Semantic Segmentation Based on Temporal Features: Learning of Temporal–Spatial Information From Time-Series SAR Images for Paddy Rice Mapping

Lingbo Yang, Ran Huang, Jingfeng Huang, Tao Lin, Limin Wang, Ruzemaimaiti Mijiti, Pengliang Wei, Chao Tang, Jie Shao, Qiangzi Li, and Xin Du

***Abstract*— Synthetic aperture radar (SAR) can be used to obtain remote sensing images of different growth stages of crops under all weather conditions. Such time-series SAR images can provide an abundance of temporal and spatial features for use in large-scale crop mapping and analysis. In this study, we propose a temporal feature-based segmentation (TFBS) model for accurate crop mapping using time-series SAR images. This model first extracts deep-seated temporal features and then learns the spatial context of the extracted temporal features for crop mapping. The results indicate that the TFBS model significantly outperforms traditional long short-term memory (LSTM), U-network, and convolutional LSTM models in crop mapping based on time-series SAR images. TFBS demonstrates better generalizability than other models in the study area, which makes it more transferable, and the results show that data augmentation can significantly improve this generalizability. The visualization of the temporal features extracted by the TFBS shows that there is a high degree of intraclass homogeneity among rice fields and interclass heterogeneity between rice fields and other features. TFBS also achieved the highest accuracy of the four deep learning models for multicrop classification in the study area. This study presents a feasible way of producing high-accuracy large-scale crop maps based on the proposed model.**

***Index Terms*— Data augmentation, feature visualization, generalization ability, temporal feature-based segmentation (TFBS), time-series images.**

Manuscript received December 14, 2020; revised March 31, 2021, May 12, 2021, and July 14, 2021; accepted July 20, 2021. Date of publication August 4, 2021; date of current version January 17, 2022. This work was supported in part by the Eramus+ Programme of the European Union under Grant 598838-EPP-1-2018-EL-EPPKA2-CBHE-JP and in part by the National Major Project of China under Grant 09-Y20A05-9001-17/18. *(Corresponding authors: Jingfeng Huang; Limin Wang.)*

Lingbo Yang, Jingfeng Huang, Ruzemaimaiti Mijiti, Pengliang Wei, and Chao Tang are with the Institute of Applied Remote Sensing and Information Technology, Zhejiang University, Hangzhou 310058, China (e-mail: yanglingbo@zju.edu.cn; hjf@zju.edu.cn; ruzi_mamat@zju.edu.cn; weipengliang@zju.edu.cn; xueshu@zju.edu.cn).

Ran Huang is with the School of Automation, Hangzhou Dianzi University, Hangzhou 310018, China (e-mail: ran_huang@hdu.edu.cn).

Tao Lin is with the College of Biosystems Engineering and Food Science, Zhejiang University, Hangzhou 310058, China (e-mail: lintao1@zju.edu.cn).

Limin Wang is with the Institute of Agricultural Resources and Regional Planning, Chinese Academy of Agricultural Sciences, Beijing 100081, China (e-mail: wanglimin01@caas.cn).

Jie Shao is with the Institut de Recherche en Informatique de Toulouse (IRIT), CNRS, University of Toulouse, 31062 Toulouse, France (e-mail: shaojie@mail.bnu.edu.cn).

Qiangzi Li and Xin Du are with the Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, China (e-mail: liqz@radi.ac.cn; duxin@radi.ac.cn).

Digital Object Identifier 10.1109/TGRS.2021.3099522

I. Introduction

WORLD production has grown significantly over the past 60 years, which has greatly reduced the proportion of hungry and undernourished people in the world. Nevertheless, a new set of challenges threatens world food security [1]–[4]. The number of hungry people worldwide has slowly risen since 2014, and a recent estimate for 2019 has revealed that an additional 60 million people have become affected by hunger during the past five years [4]. Cropland monitoring plays an important role in understanding the state of food security [5], [6], providing critical basic information for analyzing the change in the crop growing area [7], fluctuation of yield [8], and formulation of agricultural policies [9].

In recent years, with the explosive development of deep learning technologies, a series of state-of-the-art deep learning models has been developed and applied to the fine identification of crop types and other features based on remote sensing over large areas [10], [11]. Expert knowledge-based classification models and classical machine learning models, such as support vector machines (SVMs) and random forest (RF), typically require the extraction of effective features from raw data either manually or through data mining techniques before classification or modeling is performed [12]–[16]. In contrast, taking advantage of neural networks that imitate the human nervous system, deep learning models are capable of extracting massive and deep-seated features automatically from raw images [10], [17]–[19]. Through the deepening of neural networks and the subsequent increase in the number of neurons, deep learning models are capable of achieving higher performance than that of traditional machine learning models [20]–[23].

Time-series satellite images involve rich spatial contextual features and temporal features that are essential for object recognition [21], [22], [24]–[26]. Convolutional neural networks, which include a series of semantic segmentation models, focus on extracting the spatial contextual information from raw images via convolutional filters to realize end-to-end classification [27]–[30]. The semantic segmentation model, which includes the fully convolutional network (FCN), SegNet, and U-network (UNET) models, is a type of model that can classify each pixel of an image into a predefined category based on a series of convolutional and pooling lay-

ers [24], [30]–[33]. It generally contains an encoding structure followed by a decoding structure. The encoding structure can produce deep classification information, whereas the decoder can produce precise boundaries for each class. As it is capable of learning spatial contextual information from the local to the global scale, semantic segmentation models have been initially applied to cloud recognition [24], [30], [34], road detection [32], and crop classification [35], [36], demonstrating better performance than traditional classification methods.

In addition, recurrent neural networks (RNNs), represented by long short-term memory (LSTM) and gated recurrent unit (GRU) networks, can exhibit temporal dynamic information from raw time-series satellite images to carry out pixel-wise classification based on a collection of fully connected layers [37]–[39]. The same process is performed for each element of a sequence, whereby the output of each step depends on the computations of previous steps along with the current input [40]–[42]. This design gives it a memory of what information has already been presented in the sequence. LSTM is a specific RNN that remembers information over arbitrary intervals and is designed to deal with the exploding gradients or vanishing gradients problem of classical RNNs [43]. It is capable of capturing short- or long-term dependencies in sequence data, such as time-series satellite images [22], [44]. Taking advantage of its insensitivity to interval length, LSTM has achieved great success in the prediction and recognition of long sequences [42], [45]. Based on previous research conducted in recent years, LSTM is capable of efficiently and automatically learning temporal features from time-series images and has shown great potential in time-series image-based crop classification [21], [42], [46].

Deep learning models usually have a considerable number of parameters; hence, training a reliable and high-precision model requires a large amount of training data and corresponding label data [47], [48]. In recent years, a series of medium-resolution satellites, particularly satellites equipped with a synthetic aperture radar (SAR) instrument, have been launched, making it possible to obtain sufficient, cloud-free remote sensing time-series images [49], [50]. Nevertheless, obtaining the most recent accurate and large-scale label data is always a significant challenge for crop classification through deep learning models. The Cropland Data Layer (CDL), published by the United States Department of Agriculture (USDA), provides annually produced crop-specific land-cover maps for the continental United States. CDL provides sufficient, reliable label data for training and testing novel deep learning models [51], [52]. Based on multiyear, multiregional CDL data, the generalizability of deep learning models, which is critical for crop recognition in areas where reliable and large amounts of label data are difficult to obtain, can be evaluated.

Moreover, data augmentation is a potential technique for improving the generalizability of deep learning models [53]. It is a strategy that can significantly increase the diversity of training data, especially when collecting the actual data from where the model is to be applied proves difficult [54]. As deep learning models are heavily reliant on big data to overcome overfitting, exploring the data augmentation techniques and evaluating their role in improving the generalizability of deep learning models are of great importance [54]–[57].

Sentinel-1 satellites are capable of acquiring data under all weather conditions during both day and night. Moreover, the rapid revisit of Sentinel-1 can provide greater accuracies in crop mapping based on the acquisition of time-series images. Although great progress has been made in the research of spatial–temporal deep learning models in recent years and some spatial–temporal prediction models [e.g., ConvLSTM (CLSTM)] [57]–[59] have been developed. Most models first extract spatial features and then temporal features afterward; few models learn temporal features before spatial features. In addition, the performance of these models in crop recognition based on time-series remote sensing data still needs to be tested, and the development of new temporal–spatial deep learning models for crop recognition based on time-series satellite images remains important. To fully exploit and utilize the temporal and spatial features contained in time-series SAR data, a novel temporal feature-based segmentation (TFBS) model is proposed in this study and used for rice recognition of a large area located in the U.S. rice belt. Sentinel-1 SAR time-series images were fed into the model, and CDL data were used as label data. In addition, LSTM, UNET, and ConvLSTM models are trained and tested based on the same dataset for sake of comparison. The following questions are addressed in this article.

1) What is the performance of the TFBS model compared with those of the classical temporal feature-based model (LSTM), spatial feature-based model (UNET), and spatial–temporal prediction model (ConvLSTM)?
2) What is the temporal and spatial generalizability of the TFBS model compared with those of the LSTM, UNET, and ConvLSTM models?
3) What is the nature of the temporal feature used in the TFBS model and how does it improve the classification capability of the TFBS model?
4) Can data augmentation improve the generalizability of the TFBS model?
5) Is it feasible to extend the binary classification model to a multicrop classification task?

II. Study Area and Data

A. Study Area

The study areas are located in the southern-center part of the United States and the center of California (Fig. 1), including areas of Arkansas (AR), Mississippi (MS), southern Missouri (MO), western Tennessee (TN), northern Louisiana (LA), and the Sacramento Valley (SV), which represents a significant portion of the U.S. rice crop cultivation [60], [61]. According to CDL data from 2019 [62], rice production in the study area accounts for 76% of the total rice production in the United States.

The life cycle of rice cultivars in the southern-center part of the United States ranges from 105 to 145 d from germination to maturity, depending on the variety and environment [63]. Rice seeding in this area is typically performed in April and May, within a period of three to five weeks, and harvested from September to October [64], [65]. Besides, the rice fields

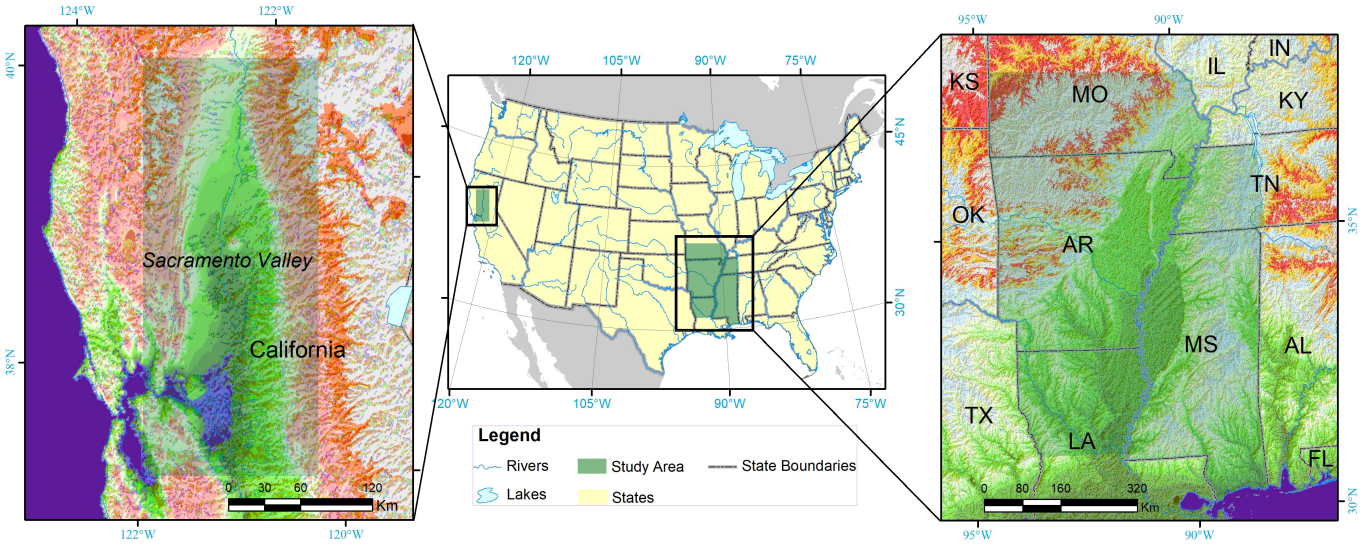


Fig. 1. Location of the study area. The study area includes AR, MS, southern MO, western TN, northern LA, and the SV of California.

in the SV are usually flooded and aerially seeded in May and harvested from September to October [66].

