remote sensing

MDPI

Article

Converging Channel Attention Mechanisms with Multilayer Perceptron Parallel Networks for Land Cover Classification

Xiangsuo Fan [1,2], Xuyang Li [1,*], Chuan Yan [1], Jinlong Fan [3], Lin Chen [1] and Nayi Wang [1]

[1] School of Automation, Guangxi University of Science and Technology, Liuzhou 545006, China
[2] Guangxi Collaborative Innovation Centre for Earthmoving Machinery, Guangxi University of Science and Technology, Liuzhou 545006, China
[3] National Satellite Meteorological Center, China Meteorological Administration, Beijing 100081, China; fanjl@cma.gov.cn
* Correspondence: 221077062@stdmail.gxust.edu.cn

Abstract: This paper proposes a network structure called CAMP-Net, which considers the problem that traditional deep learning algorithms are unable to manage the pixel information of different bands, resulting in poor differentiation of feature representations of different categories and causing classification overfitting. CAMP-Net is a parallel network that, firstly, enhances the interaction of local information of bands by grouping the spectral nesting of the band information and then proposes a parallel processing model. One branch is responsible for inputting the features, normalized difference vegetation index (NDVI) and normalized difference water index (NDWI) band information generated by grouped nesting into the ViT framework, and enhancing the interaction and information flow between different channels in the feature map by adding the channel attention mechanism to realize the expressive capability of the feature map. The other branch assists the network's ability to enhance the extraction of different feature channels by designing a multi-layer perceptron network based on the utilization of the feature channels. Finally, the classification results are obtained by fusing the features obtained by the channel attention mechanism with those obtained by the MLP to achieve pixel-level multispectral image classification. In this study, the application of the algorithm was carried out in the feature distribution of South County, Yiyang City, Hunan Province, and the experiments were conducted based on 10 m Sentinel-2 multispectral RS images. The experimental results show that the overall accuracy of the algorithm proposed in this paper is 99.00% and the transformer (ViT) is 95.81%, while the performance of the algorithm in the Sentinel-2 dataset was greatly improved for the transformer. The transformer shows a huge improvement, which provides research value for developing a land cover classification algorithm for remote sensing images.

Keywords: CAMP-Net; land use; channel attention; multilayer perceptron; parallel networks

Citation: Fan, X.; Li, X.; Yan, C.; Fan, J.; Chen, L.; Wang, N. Converging Channel Attention Mechanisms with Multilayer Perceptron Parallel Networks for Land Cover Classification. *Remote Sens.* **2023**, *15*, 3924. https://doi.org/10.3390/rs15163924

Academic Editor: Michalis Savelonas

Received: 26 June 2023
Revised: 1 August 2023
Accepted: 4 August 2023
Published: 8 August 2023

1. Introduction

Today's remote sensing (RS) technology provides us with a large amount of earth observation data, such as satellite images and LiDAR data. These data can not only provide environmental monitoring on a global scale, but can also be used in application areas such as land cover classification and change detection and natural disaster monitoring and assessment [1–4]. Additionally, improving the accuracy of land cover classification is extremely important for fine agricultural division, earth observation, regional environmental protection, and urban planning tasks [5–8].

In land cover classification, traditional classification methods mainly use shallow features such as pixel color and texture to classify images [9–13]. In addition, machine learning-based algorithms can also be used for RS image classification, including support vector machines (SVMs) [14], random forests (RFs) [15], K-means clustering (K-Means) [16], K-nearest neighbor (KNN) [17], etc. SVM utilizes the constrained optimal solution problem

for hyperplane classification, which is suitable for linearly divisible problems. RF obtains results by constructing multiple decision trees and voting. K-Means uses distance metrics for clustering analysis, but it is sensitive to abnormal samples and cannot deal with discrete features or guarantee global optimality. KNN uses the measurement of distances between different features to classify, but it cannot deal with multi-sample problems. However, these traditional machine learning algorithms still need to further improve the model's generalization ability to cope with more complex RS image classification tasks. The deep learning (DL) model plays an indispensable role in remote sensing image processing. Due to its multilevel learning property, it can accurately approximate nonlinear relationships [18], thus realizing applications such as classification, fusion, and downscaling [19–21]. Deep learning has been successfully applied to land cover classification tasks in RS images, such as U-Net [22], which is a classical fully convolutional network that can obtain good classification results with a lower amount of training data, and thus has become one of the most widely used algorithms in RS image segmentation tasks. Deng et al. introduced a method for land use and cover classification using the U-Net network. A weighted loss function was introduced to address the category imbalance in the data, and data-enhancement techniques were used to improve the robustness of the model [23]. Wu et al. introduced a deep learning method based on an improved version of the U-Net neural network model for the task of sugarcane segmentation in multispectral unmanned aerial vehicle (UAV) images [24]. Hu et al. used the U-Net neural network model based on U-Net for oil palm tree detection in high-resolution remote sensing images [25]. Schiefer et al. used a deep learning method with a convolutional neural network and U-Net for rice fall detection [26].

For complex or heterogeneous feature types, because the feature extraction problem may cause U-Net and improved algorithms to have a large bias on the prediction results since the emergence of transformer [27], it has gradually become the mainstream of RS remote sensing classification. Transformer has a powerful sequence modeling capability and the ability to perceive the global information of the input sequences at the same time, uses the attention mechanism to model the global dependencies of input and output, and uses position coding to solve the problem of how to represent the relative or absolute positional relationships of elements in the sequence. In 2020, Google proposed vision transformer (ViT) [28], which successfully applied transformer to the image classification model and became the mainstay of landmark work of transformer application in the CV field. Hong et al. [29] proposed a new backbone network SpectralFormal (SF), which is able to learn the spectral local sequence information from the neighboring bands of HS images, thus generating the component spectral embeddings and obtaining good classification results.

Although many ViTs achieve good classification results, for example, [30] proposes the ASRC-Net architecture, which fuses CNN with ViT, adopts weight sharing, and uses the same way of processing at different image locations to help the transformer utilize the data more effectively. Considering the advantages of ViT, it can be combined with MLP. ViT has the advantage of being able to consider the relationship between all pixels in an image at both small and large scales, but the disadvantages are also obvious. ViT needs to be learned through a lot of training and because of its increased computational cost, it requires a higher investment of resources.It is sensitive to positional coding: the ViT uses positional coding to assign to each position a specific marker so that the model can recognize the relative position of each location. However, position coding can be affected by operations such as rotation, scaling, and translation, which can lead to a degradation of the model's performance. Meanwhile, effectively combining ViT with MLP has several advantages. For better feature extraction capability, MLP has an excellent feature extraction capability in image processing, while ViT performs well in natural language processing. Combining them can further improve the accuracy of image processing; faster training speed: ViT uses a self-attention mechanism to process sequence data, which improves processing speed. When combined with MLP, different features in the image can be better captured, thus

speeding up the training of the model; better adaptability: MLP and ViT have each been successful in different domains, and in combining them, the model can be made more generalized. This also means that the model can be applied to tasks in different domains, thus further improving the usefulness of the model.

Therefore, in this paper, combining MLP and ViT can utilize the strengths of both to improve the feature extraction capability and training speed, while also providing better adaptability and lower computational cost. The main contributions of the paper include:

(1) The problem of multispectral image classification is revisited from the perspective of improving the feature extraction capability, and a parallel network architecture integrating the channel attention mechanism and the multilayer perceptron based on the utilization of the feature channels is proposed, which effectively exploits the global correlation information of the image and the feature information between different channels and fully integrates the spatial and channel location correlation. It also allows the lexical features to have a richer expression, so that the pixel-level image classification can be better realized. the level image classification and the comprehensive classification accuracy can both reach 99.00%.

(2) Adding short wave infrared radiometer (SWIR) bands to the commonly used RGB + NiR bands to form the input seven bands, which can deeply excavate the a priori potentially useful information, and is more conducive to the classification of specific land cover in the precise study area.

(3) The analysis of land use dynamics changes in the study area focused on the changes in the distribution area of rapeseed and vegetable during the period of 2018–2022. Land use rate and land cover changes can vividly demonstrate the relationship between local economic development and conservation of ecological diversity. These data provide strong support for highly standardized aquaculture and rationalized farmland construction, regularized land improvement projects, and effective decision-making by relevant departments.

The rest of this paper is organized as follows. Section 2 will introduce the materials used for the study and the methods used. Section 3 will present our findings and provide further explanations. Section 4 will explore the potential research directions and future perspectives of these results. Finally, Section 5 will summarize the main findings of this paper and provide suggestions for future research.

2. Materials and Methods

In this section, first, we will briefly introduce the details of the study area and data-processing operations. Subsequently, we will detail the network architecture and the optimization of the framework.

2.1. Study Area Overview

South County, which belongs to Yiyang City of Hunan Province, is located at the border of Xiang and Hubei Provinces; the hinterland of Dongting Lake area; is connected with Shishou, public security, and Songzi of Hubei Province in the north; Anxiang and Hanshou counties of Changde City in the west; Huarong County of Yueyang City in the east; Yuanjiang City of Yiyang City in the south; and several large agricultural (fishing) farms, such as Datong Lake, Beizhouzi, Jinpan, Nanwan Lake, and Qianshanhong in the southeast. It is one of the 36 border counties in Hunan Province. South County exists between 29°3′3″N and 29°31′37″N latitude and between 112°10′53″E and 112°49′6″E longitude, with a total area of 1321 km^2. It is close to Yueyang in the east, Changsha in the south, and the Yangtze River in the north. It is a subtropical transition to monsoon humid climate type, with cool winters and warm summers, four distinct seasons, sufficient heat, abundant rainfall, long sunshine hours, and short frost periods. It is very suitable for aquaculture and crop cultivation.

Xiangyin County, known as Luocheng in ancient times, is a county under the jurisdiction of Yueyang City, Hunan Province. It is located in the northeastern part of Hunan

Province, between the Xiangjiang River and the Zijang River, on the southern shore of Dongting Lake, between 28°30′13″N and 29°3′2″N latitude and between 112°30′20″E and 113°1′50″E longitude. The Xiangjiang River runs from the south to the north, dividing the county into two parts: the eastern part is hilly and the western part is a lakeside plain, with similar types of features as in South County, which are suitable for aquaculture and crop cultivation. In this experiment, the most abundant areas of South County and Xiangyin County were selected for the following experiments. The study area is shown in Figure 1.

