A CNN-Transformer Hybrid Approach for Crop Classification Using Multitemporal Multisensor Images

Zhengtao Li, Guokun Chen, and Tianxu Zhang

***Abstract*—Multitemporal Earth observation capability plays an increasingly important role in crop monitoring. As the frequency of satellite acquisition of remote sensing images becomes higher, how to fully exploit the implicit phenological laws in dense multitemporal data is of increasing importance. In this article, we propose a CNN-transformer approach to perform the crop classification, in the model, we borrow the transformer architecture from the knowledge of NLP to dig into the pattern of multitemporal sequence. First, after unifying the spatial-spectral scale of each multiband data acquired from different sensors, we obtain the scale-consistent feature and position feature of multitemporal sequence. Second, with adopting multilayer encoder modules derived from the transformer, we mine deep correlation patterns of multitemporal sequence. Finally, the feed-forward layer and softmax layer serve as output layers of the model to predict crop categories. The proposed CNN-transformer approach is illustrated in a crop-rich agricultural region in central California, where 65 multitemporal profiles from multisensor Sentinel-2 A/B and Landsat-8 are obtained in 2018. Through multiband multiresolution fusion, sequence correlation extraction of multitemporal data and category feature extraction, the classification results show that the proposed method has a significant performance improvement compared with other traditional methods.**

***Index Terms*—Crop classification, multitemporal multisensor, self-attention, transformer.**

I. Introduction

FOOD security is the foundation of world economic security. Agriculture is not only the basic condition to ensure social development, but also an important economic area for promoting social development. In order to improve the healthy development of agriculture, it is increasingly important to monitor the types of crops on the agricultural land and to control the changes of crop types. In recent years, with more and more earth observation satellites being put into use, we can obtain increasingly time-intensive remote sensing (RS) images, and that can help us to improve the classification accuracy of crops [1], [2]. Especially, by combining Sentinel-2 A,B and Landsat-8 [3], we can obtain medium-resolution RS data with a revisit cycle as short as 3–5 days [4]. Benefited from dense temporal RS data, we can monitor land use more accurately [5], [6], such as: the crop classification, phenological changes, and land classification.

Different crops have phenological differences, and the classification accuracy of crops can be improved through distinguishing different temporal spectral rules of crops. Several studies have shown the importance of multitemporal information, Devadas *et al.* [7] present an object-based classification approach using support vector machine (SVM), which is superior to the pixel-based methods for classifying different crop types in summer and winter seasons with multitemporal Landsat data. Li *et al.* [8] use SVM and decision tree supervised classification methods on multitemporal HJ satellite images, and the crop classification results indicate that HJ-1 A/B satellite had the particular advantage in extracting vegetation information because its higher temporal resolution. In [9], Melgani proposed a spatial and spectral fuzzy fusion approach for classification, and obtained the temporal information by using transition probabilities. The accuracy of agricultural land cover mapping is basically positively correlated with the number of multitemporal images, Pax *et al.* [10] found that with the increase in the number of time-series images, the estimation accuracy of land area in agricultural areas in Egypt become higher. In order to obtain sufficient cloud-free remote sensing data in humid, tropical, or subtropical regions, Useya and Chen [11] fuse multitemporal Landsat 8, Landsat 7, and Sentinel-2 data, and obtain more accurate crop maps in Zimbabwe. However, in the critical stages of crop growth, the subtle differences in phenology often have important indications for crop production, such as: the flowering of canola or the tillering stage of transplanted rice. But the sparse multitemporal RS data does not indicate the subtle differences in crop phenology, therefore, it is important to use the dense multitemporal RS data to mine the subtle phenological differences. Up to now, how to use dense temporal features of RS data to distinguish the fine phenological differences of crops is still an important challenge.

As a powerful feature extraction framework, deep learning has achieved great success in a wide range of tasks. Compared with the traditional machine learning algorithms that require complex feature engineering, deep learning can automatically learn robust feature representation and adapt to different fields and applications more easily. In remote sensing image processing tasks, deep learning also shows great potential [12]. In Data Fusion Contest organized by the Image Analysis and Data

Manuscript received July 16, 2019; revised December 29, 2019; accepted January 21, 2020. Date of publication February 14, 2020; date of current version March 17, 2020. *(Corresponding author: Zhengtao Li.)*

Zhengtao Li and Tianxu Zhang are with the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan 430074, China (e-mail: 1495403@qq.com; txzhang@hust.edu.cn).

Guokun Chen is with Wireless Technology Laboratory, Ericsson Research, 16440 Kista, Sweden (e-mail: 279283265@qq.com).

Digital Object Identifier 10.1109/JSTARS.2020.2971763

Fusion Technical Committee of the IEEE Geoscience and Remote Sensing Society, a series of deep learning techniques [13], such as CNN and super-resolution, played a significant role and achieved obvious results. Zhang *et al.* [14] proposed a unified spatial–temporal–spectral framework based on CNN to complete the missing data in remote sensing image, and the proposed approach can solve the missing information reconstruction tasks include: dead lines in MODIS band 6, thick cloud removal, etc. In addition to multisensor images, Chen *et al.* [15] introduced deep learning architecture in hyperspectral classification problem. By using stacked autoencoders, the classification results achieved competitive performance than the traditional SVM methods.

The attention mechanism is a method, which human rapidly screen out high-value information from a large amount of information with utilizing attention. It is a survival mechanism formed in the long-term evolution of human beings. Visual attention mechanism can greatly improve the efficiency and accuracy of information processing. Attention mechanism imitate the internal processing of biological observation behavior, which align internal perception and external sensation to increase the observation fineness in some areas. Attention can be interpreted as a method, which allocates available computing power to the most informative part of the signal [16]–[19]. And attention mechanism has shown its utility in many tasks, including: sequence learning [20], [21], image captioning [22], [23], machine translation [24], [25], sentence summarization [26], machine comprehension [27], and document classification [28].

In this article, we borrow the transformer architecture from the field of NLP to model the crop phenological differences. The transformer architecture proposed by Google, which based on self-attention mechanism has shown an excellent expression ability to model the sequence correlation, and it has achieved the best results in many NLP tasks, such as machine translation [29], document classification, etc. Depending on the powerful sequence modeling capability of the self-attention module, we can distinguish the subtle but important phenological differences, such as in the rice transplanting period, the subtle phenological differences contained in the dense time-series data can help us to distinguish the tillering period, and to predict the number of tillers [30]; the flag leaf growth at the booting stage of cereal crops which has subtle phenological differences also can be used to predict the grain production [31]; and due to the differences in planting dates, cultivars and soil conditions, the timing of flowering and podding among fields of canola is usually different, the dense monitoring of phenology and environmental conditions can elevate disease and insect infestations risk in canola [32]. For these cases, compared with the traditional classification methods [7], the transformer architecture can obtain more precise phenological patterns.

The transformer architecture utilizes multihead self-attention modules to represent the sequence patterns. The multihead self-attention modules abandon the traditional recurrent sequence information modeling methods such as RNN, GRU, and LSTM, shorten the length of the paths by which information between different positions in sequence traverse in the network, and can directly obtain long-range dependencies for any combination of positions in sequence. Therefore, compared with the traditional RNN and LSTM structures, the multihead self-attention modules greatly improve the ability of sequence information correlation representation. We use this excellent structure to extract the characters of sequence information, express the correlation between time positions in sequence, and use phenological differences between crops to obtain more accurate results of crops classification. First, by using the ability of CNN to express spatial and spectral correlations, we unify the scale of spatial-spectral features for Sentinel-2 and Landsat-8 sample multiband images. Second, we apply the transformer architecture on temporal-spatial-spectral tensor to obtain its temporal correlation. Finally, appending a feed-forward layer and a softmax layer, we can get the crop types.

The rest of this article is organized as follows. Section II describes, the ground truth of crop types in the study region and the data preprocessing. Section III illustrates the classification method based on transformer architecture. Section IV describes the study region and the results of experiment. Finally, further discussion and conclusion of this article are given in Section V.

II. Related Work

A. Ground Truth of Crop Types

The cropland data layer (CDL) product, produced by the United States Department of Agriculture (USDA) National Agricultural Statistics Service, is a georeferenced, rasterformatted, crop types cover classification map with a resolution of 30 m across the country. By using the CDL data, a series of cropland changes can be well assessed, such as crop intensity, rotation, epidemiology, watershed, environmental risk, and disaster response. For instance, Maxwell *et al.* [33] study the relationship between cancer and agricultural chemical exposure. Shan *et al.* [34] estimate and map the flood damage with RS images and CDL product.

