Remote Sensing of Environment 247 (2020) 111952

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

Deeply synergistic optical and SAR time series for crop dynamic monitoring

Wenzhi Zhao[a,b], Yang Qu[a,b,c,*], Jiage Chen[d], Zhanliang Yuan[c]

[a] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[b] Beijing Engineering Research Center for Global Land Remote Sensing Products, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China
[c] School of Surveying & Land Information Engineering, Henan Polytechnic University, Henan 454000, China
[d] National Geomatics Center of China, Beijing 100830, China

ARTICLE INFO

Keywords:
Multi-temporal remote sensing
Optical data
SAR data
Long-time series
Deep learning
CNN-RNN
Crops dynamic

ABSTRACT

Multi-temporal remote sensing imagery has been regarded as an effective tool to monitor cropland. But optical sensors often miss key stages for crop growth because of clouds, which poses challenges to many studies. The synergistic of SAR and optical data is expected to lift this problem, especially in areas with persistent cloud cover. However, due to the different characteristics of optical and SAR sensors, it is difficult to build a relationship between the two with most existing methods, let alone construct the long-time correlations to fill optic observation gaps using SAR data. Inspired by deep learning, this study presents a novel strategy to learn the relationship between optical and SAR time series based on the sequence of contextual information. To be specific, we extended the conventional CNN-RNN to build Multi-CNN-Sequence to Sequence (MCNN-Seq) model, and formulate the correlation between the optic and SAR time series sequences. We verified the MCNN-Seq model and found that the accuracy of the predicted optical image was determined by crop types and phenological stages, both in the spatial and temporal domain, respectively. For several crops, such as onion, winter wheat, corn, and sugar beet, our predictions are fitting well with R^2 0.9409, 0.9824,0.9157, and 0.9749, respectively. Compared to CNN and RNN, the simulation accuracy achieved by the MCNN-Seq model is much better in terms of R^2 and RMSE. In general, results demonstrate that deep learning models have the potential to synergize SAR and optical data and provide replaceable information when the optical data has a long data gap due to the persistent clouds.

1. Introduction

Cropland is the basic element for human survival and socio-economic development. Most challenges that urgently need to be resolved for humanity are related to crop production (Johnson et al., 2014). Timely and accurately acquiring crop information can improve crop management, and issues such as food production and food security. Remote sensing technology has the advantages of wide range monitoring, low cost, and short revisit periods. Thus, it provides a cost-effective, efficient, and reliable method for cropland management. Nowadays, multi-temporal optical remote sensing images have become important materials for crop information extraction (Dong et al., 2015; Wardlow and Egbert, 2008; Xiao et al., 2005).

During the past few decades, many methods have been developed to process and analyze multi-temporal optical remote sensing data. For this reason, the data have been successfully used in crop yield estimation(Claverie et al., 2017; Liu et al., 2010), forest disturbance detection (Frazier et al., 2015; White et al., 2017), land management and planning (Inglada et al., 2017). However, extensive accurate and spatially detailed crop monitoring over large areas was hampered by the lack of high resolution and dense Satellite Image Time Series (SITS). Fortunately, the resolution and repetition frequency of optical remote sensing data has greatly increased over the last decades. For instance, Sentinel-2a/b has a short revisiting period and provide a higher spatial resolution (10-20 m) than low or medium spatial satellites (Drusch et al., 2012). Therefore, current earth observation data provides the possibility to build the denser, higher-resolution SITS than before. Nevertheless, the limitation of optical imagers that rely on clear sky conditions has not been eliminated. In the task of crop monitoring, persistent cloud coverage drastically reduces the number of usable imageries in specific areas and unable to probe the development stages of the crop.

* Corresponding author at: State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing Science and Engineering, Faculty of Geographical Science, Beijing Normal University, Beijing 100875, China.
E-mail address: 211804020029@home.hpu.edu.cn (Y. Qu).

https://doi.org/10.1016/j.rse.2020.111952
Received 10 March 2020; Received in revised form 8 June 2020; Accepted 12 June 2020

To fill the observation gaps under bad weather conditions, temporal replacement methods (Zeng et al., 2013; Zhang et al., 2014; Zhang et al., 2010) and temporal filtering methods (Julien and Sobrino, 2010; Lu et al., 2007)are usually applied to multi-temporal remote sensing researches(Shen et al., 2015). In general, these methods assume that adjacent temporal images have the same vegetation type, which could get good results in areas within a short time interval. Unfortunately, cropland is a type of land cover that changes dynamically according to seasonal cycles and plant growth periods. Therefore, the assumption of adjacent temporal observations is not always applicable to crop research. To address these issues, some studies have been proposed to obtain more frequent and complete SITS by combining multi-sensor data that share similar observation properties (such as Landsat and Sentinel). However, due to the difference in spatial resolution for multi-sensor data, it is difficult to construct dense time-series image stack with data combination. To better utilize multi-sensor data, a common simple approach is to resample the high-resolution data to match the coarse-resolution data (Wang et al., 2017), such as NASA's Harmonized Landsat and Sentinel-2 (HLS) project. This approach is very helpful for researches that focused on the dynamic monitoring of large areas. However, in order to dynamically monitor the spatial-temporal evolving pattern of crop land, it is often require high spatial resolution satellite data, especially for small farmland (Ozdogan and Woodcock, 2006). In addition, all these methods are difficult to fill long data gaps (e.g., more than two months). Moreover, the reliability of these methods depends on the fraction of the cloud-free pixels. But, this prerequisite is hardly to satisfy especially for areas that covered by continuous clouds through plant growth periods.

Recently, Sentinel-1 Synthetic Aperture SAR (SAR) has attracted much attention due to its high revisit frequency and all-weather imaging capacity. A large number of studies have already demonstrated that Sentinel-1 is suitable for dynamic change information extraction, such as land cover mapping (Minh et al., 2018); crop monitoring (Bargiel, 2013; Inglada et al., 2016); urban change monitoring (Ban and Yousif, 2012; Gamba et al., 2006). Recently, how to better combine optical imagery and SAR data has become a very attractive topic in the field of remote sensing (Dusseux et al., 2014; Gao et al., 2017; Reiche et al., 2015; Sharma et al., 2018; Van Tricht et al., 2018; Whyte et al., 2018). In this area, the common practice is to use the observations of two sensors as independent features, and then feed them together into a standard machine learning model (Denize et al., 2019; Lu et al., 2018). Although these methods have achieved good results compared to a single sensor, still, the interplay between optical and SAR data is underutilized. Thus, it is urgent to propose a strategy that combines the optical and SAR data by simultaneously considering the correlation and complementarity between the two data.

Generally, both optical time series and SAR time series are considered to be able to accurately describe the vegetation growth cycle. In recent years, several studies have shown that a relationship between Sentinel-2 and Sentinel-1 time series data can be constructed(He and Yokoya, 2018; Scarpa et al., 2018; Veloso et al., 2017). The main challenge to construct such a relationship lies in the variety of SAR backscatter signals. Specifically, on the one hand, due to the sensitive nature of SAR data, the signal changes significantly when responding to different vegetation types (Bousbih et al., 2017). On the other hand, backscatter value varies at different growth stages even for the same vegetation. The ground scattering dominates in the early and late stages of vegetation growth, while vegetation scattering dominates in the middle (Mattia et al., 2003). To alleviate this difficulty, some studies have suggested that the correlation of SAR backscatter with optical vegetation descriptors is higher than single bands(Kim et al., 2011). Therefore, how to formulate the complex relationship between the SAR and optical time series is important for dense time series construction and cropland monitoring. A recent study introduced a new fusion approach for Sentinel-1 and Sentinel-2 to generate the LAI time series without data gaps (Pipia et al., 2019). Unfortunately, due to the limitation of the correlation model, it is difficult to show satisfactory performance in long temporal gap tasks. In short, filling the temporal gap of optical imagery with SAR data is a suitable solution. However, most of the current methods have limitations (e.g. the limitations of some regression models on sequence length, and the limitations of the fixed form of the predefined models and associated mathematical assumptions on model flexibility), thus it is necessary to design more robust and automatic methods. Recently, deep learning provides viable solutions for such a task(Gao et al., 2020; Wang et al., 2019). For example, (Scarpa et al., 2018) realized the direct estimation of NDVI from SAR data based on the CNN model. Although without a quantitative estimation of absolute errors, deep learning poses great potential in complex correlation construction.

Compare to physical models, deep learning is a fully data-driven approach (LeCun et al., 2015). In the field of remote sensing, the most commonly used deep learning models are convolutional neural networks (Kussul et al., 2017) (CNN) and recurrent neural networks (Bengio et al., 2013; Lyu et al., 2016) (RNN). Specifically, CNN is mainly applied to the extraction of spatial-spectral features (Zhao et al., 2019; Zhao and Du, 2016). In addition, RNN focuses on analyzing time series data(Zhong et al., 2019), and it can use temporal dependency at various time spans with hidden unit connections. Both models show encouraging results when used alone (Mou et al., 2018; Mou et al., 2017; Shao and Cai, 2018; Zhao et al., 2017). Still, a single model seems to be difficult to meet the needs in certain tasks. For instance, CNN is more suitable for estimating data gaps over short time series in the task of optical and SAR data fusion (Schmitt et al., 2018). Recently, some solutions have been proposed to combine convolutional neural networks and recurrent neural networks (Xingjian et al., 2015). Compared to single models, combined models generally provide better performance. For example, a complex model based on ConvGRU was built to realizes the coordination of the temporal and spatial characteristics of optical data and SAR data (Ienco et al., 2019). Although these efforts did not attempt to use SAR data to solve the problem of optical data missing. Still, in some way, the combination model may be a feasible solution for this particular task. Following a similar strategy, it is feasible to combine the CNN and RNN models to maximize their advantages.

With the continuous development of optical and SAR platforms, how to use SAR observation data to fill the temporal gaps of optical data has become one of the biggest challenges. In this study, we propose a novel collaborative approach to construct a relationship between Sentinel-1 and Sentinel-2. It aims to utilize SAR data to fill the optical gaps caused by clouds. To be specific, we propose an improved model based on the CNN-RNN combined architecture, where CNN is used to extract robust information from noisy SAR data, and RNN is used to establish the relationship between SAR and optical data. The application of this combined architecture in cropland monitoring is relatively rare. We name the proposed deep learning architecture as MCNN-Seq. The main contribution of this work as below:

(1) The MCNN-Seq is used to construct the relationship between multi-polarized SAR and optical time series data.
(2) An efficient SAR-NDVI estimation method is proposed, it provides reliable references for the missing optical time series.

This paper is structured as follows: the second chapter describes the study area and data; the third chapter introduces the proposed MCNN-Seq; the fourth chapter analyzes the experiments and results; the fifth chapter introduces the conclusions and future work.

2. Study areas and data preparation

2.1. Study areas

The study area is located in Imperial, Southern California, north

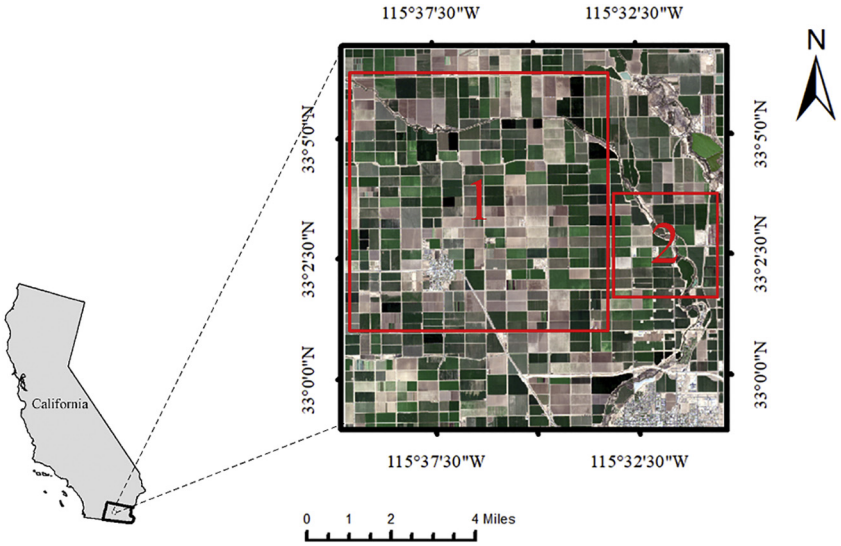

Fig. 1. Study areas located in California, USA. The RGB image derived from Sentinel-2 image acquired on 2018/03/12. The marked red subareas (1 and 2) were used for evaluating the fusion network performance. Subareas1 is with a spatial extent of 10 km × 10 km. Subareas2 is with a spatial extent of 4 km × 4 km.

latitude 32°59′N ~ 33°6′N, west longitude 115°39′W ~ 115°30′W. This area is located in the low-altitude area of the Colorado Desert. The study area has a tropical desert climate where the temperature is high but with high crop productivity. The average annual temperature is higher than 27 °C, and the temperature variation is also very large. The area is dry with little rain throughout the year, and the average annual precipitation (3 in.) is much lower than the average annual precipitation in the United States (28 in.). In California, Imperial County has one of the highest yields of crops such as alfalfa and onions (Belgiu and Csillik, 2018). Fig. 1 shows the study area with a spatial extent of approximately 200km^2. Two subareas (1 and 2) as shown in Fig. 1. The subareas1 is used to evaluate the performance of the MCNN-Seq. In order to visually compare the predicted images for different strategies, the subarea 2 with a small extent, and no cloud cover was selected.

