remote sensing

MDPI

Article

Fusion Classification of HSI and MSI Using a Spatial-Spectral Vision Transformer for Wetland Biodiversity Estimation

Yunhao Gao [1], Xiukai Song [2,*], Wei Li [1], Jianbu Wang [3], Jianlong He [2], Xiangyang Jiang [2] and Yinyin Feng [2]

[1] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China; gaoyunhao@bit.edu.cn (Y.G.); liwei089@ieee.org (W.L.)
[2] Shandong Provincial Key Laboratory of Restoration for Marine Ecology, Shandong Marine Resources and Environment Research Institute, Yantai 264006, China; hejianlong@shandong.cn (J.H.); jiangxiangyang@shandong.cn (X.J.); fengyinyin@shandong.cn (Y.F.)
[3] Lab of the Marine Physics and Remote Sensing, First Institute of Oceanography, Ministry of Natural Resources, Qingdao 266061, China; wangjianbu@fio.org.cn
* Correspondence: songxiukai@shandong.cn

Abstract: The rapid development of remote sensing technology provides wealthy data for earth observation. Land-cover mapping indirectly achieves biodiversity estimation at a coarse scale. Therefore, accurate land-cover mapping is the precondition of biodiversity estimation. However, the environment of the wetlands is complex, and the vegetation is mixed and patchy, so the land-cover recognition based on remote sensing is full of challenges. This paper constructs a systematic framework for multisource remote sensing image processing. Firstly, the hyperspectral image (HSI) and multispectral image (MSI) are fused by the CNN-based method to obtain the fused image with high spatial-spectral resolution. Secondly, considering the sequentiality of spatial distribution and spectral response, the spatial-spectral vision transformer (SSViT) is designed to extract sequential relationships from the fused images. After that, an external attention module is utilized for feature integration, and then the pixel-wise prediction is achieved for land-cover mapping. Finally, land-cover mapping and benthos data at the sites are analyzed consistently to reveal the distribution rule of benthos. Experiments on ZiYuan1-02D data of the Yellow River estuary wetland are conducted to demonstrate the effectiveness of the proposed framework compared with several related methods.

Keywords: coastal wetlands; multisource remote sensing; land-cover mapping; biodiversity estimation; spatial-spectral vision transformer

check for updates

Citation: Gao, Y.; Song, X.; Li, W.; Wang, J.; He, J.; Jiang, X.; Feng, Y. Fusion Classification of HSI and MSI Using a Spatial-Spectral Vision Transformer for Wetland Biodiversity Estimation. *Remote Sens.* **2022**, *14*, 850. https://doi.org/10.3390/rs14040850

Academic Editor: Nicola Clerici

Received: 10 January 2022
Accepted: 6 February 2022
Published: 11 February 2022

1. Introduction

Coastal wetland is a transitional area between terrestrial and marine ecosystems, which has a complex environment and monitoring elements [1]. Accurate biodiversity monitoring of coastal wetlands is of great significance in water conservation [2], biodiversity conservation [3], and blue carbon sink development [4]. Recently, natural factors and human activities have deteriorated biotope and biodiversity.

The traditional on-site monitoring receives data by stations and sections, which is time-consuming and laborious. In contrast, remote sensing technology has the advantages of large-area coverage, spatio-temporal synchronization, and high spatial-spectral resolution [5], providing highly relevant information for a wide range of wetland monitoring applications. Therefore, biodiversity estimation based on remote sensing achieves economic and real-time data collection. In recent years, a lot of works have been developed for biodiversity estimation based on remote sensing.

The limitation of remote sensing including resolution and sensors makes the biodiversity estimation applied at a coarse scale [6]. Biodiversity is mainly divided into animal diversity and plant diversity. land-cover mapping is one of the most widely used applications of optical remote sensing, which directly serves the plant diversity estimation. In addition, considering the limitations of remote sensing monitoring of animal diversity, land-cover mapping is capable of estimating animal diversity indirectly [7]. Therefore, biodiversity estimation based on remote sensing relies on the interpretation of land-cover, for which hyperspectral images (HSI) have attracted significant attention [8]. Su et al. [9] designed an elastic network based on low-rank representation to classify HSI, which is collected by a GaoFen-5 satellite, thus plant diversity estimation of coastal wetland was achieved. Hong et al. [10] combined the convolutional neural network (CNN) and graph convolutional network (GCN) to extract different types for urban land-cover classification. Zhang et al. [11] developed a transferred 3D-CNN for HSI classification, for which the overfitting problem caused by insufficient labeled samples was alleviated. Wang et al. [12] proposed a generative adversarial network (GAN) for land-cover recognition, and achieved promising results with imbalanced samples. Zhu et al. [13] designed a spatial-temporal semantic segmentation model to harness temporal dependency for land-use and land-cover (LULC) classification. Zhang et al. [14] proposed a parcel-level ensemble method for land-cover classification based on Sentinel-1 synthetic aperture radar (SAR) time series and segmentation generated from GaoFen-6 images. In [15], the land-cover in coastal wetland were classified using an object-oriented random forest algorithm. In [16], a hierarchical classification framework (HCF) was developed for wetland classification. The HCF classifies land-cover from rough classes to their subtypes based on spectral, texture, and geometric features.

Generally speaking, HSIs exhibit great advantages in land-cover classification due to carrying plenty of spectral information [17]. However, the spatial resolution of HSI is usually lower because of the requirements of signal-to-noise ratio in long exposure [18]. In addition, the existing "different body with same spectrum" or "same body with different spectrum" phenomenon on HSI deteriorates the interpretation. Therefore, joining the complementary merits of multisource data further improves the accuracy of land-cover classification [19]. In the past decade, extensive classification techniques have been successfully applied to multisource data [20–22]. Some of the machine learning methods rely on support vector machine (SVM), extreme learning machine (ELM) and random forest (RF) [23–25]. More recently, deep learning has boosted the performance of classification in the remote sensing community. In [26], the adaptive differential evolution was utilized to optimize the classification decision from different data sources. Rezaee [27] et al. employed the deep CNN to classify wetland land-cover on a large scale. In [28], a 3D-CNN was designed for multispectral image (MSI) classification to serve wetland feature monitoring. Zhao et al. [29] developed a hierarchical random walk network (HRWN) to exploit the spatial consistency of land-cover over HSI and light detection and ranging (LiDAR) data. In [30], a three-steam CNN was designed to fuse HSI and LiDAR data, in which the multi-sensor composite kernels (MCK) scheme was employed for feature integration. Xu [31] et al. developed a dual-tunnel CNN and a cascaded network, named two-branch CNN, for feature extraction, and the multisource features were stacked for fusion. Liu [32] et al. improved the two-branch CNN through interclass sparsity-based discriminative least square regression (CS_DLSR), which encouraged the feature discrimination among different land-cover. Zhang [33] et al. designed an encoder–decoder to construct the latent representation between multisource data, and then fuse them for classification. In [34], a depth feature interaction network (DFINet) was developed for HSI and MSI classification in the Yellow River estuary wetland. In [35], a hierarchy-based classifier for urban vegetation classification was designed to incorporate the canopy height features into spectral and textural data.

Despite the intense interest in multisource data classification, it remains a highly challenging problem. The primary challenges are summed up into two aspects: (1) *Data quality needs to be improved*. A higher spatial and spectral resolution is conducive to the extraction of texture and spectral features, which improves the final classification results. (2) *Sequential features need to be noticed*. The continuity of spatially distribution and spectrum curves enhances the discrimination of features and the classification performance.

To address the aforementioned challenges, a systematic framework for multisource remote sensing image processing is constructed. The typical CNN is used for data fusion, and then a spatial-spectral vision transformer (SSVit) is employed for land-cover mapping which promotes biodiversity estimation. In stage 1, a CNN-based method is utilized to fuse the HSI and MSI over the Yellow River estuary wetland, thus both the spatial and spectral resolution of the fused image is improved. In stage 2, the land-cover mapping of fused image is generated by the SSViT, which exploits the sequential relationship of spatial and spectral information, respectively. After that, an external attention module is adopted for feature integration. Finally, biodiversity estimation is achieved by land-cover mapping and benthic collection in the study area. Extensive experiments conducted on the Yellow River estuary dataset and several related methods reveal that the proposed framework provides competitive advantages in terms of data quality and classification accuracy.

The main contributions of the proposed method are summarized as follows:

- A systematic framework including stage 1 (data fusion) and stage 2 (classification) is constructed for land-cover mapping, which serves as the precondition of biodiversity estimation. In fact, the relationship between land-cover and biomass is of utmost importance for remote sensing monitoring of biodiversity. In this paper, the coarse-scale biodiversity estimation of the Yellow River estuary dataset is indirectly achieved based on land-cover mapping.
- The classification stage is crucial for information interpretation. To explore the spatial-spectral sequential features of wetland, the spatial transformer and spectral transformer both with position embedding are utilized to extract the neighborhood correlation, which encourages the discrimination between different classes. In addition, an external attention module is employed to enhance the spatial-spectral features. Different from the self-attention module, the external attention module is optimized with all training sets. Finally, the pixel-wise prediction is achieved for land-cover mapping.

The rest of this paper is organized as follows: In Section 2, the study area and considered data are described. Section 3 illustrates the proposed method in detail. Section 4 presents the experimental results on the Yellow River estuary dataset to validate the proposed method, and then the biodiversity estimation is achieved. Finally, the conclusions are presented in Section 5, respectively.

