aligned} M_s(F') &= (f^{7\times7}([AvgPool(F'); MaxPool(F')])) \\ &= (f^{7\times7}([F'^{s}_{avg}; F'^{s}_{\max}])) \end{aligned} \tag{1}$$

where the dimensions of $F's\ avg$ and $F's\ max$ are defined as C/16×H×W. The variable σ is represented by the sigmoid function, while $f^{\,7\times7}$ denotes a convolution operation with a kernel size of 7 × 7. By multiplying the resulting spatial attention weight M_s with the input feature map F', one can obtain a feature map that incorporates spatial weighting.

3.3. RCA Module

The RCA mechanism is a technique designed to enhance the adjustment of importance weights for feature maps along channels [46]. In the existing HRNet model, various branches may extract feature information at different levels and perspectives. During the final fusion of these branches, redundant information and less significant feature channels can emerge, potentially impacting both computational efficiency and the model's generalization capability. To address this issue, we incorporate the residual network principle and introduce an improved channel attention mechanism module into the model [47]. The obtained RCA is illustrated in Figure 3.

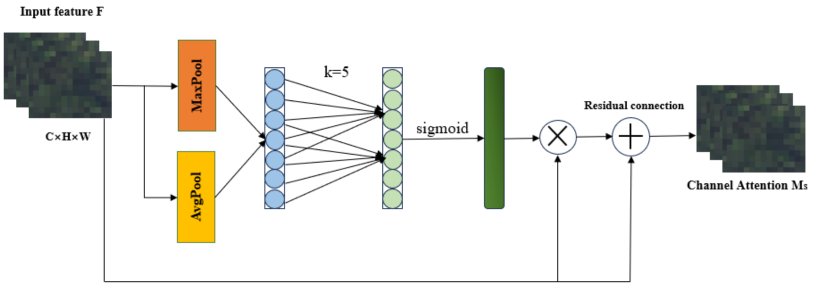


Figure 3. RCA mechanism module.

In our framework, the input feature maps undergo global average pooling, which is performed across the height and width of the feature layer to yield a 1 × 1 feature descriptor. Then, an adaptive 5 × 5 convolutional layer is applied for dimensionality reduction across all features. This is followed by another fully connected layer that further reduces the feature dimensions. Finally, each channel's weight in the feature map is normalized to a range between 0 and 1 using a sigmoid activation function, resulting in weights along the channel dimension.

The weighted feature map M_c is obtained by multiplying these normalized weights with the input feature map. The processed feature map is then derived through an additive residual connection with the original input feature map. The detailed processing process is shown in Equations (2) and (3):

$$\begin{aligned} M_c(F') &= \sigma(f^{(5\times5)}([AvgPool(F')])) \\ &= \sigma(f^{(5\times5)}([F'^{s}_{avg}])) \end{aligned} \tag{2}$$

$$k = \psi(C) = \left| \frac{\log_2(C)}{\gamma} + \frac{b}{\gamma} \right| \tag{3}$$

where $F\prime^{s}_{avg}$ represents an average pooling operation size of 1 × 1 × H, σ represents the sigmoid function, k represents the calculated size of the adaptive convolution kernel in the convolutional layer, and the feature map is processed through the $f^{\,5\times5}$ over-convolution operation to obtain the weight M_C. Ultimately, by multiplying and adding, we obtain the fused feature map F', which incorporates the channel attention mechanism represented by M_C.

3.4. Multi-Level Upsampling Module

In the upsampling process, bilinear interpolation is commonly employed to restore the low-resolution feature map. However, bilinear interpolation tends to be relatively simplistic when processing edge pixels, which results in blurred or jagged edges during image enlargement. This limitation adversely affects applications that demand high-quality edge preservation [48]. Consequently, we propose replacing bilinear interpolation with bicubic interpolation. Bicubic interpolation builds upon bilinear methods by considering a greater number of surrounding pixels and utilizing more complex polynomial functions for improved accuracy. Furthermore, in the original HRNet architecture, feature fusion involves directly amplifying features from branches that yield different resolutions before merging them. This straightforward approach leads to detail loss within the feature maps during processing, ultimately compromising the quality of the final fused output. To address this issue, we introduce a multi-level upsampling module designed to enhance detail retention during feature fusion. The processing process of this module is shown in Equation (4):

$$f(F_1, F_2, F_3, F_4) = Concat(Concat(Concat(F_3, F_4), F_2), F_1) \tag{4}$$

where F_1, F_2, F_3, F_4 is the feature map extracted from four levels.

The formula for the Bicubic difference is shown in Equation (5):

$$f(x,y) = \sum_{i=0}^{3}\sum_{j=0}^{3} p(i,j) \times W(x-i) \times W(y-i) \tag{5}$$

where p(i, j) is the value of adjacent 4 × 4 pixels. W(t) is a third-order difference function that uses a vertical variance kernel function, as shown in Equation (6):

$$W(t) = \begin{cases} 1 - 2|t|^2 + |t|^3 & if\ 0 \le |t| \le 1 \\ 1 - 2|t|^2 + |t|^3 - |t|^3 & if\ 1 \le |t| \le 2 \\ 0 & otherwise \end{cases} \tag{6}$$

In each layer, the feature maps of size H × W × C are concatenated with the feature maps obtained from the previous layer after applying Bicubic difference processing on each channel. This results in feature maps of size 2H × 2W × C/2. The feature maps F_i were merged and extracted according to their respective network levels, followed by a layer-by-layer upsampling process until the final extraction yielded feature maps of size 8H × 8W × C/8. The specific process is illustrated in Figure 4. The improved sampling mechanism during the fusion process is progressive, allowing for greater detail retention and producing clearer and smoother characteristic images that contain more information.