B. Sentinel-1 Time-Series SAR Images

Sentinel-1 is the first satellite constellation of the Copernicus Program, which was implemented by the European Space Agency (ESA) [67]. It comprises two polar-orbiting satellites, Sentinel-1 A and Sentinel-1 B, which share the same orbit [68]. The orbit has a 12-day repeat cycle and completes 175 orbits per cycle. Sentinel-1 carries a C-band SAR instrument operating in four exclusive acquisition modes with different spatial resolutions and swath widths [69]. The interferometric wide swath model is the only operational model over the study area, which offers vertical transmit, vertical receive (VV) + vertical transmit, horizontal receive (VH) polarization data with a large swath width (250 km) and moderate geometric resolution (5 m × 20 m). Taking advantage of the SAR instrument, Sentinel-1 satellites are capable of acquiring data over the study area day or night under all weather conditions, regardless of cloud cover or solar illumination [70], [71].

Corresponding to the rice phenology of the study area, all Sentinel-1 VV and VH polarized SAR images from April to October for the years 2017–2019 were selected. Sentinel-1 VV and VH 24-day-averaged composite time-series images were then produced based on the selected SAR images. Each of the composite time-series images was composed of nine channels, each representing a 24-day-averaged composite image (Table I). The time-series images were then resampled to a resolution of 30 m to ensure consistency with the CDL data.

All the Sentinel-1 data selection and preprocessing were completed on the Google Earth Engine (GEE) cloud platform [15], [72]. GEE is a planetary-scale platform for Earth observation and analysis, which archives a large catalog of satellite imagery and geospatial datasets, updated and expanded daily [73]. The Sentinel-1 dataset on GEE includes the Sentinel-1 Ground Range Detected (GRD) images. It is provided by ESA and has been preprocessed to the backscatter coefficient (σ°) in decibels (dB), which makes acquiring Sentinel-1 data over the study area quite efficient and convenient [74]. Finally, the Sentinel-1 VV/VH 24-day-averaged composite time-series images produced by GEE were downloaded to a local computer for subsequent study.

TABLE I
START AND END DATES OF THE NINE 24-DAY-AVERAGED COMPOSITE SENTINEL-1 SAR IMAGES

Index	Start date	End date
1	April 1	April 24
2	April 25	May 18
3	May 19	June 11
4	June 12	July 5
5	July 6	July 29
6	July 30	August 22
7	August 23	September 15
8	September 16	October 9
9	October 10	November 2

C. Cropland Data Layer

CDL, a 30-m resolution crop-specific land cover map produced annually for the continental United States [75], [76], is used as reference data for model training and testing. It was produced by the USDA, National Agricultural Statistics Service (NASS), Research and Development Division based on extensive agricultural ground truth and a series of moderate resolution satellite imagery, including Landsat-8, Sentinel-2, and Deimos [62].

CDL products from 2017–2019 were used in this study. The 2017 CDL product was released on January 26, 2018, the 2018 CDL product was released on February 15, 2019, and the 2019 CDL product was released on February 5, 2020. Based on the metadata corresponding to the CDL product, the rice accuracy for each state area is provided [62]. The mean kappa values of rice in the study area in 2017–2019 were 0.9058, 0.8894, and 0.9150, respectively. Meanwhile, these

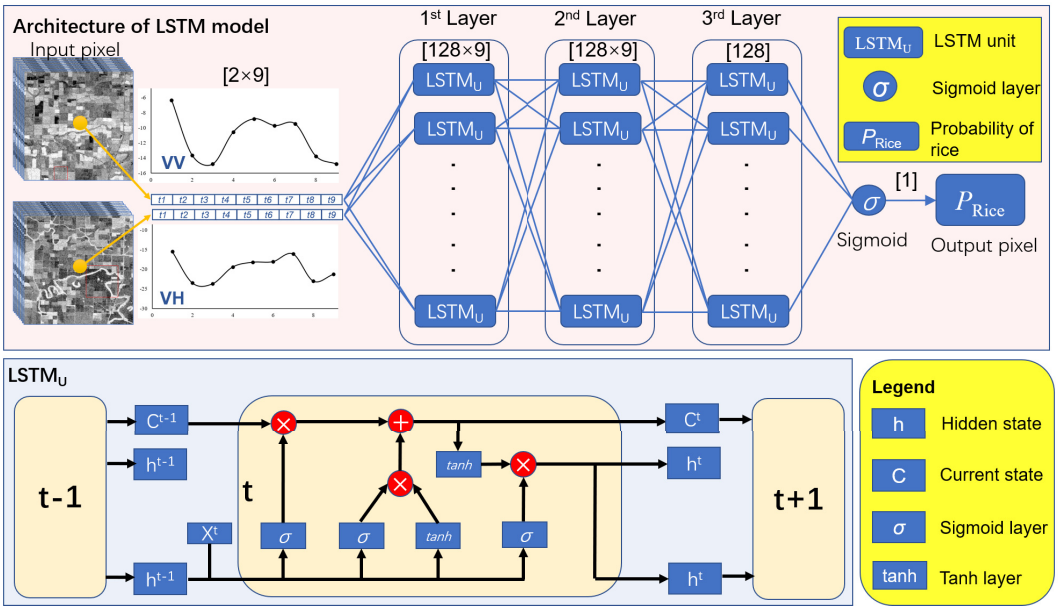


Fig. 2. Architecture of LSTM model used in this study. The first LSTM layer receives a vector with dimensions 2 × 9 from a pixel of the input VV and VH time-series images and outputs a vector with dimensions 128 × 9, which is used as the input data of the second LSTM layer. The third LSTM layer receives a vector with dimensions 128 × 9, which is output from the second LSTM layer and yields an output vector of dimensions 128 × 1. Finally, a sigmoid layer is employed to generate the probability that the pixel is classified as rice. Thus, a hard classification could be applied based on a threshold of 0.5.

years' mean user's accuracies of rice were 97.26%, 97.40%, and 97.42%, along with mean producer accuracies of rice were 91.04%, 89.48%, and 91.90%, respectively. The high accuracy of the CDL data makes it a reliable dataset for the training and testing of deep learning models [51], [52], [77].

III. Methodology

A. LSTM Model

The detailed architecture of the LSTM model used in this study is shown in Fig. 2. The LSTM model receives pixel-wise temporal satellite observations, including VV and VH polarization data as input data. In the model, three LSTM layers are employed with each layer containing 128 LSTM units. Each LSTM unit consists of a series of LSTM time steps (nine in this study, since there are nine time-series VV or VH images each year). Each time step consists of three gates, which are neurons that optionally let information through. The first gate, denoted the "forget gate," decides how much information is forgotten from the old state of the previous step. The second gate, denoted the "input gate," decides how much information is stored in the state of the current step. Meanwhile, the third gate, denoted the "output gate," decides how much information is transported to the next step. Taking advantage of these three gates, LSTM is capable of learning the dependencies of long-sequence data without the presence of gradient explosion or vanishing problems [22], [45], [46].

In this study, each LSTM unit in the first two LSTM layers returns the hidden state output for each input time step. Thus, nine hidden states are produced from each LSTM unit, and each LSTM layer returns a vector with dimensions 128 × 9. The LSTM units in the third LSTM layer only return the hidden state of the previous time step. Therefore, the third LSTM returns a vector with dimensions 128 × 1. Then, a sigmoid layer is employed to generate the probability that a pixel is classified as rice. Finally, a hard classification can be applied based on a threshold of 0.5.

B. UNET Model

UNET is a model developed for image segmentation. It labels each pixel of an image with a corresponding class and uses convolutional layers instead of fully connected layers, which makes it capable of efficiently handling images of any size. The architecture of the UNET model used in this study is shown in Fig. 3. UNET consists of a down-sampling path (also called the encoder) and an up-sampling (also called the decoder) path, which gives it a U-shaped structure. The encoder consists of a collection of successive convolutional and max-pooling layers, which is capable of understanding the local context from the satellite image on different scales. During the down-sampling pathway, local spatial information is reduced while global contextual information is highlighted. The decoder has a structure that is symmetrical to the encoder, resulting from a sequence of convolutions and up-convolutions, and the up-sampled features are concatenated with high-resolution features from the down-sampling path through skip connections between the encoder and decoder. Based on this design, abundant features and spatial information are combined to produce a precise segmentation map [31], [33], [35].

In this study, the temporal stacked image is first split into a collection of tiles, each of which contains 18 channels (nine VV and nine VH channels) and has a spatial size of 128 × 128. The encoder has four 2 × 2 max-pooling layers. Therefore, the spatial size of each tile is reduced to 8 × 8, while the count of the channel is increased to 512. The decoder has four up-convolutions; hence, the spatial size is restored to

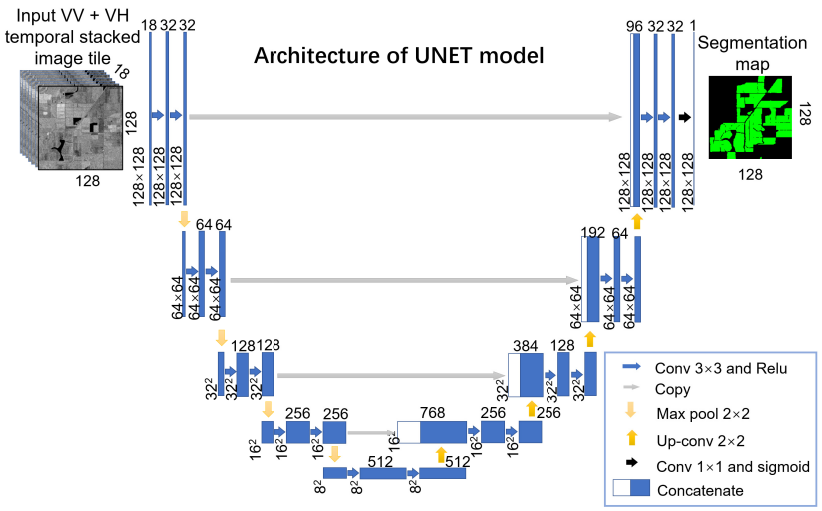


Fig. 3. Architecture of UNET model used in this study. Blue boxes represent feature images, and white boxes refer to copied feature images. The spatial size of the input data is 128 × 128 with 18 channels (nine VV and nine VH images). The spatial size of each intermediate feature image is marked at its lower left side. The count of the channel of each intermediate feature image is marked on its top side. The output data is the probability of each pixel of the image to be labeled as rice. A hard classification map could be generated based on the probability map, using a threshold of 0.5.

128 × 128, and the count of the channel is reduced to 32. Finally, a 1 × 1 convolutional layer and a sigmoid layer are used to produce a segmentation map that has the same spatial size as the input image tile.

C. Convolutional LSTM

ConvLSTM is the extension from a fully connected LSTM (FC-LSTM) and is capable of learning spatiotemporal correlations from raw time-series data. It has convolutional structures in both input-to-state and state-to-state transitions [78], [79]. In this study, the training image is split into tiles of size 128 × 128. Then, each tile is reshaped to dimensions of 9 × 2 × 128 × 128 and fed to the ConvLSTM model. Then, two ConvLSTM layers are stacked to build an end-to-end model for time-series SAR image-based crop classification. Each ConvLSTM layer consists of 64 ConvLSTM2D units. Each ConvLSTM2D unit is just like the LSTM unit, but internal matrix multiplications are replaced with convolution operations. Finally, a 1 × 1 convolutional layer and a sigmoid layer are used to produce soft classification, based on the spatial–temporal features extracted by the ConvLSTM layers; then, hard classification can be carried out based on a threshold of 0.5 (Fig. 4).

D. TFBS Model

The TFBS model contains four modules: an input module, temporal feature extraction module, segmentation module, and output module. It employs an LSTM model matrix to learn temporal features from raw time-series satellite images and produces an image consisting of these temporal features. The temporal feature image is then input into a UNET module to extract spatial context information from the temporal features and produce a segmentation image (Fig. 5).

