Phenology Alignment Network: A Novel Framework for Cross-Regional Time Series Crop Classification

Ziqiao Wang[1], ✉Hongyan Zhang[1], ✉Wei He[2], Liangpei Zhang[1]
[1]State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing,
Wuhan University, Wuhan 430079, China
[2]RIKEN Center for Advanced Intelligence Project (AIP) Tokyo, 103-0027, Japan
{wangziqiao, zhanghongyan}@whu.edu.cn, wei.he@riken.jp, zlp62@whu.edu.cn

Abstract

Timely and accurate crop type classification plays an essential role in the study of agricultural application. However, large area or cross-regional crop classification confronts huge challenges owing to dramatic phenology discrepancy among training and test regions. In this work, we propose a novel framework to address these challenges based on deep recurrent network and unsupervised domain adaptation (DA). Specifically, we firstly propose a Temporal Spatial Network (TSNet) for pixelwise crop classification, which contains stacked RNN and self-attention module to adaptively extract multi-level features from crop samples under various planting conditions. To deal with the cross-regional challenge, an unsupervised DA-based framework named Phenology Alignment Network (PAN) is proposed. PAN consists of two branches of two identical TSNet pre-trained on source domain; one branch takes source samples while the other takes target samples as input. Through aligning the hierarchical deep features extracted from two branches, the discrepancy between two regions is decreased and the pre-trained model is adapted to the target domain without using target label information. As another contribution, a time series dataset based on Sentinel-2 was annotated containing winter crop samples collected on three study sites of China. Cross-regional experiments demonstrate that TSNet shows comparable accuracy to state-of-the-art methods, and PAN further improves the overall accuracy by 5.62%, and macro average F1 score by 0.094 unsupervisedly.

1. Introduction

Time series crop classification aims to depict the type and distribution of crops over the research area accurately, which is fundamental to agricultural resources allocation and policy decision [12]. With the advent and development of machine learning algorithms, multiple temporal and spectral information can be extracted automatically from satellite imagery time series (SITS) and the classification maps are produced without labor-intensive ground survey work [5, 16, 21]. However, current researches are highly localized and confront huge challenges including invalid features and model failure problem in cross-regional crop classification.

There are mainly two reasons resulting in the model failure of cross-regional mapping. Firstly, the phenology characteristic of a same crop type considerably differs between regions owing to different climate conditions and plant patterns, referred as phenology discrepancy phenomenon. Specifically, the spectral appearance and phenology stages of a same crop change, making hand-made features and the pre-trained classifier invalid. Figure 1 illustrates this phenomenon explicitly with the visual maps and the phenological curves obtained from three study areas. Secondly, the temporal discriminability of different region's samples is inconsistent. Typically, uniform image acquisition over large area is impossible due to inevitable partial cloud, breeding noise and missing information in time sereis imageries, which aggravates the heterogeneity of crop samples from different regions. The prior knowledge learnt from one place, such as designed features and pre-trained models, cannot reflect the distribution of samples in a new region, causing enormous reduction of model's performance. In summary, it is necessary to develop a cross-regional crop classification framework to extract generalized phenological features from crop samples and adapt knowledge to new regions.

In this work, we propose a generalized model, named Temporal Spatial Network (TSNet) for time series crop classification. To further tackle the cross-domain challenges, we propose a framework named Phenology Alignment Network (PAN) on the basis of TSNet. Specifically, the TSNet is formed by stacked Gated Recurrent