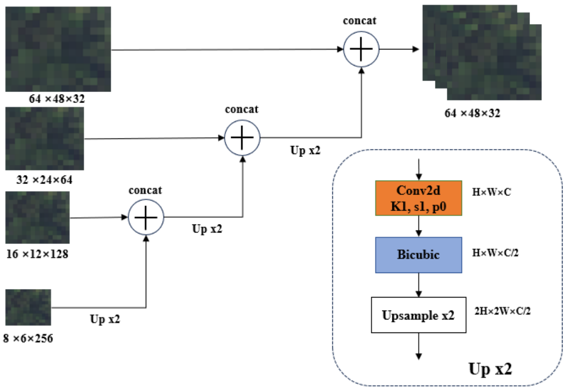


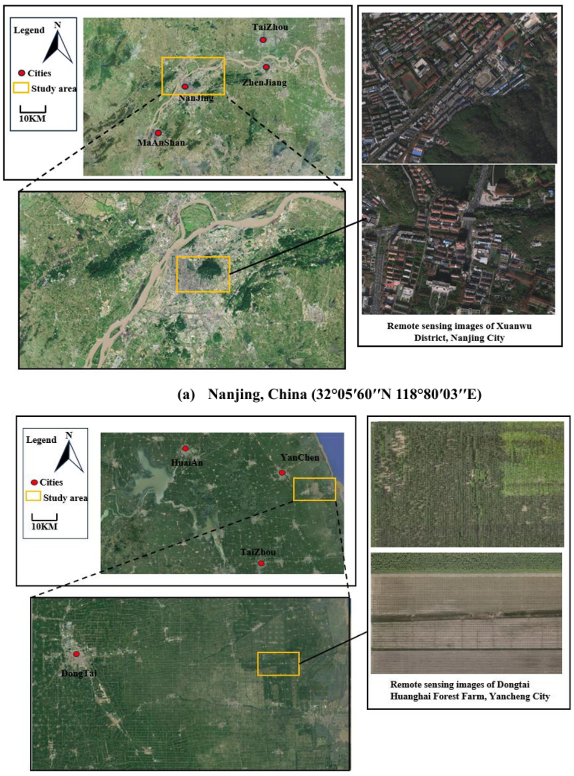
Figure 4. Multi-stage upsampling module.

4. Experimental Setting

4.1. Datasets

In terms of experiments, considering the adaptability and effectiveness of the model across various vegetation types and complex landforms, experiments were conducted utilizing a mixed dataset comprising multiple sources. LoveDA (Land-Cover dataset for Domain Adaptation) [49] is an urban-rural land cover dataset utilized for adaptive land cover selection, which will be used as the public benchmark for comparisons. Additionally, a mixed dataset was created, comprising the remote sensing urban forest from Nanjing and the Huanghai Forest Farm dataset from Yancheng, obtained through UAV aerial photography. The remote sensing images from Nanjing were employed to constitute the urban forest component of the dataset, while data from Yancheng, contributed to the artificial forest segment. Additionally, the LoveDA urban and rural dataset was utilized to incorporate diverse landform types such as natural forests, rivers, and hills to further enrich the dataset. This compiled dataset encompasses various landform characteristics, including forests, farmland, lakes, and plantations.

As shown in Figure 5a, the selected study area includes Zhongshan Mountain and Xuanwu Lake, both of which exhibit significant greening efforts and abundant urban green spaces. The dataset comprises 16 high-resolution remote sensing images with dimensions of 10,000 × 8000 pixels each, aligning well with the characteristics typical of urban forest areas. As shown in Figure 5b, in this dataset, there are two plots of poplar trees; the first plot covers an area of approximately 10,000 square meters, while the second plot spans about 3600 square meters. The equipment employed for sampling comprises the M350RTK, produced by DJI Technology Co., Ltd. in Shenzhen, China, equipped with P1 and L2 lenses (employed to collect orthophoto images and airborne laser data from the experimental area). An UAV, along with this equipment, is depicted in Figure 6. Given that deep learning semantic segmentation necessitates specific accuracy standards for input data, preprocessing was conducted on the acquired dataset. The processed remote sensing images were divided into image blocks measuring 1024 pixels by 512 pixels.


Legend
N
Cities
Study area
10KM
TaiZhou
ZhenJiang
NanJing
MaAnShan
Remote sensing images of Xuanwu District, Nanjing City
Legend
N
Cities
Study area
10KM
HuaiAn
YanChen
TaiZhou
DongTai
Remote sensing images of Dongtai Huanghai Forest Farm, Yancheng City

(a) Nanjing, China (32°05′60″N 118°80′03″E)

(b) Dongtai, China (32°51′33″N 120°50′12″E)

Figure 5. Study plots.

Figure 6. Our UAV equipment.

Since this experiment employs an end-to-end design, the input data is directly mapped to the output results; therefore, the preprocessing stage has been omitted. Labelme (version 5.1.1) image annotation software was employed for labeling purposes, resulting in a sample label dataset corresponding to order, quantity, and size specifications. Following this step, the dataset will be randomly divided into training, validation, and test sets in a ratio of 8:1:1. These subsets were designated for network training, parameter tuning during learning phases, and model evaluation, respectively.

4.2. Evaluation Metrics

We employed the Mean Intersection over Union (MIoU), Pixel Accuracy (PA), and F1-score as evaluation metrics to assess the semantic segmentation accuracy. MIoU is a widely recognized metric for evaluating semantic segmentation algorithms, as it comprehensively accounts for the prediction accuracy across different categories and provides an overall assessment. The MIoU value ranges from 0 to 1. Values closer to 1 indicate a higher degree of similarity between the predicted results and the true labels. Values approaching 0 signify lower similarity. The calculation formula is shown in Equation (7):

$$M_{Iou} = \frac{1}{k+1}\sum_{i=0}^{k}\frac{p_{ii}}{\sum_{i=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \times 100\% \tag{7}$$

PA is the proportion of correctly recognized pixels in all pixels, which is used to evaluate the accuracy of semantic segmentation. The formula is shown in Equation (8):

$$P_{PA} = \frac{\sum_{\mathrm{i}=0}^{k} p_{ii}}{\sum_{i=0}^{k}\sum_{j=0}^{k} p_{ij}} \times 100\% \tag{8}$$

In Equation (7), k + 1 is the number of classifications in the model; p_{ii} is pixel class i predicted as class i (TP), which is the pixel with correct segmentation; p_{ij} is pixel class j predicted as class i (FP); p_{ji} is pixel class i predicted as class j (FN), which represents the positive examples of incorrect prediction and negative examples of incorrect prediction.

The F1 score integrates precision and recall, making it particularly suitable for scenarios with class imbalance. It effectively reflects the overall performance of models in semantic segmentation.

