ISPRS Journal of Photogrammetry and Remote Sensing 206 (2023) 293–310

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Spatio-temporal multi-level attention crop mapping method using time-series SAR imagery

Zhu Han [a,b,c], Ce Zhang [d,e,f,*], Lianru Gao [g,**], Zhiqiang Zeng [h], Bing Zhang [a,c], Peter M. Atkinson [f,i]

[a] Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[b] International Research Center of Big Data for Sustainable Development Goals, Beijing 100094, China
[c] College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China
[d] School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK
[e] UK Centre of Ecology & Hydrology, Library Avenue, Barlrigg, Lancaster LA1 4AP, UK
[f] Lancaster Environment Centre, Lancaster University, Lancaster LA1 4YQ, UK
[g] Key Laboratory of Computational Optical Imaging Technology, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
[h] School of Electronic and Information Engineering, Beihang University, Beijing 100191, China
[i] Geography and Environmental Sciences, University of Southampton, Highfield, Southampton SO17 1BJ, UK

ARTICLE INFO

Keywords:
Crop mapping
Time-series imagery
Spatio-temporal self-attention
Multi-scale cross-attention
Deep learning

ABSTRACT

Accurate crop mapping is of great significance for crop yield forecasting, agricultural productivity development and agricultural management. Thanks to its all-time and all-weather capability, integrating multi-temporal synthetic aperture radar (SAR) for crop mapping has become essential and challenging task in remote sensing. In recent years, deep learning (DL) has demonstrated excellent crop mapping accuracy to interpret crop dynamics. However, existing DL-based methods tend to be incapable of capturing spatial and temporal features at different scales simultaneously, and this often leads to severe mis-classification due to the complex and heterogeneous distribution of crops and diverse phenological patterns. In this paper, we propose a novel spatio-temporal multi-level attention method, named as STMA, for crop mapping using time-series SAR imagery in an end-to-end fashion to increase the capability of crop phenology retrieval. Specifically, the multi-level attention mechanism is designed to aggregate multi-scale spatio-temporal representations on crops via cascaded spatio-temporal self-attention (STSA) and multi-scale cross-attention (MCA) modalities. To ensure a fine extraction of multi-granularity features, a learnable spatial attention position encoding is proposed to adaptively generate the position priors to facilitate multi-level attention learning. Experimental results on Brandenburg Sentinel-1 dataset, public PASTIS-R dataset and South Africa dataset demonstrated that STMA can achieve state-of-the-art performance in crop mapping tasks, with the accuracy of 96.54% in the Brandenburg Sentinel-1 dataset, 86.77% in the PASTIS-R dataset and 83.37% in the South Africa dataset, validating its effectiveness and superiority. Further comparison of spatio-temporal generalization capability reflected its excellent performance in spatio-temporal modeling on different crops and scenarios. This research provides a viable and intelligent spatio-temporal framework for large-area crop mapping using time-series SAR imagery in complex agricultural systems. The Brandenburg Sentinel-1 dataset and the STMA code will be publicly available at https://github.com/hanzhu97702/ISPRS_STMA.

1. Introduction

With the global demand for food increasing constantly, crop yield estimation serves a critical role in regulating the balance between food supply and demand (Thenkabail et al., 2012; Alexandratos and Bruinsma, 2012). At the same time, accurate and up-to-date crop mapping is essential for understanding agricultural land use and monitoring crop growth, enabling decision-makers to develop effective crop management practices for maintaining food security (Bargiel, 2017; Cai et al., 2018), such as water supply irrigation and targeted crop insurance (Tilman et al., 2011; Karthikeyan et al., 2020). In recent decades, satellite sensors have been used in both crop yield estimation and crop mapping due to their synoptic view, regular revisit time frequency and, in some cases, weather-independent acquisition capabilities. Based on different imaging techniques, current multi-temporal crop mapping

[*] Corresponding author at: School of Geographical Sciences, University of Bristol, Bristol BS8 1SS, UK.
[**] Corresponding author.
E-mail addresses: ce.zhang@bristol.ac.uk (C. Zhang), gaolr@aircas.ac.cn (L. Gao).

https://doi.org/10.1016/j.isprsjprs.2023.11.016
Received 23 June 2023; Received in revised form 17 November 2023; Accepted 20 November 2023
Available online 24 November 2023

methods are developed, primarily, using optical and synthetic aperture radar (SAR) data (Adrian et al., 2021). A large number of studies focus on fine or medium spatial resolution optical images for crop mapping, e.g., Sentinel-2a/b (Drusch et al., 2012), Landsat (Dong et al., 2016) and MODIS (Wardlow and Egbert, 2008). However, optical remote sensing (RS) images are limited by cloud (Julien and Sobrino, 2010; Zhao et al., 2020), haze contamination (Peng et al., 2020) and other spectral mixing perturbations (Han et al., 2020, 2022b) which cannot guarantee the acquisition of clear multi-temporal images in real applications. Synthetic aperture radar (SAR) has become a potential data source for crop mapping with its all-time and all-weather imaging capability (Amazirh et al., 2018; Zeng et al., 2021). By interacting with the crop canopy and the underlying soil, the backscattering information derived from SAR data can stably reflect the morphological structure and orientation distribution of crops at different phenological stages (McNairn et al., 2009). Thus time-series SAR can be potentially invaluable for dynamic crop mapping in complex agricultural systems (Xu et al., 2018).

With the rapid development of artificial intelligence, deep learning (DL) has been extensively explored in the field of crop mapping (Zhu et al., 2017; Zeng et al., 2022). Unlike traditional machine learning approaches, the DL-based methods can efficiently learn spatial and temporal relationships of crops in the pixel or parcel level, with vast potential for simulating complex crop phenology processes in practical applications (Wang et al., 2021; Han et al., 2022a). To synergistically utilize time-series SAR imagery for crop mapping, several recent studies have tended to develop hybrid DL network architectures based on convolution neural networks (CNNs), recurrent neural network (RNNs), and self-attention networks, owing to their excellent spatio-temporal modeling abilities (Rußwurm and Körner, 2020; Xu et al., 2021). This end-to-end feature learning mechanism of DL reduces manual intervention and further facilitates the development of high-precision and automated crop mapping. Depending on the learning mechanism, existing crop mapping approaches can be both unsupervised and supervised. The unsupervised crop mapping methods mainly focus on deep unsupervised representation and clustering strategies (Iounousse et al., 2015; Franceschi et al., 2019; Kalinicheva et al., 2020; Guo et al., 2022b) to match crop statistics with the help of clustering features and similarity measures between temporal sequences, further mitigating the lack of crop labels and human supervision when no ground data exists. Considering domain shifts of crops in different regions, unsupervised domain adaptation approaches are introduced to transfer prior knowledge from the source domain to the target domain, and further enhance the model generalization (Kwak and Park, 2022; Wang et al., 2023). However, the unsupervised methods are often vulnerable to outliers and high dimensionality, and difficult for distinguishing crops with larger difference in real applications (Wang et al., 2019; Magistri et al., 2023). The supervised crop mapping methods can effectively model or simulate ideal crop phenologies to achieve promising performance in crop mapping owing to given limited training samples (Karim et al., 2019; Garnot and Landrieu, 2021a), thus they have been extensively used to generate crop mapping relative to unsupervised approaches.

Although existing studies demonstrated the effectiveness of DL-based approaches and replaced handcrafted feature engineering, there are three problems that need to be solved in the current DL-based crop mapping frameworks. (1) Limited spatio-temporal receptive field by considering multi-scale information of crops in a single type of the convolutional network. Due to the existence of spatial variability affected by environmental conditions (e.g., climate and topography), as well as complex phenological characteristics associated with increased crop diversification and intensification (Van Bussel et al., 2011; Diao, 2019), existing approaches focus on extracting multi-scale features with the short-range joint dependency by the CNN-based operation, and lack of capturing more discriminative long-range temporal multi-scale information corresponding to crop phenological development. The self-attention-based approaches also focus on the uniform information granularity within the attention layer, and overlook the multi-scale nature of crops. As a result, the extracted SAR spatio-temporal features are easily misled by similar crop types or speckle noise, further producing unreliable and incomprehensible mapping results. The effective and multi-scale capture in both spatial and temporal dimension is vital to the performance improvement of crop mapping. (2) Irregularity temporal priors owing to acquisition obstructed by sensors. For better modeling numerical time-series RS data, state-of-the-art DL-based crop mapping methods (i.e., self-attention networks) need to input all possible calendar time to provide temporal positions for DL model training and identifying various crop types (Garnot et al., 2020; Garnot and Landrieu, 2021a). However, it is difficult to obtain complete and regular time-series data in practical applications, which affects model generalization performance (Liu et al., 2020). Some cumbersome and manual interpolation prepossessing tools are utilized to make up for this defect (Inglada et al., 2015). Therefore, exploring learnable and prior-avoiding position encoding techniques are essential for current DL-based crop mapping methods. (3) Incomplete spatial and temporal feature fusion strategies are employed for time-series SAR data. Existing DL-based methods consider to extract spatio-temporal features through the cascade form or element-wise sum weakens the contribution of vegetation (Rußwurm and Körner, 2020; Yang et al., 2021), thereby reducing spatio-temporal correlations between different polarization modes. Exploring the efficient and adaptive fusion of both spatial and temporal attention mechanism will be beneficial to suppress unimportant information and highlight the more informative spatio-temporal features.

According to the above discussion, we propose a novel spatio-temporal multi-level attention method, STMA, for crop mapping using time-series SAR imagery. STMA is a encoder–decoder CNN-transformer-based framework that efficiently leverages local feature representations of CNNs and global long-distance relationships of transformers from the multi-scale perspective. The encoder part of STMA aims at extracting multi-scale and robust spatio-temporal features by the multi-level attention mechanism, including spatio-temporal self-attention (STSA) module and multi-scale cross-attention (MCA) module. The STSA module builds a parallel structure to extract long-distance spatial and temporal characteristics respectively, and designs an adaptive feature fuse module with learnable network parameter to promote spatio-temporal feature fusion and interoperability, while the MCA module captures and aggregates long-distance correlations between multi-scale spatio-temporal features to enhance the spatio-temporal receptive field. To ensure the local information extracted by the CNN-based ResNet, a learnable spatial attention position encoding is designed to adaptively generate the position priors of time-series features, which further facilitates multi-level attention learning. Finally, a lightweight decoder is employed to reconstruct the extracted multi-level spatio-temporal features hierarchically to produce the final crop mapping result. The major contributions of this research include:

(1) We present a multi-level attention crop mapping method to efficiently aggregate multi-scale spatio-temporal representations via cascaded self-attention and cross-attention modalities, which achieves superior performance for crop mapping.
(2) The STSA module achieves multi-granularity spatio-temporal feature extraction, and the MCA module based on cosine similarity fully integrates both spatial and temporal multi-scale information in order to enlarge the spatio-temporal receptive field for long-distance dynamics, which helps to capture complex phenological characteristics of crops.
(3) We design a learnable spatial attention position encoding to enhance the multi-level attention mechanism, such that the spatial structure extracted from the CNN-based ResNet encoder is well preserved and it is applicable to different datasets and input lengths.

(4) We conduct a thorough evaluation to compare our proposed STMA method with current state-of-the-art approaches on the Brandenburg Sentinel-1 dataset, the public PASTIS-R dataset and South Africa dataset. The experimental results shows that combining multi-scale spatio-temporal information grants significant improvement in terms of spatio-temporal generalization.

The remainder of the paper is organized as follows. Section 2 introduces the related work of traditional and DL-based crop mapping approaches. The proposed STMA method is described in Section 3. The experimental results and the discussion are presented in Sections 4 and 5, respectively. The conclusion is drawn in Section 6 finally.

2. Related work

2.1. Traditional crop mapping using time-series SAR

Multiple traditional crop mapping approaches were developed using time-series SAR imagery to differentiate various crop types by characterizing the crop phenophase and distribution over the past few decades. These can be divided broadly into three categories: threshold-based, statistics-based and machine learning methods. Threshold-based methods consider primarily the optimal seasonal threshold (Satalino et al., 2013; Oyoshi et al., 2016) or variation of radar vegetation index (Trudel et al., 2012; Periasamy, 2018; Mandal et al., 2020) to separate different crop types, which can be effective for crops with distinct temporal characteristics, such as rice and soybean (Veloso et al., 2017). Statistics-based approaches focus on analyzing global statistics by building pre-defined mathematical functions or models to describe phenological dynamics or time-varying characteristics for improved crop discrimination, such as Kalman filters (Vicente-Guijalba et al., 2013), particle filters (De Bernardis et al., 2014), Hidden Markov Models (Leite et al., 2011) or Conditional Random Fields (Kenduiywoa et al., 2015; Kenduiywo et al., 2017). Although the statistical modeling of time-series SAR data can enhance mapping accuracy, the utilization of phenological knowledge is challenging in practice due to the uncertainty involved in the evolution of crop phenology caused by climate variation (Gao et al., 2021). Machine learning approaches, such as the support vector machine (SVM) (Sonobe et al., 2015), decision tree (DT) (Waske and Braun, 2009), and random forest (RF) (Sonobe et al., 2014), commonly stack time-series SAR imagery as multiple features, and adopt data mining techniques to distinguish different crop types at the pixel level. This type of method mainly relies on handcrafted features subject to expert knowledge and consider limited spatio-temporal relationships of time-series SAR imagery. In addition to the backscattering mechanism analysis, the polarization signatures has been proven to maximize the difference between different crop types in certain orientation angle and further provide discriminative features for crop mapping (Tan et al., 2011; Zhang et al., 2014; Srikanth et al., 2016; Huang et al., 2017).