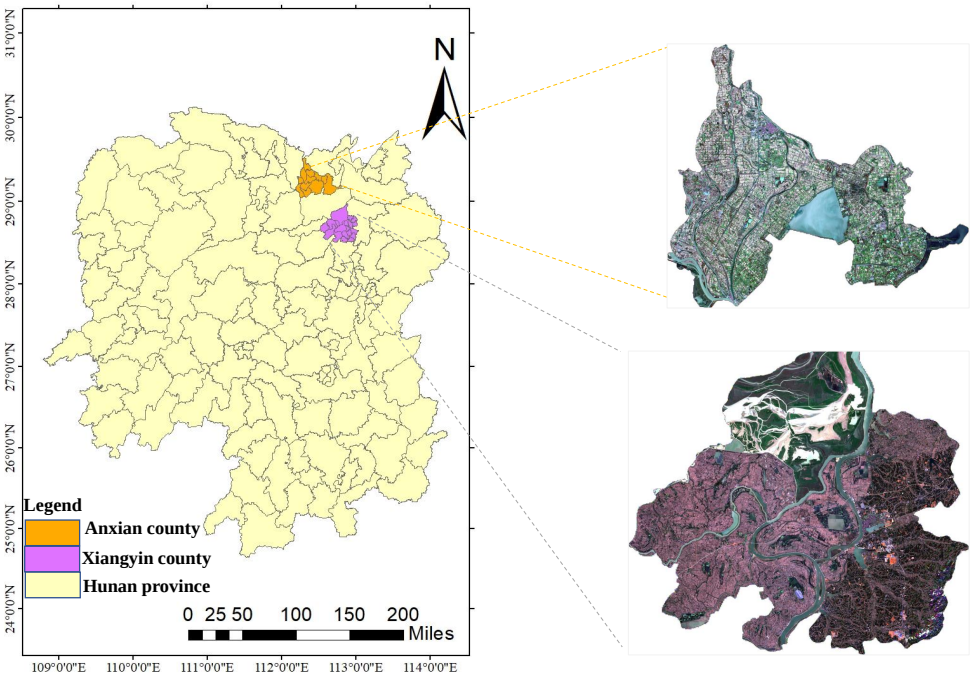

Figure 1. Location of the study area and Sentinel-2 remote sensing image (6 October 2021).

2.2. Remote Sensing Image Pre-Processing

To obtain high-quality remote sensing image training samples and accuracy verification samples, field surface cover sample point data are crucial, and the quality of these data directly determines the accuracy of classification. For study area feature types, field data were collected in the field in October 2021, and aquaculture and planting areas with areas larger than 100 m^2 were selected as sample points with priority. This helps to obtain better training and accuracy validation samples.

This paper uses multispectral images as a data source with a resolution of 10 m taken on 6 October 2021, by the Sentinel-2 satellite covering the study area. The satellite carries a multispectral instrument (MSI) containing 13 bands with pixel sizes ranging from 10 to 60 m. The blue (B2), green (B3), red (B4), and NIR (B8) bands have a resolution of 10 m, while the red end (B5), NIR (B6, B7, and B8A), and shortwave infrared SWIR (B11 and B12) are sampled on the ground at 20 m. The blue (B2), green (B3), red (B4), and near-infrared (B8) bands have a resolution of 10 m; the red end (B5), near-infrared NIR (B6, B7, and B8A) and short-wave infrared SWIR (B11 and B12) have a ground-based sampling distance of 20 m; and the pixel sizes of the coastal atmospheric aerosol (B1) and cirrus cloud bands (B10) are 60 m. The pixel size of the coastal atmospheric aerosol (B1) and cirrus cloud bands is 60 m [31] as shown in Table 1.

Table 1. Sentinel 2 image band information.

Wave Band	Resolution	Center Wavelength	Descriptive
B1	60 m	443 nm	ultramarine
B2	10 m	490 nm	blue
B3	10 m	560 nm	greener
B4	10 m	665 nm	red
B5	20 m	705 nm	VNIR
B6	20 m	740 nm	VNIR
B7	20 m	783 nm	VNIR
B8	10 m	842 nm	VNIR
B8A	20 m	865 nm	VNIR
B9	60 m	940 nm	SWIR
B10	60 m	1375 nm	SWIR
B11	20 m	1610 nm	SWIR
B12	20 m	2190 nm	SWIR

The Level 1C data were first atmospherically corrected using Sen2Cor-02.10.01-win64 software to remove atmospheric effects on the images, and steps such as image enhancement and mask extraction were performed. Then, the band resolution was resampled to 10 m resolution, and finally, eight commonly used bands (Band2, Band3, Band4, Band8, Band9, Band11, and Band12) were selected for band synthesis using ENVI software. Through a priori knowledge and field data, the region of interest (ROI) was labeled to obtain the sample bank data from South County, Yiyang City, where the sample bank contains 16,510 samples in 7 categories, divided into a training set (70%) and a test set (30%). Xiangyin County in Yueyang City is also mentioned.

2.3. The Architecture Proposed in This Paper CAMP-Net

To address the problems that traditional deep learning algorithms cannot weigh the pixel information of different bands, coupled with the high feature dimensionality of the original remote sensing images as the direct classification of them easily leads to overfitting and high computational complexity, the fused channel attention mechanism and multilayer perceptron parallel network architecture CAMP-Net proposed in this paper is shown in Figure 2, which consists of NDVI, NDWI, and MLP. The structure consists of NDVI, NDWI, MLP and the transformer with channel attention. The band information of NDVI and NDWI is directly fused with the input of the transformer; the features obtained by the transformer encoder are merged with the features obtained by MLP; the merged features are reduced in dimensionality by a fully connected layer; and finally, the classification results are obtained by activation function and MLP-head to achieve pixel-level multispectral image classification.

2.3.1. NDVI and NDWI

In the field of remote sensing, we often use vegetation index (VI) to quantitatively assess the growth of vegetation. One of the most commonly used vegetation indices is the normalized difference vegetation index (NDVI), which determines the health and coverage of vegetation by calculating the reflectance of different wavelength bands. NDVI is commonly used to monitor the growth status of various ecosystems such as large-scale agriculture, forestry, grassland, and wetlands.

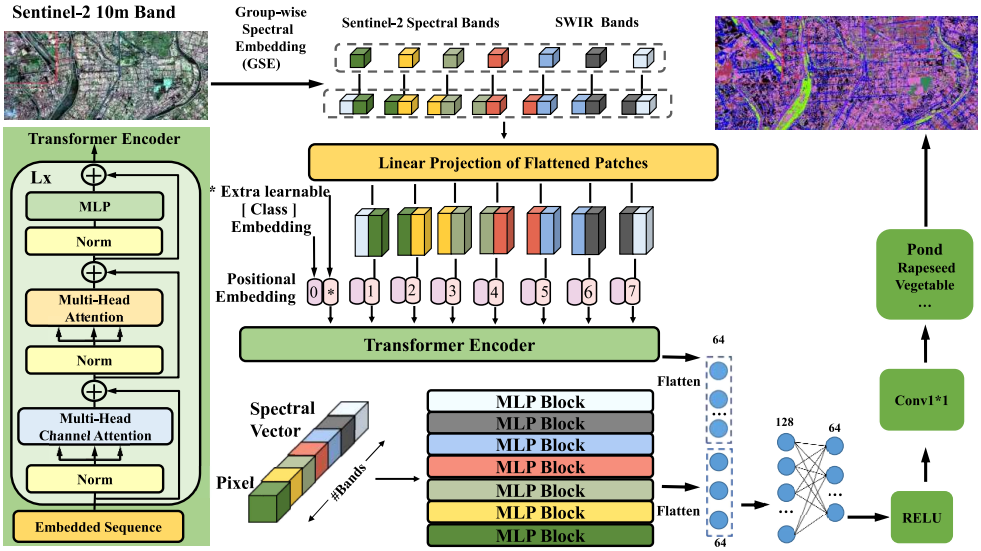

Figure 2. Schematic diagram of the CAMP-Net architecture for multispectral image classification tasks. Where * stands for multiplication.

The value of NDVI is calculated by Equation (1) to obtain the calculation formula:

$$NDVI = \frac{(NIR - RED)}{(NIR + RED)} \tag{1}$$

where NIR indicates near-infrared band reflectance and RED indicates red band reflectance. NDVI takes values between -1 and 1, with higher values representing more vegetation coverage and better health. If the NDVI is 0, it means that the vegetation coverage is 0. A negative number means that the area is mainly composed of non-vegetation types such as water bodies, buildings, and roads.

The full name of NDWI is "Normalized Difference Water Index", which means the normalized water body index. In remote sensing, this index is used to monitor water bodies and wetlands and can help detect floods, droughts, snow and ice, soil moisture content, etc. The value of NDWI is calculated by Equation (2):

$$NDWI = \frac{(Green - NIR)}{(Green + NIR)} \tag{2}$$

where NIR represents the reflectance of the NIR band and Green represents the reflectance of the green band. In images formed by remote sensing images of visible (green and blue) and infrared band signals under cloudless skies, NIR is less affected by the atmosphere and easier to obtain; there is a natural correlation between the green band and NIR band reflectance, which can be described by an equation. Higher water content corresponds to index values ranging from -1 to +1, with higher values representing more water bodies in the region and lower values if the region is all land.

The steps for fusion using NDVI and NDWI band information in the VIT framework are as follows. First, image data containing NDVI band information need to be prepared. Calculating NDVI and NDWI depends on the remote sensing data used. After calculating NDVI, it is added to the image as an additional band to use as an input to the VIT model. Usually, all bands can be superimposed into a four-dimensional array (i.e., tensor), where the first dimension represents the number of images, the second and third dimensions represent the height and width of the image, and the fourth dimension represents the number of bands (including NDVI and NDW bands). Then, NDVI and NDW are incorporated into the ViT framework as band information.

2.3.2. Multi-Layer Perceptron Based on Feature Channels

MLP based on feature channels is a common artificial neural network model that can be used for tasks such as classification and regression. It comprehends patterns in input data and uses feature channels for learning. Each feature channel corresponds to one dimension of the input data, such as R, G, and B of a color image.

The feature-channel-based multilayer perceptron is a neural network model consisting of an input layer, multiple hidden layers, and an output layer that can handle nonlinear data and adapt to complex data distributions. The model is highly flexible and performs well in different tasks because it does not restrict the relationship between features. In each hidden layer, the result of the weighted summation of the inputs is converted to the output by an activation function, and the output of the previous layer is used as input for the next layer. The output of the last hidden layer is fed to the output layer for classification or regression. With the ability to handle nonlinear data without restricting the relationship between features, it can adapt to complex data distributions, and the model is highly flexible and performs well in different tasks.

Traditional MLPs use sigmoid activation functions but may lead to gradient disappearance in backpropagation. A better choice is a variant such as GELU, which we use as well as regularization, while L1 and L2 regularization prevents overfitting and improves generalization; applying dropout reduces neuronal correlation and improves model robustness, as shown in Figure 3.

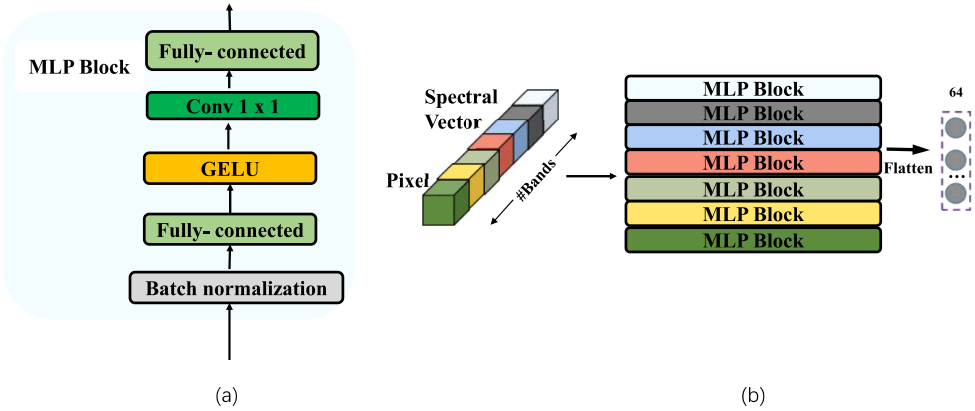

Figure 3. Schematic diagram of the structure of the multilayer perceptron based on feature channels: (**a**) MLP block. (**b**) MLP branch structure.

