remote sensing

MDPI

Article

An Adaptive Feature Fusion Network with Superpixel Optimization for Crop Classification Using Sentinel-2 Imagery

Xiangyu Tian [1,2], Yongqing Bai [1], Guoqing Li [3], Xuan Yang [4,*], Jianxi Huang [5] and Zhengchao Chen [1]

[1] Airborne Remote Sensing Center, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[2] University of Chinese Academy of Sciences, Beijing 100049, China
[3] Henan Institute of Remote Sensing and Geomatics, Zhengzhou 450003, China
[4] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[5] College of Land Science and Technology, China Agricultural University, Beijing 100083, China
* Correspondence: yangxuan@radi.ac.cn

Abstract: Crop-type mapping is the foundation of grain security and digital agricultural management. Accuracy, efficiency and large-scale scene consistency are required to perform crop classification from remote sensing images. Many current remote-sensing crop extraction methods based on deep learning cannot account for adaptation effects in large-scale, complex scenes. Therefore, this study proposes a novel adaptive feature-fusion network for crop classification using single-temporal Sentinel-2 images. The selective patch module implemented in the network can adaptively integrate the features of different patch sizes to assess complex scenes better. TabNet was used simultaneously to extract spectral information from the center pixels of the patches. Multitask learning was used to supervise the extraction process to improve the weight of the spectral characteristics while mitigating the negative impact of a small sample size. In the network, superpixel optimization was applied to post-process the classification results to improve the crop edges. By conducting the crop classification of peanut, rice, and corn based on Sentinel-2 images in 2022 in Henan Province, China, the novel method proposed in this paper was more accurate, indicated by an F1 score of 96.53%, than other mainstream methods. This indicates our model's potential for application in crop classification in large scenes.

Keywords: crop mapping; deep learning; feature fusion; multitask learning; Sentinel-2

check for updates

Citation: Tian, X.; Bai, Y.; Li, G.; Yang, X.; Huang, J.; Chen, Z. An Adaptive Feature Fusion Network with Superpixel Optimization for Crop Classification Using Sentinel-2 Imagery. *Remote Sens.* **2023**, *15*, 1990. https://doi.org/10.3390/rs15081990

Academic Editor: Dino Ienco

Received: 12 February 2023
Revised: 29 March 2023
Accepted: 6 April 2023
Published: 10 April 2023

1. Introduction

Despite being a basic guarantee of human life, grain security is at risk owing to global population growth and accelerated climate change in recent years [1]. Crop type information is fundamental for crop yield estimation [2], crop pest monitoring [3], and growth monitoring [4], which are critical for maintaining grain security. However, the extraction and acquisition of crop type information are difficult, largely because the manual methods required are costly in terms of labor and resources [5], often resulting in small sample sizes.

Remote sensing has been widely used in extracting crop-type information due to its wide monitoring range, low cost, and high timeliness [6–8]. Vegetation indices can be formed by combining visible and near-infrared bands of images, which can measure the condition of surface vegetation simply and effectively [9–11]. Therefore, most mature application methods use a vegetation index for crop classification [12]. Supported by time series of Moderate Resolution Imaging Spectroradiometer (MODIS) data, Zhang et al. [13] used the fast Fourier transform to smooth normalized difference vegetation index (NDVI) time-series curves while related parameters such as the curve's mean, phase, and amplitude were used to extract the spatial distribution data of crops in North

China. Xiao et al. [14] used the NDVI, the enhanced vegetation index (EVI), and the land surface water index (LSWI) calculated from MODIS multi-temporal images to classify rice, water, and evergreen plants. With the recent increase in the use of algorithm iterations and remote sensing data, a variety of machine learning methods such as support vector machine (SVM) [15], k-means algorithm [16], maximum likelihood method [17], extreme gradient boosting (XGBoost) [18], and random forest (RF) [12] have been successfully applied to remote sensing crop classification [19]. For example, using the RF method, Markus et al. [20] used single-temporal Sentinel-2 data to classify seven crops in Austria. They achieved an overall accuracy of 76%, establishing the foundation for applying Sentinel-2 data in crop extraction. Waldner et al. [21] used time-series images from Landsat 8 and SPOT-4 to classify wheat, corn, and sunflower using artificial neural networks, achieving the highest classification accuracy of 85%. Furthermore, Wen et al. [5] used time-series Landsat data and limited high-quality samples to achieve large-scale corn classification mapping using the RF method. As one of the more effective methods for crop classification, dynamic time warping (DTW) is widely used in crop classification [22]. Belgiu et al. [23] evaluated how a time-weighted dynamic time warping (TWDTW) method that uses Sentinel-2 time series perform when applied to pixel-based and object-based classifications of various crop types in three different study areas. However, the above classification methods require the manual design of crop features, which relies on expert knowledge and fixed scenes, resulting in the insufficient generalization ability of these methods [24]. Therefore, these methods are not dependable for large-scale intelligent crop classification with multiple scenes.

After ten years of development, the concept of deep learning was proposed in 2006, and it has since made breakthroughs in many remote sensing applications, such as land cover classification [25], building change detection [26], and complex ground object detection [27]. In the field of crop classification, deep learning has become a mainstream method with large-scale applications [28]. Compared with traditional crop classification methods, deep learning methods can automatically mine deep features from remote sensing data; can make full use of time, spatial, and spectral information in images; and have better anti-noise and generalization abilities, making them the mainstream method of large-scale crop classification [29–31].

Recurrent neural network (RNN) methods, such as long short-term memory (LSTM) [32], gated recurrent units (GRU) [33], and Conv1D-CNN [34], originated from natural language processing and are particularly dependable for sequential data. Thus, they have been applied to crop extraction methods based on time-series images. For example, Zhong et al. [35] used Landsat time series data and land use survey results from Yolo County, California, to test the classification accuracy of LSTM and Conv1D-CNN methods in a variety of summer crops. However, obtaining complete time-series data is difficult, often resulting in cloud occlusions or missing data. To address these limitations, some data-filling methods exist. For example, Zhao et al. [36] classified crops from missing Sentinel-2 time series data using the filled missing data method and GRU. However, these processes requiring multiple data sources are complex and undependable for large-scale crop extraction. Furthermore, to introduce the spatial information of an image, such as texture and planting structure, to assist crop classification [37], most studies have used a specific patch size around the point as the model input. For example, Xie et al. [38] decomposed images into patches of different sizes as inputs in a convolutional neural network (CNN) for crop classification to compare the effect of patch size on crop classification accuracy. Additionally, Seyd et al. [39] used 11 × 11 pixels patches for classification and designed a two-stream attention CNN network to extract the spatial and spectral information of crops simultaneously. These methods generally use fixed-size patches as input, but plot sizes can vary greatly in actual large-scale scenes, lowering the model's accuracy. Furthermore, crop samples are generally obtained manually from the field, which can result in a small number of samples depending on staffing and available resources. A standard solution to this issue is to use a deep network for feature extraction followed by a machine learning method as the classifier, as demonstrated by Yang et al. [40], who used a combination of CNN and RF for crop classification. However, there are significant differences in training and principles between deep neural networks (DNNs) and machine learning methods. The direct combination of these methods cannot fully exploit the advantages of deep learning. Limited by the spatial and spectral resolutions of satellite images, the spatial resolution of satellite images generally used for crop classification is low. For example, the highest resolution of Sentinel-2 is 10 m. To improve the boundary accuracy of crop extraction results, Kussul et al. [41] used plot vector data to optimize crop extraction results and proposed a new voting optimization method that could significantly improve the accuracy of the results. However, obtaining high-precision plot vector data and applying them to large-scale crop extraction is challenging. Using the images themselves for optimization may be a feasible method.

Above all, large-scale crop extraction still has the following problems:

(1) Due to inevitable conditions such as cloud cover and missing data, it is difficult to obtain complete time-series images, especially for large-scale crop classification. Moreover, there are few studies on crop extraction from single-temporal remote sensing images [42,43].

(2) Using patches as model inputs may introduce noise while introducing spatial information. As shown in Figure 1, in mixed planting or terrace scenarios, the crops are small, long, and narrow in the area. Therefore, the number of pixels different from the central pixels category in a large patch exceeds half of the total redundant information may lead to classification errors.

(3) Insufficient crop samples are usually obtained, and since deep learning methods with more parameters are prone to overfitting, there is a risk of the model having insufficient generalization ability and low accuracy in large-scale classification.

(4) Limited by the spatial resolution of multispectral images, the phenomenon of mixed pixels is consequential, resulting in inaccurate crop boundaries.

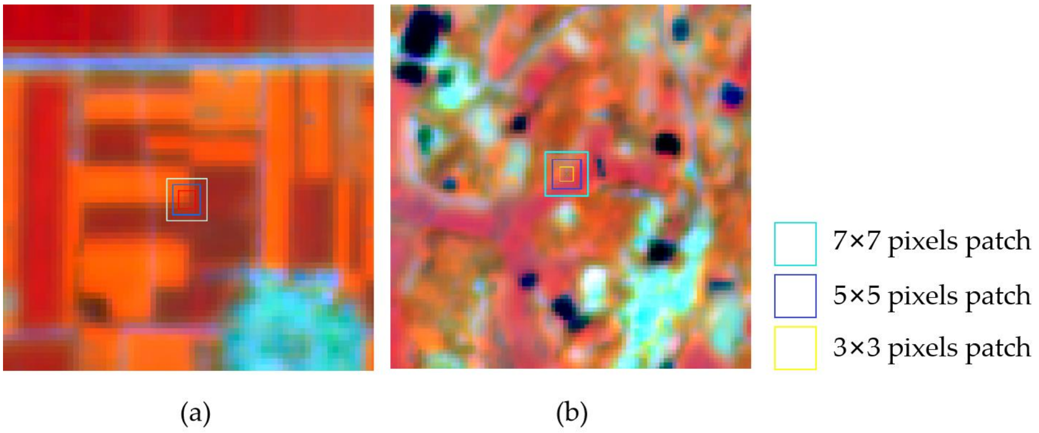


Figure 1. Coverage of different patch sizes in complex scenarios. (**a**) Mixed planting scenario, where crops grow in a random pattern. There are multiple crops in a small area. (**b**) Mountain terrace scenario, where crops grow in valleys with smaller areas.

In response to these problems, we propose a deep neural network for large-scale crop classification using single-temporal images named selective patches TabNet (SPTNet). The contributions of this study are as follows.

(1) A selective patch module that can adaptively fuse the features of patches of different sizes is designed to improve the network's ability to extract small crop plots in complex scenes.

(2) TabNet [44] and multitask learning were introduced to capture the spectral and spatial information of the central pixel to improve the weight of the central pixel during the classification process and enhance the network's generalization ability, which effectively reduced the negative impact of insufficient sample numbers.