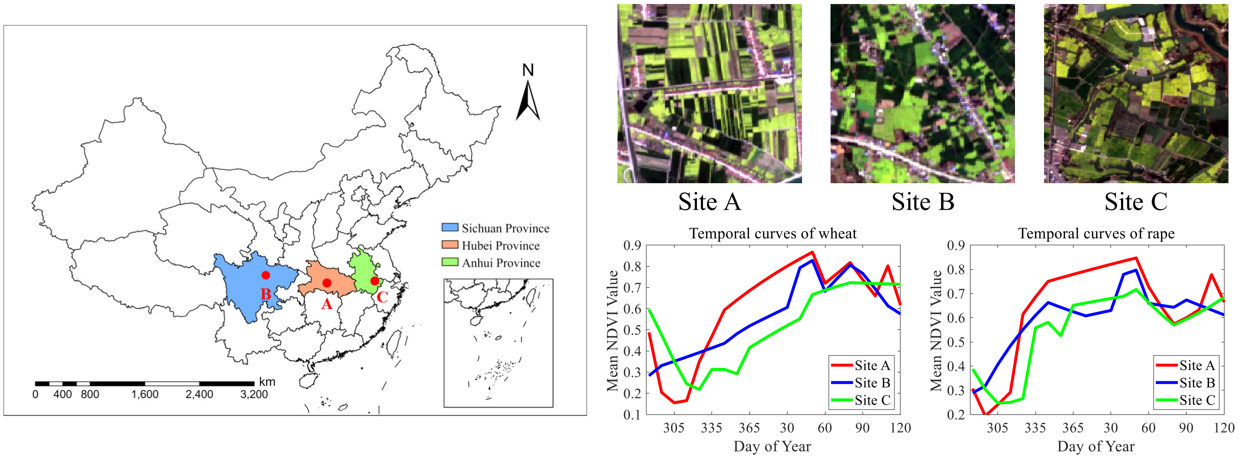


Figure 1. The locations of our study sites are denoted by red dots in the left image and huge cross-regional phenological difference can be observed. Clipped blocks of satellite images of three sites at the same date are listed on top right, where spectral differences exist in winter crop samples. Furthermore, the mean temporal NDVI curves of winter wheat and rape in three regions are also compared on lower right, showing that cross-regional phenology discrepancy exists in the same crop type.

Unit (GRU) layers combined with self-attention mechanism. The deep recurrent network is to excavate the multi-level features and temporal dependency from SITS which is often ignored in machine learning methods and the attention module is to handle heterogeneous crop samples under various environmental conditions adaptively. The PAN framework aims to improve the accuracy of the pre-trained model on the target region without additional labels. PAN consists of two branches of pre-trained TSNet with shared parameters for the source and target samples respectively as input. The hierarchical phenological features of two domains are aligned by Maximum Mean Discrepancy (MMD) loss [2] to minimize the distance of feature distribution. By doing this, the TSNet of target domain is fine-tuned by PAN and can map the target domain's data into the same feature space as that of the source domain. In summary, the proposed PAN can mitigate the regional discrepancy and improve the cross-domain classification accuracy.

It is worth noting that the proposed PAN employs a total unsupervised strategy, merely making source and target crop samples *similar* in the feature space without using any annotation information of the target domain. And we hope to offer a novel framework for large-area, national or even global agricultural applications.

Contributions of this work are three-fold:

- We propose a generalized deep model named TSNet for time series crop classification. Stacked GRU layers are utilized to extract robust phenological features automatically from SITS and self-attention mechanism is used to handle the heterogeneous input samples in a data-driven way.
- We propose an unsupervised DA-based framework named PAN to address the cross-regional crop classification challenges. MMD loss is used to decrease the discrepancy of deep features between two domains, adapting invalid models to new regions.
- Additionally, we annotated a time series crop dataset based on Sentinel-2 images, which contains winter crop samples with high intra-class variance from three geo-scattered study sites of China. The cross-regional experiments on this dataset verify the effectiveness of our TSNet and PAN to tackle the challenges aforementioned.

The rest of paper is organized as follows. In Section 2, we briefly review the related works. The detail of our annotated dataset is introduced in Section 3. In Section 4, TSNet and PAN framework are introduced systematically, and the experiments are presented in Section 5. Finally, we conclude our work in Section 6.

2. Related Works

In this section, recent crop type classification researches based on time series multi-spectral images are reviewed firstly. Subsequently, related works about cross-regional or large area classification are also summarized.

2.1. Time series crop type classification

Researches based on low and medium spatial resolution satellites (from 10m to 100m) are inclined to use pixel-based or patch-based methods instead of fully convolution network [11, 15, 37]. The reason is that the convolution kernel of large size is beyond the scope of a single field, leading to mixed pixel noise and information confusion. The lack in spatial information is compensated with temporal information [4, 8]. [10] concluded that using multi-temporal satellite images helps to achieve higher classification accuracy compared to using merely single temporal image. Hence, numerous researches aimed to exploit the rich information in SITS to map crops.

Over recent years, many traditional machine learning (ML) algorithms have been adopted by remote sensing community for the analysis and classification of SITS. Handmade features were designed and fed into ML classifier such as decision tree [6, 23], support vector machine [19] and random forest [7]. However, these methods are not designed for time series analysis task, and only take the temporal features as separate inputs without analyzing the sequential relationship between different time stamps. Thus, these methods are not robust enough for cross-regional classification. Besides, feature engineering turns out to be a challenging task and heavily relies on expertise; designed features are not representative enough when applying in large area.

Currently, with the success in all walks of research fields, deep learning (DL) also aroused great interest in remote sensing community [33, 36]. In view of its capability to extract high level structural information, the feature engineering work can be performed in a data-driven way. DL models used in researches of time series crop classification include two main architectures: convolutional neural network (CNN) [17] and recurrent neural network (RNN) [14]. [31] used modified pyramid scene parsing network (PSPNet) to conduct the land cover classification on Gaofen series satellite images. Yet under most cases, spatial convolutions are unsuitable as explained before. TempCNN model [22] applied convolutions in both temporal and spectral domains to take full advantage of temporal structure of SITS. As another common structure, RNN is specialized for comprehending sequential input data, thereby adopted more widely. [34] explored Long Short-Term Memory (LSTM)'s viability for identifying time series phenology curves derived from Landsat satellite. Deep Crop Mapping (DCM) model [30] added self-attention structure to LSTM for the in-season classification in U.S. corn belt. Besides, [20, 24] combined CNN and RNN to excavate spectral and temporal information simultaneously for classification and change detection task. Furthermore, with the boom of Transformer [28] in NLP, some up-to-date works employed such stacked multi-head self-attention model to time series crop mapping task [25, 26].

