remote sensing

MDPI

Article

Multisource High-Resolution Remote Sensing Image Vegetation Extraction with Comprehensive Multifeature Perception

Yan Li [1], Songhan Min [2], Binbin Song [1], Hui Yang [3,*], Biao Wang [1] and Yongchuang Wu [4]

[1] School of Resources and Environmental Engineering, Anhui University, Hefei 230601, China; liyan98@stu.ahu.edu.cn (Y.L.); songbinbin@stu.ahu.edu.cn (B.S.); wangbiao-rs@ahu.edu.cn (B.W.)

[2] Stony Brook Institute at Anhui University, Hefei 230601, China; songhan.min@stonybrook.edu

[3] Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China

[4] School of Artificial Intelligence, Anhui University, Hefei 230601, China; wa23101006@stu.ahu.edu.cn

* Correspondence: yanghui@ahu.edu.cn

Abstract: High-resolution remote sensing image-based vegetation monitoring is a hot topic in remote sensing technology and applications. However, when facing large-scale monitoring across different sensors in broad areas, the current methods suffer from fragmentation and weak generalization capabilities. To address this issue, this paper proposes a multisource high-resolution remote sensing image-based vegetation extraction method that considers the comprehensive perception of multiple features. First, this method utilizes a random forest model to perform feature selection for the vegetation index, selecting an index that enhances the otherness between vegetation and other land features. Based on this, a multifeature synthesis perception convolutional network (MSCIN) is constructed, which enhances the extraction of multiscale feature information, global information interaction, and feature cross-fusion. The MSCIN network simultaneously constructs dual-branch parallel networks for spectral features and vegetation index features, strengthening multiscale feature extraction while reducing the loss of detailed features by simplifying the dense connection module. Furthermore, to facilitate global information interaction between the original spectral information and vegetation index features, a dual-path multihead cross-attention fusion module is designed. This module enhances the differentiation of vegetation from other land features and improves the network's generalization performance, enabling vegetation extraction from multisource high-resolution remote sensing data. To validate the effectiveness of this method, we randomly selected six test areas within Anhui Province and compared the results with three different data sources and other typical methods (NDVI, RFC, OCBDL, and HRNet). The results demonstrate that the MSCIN method proposed in this paper, under the premise of using only GF2 satellite images as samples, exhibits robust accuracy in extraction results across different sensors. It overcomes the rapid degradation of accuracy observed in other methods with various sensors and addresses issues such as internal fragmentation, false positives, and false negatives caused by sample generalization and image diversity.

Keywords: feature fusion; multisource high-resolution imagery; vegetation extraction

check for updates

Citation: Li, Y.; Min, S.; Song, B.; Yang, H.; Wang, B.; Wu, Y. Multisource High-Resolution Remote Sensing Image Vegetation Extraction with Comprehensive Multifeature Perception. *Remote Sens.* **2024**, *16*, 712. https://doi.org/10.3390/rs16040712

Academic Editor: Abolfazl Abdollahi

Received: 27 December 2023
Revised: 10 February 2024
Accepted: 13 February 2024
Published: 18 February 2024

1. Introduction

Vegetation is an essential component of the Earth's ecosystems, holding significant importance for ecological conservation and development [1]. Vegetation extraction based on the spectral, textural, spatial, temporal, and other features of high-resolution remote sensing images plays a vital role in resource surveys, urban planning, land surveys, and forest fire monitoring [2–4]. However, when dealing with large-scale vegetation monitoring across wide areas and different sensors, current methods often face challenges such as reduced model generalization, misclassification, internal fragmentation, and unclear boundaries [5–8].

Currently, the main methods for vegetation extraction research include threshold segmentation based on vegetation indices [9,10]. This method constructs spectral indices based on the differences in reflectance between the red band and near-infrared band to extract vegetation [10]. While this method can extract vegetation from images, it may struggle to identify vegetation obscured by tall buildings. In sparsely vegetated areas, vegetation index-based methods may also struggle to accurately estimate vegetation coverage. Many machine learning methods have been used for vegetation extraction, such as maximum likelihood classification (MLC) [11], support vector machine (SVM) [12], and random forest (RF) [13]. Although these machine learning methods can achieve high accuracy in imagery, feature and threshold conditions need to be manually designed [14], making it challenging to achieve automatic extraction from broad and multisource data.

Due to its outstanding performance in the field of computer vision, deep learning has gradually been applied in vegetation monitoring applications of remote sensing images, becoming one of the important methods [15–21]. In vegetation remote sensing, deep learning methods are divided into semantic segmentation-based methods [22,23] and pixel-based methods [24–26]. These methods independently learn relevant data features in an end-to-end manner, meeting the growing demand for vegetation assessment and monitoring in diverse remote sensing data. Convolutional neural networks (CNNs) in deep learning methods can accurately reveal the spatial characteristics of vegetation from various remote sensing sensors [27]. Subsequently, many researchers have used CNN models for semantic segmentation-based vegetation extraction, including VGGNet [28], ResNet [29], and DenseNet [30]. However, due to the continuous convolutional and pooling operations, these methods lose some spatial details, making it difficult to accurately predict spatially sensitive tasks [14]. Networks with similar structures include UNet [23], HRNet [31], and others. Zou et al. [23] and others have used the UNet architecture to distinguish land cover and crop types, achieving high accuracy on retained data in the Midwest of the United States. Xu et al. [31] and others have used HRNet to classify urban green spaces with a certain level of accuracy. Although the abovementioned semantic segmentation methods achieve good accuracy in vegetation extraction, their essence is to aggregate homogeneous pixels, often ignoring some fine-grained features. They cannot consider the internal details of vegetation features and are prone to confusion with other coexisting land features [32]. Moreover, semantic segmentation methods require precise boundary information when creating labels, making it difficult to distinguish mixed pixels at the boundaries of vegetation and other land cover areas. Creating sample labels on multisource remote sensing images is challenging, which further leads to convergence difficulties in training semantic segmentation models.

Single Pixel-based classification only requires adding some labeled pixels in sample images of specific scenes, allowing us to quickly and easily adapt to new scenes [24]. Therefore, using pixel-based methods for sampling allows us to regularly retrain classifiers and enables rapid sample collection. Many scholars have also implemented vegetation and crop extraction based on pixel-based methods [33,34]. Although these methods have achieved some success in vegetation remote sensing through continuous improvement, using only spectral information can lead to issues of vegetation omission and overestimation across different data sources.

To better address these problems, methods that combine multiple features, such as spectral and vegetation indices, have been proposed and widely applied [26,34], combining different fusion methods with various classification approaches to achieve better classification results [31,35,36].

However, most remote sensing vegetation extraction methods only extract vegetation based on spatial–spectral features from a single data source or initially combine vegetation index features by concatenation [37] for input into the network for training. Using only spectral features for vegetation extraction can result in significant differences in spectral values due to factors such as temporal and radiometric variations, making it difficult to achieve high-precision vegetation extraction across different data sources. The spectral feature values differ greatly from index feature values, and when index features are introduced through concatenation, they are often overlooked, making it difficult to achieve mutual complementary correction among multiple features, leading to issues such as omission and overestimation in vegetation extraction from different data sources [38]. In addition, the selection of vegetation index features relies solely on manual experience and lacks quantitative screening methods, which makes it difficult to scientifically and reasonably verify the effectiveness of these indices [39].

In response to the above issues, this paper presents innovative solutions. Using spectral features expanded by a 5 × 5 neighborhood of target pixels combined with feature selection, parallel vegetation index branches are synchronously analyzed to narrow the differences in feature values between different sensors while expanding the receptive field. At the same time, a simplified dense connection method is introduced into the network to enhance information sharing and weight between multi-scale and multi-feature; a multi-channel enhanced feature cross-fusion method under the self-attention mechanism is constructed, that is, cross-modal feature fusion between self shallow initial features and enhanced index features, to establish the correlation of multi-feature comprehensive perception and achieve complementary global information.

The primary contributions of our work can be summarized as follows:

1. We constructed a convolutional network (MSICN) for vegetation extraction that considers the comprehensive perception of multiple features, introducing simplified dense connections and cross-attention mechanisms to enhance information sharing and weighting among multiscale and multi-feature layers, achieving multi-feature representation, enhancement, fusion, and extraction of vegetation.
2. Using random forests for the selection of vegetation index features, the impact of vegetation index features on the accuracy of vegetation extraction across different data sources was determined.
3. The universality and generalization of the network across different data sources were verified.

2. Related Work

2.1. Pixel-Based Inversion Method for Vegetation Extraction

This method, based on pixel inversion, can be divided into modeling for individual pixels and modeling for neighboring pixels for inversion. This method assigns each pixel in the image to predefined categories independently, with each pixel being considered as an independent sample, and its characteristic pixels are composed of spectral information from itself or surrounding pixels. With the rise in deep learning, it is gradually being used for vegetation extraction based on pixel inversion. The commonly used deep learning networks based on pixel inversion currently include convolutional neural networks (CNNs) [33], Recurrent Neural Networks (RNNs) [33], and Long Short-Term Memory (LSTM) Networks [38]. Mazzia et al. [33] constructed vegetation classification models and PixelR-CNN models based solely on single pixel spectral information to achieve vegetation classification and land cover and crop classification, respectively. Fang et al. [40] proposed a deep spatio-spectral feature fusion network, using CNN to learn single-pixel spectral information and LSTM to learn crop classification. However, methods based on single-pixel inversion cannot consider the spectral information of surrounding pixels and spatial contextual information, which may lead to increased salt-and-pepper noise, unclear boundaries, and other issues. Liu et al. [41] proposed a CNN-based method which crops the image into small-scale slices composed of multiple neighboring pixels to predict vegetation categories. Inspired by this, we extend the single pixel samples to 5 × 5 size slices for modeling to alleviate spatial inconsistency and boundary fragmentation issues caused by pixel inversion methods.

2.2. Vegetation Index Feature fusion Method

In order to deal with the variability between individual pixels under different sensors and the changes in vegetation conditions, methods of multi-feature fusion of spectra and vegetation indices have been proposed and widely applied [33,34], combining different fusion methods with different classification methods to achieve better classification purposes. In order to fully utilize the potential of vegetation index features, many studies directly concatenate vegetation index features in dimensions as multiple feature inputs of the network. Radke et al. [35] proposed the Y-NET network was designed using spectral features from different bands and vegetation index features combined with terrain data to monitor vegetation growth status. Chen et al. [31] used the difference value of NDVI between winter and summer combined with spectral information added to the network input layer and trained using the HRnet network for classifying urban green spaces. In order to better focus on vegetation index features [42], Wu et al. [36] integrated spectral features and vegetation index features through channel attention mechanism in the network to extract soybeans, corn, and rice in the Hulunbuir region using Sentinel-2 images. These methods are of great significance for the comprehensive utilization of spectral features and index features, but there are still certain limitations in feature complementary correction. To address this, Lee et al. [43] proposed the 'cross' attention, and Zeng et al. [44] designed a two-branch transformer structure to learn features at different scales. Inspired by this, we propose to separate the spectral branch and the index feature branch, while designing a multi-channel enhanced feature cross-fusion method based on self-attention mechanism [45], to achieve cross-modal feature cross-fusion of shallow initial features and enhanced index features, thus realizing global information complementarity.

3. Materials and Methods

This paper presents a multisource high-resolution remote sensing image vegetation extraction method that accounts for the comprehensive perception of multiple features and synchronously analyzes spectral features and vegetation index features. Due to the potential of vegetation index features to improve vegetation extraction accuracy and reduce the differences in vegetation feature values across different sensors, the approach proposed in this paper initially employs a random forest model to feature-select vegetation indices, filtering out indices that can enhance the interclass differences between vegetation and other land objects. Simultaneously, a multifeature integrated perception convolutional network (MSCIN) with multiscale feature enhancement, global information interaction, and feature cross-fusion capabilities is constructed. It combines spectral feature branches with index feature branches that have been jointly filtered and introduces a simplified dense connection module and a dual-path cross-multihead cross-attention feature fusion module to enhance the interaction and fusion of features, aiming to achieve higher-precision vegetation extraction across different sensors.

