ISPRS Journal of Photogrammetry and Remote Sensing 207 (2024) 312–325

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Self-supervised pre-training for large-scale crop mapping using Sentinel-2 time series

Yijia Xu, Yuchi Ma, Zhou Zhang *

Department of Biological Systems Engineering, University of Wisconsin–Madison, USA

ARTICLE INFO

Keywords:
Crop mapping
Remote sensing
Transformer
Self-supervised learning
Contrastive learning

ABSTRACT

Large-scale crop mapping is essential for various agricultural applications, such as yield prediction and agricultural resource management. State-of-the-art techniques for crop mapping utilize deep learning (DL) models on satellite imagery time series (SITS). Despite advancements, the efficacy of DL-based crop mapping methods is impeded by the arduous task of acquiring crop-type labels and the extensive pre-processing required on satellite data. To address these issues, we proposed a Transformer-based DL model and a self-supervised pre-training framework for the label-scarce crop mapping task. Specifically, we first developed a Transformer-based Spectral Temporal Network (STNet) which is designed to extract task-informative features from time-series remote sensing (RS) imagery via the self-attention mechanism. A self-supervised pre-training strategy, namely SITS-MoCo, was then proposed to learn robust and generalizable representations from time-series RS imagery that is invariant to spectral noise, temporal shift, and irregular-length data. To evaluate the proposed framework, experiments were conducted using Sentinel-2 time series and high-confident Cropland Data Layer (CDL) reference data on six geographically scattered study sites across the United States from 2019 to 2021. The experimental results demonstrated that the framework had superior performance in comparison to other advanced DL models and self-supervised pre-training techniques. The pre-training strategy was proven to effectively alleviate the need for complex data pre-processing and training labels for the downstream crop mapping task. Overall, this research presented a novel pipeline for improving model performance on large-scale crop mapping with limited labels and provided a viable solution to efficiently exploit available satellite data that can be easily adapted to other large-area classification tasks. The source code is available at: https://github.com/YXu556/SITS-MoCo.

1. Introduction

Crop mapping is the process of classifying crops and mapping the locations and extent of land used for crop production (Wang et al., 2021a). Accurate crop mapping is essential to food security, which helps policymakers and farmers make informed decisions about land use and crop rotation (Kluger et al., 2022). It can also assist farmers in understanding cropping patterns for agricultural resource management and market planning (Wang et al., 2019). Moreover, crop mapping can serve as an important foundation for various agricultural applications, such as crop yield prediction (Ma et al., 2023), and cover crop management (Deines et al., 2023). Therefore, crop mapping has gained significant attention within the agricultural community.

Traditionally, crop mapping is generally based on surveys and censuses, which are labor-intensive, time-consuming, and with limited scalability (Wang et al., 2019). With the increasing accessibility of satellite remote sensing (RS) images and the development of deep learning (DL) algorithms, researchers have utilized DL methods for large-scale crop mapping based on RS imagery. Specifically, satellite RS imagery provides observations on the Earth's surface with decent temporal intervals (e.g., 5 days for Sentinel-2 and 16 days for Landsat-8). Satellite imagery time series (SITS) captures the progress of crop growth, i.e., crop phenology, which can be used for monitoring the crop cover and distinguishing different crop types (Wang et al., 2019; Zhong et al., 2014). On the other hand, DL has the advantage of automatic feature learning without the need for explicit programming or knowledge of physiological mechanisms on individual crops (Ma et al., 2021). Therefore, DL-based crop mapping using SITS has made significant progress. Two of the most commonly employed DL architectures for crop mapping are Convolutional neural networks (CNN) and recurrent neural networks (RNN) (Pelletier et al., 2019; Wang et al., 2021b; Xu et al., 2020; Yu et al., 2021; Zhong et al., 2014). For example, Zhong et al.

* Corresponding author.
E-mail address: zzhang347@wisc.edu (Z. Zhang).

https://doi.org/10.1016/j.isprsjprs.2023.12.005
Received 19 April 2023; Received in revised form 5 December 2023; Accepted 11 December 2023
Available online 5 January 2024

(2019) applied convolutions on the time dimension to capture multi-level temporal features for crop classification and demonstrated the superiority of the one-dimensional convolutional (Conv1D) structure for handling the temporal domain for crop mapping on time series imagery. Xu et al. (2020) designed an RNN with the Long Short-Term Memory (LSTM) structure to integrate multi-temporal and multi-spectral RS data for corn and soybean classification in the U.S. corn belt. However, both the recurrent layers and temporal convolutions can only capture the relative relationships in the time series data, i.e., the order of observations. Due to the presence of noise caused by clouds and shadows and varying revisit periods at different sites, SITS may contain missing acquisition and invalid observations, resulting in irregular time series among SITS samples. Consequently, existing models based on CNN or RNN require temporal interpolation to impute missing values in SITS (Pelletier et al., 2019; Wang et al., 2021b; Xu et al., 2020), allowing the model to obtain correct relative relationships among observations. However, the interpolation process is not only computationally demanding, but also tends to neglect the explicit temporal information and non-linear phenomena in the original time series. The Transformer (Vaswani et al., 2017) has been promising to address the issue, which uses positional encoding (PE) to encode the position of each imagery according to the observation day of year (DoY) and uses the self-attention mechanism to focus on most informative features. This capability is crucial for SITS crop mapping since it allows the Transformer model to capture explicit temporal dependencies. Recent studies have demonstrated the superiority of the Transformer in crop mapping tasks (Rußwurm and Körner, 2020; Yuan and Lin, 2021; Garnot and Landrieu, 2020; Garnot et al., 2020). For example, Garnot et al. (2020) combined a pixel-set encoder and a modified self-attention module for object-based crop classification, which showed improved performance over RNN-based approaches. Rußwurm & Körner (2020) demonstrated that the self-attention mechanism in Transformer had the ability to suppress noise observations and performed well on crop mapping using raw satellite time series.

Despite the success, it commonly requires a large amount of data samples with ground truth label information (i.e., crop type information) to train an effective Transformer-based crop mapping model (Yuan & Lin, 2021). In practice, obtaining crop-type information can be particularly costly as it involves expensive in-site surveys or expert knowledge for photo interpretation (Mañas et al., 2021; Neumann et al., 2019). As a result, labeled data samples can be insufficient in some agricultural regions, which makes it hard to train a reliable Transformer model. Also, due to domain shifts between different geographic regions, DL models trained in one region can have poor performance if directly applied to the other region. To address domain shift and improve DL models' transferability, transfer learning (TL) is a promising strategy, in which the model is trained to transfer the knowledge learned from a data-abundant area (i.e., source domain) to tackle a task in the data-scarce area (i.e., target domain) (Pan & Yang, 2010). Fine-tuning-based transfer learning (FTL) is the most widely used TL approach, in which a model is first pre-trained in the source domain and then adapted to the target domain by fine-tuning it with a limited number of ground truth data. For example, Bazzi et al. (2020) proposed to train a teacher model in the source domain, which is then distilled into a student model and successively refined using data samples from the target domain to map irrigated areas using time series Sentinel-1 imagery. Besides, unsupervised domain adaptation (UDA) is another TL approach that has been widely used to improve the model's transferability in SITS crop mapping. The idea of UDA is to reduce domain shift by aligning the feature distributions in the source and target domain. For example, Wang et al. (2021b) utilized maximum mean discrepancy (MMD) loss to minimize the divergence of features extracted from three geo-scattered plains of China and achieved satisfactory performance on cross-regional winter crop classification. However, both FTL and UDA require a large, high-quality, labeled dataset from the source domain (Neumann et al., 2019), and the effectiveness of the model varies largely according to the domain shift between source and target domains. When source and target domains are distant from each other with different environments or crop types, either FTL or UDA could not effectively address domain shift (Nyborg et al., 2022), which limited the application of these approaches for large-scale SITS crop classification.

To reduce the reliance on the source domain with high-quality labeled datasets, self-supervised learning can be a promising alternative, in which a model is pre-trained to learn from a set of unlabeled data without any explicit supervision from labeled data. The idea behind self-supervised learning is to leverage the inherent structure or patterns in the data to train a model to learn useful representations or features of the data. The pre-trained model is then fine-tuned with labeled data samples to adapt it to the downstream tasks. Given that unlabeled data, such as RS imagery, is often easily accessible, self-supervised pre-training can be effectively used to assist the model in acquiring reliable and adaptable representations. Thus, even if the downstream task lacks labeled data, pre-trained models can prevent overfitting and deliver robust performance after fine-tuning. In self-supervised learning, contrastive learning is one of the most widely used strategies (Chen et al., 2020a; Chen et al., 2020b; He et al., 2020; van den Oord et al., 2019). The idea of contrastive learning is to learn representations that are common among similar samples (i.e., positive pairs) while distinguishing them from dissimilar samples (i.e., negative pairs). Generally, the positive pair comprises two random "views" of the same sample captured through random data augmentation (He et al., 2020), while the negative pairs come from other samples in the dataset. The learning task in contrastive learning is to minimize a contrastive loss (Hadsell et al., 2006), which brings the representation of positive pairs closer while pushing the representations of negative pairs farther away. In the field of RS, several studies have explored the use of contrastive learning for single-image applications. For example, Jung et al. and Jean et al. utilized the spatial distribution of RS images to construct positive pairs considering that features extracted from images that are closer in distance should be more similar to each other (Jean et al., 2018; Jung et al., 2022). Mañas et al. (2021) leveraged temporal information to construct positive pairs based on the idea that images from the same location at different times should encode as similar representations. However, these approaches cannot be directly applied to SITS data because image-specific augmentation methods are not applicable to temporal sequences. The application of contrastive learning to address label scarcity in SITS crop classification is a largely underexplored area.

In this paper, we proposed a contrastive learning-based pre-training/fine-tuning framework for SITS crop classification for the purpose of mitigating the need for crop-type labels. In this framework, a Transformer-based DL model, named Spectral-Temporal Network (STNet), is developed for crop classification using less-preprocessed SITS. Moreover, a self-supervised pre-training scheme is designed to learn robust and generalizable representations to cope with the problem of insufficient labeled samples. Since the pre-training strategy is inspired by MoCo (He et al., 2020), we name it SITS-MoCo. The pre-trained model is then fine-tuned on a small labeled dataset to adapt it to the crop mapping task. By leveraging the rich knowledge gained from self-supervised pre-training in SITS, the accuracy of the model can be improved in the downstream task of crop mapping, and the need for labeled crop mapping data can be reduced. We evaluated the proposed framework in the Contiguous United States (CONUS). Specifically, an unlabeled dataset of Sentinel-2 time series imagery was collected over the CONUS for self-supervised learning. Meanwhile, the labeled dataset based on the CDL was collected in six distributed counties across the U.S. for crop mapping. During the experiments, the STNet was first pre-trained on the unlabeled dataset via SITS-MoCo and then adapted to the downstream task of crop mapping via fine-tuning on the labeled dataset. To validate the effectiveness of our proposed scheme, we compared STNet and fine-tuned STNet (i.e., STNet + SITS-MoCo) with five widely used machine learning (ML) methods in SITS classification. Also, SITS-MoCo was adapted to various DL models to further evaluate

its flexibility and compare it with the SITS-BERT pre-training. Finally, the self-attention mechanism in STNet and the effectiveness of SITS-MoCo for small numbers of training data were interpreted.