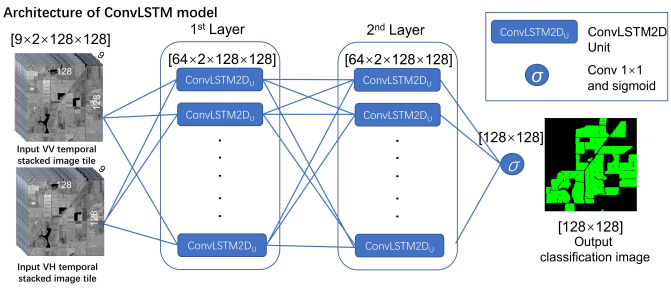


Fig. 4. Architecture of ConvLSTM model. The input image has a dimension of 9 × 2 × 128 × 128, the output feature image of each ConvLSTM layer has a dimension of 64 × 2 × 128 × 128. The output of the second ConvLSTM layer is input into a 1 × 1 convolutional layer and a sigmoid layer to produce a soft classification image.

1) *Input module:* The input data of the TFBS model is an image tile that is clipped from the original input satellite image. The dimensions of the input data are 18 × 128 × 128, which represents the channel count, width, and height. Then, a reshape layer is used to dissolve the input image into independent pixels. Each pixel has dimensions 2 × 9, which means that it has two features (VV and VH), each of which containing nine time-series images from April to October.
2) *Temporal Feature Extraction Module:* The temporal feature extraction module consists of a 128 × 128 LSTM model matrix and a reshape layer. Each LSTM model corresponds to one pixel from the input module. Each model in the LSTM matrix is exactly the same because they have the same model structure and parameter values (weights and bias). The LSTM model has one LSTM layer that consists of 64 LSTM units; thus, 64 temporal features are generated from each LSTM model (each unit returns only the hidden state of the final time step). Finally, all the temporal features from the LSTM model matrix make up a temporal feature image with dimensions 64 × 128 × 128, based on the reshape layer.

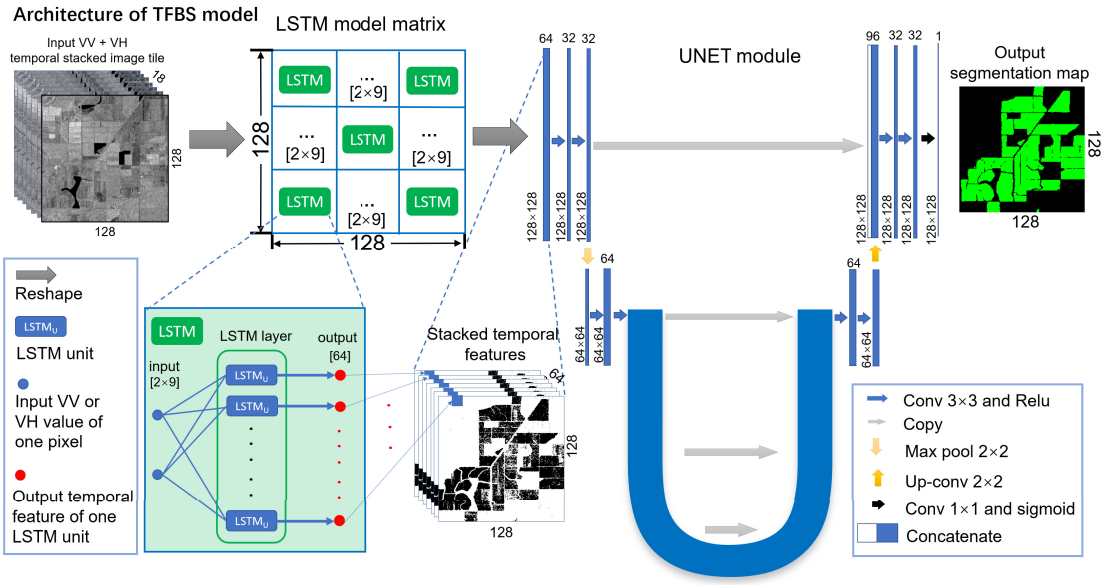


Fig. 5. Architecture of TFBS model. The input image is dissolved into independent pixels, and an LSTM model is applied to each pixel. Thus, 64 temporal features are learned from the raw time-series image and are used as the input data of a UNET segmentation model. Finally, a segmentation map is generated based on the combination of temporal and spatial contextual information.

3) *Segmentation Module:* A UNET model is employed to segment the temporal feature image in the segmentation module. The structure of the UNET model is similar to that shown in Fig. 3. The down-sampling path of the UNET model reduces the spatial size of the input data from 128 × 128 to 8 × 8 and increases the feature count from 64 to 512. Hence, during the down-sampling pathway, the local spatial information is reduced, whereas the abstract feature information is increased. During the up-sampling path, the spatial size of the feature image is restored to 128 × 128, whereas the channel size is reduced to 32. Based on skip connections and concatenations, detailed spatial information is combined with abundant feature information to produce a precise segmentation image.
4) *Output Module:* A 1 × 1 convolutional layer with only one neuron is applied to obtain a one-channel image; then, a sigmoid layer is used to map the image value between 0 and 1 and realize the soft classification of the image. A hard classification result can be obtained based on the soft classification map, using a threshold of 0.5.

E. Workflow of the Study

To fully assess the performance of the TFBS model, we divide the study area into three parts to test the performance of the TFBS model and evaluate the temporal and spatial generalizability. The first part accounts for the majority of the study area, including AR, MS, southern MO, and western TN (hereinafter referred to as ARMSMOTN). The second and third parts represent northern LA and SV of California, respectively.

The Adam optimizer is employed in all four models, and the cross-entropy metric is used as the loss function in the training process. The performances of the LSTM, UNET, ConvLSTM, and TFBS models are evaluated using datasets from ARMSMOTN in 2019, with 10-fold cross validation used as the evaluation method. This method randomly partitions the original dataset into ten equally sized subdatasets. Then, a single subdataset is retained as the validation dataset, and the remaining nine subdatasets are used as training data. This process is repeated ten times, with each of the ten subdatasets used exactly once as the validation dataset. A final estimation can then be calculated by averaging the ten results. The standard deviation of the estimation could also be calculated to assess the stability of all four deep learning models.

The growth period of crops varies slightly because the climate conditions vary from place to place and year to year. Therefore, when the model trained in a specific year and place is applied to other years or other places, its accuracy in crop classification is affected by the temporal and spatial generalizability of the model. In this study, all the data from ARMSMOTN in 2019 were used to train four deep learning models. The datasets from ARMSMOTN in 2017 and 2018 were used to evaluate the temporal generalizability of the trained deep learning models. In addition, with a view to further testing the spatial–temporal generalizability of the deep learning models, the datasets from northern LA in 2017–2019 were used as validation data to test the deep learning models trained with the ARMSMOTN 2019 dataset.

Besides, deep learning models which pretrained using the ARMSMOTN 2019 dataset are applied to the SV region for rice mapping. Due to the sowing date of rice in SV is relatively late than that in ARMSMOTN, it is not practical to map rice in SV by using the pretrained model directly. Therefore, fine-tuning method is employed to adapt pretrained models to new areas [80]. The parameters of the output layer for each pretrained deep learning model are initialized as zero, while the parameters of other layers are retained. An image tile

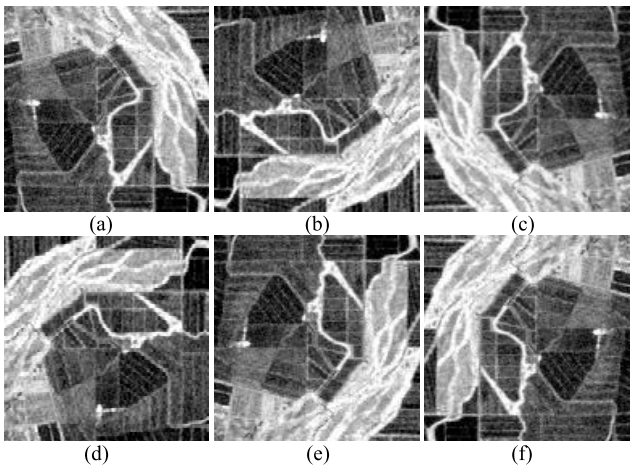


Fig. 6. Illustration of data augmentation. (a) Original SAR image. (b) Rotated 90° clockwise. (c) Rotated 180° clockwise. (d) Rotated 270° clockwise. (e) Flipped vertically. (f) Flipped horizontally.

with a spatial size of 128 × 128 pixels and its corresponding CDL image is randomly selected from the SV 2019 dataset. The selections are used to retrain the new output layer for each pretrained deep learning model. Based on the fine-tuned model, the rice classification results in the SV area from 2017 to 2019 are obtained, and the classification accuracy is evaluated by using CDL data as reference. In order to ensure the reliability of the results and the stability evaluation of the fine-tuning for each model, this process was repeated ten times. The mean and standard deviation of each classification accuracy metric are calculated and used for model evaluation.

Data augmentation is a useful technique to increase the amount and the diversity of the training data. The effects of data augmentation on the spatial and temporal generalizability of the TFBS model are evaluated in this study. Several augmentation techniques, including spatial augmentation (rotation and flipping) and temporal augmentation (random scaling of backscatter coefficient), are used to increase the size of the dataset from ARMSMOTN in 2019. The raw tile images are first randomly rotated 90°, 180°, or 270° clockwise [Fig. 6(b)–(d)]; then, the rotated images are multiplied by a random value between 0.9 and 1.1 to increase the fluctuation of the time-series curve of the sample data. In addition, each raw tile image is flipped vertically or horizontally [Fig. 6(e) and (f)], and the flipped image is subsequently multiplied by a random number between 0.9 and 1.1. Thus, the size of the training dataset is tripled.

This study also attempts to explore and visualize the intermediate temporal features, which are abstract features learned from the LSTM module of the TFBS model. Both the Jeffries–Matusita (J–M) and transformed divergence (TD) separability measures are used [81], [82] to evaluate the improvement in the classification ability of temporal features compared with the original time-series images.

Finally, the performance of multiclass classification is discussed based on a combination of binary classification models, in which four independent binary classification models are trained separately based on the augmented ARMSMOTN 2019 dataset and CDL for corn, cotton, rice, and soybeans, the primary crop types in the study area. Each model is used to produce a soft classification result for a particular crop. The soft classifications of each crop are then multiplied by a weight, which is defined by the validation precision value of each model, for which the precision value represents the probability of a correct prediction. For each pixel, according to the weighted soft classification results of four crops, the category with the largest value is selected as the final crop category. The formula is given as

$$C_{m,n} = \arg\max \left(S^i_{m,n} \times (S^i_{m,n} \geq 0.5) \times P^i \right) \tag{1}$$

where $C_{m,n}$ represents the final category of the pixel at location (m, n), i represents category i, S represents the sigmoid result, and P represents the precision value. Argmax is a function which returns the category index of the maximum value.

All the experiments were implemented based on Python 3.7. Keras with a TensorFlow backend was used to develop, train, and test the deep learning models. The geospatial data abstraction library (GDAL) is used to process raster images. The version of Keras employed in this study is 2.3.1, the version of TensorFlow is 2.2.0, and the version of GDAL is 2.3.3. All the processes were performed on a Windows 10 workstation with an AMD Ryzen Threakripper 1950X (3.4 GHz/16 Cores/32 MB Cache), 96 GB of RAM, and two NVIDIA GeForce GTX 1080 Ti graphics cards (each with 11 GB of video RAM). The dataset from ARMSMOTN in 2019 is used to evaluate the training time from each model, and the dataset from ARMSMOTN in 2018 is used to evaluate the prediction time from each deep learning model.

F. Accuracy Assessment

CDL maps of the study area in 2017–2019 were used as a reference to evaluate the accuracy of the model prediction, whereby four metrics were used to measure the accuracy of each model—Cohen's kappa coefficient, recall, precision, and F-score.