2.2. Time-series data acquisitions

2.2.1. Sentinel-1 SAR data

In this study, images of Sentinel-2 and Sentinel-1 were collected from GEE, and the obtained two data sets have a spatial resolution of 10 m. We referred to Sentinel-1 SAR images that recorded in interferometric wide (IW) swath mode (Ground Range Detected products). The SAR data processing includes:(1) thermal noise removal, (2) Apply Orbit File, (3) radiometric calibration to sigma0, (4) Range-Doppler Terrain Correction using digital elevation model data and (5) transformation of the backscatter coefficient (δ) to decibels (dB). Finally, the entire data set was projected using the UTM / WGS84 projection system. A total of 31 images from the Sentinel-1 descending orbit were collected from January 5, 2018, to December 27, 2018.

2.2.2. Sentinel-2 optical data

For the optical data set, we downloaded a total of 73 Sentinel-2 images that span from January 1, 2018 to December 27, 2018, (including Sentinel-2a and Sentinel-2b). At the same time, the preprocessing was performed to extract optical image features, including radiation calibration, cloud mask, atmospheric correction, calculation of NDVI. Meanwhile, in order to reduce the impact of noise, only cloudless or partially clouded images were included in the Sentinel-2 data set. Moreover, due to the differences in revisit periods of Sentinel-1 (12 days) and Sentinel-2 (5 days), we chose 31 Sentinel-2 images that approximately align with the Sentinel-1 data set in terms of acquisition dates, as shown in Fig. 2. Finally, the Sentinel-2 data set is projected onto the UTM / WGS84 projection system for registration with the Sentinel-1 data.

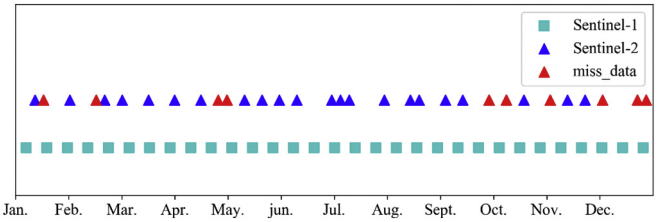

Fig. 2. Sentinel-1 and Sentinel-2 data acquisition date. A red triangle indicates that the Sentinel −2 image was affected by clouds or sensors on that date, causing some data to be missing.

2.2.3. Cropland reference data

U.S. Department of Crop (USDA) regularly maps land cover classification maps are publicly available (Boryan et al., 2011), and we used the Cropland Data Layer(*CDL*)of Imperial in 2018 to validate the experiment. This data is produced by the USDA and covers 48 states, it mainly provides the information on crop types. In this study, five major crops such as Onion, Winter Wheat, Corn, Sugar Beet, and Alfalfa were selected for analysis, other rare crops were aggregated as the “Other Crops” category (Table 1).

2.2.4. Dataset sampling strategy

MCNN-Seq can take the complete temporal profiles of individual pixels as the input samples. Therefore, the available samples are optical pixels without missing data. The Sentinel-2 data set of the entire study area is divided into two parts, which are the pixels with missing values

Table 1
Per description (pixels) statistics for subarea.

Description	Subarea1	Subarea2
Onion	84,001	9534
Winter Wheat	29,938	3086
Corn	14,613	3133
Sugar Beet	141,047	29,119
Alfalfa	272,529	28,354
Hay (Non-Alfalfa)	67,802	41,433
Non-crop land	121,442	20,527
Other Crops	228,871	9242

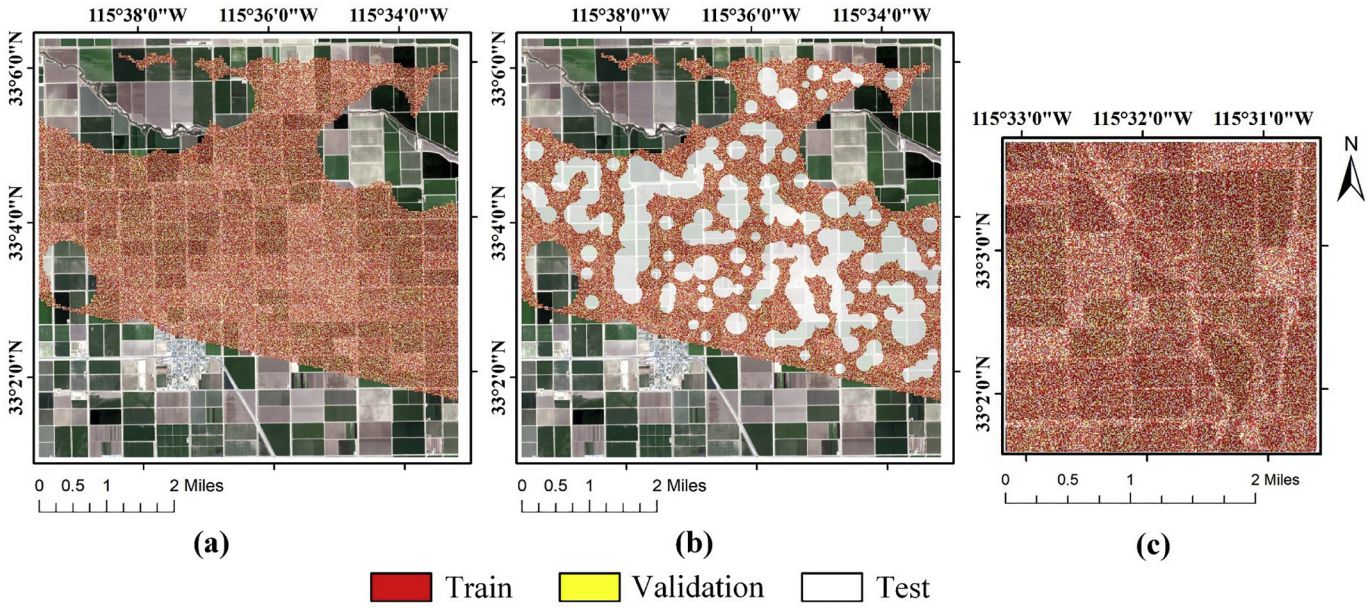

Fig. 3. Spatial distribution of training set and verification set for (a)subareas1(50%cloud), (b) subareas1(70%cloud) (c)subarea2. The pixels that did not participate in the training were not displayed (the pixels with missing values along the temporal axis).

Table 2
Per dataset statistics for subarea.

	Train samples	Validation samples	Test samples
Subarea1(50% cloud)	283,678	81,051	152,571
Subarea1(70% cloud)	175,417	52,656	289,227
Subarea2	89,040	22,260	47,700

along the temporal axis and the pixels with complete temporal profile, respectively. Based on that, we divided the available pixels into three parts: 1. Training set. It is used to train the regression algorithm of the model, accounts for about 55%. 2. Validation set. It is used to select the best hyperparameters, accounting for about 15%. 3. Testing set. It is used to verify the performance of the model, accounting for about 30% (Fig. 3). To investigate the performance of the model under high cloud cover, we manually plotted an additional 20% of the cloud cover based on subarea 1(50% cloud). This dataset is called subarea1 (70% cloud). At the same time, the training and validation sets of this dataset (70% cloud) inherit from the dataset (50% cloud), and the test set includes the inherited test set (50% cloud) and 20% of the clouds. These sets are randomly divided and independent from each other. Table 2 reports the number of pixels of the training set, verification set and test set in subarea 1 and 2.

3. Methodology

The workflow of this methodology includes the following steps: (1) preprocessing the Sentinel-1 and Sentinel-2 data (for details, see sections 2.2.1 and 2.2.2); (2) dataset processing, including the construction of VV and VH time series, generating training and validation samples; (3) training model, VV and VH input MCNN-Seq;(4) prediction, converting SAR time series data to optical time series data (Fig. 4).

The overall architecture of the MCNN-Seq has shown in Fig. 5. This architecture combines CNN and LSTM. The CNN modular is used as a front end to extract the features from the original input data and reduce the impact of noises, and the LSTM modular as the back end which receives the information output by the CNN. The CNN modular and the RNN part are connected together. The model takes VV and VH time series data as input, and processes them as two independent CNN branches. The two branches have the same neural network structure. The time series of VV and VH are equally divided into multiple subsequences, in order to extract the change characteristics of different time periods. Therefore, the CNNs of each branch will process its corresponding subsequence one by one. Then, the feature vectors of the VV and VH subsequences at the same time period are connected. Finally, it is input into the LSTM modular to fit the relationship with the target NDVI sequence. In order to help the model better adapt to the dynamic relationship of SAR and optic data, we have added an attention mechanism to the decoder. In the following, we will provide a detailed description of the CNN and LSTM modular.

3.1. CNN modular

Due to the noise adaptive ability of CNNs, useful information can be extracted even from noisy data (Qian et al., 2016). Conventionally, two-dimensional CNN (conv2D) is widely used due to its excellent capability to capture the spatial features. In contrast, one-dimensional CNN (conv1D) seems more suitable for the time series data processing (Liu et al., 2018). Since the time series can be defined as a one-dimensional vector, we choose Conv1D to extract the changing features from multi-polarized VV and VH data. In order to process multi-feature time series data, the CNN model with multiple channels is generally used. Although the features extracted by each channel are independent, some information may be lost when connecting these features (Canizo et al., 2019). To retain such information as much as possible, we designed two independent CNN branches to process the VV and VH respectively. The CNN network of these two branches sharing the same structure, which involves 2 convolutional and max-pooling layers and a full-connected layer. The convolution layer is with the trainable kernel that generates feature maps, and they are activated by the rectified linear unit (ReLU). Then the feature maps are subsampled by the max-pooling layer. Through repeating the previous steps, the noise adaptive features can be extracted layer by layer. Finally, the fully-connected layer combines these highly abstracted features into one. CNN is described by Eqs. (1), (2):

$$h_{i,j} = ReLU\left(\sum_{m=1}^{M}\left(h_{i-1,m} * w_{i,mj}\right) + b_{i,j}\right) \quad (1)$$

$$h'_{i,j}(l) = ReLU\{h_{i,j}(sl),\ldots,h_{i,j}(sl+r-1)\} \quad (2)$$

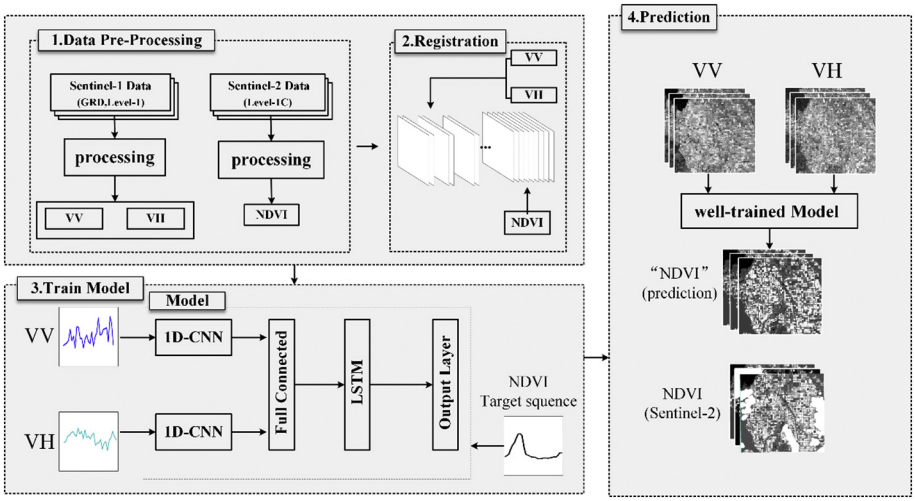

Fig. 4. Flowchart of the proposed SAR predicted optical data. It can be divided into four main steps. The first step is to process SAR and optical data. Then, the two kinds of data are registration to partition the training set and the verification set. Thirdly, the relationship between SAR data and optical data was constructed by MCNN-Seq framework. Finally, based on the well-trained MCNN-Seq framework, complete optical data can be predicted to provide a reference for missing image data.

where $h_{i,j}$ is the j output feature maps obtained by the i convolutional layer,$*$ is convolution operation,$w_{i,mj}$ represents kernel, M is number of feature maps, s is steps, and r denotes the number of outputs.