2. Materials

In this study, two datasets were organized for model training, including an unlabeled dataset for self-supervised learning and a labeled crop classification dataset for crop mapping. The unlabeled dataset was randomly sampled across the U.S. to cover a diverse range of land cover types. The labeled crop classification dataset was collected from six spatially distributed counties of the CONUS to evaluate the efficiency of the method for large-scale crop mapping. Both datasets used Sentinel-2 multi-spectral image time series. A detailed description of each dataset is provided below.

2.1. Sentinel-2 data

The Sentinel-2 L2A Bottom-Of-Atmosphere (BOA) reflectance data was used in this study (Drusch et al., 2012). Three atmospheric bands (i.e., Aerosols, Water vapor, and Cirrus) were removed and the remaining 10 spectral bands were kept as the model inputs. The detailed information on each band is listed in Table 1. We used Google Earth Engine (Gorelick et al., 2017) to process and download the Sentinel-2 imagery time series. Specifically, the built-in cloud mask band (QA60) of Sentinel-2 was first used to mask out cloud and cirrus-contaminated pixels. All available images with cloud coverage $\leq$ 80% between January 1 to December 31 of each testing year were collected. After that, we kept all available observations with corresponding DoYs to preserve explicit temporal information within the growing season, which can provide specific timing for phenological events of different crop types. It is worth noting that no linear interpolation was applied on the temporal dimension to preserve the complete non-linear crop phenology information.

2.2. Unlabeled dataset for self-supervised learning

A vast number of Sentinel-2 SITS were collected across the U.S. CONUS to create a large unlabeled dataset. This dataset is used for self-supervised learning, where the model learns rich knowledge from the RS data itself (e.g., the inherent temporal and spectral relationship). To ensure the model learns representative features, this unlabeled dataset should cover multiple land cover types from various geographic conditions. Therefore, data points were randomly sampled across the CONUS to ensure the diversity of the dataset. In this work, we followed the previous studies and collected 1 million data samples with varying sequential lengths (He et al., 2020; Mañas et al., 2021). The unlabeled dataset used to learn representations only contained data from 2019. These representations were then adapted to downstream tasks in different years to evaluate the inter-annual generalizability of the learned representations.

Table 1
Information of 10 spectral bands of Sentinel-2 data.

Band #	Name	Spatial Resolution (m)	Wavelength (nm)
B2	Blue	10	490
B3	Green	10	560
B4	Red	10	665
B5	Red Edge 1	20	705
B6	Red Edge 2	20	740
B7	Red Edge 3	20	783
B8	NIR	10	842
B8A	Red Edge 4	20	865
B11	SWIR 1	20	1610
B12	SWIR 2	20	2190

2.3. Labeled dataset for crop mapping

2.3.1. Study areas

For the crop mapping task, six geo-distributed counties in the U.S. were chosen as the study areas. As shown in Fig. 1 (a), these study sites were selected to cover the major crop production regions, including the Pacific area (site A), the Great Plains (site B, D, and F), the Corn Belt (site C), and the Mississippi Valley (site E). The six sites exhibit varying environmental conditions such as soil type, climate, and farming practices, resulting in temporal changes in crop phenology across the sites (Nyborg et al., 2022). According to Commission for Environmental Cooperation (CEC) (CEC, 1997), these study sites belong to three different ecoregions. Site A (Garfield, WA) belongs to the North American Desert (NAD) ecoregion, which is characterized by aridity and freezing temperatures in winter. The primary crop is spring wheat in Site A. Site B (Adams, ND), Site D (Harvey, KS), and Site F (Haskell, TX) are located in the Great Plains (GP) ecoregion, but their climates are diverse due to the large range from north to south. Site B is typified by large seasonal differences with severely cold winters, and thus spring wheat dominates the crop production in this site. Site D is in the middle latitude, with warm summers, mild winters, and moderate humidity. Grain farming (e.g., corn and soybean) accounts for a major part of local agriculture. Site F has a hot humid subtropical climate and belongs to the main cotton-producing area in the south. Site C (Randolph, IN) and Site E (Coahoma, MS) are in the Eastern Temperate Forests (ETF) ecoregion. Site C is close to the Lakes. It is humid and cloudy all year round, with comfortable summers and freezing winters. The main crops are corn and soybeans. Site E is located in the Mississippi Valley, where the summer is long and hot, and the rainfall is fairly even throughout the year. These differences increase the diversity of our crop classification dataset and enable the evaluation of our method in large-scale crop mapping.

2.3.2. Ground reference data

The CDL was used as ground reference data in this study. CDL is a crop-specific land cover map for the CONUS produced annually by USDA NASS. It was generated using moderate-resolution satellite imagery and extensive agricultural ground truth with a spatial resolution of 30 m. Besides the land cover classes, the CDL also provides the identification of cultivated and non-cultivated land and the predicted confidence of the given classification. The land cover classes have a long-tail distribution of more than 130 distinct classes, where the most common 13, 17, and 31 classes cover 90%, 95%, and 99% of the total pixels, respectively, according to the 2019 CDL. We used GEE to download CDL over the study sites from 2019 to 2021. To ensure the accuracy of the ground reference data, high-confidence pixels with confidence greater than 95% were reserved to form the training and testing datasets. We kept the largest top-20 classes, including Grass/Pasture (GRP), Soybeans (SOY), Corn (COR), Winter Wheat (WWT), Spring Wheat (SWT), Cotton (COT), Woody Wetlands (WWL), Evergreen Forest (EFT), Shrubland (SHR), Fallow/Idle Cropland (FIC), Canola (CNL), Sunflower (SFL), Deciduous Forest (DFT), Double Crop Winter Wheat/Soybeans (DWS), Alfalfa (ALF), Rice (RIC), Other Hay/Non Alfalfa (OHN), Sorghum (SRG), Peas (PEA), Durum Wheat (DWT). All remaining classes were aggregated into an ‘others’ category (OTH). It is worth noting that DWS represents a sequential cultivation of soybean followed by winter wheat within the same year. Therefore, both SOY and DWS categories should be considered in practical applications specific to soybeans. The class frequencies faced high imbalance as shown in Fig. 1 (c), where the y-axis is on a logarithmic scale. The imbalanced class labels can be a big challenge to the classification models.

2.3.3. Data organization

The crop mapping dataset was collected over the study sites from 2019 to 2021 and created a subset for each year. After filtering clouds in satellite imagery and pixels with low CDL confidence, the annual total of

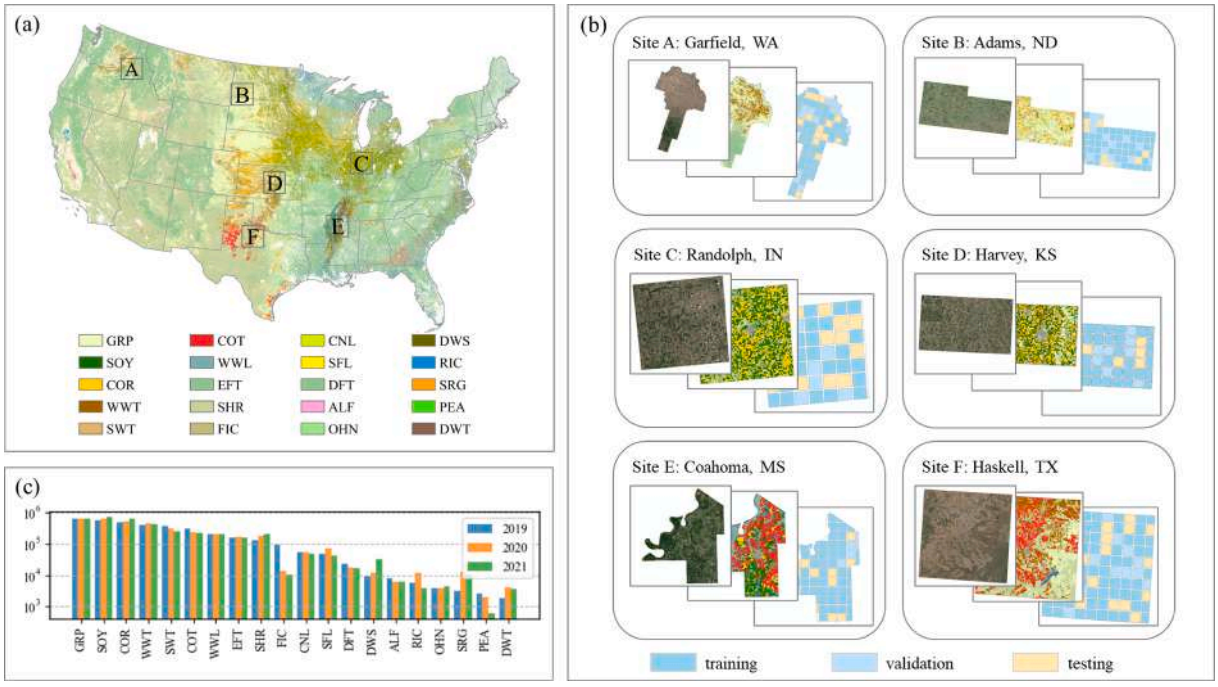


Fig. 1. Overview of the crop classification dataset. (a) Geographical positioning of the six study sites overlayed on the 2019 Cropland Data Layer (CDL) map. (b) Detailed visuals for each study site, shown from front to back: Sentinel-2 image samples, 2019 CDL map, and training-validation-testing block partitions. (c) Frequency distribution of classes from 2019 to 2021. Note: Refer to Section 2.3.2 for the abbreviations for crop types in (a) and (c).

SITS stands at approximately 3.5 million. The sequence lengths for the SITS vary between 19 and 101, indicating the diverse number of valid observations each SITS contains within our dataset. SITS was collected at a spatial resolution of 30 m from the GEE platform to match the spatial resolution of CDL. To train and evaluate models, the data was split into training, validation, and testing datasets for each year separately. According to Rußwurm & Körner (2020), the dataset partitions should be spatially separated in order to enforce independence among different datasets to avoid implicit overfitting (Rußwurm & Körner, 2017). Therefore, the six study sites were further divided into blocks of 4500 × 4500 m with a 500 m spacing between blocks, and all blocks were randomly assigned for training, validation, and testing in a ratio of 4:1:1 (Fig. 1 (b)). Finally, we got approximately 2.4 M training data and around 0.6 M validation and testing data per year.

3. Methodology

3.1. Stnet model for crop mapping

The proposed STNet mainly consists of two components, including a Transformer encoder block and an output head (Fig. 2). The objective of this network is to associate the input SITS data to the corresponding crop label. Building upon the original Transformer model, we introduced DoY positional encoding during the sequential encoding step, and incorporated NDVI-based temporal aggregation into the Seq2Vec module. These modifications enable the network to better emphasize and extract crop

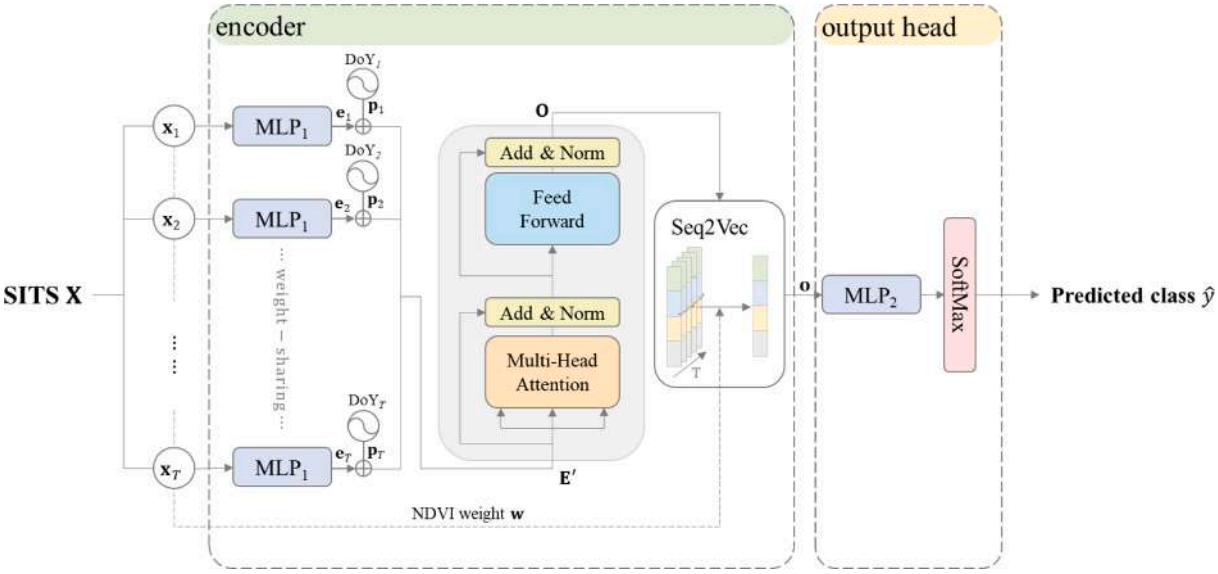


Fig. 2. The architecture of our STNet crop classification model (modified from Vaswani et al., 2017).

phenology features. The following section will provide a detailed description of this network.