2. Study Area and Data Description

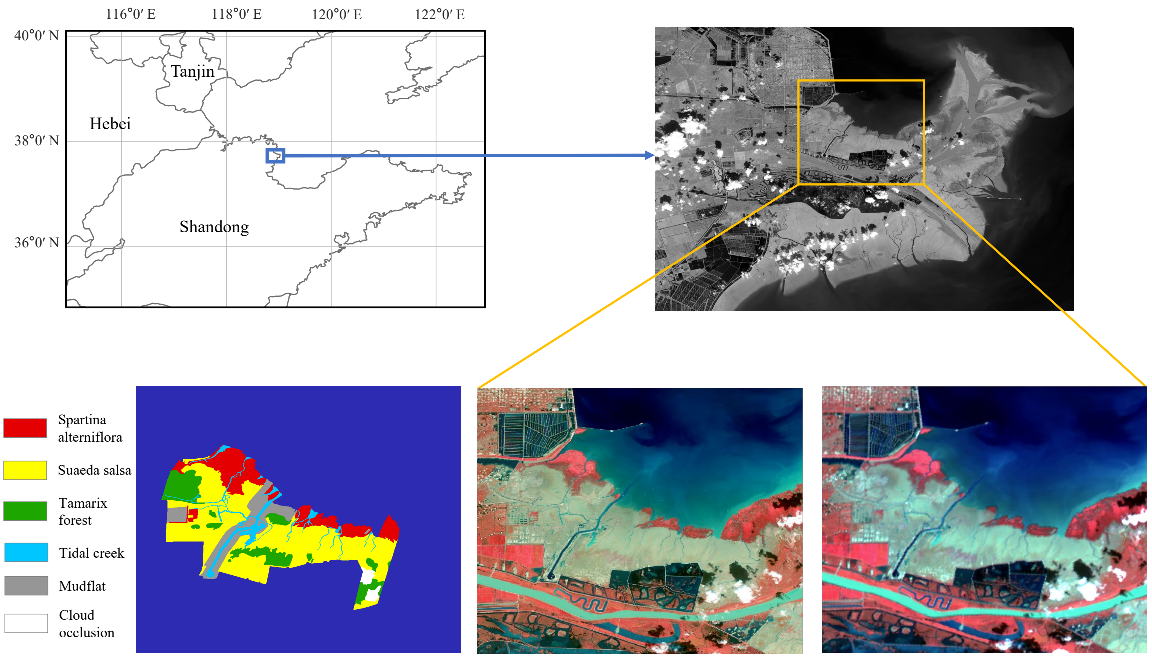
The Yellow River delta wetland locates in Shandong Province, China 118°33′–119°20′E, 37°35′–38°12′N. The Yellow River delta wetland is an important ecological functional area in the Bohai Sea, while the estuarine wetland is a typical area [36]. In particular, the intertidal zone of the Yellow River estuary wetland involves rich salt marsh vegetation and benthos. It plays an important role in biodiversity protection and ecological restoration. However, the Yellow River delta wetland is struggling to face the reduction of natural area, biodiversity, and ecosystem service function.

Therefore, the monitoring of species composition and spatio-temporal distribution in the study area promotes biodiversity estimation and protection. As shown in Figure 1, the intertidal zone in the north of the Yellow River is selected as the study area, and 11 sites are arranged for benthos collection. The coordinate and real landscape of field sites are listed in Table 1.

In this paper, a mudflat quantitative sampler with the size of 0.25 m × 0.25 m × 0.3 m is utilized to collect intertidal biological samples at the sites according to the sites and number of quadrats listed in Table 1. The size of a quadrat is 0.0625 m^2. Note that *Bullacta exarat* and *Mactra veneriformis* are usually located on the surface of the intertidal zone, which is recorded directly through field observation. Some samples at Sites A2 and A3 are scoured by tide during sampling, thus the remaining samples are recorded at the proportion of 80%. All samples retained after elutriation are transferred to the sample bottle. To obtain the information of species composition and species density, the retained samples are further analyzed quantitatively under the stereomicroscope.

Table 1. The coordinate and real landscape of field sites.

Site	Quadrats	Coordinate	Real Landscape
A1	2	119°9′45.360″E 37°47′15.324″N	*Tamarix chinensis*
A2	1	119°9′52.740″E 37°47′31.848″N	*Suaeda salsa*
A3	1	119°10′4.224″E 37°47′57.082″N	Sparse area of *Spartina alterniflora*
B1	2	119°7′35.832″E 37°47′8.052″N	Mixed area of *Suaeda salsa* and *Tamarix chinensis*
B2	4	119°7′40.620″E 37°47′47.328″N	Dense area of *Suaeda salsa*
B3	2	119°7′48.000″E 37°48′13.464″N	Sparse area of *Suaeda salsa*
C1	2	119°6′31.498″E 37°49′35.351″N	Dense area of *Spartina alterniflora*
C2	2	119°6′39.232″E 37°49′37.604″N	Mixed area of *Suaeda salsa* and *Spartina alterniflora*
C3	2	119°6′44.322″E 37°49′34.668″N	Coastal beach in invasion area of *Spartina alterniflora*
D1	1	119°4′33.575″E 37°49′6.440″N	Tidal creek in invasion area of *Spartina alterniflora*
D2	1	119°4′46.424″E 37°50′10.171″N	Mudflat near Tidal creek



(a) Ground truth (b) MSI (c) HSI

Figure 1. Location of the study area. (**a**) the ground truth image; (**b**) the MSI captured by ZY1-02D; (**c**) the HSI captured by ZY1-02D.

In addition, land-cover mapping is generated by multisource data including HSI and MSI, which are captured by ZiYuan1-02D (ZY1-02D) satellite on 26 September 2020. The MSI includes eight MSS (multi-spectral scanner) bands, with 10 m ground sample distance (GSD). The HSI is obtained by an AHSI sensor, with a spectral resolution of 10–20 nm. The parameters of ZiYuan1-02D satellite are listed in Table 2. The multisource data are preprocessed by image registration, atmospheric correction, and radiometric calibration. To be specific, the HSI and MSI are transformed into the WGS 84 geographic coordinate system, and then the considered images are registered by the automatic registration tool in ENVI. The Fast Line-of-Sight Atmospheric Analysis of Spectral Hypercubes (FLAASH) module of ENVI is employed for atmospheric correction, and the radiometric calibration is carried out by using the gain and offset coefficients. The classification system is established based on the real landscape listed in Table 1. Here, five classes are selected for land-cover classification, and the Cloud occlusion (white area in Figure 1a) is eliminated. The benthic data collected at different sites are reported in Figure 2 and Table 3. The regions of interest are selected as the training set, in which the ground truth is annotated by experts with rich knowledge in field trips as listed in Table 4. Note that Tamarix forest is mainly the distribution area of *Tamarix chinensis*, which mixed with other vegetation such as *Suaeda salsa* and *Phragmites australis*.

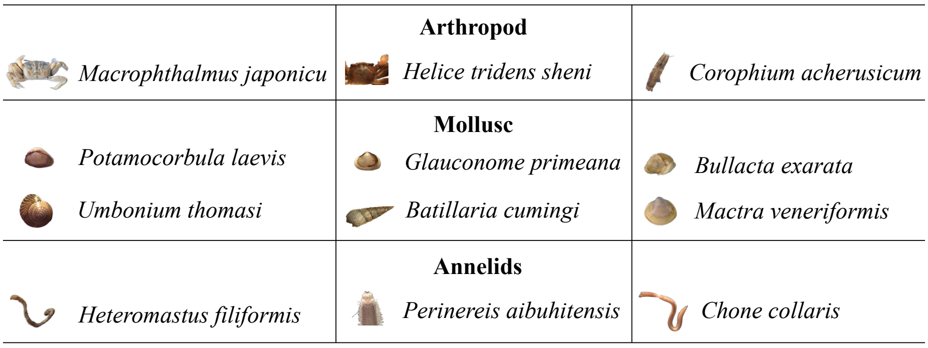


Figure 2. The population and legend of benthos in the study area.

Table 2. The parameters of ZiYuan1-02D satellite.

Type	Range of Wavelengths		Spatial Resolution	Spectral Resolution	Width
MSI	B02	452-521nm	10 m	—	115 km
	B03	522–607 nm			
	B04	635–694 nm			
	B05	776–895 nm			
	B06	416–452 nm			
	B07	591–633 nm			
	B08	708–752 nm			
	B09	871–1047 nm			
HSI	400–2500 nm		30 m	10–20 nm	60 km

Table 3. Benthic data collected at different sites.