The calculation formula is shown in Equation (9):

$$F1 = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{9}$$

$$\mathrm{Precision} = \frac{TP}{\mathrm{TP} + \mathrm{FP}} \tag{10}$$

$$\mathrm{Reacll} = \frac{TP}{\mathrm{TP} + \mathrm{FN}} \tag{11}$$

In Equation (10), TP refers to true positives (the number of correct predictions classified as positive), FP denotes false positives (the number of incorrect predictions classified as positive), and FN represents false negatives (the number of incorrect predictions classified as negative).

4.3. Parameter Sensitivity Analysis

In experiments, an RTX 4090 GPU (24 GB), manufactured by the American company NVIDIA (Santa Clara, CA, USA), along with CUDA 11.0 are employed for parallel computing while the PyTorch framework is utilized for constructing, training, and validating the deep learning network, as well as optimizing network parameters. The initial learning rate of the model is set to 0.01, with a weight decay of 0.0005 and momentum of 0.9; additionally, a batch size of 16 is adopted. Figure 7 illustrates the relationship between the epochs and Mean Intersection over Union (MIoU) during this experiment. The results indicate that the MIoU converges when the model reaches 300 epochs; therefore, a maximum training period of 300 epochs is selected.

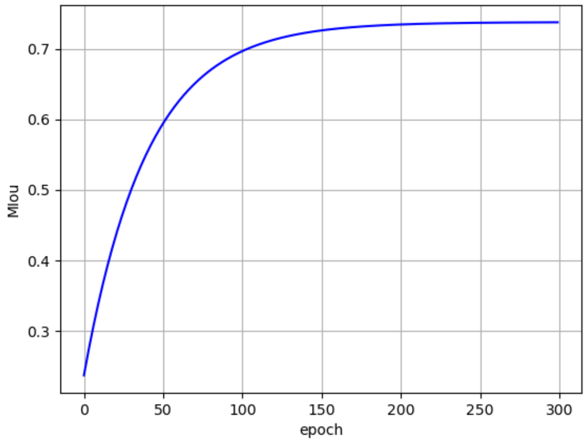


Figure 7. Epoch–MIou change trend diagram.

5. Results and Analysis

5.1. Analysis of Ablation Experiments

In order to verify the effectiveness and accuracy of the ASA mechanism module, RCA mechanism module, and multi-level upsampling module proposed in our work, we conducted an ablation experiment for validation. The results are presented in Table 1. It is evident that incorporating both the spatial attention mechanism module and the channel attention mechanism module reduces redundant information, enabling the model to better capture details and texture information present in the feature map. Compared with the original HRNet network, there is an increase of 2.08 percentage points in MIou and a rise of 4.76 percentage points in PA.

Table 1. Ablation test results.

NetWork	Multistage Upsampling	ASA	RCA	Mean Intersection Over/%	Pixel Accuracy/%
A	×	×	×	71.64	85.06
B	×	√	×	72.33	87.60
C	×	×	√	72.59	86.73
D	×	√	√	73.72	89.82
E	√	×	×	71.88	85.65
F	√	√	√	73.81	89.91

Furthermore, by introducing the multi-level upsampling module, our network retains more information during feature map fusion while effectively minimizing quality loss of the feature maps. When compared with the original model, MIou shows an improvement of 0.24 percentage points and PA increases by 0.59 percentage points.

5.2. Comparison of Semantic Segmentation Models

In order to analyze the performance of FA-HRNet in vegetation segmentation for high-resolution remote sensing images, we designed both qualitative and quantitative experimental analyses. We selected advanced semantic segmentation models with distinct network structural characteristics for comparison, including SegNet, DeepLabV3, DeepLabV3+, and HRNet. In the following experiments, these models will be tested and compared on our test dataset to validate the efficiency of FA-HRNet in vegetation segmentation.

As shown in Table 2, the mean Intersection over Union (mIoU) values for HRNet, U-Net, SegNet, DeepLabV3, and DeepLabV3+ are 68.42%, 63.97%, 62.33%, and 65.67%, respectively. It is evident that the performance of SegNet is lower than that of other network models by a margin of 3.37% compared to DeepLabV3+. HRNet demonstrates superior capability in feature map information extraction due to its multi-branch network architecture. Certain branches maintain a high resolution, facilitating the capture of detailed information, while other branches operate at a lower resolution to extract high-level features. This design endows HRNet with significant advantages over U-Net, SegNet, DeepLabV3, and other network models when addressing semantic segmentation tasks. The pixel accuracy of U-Net, SegNet, DeepLabV3, and FA-HRNet is 68.42%, 63.97%, 62.33%, and 73.81%, respectively. The experimental results indicate that FA-HRNet achieves a higher pixel accuracy (PA) value compared to the other models. Furthermore, in terms of F1-score comparison, the FA-HRNet network model attains an impressive score of 85.27, surpassing that of the other network models. Following the integration of the attention mechanism and the multi-level upsampling module, our proposed model FA-HRNet surpasses U-Net, SegNet, and DeepLabV3 in terms of mean Intersection over Union (mIoU) and pixel accuracy (PA) on the test set.

Table 2. Comparative experimental results.

NetWork	Mean Intersection Over Union/%	Pixel Accuracy/%	F1-Score/%
FA-HRNet	73.81	89.91	85.27
HRNet	71.64	85.06	82.33
U-Net	68.42	82.21	83.10
SegNet	63.97	78.88	77.94
DeepLabV3	62.33	73.12	69.23
DeepLabV3+	65.67	79.85	74.98

The results of semantic segmentation are presented in Figure 8. Images "a" and "b" were selected from urban remote sensing data of Nanjing, focusing on residential areas and road landscapes. Images "c" and "d" were obtained from the remote sensing images of Yancheng, highlighting plantation forest landscapes. Images "e" and "f" were sourced from the LoveDA dataset, showcasing rural and river landscapes. The findings indicate that the proposed network demonstrates effective segmentation performance when applied to remote sensing images featuring diverse land surface characteristics.