Sequential encoding: Given a SITS $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T]$, where T is the total number of time steps of the SITS, the input signal at each time step $\mathbf{x}_t \in \mathbb{R}^C$ is encoded separately to obtain a high-dimensional feature sequence. Specifically, each $\mathbf{x}_t$ first passed through a weight-sharing Multi-Layer Perceptron, MLP_1, yielding a sequence of high-level feature vectors $\mathrm{E} = [\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_T] \in \mathbb{R}^{T\times D}$, where D is the feature dimension. As the self-attention mechanism in Transformer block is position-agnostic, a D-dimensional positional encoding (PE) vector is added to each feature vector $\mathbf{e}_t^{'} = \mathbf{e}_t + \mathbf{p}_t$, where $\mathbf{p}_t$ is encoded by the corresponding Day of Year (DoY) at each time step, which is defined as:

$$\mathbf{p}_t = \left[\sin\left(\mathrm{DoY}_t/\tau^{2i/D}\right), \cos\left(\mathrm{DoY}_t/\tau^{2i/D}\right)\right]_{i=1}^{D/2} \tag{1}$$

where $\tau = 1000$. The use of DoY embeds absolute temporal information in the sequence, which helps to account for inconsistent time steps as well as the temporally continuous crop phenology.

Self-attention in Transformer: The time series feature vectors $\mathbf{E}^{'} = [\mathbf{e}_1^{'}, \mathbf{e}_2^{'}, \cdots, \mathbf{e}_T^{'}]$ is fed into a Transformer block to learn the temporal dependencies among SITS observations. In this step, each feature $\mathbf{e}_t^{'}$ is processed by three individual fully connected layers to obtain query vector $\mathbf{q}_t$, the key vector $\mathbf{k}_t$, and the value vector $\mathbf{v}_t$. For a given query vector, its correlation with all key vectors is calculated to indicate the correspondence among different time steps of the crop life circle. Each dot product value is subsequently scaled and passed through a softmax layer to generate the attention score α_t for time step t:

$$\alpha_t = softmax\left(\frac{1}{\sqrt{d_k}}[\mathbf{q}_t \cdot \mathbf{k}_i]_{i=1}^{T}\right) \tag{2}$$

in which d_k is the shared dimension of the vectors $\mathbf{q}_t$, $\mathbf{k}_t$, and $\mathbf{v}_t$. The attention score α_t is then used to weight the value sequence $V = [\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_T]$ and retain the most relevant information to the current time step. In multi-head attention, each head operates on its own feature subset separately to enhance the expressiveness of the model. Then, through feed-forward and skip connections, the output sequence is obtained as $\mathbf{O} = [\mathbf{o}_1, \mathbf{o}_2, \cdots, \mathbf{o}_T] \in \mathbb{R}^{T\times D}$.

NDVI aggregation: The output sequence $\mathbf{O}$ is then reduced along the temporal dimension to obtain a representation vector $\mathbf{o} \in \mathbb{R}^D$ (Seq2Vec in Fig. 2). Typically, the embedding of satellite observation sequence is achieved by applying max-pooling or average-pooling to the resulting sequence of outputs (Rußwurm & Körner, 2020; Xu et al., 2021). In our STNet, to retain valuable temporal information, we compute the NDVI-weighted sum of output sequence $\mathbf{O}$ to get the embedded feature vector $\mathbf{o}$. This is inspired by knowledge-injected BERT (Ostendorff et al., 2019; Sun et al., 2019), in which commonsense knowledge or specialized domain knowledge is integrated into the model to enhance model performance. By incorporating crop-specific knowledge (i.e., NDVI in Seq2Vec), the model is guided to focus on the time span of vegetation growth, thereby facilitating the crop classification task. Specifically, the NDVI_t is computed at each time step using Eq. (3), where the superscript of x_t^c denotes the number of the band in Table 1. For Sentinel-2 imagery, x_t^8 indicates the NIR band, and x_t^4 indicates the Red band. Then, the time series NDVI is subjected to softmax normalization, resulting in a weight vector $\mathbf{w}$ of dimension T (Eq. (4)).

$$\mathrm{NDVI}_t = \frac{x_t^8 - x_t^4}{x_t^8 + x_t^4} \tag{3}$$

$$w = softmax\left([\mathrm{NDVI}_i]_{i=1}^{T}\right) \tag{4}$$

This weight vector is subsequently used to weight the temporal mean of $\mathbf{O}$ into an embedded representation vector $\mathbf{o}$. Then the representation vector is fed into the output head, comprising an MLP_2 and a softmax layer, which projects $\mathbf{o}$ onto probabilities for each crop class. The class with the highest probability is assigned as the predicted crop type $\hat{y}$ for the corresponding input SITS $\mathbf{X}$.

In summary, the attention mechanism and NDVI-weighted sum enable the STNet to give greater attention to observations during the vegetative period, so it can perform well on less-preprocessed data with cloud contamination. In addition, the model is able to process each time step independently, and the time series are aggregated as a feature representation by Seq2Vec. Therefore, SITS with different temporal lengths T can be processed without changing the model structure, which increases the flexibility of the model for large-scale SITS crop classification.

3.2. SITS-MoCo for pre-training

Training a deep Transformer model requires large amounts of labeled data. Nevertheless, acquiring a plentiful and accurate set of ground truth labels for large-scale crop mapping can be challenging and expensive. Therefore, we proposed a two-step framework to alleviate the need for labels for large-scale crop mapping. The overall procedure of our pre-training/fine-tuning scheme is shown in Fig. 3. During pre-training, the STNet model is trained on unlabeled RS imagery time series via contrastive loss to learn general-purpose spectral-temporal representations. During fine-tuning, the pretrained network is adapted to the classification task via fine-tuning with few ground labels. We introduce this framework and its detailed implementation in this section.

3.2.1. Background

Contrastive learning aims to train the encoder to generate representations by capturing shared features of similar samples (positive pairs) and discerning distinguishable features between dissimilar samples (negative pairs). Since there are no labels to determine the sample similarity, a common approach is instance discrimination, i.e., positive pairs are generated by applying random augmentations to the same sample, while negative pairs come from different samples. Formally, a given input $\mathbf{x}$ is augmented into two "views" as a query sample $\mathbf{x}_q$ and a key sample $\mathbf{x}_k$, which are then respectively encoded by encoders f_q and f_k to get query representation $\widetilde{\mathbf{q}} = f_q(\mathbf{x}_q)$ and key representation $\widetilde{\mathbf{k}} = f_k(\mathbf{x}_k)$. A query representation $\widetilde{\mathbf{q}}$ should be similar to the positive key $\widetilde{\mathbf{k}}_+$ originated from the same $\mathbf{x}$, while dissimilar to the negative keys $\left\{\widetilde{\mathbf{k}}_-^{(1)}, \widetilde{\mathbf{k}}_-^{(2)}, \widetilde{\mathbf{k}}_-^{(3)} \cdots\right\}$ originated from different samples. In this study, the similarity is measured by dot product to form the contrastive loss function InfoNCE (Oord et al., 2019):

$$L = -\log\frac{\exp\left(\widetilde{\mathbf{q}} \cdot \widetilde{\mathbf{k}}_+ / \gamma\right)}{\sum_{i=0}^{K}\exp\left(\widetilde{\mathbf{q}} \cdot \widetilde{\mathbf{k}}_i / \gamma\right)} \tag{5}$$

where γ is a temperature hyperparameter that scales the distribution of similarity levels. The sum is over one positive and K negative samples. We emphasize that the query and key representations in this context are distinct from the query and key vectors employed in the Transformer block. Thus, the notation $\widetilde{\mathbf{q}}$ and $\widetilde{\mathbf{k}}$ are used to denote the final representations obtained from the encoding network f_q and f_k.

Furthermore, contrastive learning benefits from a large key dictionary that covers a diverse set of samples (T. Chen et al., 2020; He et al., 2020). An efficient way is to dynamically maintain the dictionary as a queue whose size can be beyond the typical batch size (He et al., 2020). However, the enormous number of key samples results in back propagation of the key encoder intractable. He et al. (2020) proposed to momentum update the key encoder f_k to ensure the key representations' consistency:

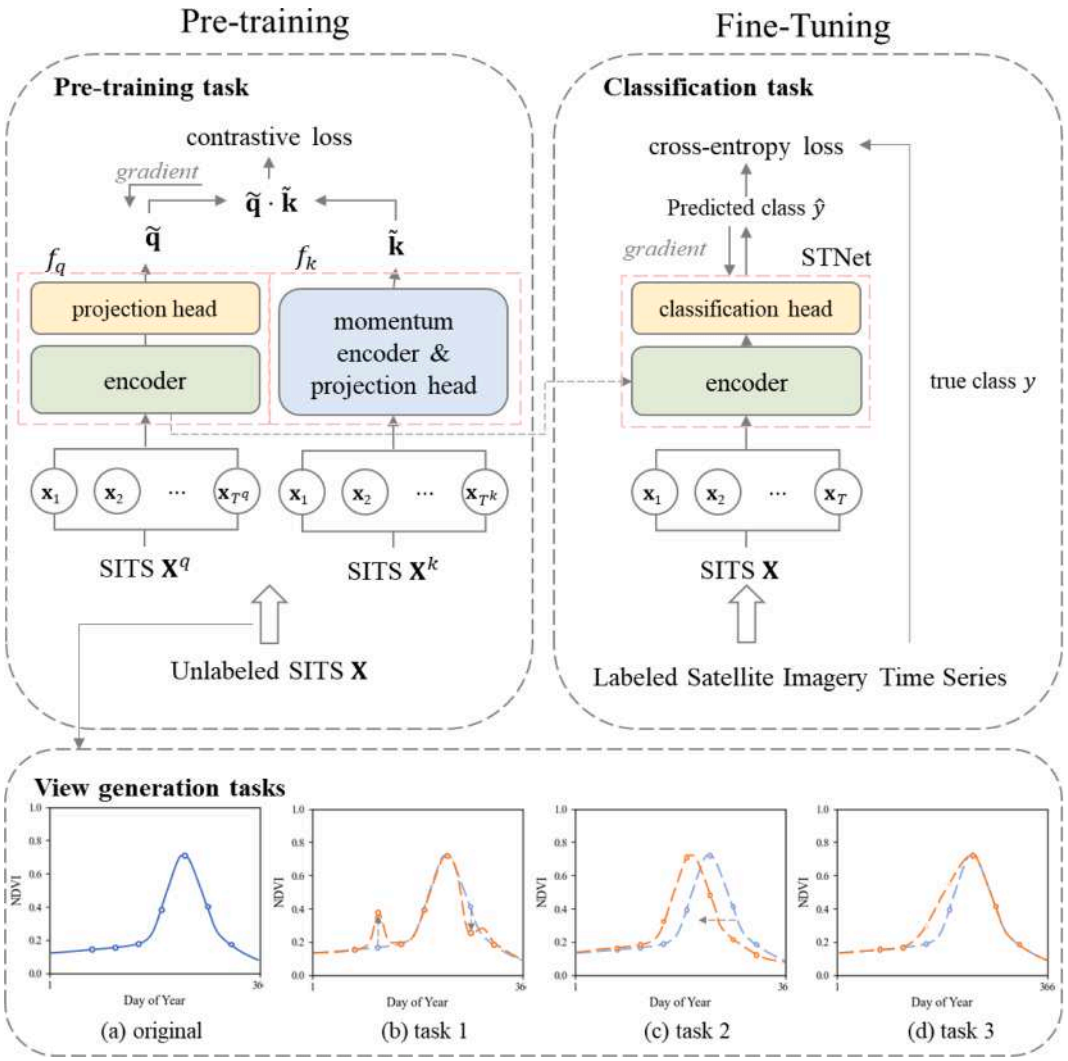


Fig. 3. Overall pre-training and fine-tuning procedures of the SITS-MoCo method for SITS crop classification. View generation tasks augmented the input SITS into two "views".

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q \tag{6}$$

where θ_k and θ_q is the parameters of f_k and f_q, respectively, and m is a hyperparameter (e.g., $m = 0.999$ as default) to control the update rate of θ_k.

3.2.2. View generation

According to the characteristics of SITS, we design three data augmentation tasks to generate two random "views" (i.e., $\mathbf{X}_q$ and $\mathbf{X}_k$) of a reference SITS (i.e., $\mathbf{X}$) to form a positive pair. By distinguishing positive pairs from negative pairs from other SITS, the pre-trained model learns features robust to these augmentations. The view generations tasks shown in Fig. 3 accordingly tackle the challenges in large-scale crop mapping, including noise caused by clouds and shadows, phenological shift, and missing observations (Fig. 4). The three tasks are elaborated below.

Task 1: The first challenge is the spectral noise caused by clouds and shadows that introduces abnormal fluctuations in the time series data. As shown in Fig. 4 (a), owing to noises caused by the clouds and shadows, positive peaks and near-zero valleys were observed in the time-series reflectance, which makes it challenging to interpret the data. Furthermore, the presence of clouds and shadows varies largely across regions, which exacerbates the spatial heterogeneity in large-scale crop mapping. In order to improve the model's robustness to such noise data, we designed the first data augmentation task as a denoising task. In this case, gaussian noises with zero mean and standard deviation 0.5 were randomly added to certain percentage of the input time series data (Fig. 3 (b)). During pre-training, the encoder was trained to minimize the representation discrepancy on positive pairs, i.e., one SITS with different noises, while maximizing the distances between negative representations, i.e., key views from different SITS. By doing so, the encoder is expected to extract representations that are both discriminative to different temporal profiles and invariant to noise caused by clouds and shadows. We referred to this pre-training task as "Data Denoising".

Task 2: The second challenge is the phenological shift between the same crop from different regions. As illustrated in Fig. 4(b), the phenology characteristic of the corn shifts between sites B and F owing to different climate conditions and management practices. This is especially evident in large-scale crop mapping covering large geographically diverse regions. It would lower the model's generalizability of the model to new regions. Thus, it is necessary to make the encoder extract representations that are invariant to the phenological shift. Therefore, we designed the second pre-training task as addressing the temporal shift in time-series data. Specifically, to introduce random shifts along the temporal dimension of the time series data, we define the shift operation as: $\text{shift}([\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T], \delta) = [\mathbf{x}_{(T-\delta+1)\%T}, \cdots, \mathbf{x}_T, \mathbf{x}_1, \cdots, \mathbf{x}_{(T-\delta)\%T}]$, where δ denotes the number of positions by which the observations are shifted. The direction of the shift, either forward or

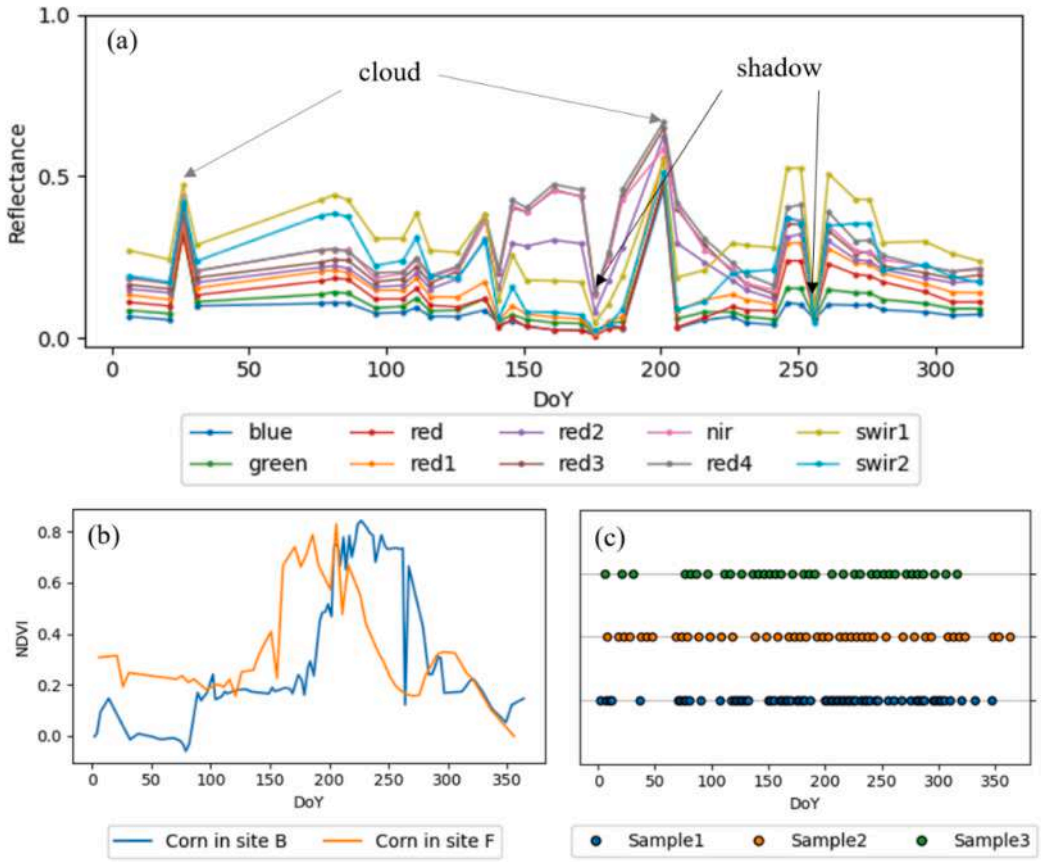


Fig. 4. Large-scale crop mapping challenges. (a) noises caused by clouds and shadows in a SITS sample, (b) the phenological shift of the same crop type from different sites, and (c) missing observations in the time series data.

backward, is determined by the sign of δ. While the order of the temporal data is adjusted, the DoY sequence remains unchanged, effectively simulating the phenological shift (Fig. 3 (c)). During pre-training, the encoder was trained to recognize SITS views with different temporal shifts, allowing the encoder to ignore temporal shifts and focus on crop growth stages. We referred to this pre-training task as "Temporal Shift Alignment".

Task 3: The third challenge is the missing observations in SITS. SITSs face varying degrees of missing temporal observations (Fig. 4(c)) due to variations in satellite revisit periods and weather conditions. The mismatched time steps of crop samples from different regions aggravate the discrepancies in the time series data, posing a challenge to accurate crop mapping at a large-scale. To enable the model to extract high-level features from the SITS dataset, even in cases where observations at certain time steps are missing, the third pre-training task was designed as data restoration. Specifically, the SITS was augmented by randomly dropping a certain proportion of observations in each time series (Fig. 3 (d)). During pre-training, the encoder was trained to maximize the similarity of representations extracted from positive pairs with different temporal absences by minimizing the contrastive loss L, making the encoder insensitive to observation missing in the time series data. We referred to this pre-training task as "Observation Completion".

Among them, the task of data denoising is spectral-related. The tasks of temporal shift alignment and observation completion are temporal-related. Contrastive learning on these tasks helps the model to learn representations that ignore the large-scale sample variation caused by spectral noise, phenological shift, and temporal data missing, respectively. Hence, the encoded representations will contain both spectral-noise invariant and temporal-noise invariant features, which can be efficiently transferred to the downstream task of crop classification. With a large unlabeled dataset for the pre-training tasks, the model can learn informative representations with high generalizability. Therefore, the pre-training model can be adapted to the downstream tasks with a small labeled training set.

3.2.3. *Stnet pre-training and fine-tuning*

In the pre-training phase, as shown in Fig. 3, the query view $\mathbf{X}^q$ and key views $\mathbf{X}^k$ are encoded by neural networks f_q and f_k, respectively, obtaining representations $\widetilde{\mathbf{q}}$ and $\widetilde{\mathbf{k}} \in \mathbb{R}^{D'}$. In our case of SITS-MoCo pre-training with STNet, f_q and f_k are modified STNet, where the Seq2Vec module in the encoder are replaced by an average pooling layer and the classification head is substituted with a projection head. The reason for employing average pooling instead of NDVI aggregation is that, during pre-training, the objective is to leverage the inherent structure within the SITS and obtain general-purpose representations that can be adaptable to various downstream tasks. The projection head is a common component used in contrastive learning. It maps the encoded representations to a D'-dimensional space where contrastive loss is applied. The output of projection head is l2-normalized, constraining the representations to lie on the unit hypersphere. f_k and f_q share the same network structure, while f_k is updated with momentum update strategy (Eq. (6)). The contrastive learning objective is optimized based on Eq. (5), where positive pairs are distinguished from negative pairs, resulting in discriminative representations that are invariant to the aforementioned augmentations.

After pre-training is completed, the pre-trained model is then fine-tuned using the labeled crop classification dataset for the crop mapping task. Specifically, the parameters of query encoder f_q are utilized to initialize the encoder in STNet. As the Seq2Vec is an unlearnable module, the average pooling is directly replaced by NDVI-weighted sum to enable the model to focus on the crop growth period and enhance its

adaptation to the crop mapping task. Furthermore, the task-specific classification head is plugged in to output the probability distribution for each category of the target crop types. During the fine-tuning stage, all the parameters are fine-tuned end-to-end by minimizing the crop classification loss, defined as the Cross-Entropy loss between predicted probability distribution and true crop class y. Finally, after convergence, the fine-tuning process is completed, and the fine-tuned model is evaluated on the testing set to assess its performance on crop classification.