2.3.3. Channel Attention

The transformer is a deep learning architecture that has demonstrated impressive performance in various tasks such as image recognition and semantic segmentation. The transformer was originally designed for sequence-by-sequence problems, and its core consists of an attention mechanism and positional coding. In the field of image processing, it is mainly used for sequence-to-sequence problems. Due to its self-attention and positional encoding, the transformer model outperforms traditional RNN models because it does not require the input and output sequences to be fully aligned, which provides significant benefits for processing sequential data. In addition, the transformer's structure can understand the interactions between different spatial locations and solve the long-term dependency problem thanks to its parallel computing capability, which significantly reduces computational resource consumption.

Our proposed algorithm uses the transformer encoder as an encoder with N encoder layers, each including multiple self-attentive modules and feedforward layers, as well as inputs including position encoding and feature embedding, with the transformer using a fixed position encoding to represent the absolute position information of the token. The most "core" channel attention structure is used, which is the epitome of parallel attention. Channel attention is not limited to calculating one-time attention since the

method fully considers the interrelationship between all spatial locations. Channel attention is used for input F with the maximum pooling layer and the average pooling layer, in the same MLP network after the output, in the sum after the activation function to obtain the final output as shown in Figure 4.

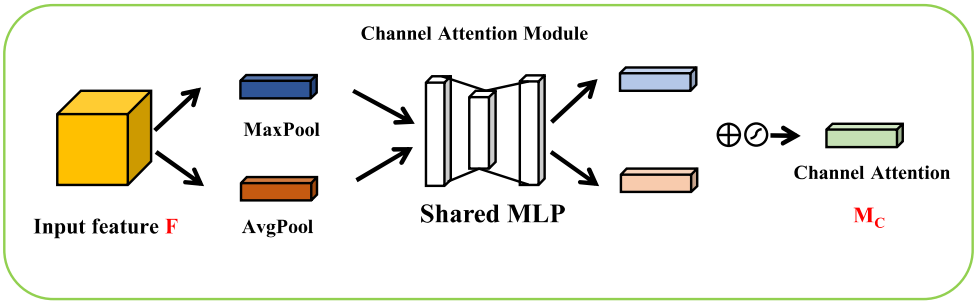

Figure 4. Channel Attention Module.

Channel attention is an attention mechanism in neural networks that enhances the interaction and information flow between different channels in a feature graph to improve the performance of the model. By applying channel attention, the relationship between different channels in a feature map can be encoded to eliminate redundant information, focus on important features, and improve the expressiveness of the feature map. The basic idea of channel attention is to calculate the importance weight of each channel in the feature map by a "compression" operation, then to weigh each channel by a "stimulation" operation, and, finally, to output the weighted feature map. This "compress-stimulate" attention mechanism is simple to implement and also has good robustness and scalability.

The benefits of changing the ViT self-attentive mechanism to channel-attentive are as follows:

(1) Better capture of image features: By using channel attention instead of the self-attention mechanism, the relationship between different channels (i.e., different color channels) can be better captured and the attention on important channels can be improved. This is important in vision tasks where different color channels may affect the image features that need to be learned.
(2) Higher computational efficiency: The self-attention mechanism needs to calculate the similarity scores between all position pairs, so it is computationally more expensive. However, using the channel attention mechanism can allocate attention to different channels, thus making the model more computationally efficient.
(3) Better generalization ability: Changing the self-attention mechanism to the channel-attention mechanism can also improve the generalization ability of the model. This is because the channel attention mechanism allows global evaluation of the whole image, thus improving the model's ability to understand more complex and variable visual scenes.

The channel attention mechanism implementation can be expressed as Equations (3)–(5):

$$z = \frac{1}{c}\sum_{i=1}^{c} x_i \tag{3}$$

$$w = \sigma(W_2 ReLU(W_1 z)) \tag{4}$$

$$\gamma = w \odot x \tag{5}$$

where x is the input tensor with dimensions (N, C, H, and W), and N denotes the batch size. c indicates the number of channels, and H and W indicate the height and width, respectively. z is the vector obtained by global average pooling of x, the average over the channel dimensions. W_1 and W_2 are the learnable weight matrix, and $ReLU$ denotes the

modified linear cell operation. σ is the sigmoid function. γ is the output tensor, which is the input tensor using attentional weighting. $\odot$ denotes multiplication by elements. The channel attention mechanism uses two fully connected layers and a sigmoid function, which can be understood as the model automatically learns the importance of each channel and enhances the signal of useful channels and cuts the information of useless channels by weighting the average, thus improving the model performance.

2.3.4. Evaluation Indicators

For the evaluation of the classification of multispectral pixels, three hybrid matrix-based evaluation metrics are used, including overall accuracy, average accuracy, and kappa. Pixel-level evaluation metrics such as accuracy, precision, and kappa are used mainly to evaluate the ability of the proposed model to classify accurately.

(1) Overall accuracy (OA): The number of correct predictions as a percentage of the number of all predictions.

$$Overall\ Accuracy = \frac{T1 + T2}{T1 + F1 + T1 + F2} \tag{6}$$

where T_1 is true positive, F_1 is false positive, T_2 is true negative, and F_2 is false negative. T_1 + T_2 denotes the total number of correctly classified samples. F_1 + F_2 denotes the total number of misclassified samples. In the vast majority of cases, a higher OA value indicates a better performance of the classifier. However, in some cases, the OA metric may be less appropriate due to the imbalance of data distribution, for example, a very small percentage of samples in a certain category leads to a high OA value when the classifier only needs to predict all the samples in that category as another category. Therefore, in this case, other metrics need to be used to evaluate the model performance comprehensively.

(2) Average accuracy (AA): The average accuracy, which is a more accurate evaluation metric. Unlike OA, AA values are calculated separately on each category and then averaged as the final score of the model. Its calculation formula is as follows.

$$Average\ Accuracy = \frac{1}{n}\frac{T1 + T2}{T1 + F1 + T1 + F2} \tag{7}$$

where n denotes the number of categories.

(3) Kappa coefficients: Kappa is a measure of the accuracy of a classifier or evaluation system, also known as Cohen's kappa coefficient, which evaluates the consistency of a classification model with its predictions in a real test. The kappa coefficient takes values in the range [−1,1]. The formula is as follows.

$$Kappa\ = \frac{P_l - P_x}{1 - P_x} \tag{8}$$

P_l denotes the exact match of the observed data, i.e., the proportion of samples for which the classifier predicts exactly the same result as the actual situation. P_x is the likelihood that the classifier predicts the classification and the actual situation to obtain consistent results.

2.3.5. Loss Function

Cross-entropy is a method used to measure the variability between probability distributions. The output in a multiclassification task is the probability that the target belongs to each category, and the sum of the probabilities of all categories is 1, in which the category with the highest probability is the one to which the target belongs. We choose cross-entropy as the loss function in our algorithm, which is useful in its multiclassification task because it allows the model to better distinguish the differences between categories, and it can be used to minimize the loss function by optimization algorithms such as gradient descent to

improve the accuracy of the model. The cross-entropy loss function compares the predicted class of each pixel with the target class and has the following expression:

$$J = -\frac{1}{X}\sum_{i=1}^{M} y_{ic}\sum_{c=1}^{x} y_{lx}\log\left(\hat{p}_{xy}\right) \quad (9)$$

where X indicates the number of sample categories; x denotes the number of categories, and y_{lx} denotes the true label of the lth sample belonging to class x. $\hat{p}_{xy}$ is the probability value indicating that the model predicts that the sample belongs to class y, between 0 and 1.

2.3.6. Experimental Environment

In the experiments, the coordinates of the region of interest are obtained using ENVI software, and the Adam optimizer [32,33] is applied in the training phase, the batch size is set to 32, the maximum number of iterations is 300, and the initialized learning rate is 5×10^4 , decaying by 0.9 at every 30 iterations. The implementation code is written using Python 3.9 in PyTorch 1.12.1 The training model was written on a Win11 + 12th Gen Intel(R) Core(TM) i5-12400F + NVIDA GeForce RTX 3060 Laptop GPU.

3. Results

The details of the sample pool in the study area are shown in Table 2, with 70% used for training and 30% for testing. The following experiments are conducted using the sample size of the dataset in Table 2.

Table 2. Sample size of the study area 2021 data set.

NO.	Class	South County		Xiangyin Xian	
		Training	Testing	Training	Testing
1	Building	1330	570	2762	1185
2	Water	775	333	3738	1603
3	Lotus	1676	719	1374	589
4	Pond	3079	1320	2920	1252
5	Wetland	1304	559	1407	603
6	Vegetable	1759	755	1768	759
7	Rapeseed	1631	700	2107	903
Total		11,554	4956	16,076	6894
		16,510		22,970	

To highlight the good performance effect improvement of our proposed algorithm, CAMP-Net is compared with SVM, RF, KNN, CNN, RNN, ViT, and SF in this paper.

3.1. Ablation Study

To evaluate the performance of the proposed network structure in this paper, ablation experiments were performed on the dataset in the study area. The experimental results are shown in Table 3. The OA of the original ViT is 95.81%, which is a satisfactory result, indicating that ViT is more suitable for handling multispectral image-classification problems. Although the OA of ViT + CA was improved to 96.76, it also proves the effectiveness of the ViT + CA scheme, which can enhance the tight connection of channels by adding channel attention. The OA of the MLP module inserted into ViT is improved by 1.10% compared with the original ViT, and AA and kappa are also improved, indicating that in multispectral data, MLP may improve ViTs classification of a subset of categories. The feature fusion of the MLP module with ViT incorporating NDVI and NDWI band information achieved an AA of 98.15, which is a 2.34% improvement relative to ViT. Kappa also improves by 1.80%

over ViT, thus verifying that the feature fusion of the MLP module with ViT incorporating NDVI and NDWI band information compares with ViT for multispectral image pixel classification task is more advantageous. However, the OA is improved when the attention of ViT is changed to channel attention and the band information of NDVI and NDWI is added. The OA improves by 98.63% when the attention of ViT is changed to channel attention with MLP first processed in parallel and then fused features, but the accuracy of CAMP-Net is 99.00 when all modules are tested together, which is an OA improvement compared to the original ViT 3.19%.AA and kappa also have a good improvement, so the achieved results are relatively good.

Table 3. Results of ablation experiments on the study area dataset using different combinations of modules of CAMP-Net.