As mentioned, we divide the time series into N segments with the same length. That means it can focus on different phases in the time series. Each subsequence will get a sequence of feature maps after convolution is applied. The features of different subsequences are independent of each other, so they are connected in series to obtain N feature maps. Then the feature maps of the two branches are connected in a sequence. Note that only the feature maps for the corresponding phases are connected. In this way, the noise adaptive feature vector is obtained and can be used as an input to the LSTM modular.

3.2. LSTM modular

RNNs have the ability to preserve the state between consecutive input data, they are generally considered to be designed specifically for processing sequential data. However, due to the problems of vanishing gradients and learning efficiency, variants of RNNs are often used in some studies, in which LSTM is the most famous variant model. LSTM will create time-varying information paths whose derivatives are robust to gradient disappearance or explosion problems. LSTM was proposed by (Hochreiter and Schmidhuber, 1997), and designed three gate units to guide the information paths, including the forget gate, input gate I_N and output gate O_N .After the original data that feed into LSTM in the form of $x = [x_1, .., x_N]^T$. First, get the input information x_j from the current step and the hidden state h_{j-1} of the previous step. Then, the combined information is passed through a sigmoid activation function to decide how much information will be thrown away from the cell state. Then, the combined information is passed to the forget gate F_j, and a sigmoid activation function is used to determine the proportion of retained information in the cell state. The sigmoid function outputs a value between 0 and 1. Then through the input gate I_j, decide what percentage of the new information $\widetilde{C}_j$ will be stored to the cell state. Among them, use the *tanh* activation function to create a candidate for updating. Besides, the new cell state C_j can be obtained by multiplying the cell state C_{j-1} of the previous step by F_j and then adding the updated information $\widetilde{C}_j$ by I_j. Finally, new hidden states h_j can be generated based on output gates O_j and new cell states C_j.

$$F_j = \sigma(W_{Fx}x_j + W_{Fh}h_{j-1} + bias_F) \tag{3}$$

$$I_j = \sigma(W_{Ix}x_j + W_{Ih}h_{j-1} + bias_I) \tag{4}$$

$$O_j = \sigma(W_{Ox}x_j + W_{Oh}h_{j-1} + bias_O) \tag{5}$$

$$\widetilde{C}_j = tanh(W_{Cx}x_j + W_{Ch}h_{j-1} + bias_C) \tag{6}$$

$$C_j = F_j * C_{j-1} + I_j * \widetilde{C}_j \tag{7}$$

$$h_j = O_j * tanh(C_j) \tag{8}$$

where W_{Fx}, W_{Fh}, W_{Ix}, W_{Ih}, W_{Ox}, W_{Oh}, W_{Cx}, W_{Ch} are the weight matrices, which are used to govern the connection from the corresponding input to the hidden layer. $bias_F$, $bias_I$, $bias_O$, $bias_C$. are the bias. The above are all trainable parameters.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{9}$$

$$tanh(x) = \frac{e^z - e^{-z}}{e^z + e^{-z}} \tag{10}$$

Considering that the dates of optical and SAR image acquisition do not always completely coincide. And in practical applications, the length of the output sequence is likely to be different from the input sequence, so the foundation LSTM model may not be applicable, So (Sutskever et al., 2014) proposed a relatively mature Sequence to Sequence (Seq2Seq) architecture to achieve flexible processing of inconsistent input and output sequences. The Seq2Seq network consists of an

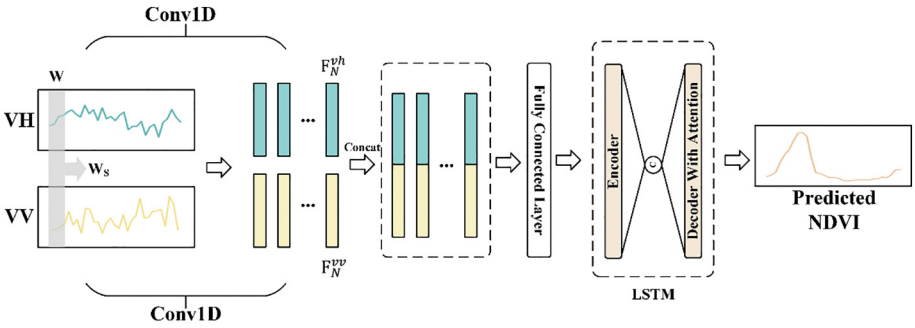

Fig. 5. MCNN-Seq architecture for conversion from SAR time series to optical time series. Starting from the left. Starting from the left, VV and VH time series data are equally divided into N subsequences of length Sub_L. Then, N feature maps are obtained for each backscatter time series. The N feature map of the VV and VH subsequences is denoted as $F_N^{vv}$and F_N^{vh}.Feature maps of VV and VH subsequences are connected in order. Finally, input the LSTM modular. Finally, the predicted NDVI is obtained. c represents a state vector.

encoding layer and a decoding layer. The encoding layer is a dynamic multi-layer LSTM. The time-series data is input into the network in order and only the last hidden state is retained. Equivalent to compressing the entire sequence together, represented by a fixed-size vector c for use by the decoding layer c represents for use by the decoding layer. In the decoding process, c will be used as the initial state of the decoder network, and the output value of the previous time step is input to the next LSTM unit. The Seq2Seq model generates the conditional probability of the target sequence $y = [y_1, .., y_t]^T$given the input sequence, defined by Eq. (11)

$$p(y_1,...,y_t \mid x_1,...,x_N) = \prod_{t=1}^{t} p(y_t \mid y_1,...,y_{t-1}, c) \tag{11}$$

There are some problems in the Seq2Seq network: the output sequence largely depends on the final hidden state of the encoder. Once the sequence is relatively long, it is easy to lose some information and severely limit the performance of the model. Although LSTM improves on this issue compared to RNN, it still performs poorly when dealing with long sequences. (Bahdanau et al., 2014) first proposed the attention mechanism to solve this problem. By adding attention vectors to each decoding step, the model can pay special attention to specific parts, so as to obtain better results. Therefore, we added the attention mechanism to Seq2Seq.

$$e_{ij} = a(s_{i-1}, h_j) \tag{12}$$

where e_{ij} indicates the degree of matching between the input of the encoder at time j and the output of the decoder at time $i - 1$. s_i is the hidden state of the decoder at time i, h_j is the output state of the Encoder hidden layer at time j. a is called alignment model.

Then normalize e_{ij} using the softmax function:

$$\alpha_{ij} = softmax(e_{ij}) \tag{13}$$

where α_{ij} indicates the importance of the input at time j to the output at time i.

Finally, calculate the context vector $\mathbf{c}_i$ at i moment

$$\mathbf{c}_i = \sum_{j=1}^{N} \alpha_{ij} h_j \tag{14}$$

After the model is built. Assume x is the SAR time series and y is the optical time series. Therefore, the training set can be expressed as $[x, y]^m$. where m is the number of training samples. The input sequence x is equally divided into N subsequences. Features are extracted one by one by CNN modular, so the feature sequence can be expressed as F_m, $m \in (1, \cdots, N)$. The LSTM modular training is for learning a mapping function $f: \hat{y} = (F; w)$,where $\hat{y}$ is the sequence of the output, and w is all parameters including weight and bias. It should be noted that we have defined the loss function and metrics, the former being Root Mean Square Error (RMSE) and the latter being R-square(R^2).

$$RMSE = \frac{\sum_{i=1}^{N} \|y(i) - \hat{y}(i)\|}{N} \tag{15}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y(i) - \hat{y}(i))^2}{\sum_{i=1}^{N} (y(i) - \overline{y}(i))^2} \tag{16}$$

4. Experiments and results

4.1. Experimental designs

In order to extract useful information from complex SAR time series, two completely independent 3-layer Conv1D frameworks were constructed for each polarization. Specifically, we divide the VV and VH time series into 31 segments with the step of 1, so the size of the input sample is set to $31 \times 1 \times 1$. For each CNN framework, the first convolutional layer contains 31 filters of 1×1. Then, the second convolutional layer includes 62 filters with size 1. After that, flatten layers are generated to produce 31×62 feature maps in each branch. Finally, the features obtained from the two CNNs are connected and converted to 31×124 feature maps.

To construct the relationship between SAR and optical data, the output features from CNNs are fed into the LSTM modular in sequences. Inside of the LSTM, firstly, the temporal relationship of input features can be extracted in the encoder. Then, the decoder was designed to capture the relationship between SAR data and the optical sequences. Finally, the relationship is used to predict the missing optical data. Moreover, in order to reduce the impact of uncertain noises, the attention mechanism was introduced in the decoder. In the training stage, the RMSprop optimizer was used for training. The learning rate was set to 0.001 and the batch size was set to 500.

4.2. Assessment of predicted images

In this section, the optical time series were predicted by SAR data with the help of the proposed MCNN-Seq. To be specific, the SAR time series was converted into optical time series with the same length. To evaluate the performance of the proposed method in terms of optical time series prediction, we selected the representative subarea 1(50% and 70% cloud) to quantitatively evaluate the test pixel. The quantitative comparison of real and predicted values is mainly achieved by error histograms and density scatter plots.

Fig. 6 shows the test accuracy of the proposed model for each image in subarea 1(50% and 70% cloud). From these metrics, it is clear to point that the predicted optical time series data has a high correlation with the original optical data. It can be found that when different data sets are used, the test accuracy of each image shows a similar trend. Among them, the R^2 and RMSE of the predicted NDVI are relatively unsatisfying on the date 2018/04/16 (The R^2 of 50% cloud and 70%

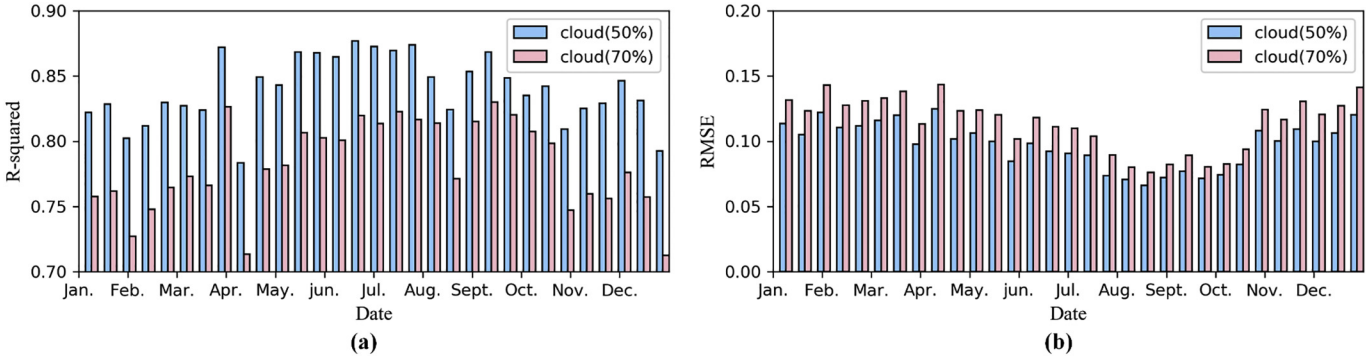

Fig. 6. From Jan. 2018 to the end of Dec. 2018, the test accuracy evaluation results of all predicted images and corresponding optical images in subarea 1 (50% and 70% cloud). (a)R^2 and (b)RMSE are calculated from the values of each test pixels and the predicted values of its corresponding pixels.