4. Experiments and results

In this section, we evaluated the proposed methodology for SITS crop classification in three testing years 2019 – 2021. The methodology was evaluated in two folds. First, we compared the proposed STNet with five widely employed methods, including RF (Breiman, 2001), TempCNN (Pelletier et al., 2019), LSTM (Hochreiter & Schmidhuber, 1997), LTAE (Garnot & Landrieu, 2020), and Transformer (Vaswani et al., 2017). Each model was trained from scratch and evaluated on the same testing set. Besides the STNet trained from scratch, we also evaluated its performance when it was pre-trained via SITS-MoCo (i.e., STNet + SITS-MoCo). Second, to further evaluate the proposed SITS-MoCo pre-training strategy, we compared it with the widely used SITS-BERT pre-training strategy (Yuan & Lin, 2021). These two pre-training methods were applied to the same model structure and evaluated on the same testing set.

4.1. Experimental setup

Model configuration. For our STNet, the MLP_1 consists of three successive fully connected layers with 32, 64, and 128 hidden units, respectively. We grid-searched the best configuration for the Transformer block and set the feature dimension to 128 and the number of heads to 16. The MLP_2 is designed to have two fully connected layers with 64 and 32 neurons and an output layer. Besides, RF is the only traditional ML method used for comparison. It was designed to have 500 estimators and the maximum depth of each estimator was set as 25. For TempCNN, we followed the original paper to stack three convolution layers with a dimensionality of 128 and a kernel size of 7, after which a dense layer was included as the output layer. LSTM starts with four bidirectional LSTM layers with a dimension of 128, and the hidden states are concatenated to process by a dense output layer. The LTAE is considered the state-of-the-art model for SITS crop classification. As it is originally designed for field parcel classification, we modified the pixel-set encoder to a pixel encoder as our MLP_1 and kept the LTAE architecture. The DoY PE in LTAE is encoded by the number of days elapsed since the beginning of the sequence, following the description in the original paper. The Transformer model has the same architecture as our STNet, but the DoY PE is replaced by the sequence order PE as in the original paper, and the Seq2Vec is an average pooling layer. Under these configurations, the number of parameters for each comparison model is listed in Table 2. For all comparison models, to obtain fixed-length time series for training models in batches, we followed (Rußwurm et al., 2020) to randomly sample each time series to a fixed length of 70 observations while maintaining the temporal order.

Implementation details. During the pre-training phase, all augmentations are applied with a probability of 15% to generate query and key samples, and the output dimension of the projection head was set as 128. As SITS-MoCo is not designed exclusively for STNet, for all the above-mentioned models, we replaced the output layer with the projection head and utilized the output representations for contrastive learning. All models are pre-trained on the unlabeled dataset with around 1 M time series where 90% was used for training and 10% for validation. The size of the dictionary queue K was set as 65,536, comprising 65,536 negative samples at a time. The temperature τ in the contrastive loss was set to the default of 0.7, and the momentum coefficient m for updating the key encoder was set to 0.999. The models were pre-trained for up to 100 epochs, with early stopping implemented if validation performance did not improve for 10 consecutive epochs. The batch size was set to 512, and the Adam optimizer was employed with a learning rate of 1e-3 and a weight decay of 1e-4. In the fine-tuning stage, we randomly selected 3,000 training examples from the training set and 300 validation examples from the validation set to better emphasize differences in the results. The Adam optimizer was employed with the same parameters as in the pre-training phase for all DL models. The models were fine-tuned for 100 epochs with a batch size of 512, and early stopping was implemented if validation performance did not improve for 10 consecutive epochs. To quantify the model performance, we employed three metrics including overall accuracy (OA), kappa coefficient (Kappa), and weighted f1 score (WF1). The results were reported as an average of five repeated trials with different random seeds for each year separately. The RF was developed with the scikit-learn library, and all the DL models were implemented using PyTorch.

Table 2
The number of parameters for comparison methods.

Method	Number of parameters
RF	N/A
TempCNN	4.83 M
LSTM	1.33 M
LTAE	0.16 M
Transformer	0.12 M
STNet	0.12 M

4.2. Model comparison

Evaluation results for each testing year of five comparison models, the proposed STNet trained from scratch (i.e., STNet), and the STNet utilizing the SITS-MoCo pre-trained model (i.e., STNet + SITS-MoCo) are reported in Table 3. The best-performing results are highlighted in bold for each year.

Among the comparison methods, RF was observed to have the worst performance, obtaining an OA of 81.01% on average. This can be attributed to the challenging nature of the dataset used in this study. As anticipated, our goal was to develop models that require minimal data preprocessing. Therefore, the dataset utilized in this study remains largely unprocessed, exhibiting spectral noise, inconsistent temporal observations, and heavily imbalanced class distribution. These factors

Table 3
Evaluation results of our methods and other comparison models of three testing years.

Year	Method	OA (%)	Kappa	WF1
2019	RF	79.38	0.7705	0.7681
	TempCNN	88.68	0.8734	0.8739
	LSTM	84.59	0.8278	0.8293
	LTAE	93.36	0.9255	0.9261
	Transformer	92.34	0.914	0.9166
	STNet	94.35	0.9364	0.9381
	STNet + SITS-MoCo	**95.62**	**0.9507**	**0.9531**
2020	RF	82.21	0.8000	0.8007
	TempCNN	87.22	0.8561	0.8582
	LSTM	85.79	0.8396	0.8461
	LTAE	93.44	0.9258	0.9259
	Transformer	94.26	0.9349	0.9373
	STNet	95.19	0.9454	0.9465
	STNet + SITS-MoCo	**95.95**	**0.9541**	**0.9556**
2021	RF	81.12	0.7859	0.7875
	TempCNN	85.63	0.8370	0.8379
	LSTM	86.10	0.8418	0.8427
	LTAE	93.11	0.9210	0.9238
	Transformer	92.56	0.9146	0.9193
	STNet	94.17	0.9331	0.9323
	STNet + SITS-MoCo	**95.17**	**0.9445**	**0.9480**

make it challenging for RF to accurately interpret and classify the data. Among the DL methods, the widely used TempCNN and LSTM demonstrated relatively lower performance, with an average OA of 87.18% and 85.49%, respectively. This is because these models can solely extract temporal information pertaining to the order of observations in SITS. However, our non-interpolated SITS contains observations from various time steps at the same location, making it a challenging task to manage. Additionally, the convolution layers in TempCNN and the recurrent layers in LSTM are characterized by a large parameter volume (Table 2), which can pose challenges when training with insufficient labeled data. In comparison, the Transformer-based methods, including LTAE, Transformer, and the proposed STNet, yielded better performance and achieved an average OA of 93.30%, 93.05%, and 94.57%, respectively. Our STNet performed the best even without pre-training via SITS-MoCo. Compared to the original Transformer, the introduction of DoY positional encoding in STNet brought a large improvement. Also, in comparison with sequence to vector strategy in LTAE, the NDVI-weighted sum in STNet enabled the model to capture the phenology of the crops and thus improve the prediction accuracy. Moreover, when pre-training STNet via SITS-MoCo, the OA was further improved by 1.27%, 0.76%, and 1.00% across three testing years, demonstrating the effectiveness of the proposed pre-training strategy.

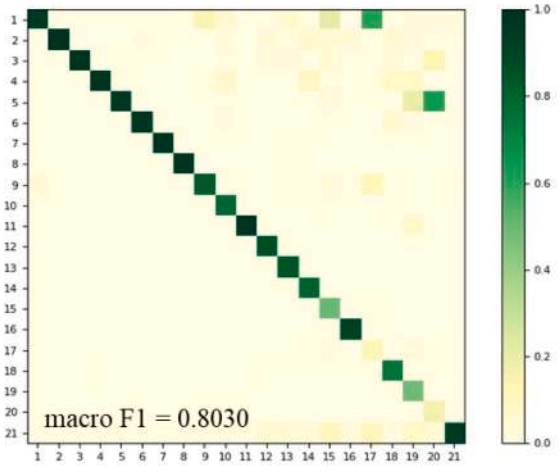
To further illustrate the performances of our STNet + SITS-MoCo in each crop type, we presented the normalized confusion matrix and F1 scores for each crop type averaged over years 2019–2021 in Fig. 5. The crop classes are arranged in descending order based on their proportion, with the exception of the others class. All crop classes within the test set were classified accurately within a macro F1 scores of 0.8030. Moreover, it was observed that misclassification concentrated in the upper-right corner in all models, indicating that all models tended to classify samples as dominated classes due to the imbalanced class distribution. For example, hay samples were tended to classify as grass/pasture for their similar texture. And durum wheat samples were commonly mistaken as spring wheat because they have similar phenology.

4.3. Pre-training strategy comparison

To further evaluate the proposed pre-training strategy SITS-MoCo, we conducted a performance comparison of five models under three different conditions: random initialization (i.e., Random init.), pre-training via SITS-BERT, and pre-training via SITS-MoCo. As shown in Table 4, the best performer is highlighted in bold for each case and each year. Note that the pre-training task in SITS-BERT requires the base model to produce temporally sequential outputs, while LTAE aggregates on temporal dimension within the LTAE module, it is not feasible to implement SITS-BERT on LTAE. Therefore, we only reported the evaluation results of LTAE under random initialization and SITS-MoCo.

Consistently across all cases, we observed that SITS-MoCo improved the model performance with a larger margin in comparison with both SITS-BERT and the random initialization. Overall, SITS-MoCo achieved an average OA of 94.10%, outperforming SITS-BERT (92.30%) and random initialization (90.72%). Specifically, when applying SITS-MoCo for pre-training TempCNN and LSTM, we observed an increase in OA by 7.28% and 5.69%, respectively, compared to random initialization. While for SITS-BERT pre-training, OA improved by 6.21% and 5.16%, respectively. Notably, in 2020, the SITS-MoCo pre-trained TempCNN achieved an OA of 95.08%, Kappa of 0.9441, and WF1 of 0.9459, which was comparable to the Transformer-based LTAE. When pre-training the Transformer-based models (i.e., LTAE, Transformer, and STNet) via SITS-MoCo, the OA was further improved by an average of 1.14% when compared to the random initialization. In contrast, the SITS-BERT pre-training showed a detrimental effect on the performance of both the Transformer and STNet, resulting in a decrease of 0.98% in OA compared to the random initialization. The subpar performance of SITS-BERT can be traced back to the discrepancy in feature distribution, such as magnitude, between the self-supervised pre-training task and the classification task performed downstream (He et al., 2020). In the Transformer-based models, where the magnitude of PE is fixed, this discrepancy in feature distribution potentially disrupts the downstream task's training process, leading to the observed decrease in accuracy. It is also noteworthy that our STNet without pre-training showed comparable results to other comparison models under SITS-MoCo pre-training, which further demonstrated the superiority of our STNet for the task of SITS crop mapping. In summary, the results demonstrate the effectiveness and flexibility of the proposed SITS-MoCo. It achieved excellent performance improvements across various models.

As the models were pre-trained solely on the unlabeled dataset collected in 2019, we then discussed the impact of crop mapping accuracy across different years. On the whole, SITS-MoCo pretraining achieved an average OA gain over random initialization of 3.29%, 3.50%, and 3.36% in 2019, 2020, 2021, respectively. There is a slight variation in performance across different years, but the improvement over random initialization remains consistent. These results imply that the advantages gained from SITS-MoCo pre-training persist over time. The pre-trained models obtained through SITS-MoCo showed inter-year generalizability and can be effectively utilized in future years, thereby



(a) Confusion matrix

ID	Crop name	F1	ID	Crop name	F1
1	Grass/Pasture	0.9337	12	Sunflower	0.9141
2	Soybeans	0.9865	13	Deciduous Forest	0.9135
3	Corn	0.9877	14	Dbl Crop WinWht / Soybeans	0.8975
4	Winter Wheat	0.9794			
5	Spring Wheat	0.9624	15	Alfalfa	0.6385
6	Cotton	0.9803	16	Rice	0.9211
7	Woody Wetlands	0.9920	17	Other Hay/Non-Alfalfa	0.2113
8	Evergreen Forest	0.9832	18	Sorghum	0.5691
9	Shrubland	0.8242	19	Peas	0.3010
10	Fallow/Idle Cropland	0.6910	20	Durum Wheat	0.2757
11	Canola	0.9876	21	Others	0.8836

(b) F1 scores

Fig. 5. (a) confusion matrices of our STNet + SITS-MoCo. The x-axis is the reference class, and the y-axis is the predicted class. Each column denotes the probability of a certain crop type being classified into each crop type. (b) crop ID, crop type names, and corresponding F1 scores.