2.2. DL-based approaches in crop mapping

In the field of crop mapping, the most popular DL-based models are CNNs (Krizhevsky et al., 2012) and RNNs (Zaremba et al., 2014), which explore the spatial and temporal feature representations from multi-temporal satellite sensor imagery. Specifically, CNNs aim at extracting spatial contextual representations via convolutional filters to realize end-to-end classification in large-area RS scenes (Gu et al., 2018), and some semantic segmentation models, such as U-network (UNet) (Ronneberger et al., 2015) and fully convolutional network (FCN) (Long et al., 2015), can perform pixel-based crop classification using a series of convolutional and pooling operations. To handle multi-temporal images, much effort has been made to achieve crop mapping based on the temporal domain of CNNs. For example, Zhong et al. (2019) adopted a one-dimensional CNN (1D-CNN) architecture to capture temporal variation in an Enhanced Vegetation Index (EVI) time-series, which demonstrated the feasibility of CNN-based architectures for temporal analysis. Adrian et al. (2021) proposed a 3D UNet crop mapping method to learn local spatial and temporal features simultaneously by applying 3D convolution kernels throughout the crop growing season, further increasing overall crop mapping accuracy compared with 2D CNN models. Li et al. (2021) considered the object-level scale sequence CNN framework to classify different crops based on SAR imagery and further ensured more precise boundaries between crop parcels. Guo et al. (2022a) designed a convolutional-autoencoder neural network (C-AENN) to achieve efficient utilization of optimal multi-temporal feature combination of time-series SAR imagery and exploited the potential of the CNN-based hybrid architecture in the task of crop mapping. As another set of DL models, RNNs are specialized in sequential data analysis and have been used widely to process multi-temporal RS data for crop mapping. Long short-term memory (LSTM) and bidirectional LSTM (Bi-LSTM) were adopted to map rice crops from time-series SAR and demonstrated their superiority in capturing temporal correlation and extracting multi-temporal features compared to traditional machine learning approaches (Crisóstomo de Castro Filho et al., 2020). To further improve the temporal modeling ability, the attention mechanism was introduced into LSTM to yield state-of-the-art classification performance (Rußwurm and Körner, 2018; Xu et al., 2020). Moreover, the hybrid architecture of CNN and RNN variants, e.g. ConvGRU and ConvLSTM, enhanced the spatial generalization for dynamic crop mapping and achieved increased accuracy (Shi et al., 2015; Chang et al., 2022).

Recently, Transformers have been developed to capture long-range dependencies and interactions with the support of self-attention mechanisms, and allow parallel computation to reduce local decreases in accuracy due to long-term context dependencies (Vaswani et al., 2017; Dosovitskiy et al., 2020). Following the adoption of the self-attention mechanism in the Transformer, Rußwurm and Körner (2020) achieved pixel-level crop recognition using optical time series, much improved compared with both RNN-based and CNN-based models. Furthermore, the combination of pixel-set encoder and lightweight temporal self-attention (PSE+LTAE) can acquire rich spatial extent and temporal patterns of crop parcels for classifying time-series optical imagery (Garnot et al., 2020; Garnot and Landrieu, 2020). Based on the celebrated self-attention architecture, Weilandt et al. (2023) proposed a multi-modal crop mapping framework by utilizing dense time series of optical and radar data to achieve multi-source feature fusion and further enhance crop mapping accuracy. In a similar vein, the U-TAE architecture introduced the spatial UNet-based architecture into LTAE to achieve pixel-level crop mapping from the semantic segmentation perspective (Garnot and Landrieu, 2021a). To overcome the dilemma of temporal shifts between different regions, thermal positional encoding (TPE) was proposed to learn invariant temporal features and improve the generalization of self-attention models (Nyborg et al., 2022). Nevertheless, owing to the limited spatio-temporal receptive field and the uniform information granularity within the attention layer, obtaining satisfactory crop mapping results through these existing crop mapping methods are difficult when facing large diversity of crops and complex time-series SAR scenarios.

3. Methodology

The overall framework of the proposed STMA crop mapping scheme is shown in Fig. 1. Given time-series SAR imagery, the encoder part of STMA integrates local spatial features from the CNN-based ResNet module and global spatio-temporal features from multi-level attention mechanism, including STSA and MCA module. To ensure multi-granularity feature extraction, a learnable spatial attention position encoding is introduced before STSA to adaptively provide the position priors for multi-level attention learning, as illustrated in Fig. 2. Afterwards, the decoder reconstructs the extracted spatio-temporal features

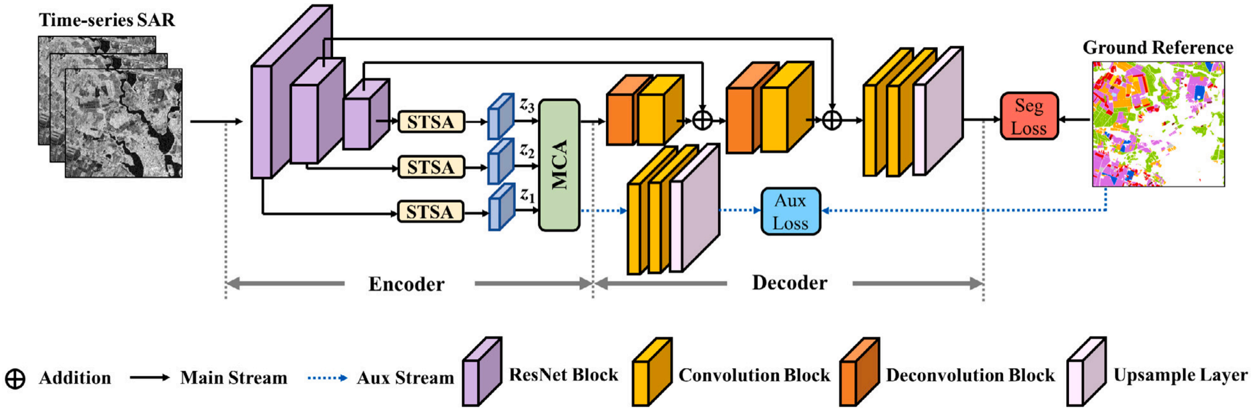


Fig. 1. The overall framework of the proposed STMA.

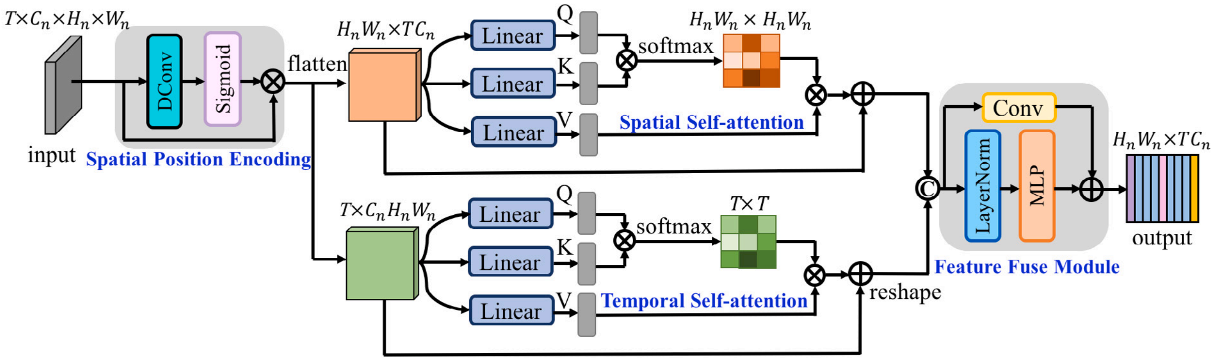


Fig. 2. The specific architecture of STSA in the proposed STMA.

hierarchically to obtain final crop mapping results with the help of principle segmentation and auxiliary supervision. In the following subsection, we provide a detailed description of each component module in the proposed STMA.

3.1. CNN-based ResNet module

Owing to strong feature extraction and noise adaptive abilities, deep CNNs have been widely applied in crop mapping tasks. However, traditional CNN-based approaches have a limited receptive field so that they usually lose some information. To retrieve more local spatial information, the Residual Network (ResNet) is often adopted as a backbone network to effectively extract the input image features by introducing the residual connection to propagate the local information from the shallow layer to the deeper layers (He et al., 2016). Therefore, in this paper, we choose the CNN-based ResNet architecture as our front-end extractor to extract the changing features from dual-polarization SAR time-series. As for the input time-series SAR imagery, we organize it into a four-dimensional tensor of shape $T \times C \times H \times W$ containing both the spatial architecture and temporal sequence, where T and C represent the sequence length and the number of channel corresponding to polarization modes, and $H \times W$ are the spatial image size. The designed ResNet network architecture consists of a standard 7×7 convolution block, a max-pooling layer and four residual blocks, where each residual block contains two 3×3 convolution operations, batch normalization (BN), rectified linear unit (ReLU) and a shortcut connection layer represented by a standard 1×1 convolution operation and BN. After experiencing different convolution and pooling operations, the spatial size of the input patch cubes is reduced by $\frac{1}{2}$, $\frac{1}{4}$, and $\frac{1}{8}$, respectively, while the channel dimension increases gradually. The adopted multi-scale features refer to the output features extracted from the CNN-based ResNet module at multiple receptive fields, corresponding to the shallow-level feature, middle-level feature and high-level feature based on different spatial resolutions. Low-level and high-level are discriminative and complementary to each other. In this process, the CNN-based ResNet module successively extracts the local spatial features for each time, and then the output features are concatenated along the temporal dimension, so that the temporal dimension information is preserved to promote the subsequent temporal self-attention module to achieve the temporal feature extraction. The output feature sequence for nth scale level is expressed as

$$\mathbf{x}_n = \left[\mathcal{F}_n(x_t)\right]_{t=1}^{T}, n \in [1,3] \tag{1}$$

where $\mathbf{x}_n \in \mathbf{R}^{T \times C_n \times H_n \times W_n}$ represents the output feature sequence with $H_n = H/2^n$ and $W_n = W/2^n$. $\mathcal{F}$ denotes the CNN-based ResNet extractor, and $[.]$ is the concatenation operator along the temporal dimension. The design of the ResNet architecture can accelerate the entire training process and prevent network degradation owing to its superior residual connection and small memory consumption.

3.2. Spatial attention position encoding

Position encoding is crucial to exploiting the order of input sequences and further helping the Transformer participate in the computation of the self-attention mechanism. The original Transformer work considers the absolute position embedding by a fixed sinusoidal encoding based on a predefined wavelength to capture positional relationships for input sequences (Vaswani et al., 2017). However, the predefined position features lack flexibility and may fail to extract important position information in a task-relevant manner, because this predefined position encoding only considers the first few dimensions of the whole embedding to store the information about the position (Kazemnejad, 2019). In addition, existing methods usually combine the position embedding with the feature map via adding operation, which is not effective to capture desired position similarity and correctly represent more distant positions in the large-area scenarios. To remedy this issue, we design a novel spatial attention position embedding module to achieve position encoding in a more flexible and data-driven way. The proposed spatial attention position embedding aims at adopting an embedded learnable function to learn complex position relationships from the feature cubes extracted by the CNN-based ResNet module, which can be represented as

$$\mathbf{x}'_n = \mathbf{x}_n \times \sigma(DConv(\mathbf{x}_n)) \tag{2}$$

where $\mathbf{x}_n \in \mathbf{R}^{H_n \times W_n \times TC_n}$ and $\mathbf{x}'_n \in \mathbf{R}^{H_n \times W_n \times TC_n}$ represent the input feature cube extracted by the CNN-based ResNet module and the output token embedding, respectively. $DConv(.)$ denotes the depth-wise convolution operation with 3 × 3 convolution kernel and padding 1, in order to extract spatial weights from the input sequences. $\sigma(.)$ is the sigmoid activation function to generate weight coefficients between 0 and 1. The depth-wise convolution applies the convolutional filter for each input channel to gather the spatial information that is conditioned on the local neighborhood of the input token and keep the translation equivalence when maintaining small trainable parameters, which provides a flexible and effective way to process input of arbitrary size without fine-tune or interpolation. By assigning different weight coefficients, each pixel in the extracted feature cube can obtain a position embedding adaptively so that the spatial information can be preserved into the output token embedding.

3.3. Spatio-temporal self-attention module

To capture the spatio-temporal relationship and phenological information of SAR time series, the STSA module is adopted to learn long-term interactions from the spatial and temporal dimension with the help of the self-attention mechanism. STSA consists of spatial self-attention, temporal self-attention and feature fuse module, as shown in Fig. 2. The spatial self-attention and temporal self-attention part independently extract the spatial and temporal information according to different flatten forms of the encoded feature cube. Then, the feature fuse module containing convolution operation, LayerNorm (LN) and multi-layer perceptron (MLP) layers, aims at integrating the concatenated features to acquire robust spatio-temporal features, avoiding the influence of manual intervention on the concatenation order of temporal and spatial features.

Firstly, the input token embedding $\mathbf{x}'_n$ is flattened along the spatial and temporal dimension into two embedded sequences $\mathbf{z}_S \in \mathbf{R}^{H_n W_n \times TC_n}$ and $\mathbf{z}_T \in \mathbf{R}^{T \times C_n H_n W_n}$. Afterwards, the flattened embedded sequences are fed to conduct the multi-head self-attention (MSA) operation along the spatial and temporal axis, in order to capture the long-term spatial and temporal information. Let $\mathbf{z}'_S \in \mathbf{R}^{H_n W_n \times TC_n}$ and $\mathbf{z}'_T \in \mathbf{R}^{T \times C_n H_n W_n}$ denote the output spatial and temporal features of the spatial and temporal self-attention, respectively. Formally,

$$\mathbf{z}'_S = MSA(LN(\mathbf{z}_S)) + \mathbf{z}_S \tag{3}$$

$$\mathbf{z}'_T = MSA(LN(\mathbf{z}_T)) + \mathbf{z}_T \tag{4}$$

$$MSA(\mathbf{z}) = Concat(head_1, \dots, head_h)\mathbf{W}^O \tag{5}$$

$$head_i = Attention(\mathbf{Q}_i, \mathbf{K}_i, \mathbf{V}_i) = softmax(\frac{\mathbf{Q}_i \mathbf{K}_i^T}{\sqrt{d_k}})\mathbf{V}_i \tag{6}$$

where h is the number of heads in MSA, and $\mathbf{W}^O$ is linear transformation matrice used for feature space transformation. $\mathbf{Q}_i = \mathbf{z}_i \mathbf{W}_i^Q$, $\mathbf{K}_i = \mathbf{z}_i \mathbf{W}_i^K$, and $\mathbf{V}_i = \mathbf{z}_i \mathbf{W}_i^V$ represent the key, query and value for each head, respectively. $\left\{ \mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V \right\}$ are linear transformation matrices whose parameters can be learned. $\sqrt{d_k}$ represents a scaling factor. The softmax activation function provides attention values from 0 to 1, and the self-attention matrix for spatial and temporal dimensions are independently acquired with size $\left[H_n W_n, H_n W_n\right]$ and $[T, T]$ to reflect the weight distribution at different pixels and times.