Sites	Species Name	Number	Species Density (#/m^2)	Weight (g)	Biomass (g/m^2)
A1	*Macrophthalmus japonicu*	2	16	1.335	10.682
A2	*Glauconome primeana*	3	60	3.127	62.538
	Macrophthalmus japonicu	1	20	2.228	44.558
	Perinereis aibuhitensis	4	80	0.865	17.304
A3	*Glauconome primeana*	4	80	3.514	70.280
	Potamocorbula laevis	4	80	0.256	5.110
	Bullacta exarata	2	2	4.776	4.776
	Macrophthalmus japonicu	1	20	3.565	71.306
	Perinereis aibuhitensis	2	40	0.858	17.166
	Mactra veneriformis	1	1	20.615	20.615
B1	*Macrophthalmus japonicu*	2	16	1.254	10.028
	Helice tridens sheni	1	8	5.226	41.810

Table 3. *Cont.*

Sites	Species Name	Number	Species Density (#/m^2)	Weight (g)	Biomass (g/m^2)
B2	*Macrophthalmus japonicu*	6	24	8.138	32.552
	Helice tridens sheni	2	8	11.242	44.968
	Perinereis aibuhitensis	5	20	0.956	3.825
	Mactra veneriformis	1	1	12.578	12.578
B3	*Glauconome primeana*	2	16	1.241	9.926
	Corophium acherusicum	12	96	0.023	0.186
	Bullacta exarata	1	0.5	1.487	0.743
	Perinereis aibuhitensis	3	24	0.151	1.204
	Umbonium thomasi	4	32	1.057	8.454
C1	*Batillaria cumingi*	2	16	2.237	17.892
	Macrophthalmus japonicu	3	24	3.435	27.483
	Perinereis aibuhitensis	5	40	0.566	4.527
C2	*Batillaria cumingi*	2	16	1.690	13.520
	Macrophthalmus japonicu	4	32	5.221	41.770
	Helice tridens sheni	1	8	0.825	6.602
	Perinereis aibuhitensis	5	40	0.882	7.056
	Umbonium thomasi	4	32	0.082	0.658
C3	*Glauconome primeana*	18	144	10.163	406.504
	Batillaria cumingi	22	176	18.350	146.802
	Bullacta exarata	1	1	1.883	1.883
	Helice tridens sheni	2	16	10.215	81.720
	Perinereis aibuhitensis	15	120	1.439	11.510
	Heteromastus filiformis	4	32	0.005	0.184
	Umbonium thomasi	11	88	0.082	0.658
D1	*Potamocorbula laevis*	4	64	0.256	4.088
	Macrophthalmus japonicu	1	16	3.251	52.016
	Perinereis aibuhitensis	2	32	0.021	0.331
	Heteromastus filiformis	8	128	0.004	0.062
D2	*Glauconome primeana*	10	160	1.900	30.395
	Chone collaris	1	16	0.005	0.082
	Corophium acherusicum	16	256	0.033	0.526
	Helice tridens sheni	3	48	6.530	104.483
	Perinereis aibuhitensis	8	128	0.840	13.442
	Heteromastus filiformis	7	112	0.063	1.013

Table 4. Number of training and testing samples for the Yellow River estuary dataset.

Class		Number of Samples	
No.	**Name**	**Training**	**Testing**
1	Spartina alterniflora	735	39,784
2	Suaeda salsa	2519	118,213
3	Tamarix forest	1069	31,044
4	Tidal creek	529	15,673
5	Mudflat	702	24,592
Total		5554	229,306

3. Proposed Classification Framework

Give a set of multisource remote sensing images, including low-resolution HSI (LR-HSI $\mathbf{X}_{hsi} \in \mathbb{R}^{H \times W \times C_h}$) and high-resolution MSI (HR-MSI $\mathbf{X}_{msi} \in \mathbb{R}^{3H \times 3W \times C_m}$) from ZY1-02D. Here, H and W are the height and width of LR-HSI. $C_h = 166$ and $C_m = 8$ are the bands' number of LR-HSI and HR-MSI, respectively.

Data fusion aims to integrate the spectral advantages of LR-HSI and the spatial advantages of HR-MSI, thus the fused image with high spatial-spectral resolution is generated. After that, the classification technology is conducted on the fused image for land-cover mapping. As shown in Figure 3, the proposed framework includes two stages: (1) In stage 1 (data fusion), the advantages of spatial-spectral information is reconstructed by the CNN-based method; (2) In stage 2 (classification), the proposed SSViT, including a spatial-spectral vision transformer and an external attention module, are employed to learn the sequential relationship of spatial-spectral information for classification.

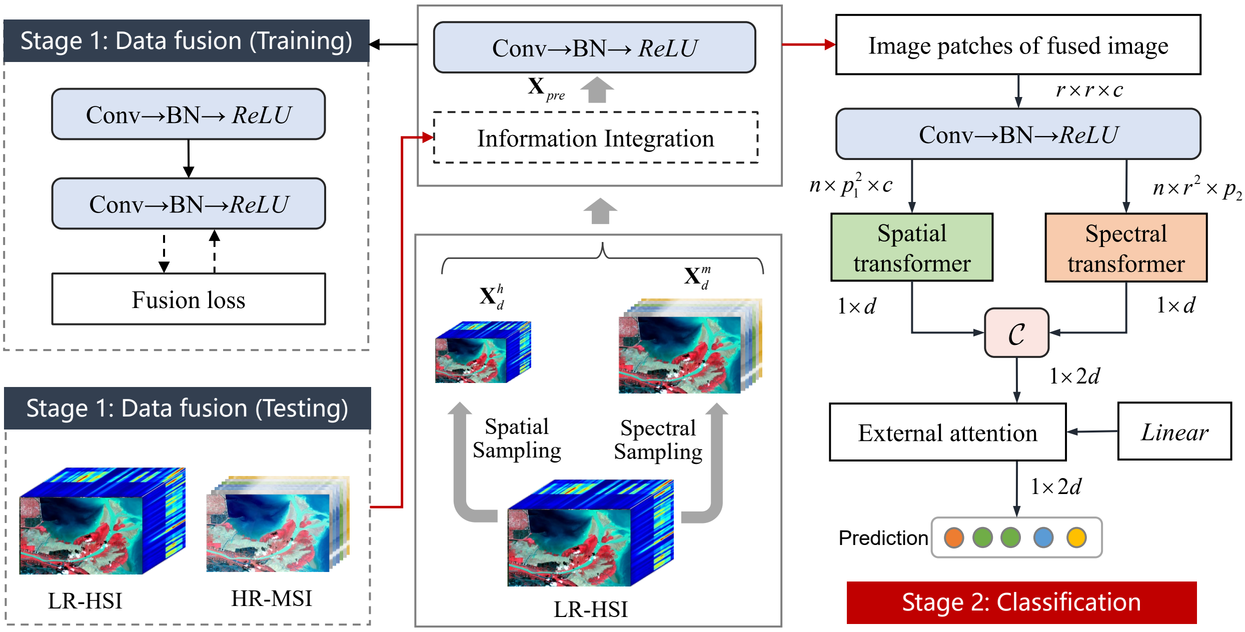


Figure 3. Framework of the proposed spatial-spectral vision transformer, which includes stage 1 (data fusion) and stage 2 (classification).

3.1. Data Fusion Based on HSI and MSI

Information fusion of HSI and MSI improves the spatial-spectral resolution of the fused image, which is of great benefit to the subsequent interpretation. In [37], several methods were conducted to fuse the HSI and MSI of ZY1-02D. Due to the unavailable spatial and spectral response functions of sensors in the application, the performance of many fusion technologies is limited, while the deep learning methods are able to alleviate this problem [38]. Therefore, a CNN-based method is introduced for information fusion of HSI and MSI. It is worth mentioning that other method options are also available for data fusion combined with the practical demands.

3.1.1. Spatial and Spectral Sampling

CNN-based fusion methods require a proportionally large number of training samples for parameters optimization. However, the reference image is scarce in the wetland scene. Thus, the spatial sampling and spectral sampling are implemented on LR-HSI to obtain the degraded images. The LR-HSI is used as the reference image, and the degraded images are the training images during parameter optimization.

The spatial sampling is implemented on LR-HSI by Gaussian blur and downsampling operation according to the scale ratio of HR-MSI and LR-HSI. The degraded MSI is generated by equal interval band selection in the visible to near-infrared bands of LR-HSI. The generation of the degraded HSI $\mathbf{X}_d^h$ and MSI $\mathbf{X}_d^m$ is calculated as:

$$\mathbf{X}_d^h = Downsampling(Gaussian(\mathbf{X}_{hsi}), 1/r), \tag{1}$$

$$\mathbf{X}_d^m = \mathbf{X}_{hsi}(b), \tag{2}$$

where $\mathbf{X}_d^h$ is the degraded HSI. $Downsampled(\cdot)$ and $Gaussian(\cdot)$ are spatially downsampling by bilinear operation and blur by Gaussian filter, respectively. $\mathbf{X}_d^m$ is sampled from original LR-HSI along the spectrum. $b = Rounding(C_h'/8) * \omega + 1, \omega = [0, 1, \cdots, C_m - 1]$, and $C_h' = 76$ is the number of bands of visible to near-infrared in $\mathbf{X}_{hsi}$. $Rounding(\cdot)$ is the operation of rounding to 0.

3.1.2. Image Fusion Based on CNN

To achieve the complementary advantages of HSI and MSI, a CNN-based approach is designed for information fusion. Firstly, the preliminary fusion is applied for information integration. The preliminary fused image is denoted as:

$$\mathbf{X}_{pre}(i) = \begin{cases} \mathbf{X}_d^m(i), & if\ i = b \\ Upsampling(\mathbf{X}_d^h(i)), & otherwise, \end{cases} \tag{3}$$

where $\mathbf{X}_{pre}$ is the preliminary fused image, $i = [1, 2, \cdots, C_h]$. $Upsampling(\cdot)$ is spatially upsampling by bilinear operation.

The preliminary fused image is filtered through 3 × 3 convolutional layer (Conv) with stride 1, batch normalization (BN), and activation layers (*ReLU*), and then the other two sequential operations with skip-connection are conducted for further fusion. The fusion loss function L_f is equipped for optimization, which is defined as:

$$L_f = \frac{1}{HWC_h} \sum_{h=1}^{H} \sum_{w=1}^{W} \sum_{c=1}^{C_h} \|\mathbf{X}_{hsi}(h, w, c) - \mathbf{X}_{fuse}(h, w, c)\|_2, \tag{4}$$

where $\mathbf{X}_{fuse}$ is the fused image, and $\|\cdot\|_2$ is *L2-norm*.

Note that three convolutional layers with a kernel size of 3 × 3 are deployed in the CNN-based method, and stride is set as 1 with padding operation. Thus, the spatial size of feature maps remains unchanged during training. More specifically, the learning rate is set to 1×10^{-4}, and the Adam is employed to train the CNN-based method, which is optimized 500 epochs.