(3) Superpixel segmentation was implemented in the post-processing of classification results to increase the boundaries of crop plots.

(4) High classification accuracy was achieved using the above modules and insufficient crop phenology information. A large-scale crop mapping of three major crops in 2022 in Henan Province, China, was produced to meet the government's demands on crop yield estimation and agricultural insurance.

2. Materials and Methods

2.1. Study Area

Crop classification and extraction experiments were carried out in Henan Province, China. As shown in Figure 2, Henan Province is located in central China, between 32°23′N–36°22′N and 110°21′E–116°38′E. With a cultivated land area of 81,500 km^2. Accounting for 6.05% of the total cultivated land area in China, Henan is an essential breadbasket in the country and has continuously produced the most grain output per province in China for many years. Located in the continental monsoon climate region of the transition from the northern subtropical to the warm temperate zone, Henan Province has two harvest seasons each year, in summer and autumn. Summer crops include mostly wheat, while autumn crops mainly include corn, rice, and peanut. This study focused on the classification and extraction of autumn crops in Henan Province in 2022.

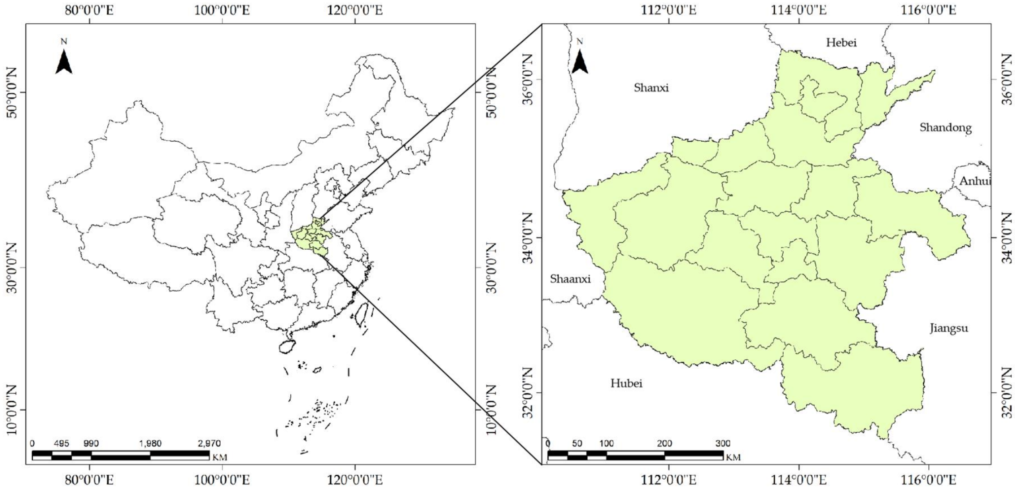


Figure 2. Spatial distribution map of the study area. The left figure shows the location of the study area in China, and the right figure shows the zoomed-in details of the study area.

2.2. Data and Processing

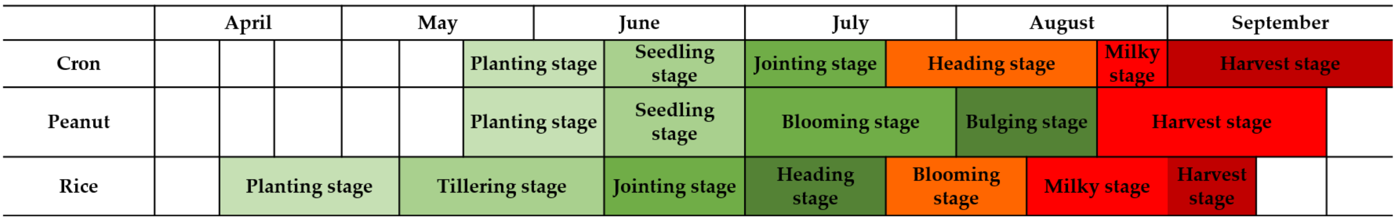
2.2.1. Remote Sensing Data

The remote sensing data used in this study were the bottom-of-atmosphere corrected reflectance products (L2A) of Sentinel-2 images, which can be downloaded from the European Space Agency (ESA) (https://scihub.copernicus.eu/, accessed on 16 March 2023). As shown in Table 1, there are 12 bands with resolutions of 10–60 m, in which the 4 vegetation red edge bands and shortwave infrared bands are sensitive to plant characteristics and are thus more dependable for crop classification [45]. Because the resolution of Sentinel-2 images varies between bands, the bands of 20 m and 60 m were pan-sharpened to 10 m, and we reprojected all images to the WGS-84 coordinate system. Considering image overlap and cloud occlusion, we processed the mosaic and standard map divisions after the cloud mask for all images.

The entire Henan Province was completely covered by 38 Sentinel-2 images. The phenology of corn, peanut, and rice is shown in Figure 3. Considering both crop phenology and image cloud cover, we selected all 38 images with the minor cloud cover between 31 July and 15 August 2022 for the experiments.

Table 1. Bands Information of Sentinel-2 L2A Images.

Sentinel-2 Bands	Central Wavelength (nm)	Resolution (m)	Description
Band 1	443.9	60	Aerosols
Band 2	496.6	10	Blue
Band 3	560.0	10	Green
Band 4	664.5	10	Red
Band 5	703.9	20	Red Edge 1
Band 6	740.2	20	Red Edge 2
Band 7	782.5	20	Red Edge 3
Band 8	835.1	10	NIR
Band 8A	864.8	20	Red Edge 4
Band 9	945.0	60	Water Vapor
Band 11	1613.7	20	SWIR 1
Band 12	2202.4	20	SWIR 2



	April	May	June	July	August	September
Cron		Planting stage	Seedling stage	Jointing stage Heading stage	Milky stage	Harvest stage
Peanut		Planting stage	Seedling stage	Blooming stage	Bulging stage Harvest stage	
Rice	Planting stage	Tillering stage	Jointing stage	Heading stage Blooming stage	Milky stage	Harvest stage

Figure 3. Phenology of corn, peanut, and rice.

2.2.2. Reference Samples

Figure 4 shows the distribution of the sample points collected through multiple field surveys across the study area. All field surveys were completed from August to September 2022. Over 4000 sample points were recorded in the field using a handheld global positioning system device (GARMIN Etrex221x; the positioning error is less than 3 m), mainly covering three crop types: corn, peanut, and rice. Additionally, a small number of other crops, such as soybean and pepper, were also recorded simultaneously and placed in a fourth category, titled *others*. In the field surveys, we abided by the following sampling rules to ensure the representativeness of the samples:

(1) Record the coordinates of sample points when the positioning signal is strong;

(2) Select the sampling position in the center of the crop plot, away from forests, green belts, rivers, and lakes, which easily affect the characteristics of the crops;

(3) Collect samples as evenly as possible in the experimental area to ensure samples are in various planting scenarios;

(4) Concentrate on collecting samples in areas with complex planting structures to increase the number of samples.

We then manually interpreted the images around the field sampling points to enrich the number of samples and supplement other background categories, such as residential areas, water, woodlands, and bare land, which were not covered in the field survey. As shown in Figure 5, all pixels in the extended area were used as samples. We then randomly divide the samples into training, validation, and test sets at a ratio of 10:1:9 to carry out the crop classification experiments. Finally, we obtained more than 300,000 sample points of the crop and non-crop types in Henan Province, of which peanut samples were the most prominent and corn samples were the least (Table 2).

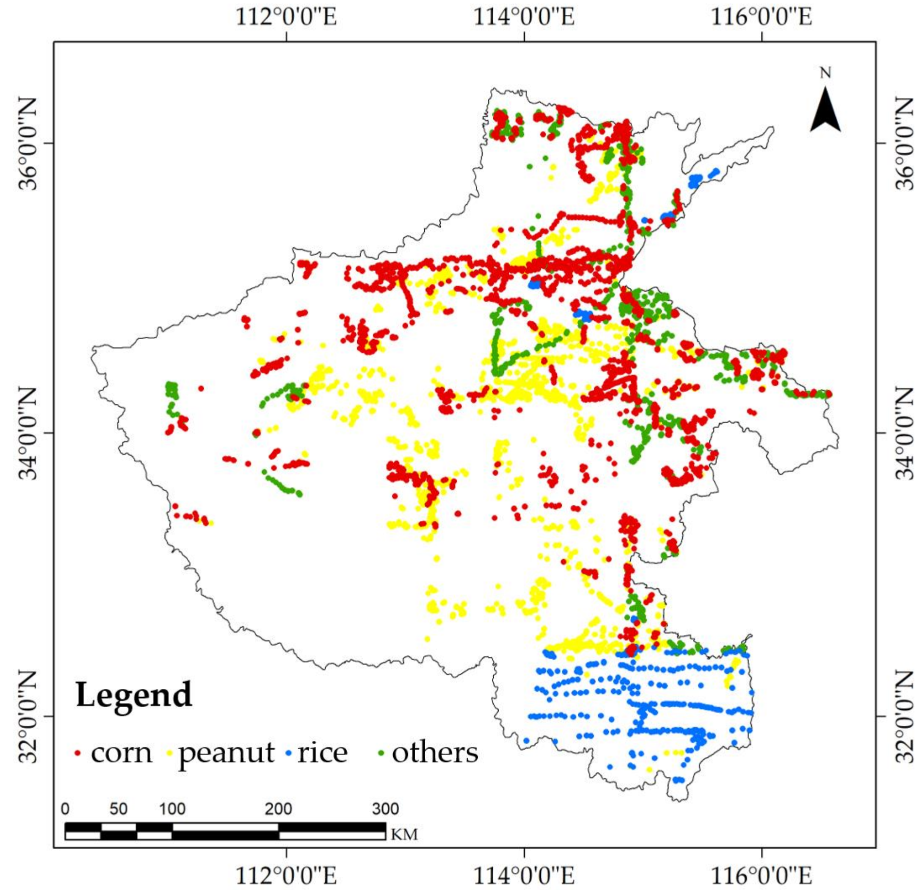


Figure 4. Distribution of sample points in the study area.