Different Methods	Different Module			Metric		
	CA	MLP	NDVI + NDWI	OA (%)	AA (%)	Kappa
ViT	×	×	×	95.81	96.05	94.98
CAMP-Net	✓	×	×	96.76	97.14	96.12
CAMP-Net	×	✓	×	96.91	97.18	96.30
CAMP-Net	×	✓	✓	98.15	98.29	97.78
CAMP-Net	✓	×	✓	98.33	98.40	98.00
CAMP-Net	✓	✓	×	98.63	98.87	98.36
CAMP-Net	✓	✓	✓	**99.00**	**99.08**	**98.81**

Where × means that the part is not used, and the check mark indicates that the part is used.

The classification results of different module pair combinations on the study area dataset are shown in Figure 5, where water, rapeseed, and pond are present. Comparing b and c shows that NDVI + NDWI classifies lotus better than ViT. The network structure of Figure 5c–e alone plus the modules does not classify pond as well as Figure 5f–h. The complete CAMP-Net outside is more distinguishable than the combination of his modules. The difficulty of the classification task for this region is the mixture of pond and vegetable in this region and the performance of the complete CAMP-Net is the best, with a good denoising effect, which can improve the performance of ViT better.

To explore the effect of the number of neighboring spectra on CAMP-Net and ViT, different numbers of neighboring spectra were taken for the experiments, and the values of the number of neighboring spectra were taken as 1, 2, 3, 4, 5, 6, 7, and the experimental results are shown in Table 4. According to the experimental results, it is known that ViT, AA, and OA, and kappa are the highest when the number of neighboring spectra is 5. For CAMP-Net, the overall results of CAMP-Net were the best when the number of neighboring spectral bands was 6. The experimental results show that the performance is not necessarily better if the number of neighboring spectral bands is higher, but the selection of the appropriate number of neighboring bands has some improvement on the model performance. Therefore, we chose the number of neighboring bands as six in the later experiments.

To evaluate the effect of small sample size training samples on the experimental results, we randomly selected 10–90% of samples from South County's sample pool for training validation, and the remaining samples were used for testing when no separate validation set was set up, and the experimental results are shown in Table 5. It is not the case that the proportion of samples increases for better results, and as the proportion of classification samples increases with the training samples, AA does not necessarily better effect and the noise generated is also randomly assigned. Since random sample selection leads to random sample point selection, randomly selected training sample points may not contain all sub-regions of ROI, leading to obvious misclassification, such as (c) misclassifying water

into a pond. so to make the best overall classification effect, we choose a sample ratio of 7:3 for the experiment.

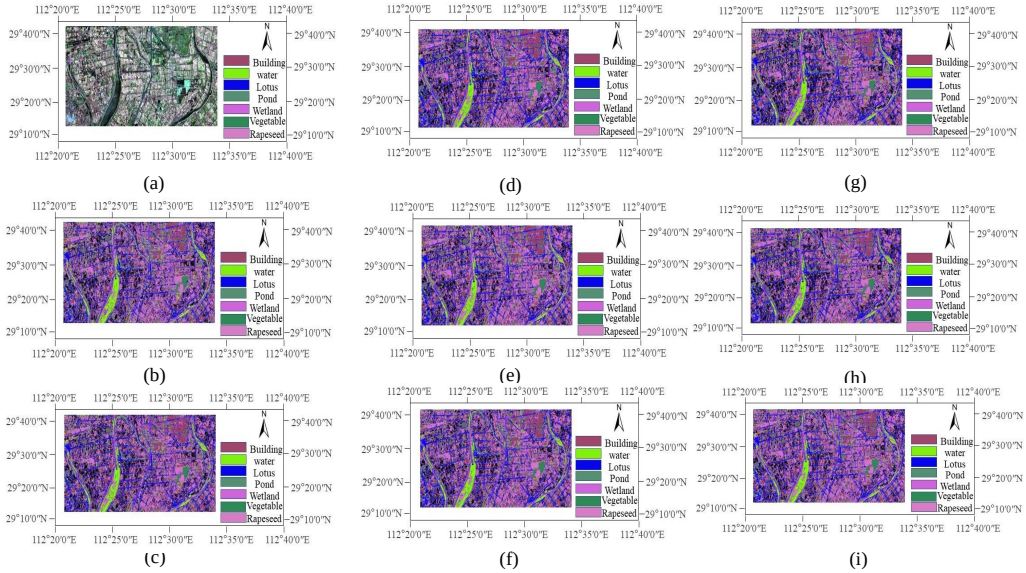

Figure 5. Plot of the classification results of CAMP-Net using different combinations of modules for the study area dataset: (**a**) Image. (**b**) ViT. (**c**) CAMP-Net (NDVI + NDWI). (**d**) CAMP-Net (MLP). (**e**) CAMP-Net (CA). (**f**) CAMP-Net (MLP + NDVI + NDWI). (**g**) CAMP-Net (CA + NDVI + NDWI). (**h**) CAMP-Net (CA + MLP). (**i**) CAMP-Net (CA + MLP + NDVI + NDWI).

Table 4. Effect of the number of neighboring bands on ViT as well as CAMP-Net.

The Number of Neighboring Bands	Class No.							Metrics		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**OA (%)**	**AA (%)**	**Kappa**
1 (ViT)	96.24	97.67	93.25	95.90	98.69	94.82	95.76	95.81	96.05	94.98
1 (CAMP)	98.04	98.96	96.59	96.55	99.46	95.84	97.30	97.22	97.54	96.67
2 (ViT)	97.51	98.58	94.27	97.85	98.46	96.53	97.73	97.20	97.28	96.64
2 (CAMP)	99.39	98.70	97.61	95.19	99.61	96.07	99.14	97.46	97.96	96.96
3 (ViT)	98.64	98.96	96.42	96.97	**99.84**	94.65	98.09	97.35	97.66	96.83
3 (CAMP)	97.51	99.22	94.92	97.85	99.46	97.27	98.52	97.67	97.83	97.21
4 (ViT)	97.96	98.96	93.61	95.29	98.77	97.49	99.20	96.88	97.33	96.27
4 (CAMP)	97.96	99.22	95.88	96.78	99.53	96.41	96.62	97.19	97.49	96.63
5 (ViT)	97.59	97.54	95.88	96.91	99.15	97.21	98.52	97.41	97.55	96.90
5 (CAMP)	98.94	98.83	95.52	97.59	98.69	96.70	95.64	97.25	97.42	96.70
6 (ViT)	98.57	97.80	96.00	96.42	98.08	97.04	95.40	96.84	97.05	96.21
6 (CAMP)	**99.92**	**99.61**	**98.21**	**99.15**	99.46	97.89	**99.35**	**99.00**	**99.08**	**98.81**
7 (ViT)	98.49	99.35	95.64	97.72	98.77	95.67	96.50	97.26	97.45	96.71
7 (CAMP)	98.87	99.35	95.28	97.10	99.53	**97.89**	98.46	97.78	98.08	97.35

To evaluate the effect of small-sample-size training samples on the experimental results, we randomly selected 10–90% of samples from South County's sample pool for training validation, and the remaining samples were used for testing when no separate validation set was set up; the experimental results are shown in Table 5. It is not the case that the proportion of samples increases for better results, and as the proportion of

classification samples increases with the training samples, AA does not necessarily have a better effect and the noise generated is also randomly assigned. Since random sample selection leads to random sample point selection, randomly selected training sample points may not contain all sub-regions of ROI, leading to obvious misclassification, such as (c) misclassifying water as a pond. Therefore, to make the best overall classification effect, we choose a sample ratio of 7:3 for the experiment.

Table 5. CAMP-Net test results using different training sample proportions in the 2021 South County dataset.

Ratio of Training	Class No.							Metrics		
	1	2	3	4	5	6	7	OA (%)	AA (%)	Kappa
10%	98.42	93.63	97.48	97.26	98.38	93.62	99.14	97.03	96.85	96.44
20%	98.68	98.19	97.07	97.61	1.00	93.62	96.13	97.15	97.33	96.59
30%	99.82	99.69	95.26	97.49	99.10	97.34	99.14	97.98	98.27	97.58
40%	99.07	99.54	97.07	97.49	**99.86**	97.11	97.74	98.00	98.28	97.60
50%	98.52	99.63	95.57	97.27	98.92	96.34	98.96	97.61	97.89	97.14
60%	98.42	99.69	91.44	97.42	99.73	**98.14**	98.78	97.38	97.66	96.87
70%	**99.92**	**99.61**	**98.21**	**99.15**	99.46	97.89	**99.35**	**99.00**	**99.08**	**98.81**
80%	99.40	**99.77**	96.60	98.23	99.59	96.12	98.06	98.05	98.26	97.66
90%	98.71	99.49	96.89	97.37	99.82	96.24	98.85	97.91	98.20	97.50

The results of training samples using different ratios are shown in Figure 6. One can see from the figure that difference between the ratios of the training samples is very small, so in the case of the results of the graph, there is not much difference. We chose to test the results of the higher accuracy of the sample ratio of 7:3 for the experiment.

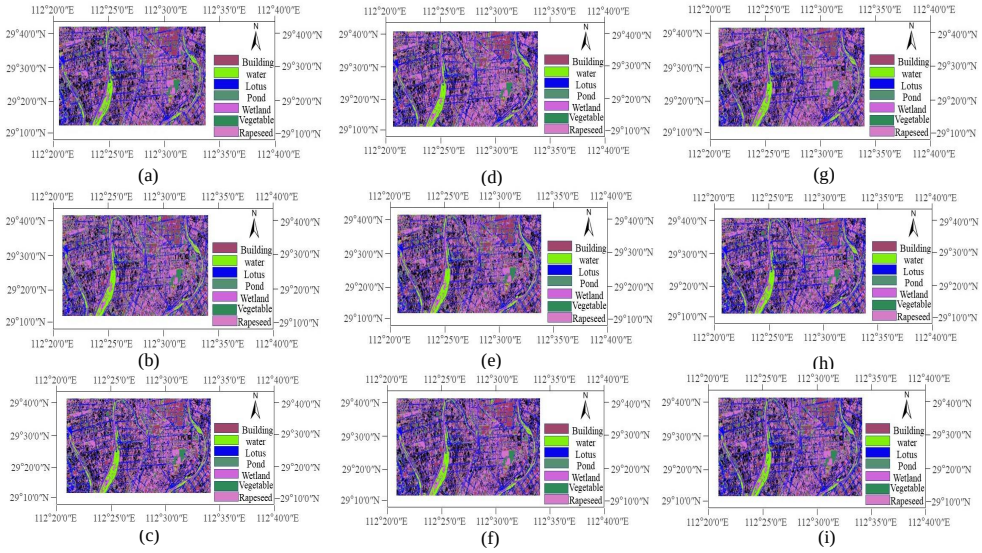

Figure 6. Plots of the results of different proportions of training samples: (**a**) 10%. (**b**) 20%. (**c**) 30%. (**d**) 40%. (**e**) 50%. (**f**) 60%. (**g**) 70%. (**h**) 80%. (**i**) 90%.