3.2. Classification Based on SSViT

Accuracy land-cover mapping is crucial for biodiversity estimation based on remote sensing. The distribution of land-cover indirectly reflects biodiversity. To exploit the sequential relationship from spatial/spectral information, attention mechanisms have achieved promising performance [39]. The vision transformer (ViT) has boosted the performance in computer vision, owing to the position embedding and self-attention [40]. Inspired by the ViT, the spatial and spectral transformer assembled by an external attention module, named SSViT, is designed for land-cover classification. More specifically, the proposed SSViT is mainly composed of the spatial transformer and spectral transformer, which are utilized to extract the sequential relationship of spatial-spectral information, respectively. The framework of spatial/spectral transformer is illustrated in Figure 4.

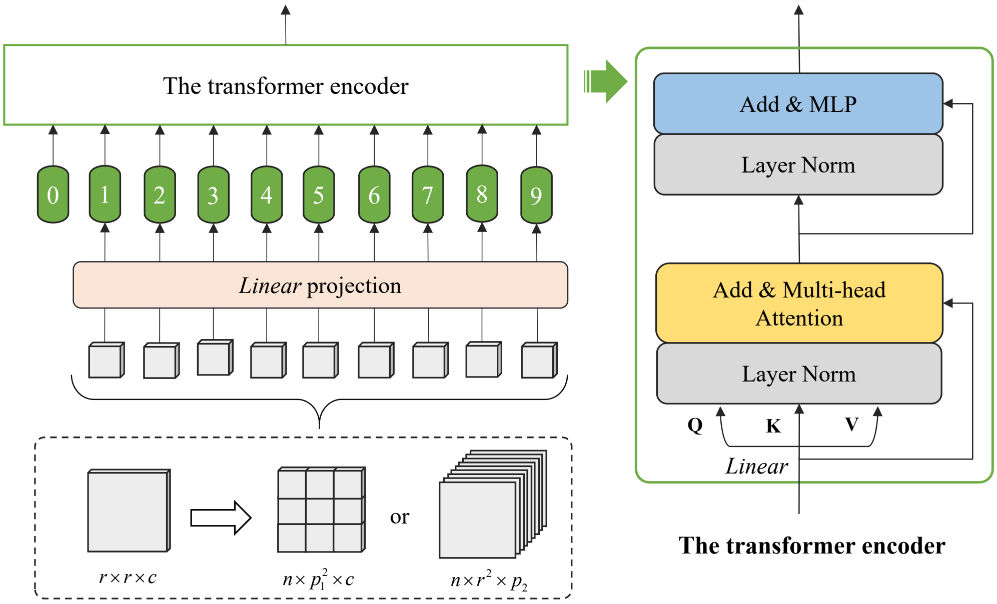


Figure 4. The framework of spatial/spectral transformer with a depth of 1.

3.2.1. The Spectral Transformer

HSIs sequentially record the information from the whole electromagnetic spectrum. Therefore, the discrimination of spectral response explicitly encourages the land-cover classification. However, the "different body with same spectrum" or "same body with different spectrum" phenomenon deteriorates the classification results. Thus, the spectral transformer with a depth of 8 is designed to extract the relationship of spectral information. The baseline of spectral transformer with a depth of 1 is exhibited in Figure 4. First, image patches centered at pixels of the fused image are fed into the spectral transformer. Give an image patch $\mathbf{X} \in \mathbb{R}^{r\times r\times C_h}$ that is filtered by sequential operations (Conv, BN, and *ReLU*) to generate the feature map $\mathbf{X}_f \in \mathbb{R}^{r\times r\times c}$, which is reshaped in several sub-patches $\mathbf{X}_{spec} \in \mathbb{R}^{n(r^2\cdot p_1)}$ where n is the number of sub-patches, which is the sequence length for the transformer, and $p_1 = c/n$. After that, a trainable *Linear* projection layer is utilized to map $\mathbf{X}_{spec}$ to d dimensions vectors. To obtain the relationship between n sub-patches, a learnable 1D position embedding is preset to represent the image through the transformer, which is illustrated in the green box of Figure 4.

The transformer encoder consists of multi-head attention, normalization layer (Layer Norm) and multilayer perceptron (MLP). The outputs of multi-head attention are calculated as:

$$\mathbf{f} = A(\mathbf{Q},\mathbf{K},\mathbf{V}) = softmax(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}})\mathbf{V}, \tag{5}$$

where $\mathbf{Q},\mathbf{K},\mathbf{V} \in \mathbb{R}^{N\times d}$ are the query set, key set, and value set, respectively, and N is batch size. $A(\cdot)$ denotes the attention function. $\mathbf{f} \in \mathbb{R}^{N\times d}$ represents the attention feature, which is generated by the weighted values $\mathbf{V}$ with respect to the attention learned from $\mathbf{Q}$ and $\mathbf{K}$. Intuitively, multi-head attention helps the network capture richer information. Multi-head attention introduces several paralleled heads in which an independent scaled dot-product attention function $A(\cdot)$ is utilized to generate the attention features. Therefore, the attention feature $\mathbf{f}$ is redefined as:

$$\mathbf{f} = [head_1, head_2, \cdots, head_h]\mathbf{W}_o, \tag{6}$$

$$head_j = A(\mathbf{Q}\mathbf{W}_j^Q, \mathbf{K}\mathbf{W}_j^K, \mathbf{V}\mathbf{W}_j^V) \quad j = [1, 2, \cdots, h], \tag{7}$$

where $\mathbf{W}_j^Q, \mathbf{W}_j^K, \mathbf{W}_j^V \in \mathbb{R}^{d\times d_h}$ are the projection matrices of j_{th} head. $\mathbf{W}_o \in \mathbb{R}^{hd_h\times d}$ is the projection matrix, and $d_h = d/h$ is the dimension of the features from each head.

3.2.2. The Spatial Transformer

The spatial distribution of wetland land-cover is continuous. The image patches cover rich spatial information, which is considered sequential. Therefore, a spatial transformer is utilized to extract the relationship of spatial information. In detail, the feature map $\mathbf{X}_f \in \mathbb{R}^{r\times r\times c}$ is reshaped in several sub-patches $\mathbf{X}_{spa} \in \mathbb{R}^{n({p_2}^2\cdot c)}$, where n is the number of sub-patches, and $p_2 = r/n$. Similarly, a trainable *Linear* projection layer is utilized to map $\mathbf{X}_{spa}$ to d dimensions vectors, and then a learnable 1D position embedding is preset to represent the image through the transformer. Finally, the relationship of spatial information is obtained by the transformer, and the detailed calculation of the transformer is described in Section 3.2.1. It is worth noting that the depth of the spatial and spectral transformers is set as 8.

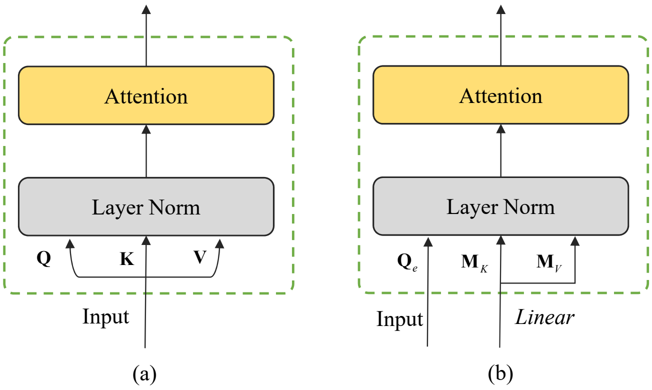


Figure 5. (**a**) Illustration of the self-attention; (**b**) illustration of the external attention.

3.2.3. The External Attention

To realize joint classification, an external attention module is utilized to integrate the feature extracted from the spatial and spectral transformer as shown in Figure 5b. Similar to the self-attention (Figure 5a) in spatial/spectral transformers, the self-correlation is obtained through $\mathbf{Q}$, $\mathbf{K}$, and $\mathbf{V}$ as computed in Equation (5). In external attention, two memory units ($\mathbf{M}_K \in \mathbb{R}^{d'\times 2d}$ and $\mathbf{M}_V \in \mathbb{R}^{d'\times 2d}$) are used to replace $\mathbf{K}$ and $\mathbf{V}$ on the baseline of self-attention, which is optimized during the whole training set. $\mathbf{Q}_e$ is generated by $\mathbf{Q}_e = Linear(\mathcal{C}(\mathbf{f}_{spec}, \mathbf{f}_{spa}))$. Here, $\mathbf{f}_{spec}$ and $\mathbf{f}_{spa}$ are the outputs from the spectral transformer and the spatial transformer, respectively. Therefore, the relationship among the whole training set is learned as follows:

$$\mathbf{f}_e = A(\mathbf{Q}_e, \mathbf{K}, \mathbf{V}) = softmax(\frac{\mathbf{Q}_e\mathbf{M}_K^T}{\sqrt{d}})\mathbf{M}_V, \tag{8}$$

where $\mathbf{Q}_e \in \mathbb{R}^{N\times 2d}$ and $d' < 2d$, and . Finally, the pixel-wise prediction is achieved by a *Linear* layer and *Softmax*.