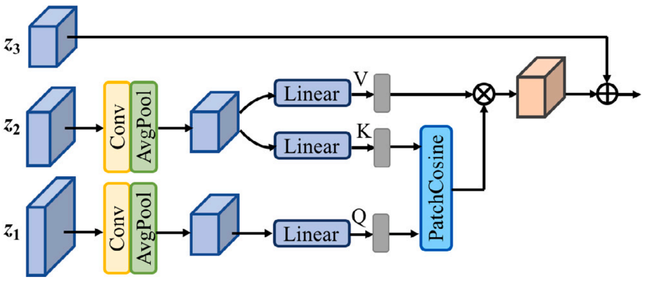


Fig. 3. The specific architecture of MCA in the proposed STMA.

To accomplish the aggregation of spatial and temporal features, the feature fuse module is designed by concatenating these extracted features to perform a feed-forward network similar to the original Transformer. The output of STSA can be written as follows

$$\mathbf{d} = MLP(LN(Concat(\mathbf{z}'_S, \mathbf{z}'_T))) + \mathbf{w}\hat{\mathbf{z}} \tag{7}$$

where $\mathbf{d} \in \mathbf{R}^{H_n W_n \times TC_n}$ is the output spatio-temporal feature. The MLP layer is composed of two fully connection layers with a Gaussian error linear unit (GELU) activation. $\mathbf{w} \in \mathbf{R}^{1\times 2}$ is the learnable network parameter for adaptive fusion. $\hat{\mathbf{z}} = [\mathbf{z}'_S, \mathbf{z}'_T]^T \in \mathbf{R}^{2\times H_n W_n \times TC_n}$ represents the concatenation results of $\mathbf{z}'_S$ and $\mathbf{z}'_T$ in the third dimension, so that the local structural information can be preserved from the convolution operation to avoid insufficient fusion. After STSA, the output spatio-temporal feature $\mathbf{d}$ is reshaped into 2D image space to generate discriminative spatio-temporal feature $\mathbf{z}_1 \in \mathbf{R}^{H_1 \times W_1 \times TC_1}$. Similarly, as for different multi-scale features obtained by the CNN-based ResNet module, the middle-level feature $\mathbf{z}_2 \in \mathbf{R}^{H_2 \times W_2 \times TC_2}$ and high-level feature $\mathbf{z}_3 \in \mathbf{R}^{H_3 \times W_3 \times TC_3}$ can also be obtained by the STSA module.

3.4. Multi-scale cross-attention module

To integrate the contributions from different scale levels, the MCA module based on cosine similarity is designed to capture long-distance correlations between multi-scale spatio-temporal features. Inspired by the cosine normalization work in the neural networks (Luo et al., 2018), we adopt cosine similarity to the cross-attention mechanism instead of the scaled dot product to capture more differences between multi-scale features and produce stable attention results. Given the multi-scale spatio-temporal features $\mathbf{z}_1$, $\mathbf{z}_2$ and $\mathbf{z}_3$, the MCA mechanism based on cosine similarity is expressed as

$$\begin{aligned}\mathbf{z}_m &= MCA(\mathbf{Q}_C, \mathbf{K}_C, \mathbf{V}_C) + \mathbf{z}_3 \\ &= PatchCosine(\mathbf{Q}_C, \mathbf{K}_C) \cdot \mathbf{V}_C + \mathbf{z}_3 = \frac{\mathbf{Q}_C \cdot \mathbf{K}_C^T}{\mathbf{M}_Q \otimes \mathbf{M}_K} \cdot \mathbf{V}_C + \mathbf{z}_3\end{aligned} \tag{8}$$

where $\mathbf{Q}_C = \mathbf{z}_1 \mathbf{W}_C^Q$, $\mathbf{K}_C = \mathbf{z}_2 \mathbf{W}_C^K$, and $\mathbf{V}_C = \mathbf{z}_2 \mathbf{W}_C^V$ represent the query, key and value for the MCA mechanism, respectively. ⊗ denotes the outer product operation. $\mathbf{M}_Q$ and $\mathbf{M}_K$ are the magnitude value of $\mathbf{Q}_C$ and $\mathbf{K}_C$. To unify the size of different multi-scale features, we adopt the 1 × 1 convolution and average pooling operations to balance the number of channels and the spatial size of $\mathbf{z}_1$ and $\mathbf{z}_2$ equal to high-level feature $\mathbf{z}_3$. Finally, by adding valuable semantic information of the high-level feature $\mathbf{z}_3$ to the generated feature, the final aggregated spatio-temporal feature $\mathbf{z}_m$ is obtained with rich multi-scale information (see Fig. 3).

3.5. Decoder

To keep the size of the extracted spatio-temporal feature map consistent with the size of ground reference for crop mapping tasks, a CNN-based decoder architecture is adopted by progressively reconstructing the feature map to maintain semantic detailed information with the help of deconvolution and standard convolution. For balancing both computational efficiency and reconstruction accuracy, we adopt one 3 × 3 deconvolution convolution block and one standard 3 × 3 convolution block in the decoder part to effectively project the high dimensional feature map into a low dimension. The purpose of the deconvolution convolution is to upscale feature maps by inserting virtual zeros between the adjacent input pixels (Dumoulin and Visin, 2016). Here, the encoded spatio-temporal feature map $\mathbf{z}_m$ is integrated with the value of previous ResNet flow by the skip connection to enlarge the spatial size of feature map and ease the training process. Finally, one standard 3 × 3 convolution block and one standard 1 × 1 convolution layer are applied to post-process the newly upscaled feature and the encoded spatio-temporal feature $\mathbf{z}_m$, respectively, to produce the final crop mapping result $\mathbf{y} \in \mathbf{R}^{K\times H\times W}$ and the auxiliary map $\mathbf{y}_{aux} \in \mathbf{R}^{K\times H\times W}$ with the number of category K.

3.6. Loss function

As stated before, the objective function of the proposed STMA is realized by minimizing the principle cross-entropy (CE) loss function and the auxiliary loss function. Specifically, the principle CE loss is adopted to train the whole STMA network to learn the mapping from the input time-series SAR to ground reference data, while the auxiliary loss is used to supervise the optimization process of the multi-level attention by preventing detailed information loss in intermediate network layers, so that these encoder–decoder CNN-transformer-based architectures can learn better multi-scale spatio-temporal feature representations, thereby improving crop mapping performance in semantic view. The overall loss of STMA can be formulated as

$$L_{All} = \alpha L_{CE} + (1-\alpha)L_{Aux} \tag{9}$$

where α is utilized to balance the trade-off between the principle CE loss and the auxiliary loss. Note that, the auxiliary loss function is calculated by minimizing the CE loss between the auxiliary map $\mathbf{y}_{aux}$ generated by the CNN-based decoder and the ground reference map $\hat{\mathbf{y}}$, and is only adopted in the training phase, not affecting the testing phase. The principle CE loss function considers the reconstruction loss between the output crop mapping result $\mathbf{y}$ and the ground reference data $\hat{\mathbf{y}}$. To be specific, the multi-class CE loss for L_{CE} and L_{Aux} is calculated as follows

$$L = -\frac{1}{N}\sum_{i=1}^{N}\hat{y}_i\log(y_i) \tag{10}$$

where N represents the number of pixels. $\hat{y}_i$ and y_i denote the one-hot encoding of the true label and the corresponding softmax output of STMA at the ith image pixel.

4. Experiments and results

4.1. Experimental datasets

To evaluate the performance of our proposed crop mapping method, we adopt three Sentinel-1 time-series SAR datasets in the experiment, including Brandenburg Sentinel-1 dataset, public PASTIS-R dataset and South Africa dataset. Different datasets are modeled separately for training, validation and testing.

The Brandenburg Sentinel-1 dataset covers two study areas, S1 and S2, in the northwest and southeast part of Brandenburg state, Germany, latitude 51°47′N∼52°58′N, longitude 12°3′E∼14°42′E, as shown in Fig. 4(a). Compared to other German states, approximately 45% of Brandenburg is dedicated to agricultural land use and a large amount of regional food production needs to be supplied, due to its geographical location close to the city of Berlin, the capital of Germany (Wolff et al., 2020). Specifically, the adopted time-series SAR imagery has 3731 × 5095 pixels with 10 m spatial resolution and its corresponding revisit time is stable at 12 days. Table 1 lists the specific acquisition data description of the selected time-series Sentinel-1A SAR data within two time phases: from 2017 to 2018 and from 2020 to 2021. In the two study areas, S1 and S2, 15 categories were investigated to validate the effectiveness of multi-class crop mapping, including some major crops (e.g., maize, wheat and rapeseed), and other land use types (e.g., fallow and residual), and Table 2 presents an overview of the ground reference data. Other rare crop and land use types were aggregated as the “Other” category. The time-series curve of six main crops are displayed in Fig. S1, which reflects a certain crop phenology difference.

The open-access PASTIS-R dataset is a benchmark dataset for panoptic and semantic segmentation of crop mapping from time-series Sentinel-1 and Sentinel-2 observations, as illustrated in Fig. 4(b). It contains 19 categories and 2433 patches within different regions of the France metropolitan territory with semantic annotations for each pixel. For each patch with size 128 × 128, there are 70 observations of Sentinel-1 sensor acquired from ascending (S1A) and descending (S1D) orbits without any speckle filtering and terrain correction processing. PASTIS-R provides the time-sereies SAR of three different modalities, including vertical polarization (VV), horizontal polarization (VH), and the ratio of vertical over horizontal polarization (VV/VH). More details about PASTIS-R dataset can refer to Garnot and Landrieu (2021b). We adopted the official 5 fold split provided in the dataset’s metadata to evaluate the performance of the proposed STMA.

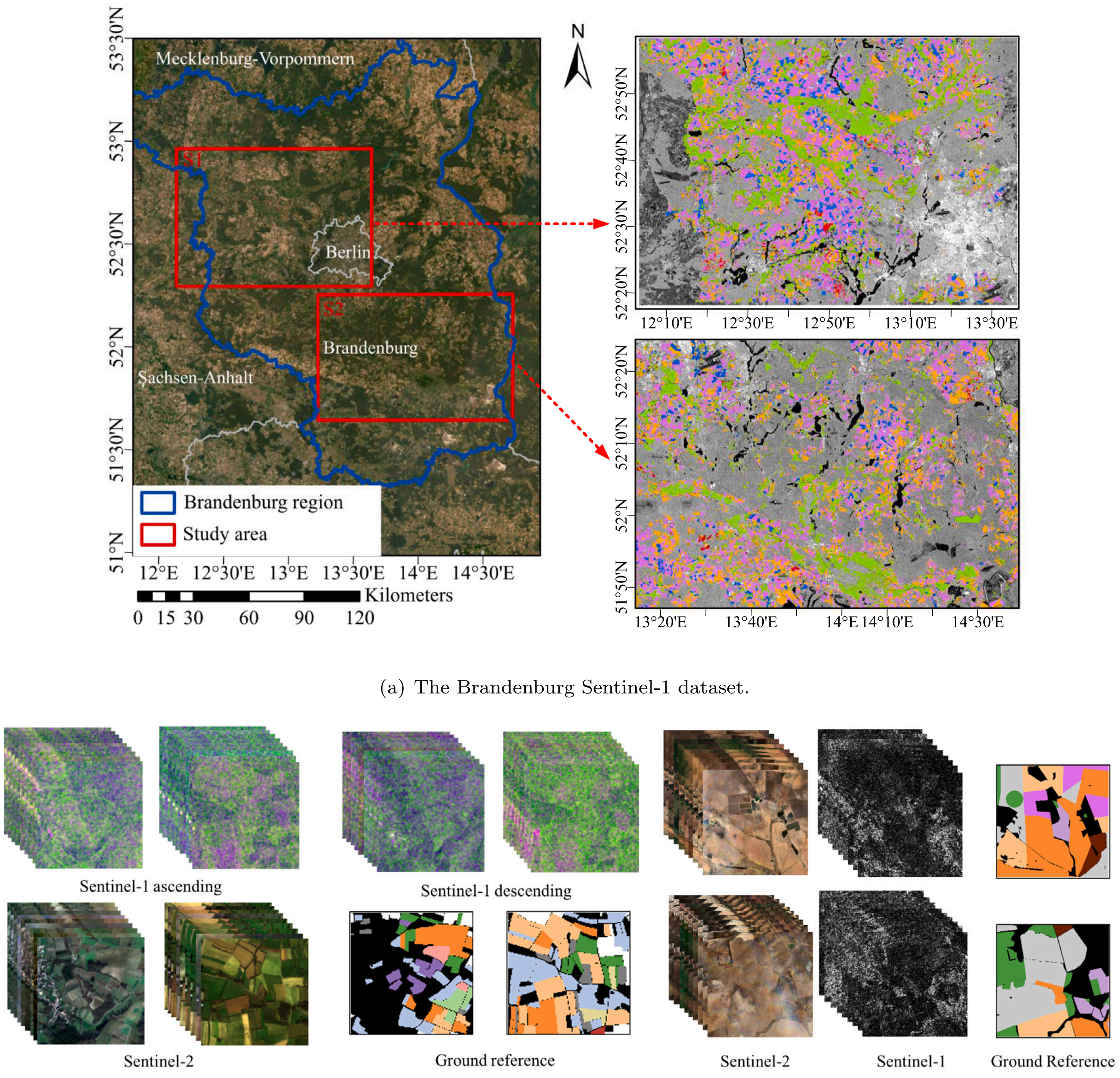
The South Africa dataset is produced as part of the Radiant Earth Spot the Crop Challenge. It collects the time-series data of Sentinel-1 and Sentinel-2 satellite from 2017 to 2018 to classify crops in the Western Cape of South Africa (Western Cape Department of Agriculture, 2021), as illustrated in Fig. 4(c). The South Africa dataset consists of small-holder farms and sparse ground truth label, and the growing season in this study area is dominated by rain and cloud cover, leading to low visibility in remote sensing imagery. In addition, the tropical climate allows for the existence of double-season or multiple cropping systems in South Africa (Waha et al., 2020), further increasing the difficulty of crop identification and mapping. There are 2650 patches for training and 1137 patches for testing with size 256 × 256, including 9 crop classes and 21 observations of VV and VH modalities in the time-series SAR imagery.