2.2. Large area or cross-regional classification

In a small area, the growth status and life cycle of a certain crop type will not change intensely, offering adequate prior to infer most samples. However, outside the labelled region, huge phenological difference exists in a same crop type owing to different soil conditions and accumulated temperatures, making cross-regional classification a challenging task. The existing works tackle the cross-regional challenge from three perspectives:

Works from perspective of model aim to modify baseline models in order to extract domain-insensitive and generalized features. [31] modified traditional PSPNet for cropland extraction under various landscape. [30] adapted LSTM to better integrate spectral and temporal information for large area dynamic crops mapping. However, when enormous discrepancy occurs between two regions, this strategy may fail given that the discrepancy problem has not been solved intrinsically.

Works from perspective of samples aim to finetune the pre-trained model with a few high-quality samples in target domain, so that the new distribution can be learnt by the original model. In [27], pseudo-labels with high confidence were used to finetune deep models for country-wide land cover classification. In [13], new samples from target domains were annotated to adapt RF classifiers by active learning. However, this approach involved in labeling a finite number of samples additionally, which is often impractical in large area researches. Besides, ample annotations are demanded to finetune deep networks, consuming considerable labor and time.

Works from perspective of features aim to map different regions' samples into the same feature subspace to reduce the gap between deep features. [18] introduced a domain adaptation method to improve the overall accuracy of cross-domain hyperspectral image classification. [1] and [35] combined DA and adversarial learning for cross-regional land cover classification, and both obtained accuracy improvement. To the best of our knowledge, we are the first to apply unsupervised DA technique to cross-regional time series crop classification.

3. Research Area

In this section, we prepare the time series Sentinel-2 imagery for the experiment. In Section 3.1, we describe the study areas with various environmental conditions, and the image acquisition and annotation procedure. The preprocessing operation on the acquired image is introduced in Section 3.2. Finally, the detailed information of our dataset is displayed in Section 3.3.

3.1. Data collection

Our study sites locate in three geo-scattered plains of China which are displayed in the left image of Figure 1. We collected Sentinel-2 imagery time series over these sites and created three subsets for each site. It is worth noting that the three sites display different environmental conditions and planting patterns, and crop samples in three sites also exhibit different phenological characteristics.

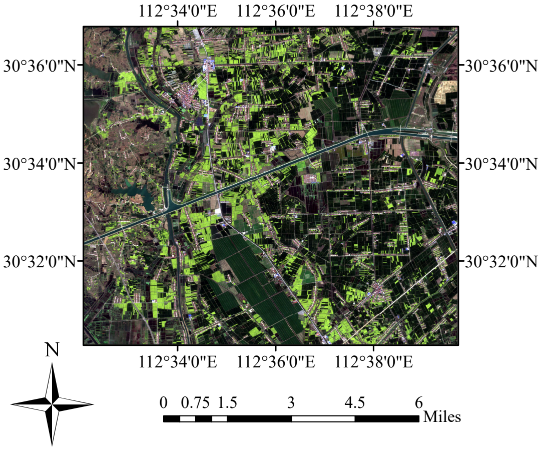


Figure 2. R-G-B composite of our study site A based on single date Sentinel-2 image.

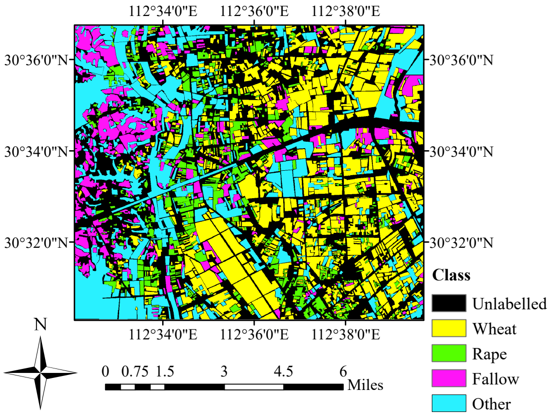


Figure 3. Corresponding annotation map of study site A. Four classes, winter wheat, winter rape, fallow and other were labelled with the help of wide ground survey and high-resolution remote sensing imagery.