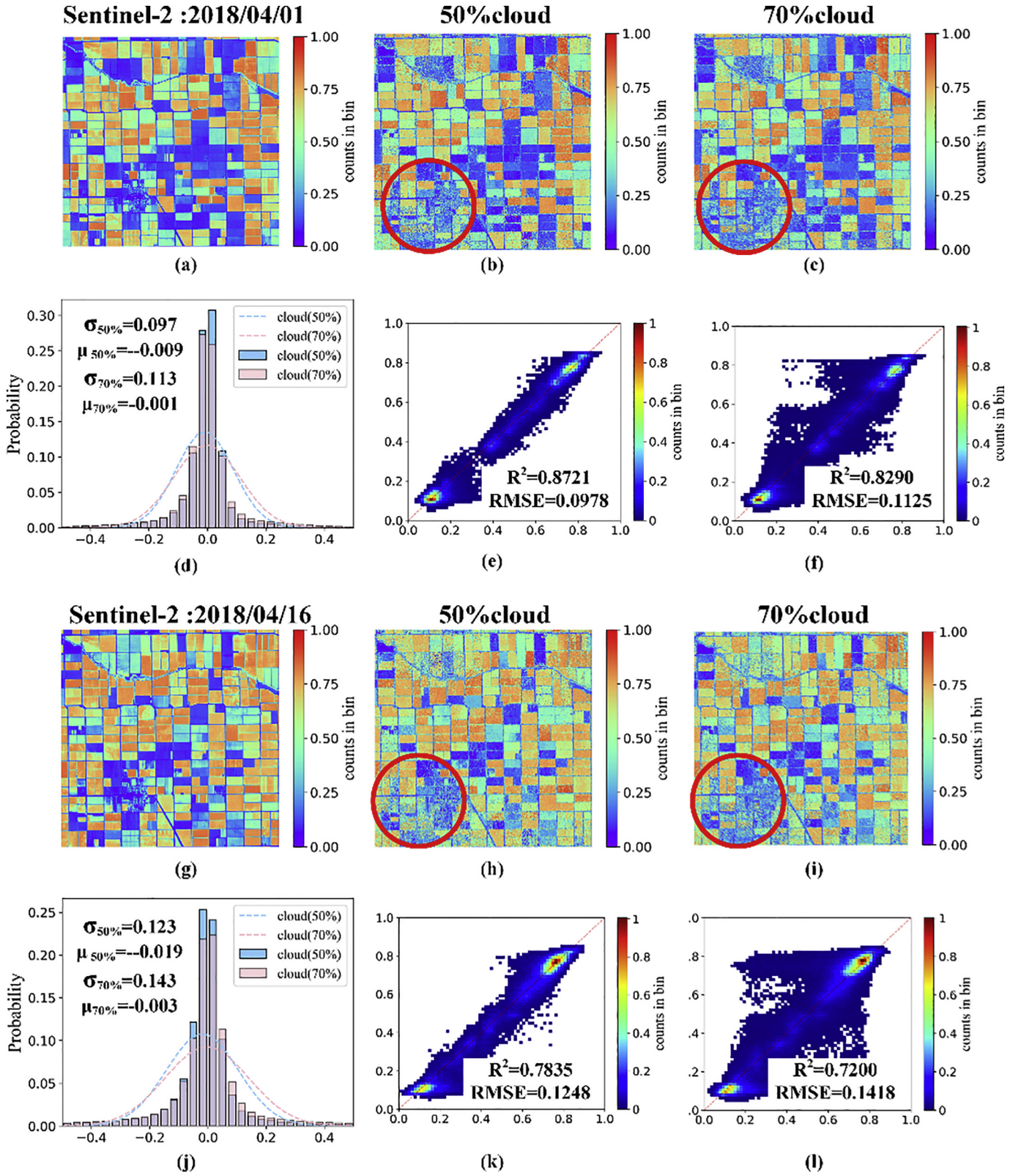

Fig. 7. Comparisons between the reference and the predicted image for dates: 2018/04/01 and 2018/04/16. (a), (g)Sentinel-2 NDVI reference and (b), (c), (h), (i) predicted images are compared visually. (d), (j)error histograms and (e), (f), (k), (l) density scatter plot compared from a statistical perspective. The red circle is the area with low accuracy.

cloud are 0.7835 and 0.7200 respectively, RMSE are 0.1248 and 1418 respectively) and 2018/12/27 (The R^2 of 50% cloud and 70% cloud are 0.7926 and 0.7167 respectively, RMSE are 0.1202 and 1403 respectively). Noticeably, the original optical acquisition on 2018/04/16 was not polluted by clouds, while the optical image on 2018/12/27 has missing data. In order to respond to this situation, we developed two scenarios to assess the spatial pattern.

Scenario 1:No missing data in optical time series. According to the

evaluation results, we selected the assessment date with the highest and the lowest accuracy (dates are 2018/04/01 and 2018/04/16, respectively). Fig. 7 shows the comparison results between the original and predicted images on 2018/04/01 and 2018/04/16. From the figure, we can conclude that all predicted images (include 50% cloud and 70% cloud) are highly similar to the optical ones in terms of spatial variation and statistical histograms. In particular, from a visual perspective, it is almost impossible to directly see the difference between the prediction images of the two data sets (50% cloud and 70% cloud). This means that dense but small cloud cover has little impact on model performance. However, the accuracy of predicted data on 2018/04/01 is the highest in 2018, still, the predicted images has more speckles in the red circle (as shown in Fig. 7) which is different from the real values. From the statistical point of view, the error histogram is almost the same. However, it can be seen from the scatter diagram that the image predicted by 50% cloud has a small degree of dispersion, confirmed by RMSE values. The prediction accuracy of 50% cloud and 70% cloud showed a similar variation trend, and the variation of 70% cloud accuracy was more significant. Thus, the 70% cloud is taken as an example to analyze the variation in the accuracy of the two dates. From the statistical point of view, for Subarea1(70% cloud), even for the lowest accuracy of predicted data on 2018/04/16, the differences between the optical and the predicted values are concentrated in the range of -0.1 and 0.1. The main reason that the prediction accuracy of 2018/04/16 is lower than that of 2018/04/01 has changed significantly within 15 days. For example, during the stem elongation of some crops (such as lettuce and broccoli) in April, as changes in the vertical structure had a greater impact on SAR backscatter.

Scenario 2: Optical time series with missing data. We selected two representative optical images (dates 2018/12/02 and 2018/12/27 respectively) and their corresponding predicted images. Among them, 2018/12/27 has the lowest accuracy, but only a small area is covered by cloud, while 2018/12/02 is not only affected by cloud pollution, but also by acquisition strips, resulting in missing data in most areas. From Fig. 8, we can conclude that all predicted images are generally consistent with their corresponding optical images, especially, the predicted images can also reflect some spatial information and crop phenology information in the data missing areas of the original images. It means that the predicted image based on SAR data has the potential to fill the gaps in optical data. Compared with 50% clouds, 70% of cloud predictions tend to overestimate low values. The main reason may be that in the region with low NDVI value, the crop biomass is less and the ground scattering takes the dominant position, leading to the unreliability of the predicted value of the model. It is worth noting that the images on these two dates also face the same problem as Scenario 1, that is, the predicted image and the optical data in the red circle contain over-fragmented small croplands. Therefore, due to the spatial variation, the predicted image contains a large number of speckles or noises which resulted in low accuracy in SAR-based optical data prediction.

4.3. Assessment of temporal profile

In Section 4. 2, we discussed the accuracy of the model in the spatial domain, but the main purpose of this research is to fill the temporal gaps in the optical time series, so the accuracy of the predictions at the temporal axis is also one of the important criteria for evaluating the prediction performance of the proposed model. In order to understand the reliability of the predicted time-series profiles, we compared the temporal behavior of valid Sentinel-2 NDVI, the Sentinel-1 VV, VH, and predicted NDVI for evaluation. With reference to the 2018 CDL, we grouped the crops into 3 categories according to their growth cycles, that are, (1) Onion and winter wheat;(2) Corn and sugar beet; (3) Alfalfa. The quantitative evaluation and analysis were performed for each type of crop. They are analyzed in the same subsection by one field. The onion field contains 5140 pixels, the cornfield contains 1336 pixels, the sugar beet field contains 2655 pixels, the winter wheat field contains 3649 pixels, and the alfalfa 3790 pixels. To describe crop growth cycles, the mean NDVI profiles along with the twice standard deviation values have been regarded as the confidence range. Meanwhile, the correlation coefficient R^2 represents the correlation between the real-time series and the predicted time series (Table 3). It should be noted that the following analysis results based on one single field may not fully represent the crop behavior of all crops in the entire study area.

4.3.1 Onion and winter wheat

The onion and winter wheat in this study area have similar phenological characteristics, with NDVI peaking between March and April and continuing to decline for the rest of the year. The difference is that winter wheat has a completely different plant structure from onions. We select an onion field or winter wheat field and its corresponding Sentinel-2 NDVI time series from the predicted time series data (Subarea1 50%cloud).

In most cases, the predicted NDVI sequences and the original NDVI profiles are basically consistent with each other, so the predicted time series has the ability to accurately describe the phenological stage of the crop (Fig. 9). Quantitative assessments of onions and winter wheat also support this view, with the $R^2 = 0.9409$ and $RMSE = 0.0432$ for winter wheat, and for the onion $R^2 = 0.9824$ and $RMSE = 0.0334$. But there are some cases where details have lost, from January to February, the NDVI of onions remained stable, and almost all the values are with the mean values. However, the predicted values have a wider range of uncertainty, and the average values are slightly lower than the observed optical values. From March to April, the NDVI value of winter wheat reached its peak, and even reached saturation of the NDVI index and remained stable at the same time. However, it is difficult for predicted values to reach such high values, for most cases they distributed around 0.9. After June, NDVI remained stable at low values after harvest of onions and winter wheat. At this time, the NDVI value was underestimated. The onion prediction values almost around 0.1, but the reference values are above 0.2. The predicted values of winter wheat during this period are almost consistent with the real values and is stable below 0.1. Interestingly, the predicted NDVI time series profiles of winter wheat and onion during this period is almost the same. After the crop harvest, the SAR backscatter signal is mainly contributed by soil, which weakens the contribution of the vegetation, resulting in similar backscatters.

4.3.2 Corn and sugar beet

Most corn and sugar beet fields in the study area will start to increase NDVI from October to November. This means that during this period, the crop regrowth or other crops are planted in these fields. To analyze whether the predicted time series can accurately describe this key information, we will discuss corn and sugar beet in the same subsection.

We can observe that the corn prediction data and the original NDVI time series have good similarity, both visually and statistically. But from January to February, the predicted time series of beets and corn showed the same problems as onion. That is, the uncertainty ranges of the predicted values are relatively large. The explanation is that there are various scattering mechanisms in the field, which cause the differences in backscatter signals and large standard deviations. At 2018/02/15, there was a sudden decrease in the NDVI of sugar beet, and the model was able to catch this mutation, but it did not produce a value that changed large enough. The reason may be that NDVI on the adjacent dates around 2018/02/15 already in a stable state, and the MCNN-Seq is not sensitive to this noise-like sudden mutation. After October, shortly after the emergence of new plants, NDVI began to increase again (Fig. 10). This key information can also be clearly described in the predicted time series. It has shown that the predicted images can not only accurately describe the growth cycle of a crop, but also accurately grasp the information of re-growth on new crops. This means that even if the type of vegetation changing, the predicted time series can fill the phenological details of the missing data within the optical time-series.