These three time-series SAR datasets are collected by Sentinel-1 observations in different regions, and various crop types are considered to achieve crop mapping. Existing crop mapping researches only focused on limited crop categories, and some crops were included as “others” to evaluate the performance of algorithm, especially for crops with similar growth characteristics, which is not persuasive for model evaluation and can also be interpreted as misclassification (Wang et al., 2023). The adopted time-series SAR datasets in this paper have various difficulties in terms of data complexity, especially for a variety of crop categories and different cropping system, and are representative for evaluating crop mapping approaches, which can validate the effectiveness of crop mapping methods from different circumstances and perspectives.

4.2. Experimental design

In this section, three experiments were performed to validate the proposed STMA in the crop mapping task, including comparative evaluation, assessment of spatio-temporal generalization, and ablation study. In addition, overall accuracy (OA), F1 score and intersection over union (IoU) were selected as evaluation metrics to evaluate the accuracy of different crop mapping approaches. The definitions of OA, precision, recall, F1 score and IoU are

$$OA = \frac{TN+TP}{TN+TP+FP+FN} \tag{11}$$



(a) The Brandenburg Sentinel-1 dataset.

(b) The PASTIS-R dataset.

(c) The South Africa dataset.

Fig. 4. Dataset introduction of Brandenburg Sentinel-1, PASTIS-R and South Africa.

Table 1
The acquisition data description of time-series Brandenburg Sentinel-1.

Description	Acquisition date/Day of year (DOY)						
Phase 1 from 2017 to 2018	2017–09–05	2017-09-17	⋯	2018–03–16	⋯	2018–12–17	2018–12–29
	248	260	⋯	75	⋯	351	363
Phase 2 from 2020 to 2021	2020–09–01	2020-09-13	⋯	2021–03–12	⋯	2021–12–13	2021–12–25
	245	257	⋯	71	⋯	347	359

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{14}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{15}$$

where TP, TN, FP and FN denote the number of true positive, true negative, false positive and false negative samples, respectively.

For all experiments presented, we divided the Bradenburg Sentinel-1 dataset into three parts based on the patch size of 128 × 128; a training dataset, validation dataset and testing dataset in the ratio of 5 to 1 to 4 by non-overlapping cropping, and the PASTIS-R dataset partition adopted 5-fold cross-validation followed the original

Table 2
Overview of ground reference data including 15 categories in S1 and S2 study area.

	Numbers of parcels in S1	Area proportion in S1	Number of parcels in S2	Area proportion in S2
Maize	15,874	20.79%	21,397	25.67%
Wheat	19,072	24.98%	20,060	24.07%
Grassland	33,501	43.87%	34,008	40.81%
Peanut	1,922	2.52%	1,469	1.76%
Potato	573	0.75%	774	0.93%
Residue	832	1.09%	706	0.85%
Fallow	603	0.79%	492	0.59%
Rapeseed	2,319	3.04%	2,406	2.89%
Vegetable	352	0.46%	586	0.70%
Legume	948	1.24%	1,108	0.02%
Herb	34	0.04%	18	0.01%
Orchard	141	0.18%	92	0.11%
Flower	66	0.09%	16	0.02%
Sugar beet	52	0.07%	137	0.16%
Other	70	0.09%	71	0.09%
Total	76,359	100%	83,340	100%

research work[1] for a fair comparison (Garnot et al., 2022). As for the South Africa dataset, we adopted the fixed training and testing partition sets in the official website.[2] Note that, the partition sets of training, validation and test are completely independent on all adopted datasets. Moreover, six classic and state-of-the-art crop mapping methods based on traditional machine learning and DL were selected for comparison, namely RF (Sonobe et al., 2014), LSTM (Crisóstomo de Castro Filho et al., 2020), CNN-LSTM (ConvLSTM) (Shi et al., 2015), 3D UNet (Adrian et al., 2021), U-TAE (Garnot and Landrieu, 2021a) and TPE-RNN (Nyborg et al., 2022). Among them, ConvLSTM, 3D UNet, U-TAE and TPE-RNN consider both spatial and temporal information in multi-temporal SAR processing, so they can provide reasonable benchmarks to compare the spatio-temporal performance of the proposed STMA. Note that, the parameter settings of six comparison approaches refer to the original literature in our experiments and are listed as follows. The number of decision tree in RF is 500 at a maximum depth of 25 and the hidden dimensionalities of LSTM are set to {32, 64, 128} with a dropout rate of 0.1. ConvLSTM has three layers with 256, 128, 64 hidden states respectively and 5 × 5 kernel size. 3D UNet consists of three encoder layers with 3 × 3 × 3 convolution and three decoder layers with 2 × 2 × 2 deconvolution. U-TAE has three spatial encoding layers with 4 × 4 convolution and one LTAE temporal encoding layer with 16 heads to extract the spatio-temporal features, and then three spatial decoding layers aim at upsampling the previous feature map with the help of 4 × 4 transposed convolution. TPE-RNN considers one GRU position encoding layer and the PSE+LTAE architecture with 16 heads to capture the spatio-temporal characteristics.

For a fair comparison, all experiments were executed using a computer with an Intel Core i7 and an NVIDIA GTX 1080Ti 11-GB GPU. The DL-based methods were implemented on the Pytorch framework, and the Adam optimizer with the learning rate of $1e-3$ was adopted to update the network parameters during the training process. The minibatch size was set to 32, and the maximum number of epoch was set to 500. The number of heads h and the trade-off loss parameter α in the STMA were empirically set to 4 and 0.5.

4.3. Comparative evaluation

Table 3 shows a quantitative comparison of results for different crop mapping approaches on Brandenburg Sentinel-1 dataset, PASTIS-R dataset and South Africa dataset. It can be obviously seen that the proposed STMA can obtain the highest OA, mF1 and mIoU among all DL and non-DL comparison methods. Besides, the STMA method takes less training time than other methods, such as 3D UNet and TPE-RNN, further illustrating its relatively lightweight model complexity. For crop types with a large proportion in the experimental scene, such as maize, wheat, grassland, rapeseed and grapevine, the STMA method can achieve more than 88% of the F1-score on Brandenburg Sentinel-1 dataset, more than 80% of the F1-score on PASTIS-R dataset, and more than 70% of the F1-score on South Africa dataset as shown in Table 4, Tables 5 and 6. The crop mapping performance of 3D UNet (94.82% and 95.98%) is superior to the other DL models (LSTM: 88.99% and 89.26%, ConvLSTM: 92.06% and 93.67%, U-TAE: 92.76% and 93.82%, and TPE-RNN: 93.51% and 94.31%) on Brandenburg Sentinel-1 dataset, since it considers the spatial and temporal information simultaneously owing to the designed 3D convolution kernel architecture. Nevertheless, the performance of 3D UNet (81.99% and 82.68%) is inferior to that of U-TAE (83.14% and 83.63%) and TPE-RNN (85.81% and 86.25%) on PASTIS-R dataset, because this local convolution model is easily disturbed by speckle noise and it is not as robust as the self-attention-based model considering the global structure of time-series SAR imagery. In addition, RF can only distinguish the main crop types, and fails to identify the crop types that account for a small proportion in the scene. The main reason is that the RF method is prone to overfit the training data when the decision tree grows too large for multi-class crop types (Jin et al., 2018), so that it cannot provide better crop mapping accuracy. Figs. 5, 6 and 7 display the corresponding crop mapping results of different comparison methods, including three regions zoomed for more detailed observation. The visualization results from Figs. 5, 6 and 7 illustrate that the proposed STMA can extract more accurate crop information in the large-area scenario, especially for crop categories with a small proportions, such as peanut and sugar beet on Brandenburg Sentinel-1 dataset, mixed cereal and sorghum on PASTIS-R dataset and weed and rooibos on South Africa dataset.

To demonstrate the effectiveness of combining spatio-temporal features for different crop mapping methods, feature separation comparison was undertaken by using t-distributed stochastic neighbor embedding (t-SNE) (Van der Maaten and Hinton, 2008) to project the extracted feature representations into a two-dimensional space, as shown in Fig. 8. Note that, the adopted features are output by the last network layer, which represents the learned spatio-temporal features in different crop mapping approaches. It can be seen from Fig. 8 that the degree of feature separation for different crop types is greatly different in the two-dimensional feature space. Compared to other state-of-the-art approaches, different crop features extracted by STMA are better grouped than the raw input representations, because STMA considers multi-scale spatial and temporal information from global and local perspectives with the help of the multi-level attention mechanism. However, other comparison methods are deficient in multi-category feature extraction for time-series SAR. For example, a small number of crop feature points are still mixed, such as legume and sugar beet.

[1] https://zenodo.org/record/5735646
[2] https://doi.org/10.34911/rdnt.j0co8q

Table 3
Quantitative comparison of results for different crop mapping approaches on Brandenburg Sentinel-1 dataset, PASTIS-R dataset and South Africa dataset, where OA, mean F1 (mF1), mean IoU (mIoU), and training time are reported. The best performance of each indicator is shown in **bold**.

Method	Brandenburg S1			Brandenburg S2			PASTIS-R S1A			PASTIS-R S1D			South Africa			Time (h)
	OA	mF1	mIoU	OA	mF1	mIoU	OA	mF1	mIoU	OA	mF1	mIoU	OA	mF1	mIoU	
RF	82.65%	23.05%	18.47%	87.81%	25.65%	19.18%	59.64%	17.45%	12.04%	58.01%	16.46%	11.01%	66.53%	11.61%	8.47%	3.16
LSTM	88.99%	50.78%	36.79%	89.26%	53.12%	40.42%	76.20%	47.46%	35.21%	78.70%	49.84%	36.86%	72.33%	41.56%	28.10%	**2.84**
ConvLSTM	92.06%	51.05%	38.64%	93.67%	56.14%	45.33%	77.74%	51.53%	38.65%	80.31%	53.64%	40.31%	73.92%	46.28%	32.12%	3.09
3D UNet	94.82%	69.43%	55.54%	95.98%	72.48%	59.32%	81.99%	54.24%	41.63%	82.68%	55.01%	43.72%	78.63%	47.37%	34.93%	5.33
U-TAE	92.76%	53.39%	40.84%	93.82%	58.51%	47.54%	83.14%	59.76%	45.71%	83.63%	60.79%	46.73%	79.88%	50.22%	36.97%	3.65
TPE-RNN	93.51%	56.44%	46.23%	94.31%	65.58%	54.27%	85.81%	61.19%	47.97%	86.25%	62.35%	48.89%	80.24%	50.83%	37.51%	4.04
STMA	**96.54%**	**74.69%**	**63.81%**	**97.59%**	**81.30%**	**70.87%**	**86.77%**	**64.97%**	**51.59%**	**87.32%**	**65.74%**	**52.62%**	**83.37%**	**57.81%**	**43.31%**	3.81

Table 4
F1-score comparison per class for different crop mapping approaches on Brandenburg Sentinel-1 dataset. The best performance of indicator is shown in **bold**.

Crop type	RF	LSTM	ConvLSTM	3D UNet	U-TAE	TPE-RNN	STMA
Maize	20.69%	66.14%	70.72%	80.16%	70.86%	71.03%	**88.63%**
Wheat	78.06%	80.50%	87.58%	91.96%	87.57%	87.55%	**95.37%**
Grassland	74.00%	81.06%	87.70%	91.59%	85.65%	83.69%	**92.62%**
Peanut	0	55.71%	57.83%	72.75%	41.16%	67.33%	**81.82%**
Potato	66.70%	69.59%	69.93%	83.37%	76.82%	80.90%	**91.56%**
Residue	0	53.19%	56.34%	72.23%	48.91%	59.01%	**74.30%**
Fallow	0	33.04%	30.24%	63.02%	24.05%	51.30%	**64.37%**
Rapeseed	91.82%	78.70%	87.01%	92.99%	92.89%	93.65%	**94.99%**
Vegetable	0	30.62%	25.43%	**50.96%**	34.25%	22.25%	49.55%
Legume	14.47%	57.98%	61.84%	75.02%	58.99%	51.10%	**87.22%**
Herb	0	36.72%	34.85%	61.48%	37.85%	58.31%	**67.09%**
Orchard	0	11.70%	0.64%	43.95%	28.32%	19.32%	**44.57%**
Flower	0	29.35%	19.65%	52.85%	39.17%	13.21%	**62.81%**
Sugar beet	0	27.26%	23.75%	**46.39%**	22.73%	28.53%	41.25%
Other	0	50.13%	52.22%	62.69%	51.61%	59.35%	**84.23%**

Table 5
F1-score comparison per class for different crop mapping approaches on PASTIS-R dataset. The best performance of indicator is shown in **bold**.