The detailed pairs are illustrated in Figure 9, the landscapes that are prone to confusion, such as lakes and roads, present challenges for DeepLabV3 and U-Net during segmentation tasks, leading to inaccuracies in vegetation segmentation. In contrast, the model proposed demonstrates superior segmentation performance and exhibits a more nuanced understanding of the easily confusable elements within remote sensing images. Figure 10 reveals that, when compared with the original HRNet, the proposed network minimizes both false detection and missed detection while identifying and segmenting canopy edges and grassland boundaries. This results in relatively complete and accurate representations of vegetation edge features. Overall, the proposed network model shows significant advancements over the original HRNet regarding detailed feature processing.

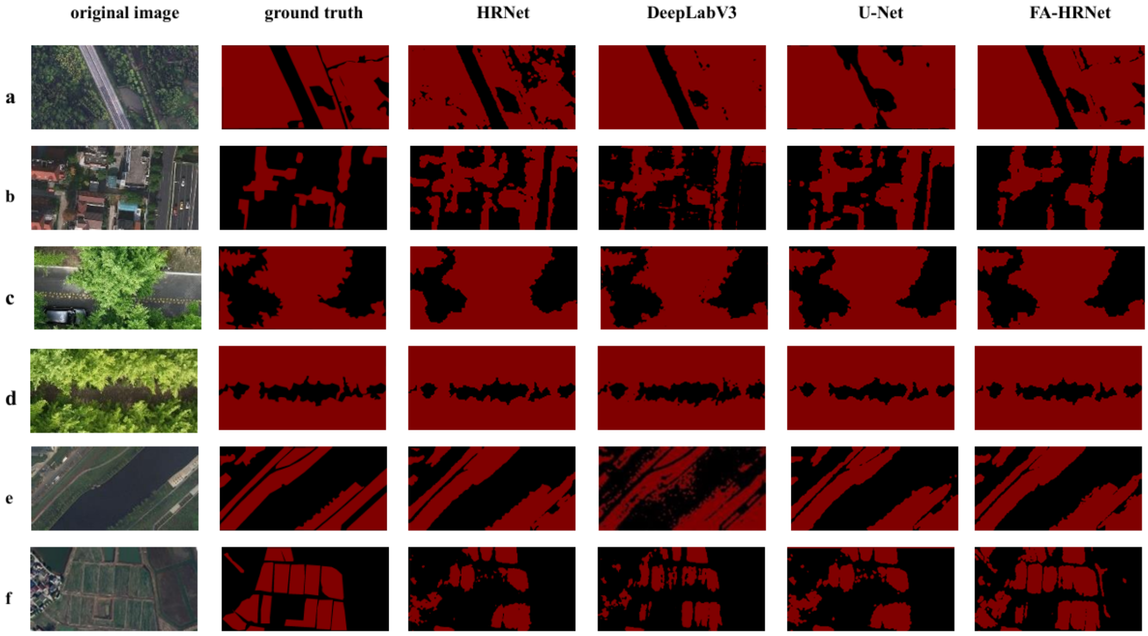


Figure 8. Comparison of vegetation segmentation visually. (**a**,**b**) Comparison of segmentation results from different semantic segmentation models on remote sensing images of Nanjing. (**c**,**d**) Comparison of segmentation results from different semantic segmentation models on remote sensing images of Yancheng. (**e**,**f**) Comparison of segmentation results from different semantic segmentation models on remote sensing images derived from the LoveDA dataset.

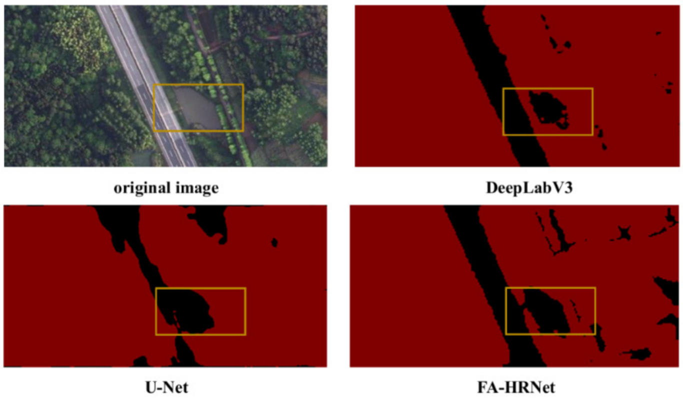


Figure 9. Comparison of DeepLabV3, U-Net, and the FA-HRNet model in urban forests. The content within the orange squares highlights the outstanding performance of the FA-HRNet model.

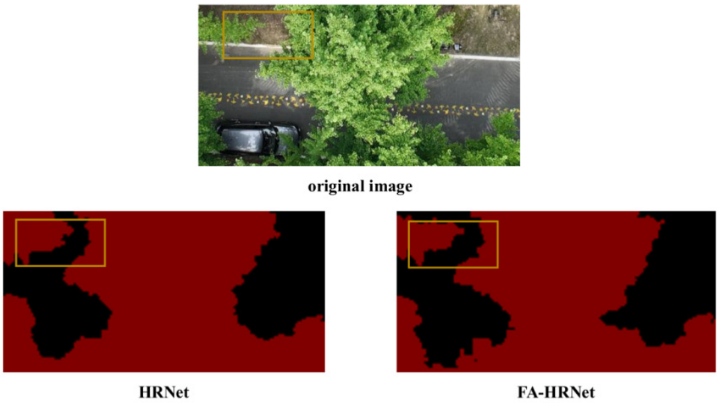


Figure 10. Comparison of HRNet and the FA-HRNet in artificial forests. The content within the orange squares highlights the outstanding performance of the FA-HRNet model.

5.3. Application of Vegetation Coverage Analysis

In order to further validate the extraction capability of the FA-HRNet model for large-scale high-resolution images, we conduct a vegetation coverage analysis on selected remote sensing images from Nanjing and Yancheng. In this study, six drone remote sensing images with a resolution of 17,004 × 26,440 pixels were selected from the Xuanwu District of Nanjing for vegetation semantic segmentation. Each remote sensing image encompasses an actual area of approximately 4.51 square kilometers. These remote sensing images capture various landform types, including buildings, lakes, wasteland, and vegetation, effectively representing the characteristics of urban forests. In contrast, the high-resolution remote sensing images obtained from Dongtai Huanghai Forest Farm in Yancheng cover an area of 10,000 square meters and possess a resolution of 54,700 × 34,800 pixels. The predominant tree species in this region are poplars; both the planting density and overall condition are favorable. Thus, these images serve as representative examples of plantation forests.