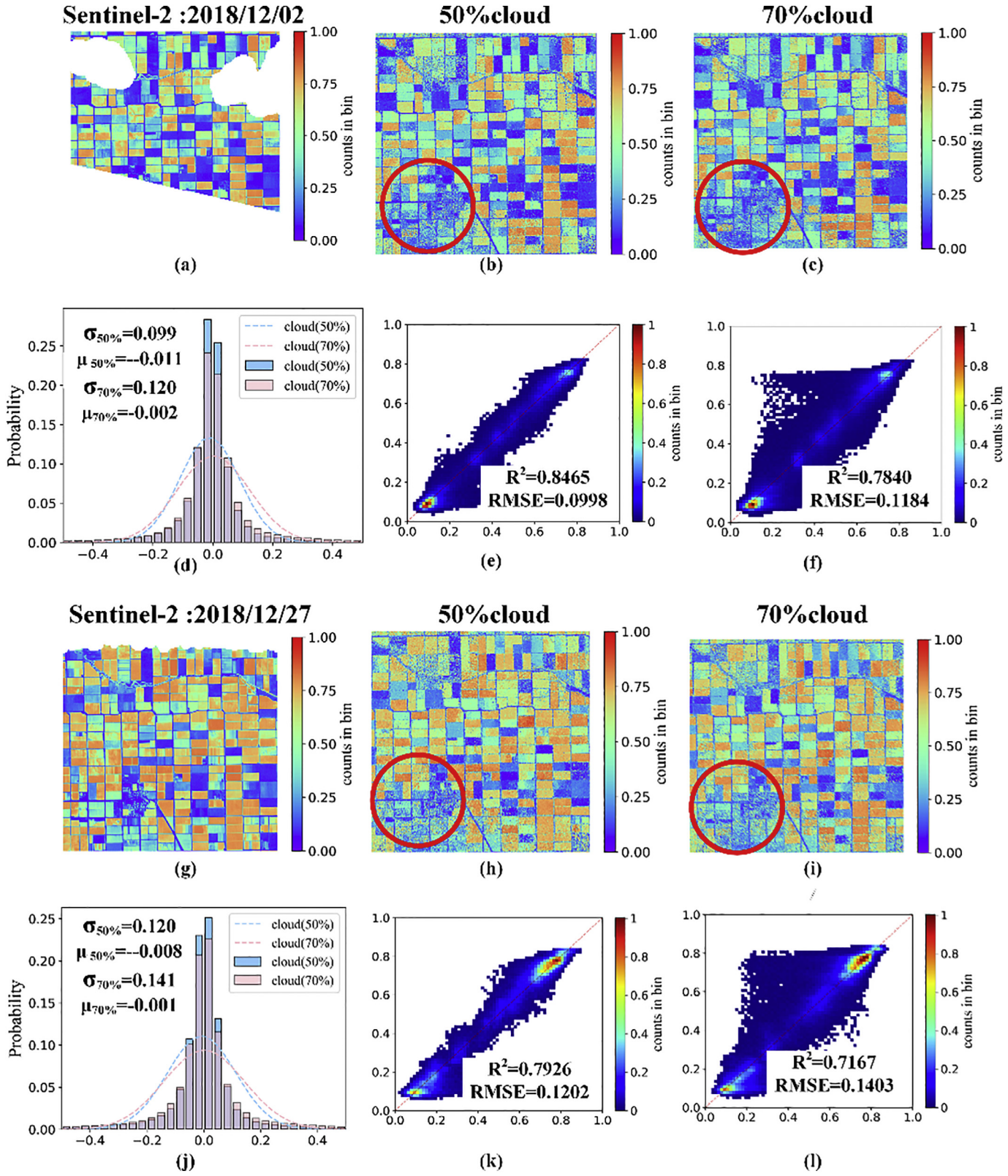

Fig. 8. Comparisons between the reference and the predicted image for dates: 2018/12/02 and 2018/12/27. (a), (g)Sentinel-2 NDVI reference and (b), (c), (h), (i) predicted images are compared visually. (d), (j)error histograms and (e), (f), (k), (l) density scatter plot compared from a statistical perspective. The red circle is the area with low accuracy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

4.3.1. Alfalfa

Alfalfa is one of the most important crops in the study area, and it is also one of the biggest challenges in the prediction process. In Fig. 11, the variation trend of the NDVI time profile on alfalfa is very complex compared with other major crops. Under this circumstance, it can be observed that the predicted data has a similar trend compared to the real NDVI (R^2 = 0.7018, RMSE = 0.0796). Their error bar of the reference data minus the predicted data is also given in Fig. 11 to

Table 3
Accuracy of the predicted time series in the selected field.

Metrics	Main crop				
	Onion	Winter wheat	Corn	Sugar beet	Alfalfa
R^2	0.9409	0.9824	0.9157	0.9749	0.7018
RMSE	0.0432	0.0334	0.0453	0.0454	0.0796

facilitate their comparison. Although there are some differences, the predicted time series can almost describe the growth cycle a year.

4.4. Comparative evaluation

In the previous sections, we have shown the structure of the proposed model and analyze the performances of the proposed model on SAR images to predict optical images. In this section, we apply the well-trained model to subarea 2 and compared it with other competing methods. Because the model is based on the time-series sequence relationship to achieve the SAR-optical converting, it is difficult to find close competitors with similar principles. Thus, we considered using the representative deep learning models such as CNN and RNN to verify the advanced nature of the MCNN-Seq. The former is often seen as one of the solutions to this task in some recent studies (Schmitt et al., 2018). The latter is the basic model of MCNN-Seq. The same training set and validation set were used to ensure a fair comparison between the three models in this section. Thus, we use Conv1D as the CNN model, which is more suitable to process time-series data.

As regards the hyperparameters of MCNN-Seq (cf. Section 4. 1). These hyperparameters have proven to be good choices in previous experiments. The hyperparameter settings of RNN is the same as the RNN branch of MCNN-Seq (include two hidden layers with 100 hidden units and a full connection layer). the CNN detailed configurations are listed in (Zhao et al., 2020). It consists of two convolution layers with 32 and 16 filters respectively, one max-pooling layer and one full connection layer. Fig. 12 shows the training accuracy curve and validation accuracy curves of the three models. It should be emphasized that the evaluation metrics here are the R^2 between the predicted sequence and the target sequence. We can observe that all three models converge at epoch = 50, and the difference in accuracy between training and validation is almost constant. The accuracy of RNN is slightly better than CNN, and the uncertainty range of the validation accuracy curve is also comparatively small. The MCNN-Seq accuracy is much higher than the other two models.

In order to analyze the advantages and disadvantages of the models from a quantitative perspective, we calculated the R^2 and RMSE of the test sequence and the reference sequence in subarea 2, as shown in Table 4. It can be observed that both CNN and RNN are relatively unstable compared to the proposed method. From January to February, the absolute value of the RNN predicted image R^2 is very low (around 0.5). The main reason is that the basic RNN is the one-to-one model that the cellular state of each hidden unit is a process of gradual accumulation. At the beginning of the time-series, the hidden unit cannot get enough context information, and resulted in low prediction accuracy. From July to December, the absolute value of RNN predicted image R^2 improved significantly. This demonstrated that it is difficult to fit SAR and optics in a one-to-one fashion without considering temporal information. Conversely, the sequence-sequence used in MCNN-Seq shows relatively stable and high accuracy in the whole time-series. It is a feasible solution to predicted optical data of SAR data based on temporal contextual information. Also, it can be noted that the prediction accuracy of CNN shows an obvious downward trend. CNN

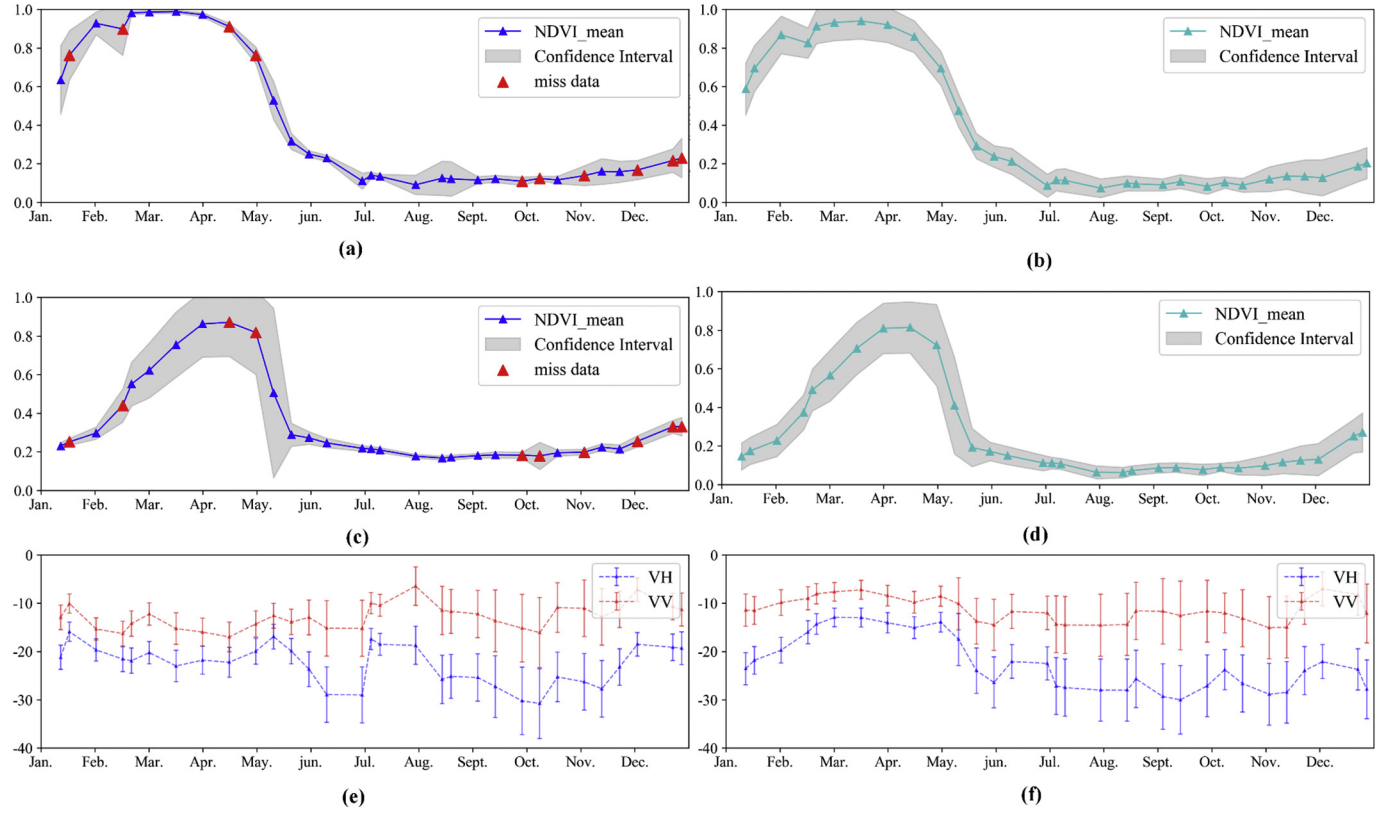

Fig. 9. Observations of the temporal behavior of NDVI for the selected fields (represented by the mean of all pixels). (a) NDVI of winter wheat (reference). (b) predicted winter wheat NDVI, the confidence interval (gray) is set to ± 2σ. (c) NDVI of onion (reference). (d) predicted onion NDVI, the confidence interval (gray) is set to ± 2σ. (e) backscatter signals of winter wheat, error bars denote σ. (f) backscatter signals of onion, error bars denote σ.A red triangle indicates that some of the optical data was missing on that date.

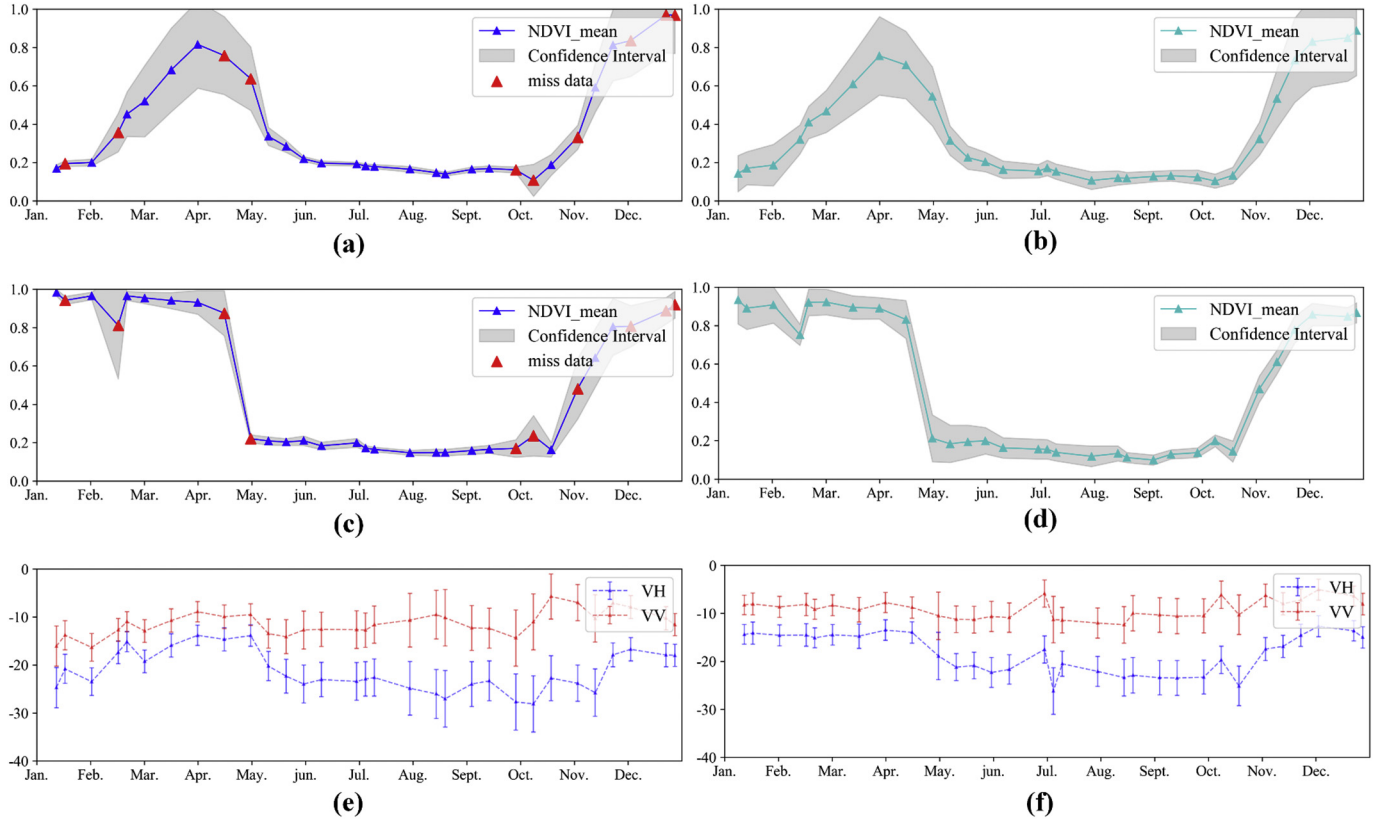

Fig. 10. Observations of the temporal behavior of NDVI for the selected fields (represented by the mean of all pixels). (a) NDVI of corn (reference). (b) predicted corn NDVI, the confidence interval (gray) is set to ± 2σ. (c) NDVI of sugar beet (reference). (d) predicted sugar beet NDVI, the confidence interval (gray) is set to ± 2σ. (e) backscatter signals of corn, error bars denote σ. (f) backscatter signals of sugar beet, error bars denote σ.A red triangle indicates that some of the optical data was missing on that date. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

mainly achieves SAR to optical prediction by extracting sequence features. With the increase of sequence length and crop rotations, the uncertainty of data is increased which leads to the failure prediction. This shows that CNN is more suitable for tasks in a short time, and it is difficult to capture stable temporal features in a long time-series.