- Site A locates in Jianghan Plain with the elevation of ∼30m. Land parcels have averaging bigger size and are more regular.
- Site B locates in Chengdu Plain with the elevation of ∼600m. Land parcels have averaging smaller size and are irregular; dataset has imbalance classes.
- Site C locates in Wuhu Plain with the elevation of ∼100m. The size of land parcels is moderate; crops samples have sparse distribution.

A rectangle of approximate 1200×1200 pixels was selected as study area at each site to ensure sufficient training samples for deep networks. We downloaded Sentinel-2 L2A products from Google Earth Engine directly to save the repeated atmospheric correction work. Clear images (cloud coverage $\leq 20\%$) between October 2019 to May 2020 were collected and cropped to the scope of our predetermined study area. We kept ten spectral bands contributing to depicting planting crops the most, which are RGB bands, four Red Edge bands, NIR band and two SWIR bands.

In each subset, four classes (winter wheat, winter rape, fallow and other) are manually annotated with the aid of wide ground surveys in study site A and B. We deliberately avoided the labeling of field boundaries and mixed pixels to ensure the high-quality and purity of samples. The fallow class are labelled for its significance to planting structure analysis. Figure 2 and 3 demonstrate the RGB composite and our annotated map of site A.

3.2. Preprocessing procedure

Owing to partial cloud noise, the temporal stamp of three study areas is mismatched, making the direct reuse of deep model impossible. Thus, in each study site, linear interpolation was applied to unify time series images to a same time interval, 10 days. Minimum composition was also conducted to reduce cloud noise if two images existing in one interval. Finally, three multiple temporal-spectral data cubes of size $(M_i, 21, 10)$ were obtained for the following cross-regional experiments, where M_i is the number of samples of site i. Each data cube has a temporal dimension of 21 and a spectral dimension of 10. All the preprocessing work was performed locally using MATLAB R2019a on a windows operating system.

3.3. Dataset details

Our dataset contains three subsets and has 2.33 million labelled samples in total. High intra-class variance of same crop type exists between subsets owing to the phenological discrepancy. Each subset is further divided into three irrelevant parts by ratio of 70% / 15% / 15% for training, validation and test. In cross-regional experiments, taking task site A $\rightarrow$ site B for example, deep models are firstly trained on training set of site A. Then the pre-trained models are adapted to site B by PAN; the accuracy of adapted models on the whole subset of site B is reported at last. The detailed information of dataset is listed in Table 1.

4. Methodology

Figure 4 illustrates our proposed model TSNet and the overall framework of PAN. TSNet serves as a robust classifier for time series crop classification, taking temporal data as input and predicted labels as output. The accuracy of pre-trained TSNet decreases on the target region due to the phenology discrepancy. To tackle this cross-regional problem, PAN finetunes the pre-trained TSNet to map the source and target domain samples into the same feature space, obtaining a high-performance model on unlabelled target regions.

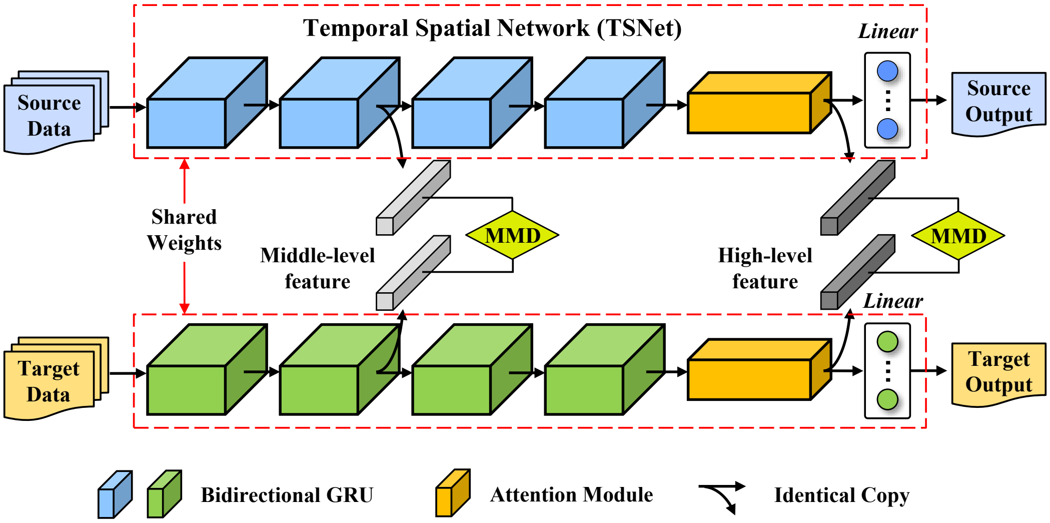


Figure 4. The overall architecture of PAN framework. TSNet is denoted within the red dotted box. Initially, PAN consists of two branches of pre-trained TSNet with same weights. Then the MMD loss is applied to hierarchical features of source and target domain, aligning the feature distribution of samples in two regions. Finetuned by PAN, the source TSNet is adapted to the target domain and can provide reliable predictions.