The vegetation segmentation outcomes for certain remote sensing images from Xuanwu District, Nanjing, are depicted in Figure 11. After conducting semantic segmentation using FA-HRNet, the extraction of vegetation cover within urban areas can be effectively accomplished. In relation to locality, as illustrated in Figure 11a, FA-HRNet exhibits a reduced incidence of misclassification for confusing landforms such as wastelands and water bodies. Furthermore, it demonstrates strong performance in segmenting urban trees that are obscured by shadows. Overall, FA-HRNet is capable of achieving relatively accurate vegetation segmentation for remote sensing images characterized by complex and diverse landforms. Subsequent analysis involving the calculation of average coverage values indicates that the vegetation coverage in Xuanwu District is approximately 29.50%. According to data released by the Nanjing Municipal Green Garden Bureau regarding forest coverage rates and green land statistics in 2024, as of the end of 2023, the forest coverage area in Nanjing is recorded at 1980.73 hectares, resulting in a forest coverage rate of 31.96% [50]. In the report, it is noted that the forest coverage rate in Nanjing stands at 31.96%, which closely aligns with our calculated vegetation coverage of 29.50% for Xuanwu District within Nanjing. Through a comparative analysis of segmentation effects and data evaluation, it can be observed that FA-HRNet demonstrates commendable performance in the semantic segmentation of remote sensing images pertaining to urban forest areas.

Furthermore, based on our analysis of plots from Huanghai Forest Farm in Dongtai, Yancheng. It is estimated that the vegetation coverage rate in the study area of Dongtai Huanghai Forest Farm is 86.55%, as illustrated in Figure 12. The Dongtai Huanghai Forest Farm falls within the Dongtai Huanghai National Forest Park located in Yancheng, Jiangsu, both exhibiting similar characteristics typical of plantation forest landforms. Consequently, vegetation data from the Dongtai Huanghai National Forest Park can be utilized to substantiate the segmentation performance discussed above. According to a special presentation by the Jiangsu Forestry Administration on 24 May 2021, the forest coverage rate within this park exceeds 80 percent [51]. This figure closely aligns with the calculated result of 86.55% presented in this work, thereby providing relevant evidence and support for our findings. In this segmentation map, various vegetation features—including trees, shrubs, and grasslands—were accurately identified by our model.

To further validate the performance of the FA-HRNet model for vegetation semantic segmentation on non-dataset images, we selected remote sensing images from Nanjing Forestry University in Jiangsu. The total area covered is 238,320 square meters, with a resolution of 36,580 × 20,100 pixels. The remote sensing image encompasses various geomorphological features including trees, shrubs, rivers, and buildings. The feature map resulting from semantic segmentation using FA-HRNet is presented in Figure 13. An analysis of this feature map reveals that the FA-HRNet network model demonstrates a commendable ability to identify vegetation obscured by buildings; furthermore, it exhibits relatively complete and clear segmentation edges. These findings substantiate that FA-HRNet performs effectively in segmenting vegetation within high-resolution remote sensing images.