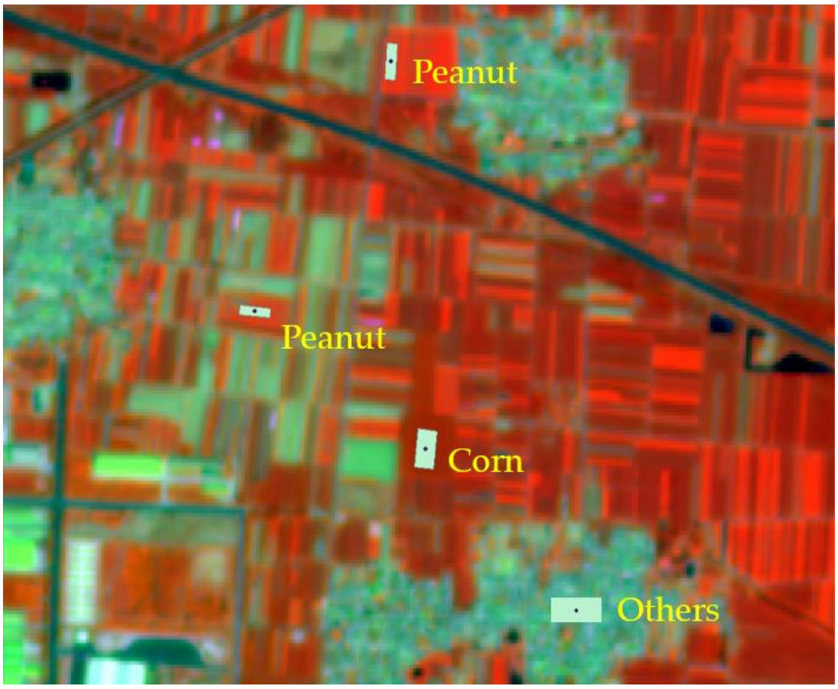


Figure 5. Example of sample expansion. The point vectors are obtained through ground surveys, and the surface vectors are obtained through manual interpretation based on the point vectors. The expanded crop samples are assumed to be manually interpreted inside crop plots.

Table 2. The number of expended reference samples that were divided into training, validation, and test samples at a ratio of 10:1:9.

Types	Number of Field Samples	Number of Expanded Samples	Training (50%)	Validation (5%)	Test (45%)
Peanut	1501	112,134	56,067	5607	50,460
Corn	1528	69,185	34,593	3459	31,133
Rice	1045	76,966	38,483	3849	34,635
Others	319	128,659	64,329	6433	57,897
Total	4392	386,945	193,471	19,349	174,125

2.3. Methods

The general workflow of crop type classification based on the proposed method is illustrated in Figure 6. The proposed classification framework was implemented in three main steps: (1) data and sample preparation, (2) model training and accuracy evaluation, and (3) prediction and crop mapping. The details of each step are discussed in the following subsections.

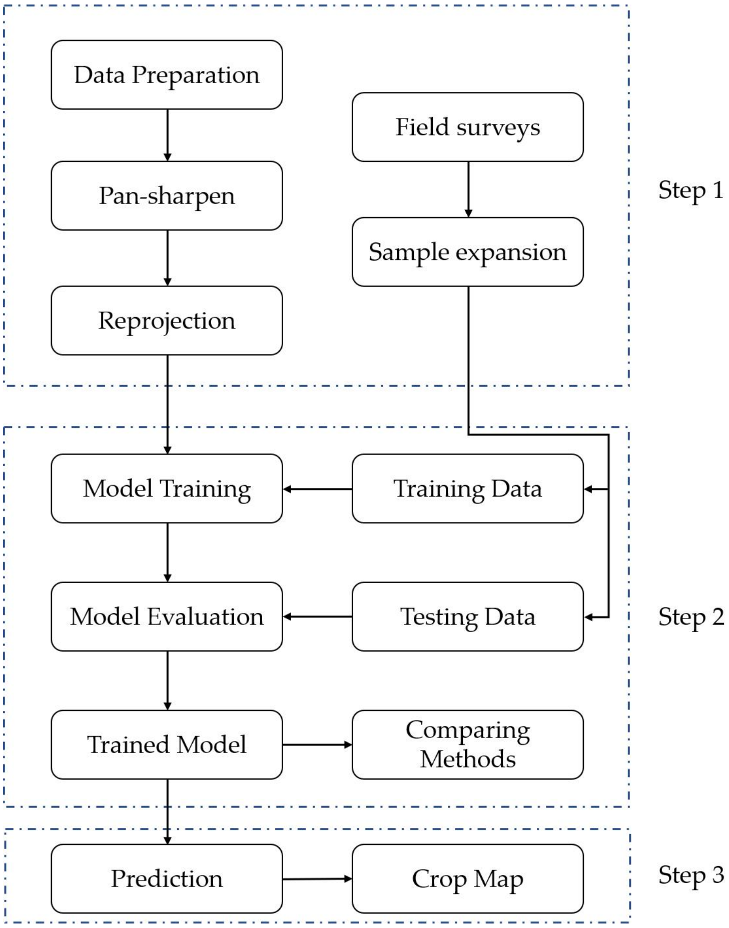


Figure 6. Overview of the proposed framework for crop mapping.

We proposed a new DNN architecture named SPTNet, depicted in Figure 7, that is characterized by a selective patch module (SPM) that adaptively acquires multi-size patch features, a TabNet branch that models the spectral information of the center point separately, and multiple loss functions. Through the SPM, the input patches are fused to select the appropriate patch size adaptively. This process balances the relationship between the obtained spatial information and noise. To increase the weight of the spectral information of the central pixels, we also used multitask learning to model the spectral information of the pixel and introduce the corresponding loss to supervise the process directly. Finally, we introduced a superpixel segmentation method for post-processing to improve the boundaries of crop plots and reduce the negative impact of mixed pixels on the model's function.

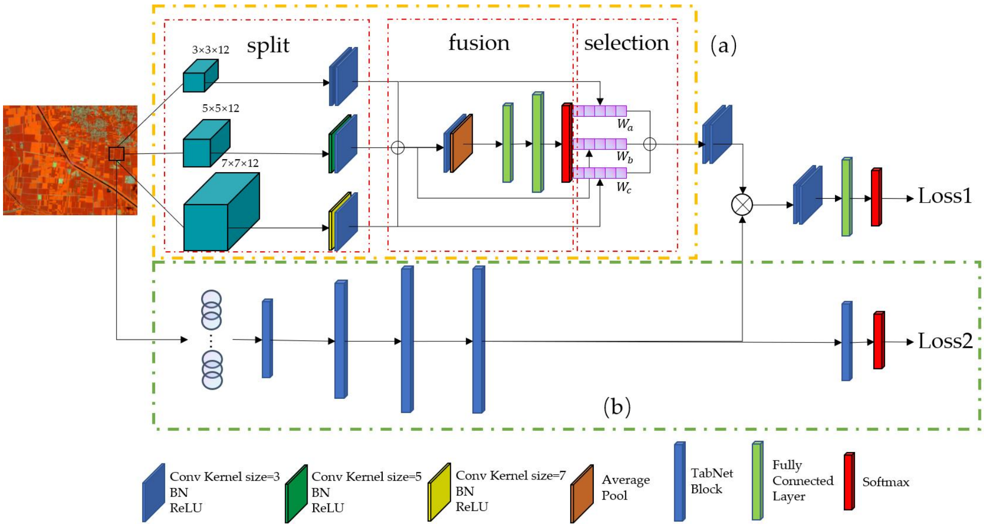


Figure 7. The proposed SPTNet's framework for crop mapping using a novel DNN architecture. (**a**) The selective patch module. (**b**) The branch of central pixel spectral feature extraction.

2.3.1. Selective Patch Module

Because crop plots vary in size in different scenarios, using oversized patches introduces too many pixels that are not in the same category as the central pixel, which may lead to some pixels being classified as noise. Conversely, using a too-small patch will lead to insufficient available spatial information, making the classification inaccurate. The above problem is similar to the receptive field problem in semantic segmentation, which requires selection according to the target size. The difference lies in selecting an appropriate receptive field or patch size. SK-Conv [46] automatically selects the convolutional receptive field. Several studies have used SK-Conv, which calculates the weight of feature maps obtained by convolution kernels of different sizes and fuses the feature maps according to the weights to realize the adaptive selection of the convolution kernel receptive field [47,48].

Inspired by the SK-Conv method, we constructed an SPM that adaptively fuses features of different patch sizes to solve scaling problems and obtain more accurate spatial information. Figure 7a shows that SPM consists of three stages: split, fusion, and selection. The division into three stages optimizes the ability to generate different patch sizes, aggregate different patch information to obtain a global representation of the selection weight, and obtain feature maps of different patch sizes based on the selection weight. First, in the split phase, the input patch is divided into new patches $\hat{P}$, $\bar{P}$, and $\tilde{P}$ according to size. In the study of crop classification in Henan Province, we divided the patches into the sizes of 3 × 3, 5 × 5, and 7 × 7 pixels due to the area of plots in Henan mainly being around 1 hectare. These patches are proportionate to the land scale in Henan Province. In the fusion phase, the patches were then convoluted twice to obtain the same size feature maps $\hat{U}$, $\bar{U}$, and $\tilde{U}$. Next, information of different branches was fused in the fusion stage $U = \hat{U} + \bar{U} + \tilde{U}$. Then, global average pooling F_{qp} is used to embed global information to generate channel statistics for $s \in R^C$. To reduce the dimensions of the channel statistics and obtain the final channel weight $z \in R^D$, the full connection layer F_{fc} is used. This process can be expressed by Formula (1):

$$z = F_{fc}(F_{qp}(\hat{U} + \bar{U} + \tilde{U})) \quad (1)$$

Finally, in the selection phase, three weight matrices, W_a, W_b, and W_c, where a , b, and c are related to different patch sizes, were generated using the final channel weight z. To ensure consistency, W_a, W_b, and W_c were generated by the softmax function such that the sum of the elements at the same position in the weight matrix is 1. Then, the weight matrix was used to weight the feature maps $\hat{P}$, $\bar{P}$, and $\tilde{P}$, and an adaptively selected spatial feature V was obtained by adding these weighted feature maps. This process can be expressed by Formulas (2) and (3):

$$W_a, W_b, W_c = F_{softmax}(z) \tag{2}$$

$$V = W_a \cdot \hat{U} + W_b \cdot \bar{U} + W_c \cdot \tilde{U} \tag{3}$$

2.3.2. Branch of Central Pixel Spectral Feature Extraction

Limited by the method used to obtain sample points, the number of samples used in crop classification is often insufficient, which forces the network to resist overfitting and generalization. However, even if patches are used as the model input, the central pixels need to be classified. Moreover, when introducing spatial information, the central pixel and other pixels in the patch are weighted equally in the classification, which may lead to classification errors. Therefore, increasing the weight of the central pixel features during the classification process is necessary.