Table 4
Evaluation results of Random initialization, SITS-BERT, and the proposed SITS-MoCo pre-training approaches on different base models.

	Base model	2019			2020			2021		
		OA (%)	Kappa	WF1	OA (%)	Kappa	WF1	OA (%)	Kappa	WF1
Random init.	TempCNN	88.68	0.8734	0.8739	87.22	0.8561	0.8582	85.63	0.8370	0.8379
SITS-BERT		92.87	0.9200	0.9213	93.45	0.9257	0.9284	92.29	0.9117	0.9142
SITS-MoCo		**94.05**	**0.9332**	**0.9356**	**95.08**	**0.9441**	**0.9459**	**94.25**	**0.9340**	**0.9380**
Random init.	LSTM	84.59	0.8278	0.8293	85.79	0.8396	0.8461	86.10	0.8418	0.8427
SITS-BERT		90.45	0.8931	0.8931	91.28	0.9014	0.9022	90.24	0.8885	0.8912
SITS-MoCo		**91.58**	**0.9054**	**0.9088**	**92.48**	**0.9148**	**0.9170**	**91.05**	**0.8976**	**0.9027**
Random init.	LTAE	93.36	0.9255	0.9261	93.44	0.9258	0.9259	93.11	0.9210	0.9238
SITS-MoCo		**94.45**	**0.9377**	**0.9386**	**95.16**	**0.9451**	**0.9454**	**93.89**	**0.9299**	**0.9311**
Random init.	Transformer	92.34	0.9140	0.9166	94.26	0.9349	0.9373	92.56	0.9146	0.9193
SITS-BERT		91.29	0.9023	0.9040	92.05	0.9101	0.9104	91.57	0.9035	0.9074
SITS-MoCo		**94.08**	**0.9334**	**0.9360**	**94.72**	**0.9402**	**0.9410**	**94.03**	**0.9314**	**0.9351**
Random init.	STNet	94.35	0.9364	0.9381	95.19	0.9454	0.9465	94.17	0.9331	0.9323
SITS-BERT		94.62	0.9395	0.9430	94.04	0.9325	0.9335	93.4	0.9243	0.9272
SITS-MoCo		**95.62**	**0.9507**	**0.9531**	**95.95**	**0.9541**	**0.9556**	**95.17**	**0.9445**	**0.9480**

reducing the costs associated with pre-training. Additionally, for STNet, SITS-MoCo reduced the interannual OA standard deviation from 0.4445 to 0.3197 compared to random initialization. This indicated that pre-training not only enhances accuracy regardless of the year but also reduces variation, leading to more stable crop mapping results.

To further verify the effectiveness of our method, the classification maps of evaluation blocks with the highest confidence of CDL in each study area were shown in Fig. 6 to visually compare the classification performance of STNet via random initialization, SITS-BERT pre-training, and SITS-MoCo pre-training. The OA indicates the consistency between the predicted map and the CDL. Specifically, in sites A and E where the crop types are relatively homogeneous, the three methods exhibited comparable and satisfactory performance. In contrast, in more challenging sites C and D the crop maps generated by STNet + SITS-MoCo also showed the highest level of consistency within the field boundaries and were generally in alignment with the CDL reference maps. Overall, the experimental results demonstrated the superiority of the proposed SITS-MoCo over other state-of-the-art pre-training strategies and effectively improved the overall accuracy of large-scale crop mapping with limited labels.

5. Discussion

5.1. Ablation study of STNet

In Section 4.2, we compared STNet with the original Transformer. The results in Table 3 demonstrated the effectiveness of incorporating the DoY PE and the NDVI aggregation in STNet, yet the individual contributions of these two strategies still need to be explored. We analyzed this problem with an ablation experiment. We considered four network configurations: TRSF-AvgPool-Position (Transformer) served as the basic network, employing avg-pooling and sequence order PE; TRSF-NDVI-Position utilized NDVI aggregation while retaining the sequence order PE; TRSF-AvgPool-DoY utilized avg-pooling along with DoY PE; TRSF-NDVI-DoY represented the complete STNet architecture with both strategies employed. The comparison results of these networks are presented in Table 5. The experiment results revealed that both the DoY PE and NDVI aggregation contributed to improvements over the baseline performance. Specifically, TRSF-AvgPool-DoY achieved a 1.09% average OA improvement compared to the TRSF baseline, and the NDVI aggregation in TRSF-NDVI-Position resulted in a slight 0.28% improvement in average OA. Notably, the proposed STNet outperformed all other network configurations in three testing years, achieving the highest performance. It exhibited a 1.52% improvement in the average OA compared to the Transformer baseline. These results validated the efficacy of our proposed network architecture, demonstrating that both the DoY PE and NDVI aggregation strategies are valuable for crop classification.

5.2. Self-attention scores

The self-attention mechanism is a key component in STNet. As described in Section 3.1, every self-attention head computes an attention matrix $A = [\alpha_1, \alpha_2, \cdots \alpha_T]$ according to Eq. (2). Each α_t represents the temporal correspondence between the observation at time step t and all observations in the time series, i.e., which observations affect the current observation at t. To provide a visual illustration of the self-attention mechanism, Fig. 7 presented the average attention matrices for two main crops in the U.S. (i.e., corn and soybean). The values in the attention matrix quantify the weight of an input temporal observation on the output temporal feature. Larger attention values suggest a larger weight on the observation. Notably, each self-attention head focuses on distinct time extents that learn different temporal features to enhance model discriminability. For example, the first attention head seems to attend to the middle-to-late phase of the time series, while the 13th head captures information from the early and late stages of the time series. Additionally, we noted that the attention heads are adaptive to different crop types. For instance, the first and last heads for corn cover a longer period than those of soybean. This can be attributed to the different distribution of these two crops in our dataset. Specifically, soybeans are mainly concentrated in Sites C, D, and E, which are geographically close to each other and tend to have similar growing periods for soybeans. As a result, the model focuses on a relatively narrow time extent to observe the growing season of soybean in these sites. Conversely, corn is scattered across distant regions that have corn planted at different times. Therefore, the model focuses on a wider temporal extent to identify corn in different regions.

5.3. Efficacy of pre-training on limited numbers of training samples

One advantage of using pre-training strategies is to reduce the need for labeled data samples for downstream tasks (Mañas et al., 2021). To better understand the efficacy of the pre-training strategy given limited numbers of training samples, we fine-tuned different models that were pre-trained via random initialization, SITS-BERT, and our SITS-MoCo method on labeled datasets with different sizes, which ranged from 300 to 30,000.

As shown in Fig. 8, models pre-trained via SITS-MoCo consistently outperformed random initialized models and models pre-trained via SITS-BERT. Particularly when the number of labeled samples was limited, the benefits of SITS-MoCo pre-training were more evident. For example, when there were only 300 training samples, randomly initialized LSTM attained an OA of 49.68%. Pre-training with SITS-BERT improved it to 66.75%, while SITS-MoCo achieved an even greater enhancement, reaching an OA of 76.52%. As the number of labeled data increased, the performance gap between SITS-MoCo pre-training and other methods narrowed down. This phenomenon demonstrated that

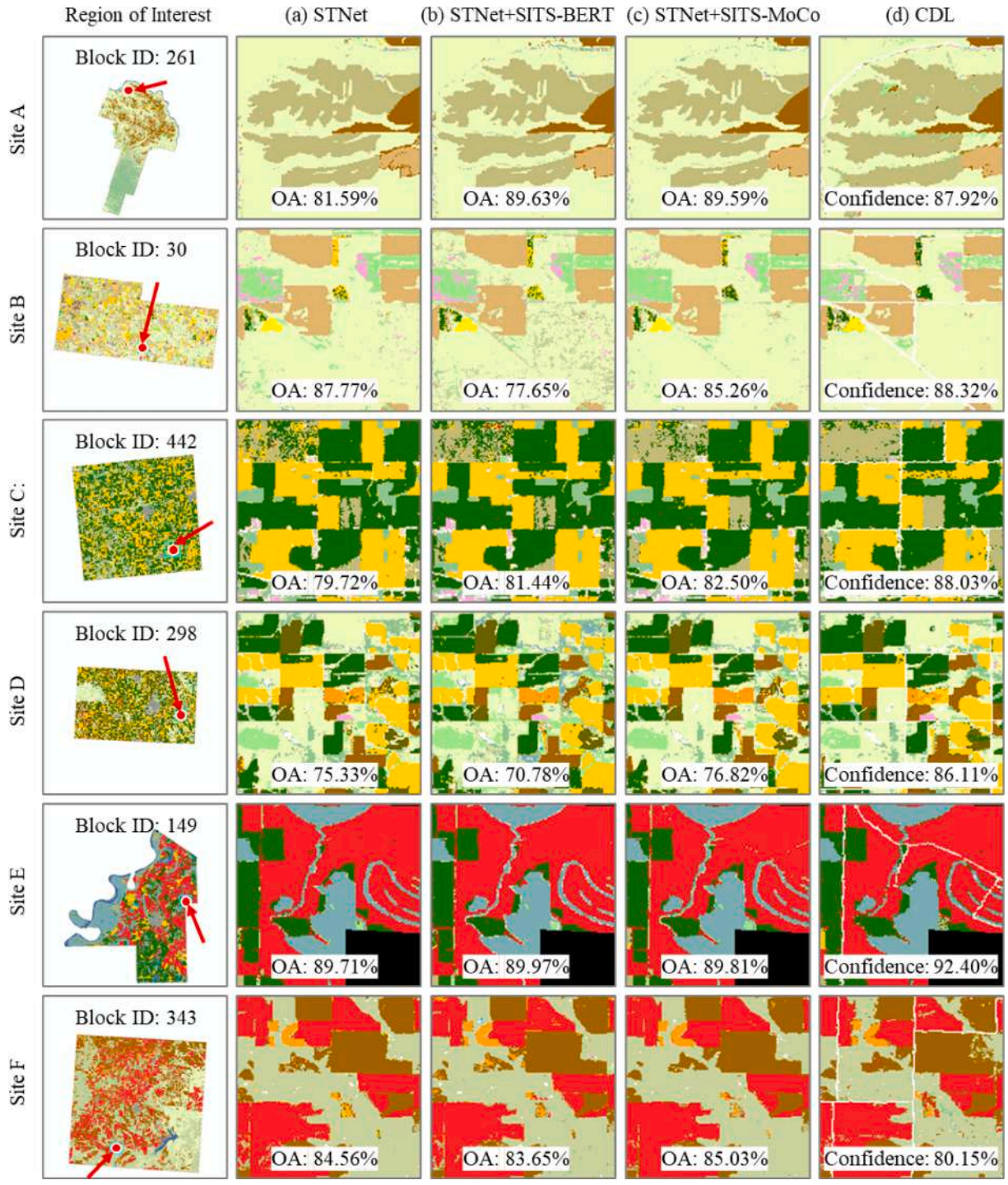


Fig. 6. Examples of crop classification maps in each study site in 2019 by STNet via (a) random initialization, (b) SITS-BERT pre-training, and (c) SITS-MoCo pre-training and the corresponding (d) CDL reference maps. Results are obtained through the majority vote of five repeated trials.