The data source for the study area taken by the Sentinel-2 satellite contains a total of 13 spectral bands, but other satellites do not necessarily contain as many. However, they all must contain RGB and NiR bands. In order to increase the versatility of the dataset, researchers usually select four common bands, RGB + NiR, from the 13 bands as the pixel sequence information for experimental studies. To determine whether it is worthwhile to prioritize fast processing and broad applicability of the four common Sentinel-2 bands and discard the a priori potentially useful information recorded by Sentinel-2, we first added

the VNIR band for the experiment, and then the short-wave infrared (SWIR) bands for comparison. This also helps to classify specific land cover in the precise study area when more Sentinel-2 bands are used. The experimental results are shown in Table 6. When all SWIR bands are added, a significant improvement is obtained with very good results, except for vegetables.

Table 6. Test results in the 2021 South County dataset using different methods in different bands.

Different Bands (Method)	Class No.							Metrics		
	1	**2**	**3**	**4**	**5**	**6**	**7**	**OA (%)**	**AA (%)**	**Kappa**
4 bands (ViT)	76.69	91.61	79.89	82.26	93.32	91.52	88.77	85.49	86.30	82.59
4 bands (CAMP-Net)	79.54	92.38	83.83	86.06	92.56	91.64	88.35	87.32	87.77	84.77
4 bands + VNIR (ViT)	91.95	94.32	92.78	89.05	97.77	96.41	93.07	92.95	93.63	91.56
4 bands + VNIR (CAMP-Net)	93.08	96.64	93.49	90.61	98.84	**98.46**	95.46	94.53	95.23	93.45
4 bands + SWIR (ViT)	96.24	97.67	93.25	95.90	98.69	94.82	95.76	95.81	96.05	94.98
4 bands + SWIR (CAMP-Net)	**99.92**	**99.61**	**98.21**	**99.15**	**99.46**	97.89	**99.35**	**99.00**	**99.08**	**98.81**

The results of different methods using different bands are shown in Figure 7, from which it can be seen that the CAMP-Net algorithm is better than the ViT algorithm for classification, while the four bands + SWIR is particularly good.

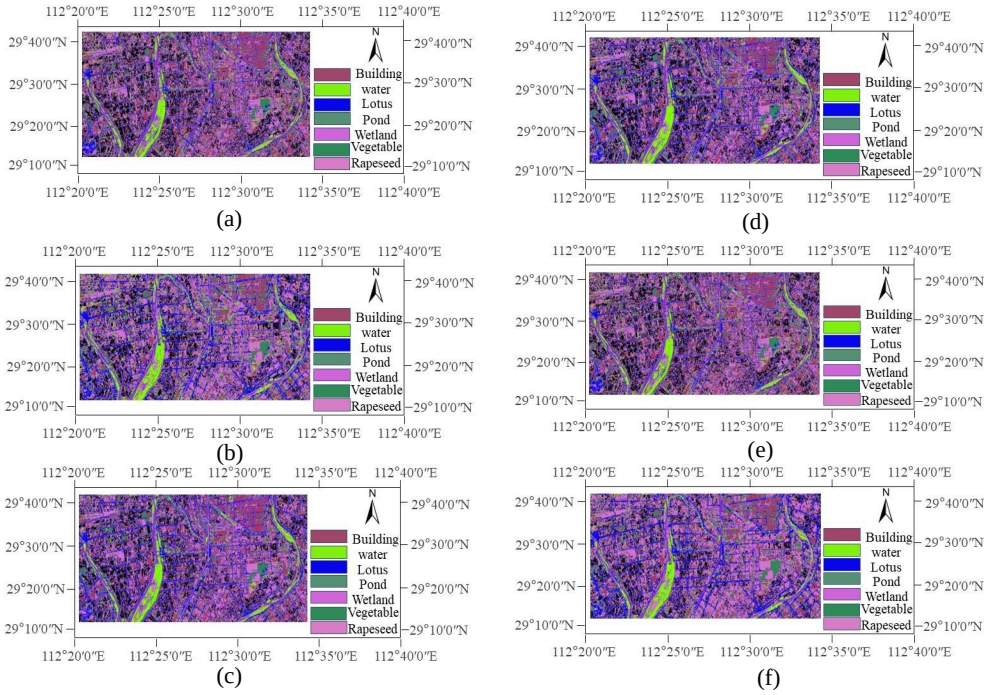

Figure 7. ViT results in different bands: (**a**) 4 bands. (**b**) 4 bands + VNIR. (**c**) 4 bands + SWIR. CAMP-Net results in different bands. (**d**) 4 bands. (**e**) 4 bands + VNIR. (**f**) 4 bands + SWIR.

3.2. Multi-Method Comparison

The quantitative classification results of OA, AA, and kappa for the study area datasets of South County and Xiangyin County and the accuracy of each category are shown in Tables 7 and 8. The overall effect of CNN is the worst, OA, AA, and kappa are lower than other models, and the accuracy for building and vegetable is only 70.22% and 72.65%, most likely It is likely because the multispectral images have fewer bands and only individual bands are used, which is not conducive to CNN learning features, while hyperspectral

images have 200 bands of information, and CNN outperforms SVM as well as KNN on hyperspectral images. Traditional classifiers SVM and KNN achieve relatively good results with OA of 89.41% and 92.19%, respectively, and SVM has a better performance on lotus, and KNN on building has poor classification ability. RNN, ViT, SF, and CAMP-Net are all spectral sequence classification methods based on deep learning, and the performance of the four models is relatively similar, proving the advantages of deep learning for sequence data processing. To better demonstrate the performance of each model, the Sentinel-2 multispectral images of the study area in 2018, 2020, and 2022 are also analyzed in Section 3.3. The classification performance of ViT and SF is well done, and the OA, AA, and Kappa of CAMP-Net is higher than the other comparable models, and the performance is even better in various categories such as building, water, and rapeseed.

The quantitative classification results of OA, AA, and kappa for the study area datasets of South County and Xiangyin County and the accuracy of each category are shown in Tables 7 and 8. The overall effect of CNN is the worst; OA, AA, and kappa are lower than other models, and the accuracy for building and vegetable is only 70.22% and 72.65%. This is likely because the multispectral images have fewer bands and only individual bands are used, which is not conducive to CNN learning features, while hyperspectral images have 200 bands of information, and CNN outperforms SVM as well as KNN on hyperspectral images. The traditional classifiers SVM and KNN achieve relatively good results with OA of 89.41% and 92.19%, respectively. SVM has a better performance on lotus, and KNN on building has poor classification ability. RNN, ViT, SF, and CAMP-Net are all spectral sequence classification methods based on deep learning, and the performance of the four models is relatively similar, proving the advantages of deep learning for sequence data processing. To better demonstrate the performance of each model, the Sentinel-2 multispectral images of the study area in 2018, 2020, and 2022 are also analyzed in Section 3.3. The classification performance of ViT and SF is well done, and the OA, AA, and kappa of CAMP-Net is higher than the other comparable models, and the performance is even better in various categories such as building, water, and rapeseed.

Table 7. Comparison of classification results of different classification methods on the South County 2021 data set.

C N.	Different Methods							
	SVM	**KNN**	**RF**	**CNN**	**RNN**	**ViT**	**SF**	**CAMP-Net**
1	87.89	82.45	89.47	70.22	97.44	96.24	99.39	**99.92**
2	89.18	96.99	96.09	87.48	97.03	97.67	99.48	**99.61**
3	85.53	90.40	91.23	77.80	86.45	93.25	95.10	**98.21**
4	89.16	92.87	92.12	78.20	94.86	95.90	97.27	**99.15**
5	91.23	97.31	94.99	78.91	97.54	98.69	98.31	**99.46**
6	92.84	93.64	92.84	72.65	95.28	94.82	97.10	**97.89**
7	90.00	92.71	91.42	85.16	95.70	95.76	97.60	**99.35**
OA (%)	89.41	92.19	92.29	78.07	94.57	95.81	97.49	**99.00**
AA (%)	89.41	92.34	92.60	78.64	94.91	96.05	97.75	**99.08**
Kappa	87.29	90.62	90.76	73.67	93.50	94.98	96.99	**98.81**

Table 8. Comparison of classification results of different classification methods on Xiangyin County 2021 data set.

C N.	Different Methods							
	SVM	**KNN**	**RF**	**CNN**	**RNN**	**ViT**	**SF**	**CAMP-Net**
1	95.18	90.12	95.52	81.13	94.60	97.24	96.70	**99.56**
2	97.69	99.62	99.56	95.31	99.27	98.74	99.33	**99.97**
3	96.26	96.77	97.45	86.17	98.03	99.63	99.34	**100.00**
4	84.66	89.93	90.33	77.39	92.94	93.18	95.58	**98.69**
5	76.11	91.87	88.88	49.18	**97.75**	91.32	94.10	97.58
6	94.33	98.02	96.31	78.05	96.38	98.35	97.62	**99.26**
7	97.34	99.55	98.00	90.36	99.47	99.19	99.71	**100.00**
OA (%)	92.47	95.13	95.52	82.26	96.27	96.92	97.61	**99.39**
AA (%)	91.66	95.13	95.16	79.66	96.07	96.81	97.49	**99.30**
Kappa	91.02	94.19	94.66	78.83	95.55	96.33	97.15	**99.27**

The classification map of South County obtained by different models is shown in Figure 8, and the study area in the square box is labeled according to the a priori knowledge and outdoor sampling data. CAMP-Net can better distinguish between wetland and water, which are less different within groups and does not misclassify vegetable, and it is obvious that CAMP-Net is clearer than the other algorithms.

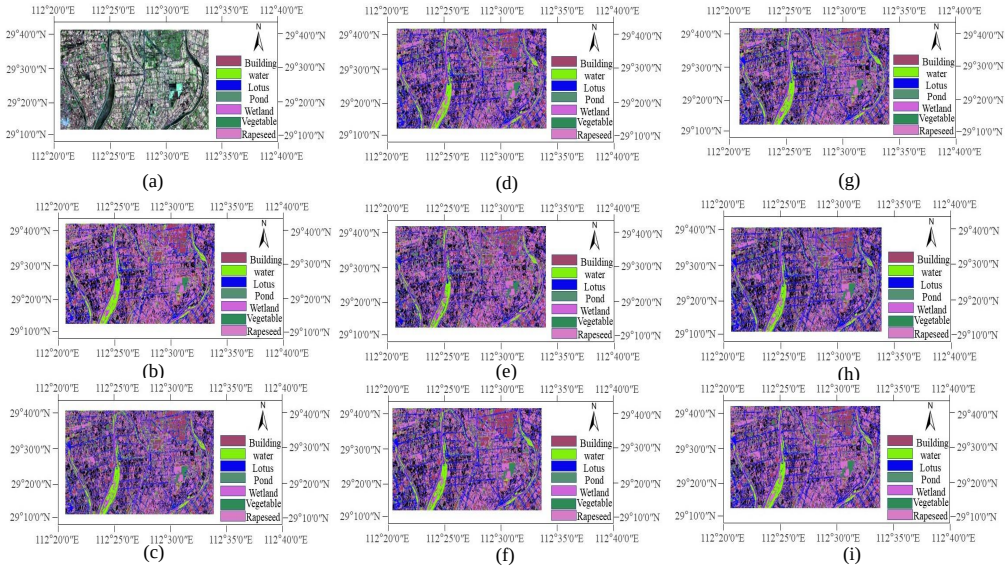

Figure 8. Classification results obtained by different models on South County data: (**a**) Image. (**b**) SVM. (**c**) KNN. (**d**) RF. (**e**) CNN. (**f**) RNN. (**g**) Transformer (ViT). (**h**) SF. (**i**) CAMP-Net.