Cohen's kappa coefficient is a measure of agreement between the reference and predictions. This indicates the proportion of agreement beyond that expected by chance [83]. For a binary classification problem, the definition of Cohen's kappa coefficient is

$$\text{kappa} = (p_o - p_e)/(1 - p_e) \tag{2}$$

$$p_o = \frac{1}{N} \sum_{i=1}^{2} n_i^{\text{corr}} \tag{3}$$

$$p_e = \frac{1}{N^2} \sum_{i=1}^{2} n_i^{\text{pred}} n_i^{\text{ref}} \tag{4}$$

where p_o is the probability of correct prediction, which can be calculated by dividing the total number of correctly classified pixels by the total number of pixels, and p_e is the probability of agreement expected by chance. N is the total number of pixels, n_i^{corr} is the number of pixels in category i that are correctly classified, n_i^{pred} is the total number of pixels in category i in the classified image, and n_i^{ref} is the total number of pixels in category i in the reference image.

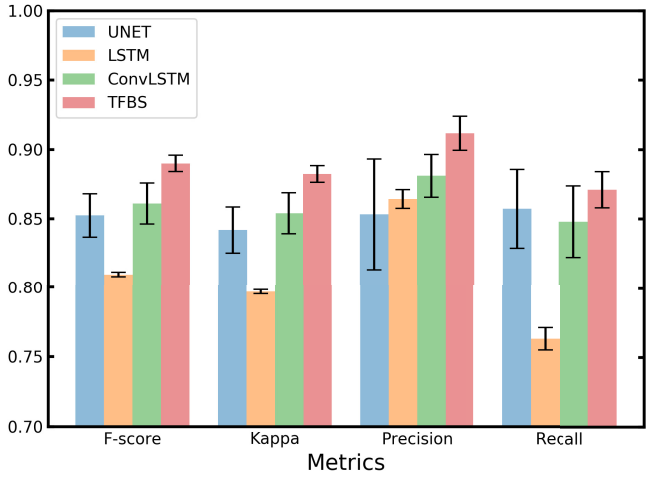


Fig. 7. Final result of the average F-score, kappa coefficient, precision, and recall of UNET, LSTM, TFBS, and ConvLSTM models, assessed by 10-fold cross validation method based on the dataset from ARMSMOTN in 2019. Error bars in the figure represent one standard deviation from the average accuracy.

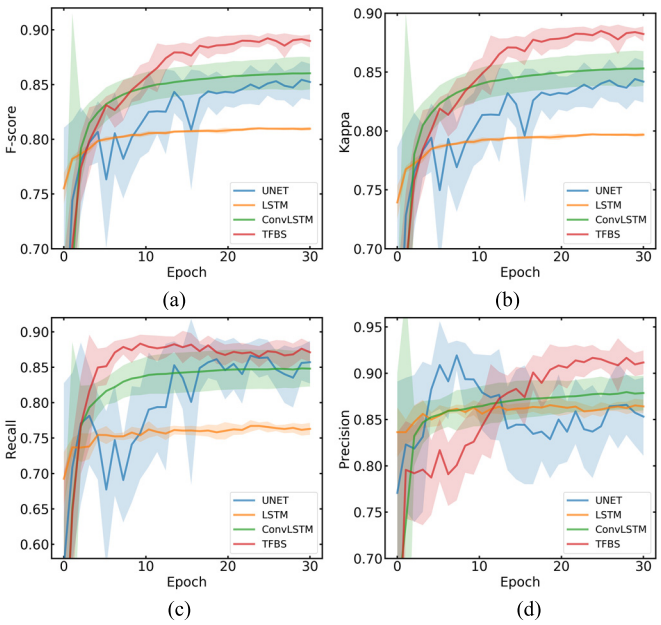


Fig. 8. Change in (a) F-score, (b) kappa coefficient, (c) recall, and (d) precision of each model across 30 epochs. All the accuracy metrics are evaluated based on 10-fold cross validation technology using the dataset from ARMSMOTN in 2019. Thus, each model is trained and tested ten times. The red lines represent the trends of the mean accuracies of the TFBS models with each epoch spanning the training process, whereas the orange lines represent those of the LSTM models, blue lines represent those of the UNET models, and green lines represent those of the ConvLSTM models. The light buffer areas along the lines refer to one standard deviation from the mean accuracy.

Recall is the proportion of correctly classified pixels of category i to the total number of pixels of category i in the reference image. Precision is the proportion of correctly classified pixels of category i to the total pixels of category i in the classified image. Recall and precision can be calculated from the following:

$$\text{Recall}_i = n_i^{\text{corr}} / n_i^{\text{ref}} \tag{5}$$

$$\text{Precision}_i = n_i^{\text{corr}} / n_i^{\text{pred}} \tag{6}$$

where Recall_i denotes the recall score of category i and Precision_i denotes the precision score of category i.

F-score is a measure of the test's accuracy, which considers both the precision score and the recall score

$$F_i = 2 \times \frac{\text{Recall}_i \times \text{Precision}_i}{\text{Recall}_i + \text{Precision}_i} \tag{7}$$

where F_i is the F-score of category i.

IV. Results and Discussion

A. Performance of TFBS Model Compared With Other Models

Fig. 7 shows that TFBS significantly outperformed UNET, LSTM, and ConvLSTM in all four accuracy metrics. The TFBS model produces the highest average kappa, F-score, precision, and recall scores of 0.8824, 0.8899, 0.9116, and 0.8711, respectively. Meanwhile, the ConvLSTM model achieves the second highest accuracy, slightly higher than the UNET model, and much higher than that of the LSTM model. This indicates that the combined use of temporal and spatial information can improve the accuracy of crop recognition based on time-series images.

The accuracy trends across the training epochs are shown in Fig. 8. This indicates that the LSTM model starts to converge at five epochs, whereas the other models start to converge at approximately 15 epochs. The standard deviation also shows that LSTM is much more stable than the UNET, TFBS, or ConvLSTM models during cross-validation. The UNET and ConvLSTM models are very sensitive to the change in samples, as the accuracy of these models trained with different samples varies greatly in the cross-validation process. The classification accuracy of the TFBS model exhibits medium stability, as the standard deviations of each metric are higher than those of the LSTM model but lower than those of the UNET and ConvLSTM models (Figs. 7 and 8).

Three typical sites from the study area are selected to illustrate the classification results of LSTM, UNET, ConvLSTM, and TFBS models (Fig. 9). The classification results produced by TFBS are much closer to the CDL maps than those of the LSTM and UNET models. Two advantages lead to the outperformance of the TFBS model over the UNET and LSTM models. First, compared with the UNET model, the TFBS model contains an LSTM module that is capable of recognizing rice pixels with weak temporal signals, as it can learn deep temporal information from the raw time-series images of SAR; in contrast, the UNET model may omit them, causing omission problems (blue rectangles marked in Fig. 9). Second, compared with the LSTM model, TFBS employs a UNET module to obtain the spatial contextual information contained in the temporal features for segmentation, which could produce complete rice fields in the classification result. The raw SAR images are naturally full of speckle noise, which could cause a strong "salt-and-pepper" effect on the result of a pixel-wise classification method such as LSTM (red rectangles marked in Fig. 9), as only the spectral or

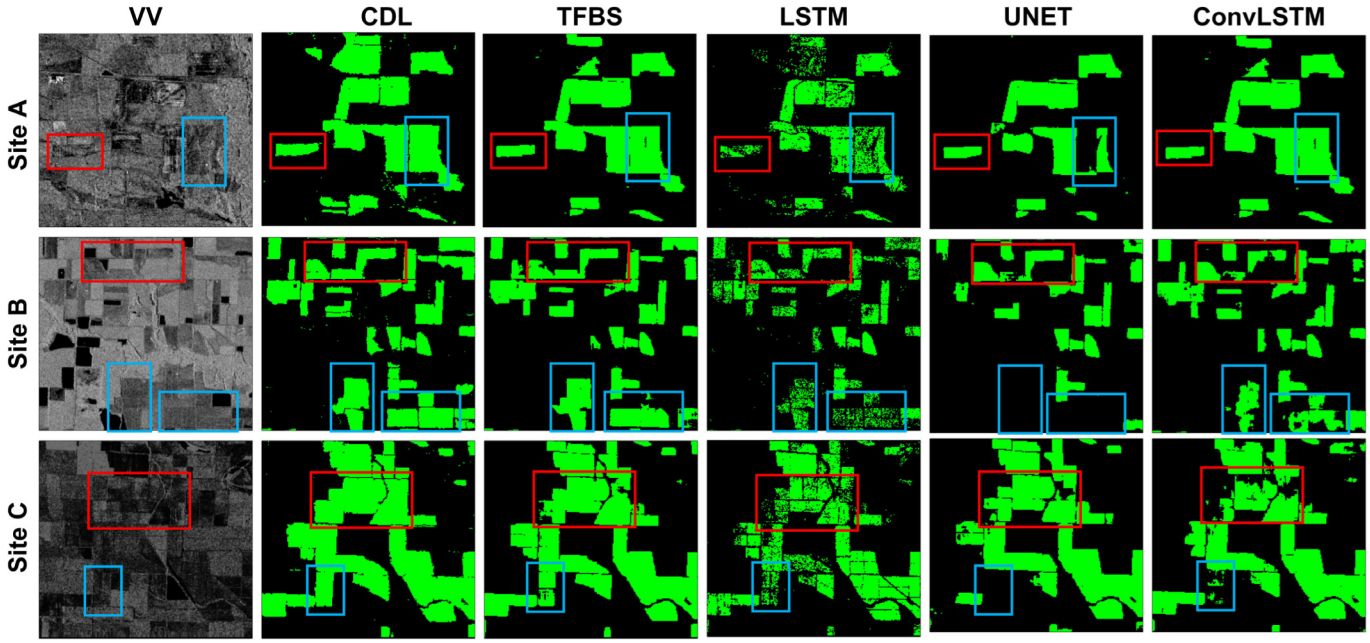


Fig. 9. Illustration of VV backscatter coefficient images (July 6, 2019) and the corresponding classification results based on different deep learning models. Blue rectangles on the maps denote omission errors of UNET model that other models do not exhibit. Red rectangles on the maps denote spackle problems of LSTM model that other models do not contain.

temporal information from each independent pixel are taken into account without consideration of its spatial context. Both the TFBS and UNET models are capable of learning spatial context on different scales based on the design of convolutional and max-pooling layers. This characteristic gives them the ability to reduce the negative impact of speckle noises on SAR data. Taking advantage of the combined use of LSTM and UNET modules, the TFBS model is a more suitable deep learning method than LSTM or UNET to perform crop recognition based on time-series SAR images. Moreover, TFBS also shows a better classification result than ConvLSTM (Fig. 9). ConvLSTM uses a convolution network instead of a full connection layer in traditional LSTM; thus, it can better extract spatiotemporal information. However, unlike clouds, water cover, or other moving targets, the change in crop plots over time is revealed spectrally rather than spatially. Therefore, to perform fine crop classification based on time-series SAR images, extracting the temporal features of each pixel first followed by the extraction of the spatial features might be more optimal.

B. Temporal Generalizability

The dataset of ARMSMOTN in 2019 is used to train TFBS, LSTM, UNET, and ConvLSTM models. Then, the trained models are used to produce the rice maps of ARMSMOTN in 2017 and 2018. CDL data were used as the reference data. The accuracies of the different models are illustrated in Fig. 10, which shows that all four models have certain levels of temporal generalizability. The TFBS model achieved the highest F-score and kappa score, implying that it is appropriate to apply the trained TFBS model to produce crop maps of the same region for other years. The ConvLSTM model achieved the second highest accuracy for both years, which shows that using spatiotemporal information performs better than that using only temporal or spatial information in crop classification based on SAR time series.

C. Spatial–Temporal Generalizability

To assess the spatial generalizability of the TFBS model compared with other models, the models trained with the ARMSMOTN dataset in 2019 are applied to northern LA in 2019. In addition, we also assessed the prediction accuracy of the trained model to northern LA in 2017 and 2018 to evaluate the generalizability of each model for different places and different years. The accuracies of all four models for 2017–2019 are shown in Fig. 11.