The proposed SSViT is deployed on an *Nvidia GTX 3080* GPU in PyTorch. The loss function is cross-entropy loss defined in Equation (9), which is optimized by stochastic gradient descent (SGD). Specifically, the learning rate is 5×10^{-4}, and the number of epochs is 500. The batch size, momentum, and weight decay are selected as 128, 0.9, and 5×10^{-4}, respectively,

$$L_c = \frac{1}{N}\sum_{i=1}^{N}\sum_{m=1}^{M} -y_i^m log\ p_i^m, \tag{9}$$

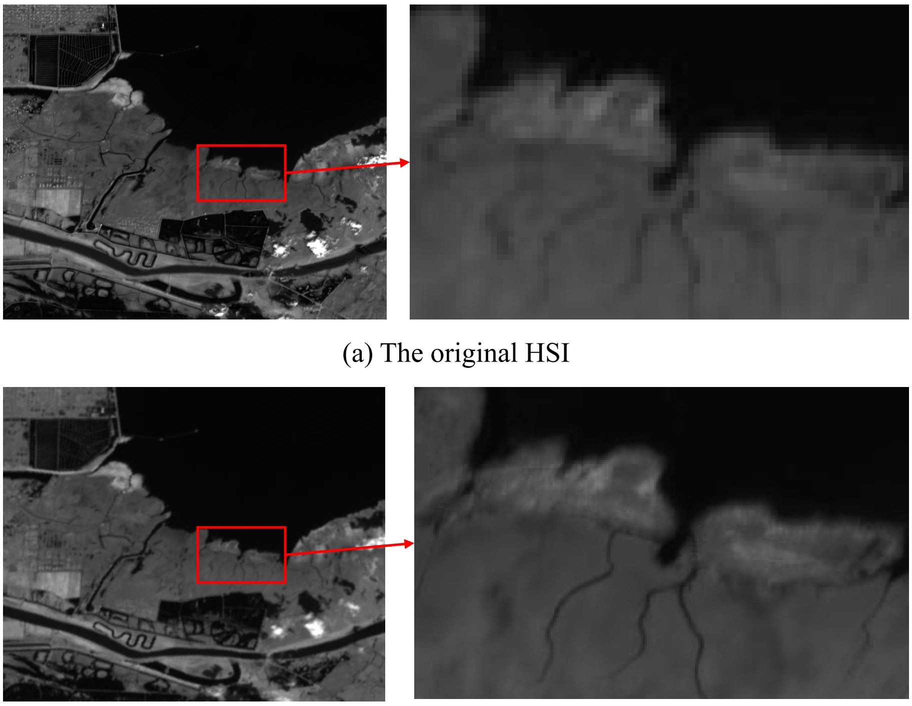
where N is batch size, M is the number of classes, and y is the real label. $y_i^m = 1$ is satisfied when the class of pixel i is m, whose prediction probability after $Softmax$ is p_i^m; otherwise, $y_i^m = 0$.

4. Experiments and Analysis

In this paper, the framework for multisource remote sensing image processing is developed for biodiversity estimation. A CNN-based method is employed for data fusion, in which the performance is evaluated by visual comparison. Next, the proposed SSViT is utilized to classify land-cover on the intertidal zone of the Yellow River estuary wetland. The superiority of the proposed SSViT is measured by the precision corresponding to some related methods. Finally, the correlation between land-cover and benthos is established by the sampling on the selected site. Thus, biodiversity estimation in the intertidal zone is achieved at a coarse scale.

4.1. The Performance of Data Fusion

HSI and MSI of ZY1-02D are utilized for collaborative land-cover classification. Firstly, the advantages of HSI and MSI are fused through the CNN-based method, and then the fused image with high spatial-spectral resolution is classified to generate land-cover mapping. In this paper, visual comparison and spectral angle mapper (SAM) are used to measure the quality of the fused image. Considering that the reference image is not available, the spectral information of the original HSI is used as the reference spectrum. Table 5 reports the SAM of each class according to the reference HSI. The visualized result of data fusion is shown in Figure 6. It is observed that the visual quality of the fused image is improved.



(a) The original HSI

(b) The fused image

Figure 6. Visualized results of data fusion using HSI and MSI.

Table 5. The SAM of each class according to the reference HSI.

Class No.	**1**	**2**	**3**	**4**	**5**
SAM	0.1227	0.0782	0.0801	0.0971	0.0797

4.2. Classification Performance

To validate the superiority of the proposed SSViT, several comparison methods are selected to conduct the experimental validation, including SVM [23], LBP-ELM [24], S2FL [41], Residual CNN [42], two-branch CNN [31], and DFINet [34]. Note that the SVM and LBP-ELM are the spectral classifiers, and other comparison methods are spatial-spectral classifiers together with the proposed SSViT. The S2FL, two-branch CNN, and DFINet are the joint classification framework through multisource feature fusion. Overall accuracy (OA), average accuracy (AA), and kappa coefficient (Kappa) are utilized for quantitative assessment.

4.2.1. Analysis of the Image Patch Size

The spatial-spectral information boosts the performance of land-cover classification, which is affected by the image patch size. The proposed SSViT is employed to extract the sequential features from image sub-patches. The number of image sub-patches is set as 9. Therefore, the image patch size r is set to a multiple of 3 ($[9, 12, 15, 18, 21]$). The relationship between OA and image patch size is shown in Figure 7. It is found that, when $r = 9$ and $r = 12$, the classification results are not satisfactory. This is because the land-cover in the intertidal zone are mixed and patchy, and a smaller image patch size leads to the fragmentation of the classification results. In contrast, when $r > 15$, the OA value decreased gradually. Excessive spatial neighborhood reduces the discrimination of spatial information. Moreover, the computational burden of the model increases. Therefore, $r = 15$ is selected as the best choice.

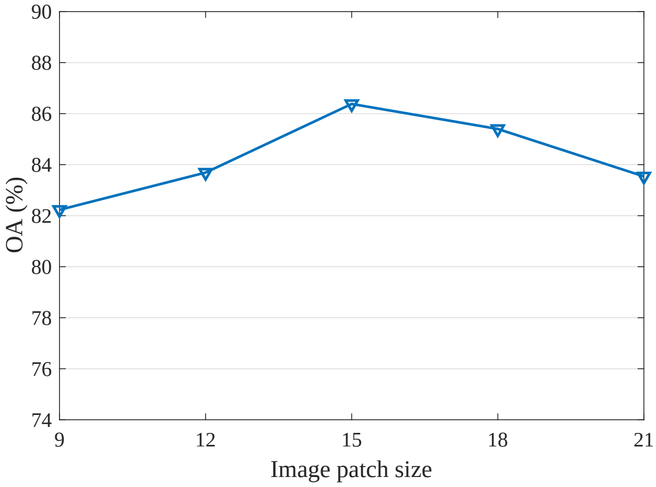


Figure 7. Relationship between OA and image patch size.

4.2.2. Ablation Experiment

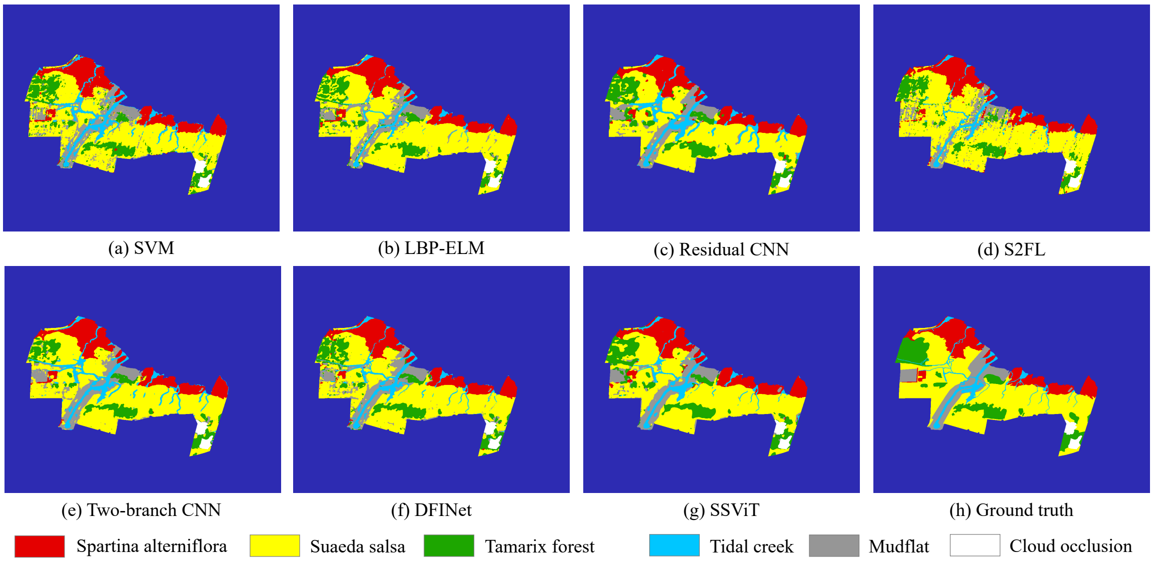
The spatial and spectral transformers are used to extract the sequential relationship of spatial and spectral information, and the external attention module is employed for feature integration. Next, the benefits of different modules on classification results are further discussed. In Table 6, spatial transformer and spectral transformer only mine the relationship of spatial or spectral information, which obtains low OA values. When combing the spatial transformer and spectral transformer, the OA value increased by at least 1.99%. In addition, the external attention module further improves the classification performance, in which the OA improvement of the full model by 0.91% is achieved. It confirms that the relationship of spatial-spectral tends to generate accurate land-cover mapping, and feature integration through external attention emphasizes the relationship among the whole training set.

Table 6. Ablation experiment of the proposed SSViT on the Yellow River estuary dataset.