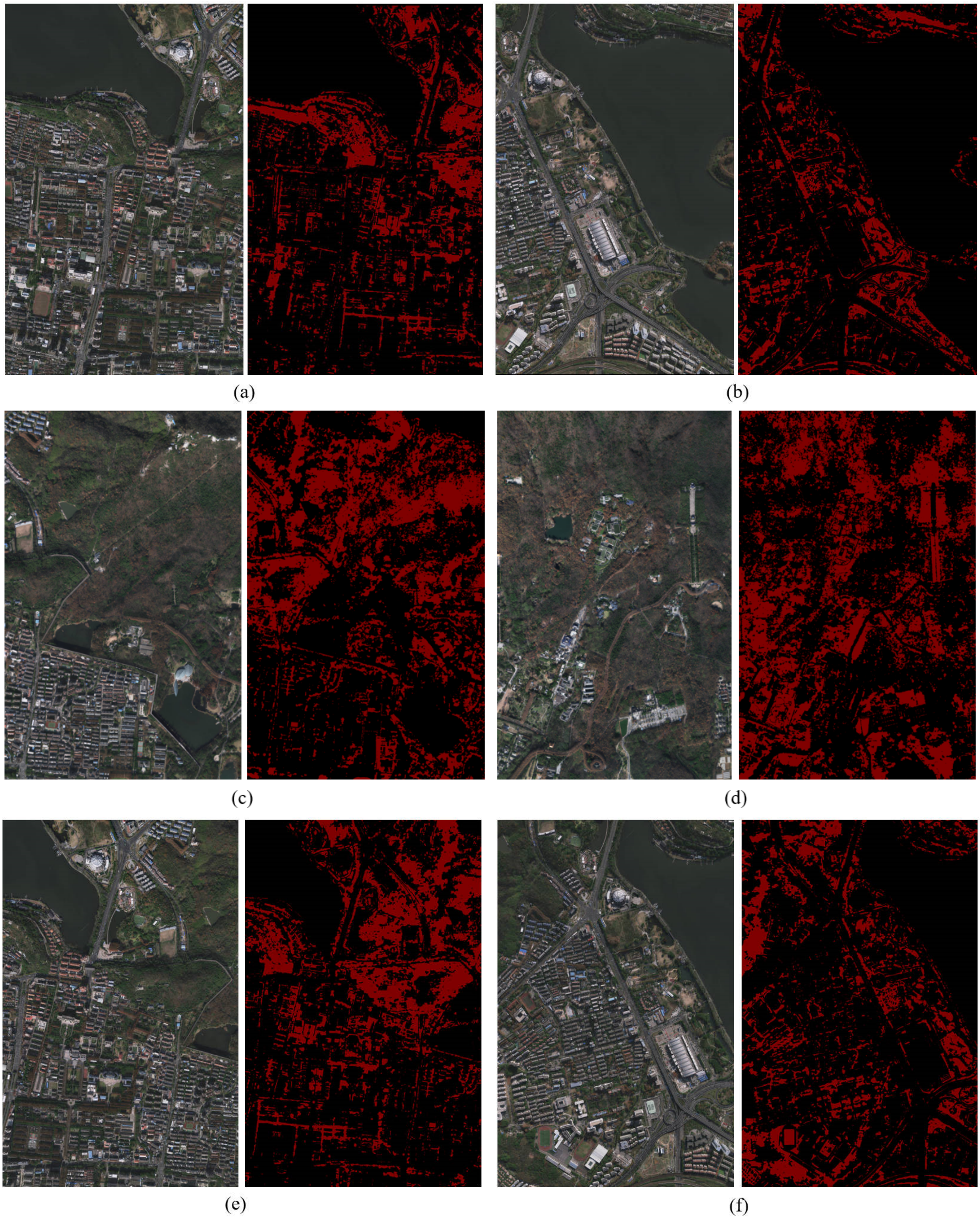


Figure 11. Segmentation results of large-scale remote sensing image in Xuanwu District, Nanjing City. (**a**,**b**) Comparison of remote sensing segmentation results in lakeside areas. (**c**,**d**) Comparison of remote sensing segmentation results in forested regions. (**e**,**f**) Comparison of remote sensing segmentation results in densely populated residential areas.

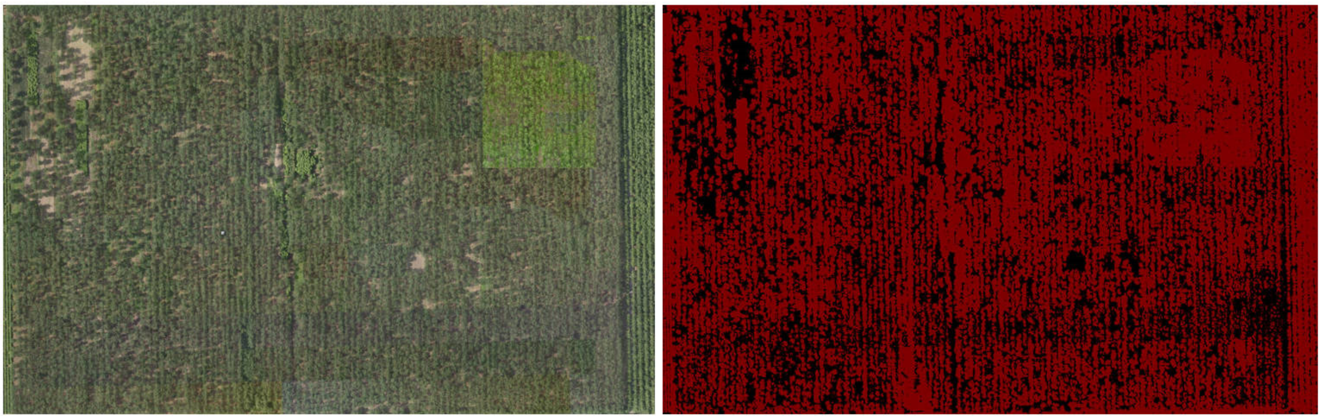


Figure 12. Segmentation results of large-scale remote sensing image in Huanghai Forest farm, Dongtai.

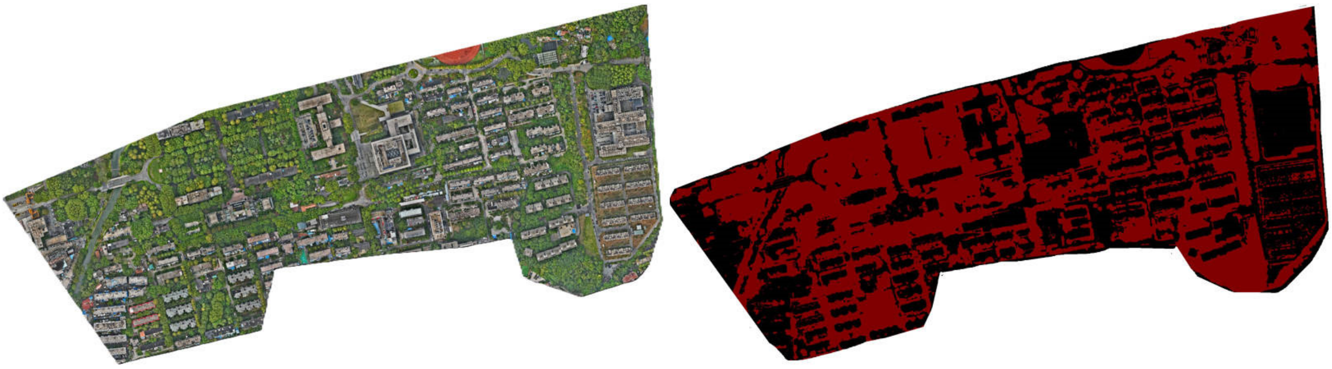


Figure 13. Segmentation results of large-scale remote sensing image in Nanjing Forestry University, Nanjing.

6. Conclusions

To achieve high-precision vegetation segmentation from remote sensing images, we fuse an optimized spatial attention mechanism module and a channel attention mechanism module. Compared with the baseline HRNet network, the mean Intersection over Union (MIoU) and pixel accuracy (PA) of our model improved by 2.08% and 4.76%, respectively. Furthermore, after incorporating the multi-level upsampling mechanism, MIoU and PA increased by an additional 2.17% and 4.85%, respectively. The vegetation segmentation performance of the proposed model is notably superior to other existing segmentation models. When compared with the U-Net network model, MIoU and PA are enhanced by 5.39% and 7.70%, respectively. When compared with DeepLabV3+, MIoU increases by 8.14% while PA rises by 10.06%.

Despite the high accuracy, experiment results show that the incorporation of multiple modules requires high hardware specifications, resulting in a relatively complex model structure and much more training times. To address these issues, future research should focus on optimizing the lightweight operation of the model and refining its internal network architecture to reduce both training duration and hardware requirements.

Author Contributions: Writing—original draft, B.H.; Writing—review & editing, B.H. and S.X.; Visualization, D.W. and S.X.; Supervision, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: This study was supported by the Scientific and Technological Innovation 2030—Major Projects (NO.2023ZD0405605).