SITS-MoCo pre-training can effectively mitigate the overfitting problem that commonly arises in DL models when facing the case of a lack of labeled samples.

Moreover, the accuracy gains varied in each model, with SITS-MoCo pre-training exhibiting particularly significant improvements when applied to TempCNN and LSTM models. Both TempCNN and LSTM models performed poorly with random initialization when less than 3000 labeled training samples were used, which is consistent with the observation in Section 4.2 regarding the overfitting potential of DL models with the large volume of parameters (Table 2). However, this problem could be largely reduced via SITS-MoCo. In particular, TempCNN with SITS-MoCo pre-training achieved an average OA of 81.1% when 300 labeled data were used, representing a significant improvement of about 20% compared to random initialization and around 10% compared to SITS-BERT. Furthermore, SITS-MoCo demonstrates greater label efficiency. For example, with only 500 labeled samples, SITS-MoCo pre-trained TempCNN achieved an average OA of 86.17%, comparable to the OA of 87.34% achieved by random initialization with 3000 labels. For attention-based models, the performance achieved through random initialization was relatively satisfactory, while our SITS-MoCo pre-training method could further improve upon this baseline performance. When utilizing 1000 labeled samples, all

Table 5
Comparison results of STNet with different strategies in three testing years.

	Method	OA (%)	Kappa	WF1
2019	TRSF-AvgPool-Position	92.34	0.9140	0.9166
	TRSF-NDVI-Position	92.78	0.9190	0.9220
	TRSF-AvgPool-DoY	93.86	0.9310	0.9326
	TRSF-NDVI-DoY	**94.35**	**0.9364**	**0.9381**
2020	TRSF-AvgPool-Position	94.26	0.9349	0.9373
	TRSF-NDVI-Position	94.41	0.9366	0.9388
	TRSF-AvgPool-DoY	94.99	0.9432	0.9447
	TRSF-NDVI-DoY	**95.19**	**0.9454**	**0.9465**
2021	TRSF-AvgPool-Position	92.56	0.9146	0.9193
	TRSF-NDVI-Position	92.81	0.9175	0.9214
	TRSF-AvgPool-DoY	93.59	0.9264	0.9272
	TRSF-NDVI-DoY	**94.17**	**0.9331**	**0.9323**

attention-based models with SITS-MoCo pre-training achieved an OA greater than 90%. In contrast, SITS-BERT pre-training did not guarantee improvement, particularly when applied to the Transformer model, where the SITS-BERT pre-trained model consistently underperformed the model with random initialization. The reason is that the objective of the pre-training task in SITS-BERT is to recognize and restore the noisy observations while the attention mechanism enables the model to ignore noisy observations. These conflicting objectives during the pre-training and fine-tuning stages of SITS-BERT might prevent it from improving the accuracy of attention-based models. Overall, SITS-MoCo has been demonstrated as a highly effective pre-training strategy that can be applied to various models, making it a promising option for addressing the issue of scarce labels in crop mapping.

5.4. In-season experiments

Early- or in-season crop mapping, that requires implementing crop mapping during the crop growth season, is crucial in agricultural practices. It tests the crop classification model's ability to extract from limited length SITS that is beneficial to distinguishing different crop types. The positional encoding module within the STNet enables applying end-of-season pre-training models to in-season downstream tasks. To assess the effectiveness of SITS-MoCo pre-training in different input temporal windows, we conducted in-season experiments in three testing years. Specifically, in the fine-tuning crop mapping task, we restricted the Sentinel-2 data SITS to span from April 1st to the month ending dates from May through September. For each in-season setting, the STNet model was re-initialized with the same whole-year pre-trained model and then fine-tuned for the specific temporal windows. This process implied that models are pre-trained to capture general and comprehensive feature representations from entire-year SITS, and subsequently fine-tuned to adapt for specific downstream tasks. The results of STNet using SITS-MoCo pre-training and random initialization are presented in Table 6.

SITS-MoCo pre-training models consistently outperformed those with random initialization by approximately 1% of OA, indicating that a whole-year pre-trained model can provide valuable insights to enhance accuracy in in-season crop mapping. From the perspective of the input length for fine-tuning, full-time series input for both initializations achieved superior overall accuracies of 95.58% and 94.57%, respectively, compared to any in-season input. This might be because our crop classification dataset incorporates winter crops, such as WWT, as well as double crops within a year, such as DWS. Excluding data from non-summer periods may result in insufficient identification information for these crop types, leading to misclassification. However, when tailoring to specific local crop phenology, the input time window can be adjusted accordingly, where the whole-year pre-trained model can still boost the accuracy of downstream crop mapping tasks. Furthermore, it can be seen that the SITS-MoCo pre-training model exceeded 90% OA by the end of July and achieved a near-optimal OA of 95.08% by the end of August. This suggested that the months of July and August are important for distinguishing the major crop types. As many crops began to be harvested in September, the usefulness of additional temporal windows diminished. Therefore, the results exhibited a stable trend with slight increases. Overall, pre-training with the full-time series model can facilitate crop classification across both the complete and varied in-season temporal windows. It indicates that there is no need for separate pre-training with different temporal lengths, which can be computationally demanding due to the extensive volume of unlabeled datasets.

6. Conclusion

In this study, we introduced a novel combination of the Transformer-based STNet model and the SITS-MoCo self-supervised pre-training strategy for cross-year crop mapping. The STNet model utilized the self-attention module to extract crop-related features from SITS data. The SITS-MoCo pre-training strategy leveraged the extensive availability of unlabeled SITS data for learning informative and transferable spectral-temporal representations. The SITS-MoCo pre-trained model was subsequently adapted to the downstream crop classification tasks. By pre-training to learn robust SITS representations, on the one hand, it can reduce the demand for crop labels, on the other hand, it can improve

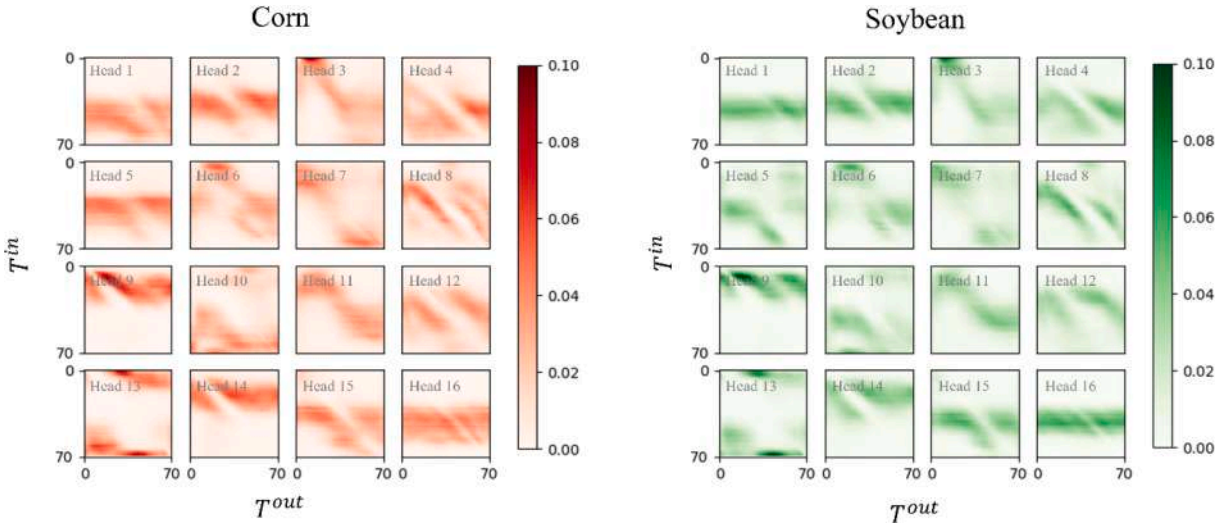


Fig. 7. Average self-attention maps of the STNet with 16 heads of class Corn (left) and Soybean (right).

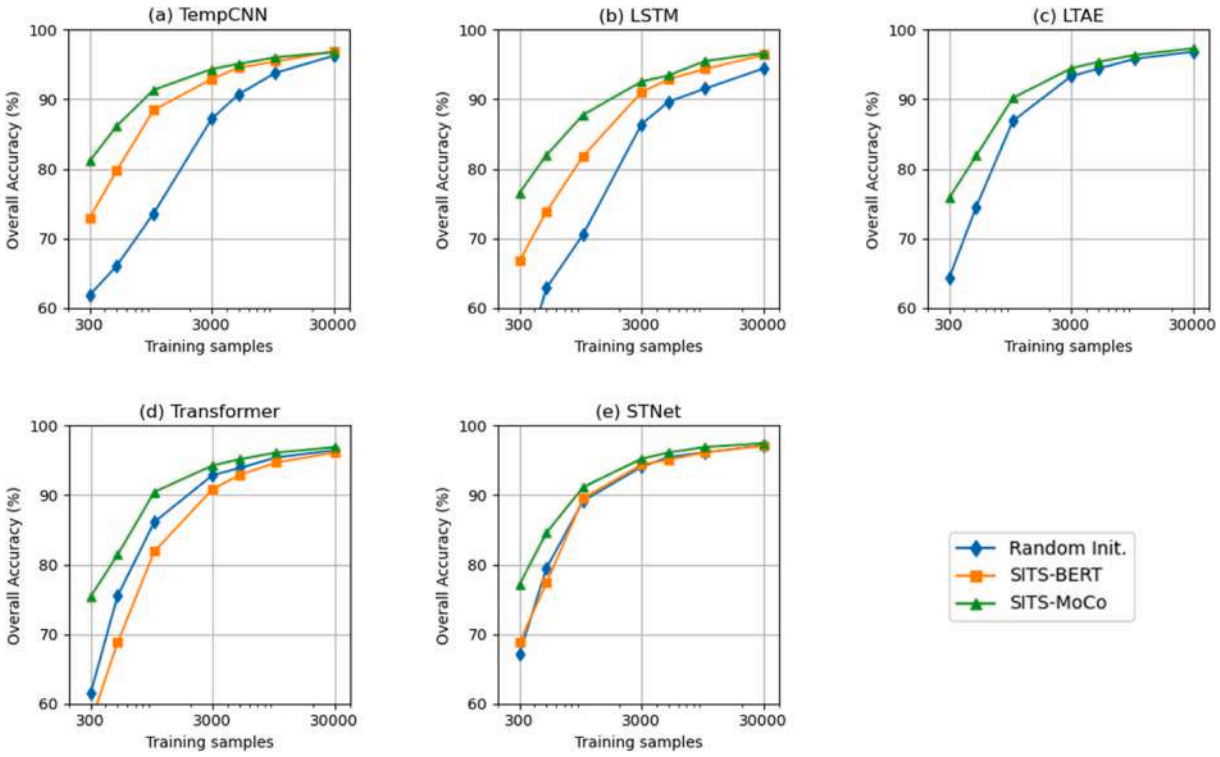


Fig. 8. Accuracy of crop classification for models (a) TempCNN, (b) LSTM, (c) LTAE, (d) Transformer, (e) STNet, given labeled datasets with different sizes ranging from 300 to 30,000. Results are averaged over three testing years and five repeated trials for each year.