During the process of CDL production, the project utilizes USDA's Farm Service Agency (FSA) Common Land Unit (CLU) data, National Land Cover Database data, remote sensing satellite data, and some other ancillary data as the input data. Among the input data, FSA CLU data set is a comprehensive agricultural ground truth data of crop types, which is updated multiple times in the growing season. Benefit from the increased available ground truth data of crop types, the CDL program has greatly increased efficiency and accuracy. By comparing the CDL crop types with independent validation data extracted from FSA CLU ground truth data, the accuracy of CDL agricultural crop types maps can be obtained. Since a large amount of agricultural survey data are used in the production of CDL products, the classification accuracy of the major crop categories is high, usually 85% to 95%. In this article, we select pure single crop regions with large areas as our crop types ground truth.

B. Preprocessing

Before the crop classification, several preprocessing steps for satellite images are implemented. For the Sentinel-2 Level-1C scenes downloaded from USGS, we first use Sen2cor atmospheric correction module to acquire the surface reflectance data derived from Level-1C TOA reflectance values, and then use the Fmask algorithm developed and improved by Zhu *et al.*

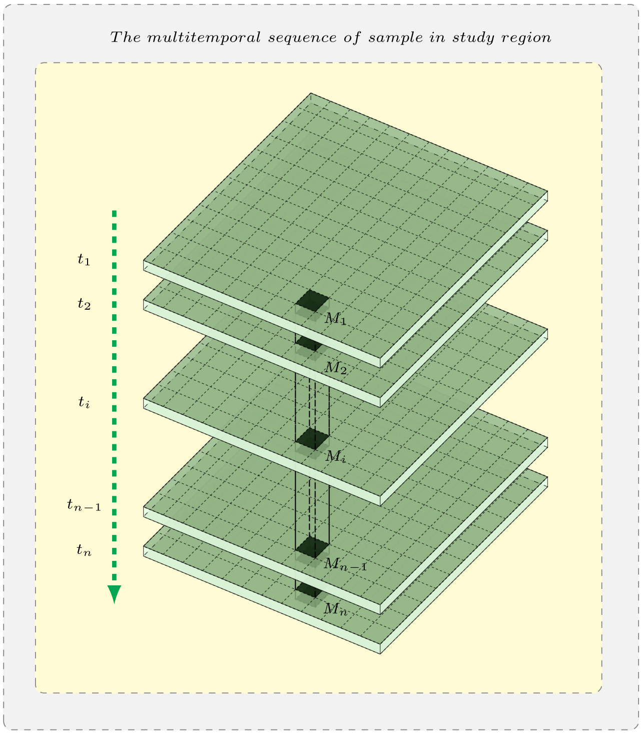


Fig. 1. Multitemporal multisensor sample in a study region. In multitemporal sample sequence, M_i indicates multiband images from Sentinel-2 or Landsat-8.

[35] to perform clouds and cloud shadows removal. Similar to Sentinel-2, for the downloaded Landsat-8 Level1 DN values, Landsat-8 images are first converted from DN values to spectral radiance, and then the spectral radiance is converted to the surface reflectance by atmospheric correction tool FLAASH which is available in ENVI. Clouds and cloud shadows removal are implemented with using the same Fmask algorithm as Sentinel-2. Next, the cloud-free scenes are mosaicked in the study region. Finally, for the surface reflectance data with missing cloudy holes, we use the self-organizing Kohonen maps (SOMs) to reconstruct the missing gap data of time-series images. The reconstruction of holes is performed for each spectral band: 1) For the time-series band images from Sentinel-2 or Landsat-8, some pixels of temporal profile contain the holes derived from the removal of clouds, and we choose the pixels of temporal profile without temporal gaps to train the SOM. 2) By SOM learning, the pixels of temporal profile without gaps can be projected into the subspace of map vectors, that is, the weight vectors of SOM indicate the temporal profiles of training sample. 3) Finally, the relevant components of the neuron-winner weight vector in SOM are used to reconstruct the missing data in the time series [36].

III. Proposed Methodology

In this article, the multitemporal multisensor images acquired by Sentinel-2 and Landsat-8 can be downloaded from the USGS EROS Center. As shown in Fig. 1, the sample in the study region can be expressed as $\mathcal{M} = [M_1, M_2, \ldots M_i \ldots, M_n]$, where n is the temporal length of multitemporal sample and M_i represents the date-corresponding multiband images.

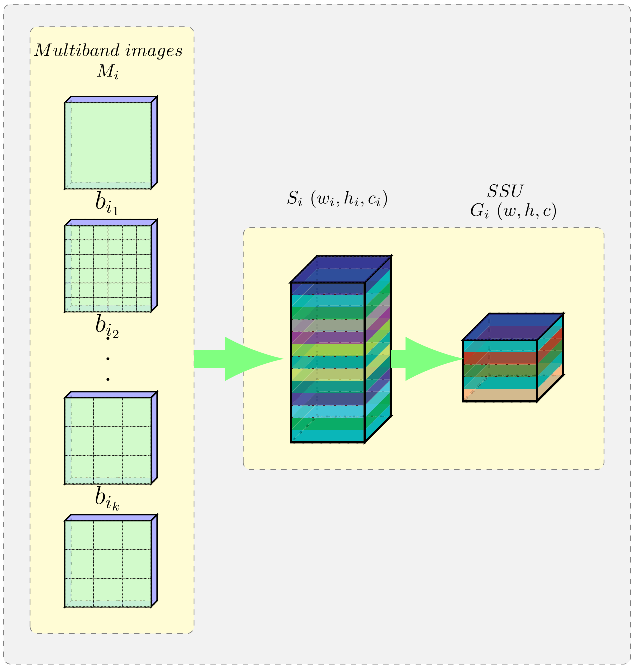


Fig. 2. Spatial and spectral fusion scheme for multiband images. The left part shows multiband images M_i, the middle part indicates the multiband features with unified spatial dimensions(w_i, h_i), and the right part expresses the spatial-spectral-unified SSU feature.

In the sample $\mathcal{M}$, the multiband images set M_i from Sentinel-2 or Landsat-8 can be expressed as:

$$M_i = \{b_{i_k} : i_k \in \text{spectral bands}\} \tag{1}$$

where i indicates the acquiring time in multitemporal sequence, k denotes the sensor band, and i_k is the sensor band on the acquiring time i, and b_{i_k} expresses the band image of i_k with dimensions $w_{i_k} \times h_{i_k}$.

In general, sensors have different numbers of multispectral bands, Sentinel-2 has 13 bands with a spatial resolution of 10, 20, and 60 m, and Landsat-8 has 11 bands with spatial resolution of 15 and 30 m. In order to obtain the normalized feature maps from different sensors, first, we normalize the spatial resolution for each band image in M_i to obtain the normalized width and height, and then normalize the channel dimension of feature maps.

A. *Multisensor Spatial-Spectral Scale Unification*

Corresponding to the multitemporal multisensor sample $\mathcal{M} = [M_1, M_2, \ldots M_i \ldots, M_n]$, we perform spatial-spectral fusion to get the spatial-spectral-unification (SSU) features $\mathcal{G} = [G_1, G_2 \ldots, G_i, \ldots G_n]$ for each images set M_i, where G_i has the unified feature dimensions (w, h, c), and w, h, c indicate the width, height, and channels number of SSU features.

As shown in Fig. 2, the spatial-spectral unification is divided into two steps, spatial unification and spectral unification.

1) Spatial Scale Unification for Each Band: In order to unify the spatial resolution of multiband images, we convolve each image b_{i_k} from the images set M_i and then concatenate all convolved features with unified scale:

$$\widetilde{b}_{i_k} = Conv(b_{i_k}) b_{i_k} \in M_i$$
$$S_i = Concat(\{\widetilde{b}_{i_k} : i_k \in 1, 2, \ldots, c_i\}). \tag{2}$$

In the formula, $Conv$ represents transposed convolution transformation on b_{i_k} to get scale-unified features, $Concat$ expresses concatenating all features data with unified scale, b_{i_k} is the image of band i_k in M_i, $\widetilde{b}_{i_k}$ is the feature of band i_k with unified spatial scale (w_i, h_i), and S_i is a 3-D tensor of dimensions (w_i, h_i, c_i), where w_i and h_i correspond to the width and height of the unified features and c_i to bands number of sensor on time i.

2) Spectral Scale Unification: Different sensors usually have different band numbers, in order to unify the multiband features from different sensors, we convolve the 3-D tensor S_i by utilizing the consistent convolution kernel numbers for the whole time sequence in the sample

$$G_i = Conv(S_i) \tag{3}$$

where $Conv$ indicates the convolution transformation and G_i is a SSU tensor with dimensions (w, h, c), in which w, h, c correspond to the width, height, and channel of spatial-spectral-unification features.