3.1. Vegetation Index Selection

This paper conducted a feature importance assessment for 12 common and widely used vegetation indices (Table 1). Feature selection was performed by calculating the various vegetation indices using the spectral values of the center pixel of the collected samples. This paper uses a random forest regression model [46] to establish a feature mapping relationship between vegetation index features and labels for feature selection. Random forest feature importance ranking refers to sorting the contribution of each feature to the prediction target in the random forest model to determine the relative importance of each feature. The model consisted of 20 decision trees [47] with a depth of 4. After fitting the data, the model assessed the importance of vegetation index features. Features with higher numerical values in the results obtained by the random forest model were considered more important for prediction accuracy. This allowed us to determine the sensitivity of vegetation index features to the model and eliminate feature sets that contributed minimally to the classification model. Through random forest feature importance ranking, it is possible to effectively determine which features in vegetation index features are most critical for extracting vegetation information. This targeted selection and utilization of these features for deep learning model training and prediction helps the model better learn vegetation characteristics, improve the performance of vegetation extraction models, reduce feature space, enhance feature expressiveness, improve model generalization ability, and enhance model interpretability, thus effectively extracting vegetation information. The importance ranking of the spectral bands and vegetation indices was determined using the method described above, as shown in Figure 1.

Table 1. Image band dataset detailed specification.

Index	Name	Formulation	Reference
NDVI	Normalized Difference VI	(NIR − Red)/(NIR + Red)	[9]
GNDVI	Green Normalized Difference VI	(NIR − Green)/(NIR + Green)	[48]
EVI	Enhanced VI	2.5 ×(NIR − Red)/(NIR + 6 × Red − 7.5 × Blue + 1)]	[49]
OSAVI	Optimized Soil Adjusted VI	(NIR − Red)/(NIR + Red + 0.16)	[50]
RVI	Ratio VI	NIR/Red	[51]
DVI	Difference VI	NIR − Red	[52]
TVI	Transform VI	0.5 × [120 × (NIR − Green) − 200 × (Red − Green)]	[53]
GVI	Green VI	(NIR/(NIR + Green)) − (Red/(Red + Green))	[54]
GI	Green Index	(NIR/Green) − 1	[55]
NDGI	Normalized Difference Green Index	(Green − SWIR)/(Green + SWIR)	[56]
MCARI	Modified Soil Adjusted VI	(1.2 × (NIR − Red) − 2.5 × (Blue − Red))/ Sqrt((2 × NIR + 1) ˆ2 − (6 × NIR − 5 × sqrt(Red)) − 0.5)	[57]
TCARI	Transform Soil Adjusted VI	3 × ((NIR − Red) − 0.2 × (NIR − Green) × (NIR/Red))	[57]

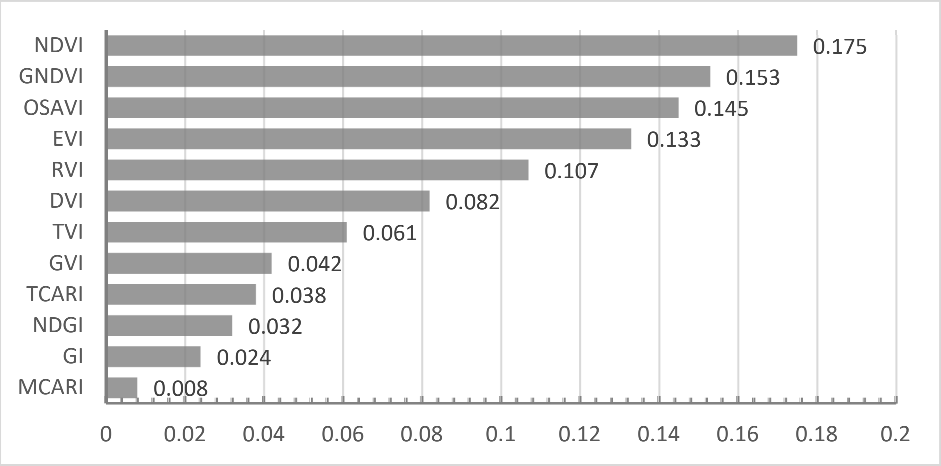

Figure 1. Vegetation index feature importance ranking chart.

From the importance ranking in Figure 1, it can be seen that this paper selects the top four vegetation index features and inputs them in parallel with the original image spectral features into the model to improve vegetation extraction accuracy.

3.2. Method

This paper combined the vegetation indices selected by the random forest regression model with the spectral features of remote sensing images for synchronous analysis. These features were input into the network in parallel, and deep feature fusion was performed after exchanging different scale features. Considering the spatial correlation between adjacent pixels, annotated pixel samples were used as central pixels, and their surrounding 5 × 5 neighborhood windows were extended as inputs to the model. The parallel multiscale network extraction method was used to extract spatial features and index features at scales of 1 × 1, 3 × 3, and 5 × 5. A simplified dense connection method was introduced, and the network was expanded in the vertical direction using downsampling to enhance feature extraction and reuse at different resolutions through multiple parallel subnetworks. Additionally, a dual-path multihead attention feature fusion module was constructed, which, under the establishment of multifeature comprehensive perception and correlation, achieved complementary interactions between spectral features and vegetation index features across channels, as shown in Figure 2.

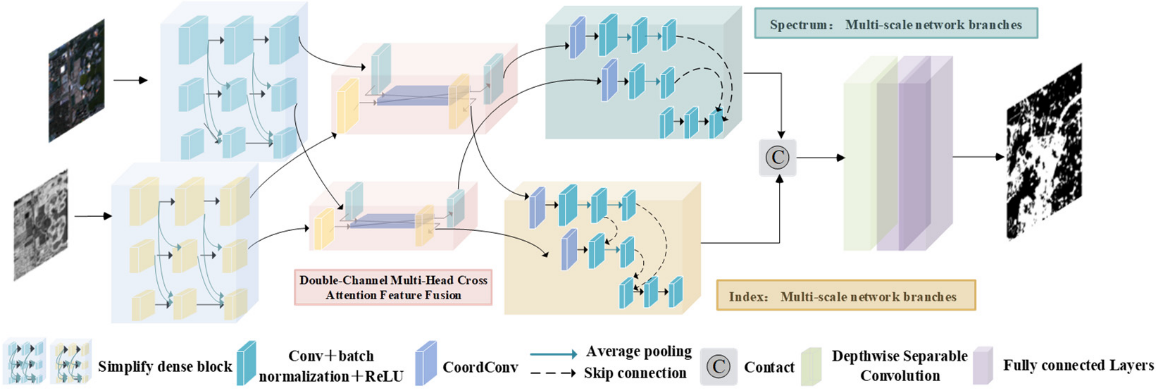
Spectrum: Multi-scale network branches
Ⓒ
Double-Channel Multi-Head Cross Attention Feature Fusion
Index: Multi-scale network branches
Simplify dense block
Conv+batch normalization+ReLU
CoordConv
Average pooling
Skip connection
Ⓒ
Contact
Depthwise Separable Convolution
Fully connected Layers

Figure 2. The presentation of MSICN-net.

Specifically, the spectrum and index 5 × 5 samples were simultaneously input into two branches, divided into three scales (1 × 1, 3 × 3, 5 × 5) centered around the target pixel. Inside the two branches, multiple parallel subnetworks were used, and a simplified dense connection module was added to simultaneously learn spectral and index features of different sizes. Multiple downsampling dense connections were used to ensure that vegetation features at different scales received information from the upper parallel subnetworks, allowing feature interaction and correction among the various features. To further enhance and improve the fusion effect of spectral and index features, this paper designed a dual-path cross-attention feature fusion module based on a visual transformer's self-attention with the maximum receptive field. This module was placed between the 5 × 5 and 3 × 3 branches and achieved cross-modal feature fusion between shallow initial features and enhanced index features. It established global information complementarity under the perception of multifeature integration and used mixed-channel embedding to produce enhanced output features. Convolutional layers were used for feature selection and integration between the two output branches, gradually integrating features at different scales. Additionally, skip connections were designed within the three scales of index branch networks to enhance the fusion of vegetation index features at different scales. Multiple parallel subnetworks are concatenated after continuous convolution, generating 1 × 1 spectral feature branches and 1 × 1 index feature branches for merging. Additionally, a depthwise separable convolution layer (DWCConv 3 × 3) was added for fine feature extraction and reduced parameter computation. Finally, the classification results of the target pixel were generated by dual-layer fully connected layers and softmax.

(1) Simplified Dense Block

Inspired by the feature reuse achieved by dense connection blocks in DenseNet [30], this paper designed a multiscale dense connection module to enhance the exchange of feature information among three scales: 5 × 5, 3 × 3, and 1 × 1. In this module, each subnetwork can receive feature information from other parallel networks, enabling information exchange at different scales and improving feature representation. Specifically, a regular convolutional network generates L connections in L layers. However, in the multifeature connection module we adopted, features from the previous layer were passed on in a stacked manner to subsequent layers. In the 5 × 5 and 3 × 3 branches, we performed downsampling in the vertical direction to expand the network's width. Features from the 5 × 5 branch were overlaid and mapped to the 3 × 3 and 1 × 1 branch networks through feature concatenation, and then the features from the 3 × 3 branch were overlaid onto the 1 × 1 branch network. Each branch network was concatenated with other branch networks in the channel dimension, retaining feature information at different scales and achieving feature fusion. This not only effectively alleviates the problem of gradient vanishing but also reduces the number of parameters, enhancing the feature extraction and reuse capabilities.

(2) Dual-Path Multihead Cross-Attention Feature Fusion

In remote sensing image analysis, multispectral data in different bands contain rich information. It can provide reflectance information of different surface bands, thereby providing strong support for research on vegetation cover, land use, and environmental changes, among other aspects. However, the information contained in the different bands of multispectral data is complex, and the complementarity between these pieces of information can be better utilized through certain information fusion strategies. Therefore, we designed a dual-path multihead cross-attention feature fusion module to enhance the interaction between multispectral data and vegetation indices, improving the accuracy and effectiveness of vegetation identification in remote sensing images, as shown in Figure 3.

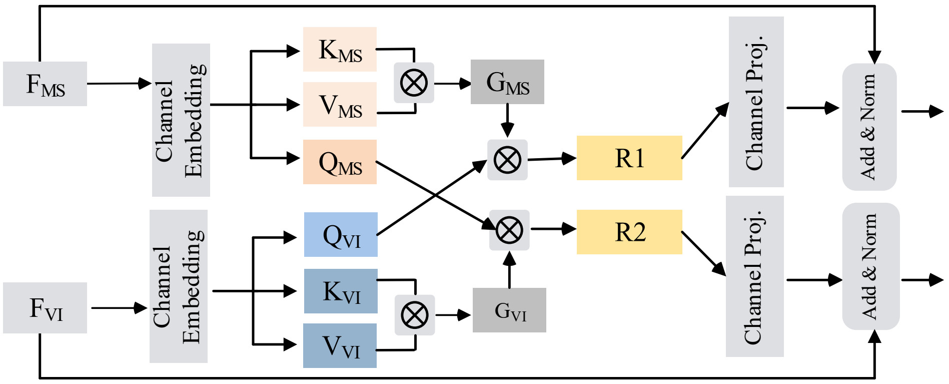

Figure 3. The presentation of MSICN-net.

Traditional self-attention mechanisms [49], when dealing with image features, first pass the input features through the PatchEmbedding layer and head operations to generate query, key, and value representations for each attention head. Then, by calculating the inner product of query (Q) and key (K), attention scores are obtained, and finally, different features are combined through weighted summation to achieve information transmission and fusion. However, this method requires significant memory and computational resources when processing high-resolution images, limiting its feasibility in practical applications. The process can be expressed as follows:

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

where Q represents the vector of query, K represents the vector of key, V represents the vector of value, and d_k represents the dimension of the key.

Changing the computation order of query Q, key K, and value V can significantly reduce the computational complexity. Inspired by this, in this module, input features F (H, W, C) were flattened to F *(N, C)*. Then, linear embedding was used to generate Q, K, and V for each attention head. Each head computed attention scores *G (MS)* and *G (VI)* for KV, which were derived from multispectral branches and vegetation index branches, respectively. After softmax normalization, the attention scores G_{VI} from vegetation index features were cross-multiplied with spectral features Q_{MS}, forcing the vegetation index branch to focus attention on spectral features. Similarly, attention scores G_{MS} from spectral features were cross-multiplied with vegetation index features Q_{VI}, forcing the spectral feature branch to focus attention on vegetation index features. Branches in the two sequences interacted complementarily, thus achieving feature fusion. This can be expressed as follows:

$$G_{MS} = K_{MS}^T V_{MS} \tag{2}$$

$$G_{vI} = K_{vI}^{T} V_{vI} \tag{3}$$

$$E_{MS}(Q_{VI}, K_{MS}, V_{MS}) = \frac{Q_{VI}}{\sqrt{n}} \left(\frac{G_{MS}}{\sqrt{n}} \right) \tag{4}$$

$$E_{VI}(Q_{MS}, K_{VI}, V_{VI}) = \frac{Q_{MS}}{\sqrt{n}} \left(\frac{G_{VI}}{\sqrt{n}} \right) \tag{5}$$

where Q_{MS} and Q_{VI}, respectively, represent the vectors of the spectrum branch and index branch for query, K_{MS} and K_{VI}, respectively, represent the vectors of the spectrum branch and index branch for key, V_{MS} and V_{VI}, respectively, represent the vectors of the spectrum branch and index branch for value, G_{MS} represents the global context vectors of the spectrum branch and index branch, and n represents the dimension of the vectors.