As shown in Figure 7b, we used a separate branch to extract the spectral information for the 12 bands of the central pixel. TabNet [44], characterized by its use of a convolutional network structure to simulate the RF operation process, was implemented for feature extraction. Compared with deep learning methods, traditional machine learning methods such as RF are more dependable when training small samples and have a better anti-overfitting ability [49]. Therefore, we used TabNet to simulate an RF for spectral feature extraction. Multitask learning involves designing multiple related tasks of a neural network using prior knowledge and then accelerating the training and convergence of the network by optimizing multiple tasks simultaneously. Multitask learning obtains more comprehensive and controllable information from data, which can improve the ability of the network to extract certain features [50]. Therefore, to improve the extraction of spectral features and increase the weight of the central pixel in the classification process, we use multitask learning alongside an auxiliary loss function to supervise the spectral information extraction process of the central pixel. Finally, the extracted spectral features of the central pixel and the spatial features obtained by SPM were multiplied to make full use of the spectral information of the central pixel, thereby improving classification accuracy.

2.3.3. Loss Function

The loss function comprises two parts of the network. In this study, cross entropy loss (*CE loss*) [51] was used as the loss function of the network. *CE loss* is the most used loss function for multiple classification tasks because its value is only related to the probability of the correct class and its derivation process is convenient. The formula for *CE loss* is:

$$Loss_{CE} = -\frac{1}{N}\sum_{i=1}^{N}(y_i log\hat{y}_i + (1 - y_i)(1 - log\hat{y}_i)) \tag{4}$$

In Formula (4), y_i is the probability that the ground truth is true, $\hat{y}_i$ is the probability that the forward propagation result is true, and N is the number of classes. The total loss function of the network consists of two branch loss functions, which can be expressed as:

$$Loss = \lambda_1 \times Loss_1 + \lambda_2 \times Loss_2 \tag{5}$$

In Formula (5), $\lambda_1 + \lambda_2 = 1$, $Loss_1$, and $Loss_2$ are both CE losses. The weight of the spectral information of the central pixels during the classification process can be adjusted by adjusting the values of λ_1 and λ_2.

2.3.4. Superpixel Optimization

Due to the low spatial resolution of Sentinel-2 images, mixed pixel phenomena are frequently observed, especially in densely distributed crops with a high variation in crop type over short distances. For example, the characteristics of pixels at the junction of corn and peanut are different from those of peanut and corn, which brings great difficulties in crop classification. Superpixel segmentation is an over-segmentation technique that obtains image objects with accurate boundaries by clustering the features of image pixels. Simple linear iterative clustering (SLIC) [52] is a classical superpixel segmentation method that converts the colors of an image from the RGB color space to the CIELab color space and clusters according to the distance of pixels in the color space to obtain an accurate boundary of image objects. To improve the boundaries of the crop plots and reduce the influence of mixed pixels, we used SLIC to optimize our classification results.

The SLIC algorithm converts RGB images into CIELab color space and clusters pixels based on the distance between pixels in the color space. In general, the greater the color difference between pixels, the more accurate the crop plots edges that can be extracted by SLIC. Sentinel-2 L2A products have 12 bands with multiple band combinations. To select the most suitable band combination as the input of the SLIC algorithm, we first counted the reflectance of the 12 bands in Sentinel-2 images of our sample. The results are shown in the box diagram [53] in Figure 8, where the reflectance of the three crops is mainly different in the four red edge bands of Band 5, Band 6, Band 7, and Band 8A and in the near-infrared band of Band 8. In some other bands such as B4, B5, and B11, the reflectance of the three crops is also partially different and to reduce the impact of mixed pixels, we prefer to use the highest resolution bands, namely, Band 2, Band 3, Band 4, and Band 8, to carry out band combination. Using our samples, we combined the above factors to calculate the average distance of different crops in the CIELab space under different band combinations. The band combinations used include color syntheses (Band 4, Band 3, Band 2), standard false color synthesis (Band 8, Band 6, Band 4), and commonly used Sentinel-2 agricultural bands (Band 8, Band 11, Band 2) [54]. We also performed a principal component analysis (PCA) [55] to map the 12 bands of the Sentinel-2 image to three principal component bands to calculate the distance of different crops in the CIELab space. Finally, the experimental results are shown in Table 3.

In the experimental results, the average distances between peanut and corn and between peanut and rice in the color space of the band combinations of Band 8, Band 11, and Band 4 were the largest at 26.5757 and 20.7087, respectively. These results have great advantages over the second-largest distance. Under this band combination, although the average distance between corn and rice in the CIELab color space is not the largest, the result of 6.9893 has no obvious disadvantage compared with the maximum of 7.8800. Therefore, when using the SLIC algorithm for post-processing, we used an RGB image composed of Band 8, Band 11, and Band 4 as input.

Table 3. The average distance of three crops (corn, peanut, and rice) in CIELab Color Space under different band combinations. The optimal accuracies are boldened, and the sub-optimal accuracies are underlined. The last row of the table features the PCA results. B1 represents Band 1 in the table.

Band Combination	Peanut–Corn	Peanut–Rice	Corn–Rice
B4, B3, B2	10.6995	3.7830	**7.8800**
B8, B4, B3	<u>14.2403</u>	7.2344	7.7433
B8, B5, B4	6.6000	13.1003	9.2887
B8, B6, B4	10.5776	<u>14.3992</u>	<u>7.8531</u>
B8, B7, B4	6.7159	5.3856	3.0826
B8, B9, B4	5.1530	4.2164	3.9393
B8, B11, B4	**26.5757**	**20.7087**	6.9893
PCA	13.5527	9.1596	2.9907

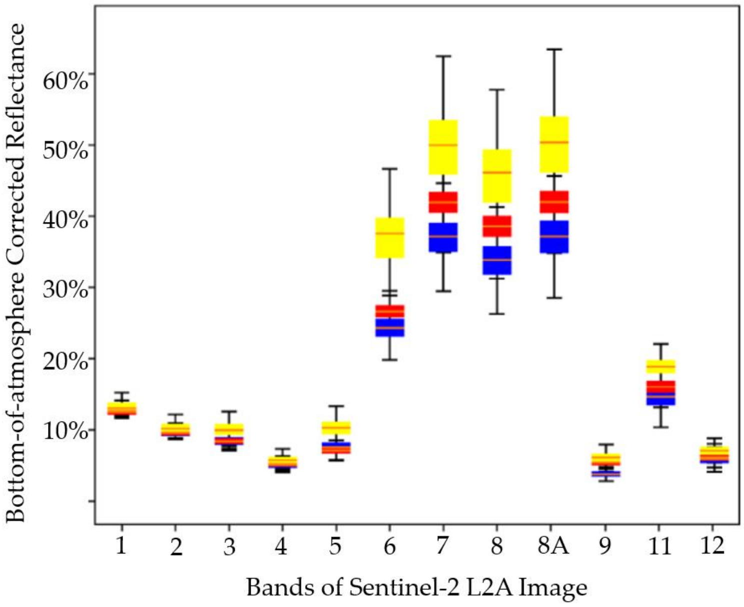


Figure 8. Reflectance of corn, peanut, and rice in different bands of Sentinel-2 images. Red represents corn, yellow represents peanut, and blue represents rice.

As shown in Figure 9, we used the RGB false color image formed by by Band 8, Band 11, and Band 4 of the Sentinel-2 image after performing image enhancement operations, such as image stretching, as input to the SLIC to obtain superpixel segmentation blocks of the image. Then, each superpixel segmentation block was traversed to calculate the area proportion of each class of pixels. If the proportion of a certain class of pixels exceeds threshold θ, all pixels in the superpixel block were modified to this category. Otherwise, the category of each pixel remained unchanged.

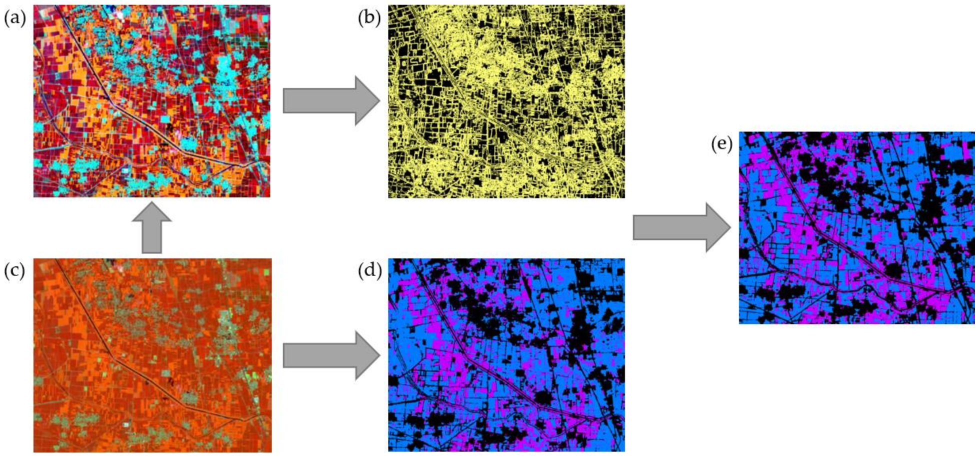


Figure 9. Schematic of the superpixel optimization process. (**a**) Enhanced Band 8, Band 11, Band 4 false color image. (**b**) SLIC extraction results. (**c**) Band 8, Band 11, Band 4 false color synthesis of the Sentinel-2 image. (**d**) Classified result. (**e**) Optimized result.

2.3.5. Evaluation Metrics

To evaluate the classification results, we used a confusion matrix. The confusion matrix is a standard format for evaluating crop classification accuracy. In the confusion matrix, the number of rows, N, represents the number of categories to be evaluated. The elements $P_{i,j}$ of i rows and j columns represent the number of pixels that are actually class i but are predicted to be class j. Through the confusion matrix, we mainly used four types of accuracy evaluation metrics. First, we used producer accuracy (PA) metrics, which is the proportion of the number of pixels correctly classified into the class to the total number of pixels in the class. Second, we used user accuracy (UA) metrics, which refers to the proportion of pixels correctly classified into the class and the total number of pixels classified into the class. Using PA and UA, we can analyze the classification ability of each type of crop and explore the reasons for the change in accuracy. The overall accuracy (OA) refers to the proportion of all correctly classified pixels in the total number of pixels. Finally, the kappa coefficient (KC) is an indicator of consistency and can also be used to measure the effect of classification. The kappa coefficient is shown to be a more discerning statistical tool for assessing the classification accuracy of different classifiers and has the added advantage of being statistically testable against the standard normal distribution [56]. In the classification problem, consistency refers to whether the model prediction results are consistent with the actual classification results.At the same time, we We also used the F1 score as an evaluation metric because the classification process focuses on more than just the recall or accuracy rates. The $F1$ score, which considers the accuracy and recall rate, is a commonly used evaluation index to evaluate classification accuracy better. The $F1$ score can be regarded as the harmonic mean of the model's accuracy and recall. The relevant formulas are as follows:

$$Precision = \frac{TP}{TP + FP} \tag{6}$$

$$Recall = \frac{TP}{TP + FN} \tag{7}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{8}$$

In Formulas (6)–(8), true positive (TP) represents pixels correctly classified as positive pixels. False positives (FP) are pixels incorrectly classified as a positive pixel. False negatives (FN) are pixels incorrectly classified as negative.