Table 6
Classification performance comparison of the STNet with SITS-MoCo pre-training and random initialization using complete SITS and five in-season input sequences. Results are averaged over three testing years.

		Temporal window after April 1, through the end of					Complete time series(whole year)
		May.	Jun.	Jul.	Aug.	Sep.	
SITS-MoCo	OA (%)	74.95	84.02	92.77	95.08	95.17	95.58
	Kappa	0.7197	0.8204	0.9181	0.9441	0.9451	0.9498
	WF1	0.7239	0.8202	0.9188	0.9464	0.9471	0.9522
Random Init.	OA (%)	73.81	83.17	91.46	94.20	94.42	94.57
	Kappa	0.7072	0.8109	0.9034	0.9341	0.9366	0.9383
	WF1	0.7139	0.8126	0.9047	0.9368	0.9395	0.9390

the performance of STNet model in label-scarce scenarios.

Experiments in six study sites in diverse regions of the U.S. showed that 1) the proposed STNet was able to extract robust features in dealing with noise in the raw SITS, outperforming the original Transformer by approximately 1.5% in OA and 2) the SITS-MoCo pre-training strategy further improved the model performance, surpassing the random initialization by approximately 1% in OA. The model analysis showed the capability of STNet to capture relevant temporal features in SITS. Besides, the comparison study on labeled datasets with different sizes ranging from 300 to 30,000 demonstrated that SITS-MoCo could largely improve the model performance in label-scarce scenarios.

Further explorations could focus on investigating the generalizability of the pre-trained models across diverse geographical regions, aiming for comprehensive adaptability in varied agricultural contexts. Another area of interest is the integration of auxiliary data sources, such as weather data or soil information. Since crop growth is closely related to the local climate and soil conditions, integrating these datasets could provide valuable contextual information to enhance the crop mapping process and improve the accuracy of crop classification. Advancing in these directions would not only improve the accuracy and robustness of crop mapping models but also broaden their applicability in practical scenarios.

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.

Acknowledgments

This work was supported by the United States Department of Agriculture (USDA) National Institute of Food and Agriculture (NIFA), Agriculture and Food Research Initiative Project under Grant 1028199 and USDA NIFA Hatch Project under Grant 7005141. We would also like to thank the NVIDIA Academic Hardware Grant Program for providing the GPU resource used in this work.

References

Bazzi, H., Ienco, D., Baghdadi, N., Zribi, M., Demarez, V., 2020. Distilling Before Refine: Spatio-Temporal Transfer Learning for Mapping Irrigated Areas Using Sentinel-1 Time Series. IEEE Geoscience and Remote Sensing Letters 17 (11), 1909–1913. https://doi.org/10.1109/LGRS.2019.2960625.

Breiman, L., 2001. Random Forests. Machine Learning 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

CEC. (1997). *Ecological Regions of North America: Toward a Common Perspective*. Commission for Environmental Cooperation.

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). *A Simple Framework for Contrastive Learning of Visual Representations* (arXiv:2002.05709). arXiv. https://doi.org/10.48550/arXiv.2002.05709.

Chen, X., Fan, H., Girshick, R., & He, K. (2020). *Improved Baselines with Momentum Contrastive Learning* (arXiv:2003.04297). arXiv. https://doi.org/10.48550/arXiv.2003.04297.

Deines, J.M., Guan, K., Lopez, B., Zhou, Q., White, C.S., Wang, S., Lobell, D.B., 2023. Recent cover crop adoption is associated with small maize and soybean yield losses in the United States. Global Change Biology 29 (3), 794–807. https://doi.org/10.1111/gcb.16489.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., Meygret, A., Spoto, F., Sy, O., Marchese, F., Bargellini, P., 2012. Sentinel-2: ESA's Optical High-Resolution Mission for GMES Operational Services. Remote Sensing of Environment 120, 25–36. https://doi.org/10.1016/j.rse.2011.11.026.

Garnot, V. S. F., & Landrieu, L. (2020). *Lightweight Temporal Self-Attention for Classifying Satellite Image Time Series* (arXiv:2007.00586). arXiv. http://arxiv.org/abs/2007.00586.

Garnot, V. S. F., Landrieu, L., Giordano, S., & Chehata, N. (2020). *Satellite Image Time Series Classification With Pixel-Set Encoders and Temporal Self-Attention*. 12325–12334. https://openaccess.thecvf.com/content_CVPR_2020/html/Garnot_Satellite_Image_Time_Series_Classification_With_Pixel-Set_Encoders_and_Temporal_CVPR_2020_paper.html.

Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: Planetary-scale geospatial analysis for everyone. Remote Sensing of Environment 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.

Hadsell, R., Chopra, S., & LeCun, Y. (2006). Dimensionality Reduction by Learning an Invariant Mapping. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2 (CVPR'06)*, *2*, 1735–1742. https://doi.org/10.1109/CVPR.2006.100.

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). *Momentum Contrast for Unsupervised Visual Representation Learning* (arXiv:1911.05722). arXiv. https://doi.org/10.48550/arXiv.1911.05722.

Hochreiter, S., Schmidhuber, J., 1997. Long Short-Term Memory. Neural Computation 9 (8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735.

Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D., & Ermon, S. (2018). *Tile2Vec: Unsupervised representation learning for spatially distributed data* (arXiv:1805.02855). arXiv. http://arxiv.org/abs/1805.02855.

Jung, H., Oh, Y., Jeong, S., Lee, C., Jeon, T., 2022. Contrastive Self-Supervised Learning With Smoothed Representation for Remote Sensing. IEEE Geoscience and Remote Sensing Letters 19, 1–5. https://doi.org/10.1109/LGRS.2021.3069799.

Kluger, D.M., Owen, A.B., Lobell, D.B., 2022. Combining randomized field experiments with observational satellite data to assess the benefits of crop rotations on yields. Environmental Research Letters 17 (4), 044066. https://doi.org/10.1088/1748-9326/ac6083.

Ma, Y., Zhang, Z., Kang, Y., Özdoğan, M., 2021. Corn yield prediction and uncertainty analysis based on remotely sensed variables using a Bayesian neural network approach. Remote Sensing of Environment 259, 112408. https://doi.org/10.1016/j.rse.2021.112408.

Ma, Y., Yang, Z., Zhang, Z., 2023. Multisource Maximum Predictor Discrepancy for Unsupervised Domain Adaptation on Corn Yield Prediction. IEEE Transactions on Geoscience and Remote Sensing 61, 1–15. https://doi.org/10.1109/TGRS.2023.3247343.

Mañas, O., Lacoste, A., Giró-i-Nieto, X., Vazquez, D., & Rodríguez, P. (2021). *Seasonal Contrast: Unsupervised Pre-Training From Uncurated Remote Sensing Data*. 9414–9423. https://openaccess.thecvf.com/content/ICCV2021/html/Manas_Seasonal_Contrast_Unsupervised_Pre-Training_From_Uncurated_Remote_Sensing_Data_ICCV_2021_paper.html.

Neumann, M., Pinto, A. S., Zhai, X., & Houlsby, N. (2019). *In-domain representation learning for remote sensing* (arXiv:1911.06721). arXiv. http://arxiv.org/abs/1911.06721.

Nyborg, J., Pelletier, C., Lefèvre, S., Assent, I., 2022. TimeMatch: Unsupervised Cross-Region Adaptation by Temporal Shift Estimation. ISPRS Journal of Photogrammetry and Remote Sensing 188, 301–313. https://doi.org/10.1016/j.isprsjprs.2022.04.018.

Ostendorff, M., Bourgonje, P., Berger, M., Moreno-Schneider, J., Rehm, G., & Gipp, B. (2019). *Enriching BERT with Knowledge Graph Embeddings for Document Classification* (arXiv:1909.08402). arXiv. https://doi.org/10.48550/arXiv.1909.08402.

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Transactions on Knowledge and Data Engineering 22 (10), 1345–1359. https://doi.org/10.1109/TKDE.2009.191.

Pelletier, C., Webb, G.I., Petitjean, F., 2019. Temporal Convolutional Neural Network for the Classification of Satellite Image Time Series. Remote Sensing 11 (5), 5. https://doi.org/10.3390/rs11050523.

Rußwurm, M., Pelletier, C., Zollner, M., Lefèvre, S., & Körner, M. (2020). *BreizhCrops: A Time Series Dataset for Crop Type Mapping* (arXiv:1905.11893). arXiv. http://arxiv.org/abs/1905.11893.

Rußwurm, M., Körner, M., 2017. Temporal Vegetation Modelling Using Long Short-Term Memory Networks for Crop Identification from Medium-Resolution Multi-spectral Satellite Images. IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2017, 1496–1504. https://doi.org/10.1109/CVPRW.2017.193.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical Satellite Time Series Classification. ISPRS Journal of Photogrammetry and Remote Sensing 169, 421–435. https://doi.org/10.1016/j.isprsjprs.2020.06.006.

Sun, Y., Wang, S., Li, Y., Feng, S., Chen, X., Zhang, H., Tian, X., Zhu, D., Tian, H., & Wu, H. (2019). *ERNIE: Enhanced Representation through Knowledge Integration* (arXiv:1904.09223). arXiv. https://doi.org/10.48550/arXiv.1904.09223.

Oord, A. van den, Li, Y., & Vinyals, O. (2019). *Representation Learning with Contrastive Predictive Coding* (arXiv:1807.03748). arXiv. http://arxiv.org/abs/1807.03748.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wang, S., Azzari, G., Lobell, D.B., 2019. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. Remote Sensing of Environment 222, 303–317. https://doi.org/10.1016/j.rse.2018.12.026.

Wang, Y., Zhang, Z., Feng, L., Ma, Y., Du, Q., 2021a. A new attention-based CNN approach for crop mapping using time series Sentinel-2 images. Computers and Electronics in Agriculture 184, 106090. https://doi.org/10.1016/j.compag.2021.106090.

Wang, Z., Zhang, H., He, W., Zhang, L., 2021b. Phenology Alignment Network: A Novel Framework for Cross-Regional Time Series Crop Classification. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW) 2021, 2934–2943. https://doi.org/10.1109/CVPRW53098.2021.00329.

Xu, J., Zhu, Y., Zhong, R., Lin, Z., Xu, J., Jiang, H., Huang, J., Li, H., Lin, T., 2020. DeepCropMapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. Remote Sensing of Environment 247, 111946. https://doi.org/10.1016/j.rse.2020.111946.

Xu, J., Yang, J., Xiong, X., Li, H., Huang, J., Ting, K.C., Ying, Y., Lin, T., 2021. Towards interpreting multi-temporal deep learning models in crop mapping. Remote Sensing of Environment 264, 112599. https://doi.org/10.1016/j.rse.2021.112599.

Yu, T., Wu, W., Gong, C., Li, X., 2021. Residual Multi-Attention Classification Network for A Forest Dominated Tropical Landscape Using High-Resolution Remote Sensing Imagery. ISPRS International Journal of Geo-Information 10 (1), Article 1. https://doi.org/10.3390/ijgi10010022.

Yuan, Y., Lin, L., 2021. Self-Supervised Pretraining of Transformers for Satellite Image Time Series Classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 14, 474–487. https://doi.org/10.1109/JSTARS.2020.3036602.

Zhong, L., Gong, P., Biging, G.S., 2014. Efficient corn and soybean mapping with temporal extendability: A multi-year experiment using Landsat imagery. Remote Sensing of Environment 140, 1–13. https://doi.org/10.1016/j.rse.2013.08.023.

Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. Remote Sensing of Environment 221, 430–443. https://doi.org/10.1016/j.rse.2018.11.032.