In this module, the multihead mechanism helped allocate appropriate attention weights in different feature subspaces, reducing the transmission and accumulation of redundant information in the network. This enables better capture of useful information for vegetation classification, enhancing the generalization of CNNs across different datasets.

4. Experiments and Results

4.1. Dataset Acquisition and Preprocessing

4.1.1. Study Area

Anhui Province covers a total area of 140,100 square kilometers, with a permanent population of 61.27 million people. Anhui Province is located between east longitude 114°54′ to 119°37′ and north latitude 29°41′ to 34°38′, as shown on the map. Anhui Province is situated in a transitional zone from a warm temperate to subtropical region. It includes 16 prefecture-level cities, and its main geographical features are plains, hills, and mountains. The annual average temperature is 17.1 degrees Celsius, with an average annual precipitation of 978 millimeters. Different climates and social factors may lead to variations in vegetation cover across different regions of Anhui Province. In this study, six random regions were selected for validation in the vegetation-rich areas of central and southern Anhui Province, as shown in Figure 4. The selected area includes different types of vegetation in urban areas, mountainous areas, and rural areas.

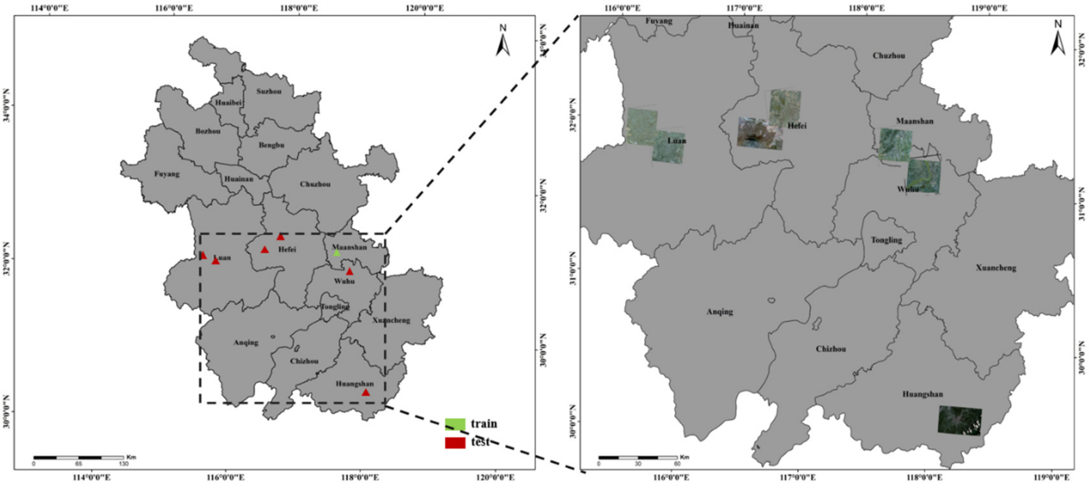

Figure 4. The study area in AnHui, China.

4.1.2. Dataset Acquisition

In this research, high-resolution remote sensing satellite imagery was primarily used. This includes vegetation extraction using GF2, GF7, and PlanetScope data. GF-2 is China's domestically developed civil optical remote sensing satellite with a spatial resolution better than 1 meter, equipped with two high-resolution sensors, one with 1-meter panchromatic and the other with 4-meter multispectral imagery. GF-7 is China's first civil submeter high-resolution optical stereo mapping satellite, providing better than 0.8-meter panchromatic stereo images and 3.2-meter multispectral imagery. PlanetScope is a satellite constellation consisting of approximately 130 satellites with a resolution of 3 meters per pixel. These high-resolution remote sensing images offer rich spectral and spatial information. Image processing for high-resolution data includes atmospheric correction, image fusion, mosaicking, and relative radiometric calibration, among other operations. Planet images undergo orthorectification and radiometric calibration to eliminate the impact of factors such as atmospheric scattering and terrain. Table 2 presents the main parameters of the GF-2, GF-7, and PlanetScope data.

Table 2. Image band dataset detailed specification.

Dataset	Band No.	Band Name	Resolution	Wavelength (nm)
GF-2	Pan	Pan	1	450–900
	B1	Blue	4	450–520
	B2	Green	4	520–590
	B3	Red	4	630–690
	B4	Near-infrared	4	770–890
GF-7	Pan	Pan	≤0.8	450–900
	B1	Blue	≥3.2	450–520
	B2	Green	≥3.2	520–590
	B3	Red	≥3.2	630–690
	B4	Near-infrared	≥3.2	770–890
Planet	B1	Blue	3	465–515
	B2	Green	3	547–593
	B3	Red	3	650–680
	B4	Near-infrared	3	845–885

4.1.3. Data Preprocessing

Considering the richness of vegetation types, this paper, based on GF2 remote sensing satellite imagery, created a training sample set for extracting typical terrain features of vegetation. This involved addressing variations in texture sparsity and spectral characteristics. By comparing samples from different vegetation categories, the paper thoroughly explored typical feature types within the images in conjunction with data quality. The sample library produced in this paper utilized data from two GF2 images in Ma’anshan City, Anhui Province, acquired in October 2022. Pixels were collected from these two images to generate training samples.

Constructing a vegetation sample set corresponding to GF2 remote sensing imagery can be divided into two steps. The first step is to better represent vegetation features using standard false-color images combined with the 4-3-2 bands, where vegetation features are characterized by the color red, delineated through visual interpretation. The second step involves iterative fine adjustments based on the initial delineation. We sampled a total of 393 patches and collected samples within these patches at their interiors and boundaries, resulting in 266,547 vegetation pixels and 435,205 background pixels. Due to GPU limitations, it was not feasible to use the entire image as input for model training. To ensure sample representativeness and model stability, after obtaining sample coordinates, we extended a 5 × 5 neighborhood window around the sampled pixels, with other pixels within the window providing spatial and spectral information to the central pixel. For the validation data, combined with various data source images, we manually visually interpreted the vegetation outlines in the images to obtain real ground truth data.

4.2. Experimental Settings

The model proposed in this study and the comparative methods were experimentally evaluated using the TensorFlow 2.4.0 framework in a Python 3.7 environment. All experiments were conducted on a machine equipped with a 1080Ti graphics card. In this paper, the Adam optimizer was employed to optimize the network. During the training process, all experiments utilized the same hyperparameters, including the number of training epochs (40 epochs), batch size (512), and initial learning rate (0.001). In the classification task of the model, the balanced cross-entropy loss function was adopted to address the issue of class imbalance by introducing weights to different class samples. The specific formula is

$$Balance\ Cross\ Entropy = -\frac{1}{N}\Sigma_{i=1}^{N} w_i(y_i log(p_i) + (1 - y_i) log(1 - p_i)) \tag{6}$$

The binary cross-entropy loss function was chosen, which measures the discrepancy between the probability distribution of model outputs and the true labels, enabling the model to more accurately predict the class of samples during the training process. The specific formula is

$$Binary\ Cross\ Entropy\ Loss = -\frac{1}{N}\Sigma_{i=1}^{N}(y_i \log(p_i) + (1 - y_i)\log(1 - p_i)) \tag{7}$$

where N is the total number of samples, w_i is the weight of sample i, y_i is the true label of sample i, and p_i is the probability predicted by the model that sample i belongs to the positive class.

The activation function of the output layer in the model for classification tasks is the Sigmoid function. It confines the output of neurons within the range of [0, 1], producing outputs similar to probabilities, and is commonly used to interpret the output as the probability of an event occurring.

4.3. Evaluation Metrics

This paper considered vegetation extraction as a binary classification problem, categorizing the prediction results into vegetation and nonvegetation. To validate the effectiveness of the method proposed in this paper for vegetation extraction, four evaluation metrics were used: overall accuracy (*OA*), the *F1* score, intersection over union (*IOU*), and precision. *OA* refers to the ratio of the sum of correctly identified pixels in the test image to the total number of pixels in all identified categories. The *F1* score is the harmonic mean of precision and recall, which can evaluate the model's ability to handle imbalanced datasets or sensitivity to misclassified categories. *IOU* measures the overlap between the model's predicted results and the ground truth annotation regions, commonly used to assess the accuracy and robustness of the model. The kappa coefficient (*Kappa*), an assessment of remote sensing interpretation accuracy, estimates the consistency between predictions and ground references. The formulas for these evaluation metrics are as follows:

$$OA = \frac{TP + TN}{TP + FP + TN + FN} \tag{8}$$

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{11}$$

$$Iou = \frac{TP}{TP + FP + FN} \tag{12}$$

$$p_e = \frac{(TP + FP)\cdot(TP + FN) + (TN + FP)\cdot(TN + FN)}{(TP + FP + TN + FN)^2} \tag{13}$$

$$Kappa = \frac{OA - p_e}{1 - p_e} \tag{14}$$

where true positive (TP) and true negative (TN) represent the number of pixels correctly predicted as positive and negative classes, respectively. False positive (FP) and false negative (FN) represent the number of nonobject pixels incorrectly classified as positive and the number of object pixels incorrectly classified as negative, respectively.

4.4. Experimental Results

By randomly selecting six areas within the study area, the qualitative and quantitative evaluation results of MSICN in these six areas are shown in Figure 5 and Table 3. The results of vegetation extraction in the GF2 image area indicate that our proposed method performs well in densely populated and complex urban areas. In particular, the extraction of boundaries between roads, buildings, shadows, and vegetation coexistence areas is relatively clear and, overall, intact.

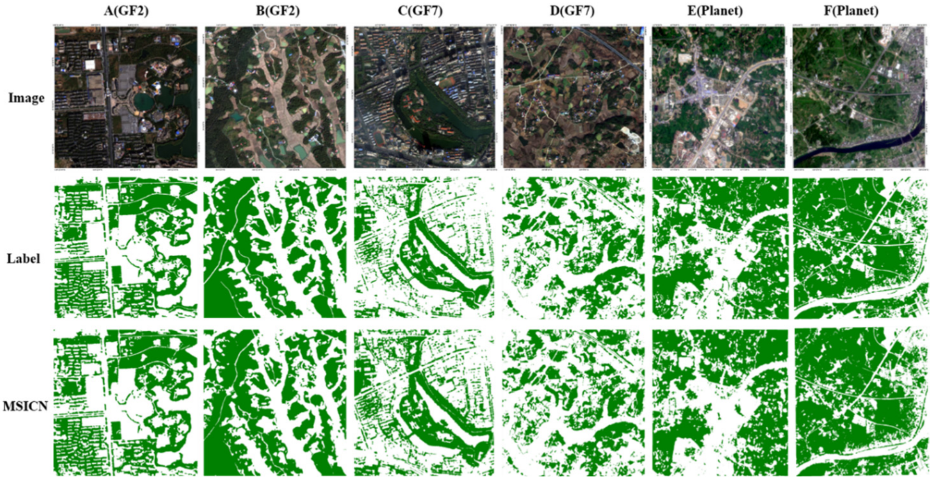

Figure 5. Visualization results of MSICN. Green represents vegetation.

Table 3. Image band dataset detailed specification.

Test Area	Imagery	Method	F1	Precision	IOU	Recall	OA	Kappa
A,B	GF2	MSICN	0.9243	0.9253	0.8646	0.9285	0.9285	0.8558
C,D	GF7	MSICN	0.9133	0.9405	0.8421	0.8734	0.9203	0.8681
E,F	Planet	MSICN	0.9219	0.9135	0.8696	0.9196	0.9182	0.8328

This is because MSICN can integrate multiscale features and effectively distinguish different types of vegetation. In the GF7 and planet image areas, the results show that our proposed method can adapt to vegetation extraction from different data sources, especially in areas with sparse vegetation, achieving good extraction results. This indicates that the auxiliary structures we designed can capture the correlations between features more deeply, effectively integrate vegetation features in different scenarios, help the model consider broader global information when integrating information, and reduce the influence of noise. In summary, our proposed method demonstrates good robustness in vegetation extraction from different data sources in various scenarios.