Crop type	RF	LSTM	ConvLSTM	3D UNet	U-TAE	TPE-RNN	STMA
Meadow	60.21%	70.02%	72.37%	77.79%	77.96%	79.50%	**81.52%**
Soft winter wheat	69.90%	80.38%	81.59%	81.31%	82.43%	82.21%	**84.21%**
Corn	47.65%	79.07%	80.69%	80.22%	81.90%	82.62%	**83.10%**
Winter barley	37.02%	64.02%	71.84%	73.9%	74.79%	77.26%	**78.61%**
Winter rapeseed	60.10%	82.29%	80.95%	83.23%	85.23%	85.23%	**86.80%**
Spring barley	0.05%	22.25%	37.11%	50.01%	59.75%	**60.77%**	60.51%
Sunflower	0.05%	45.97%	54.37%	51.94%	**71.96%**	71.77%	64.37%
Grapevine	0	56.14%	60.27%	59.07%	68.58%	73.20%	**80.46%**
Beet	45.11%	72.5%	76.49%	74.26%	80.04%	76.53%	**81.71%**
Winter triticale	0	25.21%	29.88%	45.33%	44.96%	42.51%	**49.91%**
Winter durum wheat	5.25%	50.41%	57.97%	57.41%	**63.98%**	60.31%	60.62%
Fruits	3.09%	47.57%	48.48%	49.38%	64.26%	60.67%	**65.56%**
Potatoes	0.01%	19.44%	28.89%	34.51%	26.92%	47.16%	**52.43%**
Leguminous fodder	0	20.41%	16.06%	23.60%	32.1%	31.89%	**38.45%**
Soybeans	0.34%	67.23%	69.57%	63.50%	74.55%	74.37%	**77.10%**
Orchard	0	22.29%	26.04%	30.69%	41.18%	46.70%	**59.75%**
Mixed cereal	0	12.54%	20.31%	27.20%	29.31%	26.04%	**38.54%**
Sorghum	0	25.14%	28.23%	27.37%	35.12%	38.28%	**40.81%**
Void label	4.81%	38.86%	37.96%	39.92%	40.36%	45.60%	**50.08%**

Table 6
F1-score comparison per class for different crop mapping approaches on South Africa dataset. The best performance of indicator is shown in **bold**.

Crop type	RF	LSTM	ConvLSTM	3D UNet	U-TAE	TPE-RNN	STMA
Lucerne	10.48%	42.62%	49.04%	52.38%	46.36%	47.44%	**56.98%**
Planted pasture	0	28.73%	27.76%	21.14%	32.02%	36.43%	**45.26%**
Fallow	0.12%	35.31%	42.18%	53.88%	**57.47%**	41.53%	55.03%
Wine grape	0	64.01%	71.41%	76.89%	67.42%	69.80%	**79.96%**
Weed	0	18.72%	20.65%	17.31%	38.66%	43.51%	**46.34%**
Maize	2.83%	30.49%	36.40%	34.32%	28.52%	34.65%	**42.19%**
Wheat	55.95%	73.01%	77.85%	77.16%	79.39%	76.34%	**83.74%**
Rapeseed	35.09%	52.04%	53.98%	49.70%	55.98%	57.97%	**59.62%**
Rooibos	0	29.10%	37.28%	43.59%	46.13%	49.77%	**51.21%**

4.4. Assessment of spatio-temporal generalization

Based on the proposed method, the spatio-temporal generalization analysis was analyzed using the Brandenburg Sentinel-1 dataset from 2020 to 2021 in the S2 study area, to evaluate different crop mapping approaches as shown in Table 7. The adopted training data for generalization assessment was the S1 study area from 2017 to 2018, and we only considered six main crops to assess the spatio-temporal

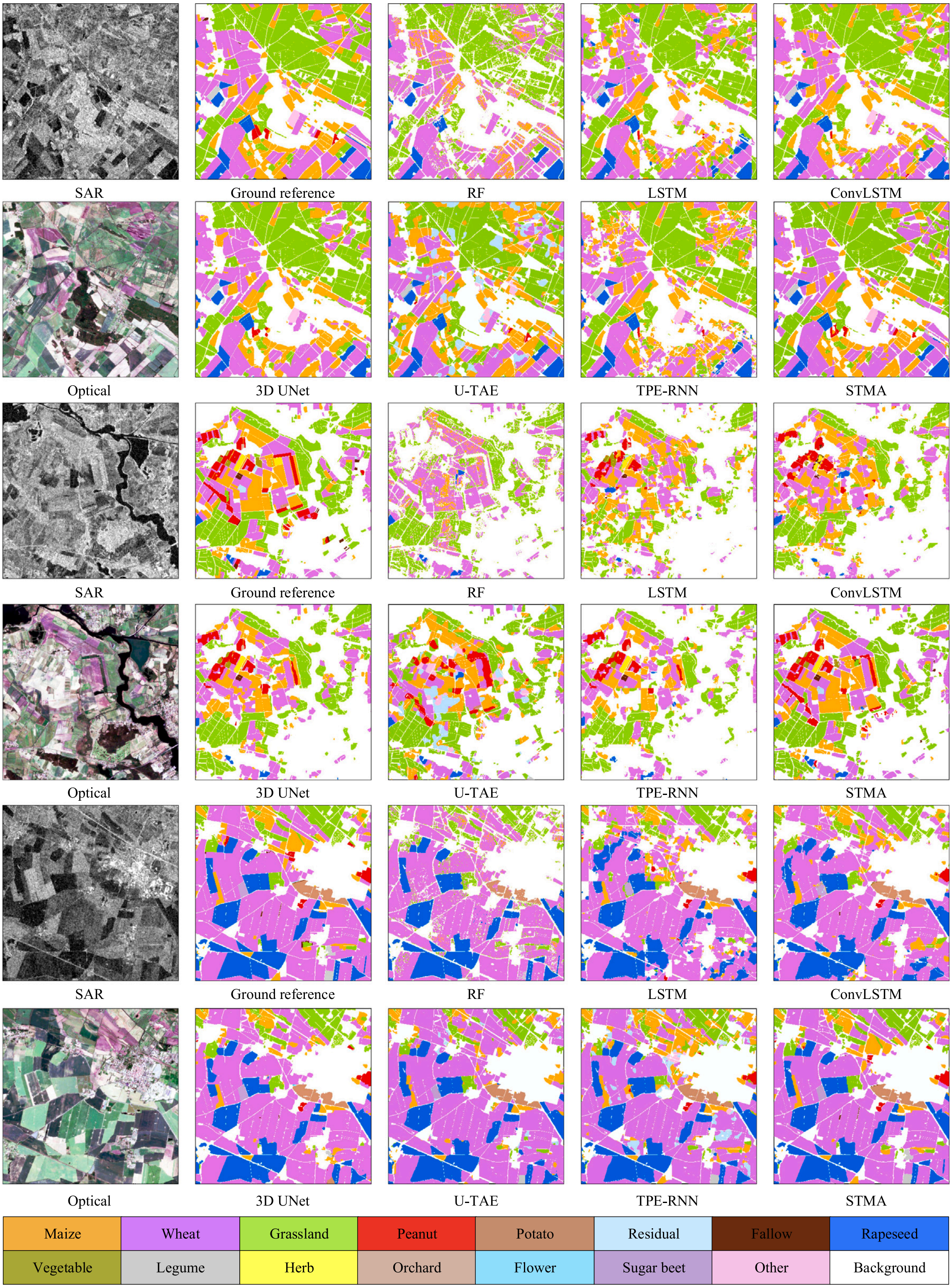


Fig. 5. Crop mapping results of different non-DL and DL methods on Brandenburg Sentinel-1 dataset, including three regions in the study area.

generalization due to imbalance crop distribution, including maize, wheat, grassland, potato, rapeseed and legume. Taking the rapeseed as example in Fig. 9, the temporal signatures of the backscattering coefficient at VH and VV are relatively similar between different areas and years, but there are still some phenology shifts. The main reason may be that the annual crop planting time in the Brandenburg study area is not exactly the same and may be affected by natural disasters to reflect the time-varying differences of the backscattering coefficients. In addition, the long-span time-series SAR data is considered as the input to train the crop mapping model, e.g., from Sep. 2017 to Dec. 2018, so that the proportion of important crop phenology changes is relatively small and increases the difficulty of spatio-temporal generalization on the Brandenburg Sentinel-1 dataset. To validate the effectiveness of cross-region generalization performance, two transferable DL-based

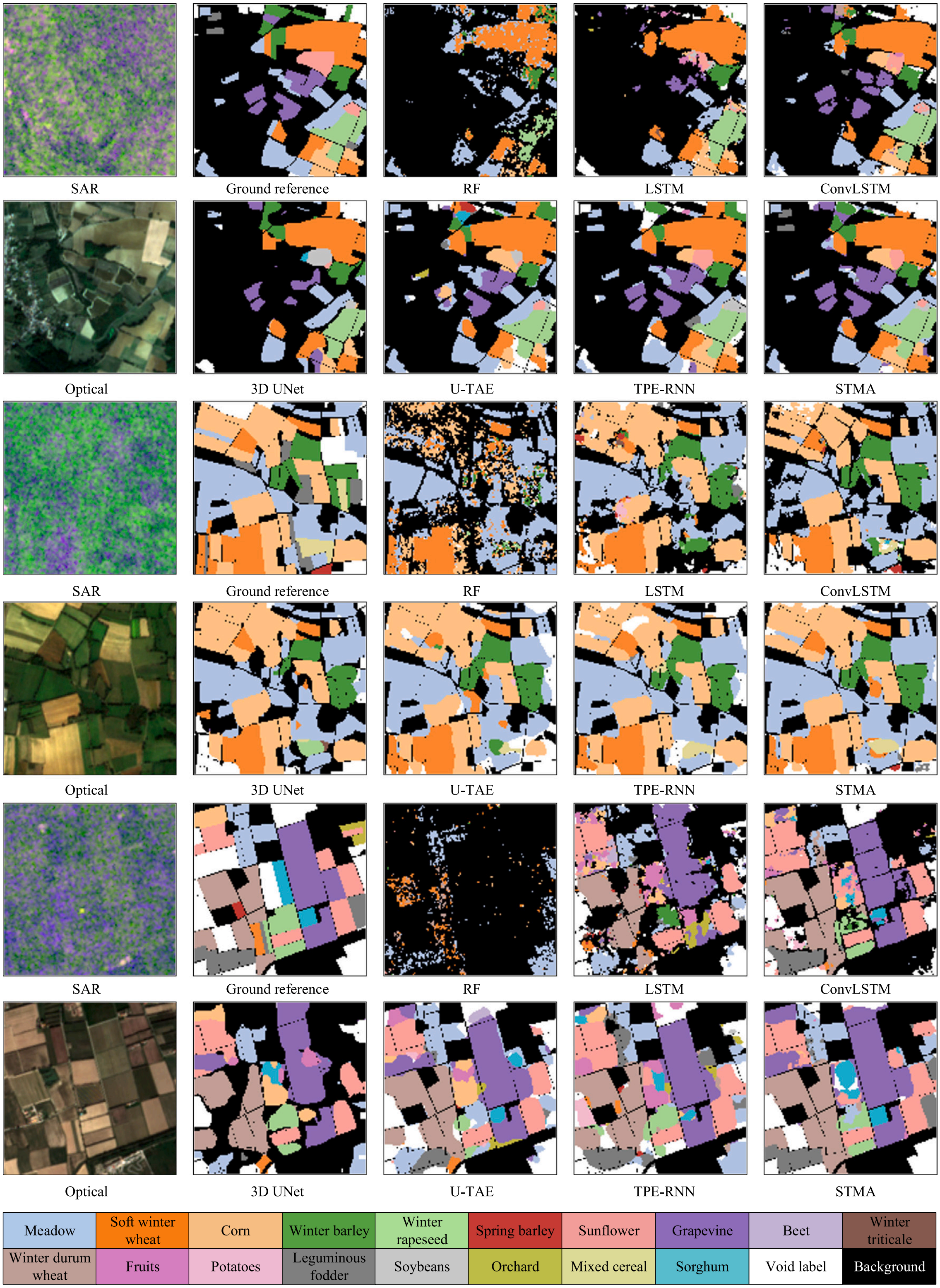


Fig. 6. Crop mapping results of different non-DL and DL methods on PASTIS-R dataset, including three regions in the study area.

crop mapping approaches were added for comparison in this section, namely transferable UNet (TUNet) (Ge et al., 2021) and phenology alignment network (PAN) (Wang et al., 2022). On the whole, it can be seen that STMA obtained the highest mF1 score of 49.83% and produced better spatio-temporal generalization performance, thanks to the capture of additional spatial and temporal information through the multi-level attention mechanism. Compared with 3D UNet and U-TAE, TPE-RNN and PAN achieved the second and third highest F1-score for a different year and place, indicating the generalization advantages of the self-attention-based models in extracting spatial and temporal information based on time-series SAR imagery. Although the time window used by TUNet maintains a certain distribution of target data, it only considers the local phenology of crops, and the complexity of SAR scattering characteristics of different crops is not

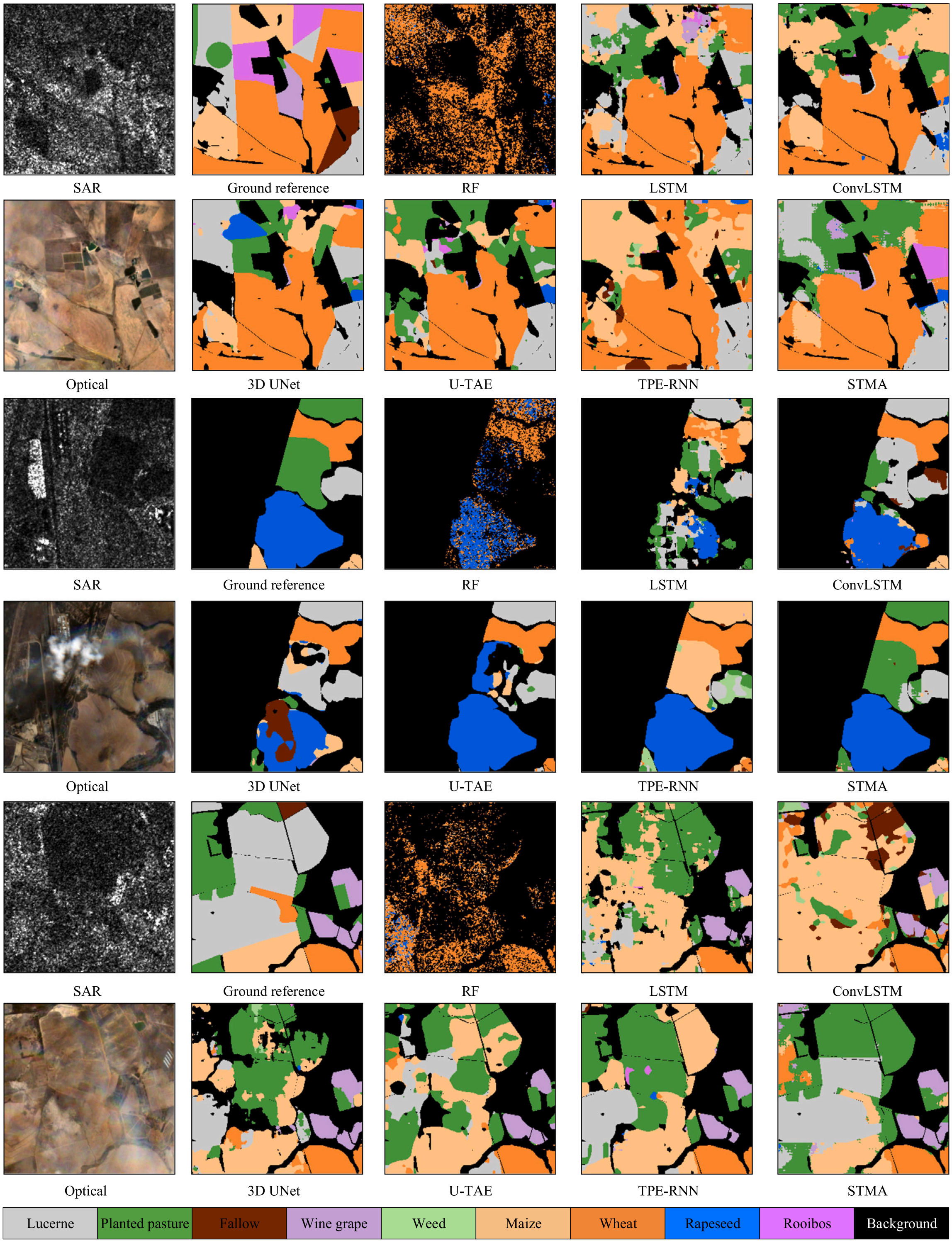


Fig. 7. Crop mapping results of different non-DL and DL methods on South Africa dataset, including three regions in the study area.