2.4. Train Details

We used the PyTorch deep learning framework to build all deep learning models used in this experiment. All training and validation samples were used in the training phase. We trained our network using one NVIDIA TITAN XP GPU (12 GB of memory). In terms of the the training parameters, the batch size was set to 256, Adam [57] was used as the optimizer, the initial learning rate was set to 0.001, and cosine annealing [58] was used to reduce the learning rate gradually. In addition, a drop rate of 0.4 was added. For the hyperparameter setting of the loss function, we set the value of λ_1 to 0.7 and λ_2 to 0.3. In the post-processing process to balance superpixel segmentation and crop extraction, we set the hyperparameter θ in the superpixel optimization to 0.6.

3. Results

3.1. Comparing Methods

We conducted quantitative and qualitative comparisons of the Henan crop dataset obtained using other mainstream crop classification methods, such as RF [12], XGBoost [18], CNN [38], CNN-RF [40], and S^3ANet [59]. RF and XGBoost are commonly used machine learning classification methods in crop classification mapping. However, they only use the spectral information of pixels for classification. In our experiments, the input of these two methods was a 1 × 12 vector of 12 band values for a single pixel. The CNN is the most basic deep learning classification network. The combined CNN-RF first extracts feature through a CNN and then uses RF for classification, which can improve the model's generalization ability by leveraging the advantages of the two methods. For fairness of comparison, the CNN network depths of these two methods are the same as those in our method. S^3ANet uses various attention methods to weight spatial, scale, and spectral information to improve crop classification accuracy. All deep learning methods in the experiment used a 7 × 7 × 12 vector as the input, which is a patch with a central pixel size of 7. The other parameter settings of all deep learning methods are the same as those described in Section 2.4. Machine learning methods were constructed in the Scikit-learn python library.

3.1.1. Quantitative Comparisons

Our comparison results are shown in Tables 4 and 5. As shown in Table 4, SPTNet has achieved competitive overall and single-crop accuracy results. Our method also has advantages in terms of the parameters and inference time. Regarding the non-deep learning algorithms, the performance of RF alone was insufficient, with an F1 score of 0.8501. In contrast, XGBoost obtained a slightly higher F1 score of 0.8741, which is close to the accuracy of the deep learning method CNN-RF. Additionally, using the deep learning method, CNN achieves an F1 score of 0.8217, which is the lowest accuracy among all of the methods and can be attributed to the many false detections of the background. It may also be due to the lack of attention to spectral information in the CNN, so some minor crop plots or backgrounds may have been missed. The accuracy of CNN-RF is significantly improved compared with that of CNN, which may be because the overfitting of CNN is reduced after classification using RF. However, it also leads to a significant increase in the number of parameters and the inference time. S^3ANet, which uses a variety of attentions for information extraction, has achieved sub-optimal accuracy in multiple metrics. Our method provides the highest classification accuracy, with an F1 score and KC of 0.9653 and 0.9531, respectively.

Table 4. The quantitative comparison between the proposed method and the mainstream method in the accuracy evaluation index on the Henan crop dataset. The optimal accuracy is bolded. The inference time is obtained by averaging the inference time of each Sentinel-2 image.

Method	Others	Peanut	Corn	Rice	mF1	KC	OA	Parameters (kb)	Inference Time (mins)
RF	0.7246	0.8719	0.9562	0.8076	0.8401	0.8199	0.8521	158,851	43
XGBoost	0.8050	0.9587	0.9230	0.7950	0.8704	0.8234	0.8741	12,456	39
CNN	0.6037	0.9131	0.7051	0.8862	0.7770	0.8217	0.7221	**3581**	**20**
CNN-RF	0.8162	0.9703	0.8199	0.9068	0.8774	0.8936	0.8801	738,547	78
S^3ANet	0.9109	0.9599	0.9329	0.9182	0.9305	0.9358	0.9305	24,250	52
SPTNet	**0.9624**	**0.9730**	**0.9664**	**0.9590**	**0.9652**	**0.9531**	**0.9656**	9731	45

Table 5. The quantitative comparison between the proposed method and the mainstream method in the PA and UA values. The optimal accuracy is bolded.

Method	Metrics	Others	Peanut	Corn	Rice
RF	PA	0.6074	0.9452	0.9705	0.8851
	UA	0.8980	0.8092	0.9424	0.7426
XGBoost	PA	0.7395	0.9646	0.9397	0.8831
	UA	0.8833	0.9530	0.9069	0.7229
CNN	PA	0.4428	0.9667	0.9653	0.9510
	UA	0.9484	0.8653	0.5554	0.8298
CNN-RF	PA	0.6974	**0.9729**	0.9662	0.9728
	UA	**0.9734**	0.9678	0.7121	0.8493
S^3ANet	PA	0.8845	0.9376	0.9533	**0.9756**
	UA	0.9391	**0.9834**	0.9135	0.8672
SPTNet	PA	**0.9639**	0.9689	**0.9696**	0.9599
	UA	0.9610	0.9773	**0.9634**	**0.9582**

3.1.2. Visualization Results

A comparison of the visualization results is shown in Figure 10. Specifically, many small objects are in the first and second rows of Figure 10. The difference is that the first row is a typical mixed planting area of corn and peanut, where the crop plots are small and broken. In the second row, although the distribution of crops was more concentrated, many roads shuttled through the field. The features of these roads are not as obvious as those of large-area crops. RF and XGBoost use only the spectral information of pixels and cannot accurately distinguish between different crops, leading to confusion around the classification of peanut and corn in their results. Other deep learning methods directly use 7 × 7 pixels patches, which are too large compared to the crop plots in these scenarios. For this reason, they have poor extraction effects on small crop plots and other small surface features such as roads and greenhouses, resulting in categorical confusion when classifying crop plots and background erosion. Our method uses SPM to adaptively fuse the features of different patch sizes for specific scenarios, which can improve the extraction ability of small surface features. Therefore, our model could accurately classify roads in fields and small crop plots. There were small and alternately planted corn and peanut where the first row was marked. Our method was able to accurately distinguish between them accurately, whereas the extraction results of other methods have obvious misclassifications. In the area marked in the second row, small roads were misclassified by most other methods, but our method accurately distinguished these roads from crops because of the better extraction of small features. Moreover, after using SLIC for post-processing, the crop plots are more regular. Other methods could not distinguish the mixed pixels, which is manifested in the images where the connected parts of different crops are classified as the background, and the extraction results of some main roads are too wide. These conditions were reduced in our results because of the use of the SLIC. In the lower-left marker of the third row, there is a river in the lower left part of the image, which is covered by aquatic plants such as cyanobacterial blooms. The image characteristics of these cyanobacterial blooms are similar to rice. Since this scenario is not common, such negative samples do not exist in our dataset. Without such negative samples, most methods misclassify them as rice. Our method is more sensitive to the spectral information of pixels because it uses TabNet and multitask learning to model and supervise the spectral information of pixels separately. The unique structure of TabNet can enhance the generalization ability of the network. Therefore, our method had the lowest number of false classifications in this scenario. In the last row, rice is scattered owing to the influence of terrain. Rice is difficult to classify because of the presence of similar grasslands and woodlands on the hills. However, compared with other methods, our method distinguished rice better from the surrounding grassland or woodland with the lowest number of false detections, and it extracted slender paddy fields between hills.

In summary, SPTNet can effectively reduce the number of missed or incorrect detections in the crop extraction results, whether in the plain where crops are mixed or in small hilly lands. Overall, RF and XGBoost based on single-pixel results exhibited obvious salt and pepper noise and could not distinguish different crop categories well, indicating that deep learning methods are not necessarily superior to traditional methods. Many details of the image appear to be embodied in small crop plots, and some road classification errors are observed because CNN uses convolution operations and lacks restrictions. The other three deep learning methods performed better than the first three methods due to crop classification improvements.

3.2. Ablation Study

Ablation experiments were conducted on the Henan crop dataset to verify the contribution of the proposed module to crop classification. We used a CNN network with the same depth as that in our model as the baseline. We then changed or added parts of the structure to our proposed modules. We divided them into the following experiments according to the different modules added: Experiment 1: Test baseline. Experiment 2: Replace the first four layers of the Baseline with SPM. Experiment 3: Add a TabNet branch to extract the spectral features of the central pixels. Experiment 4: Use a multitask learning strategy to supervise the TabNet branch. Experiment 5: Add the SLIC algorithm for post-processing.

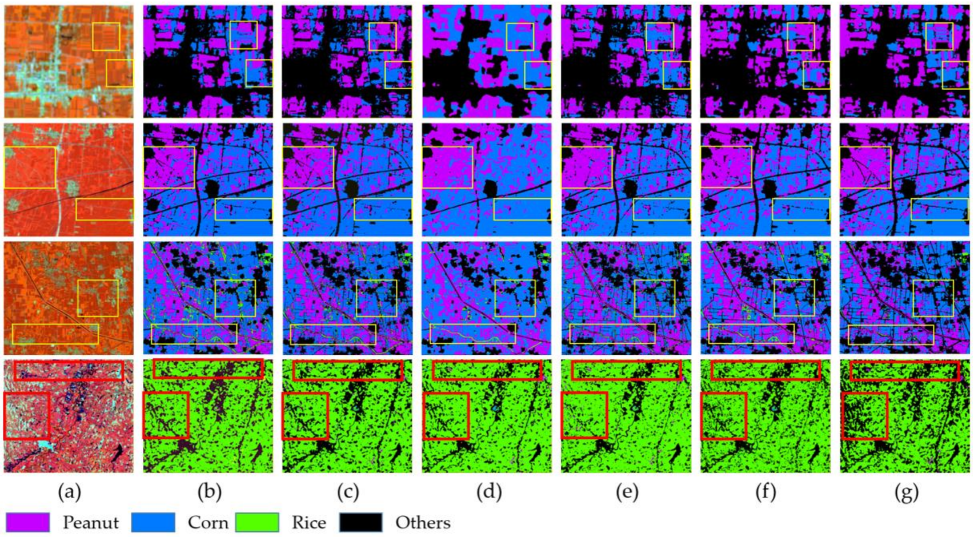


Figure 10. Some examples of the results on the dataset. From left to right: (**a**) Band 8, Band 11, Band 4 false color synthesis of Sentinel-2 image; (**b**) RF; (**c**) XGBoost; (**d**) CNN; (**e**) CNN-RF; (**f**) S^3ANet; (**g**) SPTNet.