As shown in Figure 9, the overall classification results of SVM and CNN are very poor; the remaining algorithms of CAMP-Net have clearer edge extraction and more obvious chunking for pond and vegetable, and the part between adjacent water regions is not misclassified as wetland. In general, RNN, SF, and ViT are significantly better for the classification of multispectral images than CNN and SVM. The proposed CAMP-Net, although the model structure is more complex than ViT, has some improvement in OA, AA, and kappa, and the map-forming effect is also better than ViT, which has great value.

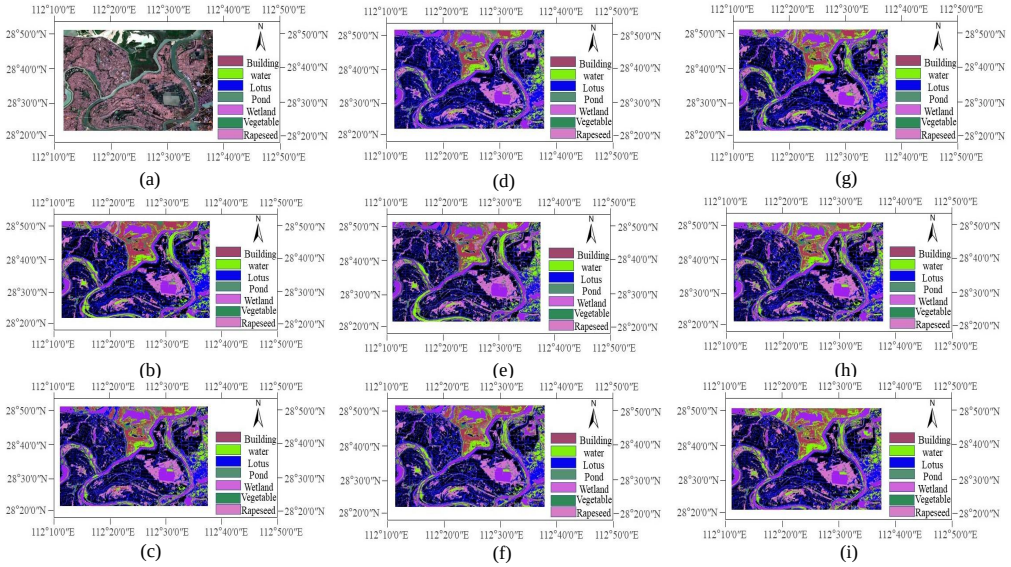

Figure 9. Classification results obtained by different models on Xiangyin County dataset: (**a**) Image. (**b**) SVM. (**c**) KNN. (**d**) RF. (**e**) CNN. (**f**) RNN. (**g**) Transformer (ViT). (**h**) SF. (**i**) CAMP-Net.

3.3. Land Use Change

This study used 17 July 2017 and 4 November 2019 Sentinel-2 series imagery downloaded from the USGS website (https://earthexplorer.usgs.gov/ (accessed on 1 March 2023)) for the study of feature classification. After pre-processing, the ROI was re-tagged using ENVI5.3 software based on the a priori knowledge accumulated from the 2021 outdoor collection data, and a .txt file containing the ROI coordinates was exported. Based on the coordinates of the .txt file, the sample set was randomly divided into a training set and a test set according to a 7:3 ratio. A variety of different models from Table 4 were also used in this paper for comparative analysis among the models. Among them, the sample pool data for 2017 and 2019 are detailed in Table 9. Finally, the CAMP-Net proposed in this paper was used to carry out a land use change analysis for the study area, focusing on the changes in rapeseed and vegetables.

Table 9. Sample sizes for the 2018, 2020, and 2022 datasets for South County.

No.	Class	2018		2020		2022	
		Training	Testing	Training	Testing	Training	Testing
1	Building	907	389	907	389	1187	509
2	Water	716	307	716	307	3466	1486
3	Lotus	611	263	611	263	1484	636
4	Pond	1200	515	1200	515	1451	622
5	Wetland	471	203	471	203	993	426
6	Vegetable	1136	488	1136	488	991	426
7	Rapeseed	783	336	783	336	1041	447
Total		5824	2501	5824	2501	10,613	4552
		8325		8325		15,165	

The classification results of different models for the sample pool data in 2018, 2019, and 2022 are shown in Tables 10–12. In the datasets of three different years, the OA of CAMP-Net is 98.25%, 97.06%, and 99.31%, which are higher than the other models. The overall classification is also better than the other models. Overall, 18, 20 and 22 years are better than the original ViT. There is an improvement of 1.70%, 4.27%, and 1.64% compared

to the original ViT. The lower level of improvement in 18 and 22 years may be due to the original VIT already being high, while the improvement in 20 years is more demonstrative compared to the 21 year-results we obtained, and the classification performance of CAMP-Net meets the requirements of conducting land use change on the study area.

Table 10. Comparison of classification results of different classification methods on the South County 2018 data set.

C N.	Different Methods							
	SVM	**KNN**	**RF**	**CNN**	**RNN**	**ViT**	**SF**	**CAMP-Net**
1	91.25	86.37	89.71	77.94	91.62	92.17	94.15	**97.68**
2	96.09	97.71	96.09	88.68	98.60	98.74	97.76	**99.44**
3	91.63	93.15	93.91	84.61	92.14	91.98	94.59	**95.09**
4	94.75	95.53	98.44	83.41	96.41	97.33	**98.33**	98.08
5	98.52	100.00	100.00	91.71	99.36	99.57	99.78	**100.00**
6	96.10	97.74	96.92	81.86	98.59	99.20	98.15	**99.47**
7	88.09	91.36	89.79	65.77	93.48	96.29	96.04	**97.70**
OA (%)	93.72	94.36	94.72	81.34	95.76	96.55	97.00	**98.25**
AA (%)	93.78	94.56	94.70	82.00	95.75	96.47	96.98	**98.21**
Kappa	92.56	93.32	93.74	77.87	94.98	95.91	96.44	**97.93**

Table 11. Comparison of classification results of different classification methods on the South County 2020 data set.

C N.	Different Methods							
	SVM	**KNN**	**RF**	**CNN**	**RNN**	**ViT**	**SF**	**CAMP-Net**
1	87.91	83.80	88.94	67.47	92.82	97.24	96.47	**99.00**
2	97.71	99.34	98.69	95.81	97.90	99.86	99.30	**100.00**
3	58.55	85.51	82.50	44.18	81.01	86.74	85.59	**94.43**
4	78.44	82.13	81.94	75.50	86.58	91.33	94.66	**97.25**
5	98.02	97.53	97.04	89.17	99.15	99.78	**100.00**	99.36
6	73.77	73.97	77.66	66.54	84.15	89.96	88.64	**95.42**
7	64.28	77.38	78.57	38.56	80.45	87.99	92.46	**94.89**
OA (%)	78.98	83.89	85.09	67.86	88.00	92.79	93.53	**97.06**
AA (%)	79.82	85.68	86.48	68.18	88.79	93.28	93.88	**97.20**
Kappa	75.01	80.94	82.34	61.57	85.78	91.46	92.33	**96.52**

Table 12. Comparison of the classification results of different classification methods on the South County 2022 data set.

C N.	Different Methods							
	SVM	**KNN**	**RF**	**CNN**	**RNN**	**ViT**	**SF**	**CAMP-Net**
1	94.49	93.51	95.67	77.50	96.96	98.73	99.49	**99.91**
2	**100.00**	99.93	**100.00**	99.82	99.65	99.79	**100.00**	99.82
3	93.55	93.39	95.28	88.20	93.73	95.95	98.45	**99.12**

Table 12. *Cont.*

C N.	Different Methods							
	SVM	**KNN**	**RF**	**CNN**	**RNN**	**ViT**	**SF**	**CAMP-Net**
4	95.17	95.17	95.33	85.25	97.31	96.07	98.82	**99.72**
5	92.01	94.36	94.60	90.43	98.99	98.99	**99.39**	98.69
6	70.18	89.90	90.84	41.97	86.78	90.11	86.17	**96.36**
7	99.32	99.32	99.32	99.32	99.80	100.00	100.00	**100.00**
OA (%)	94.22	96.13	96.79	87.38	96.96	97.67	98.22	**99.31**
AA (%)	92.11	95.09	95.87	83.22	96.18	97.10	97.48	**99.09**
Kappa	92.90	95.26	96.06	84.49	96.27	97.15	97.81	**99.16**

The spatial distribution of land use in 2018, 2020, and 2022 is shown in Figure 10 and the spatial distribution of each feature can be observed. The spatial distribution of each feature shows that vegetable is mainly distributed in the northwest and central part of the country, and rapeseed is more concentrated in the middle of the country. The pond is mainly located in the west part of the study area, surrounded by buildings, and water is mainly distributed in the southeast part. The spatial distribution of major land uses in the study area has become more balanced and rational in recent years.

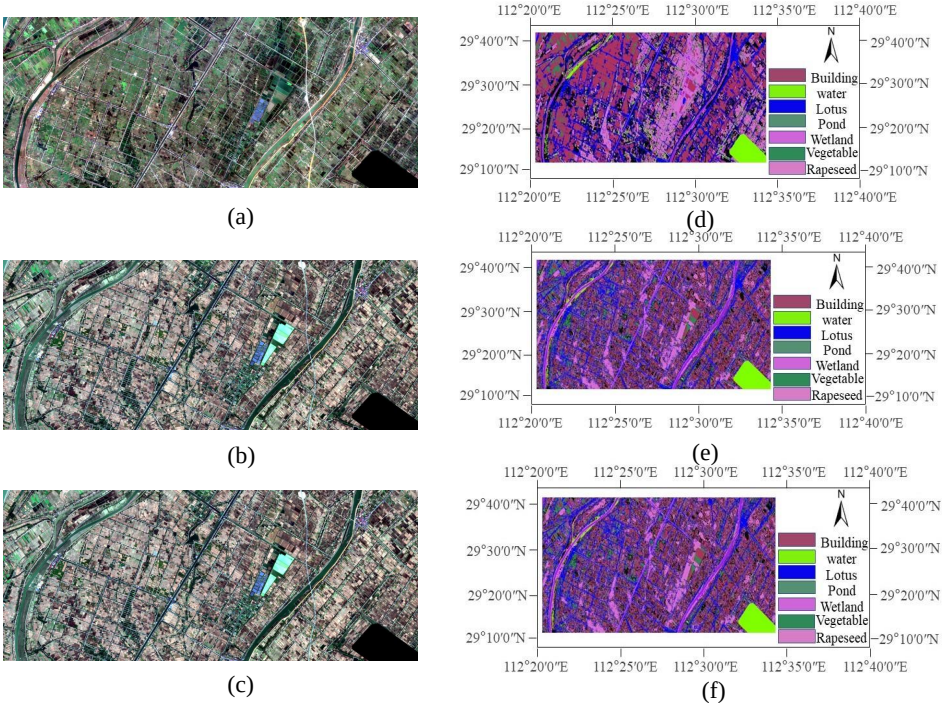

Figure 10. True color maps of the three images: (**a**) 2018, (**b**) 2020, and (**c**) 2022; spatial distribution of features in the three images: (**d**) 2018, (**e**) 2020, and (**f**) 2022.