Table 1. The elaborate information of our dataset which consists of three subsets

Site	Sentinel-2 footprint	Count of clear images	Size of image block	Number of valid pixels	Proportion of four classes (%)			
					wheat	rape	fallow	other
A	T49RFP	14	1400×1200	1049327	39.9	12.2	13.8	34.1
B	T48RVV	18	1200×1200	621402	54.6	4.6	7.7	33.1
C	T50RPV	17	1000×1300	659357	13.6	21.3	17.5	47.6

In Section 4.1, we firstly present our backbone TSNet in detail. Then in Section 4.2, we introduce the transfer procedure conducted by PAN.

4.1. The backbone TSNet

To conduct time series crop classification, TSNet is formed by four bidirectional GRU layers and a self-attention module, illustrated within the red dotted box in Figure 4.

Let $\boldsymbol{X}$ denote the time series input data and $\boldsymbol{Y}$ denote the ground truth, and each sample $\boldsymbol{x}$ can be expressed as a temporal form $[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t]$, where $\boldsymbol{x}_i$ represents input at time i. $\boldsymbol{x}_i$ can be further expanded as $[\boldsymbol{x}_{b1}, \boldsymbol{x}_{b2}, ..., \boldsymbol{x}_{bs}]$, containing multi-spectral bands information from band 1 to band s. The first GRU layer can successively accept input $[\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_t]$, and encode them into hidden states $\boldsymbol{H}_1$, which can be unfolded as $[\boldsymbol{h}_{1,1}, \boldsymbol{h}_{1,2}, ..., \boldsymbol{h}_{1,t}]$. The following layer takes the hidden states $\boldsymbol{H}_1$ as new input and encode them into $\boldsymbol{H}_2$, *i.e.* $[\boldsymbol{h}_{2,1}, \boldsymbol{h}_{2,2}, ..., \boldsymbol{h}_{2,t}]$. Finally, four stacked GRU are capable of yielding deep features containing temporal-spectral information.

In time series crop classification task, the interaction of different time nodes of crop life circle is critical, and the key phenology stage need to be paid more attention. Non-local mechanism [29] was proposed to capture long-range dependency, and widely applied in computer vision filed [9, 32]. The traditional non-local module is modified to fit our 2-D input data. The hidden state output $\boldsymbol{H}_4$ of size (T, d) is transposed as $\boldsymbol{H}_4^T$, where d denotes the dimension of hidden vector. The attention map $\boldsymbol{A}$ of size (T, T) is calculated by the dot product of $\boldsymbol{H}_4$ and $\boldsymbol{H}_4^T$. $\boldsymbol{A}$ is then sent to $softmax$ layer and dot multiplied with $\boldsymbol{H}_4$ to acquire weighted hidden state $\boldsymbol{H}_4^W$. $\boldsymbol{H}_4^W$ is added to original hidden state $\boldsymbol{H}_4$ through a residual structure, as the deep phenology features, $\boldsymbol{H}_A$, generated by TSNet. The features are then flattered and sent to linear layer for the final label predicting. The explicit production of attention map and hidden features can be formulated as (1).

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{H}_4 \boldsymbol{H}_4^T \\ \boldsymbol{H}_4^W &= softmax(\boldsymbol{A})\boldsymbol{H}_4 \\ \boldsymbol{H}_A &= \boldsymbol{H}_4 + \boldsymbol{H}_4^W \end{aligned} \tag{1}$$

In summary, stacked GRU structure is able to extract various phenology features from low-level to high-level, and attention module enables the TSNet to exploit the temporal