Next, the multitemporal transformer module will be applied on the obtained scale-unified multitemporal SSU features.

B. CNN-Transformer Architecture for Classification

In the network, the transformer architecture, which has excellent expressive ability of sequence information is introduced to model the input multitemporal features. The overall block diagram for crop classification is shown in Fig. 3:

The model architecture consists of several modules: the SSU features extraction module, position feature module, multilayer transformer $encoder$ module, feed forward module, and softmax output layer module. In this section, we will introduce these modules as below.

1) SSU Features Extraction and Position Feature Embedding: For sequence information input, the transformer adopts 1-D vector as the input for each sequence position, so the multitemporal sequence of SSU features needs to be first converted to a 1-D vector sequence. The multitemporal SSU features have the unified shape of (w, h, c), after flattening SSU features on each sequence position, we can obtain the feature sequence $W_E = [e_1, \ldots, e_n]$, and the length of 1-D feature is $w \times h \times c$.

The position feature embeddings of sequence indicate the relative or absolute position information of features in the sequence, and position features can be described by "positional encodings." The positional encodings have the same vector length d_{model} as the feature e_i of $w \times h \times c$. Here, we cite the functions in [37] to define positional feature embedding:

$$\text{PE}_{(p,2i)} = \sin(p/10\,000^{2i/d_{\text{model}}})$$
$$\text{PE}_{(p,2i+1)} = \cos(p/10\,000^{2i/d_{\text{model}}}) \tag{4}$$

where p is the position and i express the dimension of position feature.

2) Multilayer Transformer Encoder Module: The transformer architecture [37] proposed by Google, which is different from the previous RNN-like model for modeling sequence information, has shown great vitality in the field of NLP. In transformer architecture, the self-attention is a sequential encoding mechanism similar to RNN and LSTM, and it improves the expression ability of the relationship between word sequences to get better performance on various NLP tasks. In addition to the excellent expression ability of the relationship between sequence information, the self-attention is much better than the RNN-like model in parallel ability because it inputs the entire sequence at a time for training, which can greatly improve the training speed of a sequence model.

For the sequence modeling and sequence classification, there are different task paradigms. For sequence modeling, a typical task is the language modeling. For a sentence sequence $\mathcal{U} = \{u_1, \ldots, u_n\}$, the standard language modeling objective can be used to maximize the likelihood: $L(\mathcal{U}) = \sum_i \log P(u_i | u_{i-k}, \ldots, u_{i-1}; \Theta)$. When using transformer architecture to build the language model, the parallel training of sentence sequences can be achieved through forward masking mechanism which acts on the encoder and decoder. The mask mechanism shield the future word embedding, so that only leftward information of the sequence can be seen, while rightward information is blind. In contrast to the language model task, in the classification task, the output is the classification label rather than the sequence. And the label is visible to the entire sequence information, so the classification model can utilize all of the sequence information. Therefore, in the sequence classification task, we only take advantage of the encoder module in transformer without mask mechanism.

In the architecture, we use the multilayer transformer encoder module for sequence model, and it is a variant of transformer. As shown in Fig. 4, there are two sublayers in encoder module: multihead self-attention is the first part followed by the second positionwise fully connected feed-forward network. In addition, there are residual connections and layer normalizations in each encoder.

This module takes SSU features and position embeddings as input, so h_0 can be formed as

$$h_0 = W_E + W_P \tag{5}$$

where SSU feature sequence W_E is the time-series features $[e_1, \ldots, e_n]$, which includes the entire dates and position feature sequence W_P represents the position embeddings for each time position $[\text{PE}_1, \ldots, \text{PE}_n]$. The multilayer encoder module can be described as follows:

$$h_l = \text{transformer\_encoder}(h_{l-1}) \forall i \in [1, n] \tag{6}$$

where n is the number of layers.

3) Output Layer for Supervised Classification: In order to get the crop classification, we add feed-forward and softmax layers to the network. For our classification task, we get a labeled dataset $\mathcal{C}$, where each instance consists of time series features, $\mathcal{G} = [G_1, \ldots, G_n]$, along with a label y. We pass the input sequence through feature extraction layer and multilayer transformer encoder module to get the last encoder's activation

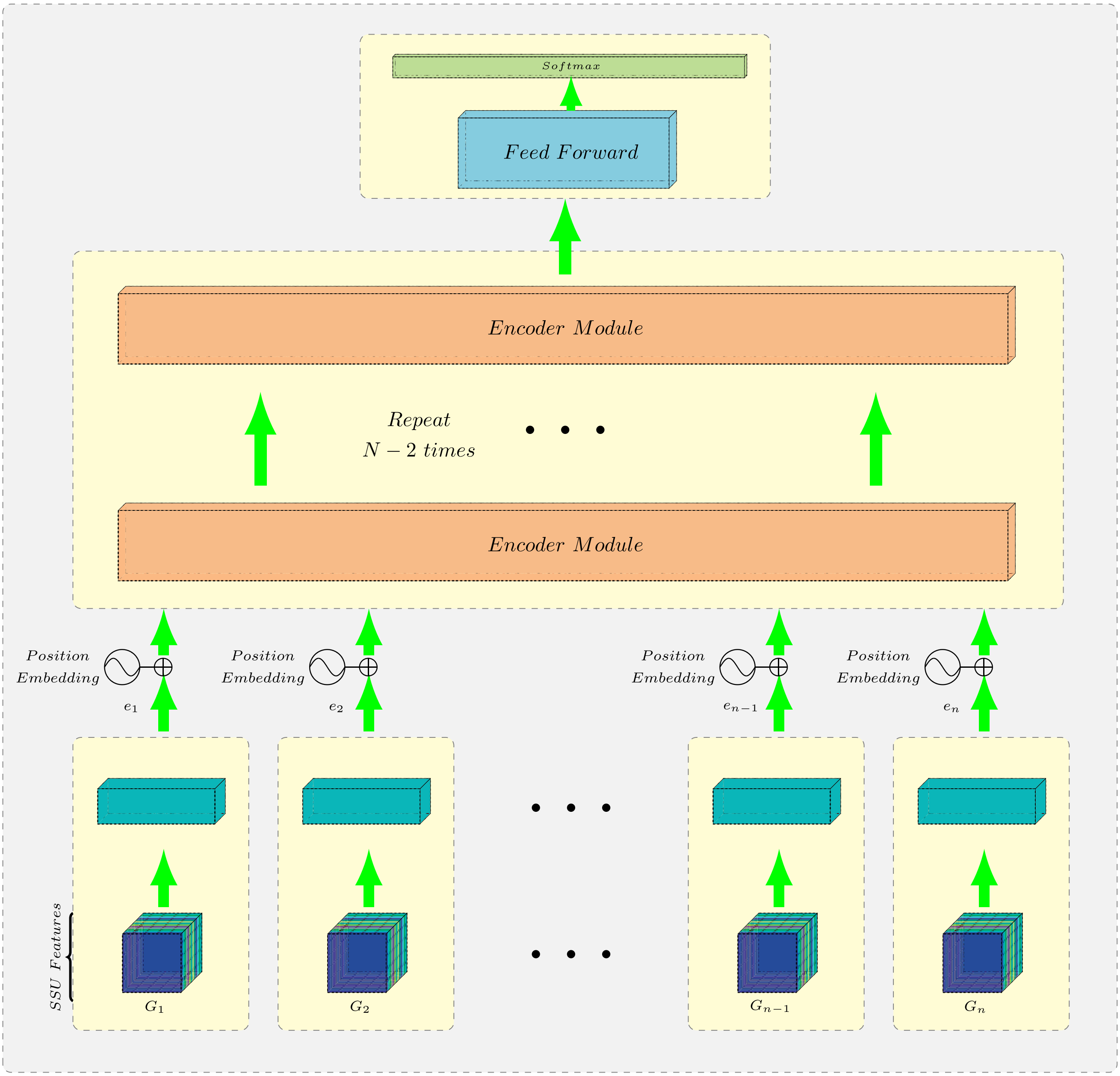


Fig. 3. Classification scheme for multitemporal multisensor images based on CNN-transformer sequence correlation extraction. The lower part of the figure indicates multitemporal feature extraction and position feature embeddings; and the upper part expresses the sequence feature extraction with multilayer encoder modules.

h_l, then add feed-forward layer with parameters W_y and softmax layer to predict y

$$P(y|G_1,\ldots,G_n) = \mathrm{softmax}(h_l W_y). \tag{7}$$

By that formula, we get the objective to maximize

$$L(\mathcal{C}) = \sum_{(\mathcal{G},y)} \log P(y|G_1,\ldots,G_n). \tag{8}$$

In the above formula, the conditional probability P is modeled using neural networks. The model can be trained using stochastic gradient descent.