5. Discussion and Analysis

5.1. Comparative Experimental Results Analysis

To validate the effectiveness of the MSICN network, we conducted comparative analyses by simply thresholding NDVI [9], RFC [12], OCBDL [35], and HRNet [31].

In this study, six regions were randomly selected from GF2, GF7, and planet images for model effectiveness evaluation. These six regions (A–F) were widely distributed in different spatial locations, collected from different images, and have relatively complex backgrounds. The quantitative results for the three data sources are listed in Table 4. Figure 6 displays the qualitative results for various scenes from these three data sources. In Figure 6, green represents vegetation areas.

Table 4. The average accuracy of each method was evaluated on three data sources.

Test Area	Imagery	Method	F1	Precision	IOU	Recall	OA	Kappa
A,B	GF2	NDVI	0.8934	0.8913	0.8343	0.9084	0.9103	0.8433
		RFC	0.8715	0.8213	0.7398	0.8503	0.8640	0.7283
		OSBDL	0.8824	0.8911	0.7911	0.8540	0.8853	0.7649
		HRnet	0.8969	0.8662	0.8139	0.9054	0.9023	0.8032
		MSICN	**0.9243**	**0.9253**	**0.8646**	**0.9285**	**0.9285**	**0.8558**
C,D	GF7	NDVI	0.8207	0.8342	0.7967	0.8136	0.8518	0.7123
		RFC	0.6399	0.6235	0.5526	0.6879	0.7234	0.5330
		OCBDL	0.7905	0.9349	0.7590	0.7672	0.8169	0.7423
		HRnet	0.7798	0.7112	0.7036	0.7289	0.8067	0.6660
		MSICN	**0.9133**	**0.9405**	**0.8421**	**0.8734**	**0.9203**	**0.8681**
E,F	Planet	NDVI	0.8998	0.8740	0.8339	0.8275	0.7868	0.7216
		RFC	0.8170	0.8909	0.6907	0.6952	0.7337	0.5714
		OCBDL	0.8802	0.8852	0.7860	0.7955	0.8190	0.6907
		HRnet	0.9171	0.9465	0.8470	0.8899	0.8579	0.7841
		MSICN	**0.9219**	**0.9135**	**0.8696**	**0.9196**	**0.9182**	**0.8328**

From the visual results, it can be seen that the five methods have certain vegetation extraction capabilities in different validation areas, but there are significant differences in the identification results across different data sources. In the NDVI threshold segmentation method, we repeatedly experimented with manual experience to select the most suitable threshold of 0.46 on GF2 imagery and applied this threshold to other data sources. As shown in Figure 6, on the GF2 imagery, the results extracted by the five methods were relatively accurate. However, on the GF7 and Planet imagery, due to differences from the sample's true label data source, the general applicability of these methods was reduced, resulting in decreased accuracy and completeness of extraction. In contrast, our MSCIN method was better able to detect vegetation information across different data sources, with fewer instances of overextraction and underextraction.

Subfigures (In Figure 6) A(3)–(7) and B(3)–(7) represent the vegetation extraction results in the GF2 image regions. Figure 6 shows that all models performed well in vegetation extraction, and they can extract vegetation quite comprehensively in mountainous areas. However, in some complex urban areas (A(3)–(7)), where urban roads, buildings, and vegetation are mixed, misclassifications are more likely to occur. When compared to other methods, our approach resulted in fewer instances of FPs and FNs, with more complete boundary extraction. Specifically, the IOU, OA, and Kappa coefficient reached 86.46%, 92.85%, and 85.58%, respectively. Subfigures (In Figure 6) C(3)–(7) and F(3)–(7) represent the extraction results in the GF7 and planet image regions. Figure 6 D(4) and F(4) show that RFC exhibits a large number of FNs and FPs on both data images. While NDVI, OCBDL and HRnet provide more complete vegetation extraction compared to RFC, they still exhibit partial FNs and FPs, particularly in sparsely vegetated areas where some omissions occur and extensive FPs occur in bare soil areas. Notably, OSBCL also exhibits the phenomenon of water being misclassified as vegetation patches in the planet images, as seen in Figure 6 F(5). Our method provides relatively complete vegetation extraction with fewer instances of FNs and FPs in both GF7 and planet images. In GF7 images, our method achieved IOU, OA, and kappa coefficients of 84.21%, 92.03%, and 86.81%, respectively. In the planet images, our method achieved IOU, OA, and kappa coefficients of 86.96%, 91.82%, and 83.28%, respectively. This indicates that MSICN maintains high internal consistency and boundary accuracy in vegetation extraction across different data sources, demonstrating strong generalization capabilities.

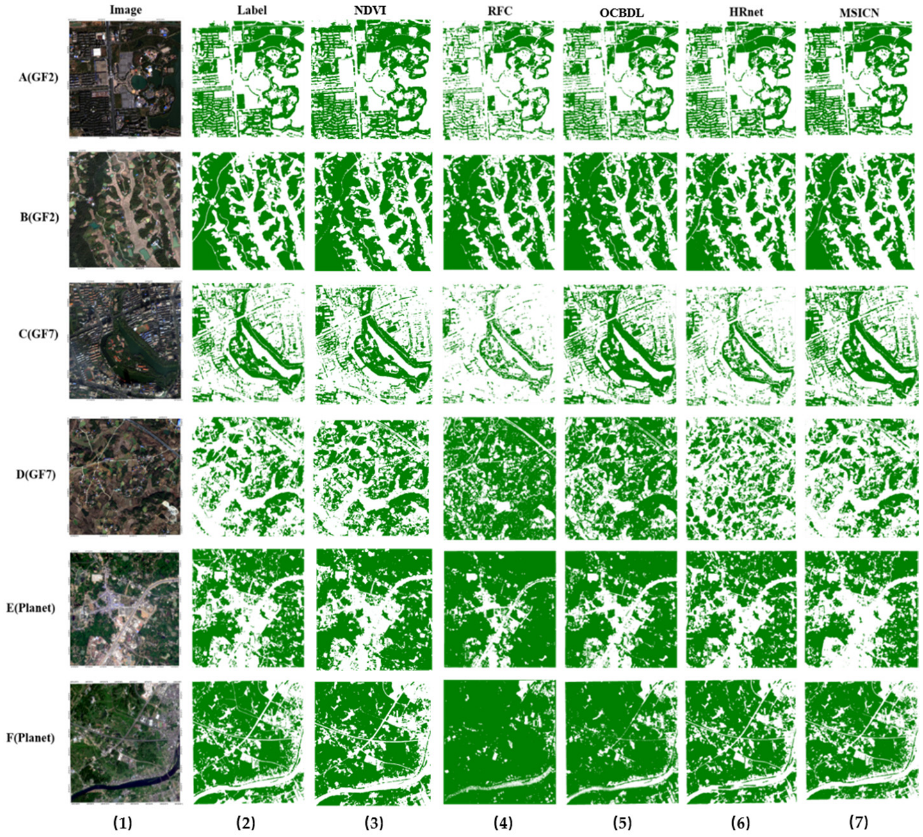

Figure 6. GF2, GF7, and Planet vegetation extraction results. From left to right, they are the original image, ground survey results, and extraction results using NDVI, RFC, OSBCL, HRNet, and MSCIN. Green represents vegetation.

Overall, the experimental results based on the three data sources show that the vegetation extraction results generated by MSICN were superior to the other four methods in terms of quantitative metric evaluation in visual presentation. Classical algorithms do not consider the complementary information between spectral features and vegetation index features, making them prone to omissions and false detections across different data sources. In contrast, MSICN performs complementary and corrective feature fusion, demonstrating good adaptability to different scene conditions. These advantages are attributed to the inclusion of the cross-feature fusion module in MSICN, which considers multilevel features, resulting in more accurate identification results.

5.2. Ablation and Analysis

The above results indicate that our method, by introducing dense connections and dual-path cross-attention mechanisms between spectral and index features, has significantly reduced extraction noise, improved vegetation extraction accuracy, and enhanced model generalization.

5.2.1. The Effectiveness of Vegetation Feature Selection

In this chapter, in order to demonstrate the improvement in vegetation extraction performance after feature selection in this study, we, respectively, input spectral features combined with 12 vegetation index features (NDVI, GNDVI, EVI, OSAVI, RVI, DVI, TVI, GVI, GI, NDGI, and TCARI) and spectral features combined with the remaining 8 unimportant vegetation indices (RVI, DVI, TVI, GVI, GI, NDGI, and TCARI) into the model. Table 5 shows the quantitative results of the two sets of ablation experiments in six regions, and Figure 7 shows the qualitative results of the two sets of ablation experiments.

Table 5. Evaluation of the accuracy of the first group of ablation experiments.

Imagery	Method	F1	IOU	Recall	OA	Kappa
GF2	Spectral+8 indices	0.8634	0.8111	0.8642	0.8613	0.8124
	Spectral+12 indices	0.9131	0.8588	0.9042	0.8921	0.8343
	Ours	**0.9243**	**0.8646**	**0.9285**	**0.9285**	**0.8558**
GF7	Spectral+8 indices	0.8612	0.7617	0.7911	0.8536	0.7008
	Spectral+12 indices	0.8903	0.8357	0.8549	0.8917	0.7731
	Ours	**0.9133**	**0.8421**	**0.8734**	**0.9203**	**0.8681**
Planet	Spectral+8 indices	0.7586	0.7125	0.7269	0.8407	0.6651
	Spectral+12 indices	0.8405	0.7682	0.8701	0.8533	0.7381
	Ours	**0.9219**	**0.8696**	**0.9196**	**0.9182**	**0.8328**

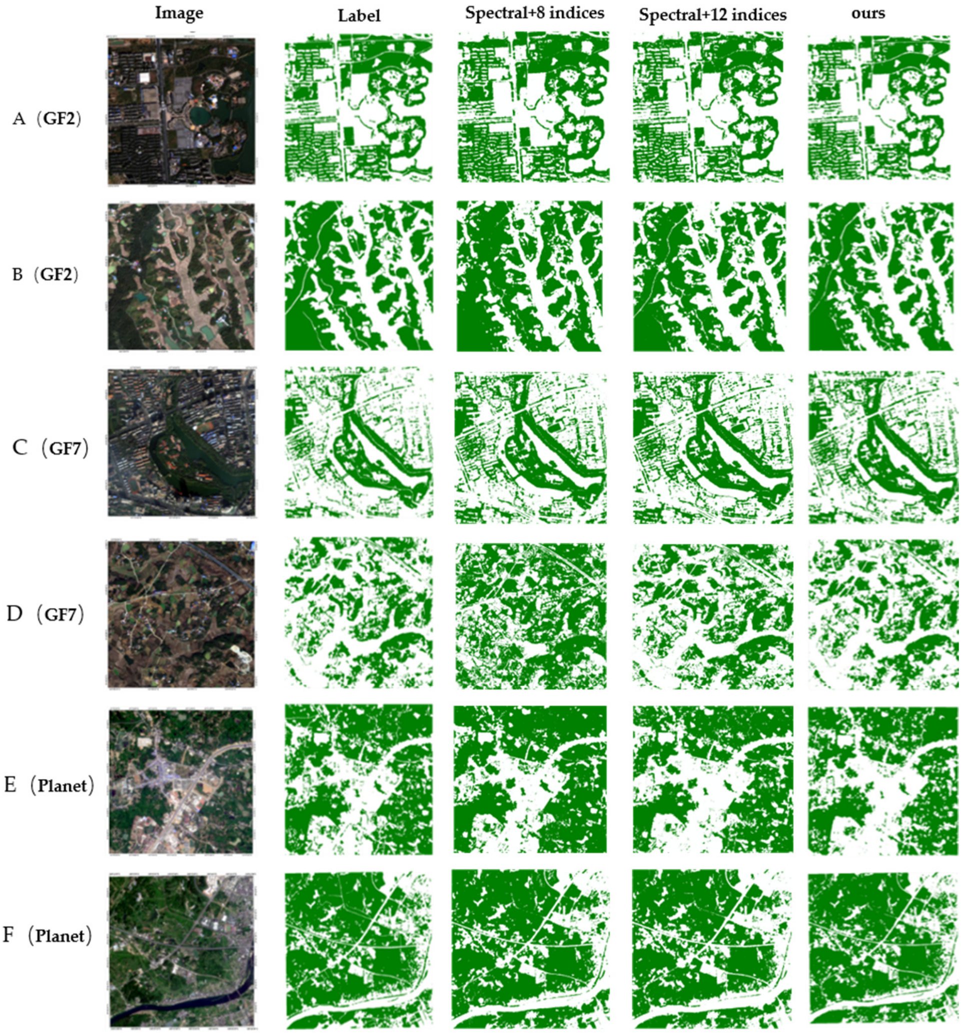

Figure 7. Results of the first group of ablation experiments in the six imaging areas. Green represents vegetation.