Method	OA (%)
Spatial transformer	83.48
Spectral transformer	83.37
Without external attention	85.47
Full model	**86.38**

4.2.3. Classification Results on the Yellow River Estuary Dataset

Figure 8 presents the land-cover mapping corresponding to the experiments reported in Table 7. The land-cover mapping obtained by SVM, LBP-ELM, and S2FL tends to be rather noisy, resulting in serious landscape fragmentation. The continuity of land-cover distribution is ignored because only the spectrum is introduced. For Residual CNN, the boundaries in different types suffer from artifacts to some extent, mainly because the spectral information is not specifically considered. Regarding the land-cover mapping produced by Two-branch CNN and DFINet, the fragmentation and artifacts are alleviated. Compared with other methods, the proposed SSViT generates better results in terms of class consistency.



(a) SVM (b) LBP-ELM (c) Residual CNN (d) S2FL

(e) Two-branch CNN (f) DFINet (g) SSViT (h) Ground truth

Figure 8. Land-cover mapping using different methods on the Yellow River estuary dataset.

Table 7. Class-specific classification accuracy (%) using different methods.

No.	SVM			LBP-ELM			Residual CNN	S2FL	Two-Branch CNN	DFINet	SSViT		
	HSI	MSI	Fused	HSI	MSI	Fused	Fused	HSI+MSI	HSI+MSI	HSI+MSI	HSI	MSI	Fused
1	90.95	83.68	93.75	90.52	83.44	**93.68**	91.66	89.74	92.82	92.89	93.23	92.92	93.45
2	82.04	90.18	83.58	86.53	**93.89**	89.38	83.53	90.35	87.82	86.10	84.33	85.36	87.96
3	71.41	48.68	62.23	71.45	44.01	65.16	**78.60**	65.80	68.04	76.20	76.24	75.23	77.78
4	79.05	71.36	72.42	69.74	64.43	63.61	80.57	68.46	**86.03**	84.40	81.89	81.62	82.43
5	70.30	36.82	78.08	69.17	21.44	75.31	74.83	57.19	80.28	78.41	77.36	77.05	**80.69**
OA (%)	80.68	76.43	81.10	82.17	75.54	83.58	83.14	81.87	85.08	85.00	83.87	84.15	**86.38**
AA (%)	78.75	66.14	78.01	77.48	61.44	77.43	81.84	74.31	83.00	83.60	82.61	82.44	**84.46**
Kappa	0.7185	0.6273	0.7209	0.7341	0.5987	0.7514	0.7559	0.7213	0.7792	0.7802	0.7649	0.7683	**0.7994**

From the results reported in Table 7, the original HSI performs the upsampling of bilinear interpolation to obtain the same spatial size of MSI. In Table 7, "HSI" denotes that the original HSI performs the upsampling of bilinear interpolation to obtain the same spatial size of MSI, and "MSI" represents the original MSI. "Fused" represents the fused image generated by the CNN-based method in Section 3.1, and "HSI+MSI" indicates that the upsampled HSI and original MSI are fed into the classifiers based on feature fusion. From the experimental results, it is possible to observe that the fused image achieves better performance than merely using HSI and MSI. For the proposed SSViT, the OA value is increased by at least 2.23%. The fused image significantly improves the classification accuracy of *Spartina alterniflora*. In addition, deep learning methods achieve better performance than traditional methods in most cases. In particular, the two-branch CNN and DFINet achieve competitive results, but the improvement is finite. Generally speaking, the proposed SSViT outperforms the precision on most classes compared with the considered methods. The comparison demonstrates that the proposed SSViT is powerful in sequential feature extraction which encourages discrimination between different classes.

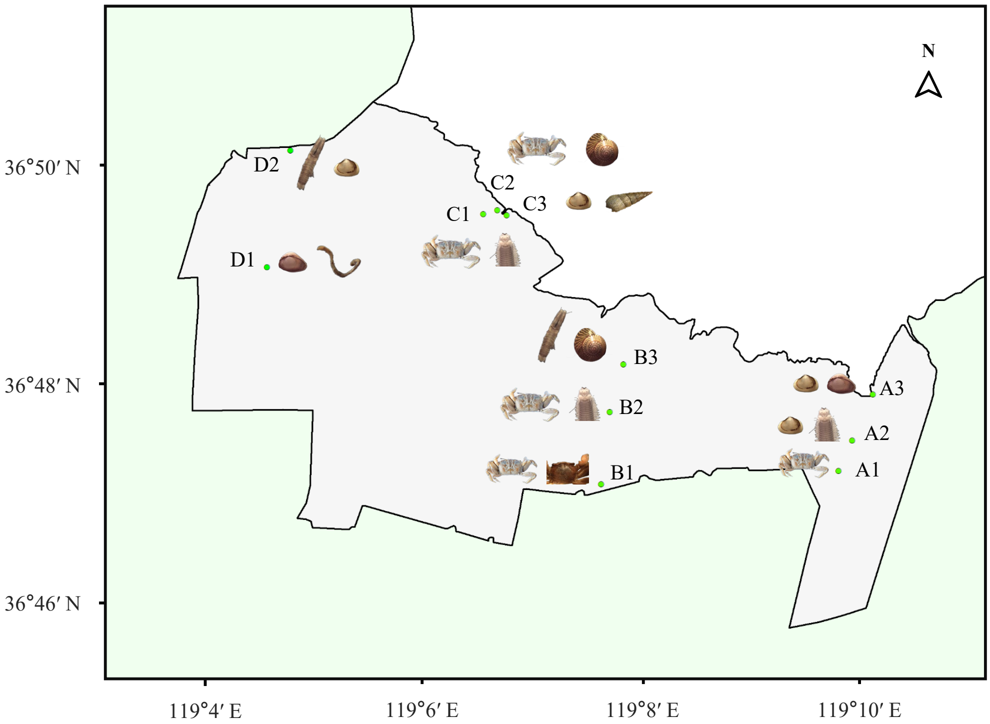


Figure 9. The dominant species of different sites in the intertidal zone of the Yellow River estuary wetland.

Table 8. Diversity index and species number of different sites.

Site	Diversity Index			Species Number		
A1–A3	0	1.406	1.914	1	3	6
B1–B3	0.918	1.568	1.665	2	4	5
C1–C3	1.485	2.419	2.298	3	5	7
D1–D2	1.640	2.256	-	4	6	-

4.3. Biodiversity Estimation

After obtaining the land-cover mapping, the biodiversity estimation is further achieved using biomass information of the 11 sites. The location and dominant species of different sites are shown in Figure 9, and the corresponding diversity index and species number are listed in Table 8.

4.3.1. Biodiversity Estimation in the Intertidal Zone

Species richness, species evenness, and diversity index are introduced for biodiversity estimation in the intertidal zone. As listed in Table 8, the sections A, B, and C are sampled from high-tide to low-tide, respectively. Considering the diversity index and dominant species, it is found that the species distribution is continuous.

Combined with the land-cover mapping, it is found that most of the Tamarix forest is located in the high-tide with lower biodiversity, and the dominant species are mainly annelids such as *Macrophthalmus japonicu* and *Helice tridens sheni*. For the *Suaeda salsa* area in the middle-tide (A2 and B2), the biodiversity is relatively high, while the *Spartina alterniflora* area distributes in the low-tide areas (A3, B3, C2, C3). The Tidal creeks are beneficial to high biodiversity. In addition, section D is selected in the intertidal zone near the Tidal creek. The diversity index and species number are relatively high, which proves that the benthos diversity is closely related to the connectivity of tidal creeks. The low-tide sites (A3, B3, C2, C3) exhibit an increasing trend of the diversity index and species number as well as the sites (D1 and D2) near the tidal creeks.

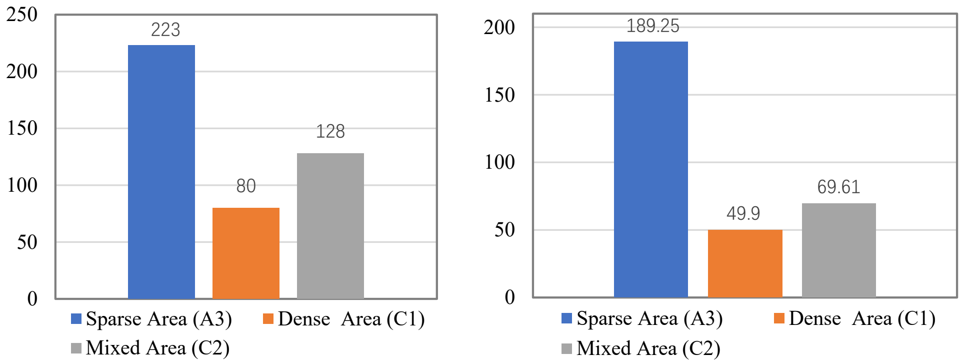
4.3.2. Biodiversity Estimation in the *Spartina alterniflora* Area

Spartina alterniflora is one of the first invasive species in China. It competes with other vegetation in the intertidal zone, resulting in the disappearance of large salt marsh plants. Moreover, *Spartina alterniflora* with developed roots destroys the habitat of offshore organisms, affecting the seawater exchange capacity and hence leading to the decline of water quality. Consequently, biodiversity in the *Spartina alterniflora* area is inevitably destroyed to a certain extent. The biodiversity estimation in the *Spartina alterniflora* area is executed according to the real landscape in Table 1.