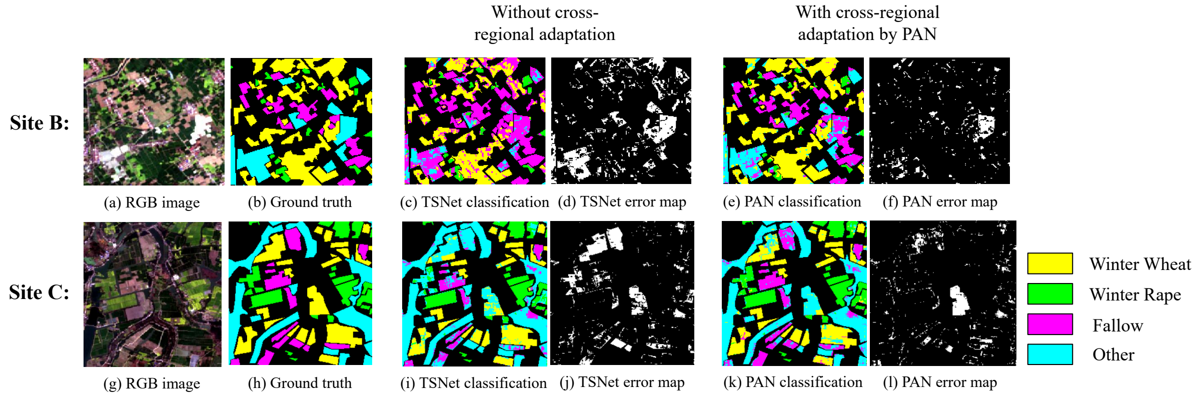


Figure 5. The results of image blocks in cross-regional experiment A→B and A→C. When the TSNet pre-trained on A is directly used to classify region B and C, wide misclassification occurs. Fine-tuned by our proposed PAN, the accuracy improves considerably and the misclassification is reduced which can be seen from subfigure (d,f,j,l)

Table 2. OA comparison between our methods and other SOTA methods (higher is better) of six transfer experiment set. The Avg. column indicates the average result of six transfer scenarios.

OA (%)	A→B	A→C	B→A	B→C	C→A	C→B	Avg.
RF [3]	88.01	76.96	82.56	73.01	86.20	88.21	82.49
Transformer [28]	81.83	75.10	73.93	76.65	80.20	75.82	77.26
TempCNN [22]	77.37	77.40	**83.70**	76.83	80.99	**89.07**	80.89
DCM [30]	78.00	79.24	79.79	73.46	82.30	85.43	79.70
TSNet (ours)	81.43	81.74	82.17	76.00	82.89	87.61	81.97
PAN (ours)	**89.18**	**83.55**	82.77	**81.66**	**88.08**	**88.97**	**85.70**

correspondence in a data-driven way, adaptively focusing on key phenology stage under various climate conditions.

4.2. Cross-regional adaptation by PAN

To adapt the source model to the unlabelled target region, PAN aims to finetune the pre-trained model by aligning the multi-level deep features extracted from two branches, mapping both source and target data into the same feature space without using the target labels. The overall structure and details of PAN are depicted in Figure 4.

Specifically, source and target data are respectively permuted in a random but unchanged order and paired and fed to two identical TSNet pre-trained on source domain.

Let $\boldsymbol{X}_S = \{\boldsymbol{x}_i^S, y_i^S | i = 1, 2, ..., n_S\}$ denote source data and labels, and $\boldsymbol{X}_T = \{\boldsymbol{x}_i^T | i = 1, 2, ..., n_T\}$ denote target data without annotation information. The two branches of PAN accept data pair $(\boldsymbol{x}_i^S, y_i^S, \boldsymbol{x}_i^T)$ as input and hierarchical features $\{\boldsymbol{H}_1^S, \boldsymbol{H}_2^S, \boldsymbol{H}_3^S, \boldsymbol{H}_4^S, \boldsymbol{H}_A^S, \}$ and $\{\boldsymbol{H}_1^T, \boldsymbol{H}_2^T, \boldsymbol{H}_3^T, \boldsymbol{H}_4^T, \boldsymbol{H}_A^T, \}$ are extracted. We deem hidden states $\boldsymbol{H}_2$ outputted from the second GRU layer as middle-level feature while the $\boldsymbol{H}_A$ outputted from the final attention module as high-level feature, and both of them have strong discriminability and complementary information for each other. MMD loss of phenology feature pair $(\boldsymbol{H}_2^S, \boldsymbol{H}_2^T)$ and $(\boldsymbol{H}_A^S, \boldsymbol{H}_A^T)$ is calculated. We define the hierarchical alignment loss by

$$\begin{aligned} \mathcal{L}_{Align} = & \left\| \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{h}_{2,i}^S) - \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{h}_{2,i}^T) \right\| \\ & + \left\| \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{h}_{A,i}^S) - \frac{1}{n} \sum_{i=1}^{n} \Phi(\boldsymbol{h}_{A,i}^T) \right\| \end{aligned} \tag{2}$$

where Φ indicates the multi-kernel gaussian function and the total loss function is defined by

$$\mathcal{L}_{Total} = \lambda \mathcal{L}_{Align} + \mathcal{L}_{Src} \tag{3}$$

where $\mathcal{L}_{Src}$ is the NLL loss calculated by source data and labels by

$$\mathcal{L}_{Src} = \frac{1}{n} \sum_{i=1}^{n} p(\boldsymbol{x}_i) \log(q(\boldsymbol{x}_i)) \tag{4}$$

In summary, by optimizing the weighted loss defined by (3), the proposed PAN aligns the feature distribution of two domains, transferring the TSNet trained on the source samples to the target domain.

5. Experiment

In this section, we compare our proposed TSNet and PAN with other SOTA on six experiments of cross-regional

Table 3. Macro F1 score comparison between our methods and other SOTA methods (higher is better) of six transfer experiment set. The Avg. column indicates the average result of six transfer scenarios.