In the first set of experiments, we combined the eight indices with low contribution after feature selection with spectral indices into the network. From the quantitative results, the accuracy of vegetation on different data sources decreased. Regarding the GF2 data, the F1, IOU, OA, and Kappa coefficients decreased by 6%, 5%, 6%, 4%; for the GF7 data, they also decreased by 5%, 8%, 7%, 16%; and for the Planet data, they decreased by 16%, 15%, 7%, 17%. In the second set of experiments, we combined the 12 vegetation indices selected by feature selection with spectral features into the network. From the quantitative results, in the GF2 data, the F1, IOU, OA, and Kappa coefficients decreased by 1%, 0.5%, 4%, 2%; in GF7 data, they also decreased by 2%, 0.6%, 3%, 10%; and in the Planet data, they decreased by 8%, 10%, 6%, 9%.

From Figure 7, it can be seen that the combination of spectral information with the eight index features, screened out by feature selection in vegetation monitoring, resulted in blurred boundaries, misclassification of small roads, and misclassification of large bare areas, especially across different data sources. The method of combining spectral information with the 12 vegetation index features also leads to misclassification of building shadows and mixed vegetation areas in complex urban landscapes, with a small amount of underclassification on GF7 and Planet.

The above results indicate that after feature selection, selecting vegetation index features with higher contribution, improves the accuracy of vegetation extraction, making the network better suited for different datasets.

5.2.2. The Effectiveness of Spectral and Index Feature Fusion

In this chapter, to demonstrate the overall effect of spectral and vegetation index fusion on vegetation extraction, we conducted ablation experiments to remove the index feature branch and spectral feature branch separately. Table 6 shows the quantitative evaluation results of the two sets of ablation experiments in six regions, and Figures 8–10 show the qualitative results of the two sets of ablation experiments for three data sources.

Table 6. Evaluation of the accuracy of the second group of ablation experiments.

Imagery	Method	F1	IOU	Recall	OA	Kappa
GF2	only image data	0.8855	0.7958	0.8548	0.8834	0.7629
	only vegetation index	0.8870	0.8005	0.8978	0.8936	0.7831
	Ours	**0.9243**	**0.8646**	**0.9285**	**0.9285**	**0.8558**
GF7	only image data	0.7759	0.6393	0.6471	0.8003	0.6145
	only vegetation index	0.8443	0.7322	0.8205	0.8744	0.7649
	Ours	**0.9133**	**0.8421**	**0.8734**	**0.9203**	**0.8681**
Planet	only image data	0.8164	0.6898	0.6943	0.7327	0.6993
	only vegetation index	0.8007	0.7254	0.8179	0.7723	0.7279
	Ours	**0.9219**	**0.8696**	**0.9196**	**0.9182**	**0.8328**

To validate the effect of spectral feature and vegetation index fusion on the original data, in the first set of experiments, we used only the original image's four-band data as a single input, removing the index feature branch. This resulted in a significant decrease in vegetation extraction accuracy across different data sources. For the GF2 data, the F1 score, IOU, OA, and kappa coefficient decreased by 3%, 7%, 5%, and 9%, respectively. For the GF7 data, the corresponding values decreased by 14%, 20%, 12%, and 25%, and for the planet data, they decreased by 10%, 18%, 19%, and 13%. In the second set of experiments, we used only the vegetation feature branch as input, adopting a pure index-based approach for vegetation extraction. For the GF2 data, the F1, IOU, OA, and kappa coefficient decreased by 4%, 6%, 3%, and 7%, respectively. For the GF7 data, the corresponding values decreased by 7%, 10%, 5%, and 10%, and for the planet data, they decreased by 12%, 14%, 15%, and 10%.

Figure 8 shows that using only image date or solely the vegetation index features resulted in relatively small overall differences in GF2 imagery. The overall performance manifested as incomplete boundaries and internal fragmentation. Figure 9 shows that in the GF7 imagery, there were significant issues with oversegmentation in vegetation extraction results, especially when using only image date, where some bare areas and shadows were misclassified as vegetation. Figure 10 demonstrates that in planet imagery, there were also numerous cases of oversegmentation in vegetation extraction results, with some bare soil, water bodies, and buildings being identified as vegetation and significant internal noise.

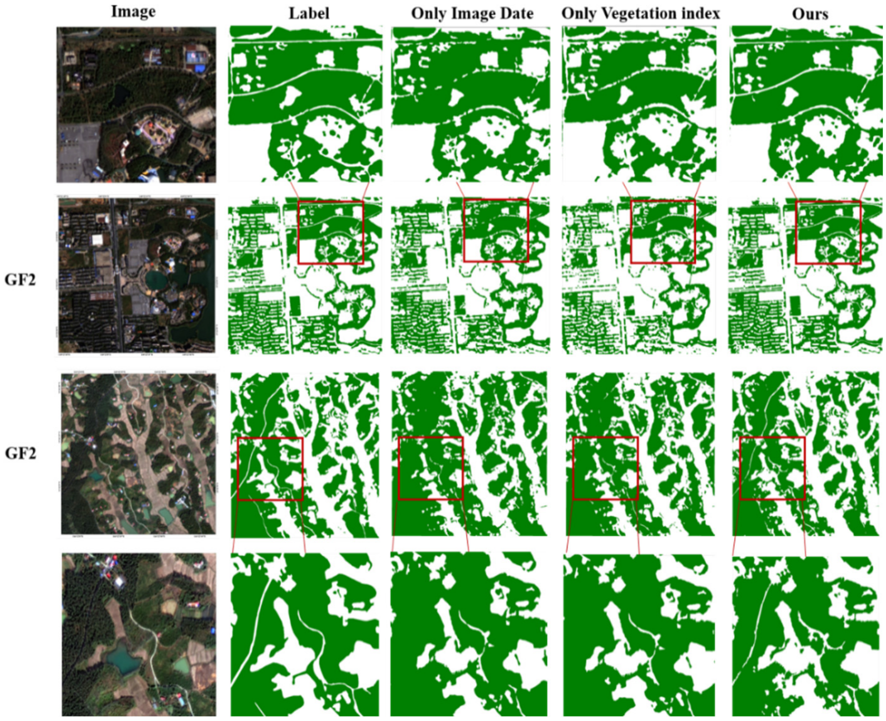

Figure 8. Results of the Second group of ablation experiments in the gf2 imaging area. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

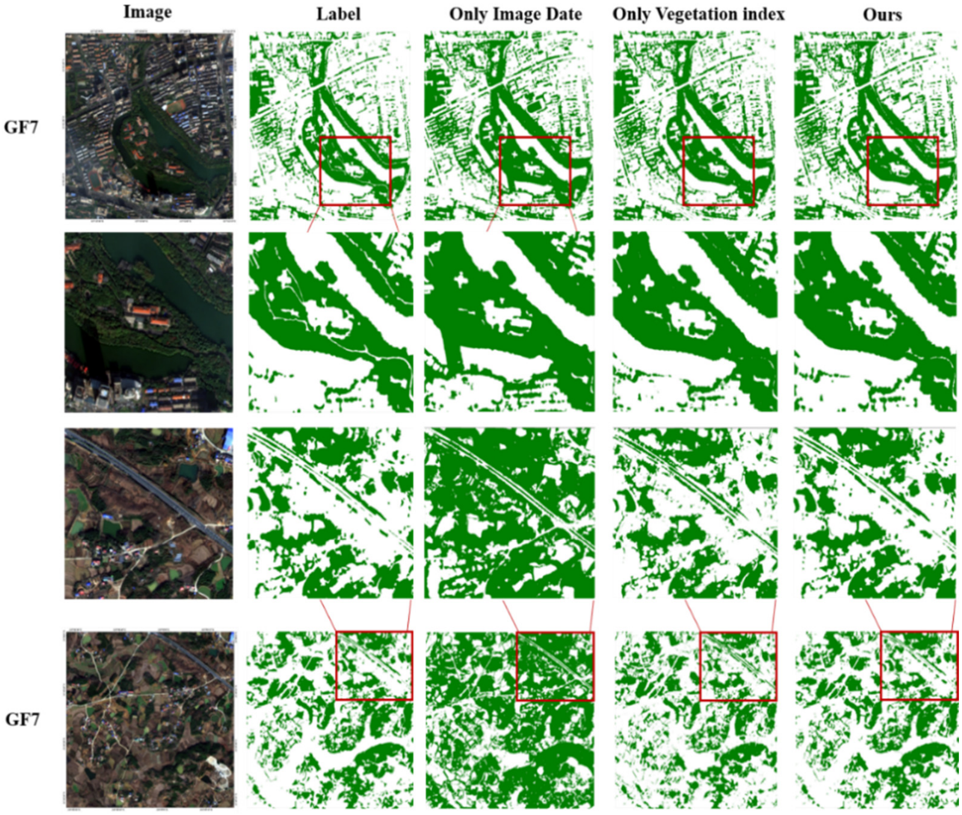

Figure 9. Results of the second group of ablation experiments in the gf7 imaging area. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

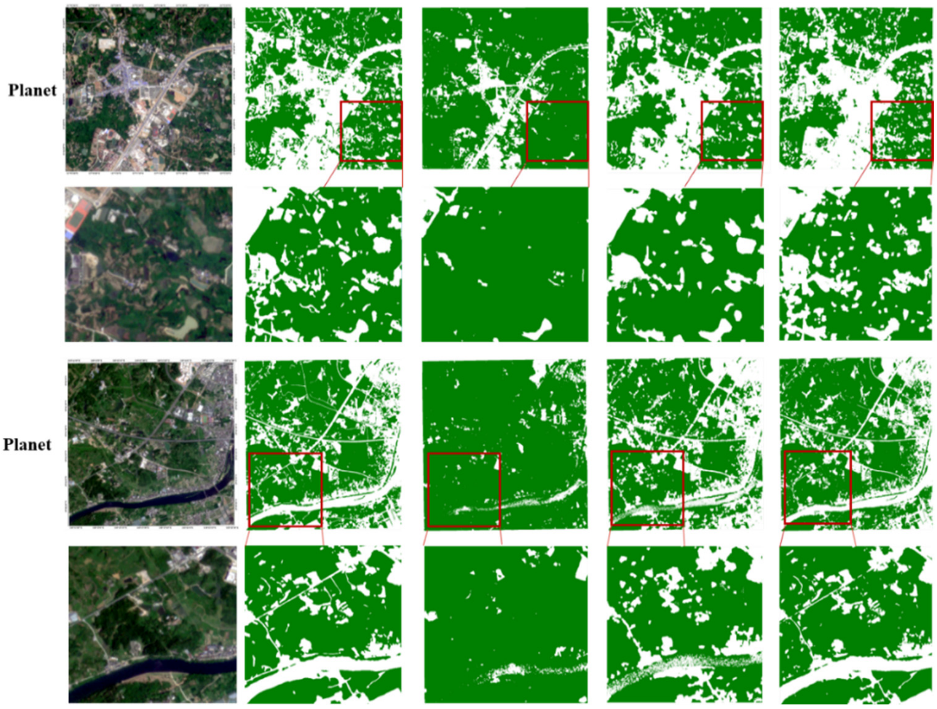

Figure 10. Results of the second group of ablation experiments in the planet imaging area. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

These results indicate that using only spectral image data alone makes it challenging to achieve high-precision vegetation extraction across different data sources, resulting in many cases of both undersegmentation and oversegmentation. While using only the vegetation index feature branch for vegetation extraction can work across different data sources, the overall extraction accuracy is lower compared to the approach of the parallel input of image data and index features. Therefore, in complex background areas of different data sources, it is only through the combination of image data and vegetation index features as dual inputs to the network that high-precision vegetation extraction can be better achieved.