In order to intuitively understand the performances of the three different models, we have selected 2 images (the dates are 2018/01/11 and 2018/12/27, respectively), as shown in Fig. 13 and Fig. 14. These two images are the first and last images of the time series. It can be found that compared with the RNN and CNN, the images predicted by the MCNN-Seq are much more similar to the real images, in terms of measuring the R^2 and RMSE. The low prediction accuracy of RNN and

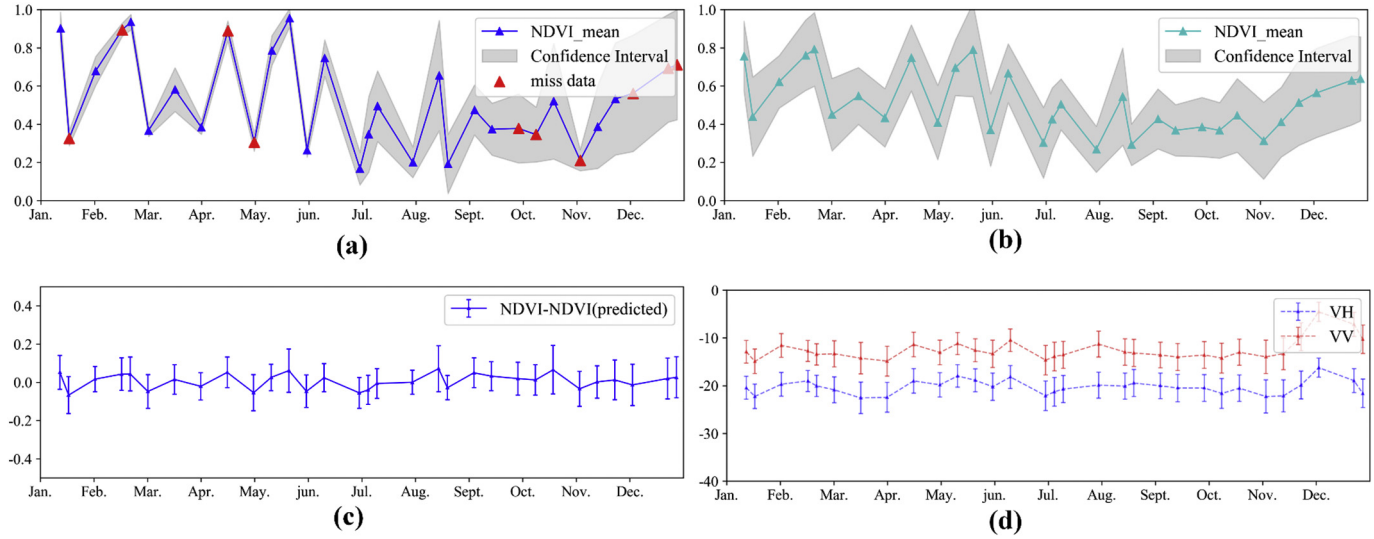

Fig. 11. Observations of the temporal behavior of NDVI for the selected fields (represented by the mean of all pixels). (a) NDVI of alfalfa (reference). (b) predicted alfalfa NDVI, the confidence interval (gray) is set to ± 2σ. (c) the reference data minus the predicted data, error bars denote σ (d) backscatter signals of alfalfa, error bars denote σ. A red triangle indicates that some of the optical data was missing on that date. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

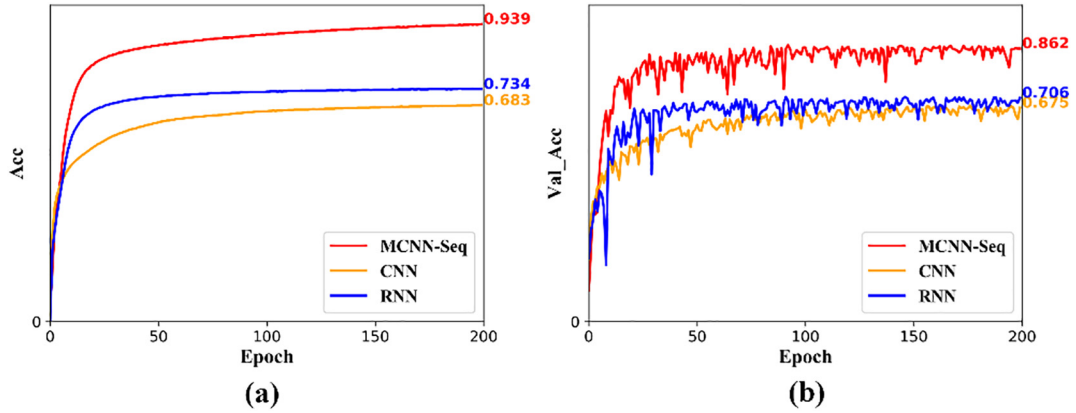

Fig. 12. Training accuracy curves and validation accuracy curves of three models: RNN, CNN and MCNN-Seq.

Table 4
R^2 and RMSE of test pixels where the images predicted by the three models compared to the reference data.

Date	RNN		CNN		MCNN-Seq	
	R^2	RMSE	R2	RMSE	R^2	RMSE
2018/01/11	0.3857	0.2168	0.6814	0.1561	0.8842	0.0941
2018/01/16	0.5084	0.1818	0.6905	0.1443	0.8915	0.0854
2018/01/31	0.5079	0.1968	0.6740	0.1602	0.8680	0.1019
2018/02/15	0.4954	0.1797	0.6511	0.1494	0.8630	0.0936
2018/02/20	0.5732	0.1724	0.6555	0.1549	0.8587	0.0991
2018/03/02	0.6266	0.1597	0.6868	0.1463	0.8711	0.0938
2018/03/17	0.6240	0.1624	0.6464	0.1575	0.8480	0.1032
2018/04/01	0.5525	0.1774	0.6533	0.1562	0.8471	0.1036
2018/04/16	0.4676	0.1947	0.5321	0.1826	0.7423	0.1353
2018/05/01	0.5046	0.1826	0.5865	0.1668	0.8321	0.1062
2018/05/11	0.6014	0.1715	0.6094	0.1698	0.8278	0.1126
2018/05/21	0.6461	0.1648	0.6145	0.172	0.8408	0.1105
2018/05/31	0.6305	0.1349	0.5981	0.1407	0.8488	0.0863
2018/06/10	0.6568	0.1448	0.5357	0.1685	0.8371	0.0997
2018/06/30	0.6974	0.1435	0.5770	0.1696	0.8515	0.1005
2018/07/05	0.7531	0.1283	0.5996	0.1634	0.8854	0.0874
2018/07/10	0.7612	0.1229	0.5972	0.1596	0.8830	0.086
2018/07/30	0.7231	0.1160	0.6759	0.1255	0.8580	0.0831
2018/08/14	0.8334	0.1099	0.7603	0.1318	0.8852	0.0913
2018/08/19	0.8234	0.1105	0.7224	0.1386	0.8755	0.0928
2018/09/03	0.7956	0.1053	0.7154	0.1242	0.8610	0.0868
2018/09/13	0.8440	0.1068	0.7380	0.1385	0.8776	0.0947
2018/09/28	0.8400	0.1181	0.7307	0.1533	0.8809	0.1019
2018/10/08	0.8607	0.1091	0.7330	0.1511	0.8779	0.1022
2018/10/18	0.8236	0.1080	0.7040	0.1399	0.8601	0.0962
2018/11/02	0.8000	0.1092	0.6565	0.1431	0.8325	0.0999
2018/11/12	0.7831	0.0989	0.6139	0.1319	0.8302	0.0875
2018/11/22	0.7883	0.1011	0.5526	0.1469	0.8141	0.0947
2018/12/02	0.7810	0.1023	0.5299	0.1499	0.8013	0.0974
2018/12/22	0.7719	0.1101	0.5141	0.1607	0.8053	0.1016
2018/12/27	0.7510	0.1157	0.5012	0.1638	0.8010	0.1034

LSTM is mainly impacted by two aspects: value differences and a large number of noises. The former is the shortcomings of CNN and RNN, as we mentioned above. The latter is caused by the high uncertainty of SAR data, and noises that reduced the formulation power of the model, which is the reason we consider CNN as the information extractor.

5. Discussion

5.1. Sources of error

In this section, we will discuss the reasons for the high uncertainty of the predicted images, especially for the low accuracy of the predicted images in the red circle in section 4. 2. Intuitively, we speculate the low prediction accuracies may be impacted by the land cover types. In order to evaluate the accuracy of each pixel for the predicted time series data, we have plotted the spatial distribution of prediction accuracy for quantitative analysis, as shown in Fig. 15.

From Fig. 15, it is easy to find that the prediction accuracy difference based on the two data sets (50% and 70% cloud) is small and the spatial distribution is similar. A possible reason for this is that those small areas of cloud cover do not fully cover crop areas with main crop types. That is, there is still enough number of representative samples for the model to learn the SAR-Optics relationship. In this way, it seems that this method can learn local knowledge and then predict missing data. By contrast, large areas of cloud cover the vast majority of arable land, limiting the performance of the model. It should be noted that the prediction images of two data sets have very low accuracy for some land cover types such as non-crop land. This suggests that the lower performances may be related to the land cover types. From this point of view, the model may not be suitable for areas where backscattering is dominated by soil scattering while the vegetation volume scattering is relatively weak. On the one hand, after the crop been harvested, the consistency between the predicted time series and the real-time series is low, largely due to the weak contribution of vegetation volume scattering. On the other hand, the accuracy of the model has been increased after the second growth cycle of the crop. As the stem elongation period is entered, both the number of stems and the length of the crop, as well as the number of crops are increased, resulting in an increasing contribution of volume scattering.

In the two prediction images, the reliability of the pixels in the black circle is very low. This mainly because more non-crop land cover and hay can be observed, while fewer crops were identified. These factors affect the model to make incorrect judgments and failed to predict a reliable NDVI time series profile. Moreover, the optical time-series data will be affected by noise, such as atmospheric correction residual errors and surface heterogeneous. These effects make the original optical data may limit the performance of the model, unless smoothing procedures are previously applied prior to the experiment. However, the process of NDVI filtering may remove the key information for crops. For example, the NDVI time series of alfalfa could be over-smoothed, where the NDVI curves of the beet could change abruptly in March.

In summary, there are two main factors that restrain the accuracy of the predicted images. First, the model cannot build a reliable relationship for non-crop land, which reduces the statistical accuracy of the entire image. The other is that the noises in the optical time series weaken the performance of the model and increase the prediction errors. For the former, non-crop land can be masked in advance. For the latter, one can choose whether to smooth the optical time series according to the characteristics of the study area or the purpose of the study.