suitable for this local time window method. CNN-based and RNN-based crop mapping approaches, such as LSTM and ConvLSTM, are difficult to handle the subtle variability of time-series SAR imagery in different spatio-temporal scenarios, because they cannot capture the universal spatio-temporal characteristics of different crops based on limited receptive field of the network. Moreover, the spatio-temporal generalization performance of our STMA model with the varying number of available training samples is reported in Fig. 10. It can be seen that the mean F1 value gradually improves with the increase in the percentage of available samples, and it is tending toward stability when more training samples (e.g., up to 80%) are involved, showing the spatio-temporal generalization stability of the proposed STMA to a great extent.

4.5. Ablation study

To validate the effectiveness of the proposed STSA module, the MCA module, spatial attention position embedding module and the

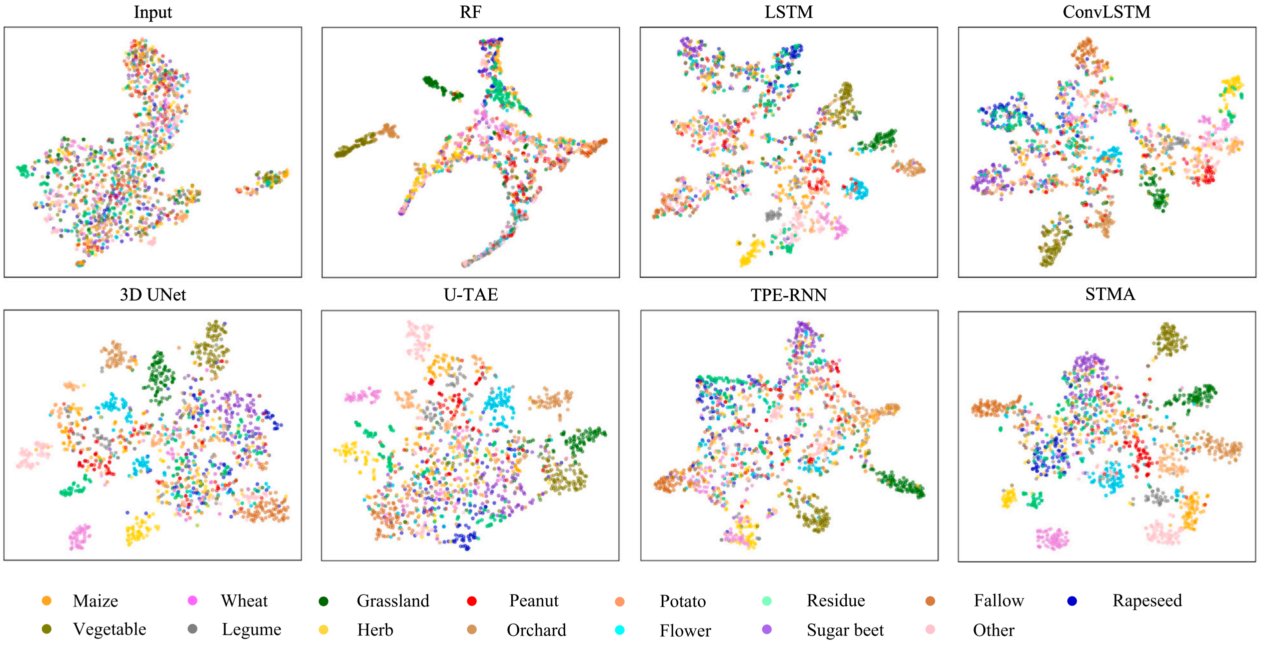


Fig. 8. Feature separation comparison of different crop mapping methods based on the t-SNE distribution.

Table 7
F1-score comparison of six main crops for spatio-temporal generalization analysis. The best performance of indicator is shown in **bold**.

Method	Maize	Wheat	Grassland	Potato	Rapeseed	Legume	mF1
RF	13.66%	17.74%	20.63%	16.77%	46.58%	6.21%	20.26%
LSTM	42.51%	35.24%	38.10%	21.42%	28.24%	10.72%	29.37%
ConvLSTM	47.65%	41.34%	48.23%	20.35%	32.65%	20.01%	35.04%
3D UNet	50.63%	42.26%	51.11%	34.39%	46.81%	31.79%	42.83%
U-TAE	49.68%	48.72%	54.79%	36.30%	49.57%	28.65%	44.62%
TPE-RNN	51.96%	47.52%	56.92%	37.91%	52.16%	30.16%	46.10%
TUNet	48.32%	43.17%	46.59%	27.21%	42.92%	24.64%	38.81%
PAN	49.12%	49.86%	55.61%	34.87%	50.99%	29.85%	45.05%
STMA	**52.55%**	**53.18%**	**61.15%**	**43.04%**	**55.27%**	**33.79%**	**49.83%**

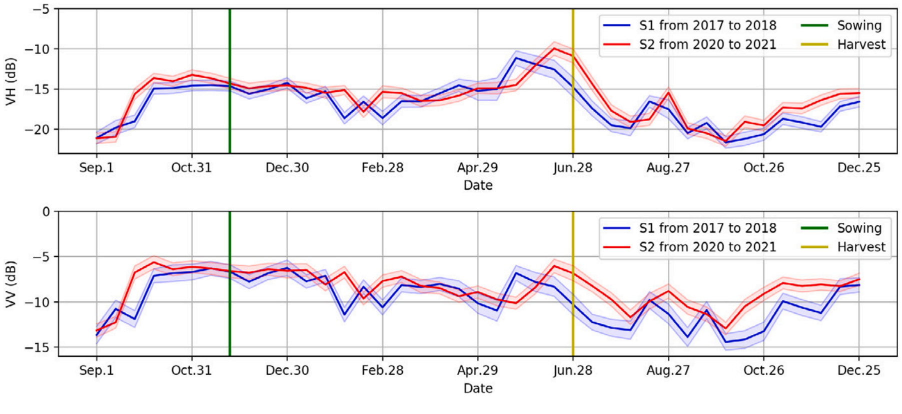


Fig. 9. Time-series curves of backscattering coefficients at VH and VV of rapeseed for different areas and years. The shaded areas refer to the standard deviation calculated from 1000 sample points.

combined loss function, we conduct the ablation experiment for STMA as shown in Table 8. Specifically, different spatio-temporal interaction, multi-scale feature integration, position embedding and loss function strategies are adopted to compare the crop mapping performance under different ablation experimental settings, including the combination extraction of spatio-temporal interactions, the scaled dot product attention and the cosine similarity attention, the absolute position embedding by sine function in the original Transformer model and the proposed spatial attention position embedding, as well as whether adopting auxiliary loss L_{Aux}. Overall, as illustrated in Table 8, we find that the combination of spatial and temporal self-attention, namely STSA, is the best design choice in our proposed STMA framework. Moreover, considering cosine similarity instead of the scaled dot product in the MCA module can increase crop mapping performance by 2% to 74.69% mF1, further illustrating the robust improvement and superiority of the proposed cosine similarity attention in the MCA module to integrate multi-scale spatio-temporal representations compared to the baseline scaled dot product attention. From Table 8, it can be seen that removing position embedding from STMA leads to a clear drop in accuracy. This is mainly because the extracted convolutional features

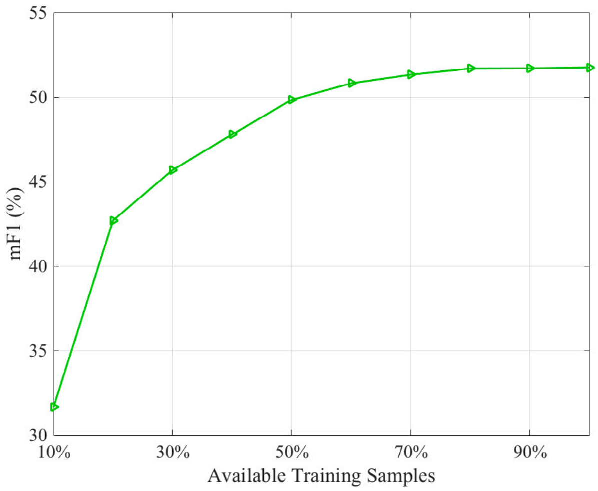


Fig. 10. Spatio-temporal generalization results (mF1) obtained by our proposed STMA with the varying number of available training samples.

Table 8
Ablation analysis of the proposed STMA with a combination of different spatio-temporal interaction, multi-scale feature integration, position embedding and loss function. The best performance of each indicator is shown in **bold**.

Model	OA	mF1	mIoU
PE_{SA} + STSA + MCA(cosine) + L_{CE} + L_{Aux} (ours)	**96.54%**	**74.69%**	**63.81%**
– STSA + spatial self-attention	94.24%	66.68%	52.98%
– STSA + temporal self-attention	95.06%	71.73%	58.22%
– MCA(cosine) + MCA(dot product)	95.13%	72.31%	59.97%
– PE_{SA}	93.14%	64.68%	49.98%
– PE_{SA} + PE_{Abs}	94.57%	68.29%	54.31%
– L_{Aux}	95.48%	71.03%	58.68%

are flatten and then sent directly to the STSA module without marking the spatial position, further resulting in a certain spatial information loss. The introduction of absolute position embedding and spatial attention position embedding can solve this dilemma by an approximately 4% and 10% increment in mF1. In addition, the auxiliary semantic loss function is capable of capturing the multi-class crop features and boosts the crop mapping performance by approximately 3.6% in mF1. As a whole, the combination of STSA, MCA based on cosine similarity, spatial attention position embedding and auxiliary loss function can yield the highest precision, further demonstrating the effectiveness of STMA in handling different types of crops in large-area scenes.

5. Discussion

Accurate crop mapping is essential for decision-makers to assess crop yields and maintain food security. Time-series SAR imagery provides repeated and stable observation of large-area crops over space and time, and yet it is challenging to capture subtle crop phenology information effectively from time-series SAR data, even with state-of-the-art DL-based methods. Previous research has tended to be incapable of capturing spatial and temporal features at different scales simultaneously. In this way, some crops in the continuous time-period of the growing season can be differentiated easily, whereas complex phenological characteristics of other crops are hard to capture and distinguish well. In this research, we propose a STMA method to consider multi-scale spatial and temporal relationships within time-series SAR data comprehensively to learn the phenology changes of crops from local to global perspectives. The STMA is fundamentally different from current DL-based crop mapping methods in two aspects, including: (1) the realization of a spatio-temporal network architecture by the multi-level attention to represent hierarchical spatial and temporal relationships of crops, achieving accurate multi-class crop mapping in the large-area landscapes; (2) better spatio-temporal generalization of STMA when applied over different years and spatial regions.

5.1. The performance of STMA for crop mapping

STMA emphasizes its feature learning advantages in both spatial and temporal dimensions to achieve multi-class crop mapping from the multi-scale perspectives, where the precise phenological characteristics of crops can be extracted and represented holistically. Traditional CNN, RNN and self-attention approaches use a limited receptive field in the single scale to extract features from raw time-series SAR data and ignore multi-dimensional spatio-temporal contextual information, since their feature representation relies on the limited perceptual space of convolutional kernels and the order of time-series at temporal scale (Canizo et al., 2019; Crisóstomo de Castro Filho et al., 2020; Lin et al., 2022). The proposed STMA addresses this critical issue by extracting joint multi-scale information from the complete time-series SAR data using the multi-level attention architecture via cascaded self-attention and cross-attention modalities, which can effectively capture precise long-range dependency coupling and remain high-level spatio-temporal information as shown in Fig. 8. The heatmap comparisons of the encoded spatio-temporal feature extracted from TPE-RNN and the proposed STMA in Fig. 11 illustrate superior long-range spatio-temporal feature extraction ability of STMA, which is beneficial for multi-category crop mapping and mixed cropping systems. This new method provides a novel perspective for separating phenological characteristics and distinguishing crop categories in the field of crop mapping. In addition, the introduction of STSA further provides reliable and robust spatial and temporal texture support for the subsequent MCA and avoids information loss during model training.