Macro F1 score	A→B	A→C	B→A	B→C	C→A	C→B	Avg.
RF [3]	0.6741	0.6947	0.7104	0.5983	0.8163	0.7205	0.7024
Transformer [28]	0.6810	0.7041	0.6390	0.6777	0.7465	0.6264	0.6791
TempCNN [22]	0.7065	0.7365	**0.7214**	0.6523	0.7552	0.8010	0.7288
DCM [30]	0.7038	0.7689	0.6827	0.6157	0.7414	0.7333	0.7076
TSNet (ours)	0.7434	0.8077	0.7118	0.6683	0.7645	0.7858	0.7469
PAN (ours)	**0.8248**	**0.8123**	0.7082	**0.7861**	**0.8517**	**0.8103**	**0.7989**

Table 4. OA comparison of three aligning strategies of six transfer experiment sets. The results before PAN (the TSNet column) are also listed.

OA (%)	TSNet	S1	S2	S3
A→B	81.43	88.78	89.18	**89.64**
A→C	81.74	82.13	83.55	83.10
B→A	82.17	**83.34**	82.77	83.27
B→C	76.00	79.73	**81.66**	80.84
C→A	82.89	87.78	**88.08**	87.83
C→B	87.61	88.25	88.97	**89.11**
Avg.	81.97	85.00	**85.70**	85.63

Table 5. Macro F1 score comparison of three aligning strategies of six transfer experiment sets. The results before PAN (the TSNet column) are also listed.

Macro F1 score	TSNet	S1	S2	S3
A→B	0.7434	0.8225	0.8248	**0.8315**
A→C	0.8077	0.7925	**0.8123**	0.8083
B→A	0.7118	**0.7328**	0.7082	0.7276
B→C	0.6683	0.7546	**0.7861**	0.7769
C→A	0.7645	0.8460	**0.8517**	0.8462
C→B	0.7858	0.7984	0.8103	**0.8149**
Avg.	0.7469	0.7911	0.7989	**0.8009**

time series crop classification, including random forest (RF) [3], Transformer [28], TempCNN [22] and DCM [30]. Firstly, we pre-train our TSNet and other state-of -the-art (SOTA) algorithms on source domain and apply them to target domain directly. Furthermore, we fine-tune the pre-trained TSNet by the unsupervised PAN framework and present the evaluation indices after adaptation.

5.1. Training setup

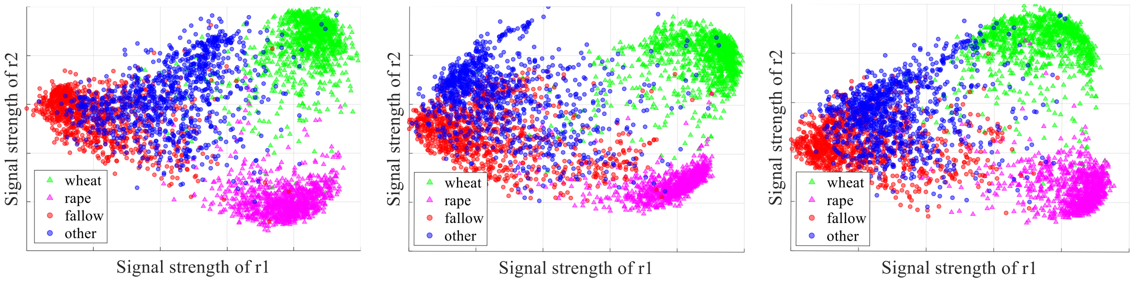
Pre-training phase. In this phase, all methods are trained solely on each site's training set to acquire pre-trained models of three regions. To be fair, every compared method was repeated trained on each subset from scratch 15 times with the same training configuration. That is, we saved 45 pre-trained models for each method. RF classifier was trained with parameter $tree\_num = 50$ and $leaf\_size = 15$. For DCM, we grid-searched the best parameters and set $hidden\_dim = 256$ and $num\_layer = 3$ for all six transfer sets. For TempCNN, we modified the $dropout$ to 0.5 and kept other parameters as the original paper. For the Transformer model, we grid-searched the best parameters and set $d_{model} = 64$ and the network inner dimensionality $d_{inner} = 128$ given that the complexity of pixelwise classification task is relatively low. The stack number of multi-head attention module was set as $N = 4$. For TSNet, the dimensionality of hidden states of GRU was 128, and with a dropout of 0.5 between each layer. The optimizer of all deep learning methods was replaced with Adam optimizer with initial learning rate of 0.001 and $\beta = (0.9, 0.998)$. The RF classifier was trained using MATLAB R2019a, and all deep models were implemented by PyTorch and trained on an NVIDIA Tesla V100 until the end of 100 epochs or convergence.