3.2.1. Quantitative Comparisons

The ablation results are summarized in Tables 6 and 7. The CNN's lack of attention to spectral information and small crop plots resulted in many false positives and a fairly low F1 score of 0.8217. After adding the SPM module, the experimental accuracy is significantly improved to 0.8989 because SPM can adaptively fuse the features of patches of different sizes for specific scenes. Therefore, a better extraction ability for small objects was obtained, and the extraction accuracy of small crop plots or roads was higher. From the quantitative results, most PA and UA values of categories increased after applying SPM. Only the peanut's PA and the background's UA decreased slightly ($0.9667 - 0.9552$, $0.9484 - 0.9115$). This result shows that using the SPM module can effectively reduce errors and missed detection in the crop classification process. After adding the TabNet branch to extract the spectral features of the central pixel, we obtained an F1 score of 0.9456. With the addition of the TabNet network, the PA of crops decreased while the UA of crops increased compared with the results of the baseline CNN and the results after adding the SPM. The network is more sensitive to the spectral information of the center point, and the UA of various crops increases, thereby enhancing the direct extraction of spectral information. Despite this, more constraints are consequently added to the crop classification process, resulting in a decrease in the PA of various crops. Furthermore, we used multitask learning to supervise the TabNet, which was implemented to optimize the central pixel spectral feature extraction process in Experiment 3. After optimizing the multitask learning strategy, the UA and PA of various crops could be balanced to improve the classification accuracy of the network. The classification accuracy of Experiment 3 reached an F1 score of 0.9523. In the last experiment, we added the SLIC algorithm for post-processing. The results in Tables 6 and 7 show that the superpixel segmentation optimization strategy improves the UA and PA of various crops. The final network accuracy has an F1 score of 0.9653 because of superpixel segmentation's more accurate boundary information.

Table 6. Ablation experiments for the network design. The optimal accuracy value in each is bolded.

Baseline	SPM	TabNet	Multitask	SLIC	Others	Peanut	Corn	Rice	mF1	KC	OA
✓					0.6037	0.9131	0.7051	0.8862	0.7770	0.8217	0.7891
✓	✓				0.8274	0.9369	0.9675	0.8723	0.9010	0.8836	0.9256
✓	✓	✓			0.9332	0.9525	0.9492	0.9434	0.9446	0.9226	0.9435
✓	✓	✓	✓		0.9411	0.9698	0.9562	0.9409	0.9520	0.9530	0.9522
✓	✓	✓	✓	✓	**0.9624**	**0.9730**	**0.9664**	**0.9590**	**0.9652**	**0.9531**	**0.9656**

Table 7. Ablation experiments for the network design evaluate in the PA and UA values. The optimal accuracy is bolded.

Baseline	SPM	TabNet	Multitask	SLIC	Metrics	Others	Peanut	Corn	Rice
✓					PA	0.4428	0.9667	0.9653	0.9510
					UA	0.9484	0.8653	0.5554	0.8298
✓	✓				PA	0.7576	0.9552	**0.9705**	0.9551
					UA	0.9115	0.9194	0.9646	0.8028
✓	✓				PA	**0.9790**	0.9303	0.9230	0.9217
					UA	0.8916	0.9758	0.9771	**0.9663**
✓	✓	✓			PA	0.9236	0.9541	0.9624	**0.9639**
					UA	0.9694	9708	0.9433	0.9228
✓	✓	✓	✓	✓	PA	0.9639	**0.9689**	0.9696	0.9599
					UA	**0.9610**	**0.9773**	**0.9634**	0.9582

3.2.2. Visualization Results

To qualitatively compare the contribution of the proposed modules, the visualization results of some ablation experiments are shown in Figure 11, which shows that the extraction results using all proposed modules are optimal and the proposed modules achieve the anticipated visualization results.

The first row of Figure 11 shows waterweeds that are easily confused with rice since many tiny roads and small crop plots travel between fields. The baseline CNN incorrectly classified some waterweeds as rice because the small crop plots could not be accurately distinguished and the roads were not successfully extracted. The SPM structure can improve the model's ability to extract small objects; therefore, some roads and other crops covered by large areas of crops can be classified. However, the model still made false detections, and more waterweeds were classified as rice, indicating that using only the patch's spatial information is insufficient to accurately classify crops. After adding TabNet to model the spectral information of the central pixels and applying multitask learning for supervision, false detection was greatly reduced. In the results, water plants were correctly classified as the background, and more roads were extracted. Finally, SLIC was used for post-processing because the superpixel segmentation algorithm can obtain more accurate edges. In the final extraction results, the shape of the crop plots was more regular, and the road results were more accurate. The difference in the classification of roads in each experiment is more obvious in the second row wherein the crops are densely distributed. There are several obvious roads in the upper right part of the image, but they are spread over large areas of peanut. In this case, the main class of pixels in the road patch is peanut in this case. Hence, the baseline CNN completely ignored them and misclassified them as peanut. After adding SPM, however, the network's ability to extract small objects was improved. Despite this, owing to the mixed pixels in the image, the road in the peanut image shows different characteristics from those of the pure road pixels. Although the network could distinguish them after SPM was added, they were still misclassified as corn. After using TabNet to extract the spectral features of the central pixels separately, the classification ability of the ground objects was improved, and some roads were extracted. The weight of the central pixels' spectral information increased after using multitask learning for supervision. Although more roads were extracted, some were still classified as corn. However, SLIC could extract the road into a superpixel block. Therefore, after SLIC post-processing, the more complete roads are distinguished. In the last row, rice is scattered owing to the influence of terrain. With the gradual addition of the proposed module, the surrounding grassland or woodland misclassified as rice was obviously reduced. Finally, our method accurately extracted the rice between hills, and the edges of the crop plots were more accurate.

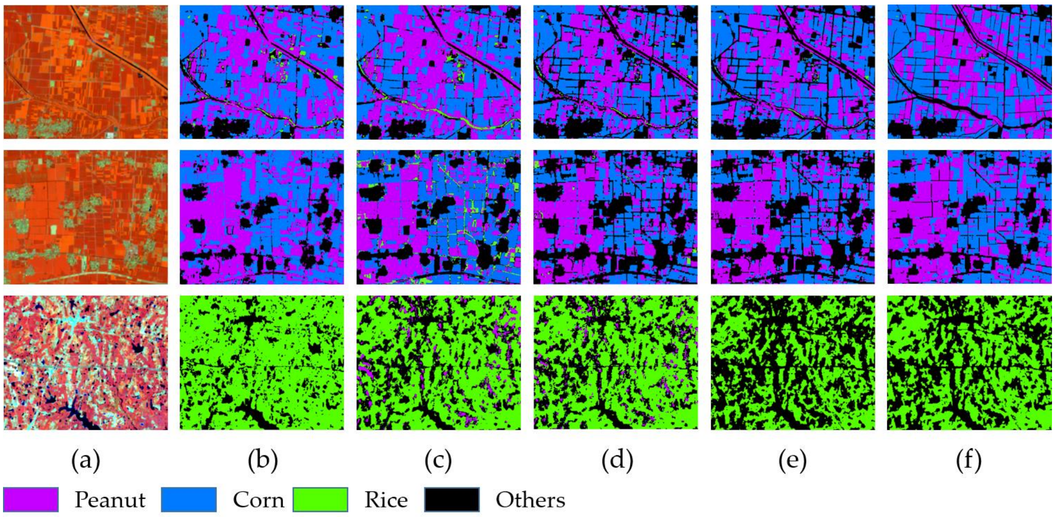


Figure 11. Some examples of crop classification ablation experiments. From left to right: (**a**) Band 8, Band 11, Band 4 false color synthesis of Sentinel-2 image; (**b**) baseline; (**c**) Experiment 1; (**d**) Experiment 2; (**e**) Experiment 3; (**f**) Experiment 4.

3.3. Crop Mapping in Henan Province

We used the proposed network structure to complete the mapping of 10 m plots of principal crops, including corn, peanut, and rice, in Henan Province, China, in 2022, and the results are shown in Figure 12.

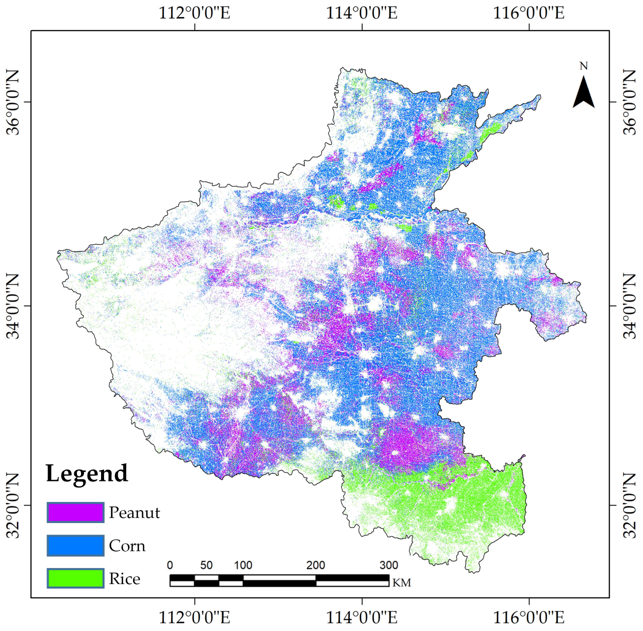


Figure 12. Mapping results of corn, peanut, and rice from Sentinel-2 images in Henan Province, China, obtained by the proposed SPTNet method.

4. Discussion

4.1. The Split Strategy Selection of SPM

In the split stage of the SK-Conv structure, the feature maps were processed using convolution kernels of different sizes to obtain the features of different receptive fields. However, our method differs from that of SK-Conv. Using our method, in the split stage, patches of different sizes were used for splitting rather than the sizes of the convolution kernels. Using the size of the convolution kernel in the split stage can obtain multi-scale receptive fields, whereas using patches can balance spatial information and noise. We conducted experiments to verify the advantages and disadvantages of these two methods. To ensure the fairness of the comparison, other structures and parameter settings of the network were consistently used across methods. The results are presented in Table 8.

The results show that the F1 score using convolutional kernel sizes in the split stage is 0.9480, which is less than that using patch sizes in the split stage. The scale of the input vector is too small to provide a multi-scale receptive field because the input of the model is a patch of 7 × 7 pixels. This eliminates the advantage of using convolutional kernels for splitting. Furthermore, using patches of the same size as the input in each branch of SPM introduces more noise, which may also be responsible for the reduced accuracy. The method of using patch sizes in the split stage introduces prior information, which can better adapt to different sizes of crop plots in complex scenes. Therefore, we used different patch sizes for the split stage in the structure.

Table 8. Ablation experiments for the network design.