IV. Experimentation

A. Study Area and Dataset Description

In our study, the northern Sacramento Valley in California was selected as the study area, which has a typical Mediterranean climate. The Mediterranean climate is sunny during the crop growing season from March to September, with less cloud cover, which is conducive to obtain more useful remote sensing images. In the area, a region of approximately 100 km × 100 km was selected as the study region, as shown in Fig. 5. This region has a diverse crop matrix that includes tomatoes, corn, rice, grapes, alfalfa, sunflower, clover, almonds, and walnuts as well as other specialty crops (e.g., watermelons, carrots, onions, peas), as

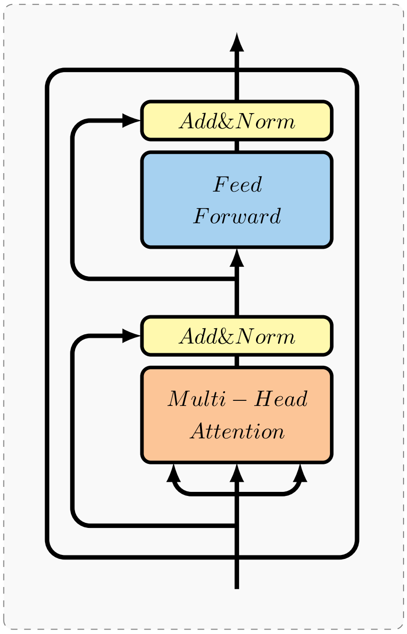


Fig. 4. Encoder module used in the proposed model.

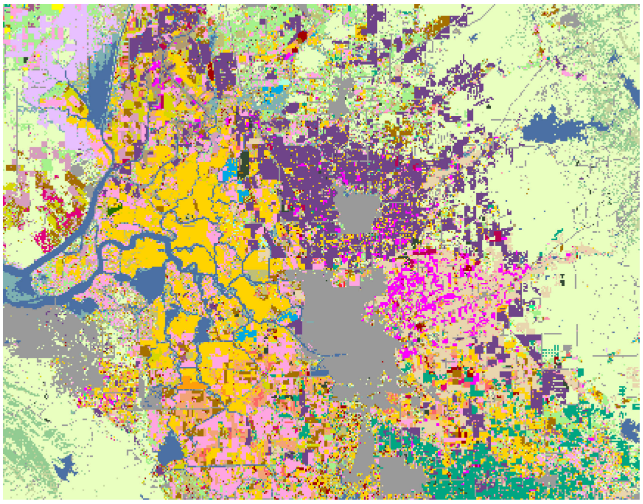


Fig. 5. Crop types product for Northern Sacramento Valley in California. In the region, there are a variety of crops, such as cotton, rice, corn, soy, etc.

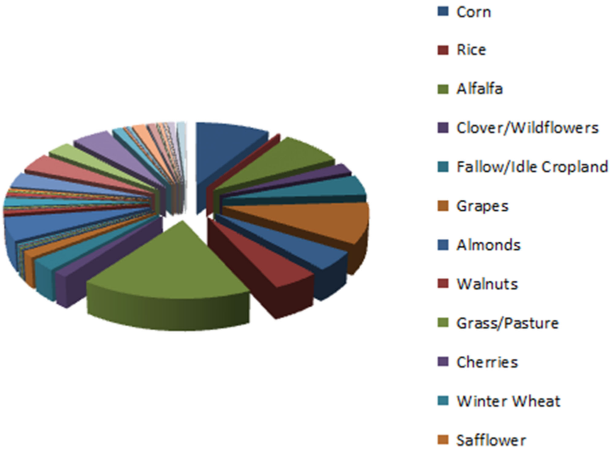


Fig. 6. Pie chart depicts the share of various crops in the study area, such as corn, rice, alfalfa, grapes, etc.

TABLE I
CROP CLASSES AND THE NUMBER OF SAMPLES IN THE DATASET

Class		*No. of Samples*	
Code	*Number*	*Train*	*Test*
1	*Corn*	32	3413
2	*Rice*	22	2262
3	*Alfalfa*	37	3775
4	*Clover/Wildflowers*	83	7312
5	*Grapes*	46	4373
6	*Almonds*	37	3483
7	*Walnuts*	38	3679
8	*Grass/Pasture*	42	5232
9	*Cherries*	26	2878
10	*Safflower*	36	3153
Total		399	39560

shown in Fig. 6. The ground truth of the crop types in the region can be obtained from CDL product.

In order to preserve the original spatial information of the multiband images acquired by the satellites, spatial resampling of multiband images with a different spatial resolutions should be avoided. Therefore, the pixel size of sample patch is set to 60 m × 60 m, and within a pixel patch, the spatial information of each band is original information acquired from satellites. Within the study area, the pure crop plantations with a large area are selected as the sample collecting area of interest for major crops. With 60 m × 60 m as the size of the sample pixels, we select approximately equivalent numbers of samples for the major crop types to form the research dataset, including corn, rice, alfalfa, clover, fallow, grapes, almonds, walnuts, pasture, cherries, winter wheat, safflower. In this experiment, the dataset contains a total of ten types of 39 560 samples, out of which 1% are used for training and the remaining 99% are used for verification, as shown in Table I.

For the study region, all the multiband remote sensing images of Sentinel-2 and Landsat-8 were downloaded from USGS for 2018 year. Then, the images with thick clouds are removed, the cloudless and less cloudy images are retained. Through several preprocessing steps to mask the cloud in images, remove the areas covered by the cloud and fill the corresponding holes, we can obtain the time-series images with a length of 65. As shown in Fig. 7, due to the fine weather, during the crop growing season from March to September, the temporal data in growth period is dense. Furthermore, abundant temporal information can help us to dig out the small differences in crop phenology.

B. Evaluation Criteria and Classification Methods

In order to evaluate the effects of different classification models, a lot of evaluation criteria have been developed. The confusion matrix can clearly express the number of correctly classified samples for each category and the detail for each misclassified category. However, from the confusion matrix, we cannot evaluate the performance of various classification models immediately. Therefore, a variety of classification accuracy indicators are derived from the confusion matrix, among which

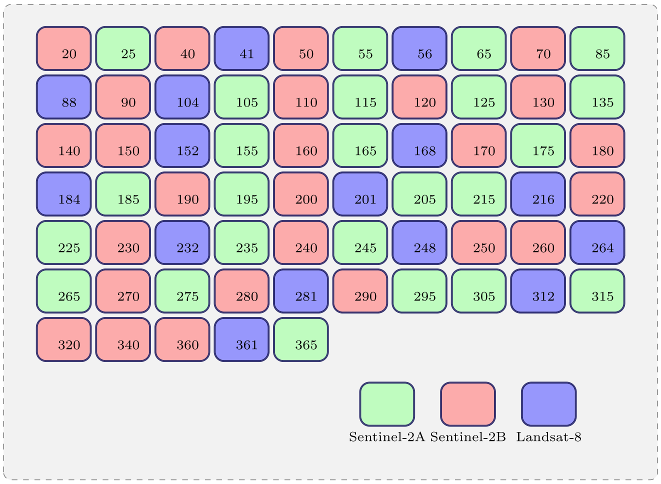


Fig. 7. Multitemporal sequence from different satellites: Sentinel-2 A, Sentinel-2B, and Landsat-8. The numbers in the icons represent the acquiring DOY(day of year), and green, orange, and violet represent the different sources: Sentinel-2 A, Sentinel-2B, and Landsat-8, respectively.

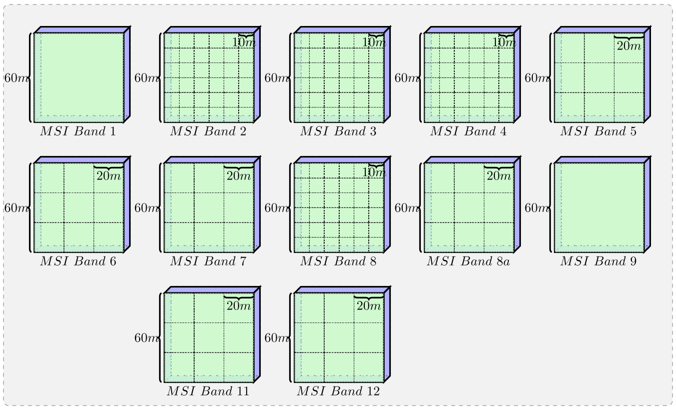


Fig. 8. Multiband MSI images from Sentinel-2 in a sample.

the overall accuracy (OA), average accuracy (AA), and Kappa coefficient are the most widely used.