Fig. 12 shows the time-series VV and VH curves of rice fields based on CDL, indicating that the time-series VV and VH curves of northern LA in 2019 are similar to those of ARMSMOTN in 2019, which can explain why the accuracies of the three models in 2019 are relatively high. The time-series VV and VH curves of rice in northern LA in 2017 and 2018 were slightly different from those of ARMSMOTN in 2019, especially in early April. The extremely low backscatter coefficients at VV and VH indicate that in early April of 2017 and 2018, most of the rice fields in northern LA were already flooded, while not until late April were the rice fields in ARMSMOTN flooded. The differences in the rice phenological period in 2017 and 2018 can be used to test the temporal generalizability of each model. The kappa and F-scores of TFBS for all three years are substantially higher than those of the LSTM and UNET models (Fig. 11), clearly indicating that the spatial and temporal generalizability of TFBS is superior to those of the other models, especially for a different place and year. ConvLSTM achieves the second highest kappa and F-scores in all three years, indicating the advantages

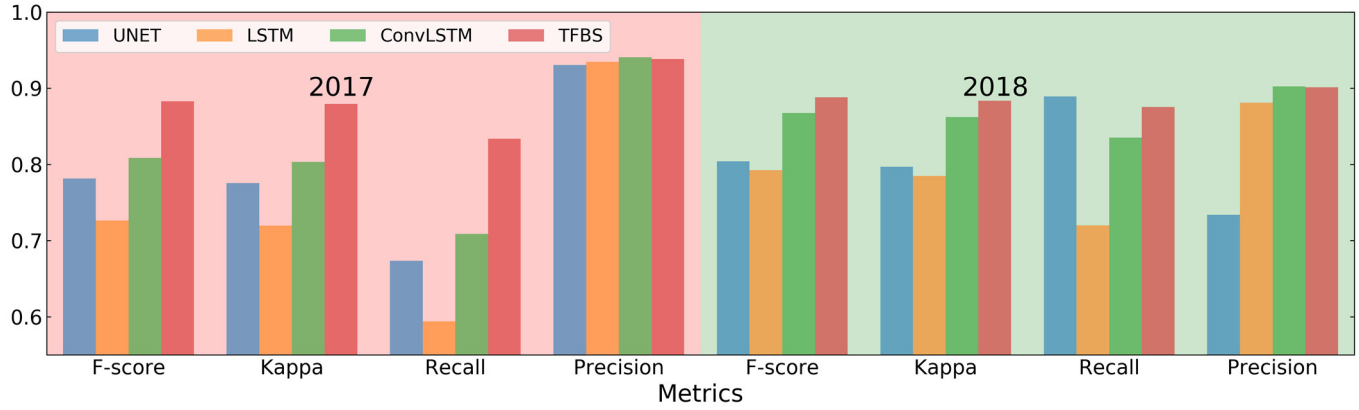


Fig. 10. Illustration of classification accuracies of UNET, LSTM, ConvLSTM, and TFBS models in 2017 (with red background) and 2018 (with green background). All the four models are trained with the ARMSMOTN dataset in 2019. CDL data is used as reference data to assess the accuracies.

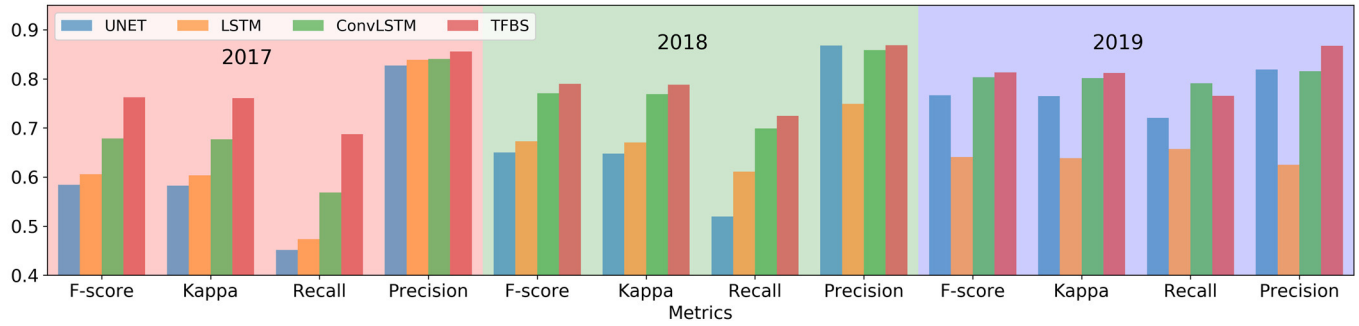


Fig. 11. F-score, kappa coefficient, precision, and recall of UNET, LSTM, ConvLSTM, and TFBS models trained by the dataset of ARMSMOTN in 2019 and tested by the dataset of northern LA in 2017–2019.

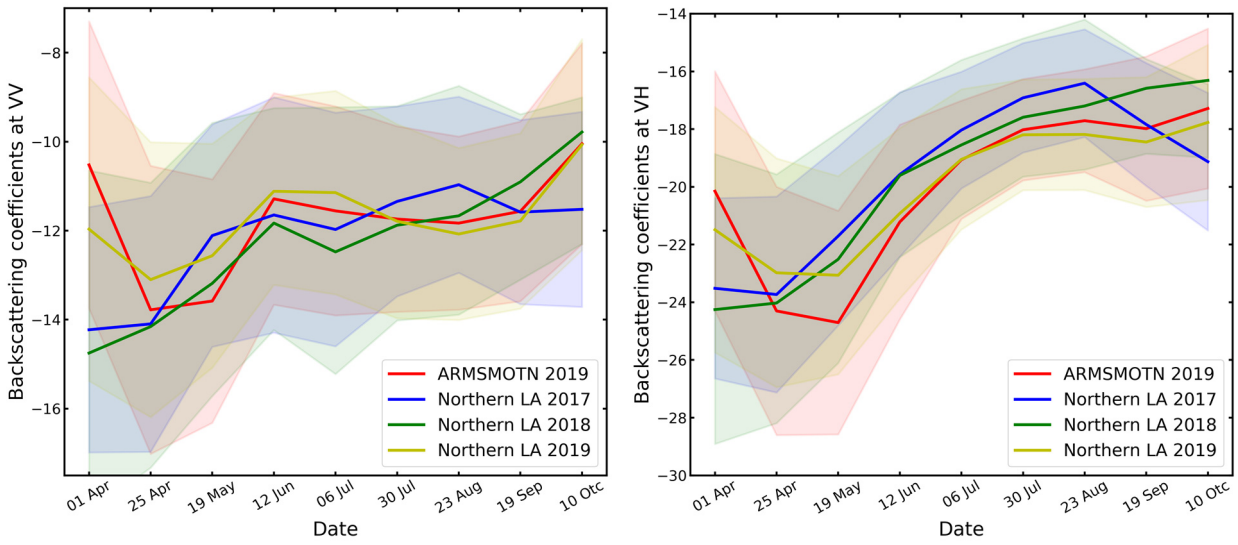


Fig. 12. Time-series curves of backscattering coefficients at VV and VH of rice for different areas and years. The lines in the figures refer to the mean backscatter coefficient values of all rice pixels at different areas and different years. The light buffer areas along the lines refer to one standard deviation from the mean value.

of the integrated use of spatial and temporal information in crop mapping based on time-series SAR images. It also shows that the accuracy of the UNET model declined sharply in 2017 and 2018 compared with 2019, which indicates that the spatial–temporal generalizability of UNET is poorer than that of the temporal feature-based classification models. LSTM, ConvLSTM, and TFBS are capable of learning deep temporal features from the raw time-series images; hence, they are capable of handling the subtle variability of the phenophase of rice in different years (Fig. 12).

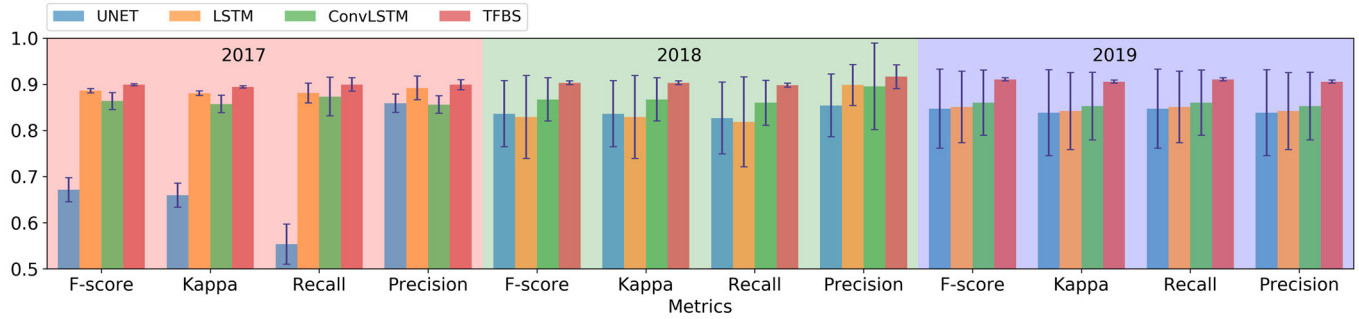


Fig. 13. Rice mapping accuracies of the fine-tuned deep learning models. The models were pretrained by ARMSMOTN 2019 dataset and transferred to SV based on fine-tuning technology. The CDL data of SV from 2017 to 2019 were used as reference data.

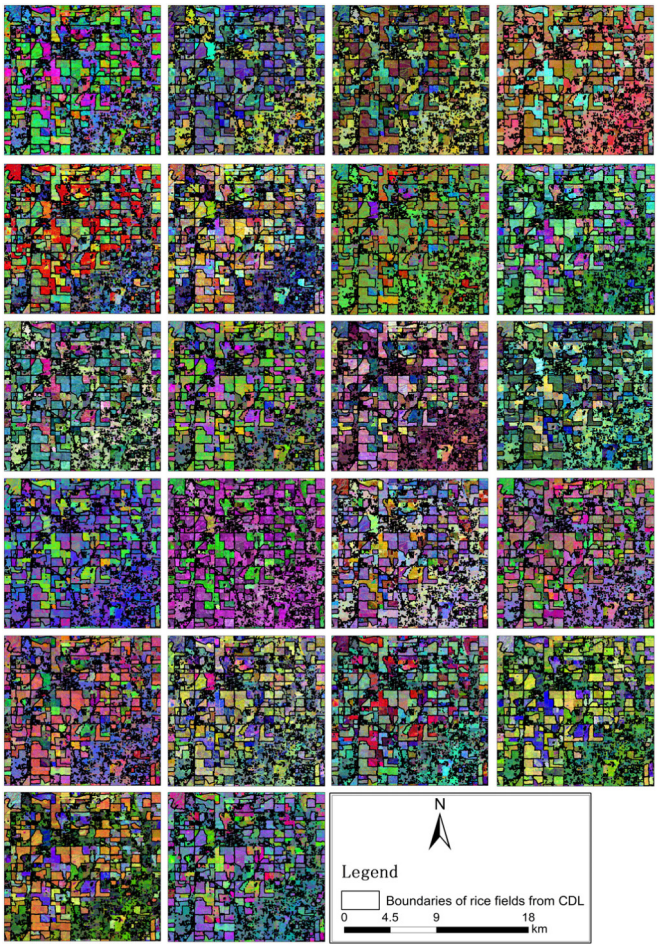


Fig. 14. Illustration of 64 deep temporal features extracted from raw time-series SAR images. Each image is composed of three temporal features in sequence as R/G/B-bands. The last image is composited of the 64th feature, the first feature, and the second feature. Polygons with black boundary denote rice fields from CDL.

D. Rice Mapping Based on Fine-Tuned Deep Learning Models

Four deep learning models were pretrained by ARMSMOTN 2019 dataset and then applied to the SV based on fine-tuning method for rice mapping from 2017 to 2019. CDL data were used to evaluate the classification accuracy from each model. Fig. 13 shows that TFBS significantly outperforms UNET, LSTM, and ConvLSTM in all three years

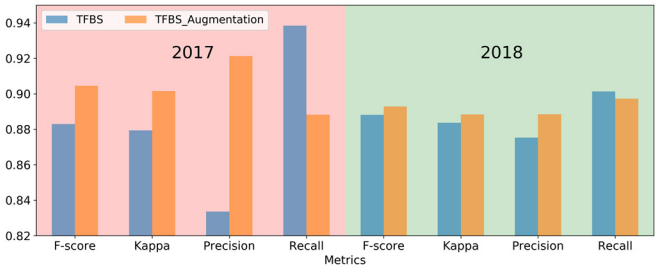


Fig. 15. Accuracies of TFBS models before (blue bars) and after data augmentation (orange bars). Dataset from ARMSMOTN in 2019 is used as training data, and the ARMSMOTN datasets from 2017 (red background) and 2018 (green background) are used as test data.