5.2.3. The Role of Simplified Dense Connections and Dual-Path Cross-Attention Mechanisms in the Network

To further explore the enhancement of simplified dense connections and dual-path multihead cross-attention feature fusion modules on vegetation extraction and the effectiveness of model design, we conducted two sets of ablation experiments. Figures 11–13 show the vegetation extraction results of the two sets of ablation experiments, and Table 7 presents the quantitative evaluation results of the two sets of ablation experiments across different data sources.

Table 7. Evaluation of the accuracy of the third group of ablation experiments.

Imagery	Method	F1	IOU	Recall	OA	Kappa
GF2	No simplified dense block	0.8963	0.8052	0.8833	0.8980	0.7937
	No cross-attention feature fusion	0.8869	0.7799	0.8339	0.8783	0.7582
	Ours	**0.9243**	**0.8646**	**0.9285**	**0.9285**	**0.8558**
GF7	No simplified dense block	0.8346	0.7168	0.7573	0.8767	0.7379
	No cross-attention feature fusion	0.7953	0.6673	0.6823	0.8237	0.6524
	Ours	**0.9133**	**0.8421**	**0.8734**	**0.9203**	**0.8681**
Planet	No simplified dense block	0.9014	0.8207	0.8488	0.8747	0.7298
	No cross-attention feature fusion	0.8649	0.7477	0.8281	0.8220	0.6150
	Ours	**0.9219**	**0.8696**	**0.9196**	**0.9182**	**0.8328**

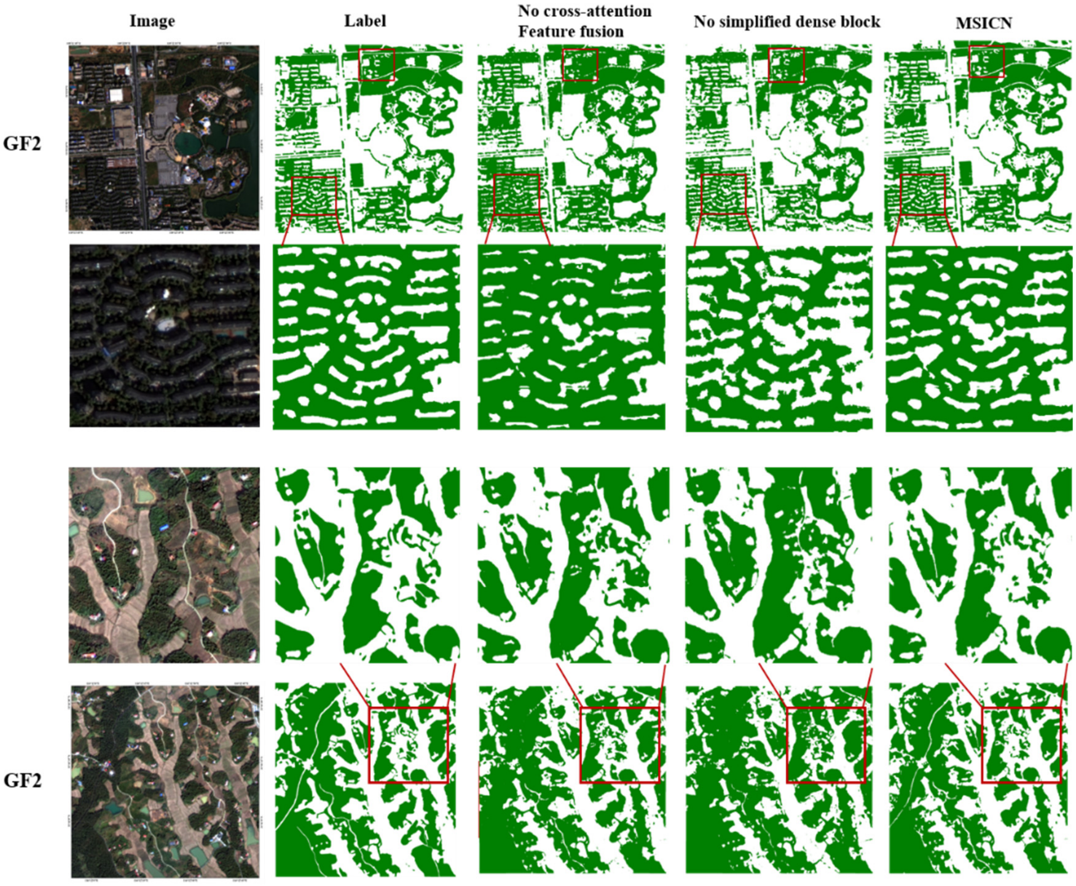

Figure 11. Results of the third group of ablation experiments in the gf2 imaging area. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

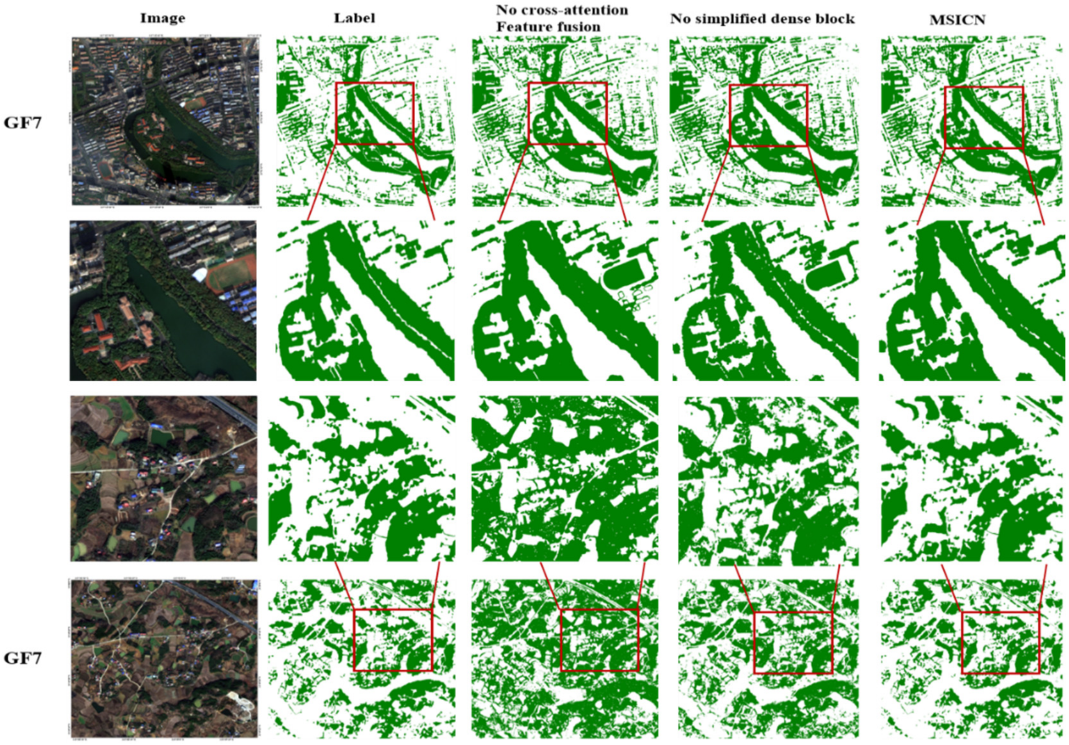

Figure 12. Results of the third group of ablation experiments in the gf7 imaging area. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

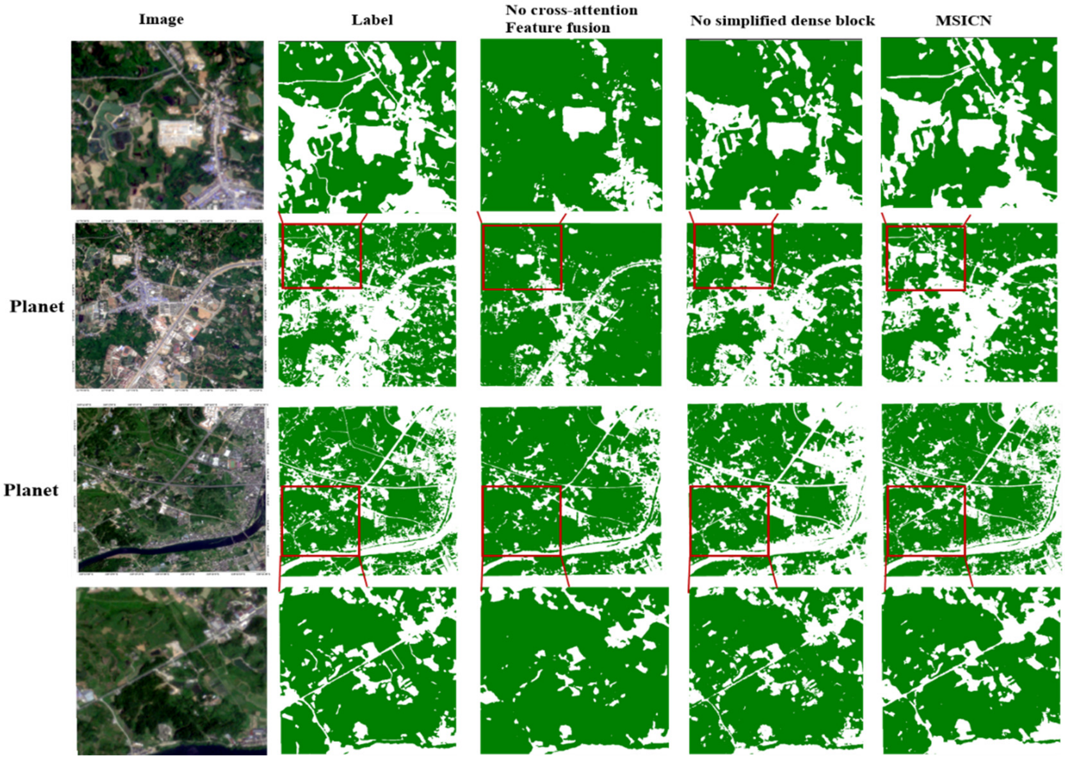

Figure 13. Results of the third group of ablation experiments in the planet imaging area. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

To validate the effectiveness of the dual-path multihead cross-attention feature fusion module, we conducted the first set of ablation experiments by removing the feature fusion module. In the GF2 image region, the F1, IOU, OA, and Kappa coefficients decreased by 4%, 8%, 5%, and 1%, respectively. In the GF7 image region, these metrics decreased by 12%, 17%, 10%, and 24%, and in the planet image region, they decreased by 6%, 12%, 9%, 22%, performing even worse than other models. Details in the magnified red rectangles in Figures 11–13 show that in different data sources, recognition results without multihead cross-attention feature fusion exhibited more FPs and incomplete boundaries compared to MSICN, especially in the GF7 and planet image regions. For example, in the GF2 image region, in complex urban areas, vegetation under large building shadows could not be accurately extracted after removing the multihead cross-attention feature fusion. In the GF7 and planet image regions, the results of removing the dual-path multihead cross-attention feature fusion included many misclassifications of farmland as vegetation. This indicates that the dual-path multihead cross-attention feature fusion module is an indispensable part of MSICN, playing a crucial role in maintaining the internal consistency of recognition results and accurate extraction across different data sources.

To validate the effectiveness of the simplified dense connections, we conducted the second set of ablation experiments by removing the dense connection module. In the GF2 image region, the removal of the simplified dense connections resulted in a decrease of 3% in F1, 6% in IOU, 5% in Recall, 3% in OA, and 6% in Kappa coefficients. In the GF7 image region, the removal of simplified dense connections led to a decrease of 8% in F1, 4% in IOU, 2% in recall, 4% in OA, and 13% in Kappa coefficients. In the planet image region, the removal of simplified dense connections resulted in a decrease of 2% in F1, 5% in IOU, 7% in Recall, 4% in OA, and 1% in Kappa coefficients. Details in the magnified rectangles in Figures 11–13 show that the removal of simplified dense connections in MSICN led to FPs, internal fragmentation, and incomplete boundaries in different data sources.

Through the above validation and analysis, we can see that due to the simplified dense connections, the deep learning neural network's ability to extract vegetation features is enhanced. The dual-path multi-head cross-attention feature fusion allows the model to better understand the correlation between different features, extract rich information from them, and simultaneously focus on pixel-level spatial information and vegetation semantic features. This enables the model to better handle the accurate identification of vegetation in complex backgrounds, resist noise and changes caused by factors such as lighting and seasonal variations, and improve the accuracy and robustness of vegetation extraction.

5.3. Universal Validation Analysis

To validate the universality of our approach, we selected three larger regions within the study area for verification and analysis. Table 8 presents the quantitative results, and Figure 14 provides the qualitative results of vegetation extraction for the three data sources. Figure 15 displays detailed vegetation extraction results. From the quantitative results, it can be observed that our method outperforms other methods in terms of accuracy when extracting vegetation across different data sources and regions.