1) Overall Accuracy (OA): The overall accuracy is the sum of the correctly classified values (values on the right diagonal) in the confusion matrix, divided by the total number of samples, which represents the ratio of the correctly predicted samples in all samples.

2) Average Accuracy (AA): The average classification accuracy is the average value of all classification accuracies.

3) Kappa Coefficient: Different from the overall accuracy, the Kappa coefficient is calculated from all the information of the confusion matrix. It is considered as the consistency measure between the ground-truth map and the final classification map, which can more accurately express the entire classification accuracy.

In order to evaluate the effect of the CNN-transformer model proposed in this article, we compare it with traditional vector-based classification methods such as support vector machine and random forest. Then, the performance of a classical deep network model such as multitemporal CNN and CNN-LSTM are further compared with the proposed CNN-transformer model.

1) *RF-200:* In the experiment, the number of decision trees in the random forest is set to 200.

2) *SVM-RBF:* We use the LIBSVM package to carry out the support vector machine classifier with RBF kernel, and take into account fivefold cross-validation to optimize the hyperplane parameters.

3) *CNN-transformer crop classifier:* The proposed CNN-transformer classifier combines CNN and transformer architecture to mine the category pattern of crops, where CNN is used to extract the spatial-spectral features of each acquisition date, and the transformer module with powerful sequence pattern extraction capability is used to express the correlation of time series.

4) *Multitemporal CNN classifier:* For the proposed CNN-transformer classifier, we replace the multilayer transformer encoder modules with regular CNN layers to obtain the Multi-Temporal CNN (MT-CNN) classifier, which is compared with CNN-Transformer classifier to represent the difference in the ability to mine sequence information.

5) *CNN-LSTM crop classifier:* The CNN-LSTM classifier combines CNN and LSTM to model the orderly and continuing multitemporal multisensor features of the sample, different from the multi-layer transformer encoder modules, this classifier utilizes the RNN-like networks to mine the sequence pattern.

Experiments are organized into two parts. The first part first discusses the spatial-spectral-unification parameters for the two satellites on each acquisition date in the proposed method, and then discusses the input tensor shape, self-attention layers, number of encoder layers, and output layer configuration in the transformer architecture, finally describes the compared multi-temporal CNN classifier which replaces the transformer architecture with CNN architecture. In the second part, the effectiveness of a CNN-transformer classifier that is based on transformer architecture is compared with the traditional vector-based models, such as random forest, SVM, regular CNN classifier which replaces the transformer encoder layers with regular CNN layers, and CNN-LSTM classifier.

C. Hyperparameter Analysis of the Proposed Network

In experiments, the pixel-based samples set is collected with size of 60 × 60 m.

1) Spatial-Spectral Unification for Sentinel-2: There are 13 multispectral images acquired by MSI on board the Sentinel-2 in VNIR and SWIR bands, and the cirrus image with band of 10 mainly indicates the distribution of cirrus clouds rather than ground reflections. Therefore, after discarding cirrus band, the remaining multiband images are as follows:

1) 4 bands at 10 m: 490 nm (B2), 560 nm (B3), 665 nm (B4), 842 nm (B8).
2) 6 bands at 20 m: 705 nm (B5), 740 nm (B6), 783 nm (B7), 865 nm (B8a), 1610 nm (B11), 2190 nm (B12).
3) 2 bands at 60 m: 443 nm (B1), 945 nm (B9).

As shown in Fig. 8, in a sample region of 60 × 60 m, there are three image resolutions: image size 6×6 for 10 m bands, image size 3×3 for 20 m bands, and image size 1×1 for 60 m bands.

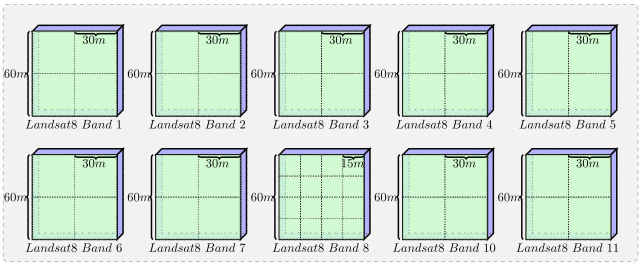


Fig. 9. Multiband OLI and TIRS images from Landsat-8 in a sample.

In order to obtain the spatial-unified features for multiband images, we use different transposed convolution kernel parameters for each different input image:

1) 10 m resolution, unconvolving a 1×1 kernel over a 6×6 input with unitary stride and no padding $(i = 6, k = 1, s = 1$ and $p = 0)$, output is a 6×6.
2) 20 m resolution, unconvolving a 4×4 kernel over a 3×3 input with unitary stride and no padding $(i = 3, k = 4, s = 1$ and $p = 0)$, output is a 6×6.
3) 60 m resolution, unconvolving a 6×6 kernel over a 1×1 input with unitary stride and no padding $(i = 1, k = 6, s = 1$ and $p = 0)$, output is a 6×6.

Finally, the multisensor spatial-spectral-unification feature G_i with unified dimensions (6, 6, 5) is obtained with the same number 5 of convolution kernels for multisensor.

2) Spatial-Spectral Unification for Landsat-8: There are 11 Landsat-8 spectral bands on the sample, of which cirrus detection image of band 9 shows the distribution of cirrus clouds rather than ground information. Therefore, excluding the cirrus band, there are ten multispectral bands with two spatial resolutions:

1) 9 multispectral bands at 30 m.
2) 1 panchromatic band at 15 m.

As shown in Fig. 9, in a sample of 60 m × 60 m, there are two image sizes for Landsat-8 multiband data: image size 4×4 for 15 m band and 2×2 for 30 m bands. With producing the same spatial output of 6×6, we have two transposed convolution kernel parameters for Landsat-8:

1) 15 m resolution, unconvolving a 3×3 kernel over a 4×4 input with unitary stride and no padding $(i = 4, k = 3, s = 1$ and $p = 0)$, output is a 6×6.
2) 30 m resolution, unconvolving a 5×5 kernel over a 2×2 input with unitary stride and no padding $(i = 2, k = 5, s = 1$ and $p = 0)$, output is a 6×6.

By concatenating the features of the same spatial scale, we can obtain the dimension (6, 6, 10), and then perform convolution with five convolution kernels, we can get the SSU feature G_i with dimension (6, 6, 5).

3) Parameter Analysis of the CNN-Transformer Network: As shown in Fig. 3, after the above spatial-spectral scale unification, 65 multitemporal spatial-spectral group features with shape of [6, 6, 5] can be obtained. Then, in order to obtain the features for transformer $encoder$ inputs, we flatten the spatial-spectral group features to generate 1-D feature with shape [180], meanwhile, the dimension of position embeddings is set to [180], which is the same as $encoder$ inputs. In the model, the number of multilayer encoder modules is experimentally set to 4, after sequence encoding for multitemporal features, the parameters of feed forward layer which consists of two fully connected layers are set to 100 and 40.

In order to evaluate the effects of the models fairly, compared with CNN-transformer network, multitemporal CNN classifier and CNN-LSTM only replace the multilayer transformer encoder modules with regular convolution layers and LSTM layers, and the other parts of the model, such as spatial-spectral unification layer and categories output layer, have the same network structure.

For the proposed model, the network with transformer architecture is trained with Adadelta algorithm, and in the experiment, we use a low learning rate of 0.01 to train the network. In the model, all the convolutional layers are appended by a batch normalization layer in which all the weight matrices and bias vectors of the convolutional layer in the model are uniformly initialized by xavier, and the convolutional weights and BN parameters are updated during training. In the experiments, the training epoch number of CNN-transformer model and compared multitemporal CNN model is set to 1000, which ensures the convergence of models.

It should be noted that in the proposed network, we utilize original spatial information of bands with different spatial resolutions, while traditional vector-based classification methods, such as random forests and SVM require a uniform 60-m spatial resolution band images that are resampled from bands data with different spatial resolutions.

D. Experimental Results

The classification confusion matrix and accuracy assessment for the multitemporal crop dataset are shown in Fig. 10 and Table II. Due to the SVM with RBF kernel handle nonlinear data more efficiently than the random forest, the classification results show that the SVM model outperforms the random forest model. Furthermore, it can be seen that the proposed CNN-Transformer model yields more accurate results than the other models. Specifically, as shown in Table II, compared with RF, CNN-transformer model increases the accuracy significantly by 6.12% of OA, 5.61% of AA, and 0.0692 of the Kappa coefficient, respectively. Compared with SVM-RBF, multitemporal CNN and CNN-LSTM, the increments of OA, AA, Kappa coefficient obtained by CNN-transformer model are 4.86%, 5.12%, 0.0548, and 1.68%, 2%, 0.0189, and 2.28%, 2.66%, 0.0257, respectively.