Compared with state-of-the-art crop mapping approaches, Figs. 5, 6 and 7 and Table 3 demonstrate the superiority and effectiveness of STMA in extracting short-term and long-term spatio-temporal characteristics in relation to crop growth and phenological status. The experimental results suggest that the information in the input time-series SAR has been fully utilized for STMA to handle spatial and temporal relationships. Unlike the pre-defined function to reflect phenological characteristics, the STMA characterizes multi-view pattern features automatically to improve the crop phenology retrieval for time-series processing. Specifically, the STMA consists of three views to establish robust spatio-temporal representations for different crops: (1) spatial–temporal view to model long-term spatial and temporal correlations by the CNN-based ResNet module, spatial attention position embedding and STSA; (2) multi-scale integration view to aggregate the contributions from different scale levels via MCA to generate feature representations with rich multi-scale information; (3) semantic view to guarantee the context and detailed information to the largest extent by learning image and feature dimensions jointly. This multi-view mechanism in STMA can realize the automatic extraction of reliable spatio-temporal features and reduce manual intervention on the basis of ensuring accurate crop extraction. Overall, the proposed STMA method can achieve better perception to phenological characteristics of crops in time-series SAR and is a feasible solution for practical applications in crop mapping.

5.2. Spatio-temporal generalization analysis

To investigate the spatio-temporal generalization performance of the proposed STMA, we assess the prediction crop mapping accuracy of the model trained from the S1 study area in the S2 study area as shown in Table 7. The self-attention-based approaches can guarantee more spatio-temporal characteristics of different crops owing to their long-distance information capabilities. Other than U-TAE and TPE-RNN, with the help of spatial attention position encoding, the introduction of the multi-level attention mechanism helps STMA to enhance the

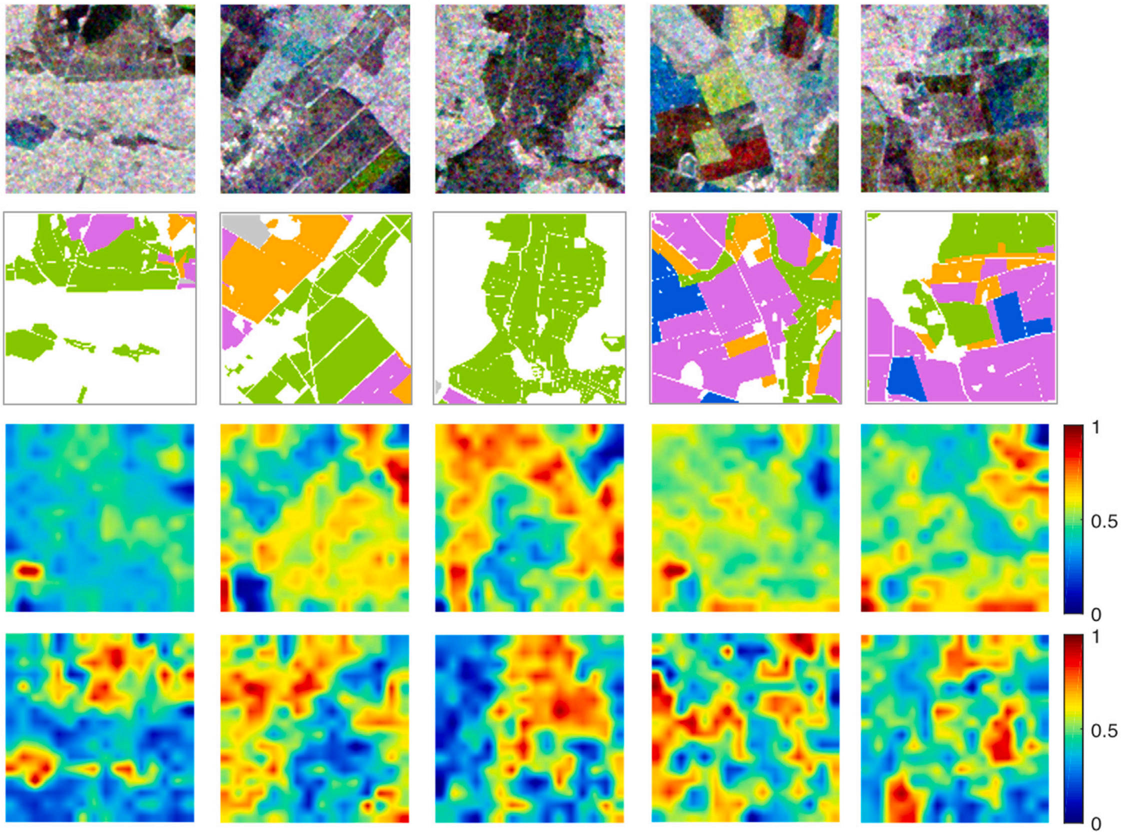


Fig. 11. Heatmap comparisons of the encoded spatio-temporal features extracted from TPE-RNN and the proposed STMA. From top to bottom, each row represents the SAR image, ground reference, TPE-RNN and STMA.

generalization ability on various years and spatial regions by considering different receptive fields to extract spatio-temporal features from the spatial and temporal dimension. Specifically, the depth-wise convolution operation in the proposed positional encoding can gather the spatial information about local neighborhood of the input token and makes the generated token embedding permutation-variant but translation-invariant, which helps the entire crop mapping network easily generalize to learn complex position relationships. Furthermore, the spatial self-attention module is adopted to extract the two-dimensional spatial correlation information crops and improve the ability to express the spatial characteristics of crops at different scales. Meanwhile, the temporal self-attention module aims at acquiring the temporal characteristics of crops at different times and capturing the dynamic changes of crop phenology. The effective integration of these two modules helps the spatio-temporal variability of crops be captured by the proposed STMA and achieve better crop mapping performance. This observation is in accordance with (Hao et al., 2019), who found that the positional encoding and self-attention techniques in the original Transformer can jointly encoder the order information to keep the sequential characteristics intact for learning. Furthermore, the multi-scale feature learning of the DL model can facilitate the generalization performance improvements (Olimov et al., 2023). In some cases, traditional machine learning approaches are limited by their auto-regressive nature during model training, such that they focus only on remembering past observations rather than generalizing training samples to new data (Katharopoulos et al., 2020). On the contrary, the combination of the proposed STSA and MCA module in the STMA forces the entire model to learn universal spatio-temporal features with rich multi-scale information that are conducive to the final crop mapping, so that the generalization performance of STMA has significant advantages in comparison with state-of-the-art benchmarks.

5.3. Future research

The proposed STMA method provides a robust and accurate strategy to achieve crop mapping using time-series SAR imagery. Although we tested a few spatio-temporal extensions on the Brandenburg Sentinel-1 dataset, the PASTIS-R dataset and the South Africa dataset, the STMA still needs to be tested on data scarce areas or more large-scale scenarios, such as the global South. Limited by cumbersome category labeling for multiple crop categories in different regions, it is difficult to implement at present. Due to the unique imaging principle of SAR, the scattering characteristics of crops in time-series SAR data could be affected by observation condition, crop irrigation and crop planting time, which also leads to poor temporal consistency and brings challenges to the application of spatio-temporal generalization. Therefore, a range of transfer learning or cross-scene techniques will be adopted in the future to integrate useful information from data-rich regions and enhance the spatio-temporal generalization performance over different scenarios, to enhance crop mapping prediction under the condition of limited RS data resources. In addition, existing crop mapping approaches simply focus on fine-grained crop identification, without considering actual application demand, such as land use and land management. Future research will construct reliable spatio-temporal relationships in different application scenarios, so that subtle and useful information in the spatial and temporal domains can be captured to enhance feature extraction capability.

6. Conclusion

This paper proposes a novel spatio-temporal multi-level attention method, named as STMA, to achieve crop mapping using time-series SAR imagery. Unlike traditional DL-based crop mapping approaches that only consider limited spatio-temporal receptive field, STMA can aggregate comprehensive and multi-scale spatio-temporal features for time-series SAR data with the support of the multi-level attention mechanism. Furthermore, we develop a learnable spatial attention position encoding to adaptively generate the position priors to facilitate the extraction of multi-granularity features. The experimental results show that the crop mapping produced by STMA achieved the highest accuracy on Brandenburg Sentinel-1 dataset, public PASTIS-R dataset and South Africa dataset, much higher than the other benchmark crop mapping approaches. Meanwhile, the STMA method exhibits an excellent generalization ability on different spatio-temporal scenarios, and provides vast potential for cross-scene application.

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 Natural Environment Research Council [grant number NE/T004002/1]. The research was partially funded by the National Natural Science Foundation of China under Grant 42325104, and also sponsored by the China Scholarship Council (Grant No. 202204910382).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.isprsjprs.2023.11.016.

References

Adrian, J., Sagan, V., Maimaitijiang, M., 2021. Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine. ISPRS J. Photogramm. Remote Sens. 175, 215–235.

Alexandratos, N., Bruinsma, J., 2012. World Agriculture Towards 2030/2050: The 2012 Revision. ESA Working Paper No. 12-03, Food and Agriculture Organization, Rome.

Amazirh, A., Merlin, O., Er-Raki, S., Gao, Q., Rivalland, V., Malbeteau, Y., Khabba, S., Escorihuela, M.J., 2018. Retrieving surface soil moisture at high spatio-temporal resolution from a synergy between Sentinel-1 radar and Landsat thermal data: A study case over bare soil. Remote Sens. Environ. 211, 321–337.

Bargiel, D., 2017. A new method for crop classification combining time series of radar images and crop phenology information. Remote Sens. Environ. 198, 369–383.

Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., Li, Z., 2018. A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach. Remote Sens. Environ. 210, 35–47.

Canizo, M., Triguero, I., Conde, A., Onieva, E., 2019. Multi-head CNN–RNN for multi-time series anomaly detection: An industrial case study. Neurocomput. 363, 246–260.

Crisóstomo de Castro Filho, H., Abílio de Carvalho Júnior, O., Ferreira de Carvalho, O.L., Pozzobon de Bem, P., dos Santos de Moura, R., Olino de Albuquerque, A., Rosa Silva, C., Guimaraes Ferreira, P.H., Fontes Guimarães, R., Trancoso Gomes, R.A., 2020. Rice crop detection using LSTM, bi-LSTM, and machine learning models from Sentinel-1 time series. Remote Sens. 12 (16), 2655.

Chang, Y.L., Tan, T.H., Chen, T.H., Chuah, J.H., Chang, L., Wu, M.C., Tatini, N.B., Ma, S.C., Alkhaleefah, M., 2022. Spatial-temporal neural network for rice field classification from SAR images. Remote Sens. 14 (8), 1929.

De Bernardis, C.G., Vicente-Guijalba, F., Martinez-Marin, T., Lopez-Sanchez, J.M., 2014. Estimation of key dates and stages in rice crops using dual-polarization SAR time series and a particle filtering approach. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 8 (3), 1008–1018.

Diao, C., 2019. Innovative pheno-network model in estimating crop phenological stages with satellite time series. ISPRS J. Photogramm. Remote Sens. 153, 96–109.

Dong, J., Xiao, X., Menarguez, M.A., Zhang, G., Qin, Y., Thau, D., Biradar, C., Moore III, B., 2016. Mapping paddy rice planting area in northeastern Asia with landsat 8 images, phenology-based algorithm and Google Earth Engine. Remote Sens. Environ. 185, 142–154.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Drusch, M., Del Bello, U., Carlier, S., Colin, O., Fernandez, V., Gascon, F., Hoersch, B., Isola, C., Laberinti, P., Martimort, P., et al., 2012. Sentinel-2: ESA's optical high-resolution mission for GMES operational services. Remote Sens. Environ. 120, 25–36.

Dumoulin, V., Visin, F., 2016. A guide to convolution arithmetic for deep learning. arXiv preprint arXiv:1603.07285.

Franceschi, J.-Y., Dieuleveut, A., Jaggi, M., 2019. Unsupervised scalable representation learning for multivariate time series. Adv. Neural Inf. Process. Syst. 32.

Gao, H., Wang, C., Wang, G., Fu, H., Zhu, J., 2021. A novel crop classification method based on ppfSVM classifier with time-series alignment kernel from dual-polarization SAR datasets. Remote Sens. Environ. 264, 112628.

Garnot, V.S.F., Landrieu, L., 2020. Lightweight temporal self-attention for classifying satellite images time series. In: Advanced Analytics and Learning on Temporal Data: 5th ECML PKDD Workshop, AALTD 2020, Ghent, Belgium, September 18, 2020, Revised Selected Papers 6. Springer, pp. 171–181.

Garnot, V.S.F., Landrieu, L., 2021a. Panoptic segmentation of satellite image time series with convolutional temporal attention networks. In: Proc. IEEE/CVF Intern. Conf. Comput. Vision. pp. 4872–4881.

Garnot, V.S.F., Landrieu, L., 2021b. Pastis-r - panopiic segmentation of radar and optical satellite image time series. http://dx.doi.org/10.5281/ZENODO.5735646, URL https://zenodo.org/record/5735646.

Garnot, V.S.F., Landrieu, L., Chehata, N., 2022. Multi-modal temporal attention models for crop mapping from satellite time series. ISPRS J. Photogramm. Remote Sens. 187, 294–305.

Garnot, V.S.F., Landrieu, L., Giordano, S., Chehata, N., 2020. Satellite image time series classification with pixel-set encoders and temporal self-attention. In: Proc. Comput. Vision Pattern Recognit.. pp. 12325–12334.

Ge, S., Zhang, J., Pan, Y., Yang, Z., Zhu, S., 2021. Transferable deep learning model based on the phenological matching principle for mapping crop extent. Int. J. Appl. Earth Obs. Geoinf. 102, 102451.

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., et al., 2018. Recent advances in convolutional neural networks. Pattern Recognit. 77, 354–377.

Guo, Z., Qi, W., Huang, Y., Zhao, J., Yang, H., Koo, V.-C., Li, N., 2022a. Identification of crop type based on C-AENN using time series Sentinel-1A SAR data. Remote Sens. 14 (6), 1379.

Guo, W., Zhang, W., Zhang, Z., Tang, P., Gao, S., 2022b. Deep temporal iterative clustering for satellite image time series land cover analysis. Remote Sens. 14 (15), 3635.

Han, Z., Hong, D., Gao, L., Yao, J., Zhang, B., Chanussot, J., 2022a. Multimodal hyperspectral unmixing: Insights from attention networks. IEEE Trans. Geosci. Remote Sens. 60, 1–13.

Han, Z., Hong, D., Gao, L., Zhang, B., Chanussot, J., 2020. Deep half-siamese networks for hyperspectral unmixing. IEEE Geosci. Remote Sens. Lett. 18 (11), 1996–2000.