Split with		Others	Peanut	Corn	Rice	mF1	KC	OA
Convolutional Kernel Size	**Patch Size**							
✓		0.9341	0.9572	0.9535	0.9470	0.9480	0.9290	0.9475
	✓	0.9624	0.9730	0.9664	0.9590	0.9653	0.9531	0.9656

4.2. The Contribution of Different Sizes of Patches to Crop Extraction

In this section, we describe the experiments conducted to explore the contribution of different patch sizes to crop extraction. Specifically, in Figure 7a, we removed the structure of one patch and retained only the other two patches. To ensure fairness in the experiment, the other parameter settings remained consistent across methods. The qualitative results of the experiments are presented in Table 9. Regardless of which patch is removed, the classification accuracy decreases. It shows that these three different patch sizes contribute to improving classification accuracy. In the case of removing the 3 × 3 pixel patches, the classification accuracy decreased the most, whereas the classification accuracy decreased the least when removing the 5 × 5 pixel patches. This shows that the 3 × 3 pixel patches have the smallest contribution. The 5 × 5 pixel patches have the largest contribution in the classification process, possibly because the overall plot size in Henan was more proportional to 5 × 5 pixel patches. Of note, the classification accuracy of rice decreased significantly in the case of removing 3 × 3 pixel patches, which may be because rice is mostly planted in smaller hilly areas.

Table 9. Experimental results for exploring the contribution of different size of patches across experiments where one plot size is removed.

3 × 3 Pixel Patch	5 × 5 Pixel Patch	7 × 7 Pixel Patch	Others	Peanut	Corn	Rice	mF1	KC	OA
	✓	✓	0.9286	0.9634	0.9534	0.9283	0.9549	0.9312	0.9502
✓		✓	0.9452	0.9667	0.9638	0.9375	0.9534	0.9374	0.9537
✓	✓		0.9332	0.9525	0.9492	0.9434	0.9556	0.9226	0.9535
✓	✓	✓	0.9624	0.9730	0.9664	0.9590	0.9653	0.9531	0.9656

4.3. Advantages and Limitations

This study proposes a dependable network for large-scale crop extraction and achieves satisfactory classification accuracy using only single-temporal Sentinel-2 images. First, the proposed structure can adaptively select and fuse the spatial information of the image according to the size of the plot in the actual scene, thereby improving its ability to extract small features. Second, the proposed structure uses TabNet and a multitask learning strategy to extract the spectral features and improve the stability of the network while modeling the spectral information of the central pixels separately. TabNet, which has a similar strcuture to a decision tree, can enhance the network's generalization ability and alleviate the overfitting problem caused by small sample sizes. Using multitask learning strategies can enhance the weight of the spectral features of pixels and further improve classification accuracy. Finally, the application of an SLIC algorithm to post-process the extraction results could improve the accuracy of the classification result boundary.

The SPTNet obtained the highest extraction accuracy than other methods, regardless of the crop. In the qualitative comparison, the SPTNet greatly improved the extraction ability of small crop plots, and the edge of crops has also been ameliorated. These comparisons prove the advantages of our method.

Although our method has obvious advantages for Henan crop extraction, there are still some problems. To avoid the problem of missing time-series images, we used single-temporal images as data sources, for which the extraction accuracy of staple crops, such as corn, peanut, and rice, is generally high. However, the classification of certain characteristic economic crops in the absence of crop phenology information remains a limiting factor. The spectral characteristics of some economic crops, such as soybeans and grapeseed, are relatively similar. Thus, using only single-temporal images to distinguish them is a challenge. Furthermore, it is difficult to overcome image resolution limitations, evidenced by the fact that the classification accuracy did not significantly improve after applying the SLIC algorithm, a superpixel segmentation method based on pixel spectral features.

In future research, we will attempt to use time-series remote sensing images for crop classification. To ensure the integrity of the data and make full use of the high-precision, single-temporal extraction method proposed in this paper, we will attempt to extract the key time points of crop phenology and classify crops under the premise of using the shortest time series data. We will also attempt to combine Sentinel-2 images with high-resolution data for crop extraction. Sentinel-2 images with high spectral resolution were used for crop extraction, and high-resolution images with high spatial resolution were used for plot boundary extraction. We hope to optimize the boundaries of the crop extraction results through high-spatial-resolution plot extraction results, which can improve the overall accuracy of crop classification. Finally, we will try to add SAR data for crop classification. SAR satellites can image in any weather condition, which can make up for the lack of optical data. Furthermore, it is difficult to overcome image resolution limitations, evidenced by the fact that the classification accuracy did not significantly improve after applying the SLIC algorithm, a superpixel segmentation method based on pixel spectral features.

5. Conclusions

Large-scale crop extraction is essential for grain security and sustainable agriculture. This paper proposes a crop classification method using single-temporal sentinel images to better assess large-scale and high-precision crop extraction, which has been successfully applied to three main autumn crops in Henan Province: corn, peanut, and rice. To improve the extraction ability of small crop plots in complex scenes and mitigate the negative impact of insufficient samples, this paper proposes SPM, which can adaptively fuse the features of different patch sizes. In addition, to improving the classification ability of the central pixels, we used the TabNet network to extract spectral features and improve the stability of the network. Furthermore, we use multitask learning to introduce auxiliary loss and increase the weight of the central pixels' spectral information in the classification process. Finally, we introduce the SLIC superpixel segmentation method for post-processing to reduce the impact of mixed pixels and ameliorate the boundaries of crop plots. Qualitative and quantitative analysis shows that compared with different mainstream classification methods, SPTNet achieves state-of-the-art performance with F1 scores of 96.53% on our Henan crop classification dataset, which indicates that our proposed structure effectively improves the large-scale crop classification process. Large-scale crop classification mapping of corn, peanut, and rice in Henan Province also proves the effectiveness of our method in practical application.

In future work, we will attempt to use high-resolution and Sentinel-2 images for crop classification and combine the advantages of these two images to obtain finer crop extraction results. We will also attempt to incorporate phenological information at a minimum cost to enhance the extraction of more crop varieties.

Author Contributions: Conceptualization, J.H. and Z.C.; methodology, J.H. and Y.B.; validation, X.Y. and Y.B.; formal analysis, G.L.; resources, Z.C.; data curation, Y.B.; writing—original draft preparation, G.L. and X.T.; writing—review and editing, X.T.; visualization, X.T.; supervision, X.Y.; project administration, Z.C.; funding acquisition, Z.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the National Natural Science Foundation of China (Grant No. 42071407, 42071321).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not publicly available.

Acknowledgments: The authors are grateful to the editors and anonymous reviewers for their informative suggestions.

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

Abbreviations

The following abbreviations are used in this manuscript:

CE	cross entropy
CNN	convolutional neural network
DNN	deep neural network
EVI	enhanced vegetation index
FN	false negative
FP	false positive
GRU	gated recurrent unit
KC	kappa coefficient
LSTM	long short-term memory
LSWI	land surface water index
NDVI	normalized difference vegetation index
PA	producer accuracy
PCA	principal component analysis
RF	random forest
RNN	recurrent neural network
SLIC	simple linear iterative clustering
SPM	selective patch module
TP	true positive
UA	user accuracy

References

1. Chen, X.; Cui, Z.; Fan, M.; Vitousek, P.; Zhao, M.; Ma, W.; Wang, Z.; Zhang, W.; Yan, X.; Yang, J.; et al. Producing more grain with lower environmental costs. *Nature* **2014**, *514*, 486–489. [CrossRef]
2. Kuzman, B.; Petković, B.; Denić, N.; Petković, D.; Ćirković, B.; Stojanović, J.; Milić, M. Estimation of optimal fertilizers for optimal crop yield by adaptive neuro fuzzy logic. *Rhizosphere* **2021**, *18*, 100358. [CrossRef]