Data Availability Statement: The original contributions presented in the study are included in the article, further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Alam, A.; Bhat, M.S.; Maheen, M. Using Landsat satellite data for assessing the land use and land cover change in Kashmir valley. *GeoJournal* **2020**, *85*, 1529–1543. [CrossRef]
2. Gao, Y.; Shao, Y.; Jiang, R.; Yang, X.; Zhang, L. Satellite image cloud automatic annotator with uncertainty estimation. *Fire* **2024**, *7*, 212. [CrossRef]
3. Lechner, A.M.; Foody, G.M.; Boyd, D.S. Applications in remote sensing to forest ecology and management. *One Earth* **2020**, *2*, 405–412. [CrossRef]
4. Ramankutty, N.; Foley, J.A. Estimating historical changes in global land cover: Croplands from 1700 to 1992. *Glob. Biogeochem. Cycles* **1999**, *13*, 997–1027. [CrossRef]
5. Li, S.; Xu, L.; Jing, Y.; Yin, H.; Li, X.; Guan, X. High-quality vegetation index product generation: A review of NDVI time series reconstruction techniques. *Int. J. Appl. Earth Obs. Geoinform.* **2021**, *105*, 102640. [CrossRef]
6. Kamiyama, M.; Taguchi, A. Color conversion formula with saturation correction from HSI color space to RGB color space. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.* **2021**, *E104A*, 1000–1005. [CrossRef]
7. Neyns, R.; Canters, F. Mapping of urban vegetation with high-resolution remote sensing: A Review. *Remote Sens.* **2022**, *14*, 1031. [CrossRef]
8. Bradter, U.; O'Connell, J.; Kunin, W.E.; Boffey, C.W.; Ellis, R.J.; Benton, T.G. Classifying grass-dominated habitats from remotely sensed data: The influence of spectral resolution, acquisition time and the vegetation classification system on accuracy and thematic resolution. *Sci. Total. Environ.* **2020**, *711*, 134584. [CrossRef]
9. Prakash, N.; Manconi, A.; Loew, S. Mapping landslides on EO data: Performance of deep learning models vs. traditional machine learning models. *Remote Sens.* **2020**, *12*, 346. [CrossRef]
10. Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; et al. Deep High-resolution representation learning for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *43*, 3349–3364. [CrossRef]
11. Gu, Y.; Zhang, H.; Zhang, Z.; Ye, Q. Unsupervised deep triplet hashing with pseudo triplets for scalable image retrieval. *Multimedia Tools Appl.* **2020**, *79*, 35253–35274. [CrossRef]
12. Chen, Q.; Zhang, H.; Ye, Q.; Zhang, Z.; Yang, W. Learning discriminative feature via a generic auxiliary distribution for unsupervised domain adaptation. *Int. J. Mach. Learn. Cybern.* **2022**, *13*, 175–185. [CrossRef]
13. Huang, X.; Jensen, J.R. A machine-learning approach to automated knowledge-base building for remote sensing image analysis with GIS data. *Photogramm. Eng. Remote Sens.* **1997**, *63*, 1185–1193.
14. Nath, S.S.; Mishra, G.; Kar, J.; Chakraborty, S.; Dey, N. A survey of image classification methods and techniques. In Proceedings of the 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), Kanyakumari District, India, 10–11 July 2014; pp. 554–557.
15. Zhu, L.; Ren, R.; Chen, D.; Song, A.; Liu, J.; Ye, N.; Yang, Y. Feel the inside: A haptic interface for navigating stress distribution inside objects. *Vis. Comput.* **2020**, *36*, 2445–2456. [CrossRef]
16. Kurita, T. Principal component analysis (PCA). In *Computer Vision: A Reference Guide*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 1–4.
17. Blundell, R.; Bond, S.; Windmeijer, F. Estimation in dynamic panel data models: Improving on the performance of the standard GMM estimator. *Nonstationary Panels Panel Cointegration Dyn. Panels* **2001**, *15*, 53–91.
18. Lei, T.; Jia, X.; Zhang, Y.; Liu, S.; Meng, H.; Nandi, A.K. Superpixel-based fast fuzzy C-means clustering for color image segmentation. *IEEE Trans. Fuzzy Syst.* **2018**, *27*, 1753–1766. [CrossRef]
19. Lv, Z.; Liu, T.; Shi, C.; Benediktsson, J.A.; Du, H. Novel land cover change detection method based on k-means clustering and adaptive majority voting using bitemporal remote sensing images. *IEEE Access* **2019**, *7*, 34425–34437. [CrossRef]
20. Ye, Q.; Huang, P.; Zhang, Z.; Zheng, Y.; Fu, L.; Yang, W. Multiview learning with robust double-sided twin SVM. *IEEE Trans. Cybern.* **2021**, *52*, 12745–12758. [CrossRef]
21. Li, M.; Zang, S.; Zhang, B.; Li, S.; Wu, C. A Review of remote sensing image classification techniques: The role of spatio-contextual Information. *Eur. J. Remote Sens.* **2014**, *47*, 389–411. [CrossRef]
22. Wang, C.; Ye, Q.; Luo, P.; Ye, N.; Fu, L. Robust capped L1-norm twin support vector machine. *Neural Netw.* **2019**, *114*, 47–59. [CrossRef]
23. Wang, Y.; Yang, X.; Zhang, L.; Fan, X.; Ye, Q.; Fu, L. Individual tree segmentation and tree-counting using supervised clustering. *Comput. Electron. Agric.* **2023**, *205*, 107629. [CrossRef]
24. Rigatti, S.J. Random forest. *J. Insur. Med.* **2017**, *47*, 31–39. [CrossRef] [PubMed]
25. Weng, L.; Qian, M.; Xia, M.; Xu, Y.; Li, C. Land use/land cover recognition in arid zone using A multi-dimensional multi-grained residual Forest. *Comput. Geosci.* **2020**, *144*, 104557. [CrossRef]
26. Sheykhmousa, M.; Mahdianpari, M.; Ghanbari, H.; Mohammadimanesh, F.; Ghamisi, P.; Homayouni, S. Support vector machine versus random forest for remote sensing image classification: A meta-analysis and systematic review. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 6308–6325. [CrossRef]
27. Linhui, L.; Weipeng, J.; Huihui, W. Extracting the forest type from remote sensing images by random forest. *IEEE Sens. J.* **2020**, *21*, 17447–17454. [CrossRef]