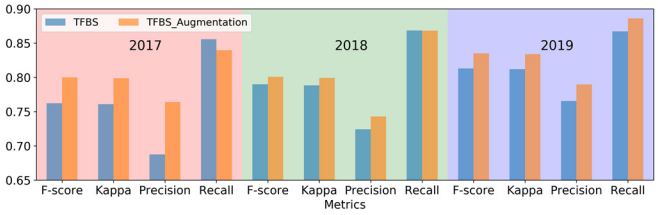


Fig. 16. Accuracies of TFBS models before (blue bars) and after data augmentation (orange bars). Dataset from ARMSMOTN in 2019 is used as training data, and the northern LA datasets from 2017 (red background), 2018 (green background), and 2019 (blue background) are used as test data.

in SV's rice mapping, the average F-score of which is 15.2%, 5.7%, and 4.7% higher than that of UNET, LSTM, and ConvLSTM, respectively. Meanwhile, the stability of TFBS model is also higher than the other three of which shows a much lower standard deviation of classification accuracy. The results show that it is feasible to apply a pretrained model to areas with different rice phenology for rice mapping based on fine-tuning technology and a small number of samples.

E. Exploring the Temporal Features Extracted From Raw Time-Series Images

Although the temporal features extracted by the LSTM matrix module of the TFBS model are abstract long- or short-term dependencies of rice fields existing in the time-series SAR images, they can still be visualized to explore how they work based on qualitative and quantitative analysis. An example of the intermediate temporal features mined from the raw time-series SAR images is shown in Fig. 14. The separability indicators of rice in the raw time-series SAR

TABLE II

SEPARABILITY INDICATORS OF RICE IN RAW TIME-SERIES SAR IMAGES AND INTERMEDIATE TEMPORAL FEATURES MINED BY TFBS MODEL

	ARMSMOTN			Northern LA			Mean	Std.
	2017	2018	2019	2017	2018	2019		
J-M values of raw time-series images	1.34	1.36	1.22	1.43	1.39	1.36	1.350	0.065
J-M values of intermediate temporal features	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0
TD of raw time-series images	1.46	1.47	1.37	1.57	1.54	1.5	1.485	0.064
TD of intermediate temporal features	2.00	2.00	2.00	2.00	2.00	2.00	2.00	0

J-M represents Jeffries-Matusita and TD represents Transformed Divergence indicator. ARMSMOTN represents Arkansas (AR), Mississippi (MS), southern Missouri (MO) and western Tennessee (TN). LA represents Louisiana.

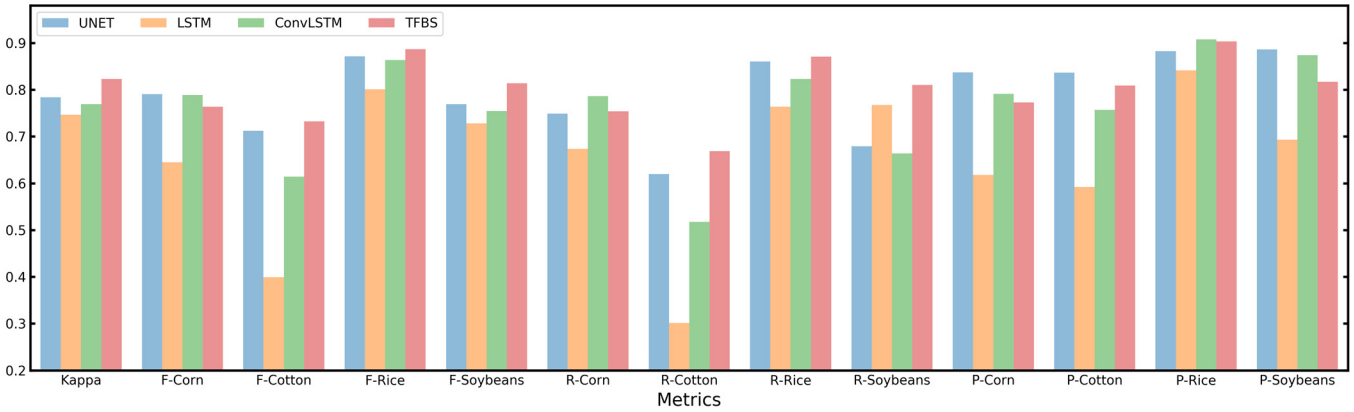


Fig. 17. Classification accuracies of UNET, LSTM, TFBS, and ConvLSTM models. The models are trained with the augmented ARMSMOTN 2019 dataset and tested with the ARMSMOTN 2018 dataset. F represents the F-score, R represents the recall, and P represents precision.

images and the intermediate temporal features are shown in Table II.

Fig. 14 clearly demonstrates the high degree of intra-class homogeneity of the rice fields and the high degree of interclass heterogeneity between the rice fields and other categories. Table II shows that the separability of temporal features is much higher than that of the raw time-series SAR images. Both the J–M distance and TD value obtained from the extracted temporal features reached a maximum of 2.0, outperforming the values for the raw SAR images, which were 1.350 and 1.485, respectively. Taking advantage of these abundant intermediate temporal features mined from the raw time-series images, the TFBS model is capable of producing higher-accuracy segmentation maps than the UNET model.

F. Improvement of Spatial–Temporal Generalizability of TFBS by Data Augmentation

Data from ARMSMOTN 2019 were used as the training dataset. The raw training dataset contained 6607 image tiles (each consisting of 128 × 128 pixels), whereas the augmented training dataset contained 19 821 image tiles. The results show that data augmentation can significantly improve the spatial and temporal generalizability of the TFBS model (Figs. 15 and 16). The prediction accuracy was significantly improved after the augmentation of training data in both ARMSMOTN and northern LA, except in 2019.

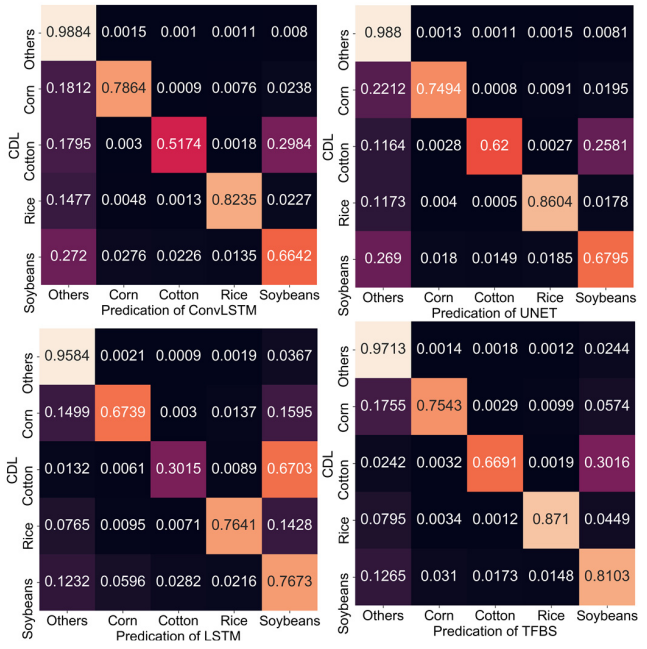


Fig. 18. Confusion matrixes of UNET, LSTM, TFBS, and ConvLSTM models. The models are trained by the augmented ARMSMOTN 2019 dataset and tested by the ARMSMOTN 2018 dataset.

The kappa coefficient increased by an average of 2.40% after augmentation. The F-score increased by an average of 2.43% after augmentation.

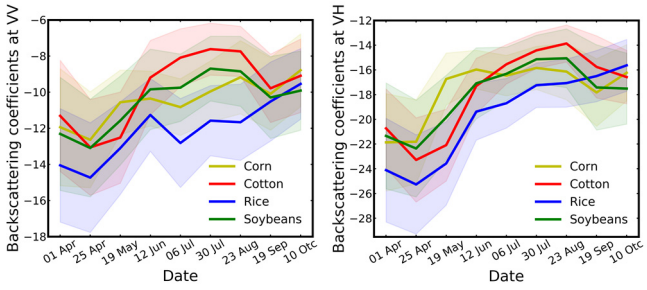


Fig. 19. Time-series curves of backscattering coefficients at VV and VH of corn, cotton, rice, and soybeans in ARMSMOTN in 2018. The lines refer to the mean backscatter coefficient values of all pixels of different crops in the study area. The light buffer areas along the lines refer to one standard deviation from the mean value.

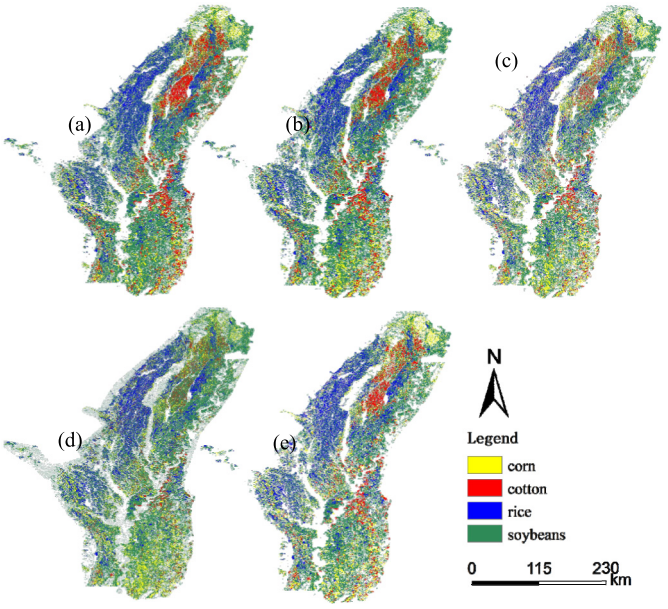


Fig. 20. (a) CDL and multicrop classification results produced by (b) TFBS, (c) ConvLSTM, (d) LSTM, and (e) UNET models in ARMSMOTN in 2018. ARMSMOTN represents the region of AR, MS, southern MO, and western TN.

G. Performance of TFBS Model on Multicrop Classification

Figs. 17 and 18 indicate that TFBS significantly outperforms the UNET, LSTM, and ConvLSTM models for multiclass classification. The kappa value of the TFBS model reaches 0.8230, whereas those for UNET, LSTM, and ConvLSTM reach 0.7840, 0.7465, and 0.7691, respectively. The TFBS model also achieved the highest accuracy for three of the four crops—cotton, rice, and soybeans. Among all four crops, the classification accuracy of rice is significantly higher than that of the other three crops, which is mainly due to the easy recognition of irrigation characteristics during the rice transplanting period. During the entire growth period, the VV and VH values of rice are generally lower than those of other crops (Fig. 19). Cotton is easy to confuse with soybeans, resulting in low accuracy. This is mainly because the time series curve of cotton is very close to that of soybeans (Fig. 19). The CDL data and multicrop classification results of each model are shown in Fig. 20. Based on the above analysis, multi-crop classification based on time-series SAR and the TFBS binary classification model is feasible and can obtain high classification accuracy.

TABLE III
TRAINING AND PREDICTION TIME OF FOUR DEEP LEARNING MODELS

	Training	Prediction
UNET	2,883 s	97 s
LSTM	43,120 s	15,709 s
ConvLSTM	15,199 s	376 s
TFBS	4,426 s	164 s

Dataset from ARMSMOTN in 2019 was used for training and dataset from ARMSMOTN in 2018 was used for prediction.

H. Comparison of the Efficiency of Each Model

Table III shows that UNET has the highest efficiency compared with other models, followed by TFBS, and then ConvLSTM, while LSTM is significantly higher than other models in both training and prediction time. TFBS has relatively high efficiency and the best performance in classification accuracy, which is significantly great for large-area crop mapping.