The land use and land use change rate of the study area in the last 4 years are shown in Table 13. In the three images, the largest proportion of the coverage area in the study area is vegetable, and the largest change is lotus. Except for lotus, the change in other species in the last four years is relatively small, due to the increase in pond. Aquatic and crops are relatively increased while building and wetland are relatively decreased. The area is also increased, and the aquatic category is naturally valued by the people of Hunan as a specialty food and the natural geographical advantage of Hunan, so it has been

growing continuously. Various data during the 4 years indicate that the study area meets the sustainable development strategy.

Table 13. Table of land use type changes in the study area for different periods from 2018 to 2022.

Class	Area (km^2)			Area Change Rate (%)		
	2018	2020	2022	2018–2020	2020–2022	2018–2022
Building	26.79	25.86	26.15	−0.03	0.01	−0.02
Water	4.02	4.00	3.65	−0.01	−0.09	−0.09
Lotus	3.62	11.85	23.31	2.27	0.97	5.43
Pond	24.42	24.82	25.27	0.02	0.02	0.03
Wetland	7.57	5.49	6.93	−0.27	0.26	−0.08
Vegetable	41.83	37.94	48.43	−0.09	0.28	0.16
Rapeseed	20.69	18.99	19.72	−0.08	0.04	−0.05

Since rapeseed and vegetables are the main aquatic species and crops in the study area, this paper focuses on the spatial distribution of rapeseed and vegetables, whose spatial distribution map is shown in Figure 11. From the overall observation, the area where pond is located is mainly scattered near the river in the central region. The rich water resources in the area are conducive to the development of aquaculture as well as cultivation, and the good sales of seafood products have promoted the local people to vigorously develop aquaculture. Rapeseed and vegetable are scattered in various areas of the study area, and these two species play a very important role in making full use of natural and labor resources to increase food production, so it is reasonable for the government to carry out the full aspect of agricultural farming.

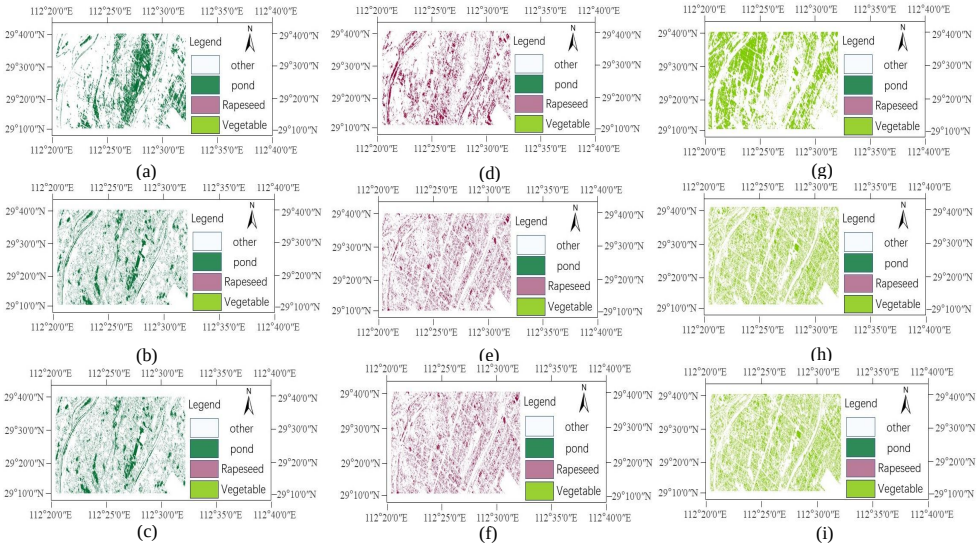

Figure 11. Spatial distribution of pond in three images: (**a**) 2018, (**b**) 2020, and (**c**) 2022; spatial distribution of rapeseed in three images (**d**) 2018, (**e**) 2020, and (**f**) 2022; spatial distribution of vegetable in three images (**g**) 2018, (**h**) 2020, and (**i**) 2022.

The dynamics of rapeseed and vegetables from 2018 to 2022 are shown in Figure 12. Rapeseed's farming area was 20.69 km^2 in 2018, 18.99 km^2 in 2020, and 19.72 km^2 in 2022. In 2020, 2020–2022, and 2018–2022, respectively, the area decreased by 8%, increased by 4%, and decreased by 5%. The dynamics of total rapeseed and vegetable from 2018 to 2022 are shown in Figure 12. Rapeseed's farming area was 20.69 km^2 in 2018, 18.99 km^2 in 2020, 19.72 km^2 in 2022, and 19.72 km^2 in 2018. In 2020, 2020-2022, and 2018–2022, respectively,

the area decreased by 8%, increased by 4%, and decreased by 5%. The is an overall decreasing trend from 2018 to 2022, but 2020–2022 is a rebound state, probably because rapeseed farmers can obtain more income. At the same time, the climate of the Hunan region is also more suitable for planting, leading farmers to increase planting rapeseed in recent years. The increase in vegetable planting area is larger, especially from 2020 to 2022, which is expected. Because of the fertile soil and water in the region, there is a natural advantage of growing more vegetables to obtain greater returns, and the distribution of the various species planted may be due to some coordination by the national government to achieve joint development in many aspects. Vegetable planting area in 2018 was 41.83 km^2, in 2020 it was 37.94 km^2, and 2020 it was 37.94 km^2. The increase in vegetable planting area, especially from 2020 to 2022, was expected because the area is fertile and has natural advantages for cultivation. There is a natural advantage of growing more vegetables for greater profitability, and the distribution of the various species planted may be due to some coordination by the national government to achieve a multifaceted co-development.

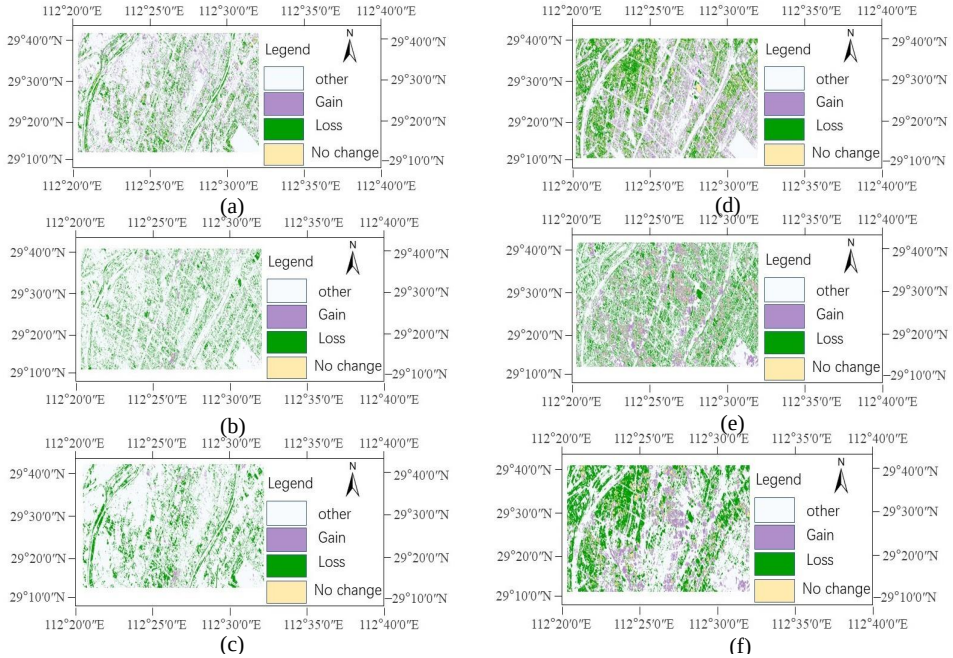

Figure 12. Dynamics of rapeseed in 2018–2020 (**a**), 2020–2022 (**b**), and 2018–2022 (**c**). Dynamics of vegetable in 2018–2020 (**d**), 2020–2022 (**e**), and 2018–2022 (**f**).

4. Discussion

In this work, we propose a new novel network for land cover classification, which is different and innovative from previous studies because of the following points:

(1) Our work adds NDVI and NDWI at the beginning neighboring bands and then uses Gse spectral feature module and selects every two sequences for band fusion in the transformer branch, while it only uses half of every two sequences for fusion in the experiment.

(2) Our paper uses two layers of attention stacked in the transformer branch and a channel MLP branch for feature extraction, while the other papers use a CNN branch and choose to cascade the two convolutions to form a DenseNet structure.

(3) Our paper merges the features obtained by a transformer encoder with the features obtained by channel MLP, while other papers use the features obtained by the transformer encoder with the features obtained by CNN to merge in order to produce the features in a parallel manner.

(4) Our paper merges the resulting features to reduce the dimensionality through the fully connected layer and finally classifies them through the activation function and

conv1*1 and the MLP head in the vision transformer (ViT), while it uses only the activation function and conv1*1 to obtain the classification results.

Because SWIR bands cover a wider spectral range, more optical information can be captured. Although the bandwidth of SWIR technology is narrower than that of RGB bands, it possesses a strong lateral extension capability, which enables it to cover weaker areas and areas that cannot be sensed by RGB bands, thus improving the accuracy of feature information. SWIR bands have a strong penetration capability, which allows them to penetrate clouds, smoke, and haze, and maintain a good performance even in the presence of thicker cloud layers. Since SWIR wavelengths are closely related to the structure and composition of substances, the unique spectral characteristics of certain substances or objects can be captured. Therefore, instead of adopting the RGB + NiR four bands, which is significantly applicable, the four bands + SWIR method was used to carry out the experiments, which helps to enhance the accuracy of the study and the credibility of the results.

Compared to traditional classification methods that cannot comprehensively consider the interrelationships between pixels in an image, the classification effect may be less than ideal in some complex application scenarios. In contrast, multispectral data provide a wider range of spectral information and take into account the interaction between various bands. This makes land cover mapping techniques based on multispectral data more accurate and reliable. Image classification based on multispectral data can identify land cover types by analyzing the correlation between different bands. In addition, multispectral data can provide information on many other surface characteristics, such as vegetation index, soil moisture content, etc., which are important for land management and environmental monitoring. This experiment compares the advantages and disadvantages of the traditional machine learning algorithms SVM, KNN, and RF and the deep learning algorithms CNN, RNN, transformer (ViT), SF, and CAMP-Net (a total of eight methods), with respect to the multispectral images, and the results show that our proposed algorithm, CAMP-Net, has a great advantage in the classification performance. Among the traditional methods, SVM, KNN, and RF have a high efficiency with a short training time, but their ability to capture band information is not as good as the RNN model, which is good at processing sequence information. Since the information of multispectral images is not as good as hyperspectral images, it means the effect of classification accuracy of CNN is not as good as expected, but unlike RNN, which is good at dealing with sequence data, the classification performance of RNN is better than that of CNN, but it is difficult for it to learn the long-term dependency of the input and output like transformer. This is because the transformer can capture global sequence information through positional encoding. The classification performance of SF on multispectral images is better than all models except CAMP-Net in all datasets.