According to the real landscape reported in Table 1, three typical sites A3, C1, and C2 are selected, which are located in sparse area, dense area, and mixed area of *Spartina alterniflora* and *Suaeda salsa* (mixed area). As illustrated in Figure 10, the lowest species density and biomass are collected in the dense area. In the sparse area, the species density and biomass are relatively high. The diversity index of the mixed area was slightly higher than that of a dense area, mainly because *Suaeda salsa* played a positive role in soil remediation [43]. Furthermore, the species richness, species evenness, and diversity index of *Spartina alterniflora* area are further discussed. As shown in Figure 11, the species richness and diversity index of considered sites conform to the state in Figure 10. Note that site C1 has the highest species evenness, which is used to measure the stability of biological communities. This is because the lower benthos diversity in the dense area is difficult to change. Conclusively, *Spartina alterniflora* damages biodiversity, whose growth density is negatively correlated with biodiversity. Monitoring the expansion and management of *Spartina alterniflora* by remote sensing is of great significance for wetland biodiversity protection.

4.3.3. Biodiversity Estimation in the *Suaeda salsa* Area

Suaeda salsa is widely distributed in the intertidal zone of the Yellow River estuary wetland, which is conducive to the restoration of the ecological environment. As shown in Figure 12, the biodiversity estimation in the *Suaeda salsa* area is executed according to the real landscape in Table 1, from which sites A2, B2, and B3 are utilized for diversity analysis. Site B2 has the largest coverage, followed by site A2, and the lowest is site B3. Both sites A2 and B3 are located in the middle-tide area, and it is found that the dense area of *Suaeda salsa* is conducive to the distribution of benthos. For the sites B2 and B3 in section B, site B3 at the low-tide area does not realize the expected increase in the diversity index, which indicates that the distribution of *Suaeda salsa* has a positive effect on the distribution of benthos.



(a) The species density

(b) The biomass

Figure 10. The species density and biomass (g/m^2) of the *Spartina alterniflora* area.

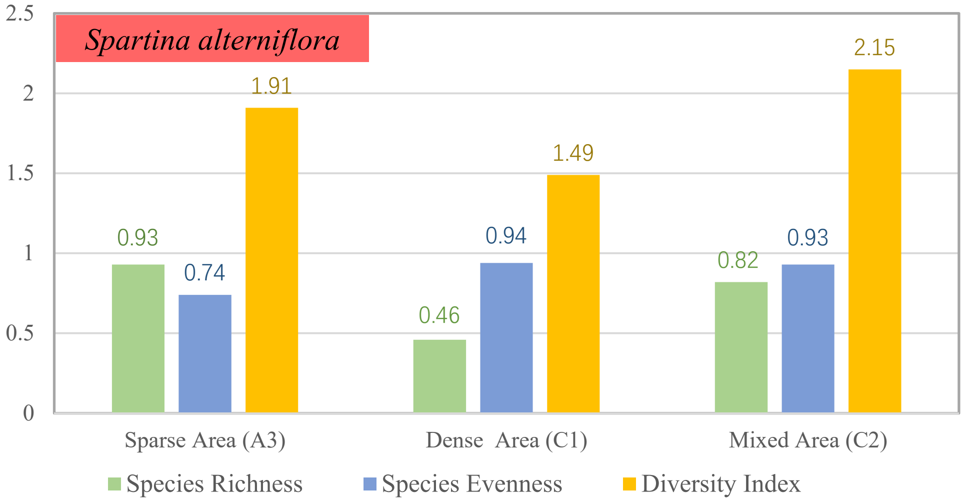


Figure 11. The species richness, species evenness, and the diversity index of the *Spartina alterniflora* area.

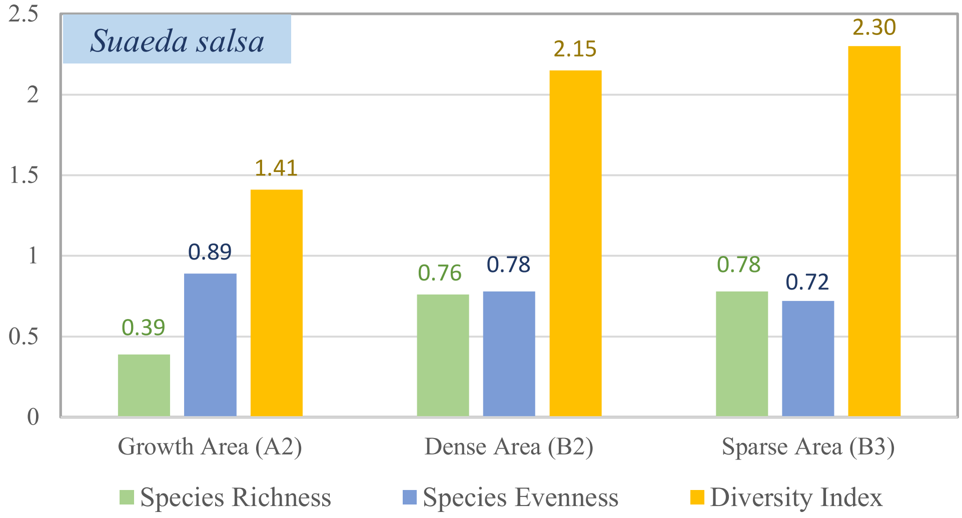


Figure 12. The species richness, species evenness, and diversity index of the *Suaeda salsa* area.

4.4. Discussion

Considering the complex environment in the wetlands, recognizing land-cover and estimating biodiversity based on remote sensing is full of challenges. In this paper, the designed remote sensing image processing framework is used to realize the classification of land-cover in the study area. The benthos diversity of the study area is estimated by integrating the collected data of benthos and land-cover mapping. In general, the proposed framework is capable of mining the sequential features of spatial-spectral information, which improves the precision of land-cover classification. Due to the mixed pixels of medium resolution remote sensing data, how to extract the subspace information for classification needs to be specially studied.

Different from [44,45], this paper realizes the coarse-scale monitoring of biodiversity by the distribution law between benthos and land-cover. In other words, biodiversity estimation is indirectly achieved by land-cover mapping. It is observed that most of the *Spartina alterniflora* and mudflat are distributed in the low-tide area with developed tidal creeks. Therefore, the diversity of benthos is higher. The middle-tide area is also covered by the semidiurnal tide, and the land-cover are mainly salt-tolerant *Suaeda salsa*. Due to the low frequency of tide covered in the high-tide area, *Tamarix chinensis* begins to grow. The diversity of benthos in a high-tide area is the lowest because of the little tidal, and the dominant species is arthropods such as crabs. In addition, it is found that *Spartina alterniflora* deteriorates the ecological environment and the biodiversity. In the future, the fine classification and time-series monitoring of land-cover combined with high-resolution remote sensing will become the focus of works.

5. Conclusions

This paper constructed a systematic framework for a multisource remote sensing image process in the wetland scene. The proposed framework benefits from two aspects. On the one hand, a CNN-based fusion method has been conducted over multisource data, thus the complementary merits have been assembled into the fused image. High-quality fused images consequently serve wetland monitoring. On the other hand, a spatial-spectral vision transformer (SSViT) has been designed for land-cover mapping. The sequential features of the fused image in spatial and spectral dimensions are extracted by the spatial transformer and spectral transformer, respectively. After that, the external attention module is utilized to integrate the spatial-spectral features. In addition, the biodiversity estimation of the study area is further achieved by combining the benthic data and land-cover mapping.

Extensive experiments are conducted on the Yellow River estuary dataset, which reveal the effectiveness of the established systematic framework for multisource remote sensing images. The proposed framework has superior performance in terms of improving data quality and classification accuracy. Combined with the benthic data, the biodiversity of the study area is achieved.

Author Contributions: Conceptualization, Y.G., X.S. and W.L.; methodology and validation, Y.G., W.L. and J.W.; formal analysis, X.S., J.H. and X.J.; investigation, X.S. and Y.F.; writing—original draft preparation, Y.G. and W.L.; writing—review and editing, Y.G. and W.L. All authors have read and agreed to the published version of the manuscript.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: This work is supported by the Beijing Natural Science Foundation (Grant No. JQ20021), the National Natural Science Foundation of China (Grant Nos. 61922013 and 62001023), and the China Postdoctoral Science Foundation (Grant No. BX20200058).