3. Jez, J.M.; Topp, C.N.; Silva, G.; Tomlinson, J.; Onkokesung, N.; Sommer, S.; Mrisho, L.; Legg, J.; Adams, I.P.; Gutierrez-Vazquez, Y.; et al. Plant pest surveillance: From satellites to molecules. *Emerg. Top. Life Sci.* **2021**, *5*, 275–287. [CrossRef]
4. Rasti, S.; Bleakley, C.J.; Holden, N.; Whetton, R.; Langton, D.; O'Hare, G. A survey of high resolution image processing techniques for cereal crop growth monitoring. *Inf. Process. Agric.* **2021**, *9*, 300–315. [CrossRef]
5. Wen, Y.; Li, X.; Mu, H.; Zhong, L.; Chen, H.; Zeng, Y.; Miao, S.; Su, W.; Gong, P.; Li, B.; et al. Mapping corn dynamics using limited but representative samples with adaptive strategies. *ISPRS J. Photogramm. Remote Sens.* **2022**, *190*, 252–266. [CrossRef]
6. Gallego, J.; Carfagna, E.; Baruth, B. Accuracy, objectivity and efficiency of remote sensing for agricultural statistics. In *Agricultural Survey Methods*; John Wiley & Sons, Inc.: Hoboken, NJ, USA, 2010; pp. 193–211.
7. Kussul, N.; Lavreniuk, M.; Skakun, S.; Shelestov, A. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geosci. Remote Sens. Lett.* **2017**, *14*, 778–782. [CrossRef]
8. Bargiel, D. A new method for crop classification combining time series of radar images and crop phenology information. *Remote Sens. Environ.* **2017**, *198*, 369–383. [CrossRef]
9. Haboudane, D.; Miller, J.R.; Tremblay, N.; Zarco-Tejada, P.J.; Dextraze, L. Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture. *Remote Sens. Environ.* **2002**, *81*, 416–426. [CrossRef]
10. Zhang, L.; Gao, L.; Huang, C.; Wang, N.; Wang, S.; Peng, M.; Zhang, X.; Tong, Q. Crop classification based on the spectrotemporal signature derived from vegetation indices and accumulated temperature. *Int. J. Digit. Earth* **2022**, *15*, 626–652. [CrossRef]
11. Huang, X.; Huang, J.; Li, X.; Shen, Q.; Chen, Z. Early mapping of winter wheat in Henan province of China using time series of Sentinel-2 data. *GIScience Remote Sens.* **2022**, *59*, 1534–1549. [CrossRef]
12. Yang, N.; Liu, D.; Feng, Q.; Xiong, Q.; Zhang, L.; Ren, T.; Zhao, Y.; Zhu, D.; Huang, J. Large-scale crop mapping based on machine learning and parallel computation with grids. *Remote Sens.* **2019**, *11*, 1500. [CrossRef]
13. Mingwei, Z.; Qingbo, Z.; Zhongxin, C.; Jia, L.; Yong, Z.; Chongfa, C. Crop discrimination in Northern China with double cropping systems using Fourier analysis of time-series MODIS data. *Int. J. Appl. Earth Obs. Geoinf.* **2008**, *10*, 476–485. [CrossRef]
14. Xiao, X.; Boles, S.; Frolking, S.; Li, C.; Babu, J.Y.; Salas, W.; Moore, B., III. Mapping paddy rice agriculture in South and Southeast Asia using multi-temporal MODIS images. *Remote Sens. Environ.* **2006**, *100*, 95–113. [CrossRef]
15. Hearst, M.A.; Dumais, S.T.; Osuna, E.; Platt, J.; Scholkopf, B. Support vector machines. *IEEE Intell. Syst. Their Appl.* **1998**, *13*, 18–28. [CrossRef]
16. Ahmed, M.; Seraj, R.; Islam, S.M.S. The k-means algorithm: A comprehensive survey and performance evaluation. *Electronics* **2020**, *9*, 1295. [CrossRef]
17. Nitze, I.; Schulthess, U.; Asche, H. Comparison of machine learning algorithms random forest, artificial neural network and support vector machine to maximum likelihood for supervised crop type classification. In Proceedings of the 4th GEOBIA, Rio de Janeiro, Brazil, 7–9 May 2012; Volume 79, p. 3540.
18. Chen, T.; He, T.; Benesty, M.; Khotilovich, V.; Tang, Y.; Cho, H.; Chen, K. Xgboost: Extreme Gradient Boosting. R Package Version 0.4-2; 2015; pp. 1–4. Available online: http://cran.fhcrc.org/web/packages/xgboost/vignettes/xgboost.pdf (accessed on 5 April 2023).
19. You, N.; Dong, J.; Li, J.; Huang, J.; Jin, Z. Rapid early-season maize mapping without crop labels. *Remote Sens. Environ.* **2023**, *290*, 113496. [CrossRef]
20. Immitzer, M.; Vuolo, F.; Atzberger, C. First experience with Sentinel-2 data for crop and tree species classifications in central Europe. *Remote Sens.* **2016**, *8*, 166. [CrossRef]
21. Waldner, F.; Lambert, M.J.; Li, W.; Weiss, M.; Demarez, V.; Morin, D.; Marais-Sicre, C.; Hagolle, O.; Baret, F.; Defourny, P. Land cover and crop type classification along the season based on biophysical variables retrieved from multi-sensor high-resolution time series. *Remote Sens.* **2015**, *7*, 10400–10424. [CrossRef]
22. Bagnall, A.; Lines, J.; Bostrom, A.; Large, J.; Keogh, E. The great time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Min. Knowl. Discov.* **2017**, *31*, 606–660. [CrossRef]
23. Belgiu, M.; Csillik, O. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. *Remote Sens. Environ.* **2018**, *204*, 509–523. [CrossRef]
24. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef]
25. Yuan, X.; Shi, J.; Gu, L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Syst. Appl.* **2021**, *169*, 114417. [CrossRef]
26. Wen, D.; Huang, X.; Bovolo, F.; Li, J.; Ke, X.; Zhang, A.; Benediktsson, J.A. Change detection from very-high-spatial-resolution optical remote sensing images: Methods, applications, and future directions. *IEEE Geosci. Remote Sens. Mag.* **2021**, *9*, 68–101. [CrossRef]
27. Sun, X.; Wang, P.; Wang, C.; Liu, Y.; Fu, K. PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery. *ISPRS J. Photogramm. Remote Sens.* **2021**, *173*, 50–65. [CrossRef]
28. Gu, L.; He, F.; Yang, S. Crop classification based on deep learning in northeast China using sar and optical imagery. In Proceedings of the 2019 SAR in Big Data Era (BIGSARDATA), Beijing, China, 5–6 August 2019; IEEE: New York, NY, USA, 2019; pp. 1–4.
29. Yuan, Q.; Shen, H.; Li, T.; Li, Z.; Li, S.; Jiang, Y.; Xu, H.; Tan, W.; Yang, Q.; Wang, J.; et al. Deep learning in environmental remote sensing: Achievements and challenges. *Remote Sens. Environ.* **2020**, *241*, 111716. [CrossRef]

30. Wang, L.; Wang, J.; Liu, Z.; Zhu, J.; Qin, F. Evaluation of a deep-learning model for multispectral remote sensing of land use and crop classification. *Crop J.* **2022**, *10*, 1435–1451. [CrossRef]
31. Xu, J.; Yang, J.; Xiong, X.; Li, H.; Huang, J.; Ting, K.; Ying, Y.; Lin, T. Towards interpreting multi-temporal deep learning models in crop mapping. *Remote Sens. Environ.* **2021**, *264*, 112599. [CrossRef]
32. Yu, Y.; Si, X.; Hu, C.; Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **2019**, *31*, 1235–1270. [CrossRef]
33. Ren, L.; Cheng, X.; Wang, X.; Cui, J.; Zhang, L. Multi-scale dense gate recurrent unit networks for bearing remaining useful life prediction. *Future Gener. Comput. Syst.* **2019**, *94*, 601–609. [CrossRef]
34. Pullanagari, R.; Dehghan-Shoar, M.; Yule, I.J.; Bhatia, N. Field spectroscopy of canopy nitrogen concentration in temperate grasslands using a convolutional neural network. *Remote Sens. Environ.* **2021**, *257*, 112353. [CrossRef]
35. Zhong, L.; Hu, L.; Zhou, H. Deep learning based multi-temporal crop classification. *Remote Sens. Environ.* **2019**, *221*, 430–443. [CrossRef]
36. Zhao, H.; Duan, S.; Liu, J.; Sun, L.; Reymondin, L. Evaluation of five deep learning models for crop type mapping using sentinel-2 time series images with missing information. *Remote Sens.* **2021**, *13*, 2790. [CrossRef]
37. Zhao, J.; Zhong, Y.; Hu, X.; Wei, L.; Zhang, L. A robust spectral-spatial approach to identifying heterogeneous crops using remote sensing imagery with high spectral and spatial resolutions. *Remote Sens. Environ.* **2020**, *239*, 111605. [CrossRef]
38. Xie, B.; Zhang, H.K.; Xue, J. Deep convolutional neural network for mapping smallholder agriculture using high spatial resolution satellite image. *Sensors* **2019**, *19*, 2398. [CrossRef]
39. Seydi, S.T.; Amani, M.; Ghorbanian, A. A Dual Attention Convolutional Neural Network for Crop Classification Using Time-Series Sentinel-2 Imagery. *Remote Sens.* **2022**, *14*, 498. [CrossRef]
40. Yang, S.; Gu, L.; Li, X.; Jiang, T.; Ren, R. Crop classification method based on optimal feature selection and hybrid CNN-RF networks for multi-temporal remote sensing imagery. *Remote Sens.* **2020**, *12*, 3119. [CrossRef]
41. Kussul, N.; Lemoine, G.; Gallego, F.J.; Skakun, S.V.; Lavreniuk, M.; Shelestov, A.Y. Parcel-based crop classification in Ukraine using Landsat-8 data and Sentinel-1A data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2016**, *9*, 2500–2508. [CrossRef]
42. Wang, D.; Cao, W.; Zhang, F.; Li, Z.; Xu, S.; Wu, X. A review of deep learning in multiscale agricultural sensing. *Remote Sens.* **2022**, *14*, 559. [CrossRef]
43. Orynbaikyzy, A.; Gessner, U.; Conrad, C. Crop type classification using a combination of optical and radar remote sensing data: A review. *Int. J. Remote Sens.* **2019**, *40*, 6553–6595. [CrossRef]
44. Arik, S.Ö.; Pfister, T. Tabnet: Attentive interpretable tabular learning. *Proc. Aaai Conf. Artif. Intell.* **2021**, *35*, 6679–6687. [CrossRef]
45. Sun, Y.; Qin, Q.; Ren, H.; Zhang, T.; Chen, S. Red-edge band vegetation indices for leaf area index estimation from sentinel-2/msi imagery. *IEEE Trans. Geosci. Remote Sens.* **2019**, *58*, 826–840. [CrossRef]
46. Li, X.; Wang, W.; Hu, X.; Yang, J. Selective kernel networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 510–519.
47. Li, R.; Zheng, S.; Zhang, C.; Duan, C.; Su, J.; Wang, L.; Atkinson, P.M. Multiattention network for semantic segmentation of fine-resolution remote sensing images. *IEEE Trans. Geosci. Remote Sens.* **2021**, *60*, 1–13. [CrossRef]
48. Wang, D.; Chen, X.; Jiang, M.; Du, S.; Xu, B.; Wang, J. ADS-Net: An Attention-Based deeply supervised network for remote sensing image change detection. *Int. J. Appl. Earth Obs. Geoinf.* **2021**, *101*, 102348.
49. Buskirk, T.D. Surveying the forests and sampling the trees: An overview of classification and regression trees and random forests with applications in survey research. *Surv. Pract.* **2018**, *11*, 1–13. [CrossRef]
50. Thung, K.H.; Wee, C.Y. A brief review on multi-task learning. *Multimed. Tools Appl.* **2018**, *77*, 29705–29725. [CrossRef]
51. Zhang, Z.; Sabuncu, M.R. Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels. *arXiv* **2018**, arXiv:1805.07836.
52. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282. [CrossRef]
53. Kumar, Y.J.N.; Spandana, V.; Vaishnavi, V.; Neha, K.; Devi, V. Supervised machine learning approach for crop yield prediction in agriculture sector. In Proceedings of the 2020 5th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 10–12 June 2020; IEEE: New York, NY, USA, 2020; pp. 736–741.
54. Verrelst, J.; Rivera, J.P.; Veroustraete, F.; Muñoz-Marí, J.; Clevers, J.G.; Camps-Valls, G.; Moreno, J. Experimental Sentinel-2 LAI estimation using parametric, non-parametric and physical retrieval methods—A comparison. *ISPRS J. Photogramm. Remote Sens.* **2015**, *108*, 260–272. [CrossRef]
55. Estornell, J.; Martí-Gavilá, J.M.; Sebastiá, M.T.; Mengual, J. Principal component analysis applied to remote sensing. *Model. Sci. Educ. Learn.* **2013**, *6*, 83–89. [CrossRef]
56. Fitzgerald, R.; Lees, B. Assessing the classification accuracy of multisource remote sensing data. *Remote Sens. Environ.* **1994**, *47*, 362–368. [CrossRef]
57. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.

58. Loshchilov, I.; Hutter, F. Sgdr: Stochastic gradient descent with warm restarts. *arXiv* **2016**, arXiv:1608.03983.
59. Hu, X.; Wang, X.; Zhong, Y.; Zhang, L. S3ANet: Spectral-spatial-scale attention network for end-to-end precise crop classification based on UAV-borne H2 imagery. *ISPRS J. Photogramm. Remote Sens.* **2022**, *183*, 147–163. [CrossRef]

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