28. Zhang, L.; Wang, M.; Liu, M.; Zhang, D. A survey on deep learning for neuroimaging-based brain disorder analysis. *Front. Neurosci.* **2020**, *14*, 779. [CrossRef] [PubMed]
29. Fan, X.; Tjahjadi, T. Fusing dynamic deep learned features and handcrafted features for facial expression recognition. *J. Vis. Commun. Image Represent.* **2019**, *65*, 102659. [CrossRef]
30. Gao, D.; Ou, L.; Liu, Y.; Yang, Q.; Wang, H. DeepSpoof: Deep reinforcement learning-based spoofing attack in cross-technology multimedia communication. *IEEE Trans. Multimedia* **2024**, 1–13. [CrossRef]
31. Fan, X.; Luo, P.; Mu, Y.; Zhou, R.; Tjahjadi, T.; Ren, Y. Leaf image based plant disease identification using transfer learning and feature fusion. *Comput. Electron. Agric.* **2022**, *196*, 106892. [CrossRef]
32. Gao, D.; Wang, H.; Guo, X.; Wang, L.; Gui, G.; Wang, W.; Yin, Z.; Wang, S.; Liu, Y.; He, T. Federated learning based on CTC for heterogeneous internet of things. *IEEE Internet Things J.* **2023**, *10*, 22673–22685. [CrossRef]
33. Yu, T.; Hu, C.; Xie, Y.; Liu, J.; Li, P. Mature pomegranate fruit detection and location combining improved F-PointNet with 3D point cloud clustering in orchard. *Comput. Electron. Agric.* **2022**, *200*, 107233. [CrossRef]
34. Zhu, Y.; Sun, W.; Cao, X.; Wang, C.; Wu, D.; Yang, Y.; Ye, N. TA-CNN: Two-way attention models in deep convolutional neural network for plant recognition. *Neurocomputing* **2019**, *365*, 191–200. [CrossRef]
35. Liu, X.; Hu, C.; Li, P. Automatic segmentation of overlapped poplar seedling leaves combining Mask R-CNN and DBSCAN. *Comput. Electron. Agric.* **2020**, *178*, 105753. [CrossRef]
36. Lu, C.; Xia, M.; Lin, H. Multi-scale strip pooling feature aggregation network for cloud and cloud shadow segmentation. *Neural Comput. Appl.* **2022**, *34*, 6149–6162. [CrossRef]
37. Ahmed, K.; Torresani, L. Connectivity learning in multi-branch networks. *arXiv* **2017**, arXiv:1709.09582.
38. Mahendra, H.N.; Mallikarjunaswamy, S.; Subramoniam, S.R. An assessment of vegetation cover of Mysuru City, Karnataka State, India, using deep convolutional neural networks. *Environ. Monit. Assess.* **2023**, *195*, 1–20. [CrossRef]
39. Liu, M.; Fu, B.; Xie, S.; He, H.; Lan, F.; Li, Y.; Lou, P.; Fan, D. Comparison of multi-source satellite images for classifying marsh vegetation using DeepLabV3 Plus deep learning algorithm. *Ecol. Indic.* **2021**, *125*, 107562. [CrossRef]
40. Rong, Q.; Hu, C.; Hu, X.; Xu, M. Picking point recognition for ripe tomatoes using semantic segmentation and morphological processing. *Comput. Electron. Agric.* **2023**, *210*, 107923. [CrossRef]
41. Ferchichi, A.; Ben Abbes, A.; Barra, V.; Farah, I.R. Forecasting vegetation indices from spatio-temporal remotely sensed data using deep learning-based approaches: A systematic literature review. *Ecol. Inform.* **2022**, *68*, 101552. [CrossRef]
42. Che, Z.; Shen, L.; Huo, L.; Hu, C.; Wang, Y.; Lu, Y.; Bi, F. MAFF-HRNet: Multi-attention feature fusion HRNet for building segmentation in remote sensing images. *Remote Sens.* **2023**, *15*, 1382. [CrossRef]
43. Li, L.; Tian, T.; Li, H.; Wang, L. SE-HRNet: A deep high-resolution network with attention for remote sensing scene classification. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Virtual, 26 September–2 October 2020; pp. 533–536.
44. Chu, X.; Tian, Z.; Wang, Y.; Zhang, B.; Ren, H.; Wei, X.; Xia, H.; Shen, C. Twins: Revisiting the design of spatial attention in vision transformers. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 9355–9366.
45. Jiang, S.; Lu, M.; Hu, K.; Wu, J.; Li, Y.; Weng, L.; Xia, M.; Lin, H. Personalized federated learning based on multi-head attention algorithm. *Int. J. Mach. Learn. Cybern.* **2023**, *14*, 3783–3798. [CrossRef]
46. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, 14–19 June 2020.
47. Qiao, Z.; Wang, K.; Xia, M.; Xu, Y.; Liu, W.; Weng, L. Multi-scale residual network for energy disaggregation. *Int. J. Sens. Networks* **2019**, *30*, 172. [CrossRef]
48. Yang, L.; Hong, S. Unsupervised time-series representation learning with iterative bilinear temporal-spectral fusion. In Proceedings of the International Conference on Machine Learning, PMLR, Baltimore, MA, USA, 17–23 July 2022.
49. Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; Zhong, Y. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv* **2021**, arXiv:2110.08733.
50. Nanjing City Greening and Horticulture Bureau. Nanjing City Forestry and Parks Bureau Announces Release of Related Data on Forest Coverage Rate and Green Space Ratio in Nanjing City. 15 May 2024. Available online: https://ylj.nanjing.gov.cn/njslhylj/202405/t20240515_4666645.html (accessed on 13 September 2024).
51. Jiangsu Forestry Bureau. Dongtai Huanghai National Forest Park: Transforming Saline-Alkaline Land into the Largest Plain Forest Along the East Coast. 25 May 2021. Available online: https://lyj.jiangsu.gov.cn/art/2021/5/24/art_7085_9821002.html (accessed on 14 September 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.