V. CONCLUSION

In this study, a novel TFBS model is proposed to perform paddy rice classification based on the combined use of temporal features from raw time-series SAR images and their spatial contextual information. The performance of the TFBS model was evaluated using the time-series Sentinel-1 VV and VH images acquired over the study area in AR, MS, southern MO, western TN, northern LA, and SV, together with three state-of-the-art deep learning models (LSTM, UNET, and ConvLSTM). The results show that the TFBS model outperforms LSTM, UNET, and ConvLSTM models in the study area. The kappa, F-score, precision, and recall scores of the TFBS model are significantly higher than those of the other models. The abundant temporal features mined from raw time-series SAR images by the LSTM module of TFBS achieve higher J–M and TD values than raw images, which makes TFBS more suitable for time-series image segmentation than UNET. TFBS also shows the best spatial and temporal generalizability when the trained models are applied to a different year or place. This study also evaluates the effect of data augmentation on the accuracy of the TFBS model, revealing that the generalizability of the TFBS model demonstrates substantial improvement after data augmentation, showing higher accuracies than the TFBS model without data augmentation. TFBS also shows the best performance in multiclass classification in the study area, showing the potential of the TFBS model in multicrop classification based on time-series SAR data.

Several aspects still need to be investigated based on the current study. It would be useful to apply the TFBS model trained with CDL data to other countries/regions plagued by insufficient sample data. Data augmentation technology and fine-tuning method are two feasible ways to achieve that goal. Meanwhile, with the increase in available remote

sensing data, the temporal resolution of time series data will be further improved, and multisource satellite data will be used to improve the ability of large-scale crop mapping based on deep learning technology.

APPENDIX

The code of the TFBS model used in this study together with the dataset of ARMSMOTN in 2019 are available at https://github.com/younglimpo/TFBSmodel.

ACKNOWLEDGMENT

The authors would like to thank USDA-NASS for providing the CDL dataset. They also thank the anonymous reviewers for their constructive comments and advice.

REFERENCES

[1] H. C. J. Godfray *et al.*, "Food security: The challenge of feeding 9 billion people," *Science*, vol. 327, no. 5967, pp. 812–818, Feb. 2010.

[2] D. Boddiger, "Boosting biofuel crops could threaten food security," *Lancet*, vol. 370, no. 9591, pp. 923–924, Sep. 2007.

[3] C. C. Funk and M. E. Brown, "Declining global per capita agricultural production and warming oceans threaten food security," *Food Secur.*, vol. 1, no. 3, pp. 271–289, Sep. 2009.

[4] *The State of Food Security and Nutrition in the World 2020. Transforming Food Systems for Affordable Healthy Diets*, FAO, IFAD, UNICEF, WFP, and WHO, Rome, Italy, 2020.

[5] P. Thenkabail, M. Hanjra, V. Dheeravath, and M. Gumma, "A holistic view of global croplands and their water use for ensuring global food security in the 21st century through advanced remote sensing and non-remote sensing approaches," *Remote Sens.*, vol. 2, no. 1, pp. 211–261, Jan. 2010.

[6] M. E. Brown, "Remote sensing technology and land use analysis in food security assessment," *J. Land Use Sci.*, vol. 11, no. 6, pp. 623–641, Nov. 2016.

[7] Y. Xu *et al.*, "Tracking annual cropland changes from 1984 to 2016 using time-series Landsat images with a change-detection and post-classification approach: Experiments from three sites in Africa," *Remote Sens. Environ.*, vol. 218, pp. 13–31, Dec. 2018.

[8] D. B. Lobell, D. Thau, C. Seifert, E. Engle, and B. Little, "A scalable satellite-based crop yield mapper," *Remote Sens. Environ.*, vol. 164, pp. 324–333, Jul. 2015.

[9] L. Yang, L. Wang, J. Huang, L. R. Mansaray, and R. Mijiti, "Monitoring policy-driven crop area adjustments in northeast China using Landsat-8 imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 82, Oct. 2019, Art. no. 101892.

[10] Q. Yuan *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, May 2020, Art. no. 111716.

[11] X. X. Zhu *et al.*, "Deep learning in remote sensing: A comprehensive review and list of resources," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 4, pp. 8–36, Dec. 2017.

[12] M. Boschetti *et al.*, "PhenoRice: A method for automatic extraction of spatio-temporal information on rice crops using satellite data time series," *Remote Sens. Environ.*, vol. 194, pp. 347–365, Jun. 2017.

[13] G. Azzari and D. B. Lobell, "Landsat-based classification in the cloud: An opportunity for a paradigm shift in land cover monitoring," *Remote Sens. Environ.*, vol. 202, pp. 64–74, Dec. 2017.

[14] X. Xiao *et al.*, "Mapping paddy rice agriculture in southern China using multi-temporal MODIS images," *Remote Sens. Environ.*, vol. 95, no. 4, pp. 480–492, 2005.

[15] J. Dong *et al.*, "Mapping paddy rice planting area in northeastern Asia with Landsat 8 images, phenology-based algorithm and Google Earth Engine," *Remote Sens. Environ.*, vol. 185, pp. 142–154, Nov. 2016.

[16] L. Yang, L. Mansaray, J. Huang, and L. Wang, "Optimal segmentation scale parameter, feature subset and classification algorithm for geographic object-based crop recognition using multisource satellite imagery," *Remote Sens.*, vol. 11, no. 5, p. 514, Mar. 2019.

[17] J. E. Ball, D. T. Anderson, and C. S. Chan, "Comprehensive survey of deep learning in remote sensing: Theories, tools, and challenges for the community," *J. Appl. Remote Sens.*, vol. 11, no. 4, Sep. 2017, Art. no. 042609.

[18] W. Li, Z. Niu, R. Shang, Y. Qin, L. Wang, and H. Chen, "High-resolution mapping of forest canopy height using machine learning by coupling ICESat-2 LiDAR with Sentinel-1, Sentinel-2 and Landsat-8 data," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 92, Oct. 2020, Art. no. 102163.

[19] W. Zhao, Z. Guo, J. Yue, X. Zhang, and L. Luo, "On combining multiscale deep learning features for the classification of hyperspectral remote sensing imagery," *Int. J. Remote Sens.*, vol. 36, no. 13, pp. 3368–3379, Jul. 2015.

[20] T. Liu, A. Abd-Elrahman, J. Morton, and V. L. Wilhelm, "Comparing fully convolutional networks, random forest, support vector machine, and patch-based deep convolutional neural networks for object-based wetland mapping using images from small unmanned aircraft system," *GISci. Remote Sens.*, vol. 55, no. 2, pp. 243–264, 2018.

[21] J. Xu *et al.*, "DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping," *Remote Sens. Environ.*, vol. 247, Sep. 2020, Art. no. 111946.

[22] H. Zhao, Z. Chen, H. Jiang, W. Jing, L. Sun, and M. Feng, "Evaluation of three deep learning models for early crop classification using Sentinel-1A imagery time series—A case study in Zhanjiang, China," *Remote Sens.*, vol. 11, no. 22, p. 2673, Nov. 2019.

[23] M. Reichstein *et al.*, "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, pp. 195–204, Feb. 2019.

[24] D. Chai, S. Newsam, H. K. Zhang, Y. Qiu, and J. Huang, "Cloud and cloud shadow detection in Landsat imagery based on deep convolutional neural networks," *Remote Sens. Environ.*, vol. 225, pp. 307–316, May 2019.

[25] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, "Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN)," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111446.

[26] Z. Wu, X. Wang, Y. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. ICM*, Brisbane, QLD, Australia, 2015, pp. 461–470.

[27] N. Lang, K. Schindler, and J. D. Wegner, "Country-wide high-resolution vegetation height mapping with Sentinel-2," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111347.

[28] X.-Y. Tong *et al.*, "Land-cover classification with high-resolution remote sensing images using transferable deep models," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111322.

[29] C. Zhang, P. A. Harrison, X. Pan, H. Li, I. Sargent, and P. M. Atkinson, "Scale sequence joint deep learning (SS-JDL) for land use and land cover classification," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111593.

[30] J. H. Jeppesen, R. H. Jacobsen, F. Inceoglu, and T. S. Toftegaard, "A cloud detection algorithm for satellite imagery based on deep learning," *Remote Sens. Environ.*, vol. 229, pp. 247–259, May 2019.

[31] N. He, L. Fang, and A. Plaza, "Hybrid first and second order attention UNet for building segmentation in remote sensing images," *Sci. China Inf. Sci.*, vol. 63, no. 4, Mar. 2020, Art. no. 140305.

[32] X. Yang, X. Li, Y. Ye, R. Y. K. Lau, X. Zhang, and X. Huang, "Road detection and centerline extraction via deep recurrent convolutional neural network U-Net," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 7209–7220, Sep. 2019.

[33] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet++," *Remote Sens.*, vol. 11, no. 11, p. 1382, 2019.

[34] L. Jiao, L. Huo, C. Hu, and P. Tang, "Refined UNet: UNet-based refinement network for cloud and shadow precise segmentation," *Remote Sens.*, vol. 12, no. 12, p. 2001, Jun. 2020.

[35] Z. Du, J. Yang, C. Ou, and T. Zhang, "Smallholder crop area mapped with a semantic segmentation deep learning method," *Remote Sens.*, vol. 11, no. 7, p. 888, Apr. 2019.

[36] Zhao *et al.*, "Use of unmanned aerial vehicle imagery and deep learning UNet to extract rice lodging," *Sensors*, vol. 19, no. 18, p. 3859, Sep. 2019.

[37] C. Xiao, N. Chen, C. Hu, K. Wang, J. Gong, and Z. Chen, "Short and mid-term sea surface temperature prediction using time-series satellite data and LSTM-AdaBoost combination approach," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111358.

[38] C. Liu, J. Liu, J. Wang, S. Xu, H. Han, and Y. Chen, "An attention-based spatiotemporal gated recurrent unit network for point-of-interest recommendation," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 8, p. 355, Aug. 2019.

[39] A. Sharma, X. Liu, and X. Yang, "Land cover classification from multi-temporal, multi-spectral remotely sensed imagery using patch-based recurrent neural networks," *Neural Netw.*, vol. 105, pp. 346–355, Sep. 2018.

[40] Q. Liu, F. Zhou, R. Hang, and X. Yuan, "Bidirectional-convolutional LSTM based spectral-spatial feature learning for hyperspectral image classification," *Remote Sens.*, vol. 9, no. 12, p. 1330, Dec. 2017.

[41] K. Bakhti, K. Djerriri, M. E. A. Arabi, S. Chaib, and M. S. Karoui, "Improvememt of multi-temporal vegetation modeling using hybrid deep neural networks of multispectral remote sensing images," in *Proc. IGARSS*, Yokohama, Japan, 2019, pp. 1–4.

[42] L. Zhong, L. Hu, and H. Zhou, "Deep learning based multi-temporal crop classification," *Remote Sens. Environ.*, vol. 221, pp. 430–443, Feb. 2019.

[43] F. Karim, S. Majumdar, H. Darabi, and S. Chen, "LSTM fully convolutional networks for time series classification," *IEEE Access*, vol. 6, pp. 1662–1669, 2017.

[44] Q. Wang, S. Liu, J. Chanussot, and X. Li, "Scene classification with recurrent attention of VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 2, pp. 1155–1167, Feb. 2019.

[45] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[46] Y. Zhou, J. Luo, L. Feng, Y. Yang, Y. Chen, and W. Wu, "Long-short-term-memory-based crop classification using high-resolution optical images and multi-temporal SAR data," *GISci. Remote Sens.*, vol. 56, no. 8, pp. 1170–1191, Nov. 2019.

[47] J. G. A. Barbedo, "Impact of dataset size and variety on the effectiveness of deep learning and transfer learning for plant disease classification," *Comput. Electron. Agricult.*, vol. 153, pp. 46–53, Oct. 2018.

[48] M. Vakalopoulou, K. Karantzalos, N. Komodakis, and N. Paragios, "Building detection in very high resolution multispectral data with deep learning features," in *Proc. IGARSS*, Milan, Italy, 2015, pp. 1873–1876.

[49] 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, Mar. 2017.

[50] D. Ienco, R. Interdonato, R. Gaetano, and D. H. T. Minh, "Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture," *ISPRS J. Photogramm. Remote Sens.*, vol. 158, pp. 11–22, Dec. 2019.

[51] L. Zhong, L. Hu, H. Zhou, and X. Tao, "Deep learning based winter wheat mapping using statistical data as ground references in Kansas and northern Texas, U.S," *Remote Sens. Environ.*, vol. 233, Nov. 2019, Art. no. 111411.