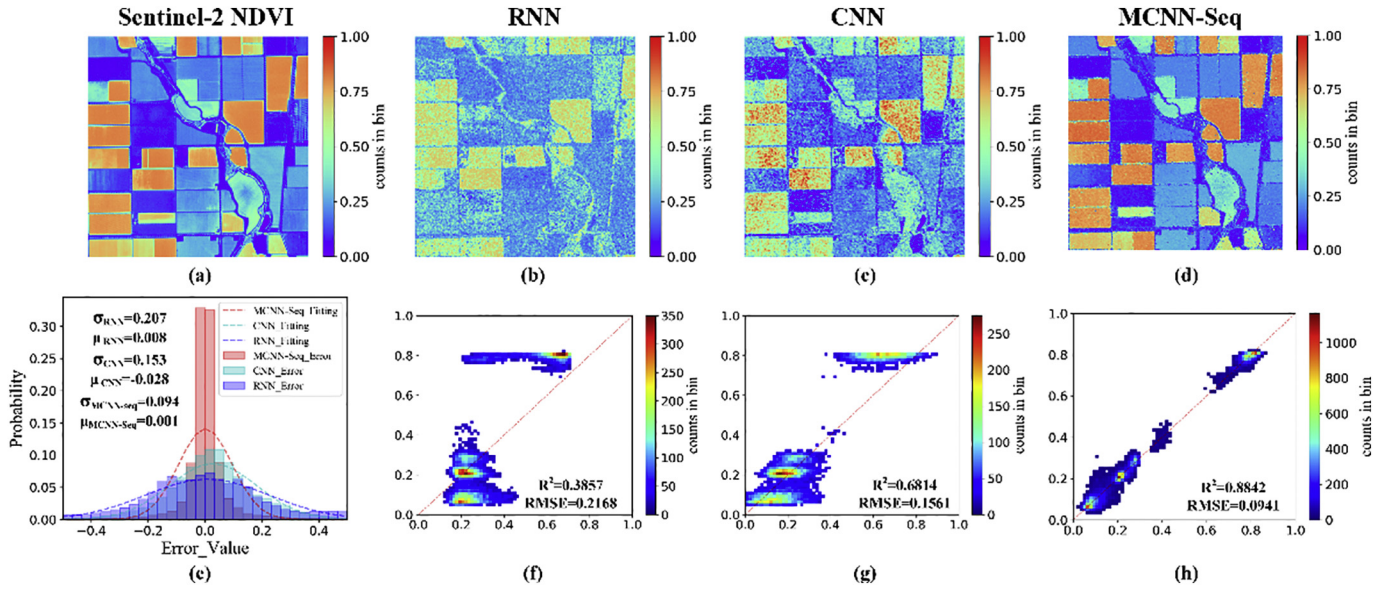

Fig. 13. Comparisons between the reference and the predicted image by RNN, CNN and MCNN-Seq for dates: 2018/01/11. (a), Sentinel-2 NDVI reference and(b), (c), (d)predicted images are compared visually, (e) error histograms and (f), (g), (h)density scatter plot compared from statistical perspective.

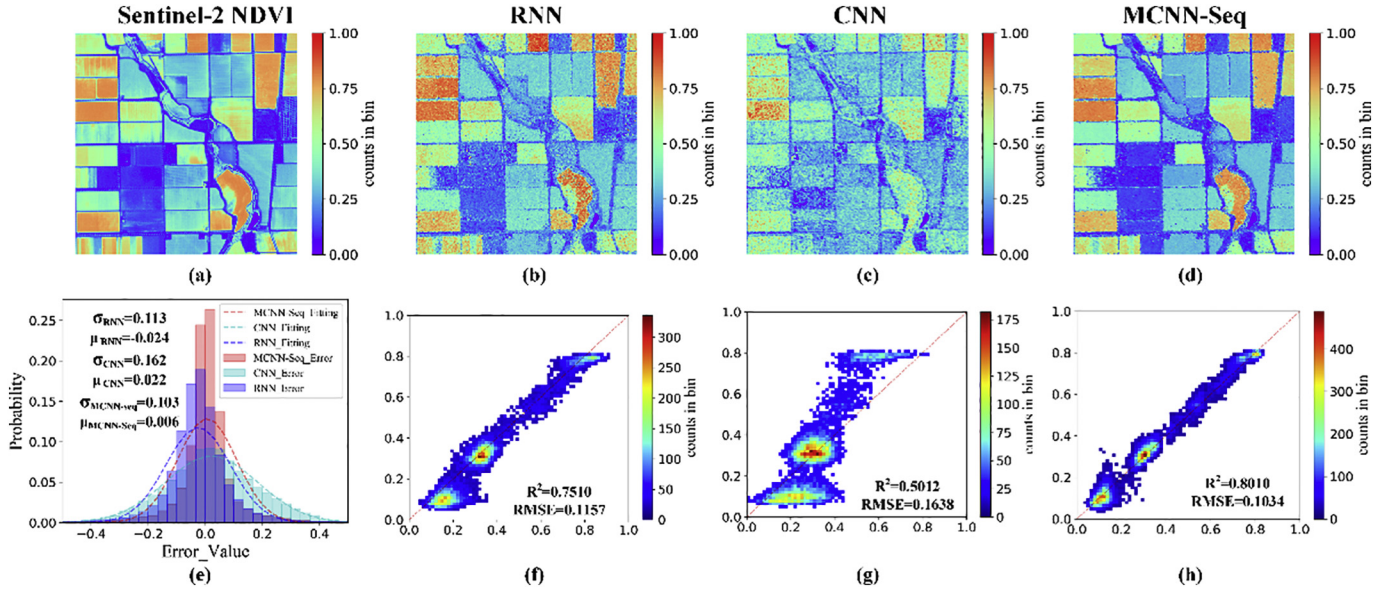

Fig. 14. Comparisons between the reference and the predicted images by RNN, CNN and MCNN-Seq for dates: 2018/12/27. (a) Sentinel-2 NDVI reference and(b), (c), (d)predicted images are compared visually, (e) error histograms and (f), (g), (h) density scatter plot compared from statistical perspective.

5.2. Pros and cons

In recent years, researches have focused on the coordination of SAR data and optical images which aim to convert one sequence to another. To achieve this purpose, the optical and SAR data on the same or adjacent dates have been utilized for regression fitting. Although these methods have achieved the conversion of SAR data to optical data. However, due to the complexity of crop areas, the results are not always satisfactory. Differently, this study proposes a deep learning method for the synergy between optical imagery and SAR data.

The previous studies mainly focused on building the translation model for optical image and SAR data converting, while neglecting the temporal relationship inside of time series data. As the cropland evolving along with time, the existing models cannot capture useful information in the temporal domain, but only increase the pressure of model calculation. This means that these methods are difficult to apply time series conversion with the temporal axis, while this application demand is mostly needed in researches such as change detection and land cover classification. In contrast, MCNN-Seq can extract the underlying temporal relationships inside time-series. From this point of view, the proposed model is more suitable for applications that requiring longer time series images than the previous methods. In addition, due to the complexity of SAR data, it is difficult to construct the reliable relationship between optical and SAR data. Thanks to the stable feature extraction capability provided by the MCNN-Seq, it is feasible to handle the mentioned problem.

In the previous section, we proved that it is feasible to build the relationship between optical and SAR time-series by MCNN-Seq. However, there are some shortcomings in this work. The first is that the high calculation cost of the model. For a large study area, the proposed model will take a lot of time to formulate SAR-NDVI relationship. This mainly contributed by the computational complexity of the model, and

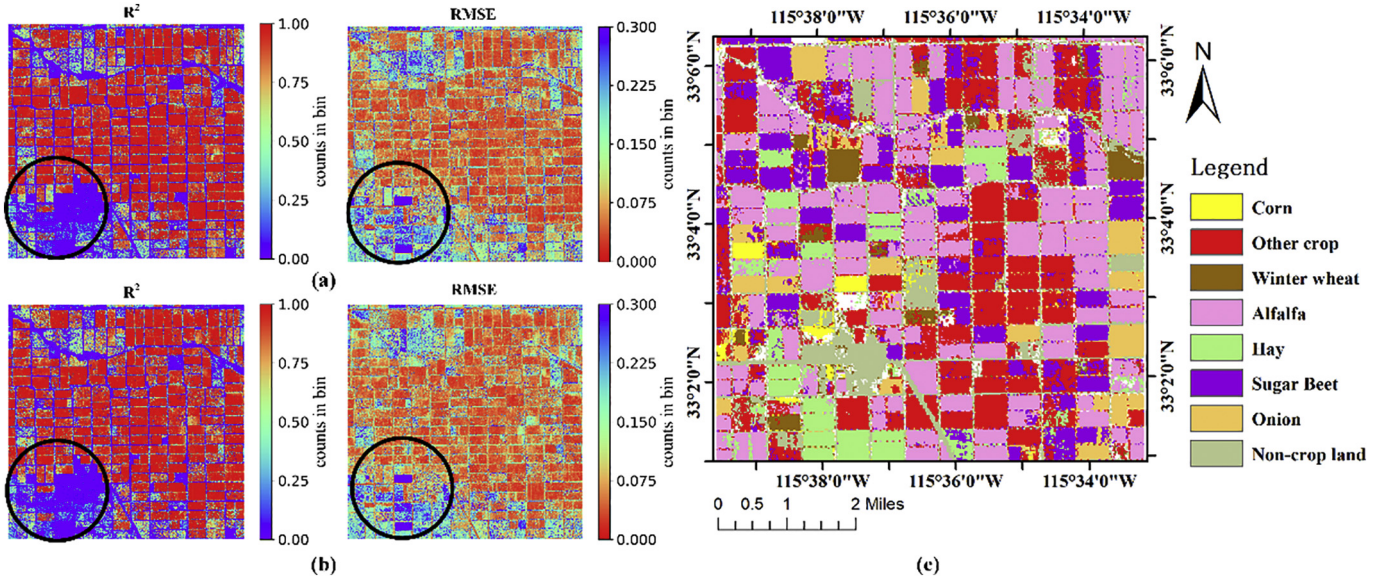

Fig. 15. Spatial distribution of estimation errors. (a) the estimation error maps of the time series are predicted by 50% cloud, (b) the estimation error maps of the time series are predicted by 70%cloud. (c) the CDL data of 2018 by USDA NASS. The black circle area with low accuracy

too many types of crops also make the model difficult to run. Another weakness of the model is that it requires a large number of training samples, which means that the area of cloud cover in the applicable area should not be too large. Moreover, the training samples need to be highly representative, covering as many types as possible. The purpose of this study is trying to fill the time gaps of optical data by using SAR data, which makes us focus on temporal profile conversion and ignores the spatial correlation between the two data sources. In future works, we should explore how to take advantage of the temporal and spatial complementarity of SAR and optical data at the same time in order to obtain more stable model performances.

6. Conclusion

This study proposed a strategy to predict optical time series using SAR data when the optical data was missing. Unlike most studies, the model was designed to realize the prediction of SAR data to optical data by constructing a relationship between two time-series. The experiment results demonstrated that predicted images have reasonable accuracies (i.e., $R^2 > 0.7$, RMSE < 0.15). Each predicted time series can accurately describe the growth cycle of most crops, especially onion, winter wheat, corn, and sugar beet. Therefore, the predicted images have the ability to provide reliable replaceable information when the optical data has a long data gap due to the persistent cloud cover. However, it is important to note that large cloud cover may affect the performance of the model. Because too much cloud may cover most crops, it is difficult for the model to learn enough knowledge. In addition, comparative experiments demonstrate the importance of contextual information. For instance, the prediction accuracy of 2018/01/11 (R^2 of 0.3857, RMSE of 0.2168) is much lower than that of 2018/12/11(R^2 of 0.7510, RMSE of 0.1157) for the RNN. In summary, this study provides new solutions for the translation of optical and SAR data. However, there are still some problems that need to be further improved, such as the predicted image contains more speckles, and a larger number of representative training data is needed. Meanwhile, it is hoped that the accuracy of prediction can be improved by combining spatial information and temporal information.

Authorship contributions

Dr. Wenzhi Zhao conceived the idea of MCNN-seq for synergistic optical and SAR time series; Mr. Yang Qu constructed the experimental framework and gathered data sets of the study area; Dr. Jiage Chen helped with the experiments and results analysis; Prof. Zhanliang Yuan offered help in paper revision and language proof-reading.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This research was supported by the National Key Research and Development Program of China (Grant No. 2016YFB0501502), and the Fundamental Research Funds for the Central Universities (Grant No. 2018NTST01).