According to the analysis of classification accuracies, RF and SVM-RBF are not efficient in distinguishing between similar crops such as corn and rice, walnuts and cherries because of their similar phenological cycles. Compared with RF and SVM-RBF, deep-learning based multitemporal CNN model which has powerful feature expression ability can mine the phenological features of similar crops to improve the crop discrimination. Compared with RF and SVM-RBF, the misclassification rate of multitemporal CNN and CNN-LSTM model for walnuts and cherries is significantly reduced. CNN-transformer model proposed in this article is better than multitemporal CNN model in expressing the sequence information and distinguishing the

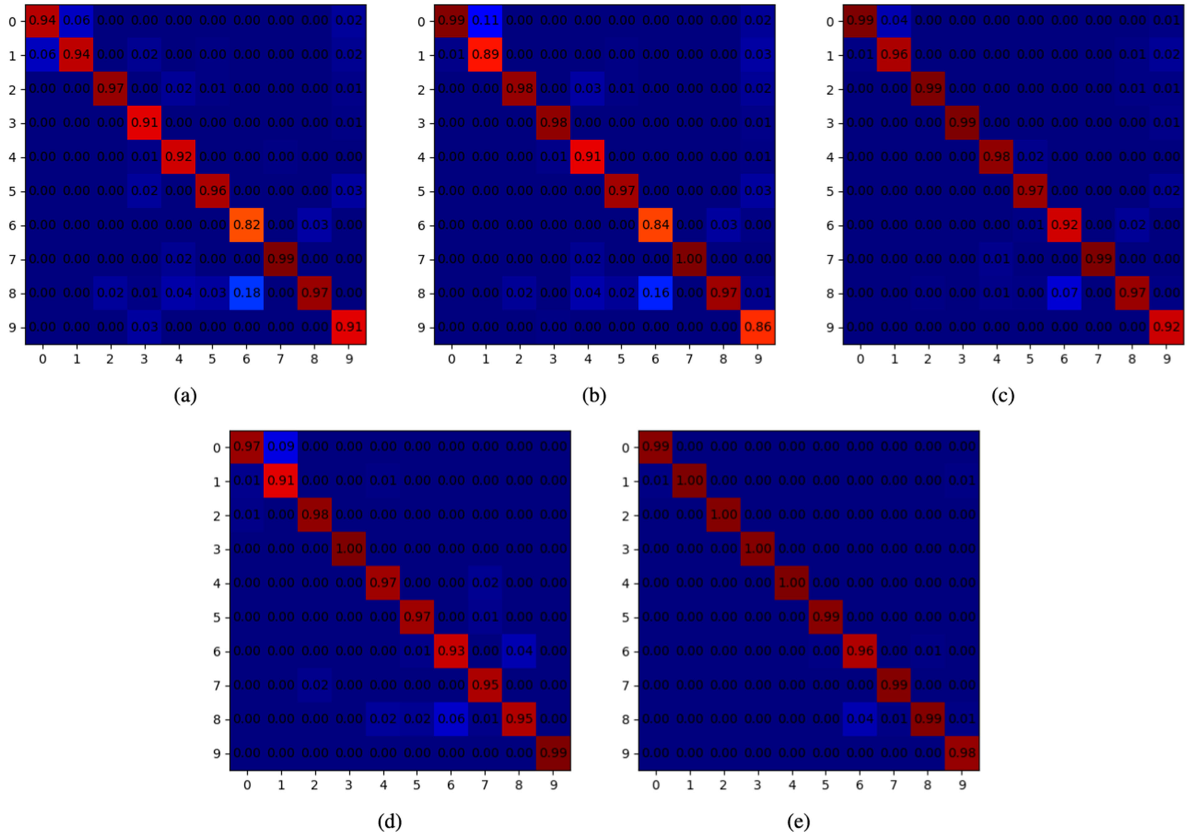


Fig. 10. Confusion matrix of different methods for the study dataset. (a) RF-200. (b) SVM-RBF. (c) Multitemporal CNN classifier. (d) CNN-LSTM classifier. (e) CNN-Transformer classifier.

TABLE II
CLASSIFICATION ACCURACIES OF DIFFERENT TECHNIQUES IN PERCENTAGES FOR TEST SAMPLES

Class No.	*Class Name*	*RF-200*	*SVM-RBF*	*Multi-Temporal CNN*	*CNN-LSTM*	*CNN-Transformer*
1	*Corn*	94.43	99.20	**99.27**	97.10	99.20
2	*Rice*	93.83	88.94	95.52	91.08	**100.00**
3	*Alfalfa*	97.32	97.60	99.28	97.89	**99.57**
4	*Clover/Wildflowers*	90.62	98.03	99.33	**99.71**	99.54
5	*Grapes*	92.24	90.79	97.61	96.68	**99.84**
6	*Almonds*	95.51	96.56	97.11	96.72	**99.02**
7	*Walnuts*	82.31	83.73	92.38	93.19	**95.55**
8	*Grass/Pasture*	99.42	**99.90**	99.44	95.49	99.40
9	*Cherries*	96.76	97.12	96.78	95.36	**99.24**
10	*Safflower*	90.67	86.14	92.46	**99.33**	97.88
OA	-	92.85	94.11	97.29	96.69	**98.97**
AA	-	93.31	93.80	96.92	96.26	**98.92**
Kappa	-	0.9192	0.9336	0.9695	0.9627	**0.9884**

The best accuracy in each row is shown in bold.

subtle differences of the sequence, compared with multitemporal CNN, the misclassification rate of corn and rice is reduced by 4%, and walnuts and cherries by 3%.

In the experiments, we train the proposed CNN-transformer model on a machine with four NVIDIA 1080ti GPUs and TensorFlow library. The training times of different methods are shown in Table III. It is expected that the training time of deep neural network model is longer than the traditional methods, however, as shown in Table IV, the deep neural networks have an advantage in testing time.

TABLE III
STATISTICS OF TRAINING TIME (MIN)

Methods	*RF-200*	*SVM-RBF*	*MT-CNN*	*CNN-LSTM*	*Ours*
Time	0.1	0.5	9.7	30.7	17.4

TABLE IV
STATISTICS OF TESTING EFFICIENCY (PIXELS/S)

Methods	*RF-200*	*SVM-RBF*	*MT-CNN*	*CNN-LSTM*	*Ours*
Efficiency	221	257	573	421	497