Han, Z., Hong, D., Gao, L., Zhang, B., Huang, M., Chanussot, J., 2022b. AutoNAS: Automatic neural architecture search for hyperspectral unmixing. IEEE Trans. Geosci. Remote Sens. 60, 1–14.

Hao, J., Wang, X., Yang, B., Wang, L., Zhang, J., Tu, Z., 2019. Modeling recurrence for transformer. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers).

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Proc. Comput. Vision Pattern Recognit.. pp. 770–778.

Huang, X., Wang, J., Shang, J., Liao, C., Liu, J., 2017. Application of polarization signature to land cover scattering mechanism analysis and classification using multi-temporal C-band polarimetric RADARSAT-2 imagery. Remote Sens. Environ. 193, 11–28.

Inglada, J., Arias, M., Tardy, B., Hagolle, O., Valero, S., Morin, D., Dedieu, G., Sepulcre, G., Bontemps, S., Defourny, P., et al., 2015. Assessment of an operational system for crop type map production using high temporal and spatial resolution satellite optical imagery. Remote Sens. 7 (9), 12356–12379.

Iounousse, J., Er-Raki, S., El Motassadeq, A., Chehouani, H., 2015. Using an unsupervised approach of probabilistic neural network (PNN) for land use classification from multitemporal satellite images. Appl. Soft Comput. 30, 1–13.

Jin, Y., Liu, X., Chen, Y., Liang, X., 2018. Land-cover mapping using random forest classification and incorporating NDVI time-series and texture: A case study of central Shandong. Int. J. Remote Sens. 39 (23), 8703–8723.

Julien, Y., Sobrino, J.A., 2010. Comparison of cloud-reconstruction methods for time series of composite NDVI data. Remote Sens. Environ. 114 (3), 618–625.

Kalinicheva, E., Sublime, J., Trocan, M., 2020. Unsupervised satellite image time series clustering using object-based approaches and 3d convolutional autoencoder. Remote Sens. 12 (11), 1816.

Karim, F., Majumdar, S., Darabi, H., Harford, S., 2019. Multivariate LSTM-FCNs for time series classification. Neural Netw. 116, 237–245.

Karthikeyan, L., Chawla, I., Mishra, A.K., 2020. A review of remote sensing applications in agriculture for food security: Crop growth and yield, irrigation, and crop losses. J. Hydrol. 586, 124905.

Katharopoulos, A., Vyas, A., Pappas, N., Fleuret, F., 2020. Transformers are rnns: Fast autoregressive transformers with linear attention. In: Int. Conf. Mach. Learn.. PMLR, pp. 5156–5165.

Kazemnejad, A., 2019. Transformer architecture: The positional encoding. kazemnejad.com. URL https://kazemnejad.com/blog/transformer_architecture_positional_encoding/.

Kenduiywo, B.K., Bargiel, D., Soergel, U., 2017. Higher order dynamic conditional random fields ensemble for crop type classification in radar images. IEEE Trans. Geosci. Remote Sens. 55 (8), 4638–4654.

Kenduiywoa, B., Bargiel, D., Soergel, U., 2015. Spatial-temporal conditional random fields crop classification from Terrasar-X images. ISPRS Annals Photogramm. Remote Sens. Spat. Inf. Sci. 2, 79–86.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. Adv. Neural Inf. Proces. Syst. 2, 1097–1105.

Kwak, G.H., Park, N.W., 2022. Unsupervised domain adaptation with adversarial self-training for crop classification using remote sensing images. Remote Sens. 14 (18), 4639.

Leite, P.B.C., Feitosa, R.Q., Formaggio, A.R., da Costa, G.A.O.P., Pakzad, K., Sanches, I.D., 2011. Hidden Markov models for crop recognition in remote sensing image sequences. Pattern Recognit. Lett. 32 (1), 19–26.

Li, H., Zhang, C., Zhang, Y., Zhang, S., Ding, X., Atkinson, P.M., 2021. A scale sequence object-based convolutional neural network (SS-OCNN) for crop classification from fine spatial resolution remotely sensed imagery. Int. J. Digit. Earth 14 (11), 1528–1546.

Lin, Z., Zhong, R., Xiong, X., Guo, C., Xu, J., Zhu, Y., Xu, J., Ying, Y., Ting, K., Huang, J., et al., 2022. Large-scale rice mapping using multi-task spatiotemporal deep learning and sentinel-1 SAR time series. Remote Sens. 14 (3), 699.

Liu, X., Yu, H.F., Dhillon, I., Hsieh, C.J., 2020. Learning to encode position for transformer with continuous dynamical model. In: Int. Conf. Mach. Learn.. PMLR, pp. 6327–6335.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. Comput. Vision Pattern Recognit.. pp. 3431–3440.

Luo, C., Zhan, J., Xue, X., Wang, L., Ren, R., Yang, Q., 2018. Cosine normalization: Using cosine similarity instead of dot product in neural networks. In: Artificial Neural Networks and Machine Learning–ICANN 2018: 27th International Conference on Artificial Neural Networks, Rhodes, Greece, October 4-7, 2018, Proceedings, Part I 27. Springer, pp. 382–391.

Van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Mach. Learn. Res. 9 (11).

Magistri, F., Weyler, J., Gogoll, D., Lottes, P., Behley, J., Petrinic, N., Stachniss, C., 2023. From one field to another—Unsupervised domain adaptation for semantic segmentation in agricultural robotics. Comput. Electron. Agric. 212, 108114.

Mandal, D., Kumar, V., Ratha, D., Dey, S., Bhattacharya, A., Lopez-Sanchez, J.M., McNairn, H., Rao, Y.S., 2020. Dual polarimetric radar vegetation index for crop growth monitoring using sentinel-1 SAR data. Remote Sens. Environ. 247, 111954.

McNairn, H., Shang, J., Jiao, X., Champagne, C., 2009. The contribution of ALOS PALSAR multipolarization and polarimetric data to crop classification. IEEE Trans. Geosci. Remote Sens. 47 (12), 3981–3992.

Nyborg, J., Pelletier, C., Assent, I., 2022. Generalized classification of satellite image time series with thermal positional encoding. In: Proc. Comput. Vision Pattern Recognit.. pp. 1392–1402.

Olimov, B., Subramanian, B., Ugli, R.A.A., Kim, J.-S., Kim, J., 2023. Consecutive multiscale feature learning-based image classification model. Sci. Rep. 13 (1), 3595.

Oyoshi, K., Tomiyama, N., Okumura, T., Sobue, S., Sato, J., 2016. Mapping rice-planted areas using time-series synthetic aperture radar data for the Asia-RiCE activity. Paddy Water Environ. 14, 463–472.

Peng, Z., Liu, W., An, S., 2020. Haze pollution causality mining and prediction based on multi-dimensional time series with PS-FCM. Information Sci. 523, 307–317.

Periasamy, S., 2018. Significance of dual polarimetric synthetic aperture radar in biomass retrieval: An attempt on sentinel-1. Remote Sens. Environ. 217, 537–549.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. In: Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervention. Springer, pp. 234–241.

Rußwurm, M., Körner, M., 2018. Multi-temporal land cover classification with sequential recurrent encoders. ISPRS Int. J. Geo Inf. 7 (4), 129.

Rußwurm, M., Körner, M., 2020. Self-attention for raw optical satellite time series classification. ISPRS J. Photogramm. Remote Sens. 169, 421–435.

Satalino, G., Balenzano, A., Mattia, F., Davidson, M.W., 2013. C-band SAR data for mapping crops dominated by surface or volume scattering. IEEE Geosci. Remote Sens. Lett. 11 (2), 384–388.

Shi, X., Chen, Z., Wang, H., Yeung, D.Y., Wong, W.K., Woo, W.c., 2015. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. Adv. Neural Inf. Proces. Syst. 28.

Sonobe, R., Tani, H., Wang, X., Kobayashi, N., Shimamura, H., 2014. Random forest classification of crop type using multi-temporal TerraSAR-X dual-polarimetric data. Remote Sens. Lett. 5 (2), 157–164.

Sonobe, R., Tani, H., Wang, X., Kobayashi, N., Shimamura, H., 2015. Discrimination of crop types with TerraSAR-X-derived information. Phys. Chem. Earth Parts A/B/C 83, 2–13.

Srikanth, P., Ramana, K., Deepika, U., Kalyan Chakravarthi, P., Sesha Sai, M., 2016. Comparison of various polarimetric decomposition techniques for crop classification. J. Indian Soc. Remote Sens. 44, 635–642.

Tan, C.P., Ewe, H.T., Chuah, H.T., 2011. Agricultural crop-type classification of multi-polarization SAR images using a hybrid entropy decomposition and support vector machine technique. Int. J. Remote Sens. 32 (22), 7057–7071.

Thenkabail, P., Knox, J., Ozdogan, M., Gumma, M., Congalton, R., Wu, Z., Milesi, C., Finkral, A., Marshall, M., Mariotto, I., et al., 2012. Assessing future risks to agricultural productivity. Water Resourc. Food Secur.: How Can Remote Sens. Help 78, 773–782.

Tilman, D., Balzer, C., Hill, J., Befort, B.L., 2011. Global food demand and the sustainable intensification of agriculture. Proc. Natl. Acad. Sci. 108 (50), 20260–20264.

Trudel, M., Charbonneau, F., Leconte, R., 2012. Using RADARSAT-2 polarimetric and ENVISAT-ASAR dual-polarization data for estimating soil moisture over agricultural fields. Can. J. Remote Sens. 38 (4), 514–527.

Van Bussel, L., Ewert, F., Leffelaar, P., 2011. Effects of data aggregation on simulations of crop phenology. Agric. Ecosyst. Environ. 142 (1–2), 75–84.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Adv. Neural Inf. Proces. Syst. 30.

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.

Vicente-Guijalba, F., Martinez-Marin, T., Lopez-Sanchez, J.M., 2013. Crop phenology estimation using a multitemporal model and a Kalman filtering strategy. IEEE Geosci. Remote Sens. Lett. 11 (6), 1081–1085.

Waha, K., Dietrich, J.P., Portmann, F.T., Siebert, S., Thornton, P.K., Bondeau, A., Herrero, M., 2020. Multiple cropping systems of the world and the potential for increasing cropping intensity. Global Environ. Change 64, 102131.

Wang, S., Azzari, G., Lobell, D.B., 2019. Crop type mapping without field-level labels: Random forest transfer and unsupervised clustering techniques. Remote Sens. Environ. 222, 303–317.

Wang, Y., Feng, L., Zhang, Z., Tian, F., 2023. An unsupervised domain adaptation deep learning method for spatial and temporal transferable crop type mapping using Sentinel-2 imagery. ISPRS J. Photogramm. Remote Sens. 199, 102–117.

Wang, Y., Gao, L., Hong, D., Sha, J., Liu, L., Zhang, B., Rong, X., Zhang, Y., 2021. Mask DeepLab: End-to-end image segmentation for change detection in high-resolution remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 104, 102582.

Wang, Z., Zhang, H., He, W., Zhang, L., 2022. Cross-phenological-region crop mapping framework using Sentinel-2 time series imagery: A new perspective for winter crops in China. ISPRS J. Photogramm. Remote Sens. 193, 200–215.

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 (3), 1096–1116.

Waske, B., Braun, M., 2009. Classifier ensembles for land cover mapping using multitemporal SAR imagery. ISPRS J. Photogramm. Remote Sens. 64 (5), 450–457.

Weilandt, F., Behling, R., Goncalves, R., Madadi, A., Richter, L., Sanona, T., Spengler, D., Welsch, J., 2023. Early crop classification via multi-modal satellite data fusion and temporal attention. Remote Sens. 15 (3), 799.

Western Cape Department of Agriculture, 2021. Crop type classification dataset for Western Cape, South Africa. Version 1.0, Radiant MLHub.

Wolff, S., Hüttel, S., Nendel, C., Lakes, T., 2020. Identifying agricultural landscape types for brandenburg, Germany using IACS data. Technical Report, FORLand-Working Paper.

Xu, J., Yang, J., Xiong, X., Li, H., Huang, J., Ting, K., Ying, Y., Lin, T., 2021. Towards interpreting multi-temporal deep learning models in crop mapping. Remote Sens. Environ. 264, 112599.

Xu, L., Zhang, H., Wang, C., Zhang, B., Liu, M., 2018. Crop classification based on temporal information using sentinel-1 SAR time-series data. Remote Sens. 11 (1), 53.

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 Sens. Environ. 247, 111946.

Yang, L., Huang, R., Huang, J., Lin, T., Wang, L., Mijiti, R., Wei, P., Tang, C., Shao, J., Li, Q., et al., 2021. Semantic segmentation based on temporal features: Learning of temporal–spatial information from time-series SAR images for paddy rice mapping. IEEE Trans. Geosci. Remote Sens. 60, 1–16.

Zaremba, W., Sutskever, I., Vinyals, O., 2014. Recurrent neural network regularization. arXiv preprint arXiv:1409.2329.

Zeng, Z., Sun, J., Han, Z., Hong, W., 2022. SAR automatic target recognition method based on multi-stream complex-valued networks. IEEE Trans. Geosci. Remote Sens. 60, 1–18.

Zeng, Z., Sun, J., Xu, C., Wang, H., 2021. Unknown SAR target identification method based on feature extraction network and KLD–RPA joint discrimination. Remote Sens. 13 (15), 2901.

Zhang, L., Sun, L., Zou, B., Moon, W.M., 2014. Fully polarimetric SAR image classification via sparse representation and polarimetric features. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens. 8 (8), 3923–3932.

Zhao, W., Qu, Y., Chen, J., Yuan, Z., 2020. Deeply synergistic optical and SAR time series for crop dynamic monitoring. Remote Sens. Environ. 247, 111952.

Zhong, L., Hu, L., Zhou, H., 2019. Deep learning based multi-temporal crop classification. Remote Sens. Environ. 221, 430–443.

Zhu, X.X., Tuia, D., Mou, L., Xia, G.S., Zhang, L., Xu, F., Fraundorfer, F., 2017. Deep learning in remote sensing: A comprehensive review and list of resources. IEEE Geosci. Remote Sens. Mag. 5 (4), 8–36.