Table 8. Visualized results of universality verification of different data sources.

Imagery	Method	F1	Precision	IOU	Recall	OA	Kappa
GF2	RFC	0.7720	0.7546	0.6887	0.8109	0.8237	0.7283
	OCBDL	0.8502	0.7908	0.7395	**0.9014**	0.8530	0.7411
	HRet	0.8479	0.8217	0.7666	0.8594	0.8859	0.7680
	MSICN	**0.8862**	**0.8842**	**0.7957**	0.8882	**0.8965**	**0.8113**
GF7	RFC	0.6318	0.6617	0.5618	0.6704	0.7103	0.5217
	OCBDL	0.7211	0.9145	0.7590	0.7155	0.7325	0.7390
	HRet	0.8339	0.7704	0.7564	0.7010	0.8307	0.7219
	MSICN	**0.9058**	**0.8805**	**0.8972**	**0.8995**	**0.9072**	**0.9167**
Planet	RFC	0.7745	0.7868	0.7770	0.7851	0.7106	0.6778
	OCBDL	0.8109	0.8843	0.7820	0.7927	0.7224	0.6923
	HRnet	0.8241	**0.9465**	0.8108	0.8669	0.8579	0.7841
	MSICN	**0.8964**	0.8883	**0.8740**	**0.8948**	**0.8702**	**0.8258**

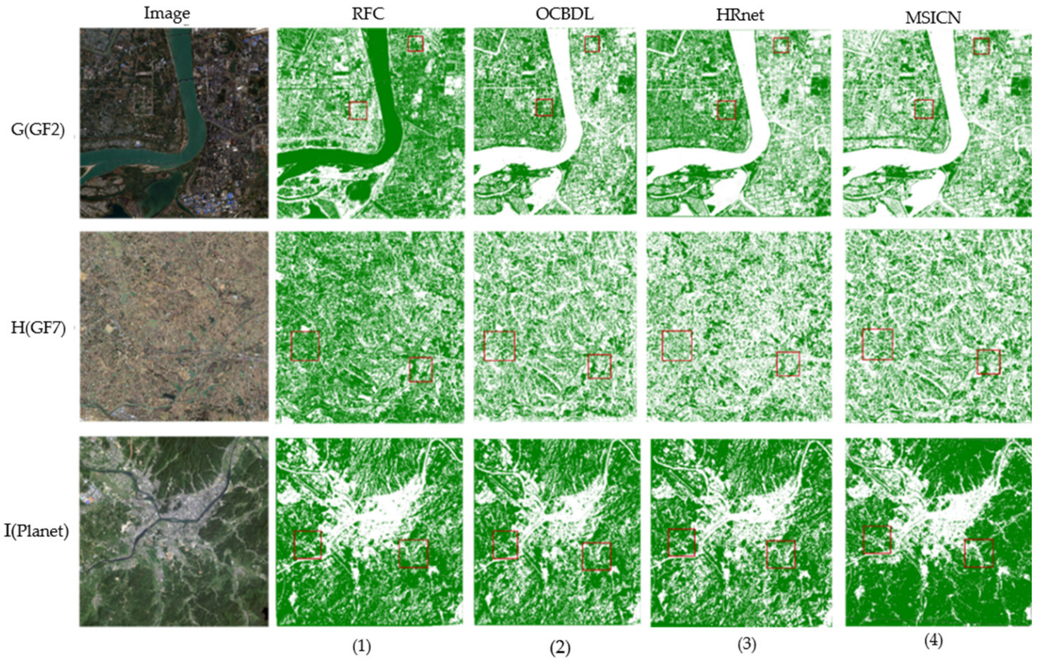

Figure 14. Universal verification area extraction results. Green represents vegetation. The red box represents a zoomed in detail image of the local area.

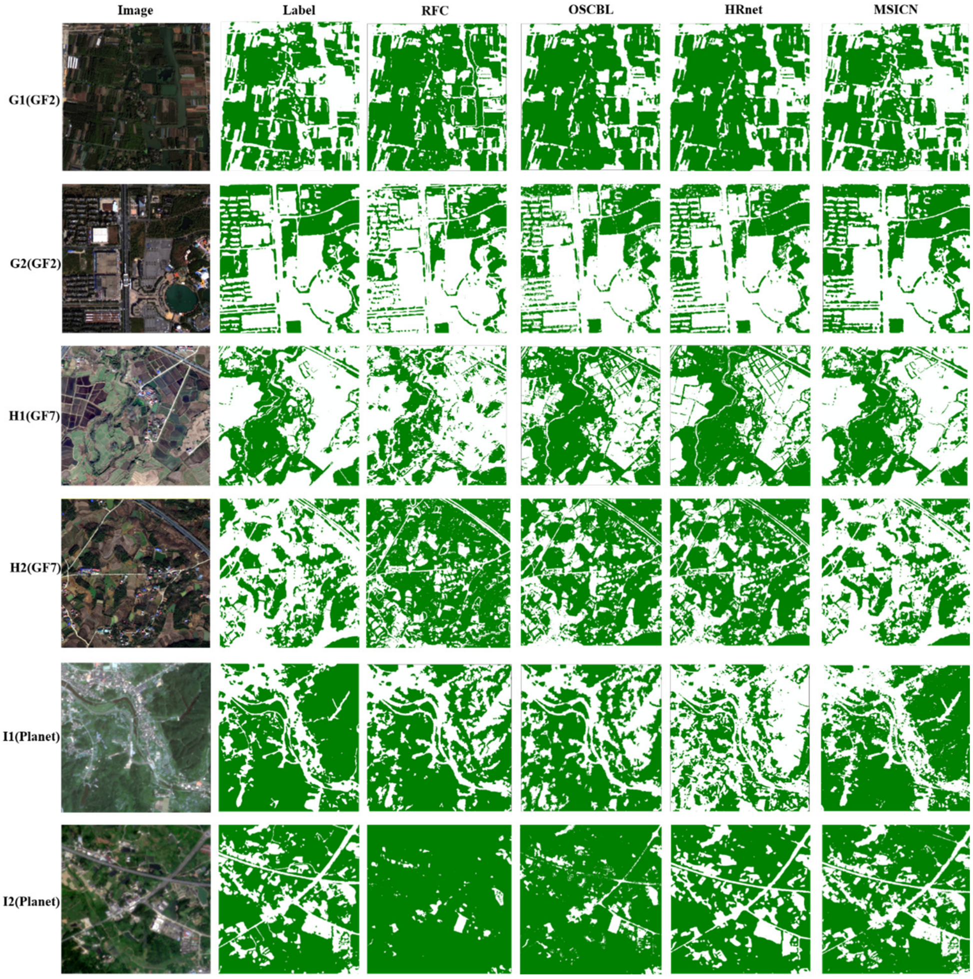

Figure 15. Universal verification area extraction results. Green represents vegetation.

Due to differences in feature values between different sensors, other methods often suffer from a significant number of omissions and FPs when applied to slightly larger regions across different data sources. As shown in Subfigure G(1) (In Figure 15), the random forest method misclassified water bodies as vegetation, and in Subfigures G(1)–(3), H(1)–(3), and I(1)–(3) (In Figure 15), there were numerous instances where bare soil areas were incorrectly identified as vegetation. In densely vegetated mountainous areas, there were also considerable omissions of vegetation shadows. Our method, which combines vegetation index features and incorporates auxiliary modules to enhance deep feature interaction, to some extent, reduces the differences in feature values. This allows the model to maintain robustness in larger regions across different data sources. As shown in Subfigures G(4), H(4), and I(4) (In Figure 15), our method achieved good vegetation extraction results in all three data sources, with fewer instances of omissions and FPs.

6. Conclusions

To overcome the challenges of achieving high-precision remote sensing vegetation extraction across wide-ranging and diverse data sources and to address the issues of omission and misclassification among different remote sensing data sources, this paper proposed a multifeature integrated perception method for high-resolution remote sensing vegetation extraction to enhance the ability to capture vegetation features and improve the vegetation extraction accuracy by fusing spectral features and vegetation index features. A multifeature integrated perception convolutional network (MSCIN) was constructed, which enhances multiscale feature information, global information interaction, and feature cross-fusion. In this model, spectral features and vegetation index features selected through a random forest model are used as dual parallel inputs to reduce differences in land feature values between different sensors. Additionally, a multiscale convolutional module was constructed to capture spectral and index features in the target pixel domain. A simplified dense connection and multihead cross-feature fusion module were designed to strengthen the global information interaction among multiple features, enabling the multifeature expression, enhancement, fusion, and extraction of vegetation. To validate the vegetation extraction capability of this method across different remote sensing data sources, the model, trained with high-resolution satellite images from GF-2, was tested on test areas selected from GF-2, GF-7, and planet images in different regions. A comparison with three typical methods, NDVI, RFC, OCBDL, and HRNet, was conducted. The results show that compared to other methods, this paper's method achieved robust vegetation extraction with the ability to overcome issues such as internal fragmentation and unclear boundaries. It also reduced omission and commission errors in vegetation regions, exhibiting superior performance metrics, including the F1 score, IoU, recall, OA, and Kappa coefficient. Two sets of ablation experiments showed the critical role played by the fusion of vegetation index features and spectral features, as well as the simplified dense connection and dual-path cross-attention mechanism in enhancing vegetation extraction precision and completeness. The method proposed in this paper demonstrates high adaptability and generalization capabilities in vegetation extraction, enabling high-precision vegetation extraction across single data source samples.

Author Contributions: Conceptualization, Y.L., S.M. and B.S.; data curation, Y.L., S.M., B.S., B.W. and Y.W.; funding acquisition, H.Y. and B.W.; investigation, S.M., B.S., B.W. and Y.W.; methodology, Y.L.; resources, H.Y.; validation, Y.L., S.M. and B.W.; visualization, Y.W.; writing—original draft, Y.L.; writing—review and editing, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of China (grant No.42101381, 41901282 and 41971311), the International Science and Technology Cooperation Special (grant number 202104b11020022), the Anhui Provincial Natural Science Foundation, (grant number 2308085MD126), and the Hefei Municipal Natural Science Foundation (grant number 202323). (Corresponding author: Hui Yang).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to the permissions issues.

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

References

1. Wu, J.; Zhang, L.; Zhao, B.; Yang, N.; Gao, P. Remote sensing monitoring of vegetation and its resilience based on the critical slowdown model and GLASS LAI: A case study of the three gorges reservoir area. *Acta Ecol. Sin.* **2023**, *12*, 1–12.
2. Guo, J.; Xu, Q.; Zeng, Y.; Liu, Z.; Zhu, X.X. Nationwide urban tree canopy mapping and coverage assessment in Brazil from high-resolution remote sensing images using deep learning. *ISPRS J. Photogramm. Remote Sens.* **2023**, *198*, 1–15. [CrossRef]
3. Zhang, Z.; Liu, X.; Zhu, L.; Li, J.; Zhang, Y. Remote Sensing Extraction Method of *Illicium verum* Based on Functional Characteristics of Vegetation Canopy. *Remote Sens.* **2022**, *14*, 6248. [CrossRef]
4. Askam, E.; Nagisetty, R.M.; Crowley, J.; Bobst, A.L.; Shaw, G.; Fortune, J. Satellite and sUAS Multispectral Remote Sensing Analysis of Vegetation Response to Beaver Mimicry Restoration on Blacktail Creek, Southwest Montana. *Remote Sens.* **2022**, *14*, 6199. [CrossRef]
5. Zhang, Y.; Wang, H.; Li, H.; Sun, J.; Liu, H.; Yin, Y. Optimization Model of Signal-to-Noise Ratio for a Typical Polarization Multispectral Imaging Remote Sensor. *Sensors* **2022**, *22*, 6624. [CrossRef]
6. Wang, R.; Shi, F.; Xu, D. The Extraction Method of Alfalfa (*Medicago sativa* L.) Mapping Using Different Remote Sensing Data Sources Based on Vegetation Growth Properties. *Land* **2022**, *11*, 1996. [CrossRef]
7. Zhang, J.; Li, J.; Wang, X.; Wu, P.; Liu, X.; Ji, Y. Remote sensing imaging: A useful method for assessing wetland vegetation evolution processes in the Nanjishan Wetland National Nature Reserve, Lake Poyang. *IOP Conf. Ser. Earth Environ. Sci.* **2019**, *349*, 012011. [CrossRef]