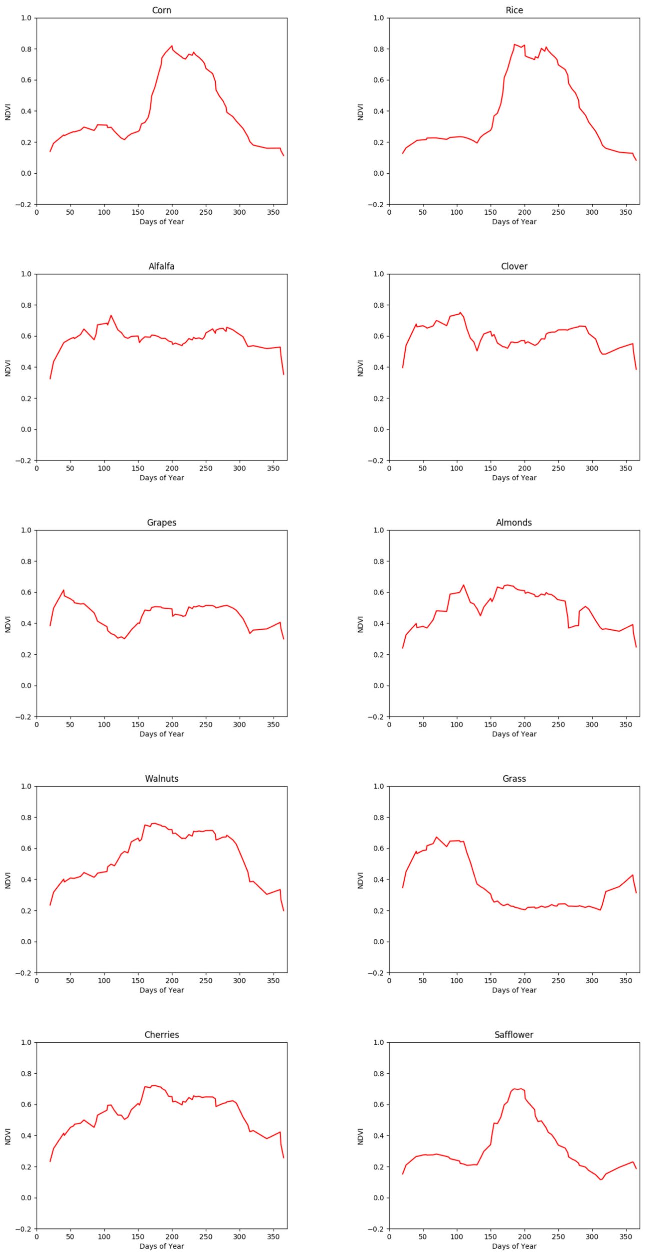
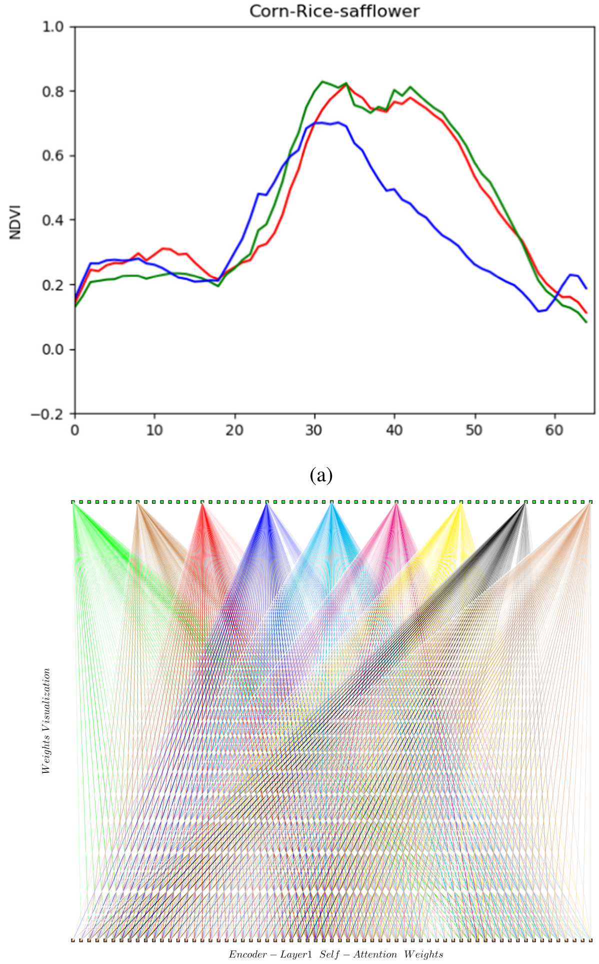


Fig. 11. Multitemporal NDVI profiles across one year for various crops: corn, rice, alfalfa, clover, grapes, almonds, walnuts, grass, cherries, the safflower.



(a)

(b)

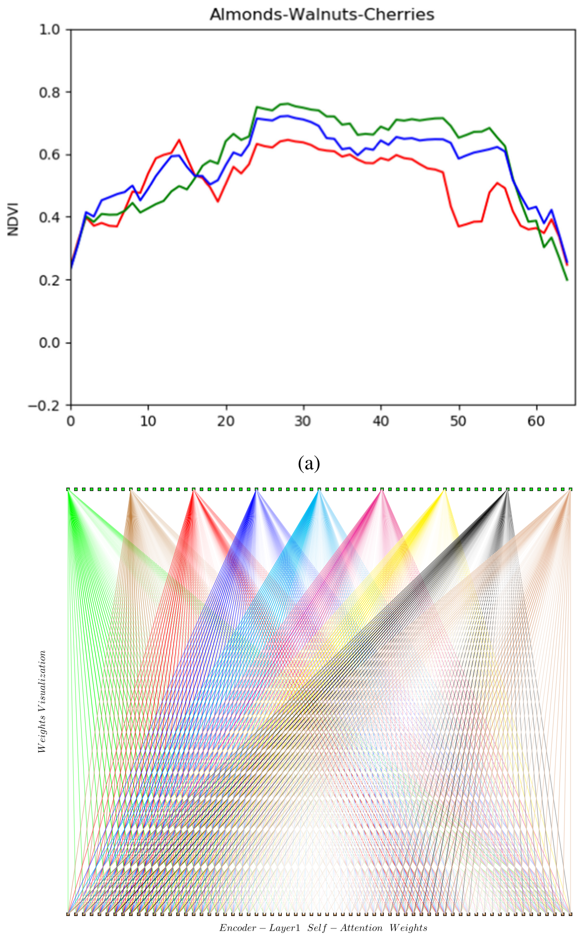
Fig. 12. (a) Comparison of similar phenological cycles: the NDVI profiles of corn, rice, and safflower. (b) Self-attention weights of the first encoder layer.

V. Discussion and Conclusion

Multitemporal profiles of NDVI represent the growth and evolution of crops, and different profiles correspond to different crop types and phenology. The NVDI profiles of a certain category in sample set are roughly consistent with same phenological cycles, so, to compute the multitemporal NDVI profile of each sample and further average the sample curves for each category, we can obtain the average temporal NVDI profiles for each crop type, as shown in Fig. 11. Among these crops, corn, rice, and safflower are summer or autumn crops, the NDVI profiles of these crops have obvious growth and development stages in spring, and a drying stage in autumn. However, almonds, walnuts, and cherries, these deciduous or semievergreen trees have no obvious growth period, but the NVDI profiles also reflect the green condition of these trees. In addition, alfalfa and clover, these herbaceous plants have no sudden changes in the multitemporal NDVI profiles throughout the whole year. In summary, all the differences of crops can be described by the profiles of Fig. 11.

To further analyze crops with similar phenological cycles, as shown in Figs. 12 and 13, we perform a comparison of corn, rice, and safflower, as well as the comparison of almonds, walnuts, and cherries.

Fig. 12(a) shows the phenological differences of corn, rice, and safflower. With using these crops with similar phenology to train the CNN-transformer, the model can get the differences between these crop phenology. For a test crop sample, Fig. 12(b) indicates the self-attention weights of the first encoder layer. The lower sequence in the figure (b) represents the multitemporal input of the encoder layer1, and the upper sequence indicates the self-attention encoding output. In order to further analyze the self-attention weights between the input and output sequences, we visualize part of the self-attention weights of the output sequence. Self-attention weights between output sequences can



(a)

(b)

Fig. 13. (a) Comparison of similar phenological cycles: the NDVI profiles of almonds, walnuts, and cherries. (b) Self-attention weights of first encoder layer.

be expressed in different colors, the dark colors represent larger weights, while light colors mean smaller weights. From the weight visualization in figure (b), it can be seen that the weights of the self-attention have the larger values in the temporal periods: 0–16, 24–40, and 50–60, which reflects the most obvious phenological differences between crops in these temporal periods. That is, the weights reflect the attention to phenological differences, which can improve the discrimination accuracy of similar crops.

Fig. 13(a) represents the phenological differences of almonds, walnuts, and cherries. Similar to Fig. 12(b), Fig. 13(b) also expresses self-attention weights of first encoder layer. And temporal periods of 0–20 and 40–60 have the larger weight values, which indicates the subtle phenological differences are mainly in these periods.

From the above analysis, we find that the phenological differences of some crops are mainly manifested in subtle periods. Benefitting from the tense multitemporal data of 65 in one year, as well as the powerful self-attention module to express the temporal sequence attention, we can get better classification results. In other words, the attention periods contribute more to the classification results, and the discrimination of subtle differences can be realized through self-attention mechanism.

In summary of this article, we propose the CNN-transformer networks to perform crop classification on multitemporal multispectral dataset. As more and more Earth observation satellites are put into use, we can obtain dense multitemporal remote sensing images. A study dataset with 65 acquiring dates is collected from Sentinel-2 A/B and Landsat-8 in the article. The multiband images of sensors have different spatial resolutions, and then we use transposed convolution which can extract spatial structure information of multibands to fuse the spatial-spectral features. Multitemporal data can be viewed as a sequence of features. To deal with the sequence information, we borrow the transformer architecture from the knowledge of NLP, which has the powerful modeling capability of sequence information, to model the correlation between time series in crop classification. In the networks, after first obtaining the unified multitemporal features, we get the spatial-spectral features and position embeddings of the sequence information. Second, the encoder module derived from transformer is used to express the correlation of the sequence, and by stacking encoder modules with four layers, we can obtain the depth pattern characteristics of the sequence. Third, we use the feed-forward layer to extract the category features of crops. Finally, the crop label is predicted by softmax output layer. In the experiments, the results have proved that the proposed CNN-transformer method can achieve the excellent performance compared with other traditional methods.