References

Bahdanau, D., Cho, K., Bengio, Y., 2014. Neural Machine Translation by Jointly Learning to Align and Translate. (arXiv preprint arXiv:1409.0473).
Ban, Y., Yousif, O.A., 2012. Multitemporal spaceborne SAR data for urban change detection in China. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 5, 1087–1094.
Bargiel, D., 2013. Capabilities of high resolution satellite radar for the detection of semi-natural habitat structures and grasslands in agricultural landscapes. Ecol. Inform. 13, 9–16.
Belgiu, M., Csillik, O., 2018. Sentinel-2 cropland mapping using pixel-based and object-based time-weighted dynamic time warping analysis. Remote Sens. Environ. 204, 509–523.
Bengio, Y., Courville, A., Vincent, P., 2013. Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 35, 1798–1828.
Boryan, C., Yang, Z., Mueller, R., Craig, M., 2011. Monitoring US agriculture: the US department of agriculture, national agricultural statistics service, cropland data layer program. Geocarto Int. 26, 341–358.
Bousbih, S., Zribi, M., Lili-Chabaane, Z., Baghdadi, N., El Hajj, M., Gao, Q., Mougenot, B., 2017. Potential of Sentinel-1 radar data for the assessment of soil and cereal cover parameters. Sensors. 17, 2617.
Canizo, M., Triguero, I., Conde, A., Onieva, E., 2019. Multi-head CNN–RNN for multi-time series anomaly detection: an industrial case study. Neurocomputing. 363, 246–260.
Claverie, M., Masek, J.G., Ju, J., Dungan, J.L., 2017. Harmonized landsat-8 sentinel-2 (HLS) product user's guide. National Aeronautics and Space Administration (NASA), Washington, DC, USA.
Denize, J., Hubert-Moy, L., Betbeder, J., Corgne, S., Baudry, J., Pottier, E., 2019. Evaluation of using sentinel-1 and-2 time-series to identify winter land use in agricultural landscapes. Remote Sens. 11, 37.
Dong, J., Xiao, X., Kou, W., Qin, Y., Zhang, G., Li, L., Jin, C., Zhou, Y., Wang, J., Biradar,

C., 2015. Tracking the dynamics of paddy rice planting area in 1986–2010 through time series Landsat images and phenology-based algorithms. Remote Sens. Environ. 160, 99–113.
Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.
Dusseux, P., Corpetti, T., Hubert-Moy, L., Corgne, S., 2014. Combined use of multi-temporal optical and radar satellite images for grassland monitoring. Remote Sens. 6, 6163–6182.
Frazier, R.J., Coops, N.C., Wulder, M.A., 2015. Boreal shield forest disturbance and recovery trends using Landsat time series. Remote Sens. Environ. 170, 317–327.
Gamba, P., Dell'Acqua, F., Lisini, G., 2006. Change detection of multitemporal SAR data in urban areas combining feature-based and pixel-based techniques. IEEE Trans. Geosci. Remote Sens. 44, 2820–2827.
Gao, Q., Zribi, M., Escorihuela, M.J., Baghdadi, N., 2017. Synergetic use of Sentinel-1 and Sentinel-2 data for soil moisture mapping at 100 m resolution. Sensors. 17, 1966.
Gao, J., Yuan, Q., Li, J., Zhang, H., Su, X., 2020. Cloud removal with fusion of high resolution optical and SAR images using generative adversarial networks. Remote Sens. 12, 191.
He, W., Yokoya, N., 2018. Multi-temporal sentinel-1 and-2 data fusion for optical image simulation. ISPRS Int. Geo-Inf. 7, 389.
Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.
Ienco, D., Interdonato, R., Gaetano, R., Minh, D.H.T., 2019. 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. 158, 11–22.
Inglada, J., Vincent, A., Arias, M., Marais-Sicre, C., 2016. Improved early crop type identification by joint use of high temporal resolution SAR and optical image time series. Remote Sens. 8, 362.
Inglada, J., Vincent, A., Arias, M., Tardy, B., Morin, D., Rodes, I., 2017. Operational high resolution land cover map production at the country scale using satellite image time series. Remote Sens. 9, 95.
Johnson, J.A., Runge, C.F., Senauer, B., Foley, J., Polasky, S., 2014. Global agriculture and carbon trade-offs. Proc. Natl. Acad. Sci. 111, 12342–12347.
Julien, Y., Sobrino, J.A., 2010. Comparison of cloud-reconstruction methods for time series of composite NDVI data. Remote Sens. Environ. 114, 618–625.
Kim, Y., Jackson, T., Bindlish, R., Lee, H., Hong, S., 2011. Radar vegetation index for estimating the vegetation water content of rice and soybean. IEEE Geosci. Remote Sens. Lett. 9, 564–568.
Kussul, N., Lavreniuk, M., Skakun, S., Shelestov, A., 2017. Deep learning classification of land cover and crop types using remote sensing data. IEEE Geosci. Remote Sens. Lett. 14, 778–782.
LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. nature. 521, 436–444.
Liu, J., Pattey, E., Miller, J.R., McNairn, H., Smith, A., Hu, B., 2010. Estimating crop stresses, aboveground dry biomass and yield of corn using multi-temporal optical data combined with a radiation use efficiency model. Remote Sens. Environ. 114, 1167–1177.
Liu, H., Mi, X., Li, Y., 2018. Smart deep learning based wind speed prediction model using wavelet packet decomposition, convolutional neural network and convolutional long short term memory network. Energy Conv. Manag. 166, 120–131.
Lu, X., Liu, R., Liu, J., Liang, S., 2007. Removal of noise by wavelet method to generate high quality temporal data of terrestrial MODIS products. Photogramm. Eng. Remote Sens. 73, 1129–1139.
Lu, L., Tao, Y., Di, L., 2018. Object-based plastic-mulched landcover extraction using integrated Sentinel-1 and Sentinel-2 data. Remote Sens. 10, 1820.
Lyu, H., Lu, H., Mou, L., 2016. Learning a transferable change rule from a recurrent neural network for land cover change detection. Remote Sens. 8, 506.
Mattia, F., Le Toan, T., Picard, G., Posa, F.I., D'Alessio, A., Notarnicola, C., Gatti, A.M., Rinaldi, M., Satalino, G., Pasquariello, G., 2003. Multitemporal C-band radar measurements on wheat fields. IEEE Trans. Geosci. Remote Sens. 41, 1551–1560.
Minh, D.H.T., Ienco, D., Gaetano, R., Lalande, N., Ndikumana, E., Osman, F., Maurel, P., 2018. Deep recurrent neural networks for winter vegetation quality mapping via multitemporal SAR Sentinel-1. IEEE Geosci. Remote Sens. Lett. 15, 464–468.
Mou, L., Ghamisi, P., Zhu, X.X., 2017. Unsupervised spectral–spatial feature learning via deep residual Conv–Deconv network for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 56, 391–406.
Mou, L., Bruzzone, L., Zhu, X.X., 2018. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Trans. Geosci. Remote Sens. 57, 924–935.
Ozdogan, M., Woodcock, C.E., 2006. Resolution dependent errors in remote sensing of cultivated areas. Remote Sens. Environ. 103, 203–217.
Pipia, L., Muñoz-Marí, J., Amin, E., Belda, S., Camps-Valls, G., Verrelst, J., 2019. Fusing optical and SAR time series for LAI gap fillingwith multioutput Gaussian processes. Remote Sens. Environ. 235, 111452.
Qian, Y., Bi, M., Tan, T., Yu, K., 2016. Very deep convolutional neural networks for noise robust speech recognition. IEEE-ACM Trans. Audio Speech Lang. 24, 2263–2276.
Reiche, J., Verbesselt, J., Hoekman, D., Herold, M., 2015. Fusing Landsat and SAR time series to detect deforestation in the tropics. Remote Sens. Environ. 156, 276–293.
Scarpa, G., Gargiulo, M., Mazza, A., Gaetano, R., 2018. A cnn-based fusion method for feature extraction from sentinel data. Remote Sens. 10, 236.
Schmitt, M., Hughes, L.H., Zhu, X.X., 2018. The SEN1–2 Dataset for Deep Learning in SAR-Optical Data Fusion. (arXiv preprint arXiv:1807.01569).
Shao, Z., Cai, J., 2018. Remote sensing image fusion with deep convolutional neural network. IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens. 11, 1656–1669.
Sharma, R.C., Hara, K., Tateishi, R., 2018. Developing forest cover composites through a combination of landsat-8 optical and sentinel-1 SAR data for the visualization and extraction of forested areas. Journal of Imaging. 4, 105.
Shen, H., Li, X., Cheng, Q., Zeng, C., Yang, G., Li, H., Zhang, L., 2015. Missing information reconstruction of remote sensing data: a technical review. IEEE Geosci. Remote Sens. Mag. 3, 61–85.
Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. In. In: Advances in Neural Information Processing Systems, pp. 3104–3112.
Van Tricht, K., Gobin, A., Gilliams, S., Piccard, I., 2018. Synergistic use of radar Sentinel-1 and optical Sentinel-2 imagery for crop mapping: a case study for Belgium. Remote Sens. 10, 1642.
Veloso, A., Mermoz, S., Bouvet, A., Le Toan, T., Planells, M., Dejoux, J.-F., Ceschia, E., 2017. Understanding the temporal behavior of crops using Sentinel-1 and Sentinel-2-like data for agricultural applications. Remote Sens. Environ. 199, 415–426.
Wang, Q., Blackburn, G.A., Onojeghuo, A.O., Dash, J., Zhou, L., Zhang, Y., Atkinson, P.M., 2017. Fusion of Landsat 8 OLI and Sentinel-2 MSI data. IEEE Trans. Geosci. Remote Sens. 55, 3885–3899.
Wang, L., Xu, X., Yu, Y., Yang, R., Gui, R., Xu, Z., Pu, F., 2019. SAR-to-optical image translation using supervised cycle-consistent adversarial networks. IEEE Access. 7, 129136–129149.
Wardlow, B.D., Egbert, S.L., 2008. Large-area crop mapping using time-series MODIS 250 m NDVI data: an assessment for the US central Great Plains. Remote Sens. Environ. 112, 1096–1116.
White, J.C., Wulder, M.A., Hermosilla, T., Coops, N.C., Hobart, G.W., 2017. A nationwide annual characterization of 25 years of forest disturbance and recovery for Canada using Landsat time series. Remote Sens. Environ. 194, 303–321.
Whyte, A., Ferentinos, K.P., Petropoulos, G.P., 2018. A new synergistic approach for monitoring wetlands using Sentinels-1 and 2 data with object-based machine learning algorithms. Environ. Model. Softw. 104, 40–54.
Xiao, X., Boles, S., Liu, J., Zhuang, D., Frolking, S., Li, C., Salas, W., Moore III, B., 2005. Mapping paddy rice agriculture in southern China using multi-temporal MODIS images. Remote Sens. Environ. 95, 480–492.
Xingjian, S., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., Woo, W.-c, 2015. Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting. In, Advances in neural information processing systems. pp. 802–810.
Zeng, C., Shen, H., Zhang, L., 2013. Recovering missing pixels for Landsat ETM+ SLC-off imagery using multi-temporal regression analysis and a regularization method. Remote Sens. Environ. 131, 182–194.
Zhang, X., Qin, F., Qin, Y., 2010. Study on the thick cloud removal method based on multi-temporal remote sensing images. In: 2010 International Conference on Multimedia Technology. IEEE, pp. 1–3.
Zhang, J., Clayton, M.K., Townsend, P.A., 2014. Missing data and regression models for spatial images. IEEE Trans. Geosci. Remote Sens. 53, 1574–1582.
Zhao, W., Du, S., 2016. Spectral–spatial feature extraction for hyperspectral image classification: a dimension reduction and deep learning approach. IEEE Trans. Geosci. Remote Sens. 54, 4544–4554.
Zhao, W., Du, S., Wang, Q., Emery, W.J., 2017. Contextually guided very-high-resolution imagery classification with semantic segments. ISPRS-J. Photogramm. Remote Sens. 132, 48–60.
Zhao, W., Bo, Y., Chen, J., Tiede, D., Thomas, B., Emery, W.J., 2019. Exploring semantic elements for urban scene recognition: deep integration of high-resolution imagery and OpenStreetMap (OSM). ISPRS-J. Photogramm. Remote Sens. 151, 237–250.
Zhao, X., Jiang, N., Liu, J., Yu, D., Chang, J., 2020. Short-term average wind speed and turbulent standard deviation forecasts based on one-dimensional convolutional neural network and the integrate method for probabilistic framework. Energy Conv. Manag. 203, 112239.
Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. Remote Sens. Environ. 221, 430–443.