[52] Z. Sun, L. Di, and H. Fang, "Using long short-term memory recurrent neural network in land cover classification on Landsat and cropland data layer time series," *Int. J. Remote Sens.*, vol. 40, no. 2, pp. 593–614, Jan. 2019.

[53] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019.

[54] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *Convolutional Neural Netw. Vis. Recognit*, vol. 11, pp. 1–8, Dec. 2017.

[55] W. Li, C. Chen, M. Zhang, H. Li, and Q. Du, "Data augmentation for hyperspectral image classification with deep CNN," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 593–597, Apr. 2019.

[56] Y. Yan, Z. Tan, and N. Su, "A data augmentation strategy based on simulated samples for ship detection in RGB remote sensing images," *ISPRS Int. J. Geo-Inf.*, vol. 8, no. 6, p. 276, Jun. 2019.

[57] N. Teimouri, M. Dyrmann, and R. N. Jørgensen, "A novel spatio-temporal FCN-LSTM network for recognizing various crop types using multi-temporal radar images," *Remote Sens.*, vol. 11, no. 8, p. 990, Apr. 2019.

[58] M. M. G. de Macedo, A. B. Mattos, and D. A. B. Oliveira, "Generalization of convolutional LSTM models for crop area estimation," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 1134–1142, Mar. 2020.

[59] J. Sun, L. Di, Z. Sun, Y. Shen, and Z. Lai, "County-level soybean yield prediction using deep CNN-LSTM model," *Sensors*, vol. 19, no. 20, p. 4363, Oct. 2019.

[60] N. Childs, "U.S. Outlook for 2008/09," in *Rice Situation and Outlook Yearbook*. Collingdale, PA, USA: Diane Publishing, 2009, pp. 6–17.

[61] C. R. Adair, "Distribution of rice in the United States," in *Rice in the United States: Varieties and Production: Agricultural Research Service*. Washington, DC, USA: U.S. Department of Agriculture, 1973, pp. 3–4.

[62] *USDA National Agricultural Statistics Service Cropland Data Layer Published Crop-Specific Data Layer*, NASS, Manama, Bahrain, 2020.

[63] *Arkansas Rice Production Handbook*, Univ. Arkansas Division Agricult., Little Rock, AR, USA, 2013, pp. 1–20. [Online]. Available: https://www.uaex.edu/publications/pdf/mp192/mp192.pdf

[64] C. E. Wilson, Jr., S. K. Runsick, and R. Mazzanti, "Trends in Arkansas rice production," *Res. Ser.*, vol. 550, pp. 13–22, Aug. 2007.

[65] M. Shipp, "Rice crop timeline for the southern states of Arkansas, Louisiana, and Mississippi," NSF Center Integr. Pest Manage., Raleigh, NC, USA, Tech. Rep., 2005, pp. 1–67. [Online]. Available: https://ipmdata.ipmcenters.org/documents/timelines/Rice.pdf

[66] J. E. Hill, J. F. Williams, R. G. Mutters, and C. A. Greer, "The California rice cropping system: Agronomic and natural resource issues for long-term sustainability," *Paddy Water Environ.*, vol. 4, no. 1, pp. 13–19, Mar. 2006.

[67] R. Torres *et al.*, "GMES Sentinel-1 mission," *Remote Sens. Environ.*, vol. 120, pp. 9–24, May 2012.

[68] C. Yang *et al.*, "Ground deformation revealed by Sentinel-1 MSBAS-InSAR time-series over Karamay oilfield, China," *Remote Sens.*, vol. 11, no. 17, p. 2027, Aug. 2019.

[69] D. Geudtner, R. Torres, P. Snoeij, M. Davidson, and B. Rommen, "Sentinel-1 system capabilities and applications," in *Proc. IEEE GARSS*, Quebec City, QC, Canada, Jul. 2014, pp. 1457–1460.

[70] S. Abdikan, F. B. Sanli, M. Ustuner, and F. Calò, "Land cover mapping using Sentinel-1 SAR data," *ISPRS Arch.*, vol. 41, p. 757, Feb. 2016.

[71] A. A. Bayanudin and R. H. Jatmiko, "Orthorectification of Sentinel-1 SAR (Synthetic Aperture Radar) data in some parts of South-eastern Sulawesi using Sentinel-1 toolbox," in *Proc. IOP Conf. Ser. Earth Environ. Sci.*, Yogyakarta, Indonesia, 2016, Art. no. 012007.

[72] N. Gorelick, M. Hancher, M. Dixon, S. Ilyushchenko, D. Thau, and R. Moore, "Google Earth Engine: Planetary-scale geospatial analysis for everyone," *Remote Sens. Environ.*, vol. 202, pp. 18–27, Dec. 2017.

[73] L. Kumar and O. Mutanga, "Google Earth Engine applications since inception: Usage, trends, and potential," *Remote Sens.*, vol. 10, no. 10, p. 1509, Sep. 2018.

[74] C. M. Arellano *et al.*, "Multi-temporal analysis of dense and sparse forests' radar backscatter using Sentinel-1A collection in Google Earth Engine," *ISPRS Arch.*, vol. XLII-4/W19, pp. 23–30, Nov. 2019.

[75] C. Boryan, Z. Yang, R. Mueller, and M. Craig, "Monitoring U.S. agriculture: The U.S. Department of Agriculture, National Agricultural Statistics Service, Cropland Data Layer Program," *Geocarto Int.*, vol. 26, no. 5, pp. 341–358, 2011.

[76] Y. Shao, R. S. Lunetta, B. Wheeler, J. S. Iiames, and J. B. Campbell, "An evaluation of time-series smoothing algorithms for land-cover classifications using MODIS-NDVI multi-temporal data," *Remote Sens. Environ.*, vol. 174, pp. 258–265, Mar. 2016.

[77] S. Wang, W. Chen, S. M. Xie, G. Azzari, and D. B. Lobell, "Weakly supervised deep learning for segmentation of remote sensing imagery," *Remote Sens.*, vol. 12, no. 2, p. 207, Jan. 2020.

[78] X. Shi, Z. Chen, H. Wang, D. Yeung, W. Wong, and W. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Proc. Adv. Neural Inf. Process Syst.*, Istanbul, Turkey, 2015, pp. 802–810.

[79] H.-C. Li, W.-S. Hu, W. Li, J. Li, Q. Du, and A. Plaza, "A^3CLNN: Spatial, spectral and multiscale attention ConvLSTM neural network for multisource remote sensing data classification," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Oct. 21, 2020, doi: 10.1109/TNNLS.2020.3028945.

[80] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Convolutional neural networks for large-scale remote-sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 2, pp. 645–657, Feb. 2017.

[81] J. A. Richards and J. A. Richards, *Remote Sensing Digital Image Analysis*. Berlin, Germany: Springer, 1999, pp. 10–38.

[82] M. Dabboor, S. Howell, M. Shokr, and J. Yackel, "The Jeffries–Matusita distance for the case of complex Wishart distribution as a separability criterion for fully polarimetric SAR data," *Int. J. Remote Sens.*, vol. 35, no. 19, pp. 6859–6873, Oct. 2014.

[83] J. Sim and C. C. Wright, "The kappa statistic in reliability studies: Use, interpretation, and sample size requirements," *Phys. Therapy*, vol. 85, no. 3, pp. 257–268, Mar. 2005.

Lingbo Yang received the B.S. degree in geographic information system from the Hefei University of Technology, Hefei, China, in 2010. He is pursuing the Ph.D. degree with Zhejiang University, Hangzhou, China.

He was with the Food and Agriculture Organization of the United Nations (FAO), Rome, Italy, as a Geospatial Specialist in 2019. His research interests include geographic information systems (GISs), remote sensing, deep learning, and their applications in agriculture.

Ran Huang received the Ph.D. degree from the College of Information and Electrical Engineering, China Agricultural University, Beijing, China, in 2019.

She is an Assistant Professor with the School of Artificial Intelligence, Hangzhou Dianzi University, Hangzhou, Zhejiang, China. Her main research interests include agricultural remote sensing, agrometeorological disaster, and crop classification.

Jingfeng Huang was born in Zhangzhou, China, in 1963. He received the B.S. and M.S. degrees in applied meteorology from the Nanjing Institute of Meteorology, Nanjing, China, in 1985 and 1990, respectively, and the Ph.D. degree in applied remote sensing from Zhejiang University, Hangzhou, China, in 2000.

From 1985 to 1997, he was a Research Scientist with the Meteorological Institute of Xinjiang, Ürümqi, China. Since 2001, he has been a Professor with the College of Environmental and Resource Sciences, Zhejiang University. He has authored 8 books and over 250 research articles. His research interests include remote sensing applications in natural resources and the environment.

Tao Lin received the M.S. and Ph.D. degrees in agricultural and biological engineering from the University of Illinois at Urbana–Champaign Champaign, IL, USA, in 2010 and 2013, respectively.

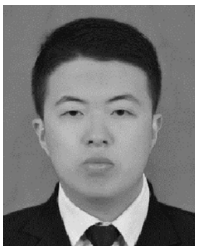
He joined the faculty of the Biosystems Engineering Department, 100 Talents Program, Zhejiang University, Hangzhou, China. He has published more than ten peer-reviewed journal articles. His research focuses on agricultural big data systems informatics and analytics, ranging from spatiotemporal analysis, geographic information system (GIS), optimization modeling analysis, to high performance cyberinfrastructure enabled decision support systems.

Limin Wang was born in Inner Mongolia, China, in 1968. He received the B.S. degree in zoology from Inner Mongolia University, Huhhot, China, in 1989, the M.S. degree in feed science from the Chinese Academy of Agricultural Science (CAAS), Beijing, China, in 1992, and the Ph.D. degree in ecology from Inner Mongolia University, in 1997.

From 1992 to 1994, he was a Research Scientist with the Chinese Academy of Tropical Agricultural Sciences, Danzhou, China. From 1997 to 2002, he was an Associate Professor with the Inner Mongolia University. Since 2003, he has been a Senior Researcher with the Institute of Agricultural Resources and Regional Planning, CAAS. He has published 5 books and over 50 research articles. His research interests include remote sensing applications in agricultural resources and sustainable agriculture.

Ruzemaimaiti Mijiti received the master's degree from the College of Resources and Environmental Science, Xinjiang University, Ürümqi, China, in 2018. He is pursuing the Ph.D. degree with the Institute of Applied Remote Sensing and Information Technology, Zhejiang University, Hangzhou, China.

His main research interests include agricultural remote sensing, especially remote sensing application in crop mapping and crop yield estimation.

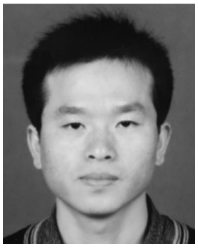


Pengliang Wei received the master's degree from the College of Mechanical and Electronic Engineering, Northwest A&F University, Xianyang, China, in 2019. He is pursuing the Ph.D. degree with the Institute of Applied Remote Sensing and Information Technology, Zhejiang University, Hangzhou, China.

His main research interests include radar remote sensing and polarimetric synthetic aperture radar signal processing.

Chao Tang is pursuing the master's degree with Zhejiang University, Zhejiang, China.

His research interests focus on the application of remote sensing and deep learning technologies in agriculture.



Jie Shao received the B.S. degree in geographic information system (GIS) from Anhui Normal University, Wuhu, China, in 2010. He is pursuing the Ph.D. degree in cartography and GIS with Beijing Normal University, Beijing, China.

His main research interests include radar remote sensing and polarimetric synthetic aperture radar signal processing.

Qiangzi Li received the M.S. degree in geoinformation science from Beijing Normal University, Beijing, China, in 1998, and the Ph.D. degree in cartography and geographic information system (GIS) from the Graduate University of Chinese Academy of Sciences, Beijing, in 2008.

He is an Associate Professor with the Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing. His research interests include crop acreage and yield estimation using remote sensing.

Xin Du received the Ph.D. degree in cartography and geographic information system (GIS) from the Graduate University of Chinese Academy of Sciences, Beijing, China, in 2010.

He is an Assistant Professor with Institute of Remote Sensing Applications, Chinese Academy of Sciences, Beijing. His research interests include biomass estimation using remote sensing data.