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

References

1. Brisco, B.; Ahern, F.; Murnaghan, K.; White, L.; Canisus, F.; Lancaster, P. Seasonal Change in Wetland Coherence as an Aid to Wetland Monitoring. *Remote Sens.* **2017**, *9*, 158. [CrossRef]
2. Xia, Y.; Fang, C.; Lin, H.; Li, H.; Wu, B. Spatiotemporal Evolution of Wetland Eco-Hydrological Connectivity in the Poyang Lake Area Based on Long Time-Series Remote Sensing Images. *Remote Sens.* **2021**, *13*, 4812. [CrossRef]
3. López-Tapia, S.; Ruiz, P.; Smith, M.; Matthews, J.; Zercher, B.; Sydorenko, L.; Varia, N.; Jin, Y.; Wang, M.; Dunn, J.B.; et al. Machine learning with high-resolution aerial imagery and data fusion to improve and automate the detection of wetlands. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102581. [CrossRef]
4. Sun, S.; Wang, Y.; Song, Z.; Chen, C.; Zhang, Y.; Chen, X.; Chen, W.; Yuan, W.; Wu, X.; Ran, X.; et al. Modelling Aboveground Biomass Carbon Stock of the Bohai Rim Coastal Wetlands by Integrating Remote Sensing, Terrain, and Climate Data. *Remote Sens.* **2021**, *13*, 4321. [CrossRef]
5. Ma, Y.; Wei, J.; Tang, W.; Tang, R. Explicit and stepwise models for spatiotemporal fusion of remote sensing images with deep neural networks. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102611. [CrossRef]
6. Wang, R.; Gamon, J.A. Remote sensing of terrestrial plant biodiversity. *Remote Sens. Environ.* **2019**, *231*, 111218. [CrossRef]
7. Filipponi, F.; Valentini, E.; Nguyen Xuan, A.; Guerra, C.A.; Wolf, F.; Andrzejak, M.; Taramelli, A. Global MODIS Fraction of Green Vegetation Cover for Monitoring Abrupt and Gradual Vegetation Changes. *Remote Sens.* **2018**, *10*, 653. [CrossRef]
8. Hong, D.; Han, Z.; Yao, J.; Gao, L.; Zhang, B.; Plaza, A.; Chanussot, J. SpectralFormer: Rethinking Hyperspectral Image Classification with Transformers. *IEEE Trans. Geosci. Remote Sens.* **2021**, 1. [CrossRef]
9. Su, H.; Yao, W.; Wu, Z.; Zheng, P.; Du, Q. Kernel low-rank representation with elastic net for China coastal wetland land cover classification using GF-5 hyperspectral imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *171*, 238–252. [CrossRef]
10. Hong, D.; Gao, L.; Yao, J.; Zhang, B.; Plaza, A.; Chanussot, J. Graph Convolutional Networks for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5966–5978. [CrossRef]
11. Zhang, H.; Li, Y.; Jiang, Y.; Wang, P.; Shen, Q.; Shen, C. Hyperspectral Classification Based on Lightweight 3D-CNN With Transfer Learning. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 5813–5828. [CrossRef]
12. Wang, Q.; Li, Q.; Li, X. A Fast Neighborhood Grouping Method for Hyperspectral Band Selection. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5028–5039. [CrossRef]
13. Zhu, Y.; Geiß, C.; Thus, E.; Jin, Y. Multitemporal Relearning with Convolutional LSTM Models for Land Use Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 3251–3265. [CrossRef]
14. Zhang, M.; Lin, H. Wetland Classification Using Parcel-level Ensemble Algorithm based on GaoFen-6 Multispectral Imagery and Sentinel-1 Dataset. *J. Hydrol.* **2022**, 127462. [CrossRef]
15. Zhang, X.; Xu, J.; Chen, Y.; Xu, K.; Wang, D. Coastal Wetland Classification with GF-3 Polarimetric SAR Imagery by Using Object-Oriented Random Forest Algorithm. *Sensors* **2021**, *21*, 3395. [CrossRef]
16. Jiao, L.; Sun, W.; Yang, G.; Ren, G.; Liu, Y. A Hierarchical Classification Framework of Satellite Multispectral/Hyperspectral Images for Mapping Coastal Wetlands. *Remote Sens.* **2019**, *11*, 2238. [CrossRef]
17. Li, S.; Song, W.; Fang, L.; Chen, Y.; Ghamisi, P.; Benediktsson, J.A. Deep Learning for Hyperspectral Image Classification: An Overview. *IEEE Trans. Geosci. Remote Sens.* **2019**, *57*, 6690–6709. [CrossRef]
18. Zhang, X.; Huang, W.; Wang, Q.; Li, X. SSR-NET: Spatial–Spectral Reconstruction Network for Hyperspectral and Multispectral Image Fusion. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 5953–5965. [CrossRef]
19. Zhang, M.; Li, W.; Tao, R.; Li, H.; Du, Q. Information Fusion for Classification of Hyperspectral and LiDAR Data Using IP-CNN. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5506812. [CrossRef]
20. Meng, Y.; Rigall, E.; Chen, X.; Gao, F.; Dong, J.; Chen, S. Physics-Guided Generative Adversarial Networks for Sea Subsurface Temperature Prediction. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, 1–14. [CrossRef]
21. Sahour, H.; Kemink, K.M.; O'Connell, J. Integrating SAR and Optical Remote Sensing for Conservation-Targeted Wetlands Mapping. *Remote Sens.* **2022**, *14*, 159. [CrossRef]
22. Zhou, R.; Yang, C.; Li, E.; Cai, X.; Yang, J.; Xia, Y. Object-Based Wetland Vegetation Classification Using Multi-Feature Selection of Unoccupied Aerial Vehicle RGB Imagery. *Remote Sens.* **2021**, *13*, 4910. [CrossRef]
23. Han, X.; Pan, J.; Devlin, A. Remote sensing study of wetlands in the Pearl River Delta during 1995–2015 with the support vector machine method. *Front. Earth Sci.* **2017**, *12*, 521–531. [CrossRef]
24. Li, W.; Chen, C.; Su, H.; Du, Q. Local Binary Patterns and Extreme Learning Machine for Hyperspectral Imagery Classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *53*, 3681–3693. [CrossRef]
25. Zhang, Y.; Cao, G.; Li, X.; Wang, B. Cascaded Random Forest for Hyperspectral Image Classification. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 1082–1094. [CrossRef]
26. Zhong, Y.; Cao, Q.; Zhao, J.; Ma, A.; Zhao, B.; Zhang, L. Optimal Decision Fusion for Urban Land-Use/Land-Cover Classification Based on Adaptive Differential Evolution Using Hyperspectral and LiDAR Data. *Remote Sens.* **2017**, *9*, 868. [CrossRef]
27. Rezaee, M.; Mahdianpari, M.; Zhang, Y.; Salehi, B. Deep convolutional neural network for complex wetland classification using optical remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 3030–3039. [CrossRef]
28. Pan, H. A feature sequence-based 3D convolutional method for wetland classification from multispectral images. *Remote Sens. Lett.* **2020**, *11*, 837–846. [CrossRef]

29. Zhao, X.; Tao, R.; Li, W.; Li, H.C.; Du, Q.; Liao, W.; Philips, W. Joint Classification of Hyperspectral and LiDAR Data Using Hierarchical Random Walk and Deep CNN Architecture. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 7355–7370. [CrossRef]
30. Li, H.; Ghamisi, P.; Soergel, U.; Zhu, X. Hyperspectral and LiDAR Fusion Using Deep Three-Stream Convolutional Neural Networks. *Remote Sens.* **2018**, *10*, 1649. [CrossRef]
31. Xu, X.; Li, W.; Ran, Q.; Du, Q.; Gao, L.; Zhang, B. Multisource Remote Sensing Data Classification Based on Convolutional Neural Network. *IEEE Trans. Geosci. Remote Sens.* **2018**, *56*, 937–949. [CrossRef]
32. Liu, C.; Tao, R.; Li, W.; Zhang, M.; Sun, W.; Du, Q. Joint Classification of Hyperspectral and Multispectral Images for Mapping Coastal Wetlands. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 982–996. [CrossRef]
33. Zhang, M.; Li, W.; Du, Q.; Gao, L.; Zhang, B. Feature Extraction for Classification of Hyperspectral and LiDAR Data Using Patch-to-Patch CNN. *IEEE Trans. Cybern.* **2020**, *50*, 100–111. [CrossRef] [PubMed]
34. Gao, Y.; Li, W.; Zhang, M.; Wang, J.; Sun, W.; Tao, R.; Du, Q. Hyperspectral and Multispectral Classification for Coastal Wetland Using Depthwise Feature Interaction Network. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 5512615. [CrossRef]
35. Zhao, S.; Jiang, X.; Li, G.; Chen, Y.; Lu, D. Integration of ZiYuan-3 multispectral and stereo imagery for mapping urban vegetation using the hierarchy-based classifier. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *105*, 102594. [CrossRef]
36. Li, C.; Zhu, L.; Dai, Z.; Wu, Z. Study on Spatiotemporal Evolution of the Yellow River Delta Coastline from 1976 to 2020. *Remote Sens.* **2021**, *13*, 4789. [CrossRef]
37. Lu, H.; Qiao, D.; Li, Y.; Wu, S.; Deng, L. Fusion of China ZY-1 02D Hyperspectral Data and Multispectral Data: Which Methods Should Be Used? *Remote Sens.* **2021**, *13*, 2354. [CrossRef]
38. Zheng, K.; Gao, L.; Liao, W.; Hong, D.; Zhang, B.; Cui, X.; Chanussot, J. Coupled Convolutional Neural Network with Adaptive Response Function Learning for Unsupervised Hyperspectral Super Resolution. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 2487–2502. [CrossRef]
39. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
40. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
41. Hong, D.; Hu, J.; Yao, J.; Chanussot, J.; Zhu, X.X. Multimodal remote sensing benchmark datasets for land cover classification with a shared and specific feature learning model. *ISPRS J. Photogramm. Remote Sens.* **2021**, *178*, 68–80. [CrossRef]
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
43. Yin, H.; Hu, Y.; Liu, M.; Li, C.; Chang, Y. Evolutions of 30-Year Spatio-Temporal Distribution and Influencing Factors of Suaeda salsa in Bohai Bay, China. *Remote Sens.* **2022**, *14*, 138. [CrossRef]
44. Fogarin, S.; Madricardo, F.; Zaggia, L.; Sigovini, M.; Montereale-Gavazzi, G.; Kruss, A.; Lorenzetti, G.; Manfé, G.; Petrizzo, A.; Molinaroli, E.; et al. Tidal Inlets in the Anthropocene: Geomorphology and Benthic Habitats of the Chioggia Inlet, Venice Lagoon (Italy). *Earth Surf. Process. Landforms* **2019**, *44*, 2297–2315. [CrossRef]
45. Trzcinska, K.; Tegowski, J.; Pocwiardowski, P.; Janowski, L.; Zdroik, J.; Kruss, A.; Rucinska, M.; Lubniewski, Z.; Schneider von Deimling, J. Measurement of Seafloor Acoustic Backscatter Angular Dependence at 150 kHz Using A Multibeam Echosounder. *Remote Sens.* **2021**, *13*, 4771. [CrossRef]