This method utilizes the final output sequence of the multilayer encoder module to extract the category feature. The outputs of different encoder layers can express the sequence correlation of different levels. In the future, it can be considered to fuse the output feature sequences of different layers, which can better utilize the relationship between the various levels of sequence information and improve the performance of the model.

Acknowledgment

The authors would like to thank the anonymous reviewers for their careful reading and valuable comments.

References

[1] E. Zillmann *et al.*, “Pan-European grassland mapping using seasonal statistics from multisensor image time series,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 8, pp. 3461–3472, Aug. 2014.

[2] N. Kussul, M. Lavreniuk, S. Skakun, and A. Shelestov, “Deep learning classification of land cover and crop types using remote sensing data,” *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 778–782, May 2017.

[3] D. P. Roy *et al.*, “Landsat-8: Science and product vision for terrestrial global change research,” *Remote Sens. Environ.*, vol. 145, no. 145, pp. 154–172, 2014.

[4] Q. Wang, G. A. Blackburn, A. O. Onojeghuo, J. Dash, and P. M. Atkinson, “Fusion of landsat 8 oli and sentinel-2 MSI data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 7, pp. 3885–3899, Jul. 2017.

[5] J. Xiong *et al.*, “Nominal 30-m cropland extent map of continental Africa by integrating pixel-based and object-based algorithms using Sentinel-2 and Landsat-8 data on Google earth engine,” *Remote Sens.*, vol. 9, no. 10, 2017, Art. no. 1065. [Online]. Available: http://www.mdpi.com/2072-4292/9/10/1065

[6] X. P. Song *et al.*, “National-scale soybean mapping and area estimation in the united states using medium resolution satellite imagery and field survey,” *Remote Sens. Environ.*, vol. 190, pp. 383–395, 2017.

[7] R. Devadas, R. J. Denham, and M. Pringle, "Support vector machine classification of object-based data for crop mapping, using multi-temporal landsat imagery," *ISPRS—Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. XXXIX-B7, no. 4, pp. 185–190, 2012.

[8] X. Li *et al.*, "Crop classification recognition based on time-series images from HJ satellite," *Trans. Chin. Soc. Agricultural Eng.*, vol. 29, no. 2, 2013, Art. no. 2013.

[9] F. Melgani, "Classification of multitemporal remote-sensing images by a fuzzy fusion of spectral and spatio-temporal contextual information," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 18, no. 2, pp. 143–156, 2004.

[10] M. Pax-Lenney and C. E. Woodcock, "Monitoring agricultural lands in egypt with multitemporal landsat TM imagery: How many images are needed?" *Remote Sens. Environ.*, vol. 59, no. 3, pp. 522–529, 1997.

[11] J. Useya and S. Chen, "Comparative performance evaluation of pixel-level and decision-level data fusion of landsat 8 OLI, Landsat 7 ETM+ and Sentinel-2 MSI for crop ensemble classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 11, pp. 4441–4451, Nov. 2018.

[12] L. Zhang, L. Zhang, and D. Bo, "Deep learning for remote sensing data: A technical tutorial on the state-of-the-art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, Jun. 2016.

[13] N. Yokoya *et al.*, "Open data for global multimodal land use classification: Outcome of the 2017 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 5, pp. 1363–1377, May 2018.

[14] Z. Qiang, Q. Yuan, Z. Chao, X. Li, and Y. Wei, "Missing data reconstruction in remote sensing image with a unified spatial-temporal-spectral deep convolutional neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 1–15, Aug. 2018.

[15] Y. Chen, Z. Lin, Z. Xing, W. Gang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, Jun. 2017.

[16] B. A. Olshausen, C. H. Anderson, and V. Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *J. Neurosci.*, vol. 13, no. 11, pp. 4700–4719, 1993.

[17] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[18] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.

[19] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, (NIPS'14). MIT Press, Cambridge, MA, USA, vol. 2, 2014, pp. 2204–2212.

[20] T. Bluche, "Joint line segmentation and transcription for end-to-end handwritten paragraph recognition," *Advances Neural Inf. Process. Syst.*, vol. 29, 2016.

[21] A. Miech, I. Laptev, and J. Sivic, "Learnable pooling with context gating for video classification," 2017, *arXiv:1706.06905*.

[22] C. Long, H. Zhang, J. Xiao, L. Nie, and T. S. Chua, "SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning," in *Proc. IEEE Conf. Comput. Vision Pattern Recognit. (CVPR). IEEE Comput. Soc.,* 2017.

[23] K. Xu, J. Ba, R. Kiros, K. Cho, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," *Proc. 32nd Int. Conf. Mach. Learn.*, 2015, pp. 2048–2057.

[24] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," ICLR 2014.

[25] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1412–1421.

[26] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 379–389. [Online]. Available: http://arxiv.org/abs/1509.00685

[27] W. Yin, S. Ebert, and H. Schütze, "Attention-based convolutional neural network for machine comprehension," in *Proc. NAACL Human-Comput. Question Answering Workshop*, 2016, pp. 15–21.

[28] Z. Yang, D. Yang, C. Dyer, X. He, A. J. Smola, and E. H. Hovy, "Hierarchical attention networks for document classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Human Lang. Technol.*, 2016, pp. 1480–1489.

[29] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://blog.openai.com/language-unsupervised

[30] D. Mandal, V. Kumar, A. Bhattacharya, Y. S. Rao, P. Siqueira, and S. Bera, "Sen4rice: A processing chain for differentiating early and late transplanted rice using time-series Sentinel-1 SAR data with Google earth engine," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 12, pp. 1947–1951, Dec. 2018.

[31] A. Veloso *et al.*, "Understanding the temporal behavior of crops using sentinel-1 and sentinel-2-like data for agricultural applications," *Remote Sens. Environ.*, vol. 199, pp. 415–426, 2017.

[32] H. McNairn, X. Jiao, A. Pacheco, A. Sinha, W. Tan, and Y. Li, "Estimating Canola phenology using synthetic aperture radar," *Remote Sens. Environ.*, vol. 219, pp. 196–205, 2018.

[33] S. K. Maxwell, J. R. Meliker, and P. Goovaerts, "Use of land surface remotely sensed satellite and airborne data for environmental exposure assessment in cancer research," *J. Expo. Sci. Environ. Epidemiol*, vol. 20, no. 2, pp. 176–185, 2010.

[34] J. Shan, E. Hussain, K. H. Kim, Biehl, and Larry, "Flood mapping with satellite images and its web service," *PE & RS Photogrammetric Eng. Remote Sens.*, vol. 76, no. 2, pp. 102–105, 2010.

[35] Z. Zhe, S. Wang, and C. E. Woodcock, "Improvement and expansion of the fmask algorithm: Cloud, cloud shadow, and snow detection for landsats 47, 8, and sentinel 2 images," *Remote Sens. Environ.*, vol. 159, pp. 269–277, 2015.

[36] S. V. Skakun and R. M. Basarab, "Reconstruction of missing data in time-series of optical satellite images using self-organizing Kohonen maps," *J. Autom. Inf. Sci.*, vol. 46, pp. 19–26, 2014.

[37] A. Vaswani *et al.*, "Attention is all you need," *Advances Neural Inf. Process. Syst.*, vol. 30, pp. 5998–6008, 2017.

Zhengtao Li received the B.S. degree in electrical engineering and automation from Yanshan University, Qinhuangdao, China, in 2003, and the M.S. degree in pattern recognition and intelligent systems from Northeastern University, Shenyang, China, in 2008. He is currently a doctoral candidate with the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology, Wuhan, China, since 2014.

His research interests include deep learning methods for remote sensing, infrared target recognition and geometric deep learning.

Guokun Chen received the B.S. and M.S. degrees in measurement and control technology and instrument and pattern recognition and intelligent systems from Northeastern university, Shenyang, China, in 2007 and 2009, respectively.

Her research interests include deep learning methods for remote sensing, target recognition, and multimodel classification of remote sensing data.

Tianxu Zhang received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 1970. He received the M.S. degree in computer science and technology from Huazhong University of Science and Technology, Wuhan, China, in 1981 and the Ph.D. degree in optical engineering from Zhejiang University, Hangzhou, China, in 1989.

He is currently a Professor with the Institute for Pattern Recognition and Artificial Intelligence (IPRAI), Huazhong University of Science and Technology, as the Director of IPRAI. His research interests include image processing, computer vision, pattern recognition, and medical imaging. He has published more than 200 research papers.