8. Moesinger, L.; Zotta, R.-M.; van der Schalie, R.; Scanlon, T.; de Jeu, R.; Dorigo, W. Monitoring vegetation condition using microwave remote sensing: The standardized vegetation optical depth index (SVODI). *Biogeosciences* **2022**, *19*, 5107–5123. [CrossRef]
9. El-Mezouar, M.C.; Taleb, N.; Kpalma, K.; Ronsin, J. Vegetation extraction from IKONOS imagery using high spatial resolution index. *J. Appl. Remote Sens.* **2011**, *5*, 053543. [CrossRef]
10. Yao, F.; Luo, J.; Shen, Z.; Dong, D.; Yang, K. Automatic urban vegetation extraction method using high resolution imagery. *J. Geo Inf. Sci.* **2018**, *18*, 248–254.
11. Li, W.; Fu, H.; Yu, L.; Gong, P.; Feng, D.; Li, C.; Clinton, N. Stacked Autoencoder-based deep learning for remote-sensing image classification: A case study of African land-cover mapping. *Int. J. Remote Sens.* **2016**, *37*, 5632–5646. [CrossRef]
12. Feng, S.; Zhao, J.; Liu, T.; Zhang, H.; Zhang, Z.; Guo, X. Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2019**, *12*, 3295–3306. [CrossRef]
13. Xi, Y.; Ren, C.; Tian, Q.; Ren, Y.; Dong, X.; Zhang, Z. Exploitation of Time Series Sentinel-2 Data and Different Machine Learning Algorithms for Detailed Tree Species Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 7589–7603. [CrossRef]
14. Gašparović, M.; Dobrinić, D. Comparative Assessment of Machine Learning Methods for Urban Vegetation Mapping Using Multitemporal Sentinel-1 Imagery. *Remote Sens.* **2020**, *12*, 1952. [CrossRef]
15. Hoeser, T.; Kuenzer, C. Object Detection and Image Segmentation with Deep Learning on Earth Observation Data: A Review-Part I: Evolution and Recent Trends. *Remote Sens.* **2020**, *12*, 1667. [CrossRef]
16. He, H.; Qian, H.; Xie, L.; Duan, P. Interchange recognition method based on CNN. *Acta Geod. Cartogr. Sin.* **2018**, *47*, 385.
17. Barbosa, A.; Trevisan, R.; Hovakimyan, N.; Martin, N.F. Modeling yield response to crop management using convolutional neural networks. *Comput. Electron. Agric.* **2020**, *170*, 105197. [CrossRef]
18. Fricker, G.A.; Ventura, J.D.; Wolf, J.A.; North, M.P.; Davis, F.W.; Franklin, J. A Convolutional Neural Network Classifier Identifies Tree Species in Mixed-Conifer Forest from Hyperspectral Imagery. *Remote Sens.* **2019**, *11*, 2326. [CrossRef]
19. Torres, D.L.; Feitosa, R.Q.; Happ, P.N.; La Rosa, L.E.C.; Junior, J.M.; Martins, J.; Bressan, P.O.; Gonçalves, W.N.; Liesenberg, V. Applying Fully Convolutional Architectures for Semantic Segmentation of a Single Tree Species in Urban Environment on High Resolution UAV Optical Imagery. *Sensors* **2020**, *20*, 563. [CrossRef] [PubMed]
20. Freudenberg, M.; Nölke, N.; Agostini, A.; Urban, K.; Wörgötter, F.; Kleinn, C. Large Scale Palm Tree Detection in High Resolution Satellite Images Using U-Net. *Remote Sens.* **2019**, *11*, 312. [CrossRef]
21. López-Jiménez, E.; Vasquez-Gomez, J.I.; Sanchez-Acevedo, M.A.; Herrera-Lozada, J.C.; Uriarte-Arcia, A.V. Columnar cactus recognition in aerial images using a deep learning approach. *Ecol. Inform.* **2019**, *52*, 131–138. [CrossRef]
22. Liu, R.; Lehman, J.; Molino, P.; Petroski Such, F.; Frank, E.; Sergeev, A.; Yosinski, J. An intriguing failing of convolutional neural networks and the coordconv solution. In Proceedings of the 32nd Annual Conference on Neural Information Processing Systems (NeurIPS2018), Montréal, QC, Canada, 2–8 December 2018; pp. 9628–9639.
23. Zou, J.; Dado, W.T.; Pan, R. Early Crop Type Image Segmentation from Satellite Radar Imagery. 2020. Available online: https://api.semanticscholar.org/corpusid:234353421 (accessed on 25 December 2023).
24. Mo, Y.; Wu, Y.; Yang, X.; Liu, F.; Liao, Y. Review the state-of-the-art technologies of semantic segmentation based on deep learning. *Neurocomputing* **2022**, *493*, 626–646. [CrossRef]
25. Rußwurm, M.; Körner, M. Multi-Temporal Land Cover Classification with Sequential Recurrent Encoders. *ISPRS Int. J. Geo-Inf.* **2018**, *7*, 129. [CrossRef]
26. Sun, W.; Wang, R. Fully Convolutional Networks for Semantic Segmentation of Very High Resolution Remotely Sensed Images Combined With DSM. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 474–478. [CrossRef]
27. Kattenborn, T.; Leitloff, J.; Schiefer, F.; Hinz, S. Review on Convolutional Neural Networks (CNN) in vegetation remote sensing. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 24–49. [CrossRef]
28. Jin, X.; Bagavathiannan, M.; Maity, A.; Chen, Y.; Yu, J. Deep learning for detecting herbicide weed control spectrum in turfgrass. *Plant Methods* **2022**, *18*, 1–11. [CrossRef]
29. de Camargo, T.; Schirrmann, M.; Landwehr, N.; Dammer, K.-H.; Pflanz, M. Optimized Deep Learning Model as a Basis for Fast UAV Mapping of Weed Species in Winter Wheat Crops. *Remote Sens.* **2021**, *13*, 1704. [CrossRef]
30. Jegou, S.; Drozdzal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The one hundred layers tiramisu: Fully convolutional DenseNets for semantic segmentation. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 1175–1183.
31. Xu, Z.; Zhou, Y.; Wang, S.; Wang, L.; Li, F.; Wang, S.; Wang, Z. A Novel Intelligent Classification Method for Urban Green Space Based on High-Resolution Remote Sensing Images. *Remote Sens.* **2020**, *12*, 3845. [CrossRef]
32. Chen, L.-C.; Hermans, A.; Papandreou, G.; Schroff, F.; Wang, P.; Adam, H. MaskLab: Instance Segmentation by Refining Object Detection with Semantic and Direction Features. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4013–4022.
33. Mazzia, V.; Khaliq, A.; Chiaberge, M. Improvement in Land Cover and Crop Classification based on Temporal Features Learning from Sentinel-2 Data Using Recurrent-Convolutional Neural Network (R-CNN). *Appl. Sci.* **2019**, *10*, 238. [CrossRef]
34. Chen, S.T.; Yu, H. High resolution remote sensing image classification based on multi-scale and multi-feature fusion. *Chin. J. Quantum Electron.* **2016**, *33*, 420–426.

35. Radke, D.; Radke, D.; Radke, J. Beyond measurement: Extracting vegetation height from high resolution imagery with deep learning. *Remote Sens.* **2020**, *12*, 3797. [CrossRef]
36. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
37. Kalita, I.; Kumar, R.N.S.; Roy, M. Deep learning-based cross-sensor domain adaptation under active learning for land cover classification. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
38. Dobrinić, D.; Gašparović, M.; Medak, D. Evaluation of Feature Selection Methods for Vegetation Mapping Using Multitemporal Sentinel Imagery. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2022**, *43*, 485–491. [CrossRef]
39. Luo, C.; Meng, S.; Hu, X.; Wang, X.; Zhong, Y. Cropnet: Deep spatial-temporal-spectral feature learning network for crop classification from time-series multi-spectral images. In Proceedings of the IGARSS 2020—2020 IEEE International Geoscience and Remote Sensing Symposium, Waikoloa, HI, USA, 26 September–2 October 2020; pp. 4187–4190.
40. Ju, Y.; Bohrer, G. Classification of wetland vegetation based on NDVI time series from the HLS dataset. *Remote Sens.* **2022**, *14*, 2107. [CrossRef]
41. Fang, S.; Li, K.; Shao, J.; Li, Z. SNUNet-CD: A densely connected Siamese network for change detection of VHR images. *IEEE Geosci. Remote Sens. Lett.* **2021**, *19*, 1–5. [CrossRef]
42. Liu, Y.; Pang, C.; Zhan, Z.; Zhang, X.; Yang, X. Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model. *IEEE Geosci. Remote Sens. Lett.* **2020**, *18*, 811–815. [CrossRef]
43. Huang, Z.; Wang, X.; Huang, L.; Huang, C.; Wei, Y.; Liu, W. Ccnet: Criss-cross attention for semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 603–612.
44. Lee, P.H.; Yu, P.L. Distance-based tree models for ranking data. *Comput. Stat. Data Anal.* **2010**, *54*, 1672–1682. [CrossRef]
45. Zeng, F.; Yang, B.; Zhao, M.; Xing, Y.; Ma, Y. MASANet: Multi-Angle Self-Attention Network for Semantic Segmentation of Remote Sensing Images. *Teh. Vjesn.* **2022**, *29*, 1567–1575.
46. Abraham, A.; Pedregosa, F.; Eickenberg, M.; Gervais, P.; Mueller, A.; Kossaifi, J.; Gramfort, A.; Thirion, B.; Varoquaux, G. Machine learning for neuroimaging with scikit-learn. *Front. Neurosci.* **2014**, *8*, 14. [CrossRef]
47. Wójcik-Gront, E.; Gozdowski, D.; Stępień, W. UAV-Derived Spectral Indices for the Evaluation of the Condition of Rye in Long-Term Field Experiments. *Agriculture* **2022**, *12*, 1671. [CrossRef]
48. Bernardes, T.; Moreira, M.A.; Adami, M.; Giarolla, A.; Rudorff, B.F.T. Monitoring Biennial Bearing Effect on Coffee Yield Using MODIS Remote Sensing Imagery. *Remote Sens.* **2012**, *4*, 2492–2509. [CrossRef]
49. Manna, S.; Mondal, P.P.; Mukhopadhyay, A.; Akhand, A.; Harza, S.; Mitra, D. Vegetation cover change analysis from multi-temporal satellite data in Jharkhali Island, Sundarbans, India. *Indian J. Geo Mar. Sci.* **2013**, *42*, 331–342.
50. Szigarski, C.; Jagdhuber, T.; Baur, M.; Thiel, C.; Parrens, M.; Wigneron, J.-P.; Piles, M.; Entekhabi, D. Analysis of the Radar Vegetation Index and Potential Improvements. *Remote Sens.* **2018**, *10*, 1776. [CrossRef]
51. Naji, T.A.H. Study of vegetation cover distribution using DVI, PVI, WDVI indices with 2D-space plot. *J. Phys. Conf. Ser.* **2018**, *1003*, 012083. [CrossRef]
52. Meivel, S.; Maheswari, S. Remote Sensing Analysis of Agricultural Drone. *J. Indian Soc. Remote Sens.* **2020**, *49*, 689–701. [CrossRef]
53. Suppakittpaisarn, P.; Jiang, B.; Slavenas, M.; Sullivan, W.C. Does density of green infrastructure predict preference? *Urban For. Urban Green.* **2018**, *40*, 236–244. [CrossRef]
54. Nedkov, R. Normalized differential greenness index for vegetation dynamics assessment. *Comptes Rendus L'Academie Bulg. Sci.* **2017**, *70*, 1143–1146.
55. Meng, Q.Y.; Dong, H.; Qin, Q.M.; Wang, J.L.; Zhao, J.H. MTCARIA kind of vegetation index monitoring vegetation leaf chlorophyll content based on hyperspectral remote sensing. *Spectrosc. Spectr. Anal.* **2012**, *32*, 2218–2222. [CrossRef]
56. Ingram, R.E. Self-focused attention in clinical disorders: Review and a conceptual model. *Psychol. Bull.* **1990**, *107*, 156–176. [CrossRef]
57. Wu, B.; Huang, W.; Ye, H.; Luo, P.; Ren, Y.; Kong, W. Using multi-angular hyperspectral data to estimate the vertical distribution of leaf chlorophyll content in wheat. *Remote Sens.* **2021**, *13*, 1501. [CrossRef]

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.