Interestingly, CNN with the addition of SWIR band information also fails to capture the spatial information well. This may be due to the low dimensionality of the spatial kernel and the band information is just too little. It would be better if a dataset of hyperspectral images is used. Adding the attention mechanism or increasing the network depth can further improve the image-classification performance. The scaled 1D sequence data is transformed from a Cartesian coordinate system to a polar coordinate system, and the temporal correlation of different time points is recognized by considering the angle sum/difference between different points. GASF (corresponding to angular sum) or GADF (corresponding to angular difference) can be utilized to do the angles, and the implementation method depends on the specific requirements. Adding NDVI and NDWI band information for grouped spectral nesting, a multilayer perceptron network based on feature channels, and CAMP-Net with the transformer, which adds channel attention and performs best on five different datasets. The fusion of parallel networks can combine the features learned by transformer and MLP, respectively, to enhance the feature expression ability and thus obtain better classification results. Then, the land use dynamic change analysis can be

carried out to obtain a clearer understanding of the land cover classification changes in the study area in recent years.

5. Conclusions

In this paper, a new novel network for land cover classification is proposed, which employs a network architecture that incorporates a fusion channel attention mechanism in parallel with a multilayer perceptron based on feature channel utilization. In order to enhance the band local information interaction, grouped spectral nesting is used to process the features after adding NDVI and NDWI band information. Then, in the ViT branch, the information between spectral features is deeply mined by incorporating the channel attention mechanism; in the other branch, a multilayer perceptron network based on feature channel utilization is designed to improve the network's ability to extract features from different feature channels. The feature information obtained from both is fused, the dimensionality of the features is reduced by a fully connected layer and the classification results are obtained by activation function and MLP-head. The integrated classification accuracy of Nanxian in 2021 in the study area can reach 99.00%, which is a 3.19% improvement in OA compared to the original ViT. The classification accuracy of Xiangyin County in 2021 also reaches 99.39%, which is a 2.47% improvement compared to the ViT's 96.92. In the Nanxian dataset, in 18, 20, and 22 years the OA of CAMP-Net is, respectively, 98.25%, 97.06%, and 99.31%, which are 1.70%, 4.27%, and 1.64% improved compared to the original ViT. This huge improvement proves that pixel-level multispectral image classification is achievable. Finally, in future works, we will further improve the pixel data sequences in order to explore more information of the information features to be used. We believe that a useful way to continue this research in future works is to explore an enhanced network architecture that can utilize more information from the pixel data to improve the performance and allow for wider applications to multiple fields and domains to improve efficiency.

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

Funding: This work is the result of the research project funded by the National Natural Science Foundation of China (62261004 and 62001129) and the Innovation Project of Guangxi University of Science and Technology Graduate Education (GKYC202321).

Data Availability Statement: The data are available at https://github.com/lixuyaaaaa/CAMP-Net accessed on 1 September 2023.

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

Abbreviations

The following abbreviations are used in this manuscript:

CAMP-Net	Combines the channel attention mechanism and the multi-layer perceptron parallel algorithm.
ViT	Vision transformer
RS	Remote sensing
PolSAR	Polarized synthetic aperture radar
CNN	Convolutional neural network
SF	SpectralFormal
SVM	Support vector machine
RF	Random forest
CN.	Class number
GSE	Grouped spectral embedding
K-Means	K-means clustering

SWIR	Short wave infrared radiometer
ROI	Region of interest
NDVI	Normalized difference vegetation index
NDWI	Normalized difference water index
MLP	Multi-Layer Perceptron
OA	Overall Accuracy
AA	Average Accuracy
UAV	Unmanned aerial vehicle

References

1. Huang, C.; Davis, L.; Townshend, J. An assessment of support vector machines for land cover classification. *Int. J. Remote Sens.* **2002**, *23*, 725–749.
2. Duro, D.C.; Franklin, S.E.; Dubé, M.G. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery. *Remote Sens. Environ.* **2012**, *118*, 259–272. [CrossRef]
3. Wang, J.; Wang, S.; Wang, F.; Zhou, Y.; Wang, Z.; Ji, J.; Xiong, Y.; Zhao, Q. FWENet: A deep convolutional neural network for flood water body extraction based on SAR images. *Int. J. Digit. Earth* **2022**, *15*, 345–361. [CrossRef]
4. Rey, S.J.; Anselin, L.; Li, X.; Pahle, R.; Laura, J.; Li, W.; Koschinsky, J. Open geospatial analytics with PySAL. *ISPRS Int. J. Geo. Inf.* **2015**, *4*, 815–836. [CrossRef]
5. Jones, H.G. Use of infrared thermometry for estimation of stomatal conductance as a possible aid to irrigation scheduling. *Agric. For. Meteorol.* **1999**, *95*, 139–149. [CrossRef]
6. Schofield, R.; Thomas, D.S.; Kirkby, M.J. Causal processes of soil salinization in Tunisia, Spain and Hungary. *Land Degrad. Dev.* **2001**, *12*, 163–181. [CrossRef]
7. Peng, J.; Pan, Y.; Liu, Y.; Zhao, H.; Wang, Y. Linking ecological degradation risk to identify ecological security patterns in a rapidly urbanizing landscape. *Habitat Int.* **2018**, *71*, 110–124. [CrossRef]
8. Long, H.; Li, Y.; Liu, Y.; Woods, M.; Zou, J. Accelerated restructuring in rural China fueled by 'increasing vs. decreasing balance' land-use policy for dealing with hollowed villages. *Land Use Policy* **2012**, *29*, 11–22. [CrossRef]
9. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981.
10. Planinsic, P.; Singh, J.; Gleich, D. SAR image categorization using parametric and nonparametric approaches within a dual tree CWT. *IEEE Geosci. Remote Sens. Lett.* **2014**, *11*, 1757–1761. [CrossRef]
11. Papadomanolaki, M.; Vakalopoulou, M.; Karantzalos, K. A novel object-based deep learning framework for semantic segmentation of very high-resolution remote sensing data: Comparison with convolutional and fully convolutional networks. *Remote Sens.* **2019**, *11*, 684. [CrossRef]
12. Guo, X.; Chen, Z.; Wang, C. Fully convolutional DenseNet with adversarial training for semantic segmentation of high-resolution remote sensing images. *J. Appl. Remote Sens.* **2021**, *15*, 016520. [CrossRef]
13. Poursanidis, D.; Chrysoulakis, N.; Mitraka, Z. Landsat 8 vs. Landsat 5: A comparison based on urban and peri-urban land cover mapping. *Int. J. Appl. Earth Obs. Geoinf.* **2015**, *35*, 259–269. [CrossRef]
14. Melgani, F.; Bruzzone, L. Classification of hyperspectral remote sensing images with support vector machines. *IEEE Trans. Geosci. Remote Sens.* **2004**, *42*, 1778–1790. [CrossRef]
15. Ayerdi, B.; Romay, M.G. Hyperspectral image analysis by spectral–spatial processing and anticipative hybrid extreme rotation forest classification. *IEEE Trans. Geosci. Remote Sens.* **2015**, *54*, 2627–2639. [CrossRef]
16. Lin, T.H.; Li, H.T.; Tsai, K.C. Implementing the Fisher's Discriminant Ratio in ak-Means Clustering Algorithm for Feature Selection and Data Set Trimming. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 76–87. [CrossRef]
17. Alimjan, G.; Sun, T.; Liang, Y.; Jumahun, H.; Guan, Y. A new technique for remote sensing image classification based on combinatorial algorithm of SVM and KNN. *Int. J. Pattern Recognit. Artif. Intell.* **2018**, *32*, 1859012. [CrossRef]
18. Schmidhuber, J. Deep learning in neural networks: An overview. *Neural Netw.* **2015**, *61*, 85–117. [CrossRef]
19. Lu, H.; Zhang, M.; Xu, X.; Li, Y.; Shen, H.T. Deep fuzzy hashing network for efficient image retrieval. *IEEE Trans. Fuzzy Syst.* **2020**, *29*, 166–176. [CrossRef]
20. Wang, J.; Wang, D.; Wang, S.; Li, W.; Song, K. *Fault Diagnosis of Bearings Based on Multi-Sensor Information Fusion and 2D Convolutional Neural Network*; IEEE: New York, NY, USA, 2021; Volume 9, pp. 23717–23725.
21. Qiu, S.; Zhao, H.; Jiang, N.; Wang, Z.; Liu, L.; An, Y.; Zhao, H.; Miao, X.; Liu, R.; Fortino, G. *Multi-Sensor Information Fusion Based on Machine Learning for Real Applications in Human Activity Recognition: State-of-the-Art and Research Challenges*; Elsevier: Amsterdam, The Netherlands, 2022; Volume 80, pp. 241–265.
22. Zhang, W.; Tang, P.; Zhao, L.; Huang, Q. A comparative study of U-nets with various convolution components for building extraction. In Proceedings of the 2019 Joint Urban Remote Sensing Event (JURSE), Vannes, France, 22–24 May 2019; IEEE: New York, NY, USA, pp. 1–4.
23. Deng, Z.; Zhu, X.; He, Q.; Tang, L. Land use/land cover classification using time series Landsat 8 images in a heavily urbanized area. *Adv. Space Res.* **2019**, *63*, 2144–2154. [CrossRef]

24. Wu, C.; Jia, W.; Yang, J.; Zhang, T.; Dai, A.; Zhou, H. Economic Fruit Forest Classification Based on Improved U-Net Model in UAV Multispectral Imagery. *Remote Sens.* **2023**, *15*, 2500. [CrossRef]
25. Schiefer, F.; Kattenborn, T.; Frick, A.; Frey, J.; Schall, P.; Koch, B.; Schmidtlein, S. Mapping forest tree species in high resolution UAV-based RGB-imagery by means of convolutional neural networks. *ISPRS J. Photogramm. Remote Sens.* **2020**, *170*, 205–215. [CrossRef]
26. Li, G.; Han, W.; Huang, S.; Ma, W.; Ma, Q.; Cui, X. Extraction of sunflower lodging information based on UAV multi-spectral remote sensing and deep learning. *Remote Sens.* **2021**, *13*, 2721. [CrossRef]
27. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*.
28. 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 16 × 16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
29. 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**, *60*, 1–15. [CrossRef]
30. Fan, X.; Li, X.; Yan, C.; Fan, J.; Yu, L.; Wang, N.; Chen, L. MARC-Net: Terrain Classification in Parallel Network Architectures Containing Multiple Attention Mechanisms and Multi-Scale Residual Cascades. *Forests* **2023**, *14*, 1060. [CrossRef]
31. Du, J.; Zhou, H.; Jacinthe, P.A.; Song, K. Retrieval of lake water surface albedo from Sentinel-2 remote sensing imagery. *J. Hydrol.* **2023**, *617*, 128904. [CrossRef]
32. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
33. Reddi, S.J.; Kale, S.; Kumar, S. On the convergence of adam and beyond. *arXiv* **2019**, arXiv:1904.09237.