Fine-tuning phase. Three study sites are combined in two pairs to get six transfer settings: A→B, A→C, B→A, B→C, C→A, C→B. Taking the task A→B for example, data pair containing samples from A and B is sent to two TSNet which are pre-trained on A's training set. Then the overall loss defined by (3) is back-propagated using Adam optimizer with initial learning rate of 0.0001 and $\beta = (0.9, 0.998)$. By the end of 20 epochs, the source TSNet is finetuned to target domain, site B, and the predicting labels of B are reported. We need to emphasize that in this procedure, we didn't use the label information from site B.

5.2. Results

Two indexes, overall accuracy (OA) and macro F1 score, are used to evaluate the performance of proposed TSNet and PAN framework. OA is calculated by the ratio of number of samples predicting right by number of all samples. Macro F1 score is the average F1 scores of all classes, which evaluates the general model performance on each class and is commonly adopted by dataset with imbalance classes.

Table 2 and 3 displays the OA and macro F1 score results of our method compared to SOTA. In inference stage, our TSNet shows OA increase over deep learning methods



(a) Samples of site C inferred on models pre-trained on site C (OA: 94.39%)

(b) Samples of site A inferred on models pre-trained on site C (OA: 81.76%)

(c) Samples of site A inferred on fine-tuned models by PAN (OA: 87.84%)

Figure 6. Visual maps of deep features before and after DA. The final OA are also presented.

and macro F1 increase over all compared methods, which demonstrates that our proposed backbone can extract generalized and domain-robust features from SITS. It is worth noting that the RF acquires a relative higher OA but a relative lower macro F1 score. By analyzing the confusion matrix, we find that the RF has higher accuracy on *other* class which occupies a large proportion in total, but has lower accuracy on crop types, resulting in lower F1 score index. In fine-tuning stage, our PAN framework further improves the overall accuracy by 3.73% and macro F1 score by 0.052 without using any label information in target domain. In case A→B and B→C, the average increase of OA even reaches 7.75% and 5.66%. To verify the effectiveness more intuitively, visual results of experiment set A→B and A→C are listed in Figure 5. The experiment results demonstrate that 1) the regional phenology discrepancy is mitigated by aligning the hierarchical phenology features; 2) our unsupervised DA-based framework improves the overall accuracy of cross-regional crop type classification, and has the potential to be applied to other large area classification or mapping tasks.

5.3. Ablation study of the PAN framework

The TSNet can extract hierarchical deep features, yet the contribution of different levels of phenology features to the transfer procedure need to be explored. We analyze this problem with an ablation experiment. Scenario 1 (S1): calculate MMD loss of only high-level feature, *i.e.* $\boldsymbol{H}_A$. Scenario 2 (S2): calculate MMD loss of the high-level feature and the mid-level feature, *i.e.* $\boldsymbol{H}_2 + \boldsymbol{H}_A$. Scenario 3 (S3): calculate MMD loss of low, middle and high-level features which are outputted by every layer of GRU and the attention module, *i.e.* $\boldsymbol{H}_1 + \boldsymbol{H}_2 + \boldsymbol{H}_3 + \boldsymbol{H}_4 + \boldsymbol{H}_A$.

Table 4 and 5 report the indices comparison of three scenarios. To better exhibit the improvement, the results of pre-trained models without DA are also listed. Experiment results show that all three strategies gain considerable improvement to baseline. The accuracy difference between S2 and S3 is negligible while the latter brings extra computational cost and is easier to overfit. Thus, we adopt S2 as the default aligning strategy in proposed PAN framework.

5.4. Feature map visualization

To demonstrate how our framework alleviates the cross-regional discrepancy problem more intuitively, we visualize the hidden features before and after the transfer procedure. We take transfer set C→A for example. 1000 samples in site C and A are randomly selected and fed into pre-trained and fine-tuned models. The final features yielded by attention module are saved to local. The dimensionality of hidden features is reduced from 128 to 2 for visualization, and the overall accuracy is also reported in Figure 6. The visual maps show that when applying model trained on C directly to samples of A, confusions tend to occur and the distance between rape and fallow class is close. Fine-tuned by the PAN framework, the clustering centers of wheat, rape and fallow class are more discriminative. The accuracy is prominently improved even though certain confusion between fallow and other class still exists.

6. Conclusion

In this paper, we firstly propose a TSNet for time series crop classification to extract phenological features and temporal dependencies from multi-temporal input data. To address the cross-domain challenges, a novel framework named PAN is further proposed to fine-tune the source model to the target regions, improving the accuracy of cross-regional classification based on an unsupervised strategy. Besides, a challenging time series crop dataset is annotated to verify the effectiveness of our methods. The experiment results of six transfer sets demonstrate that 1) the TSNet is capable of extracting generalized and domain-robust features and outperforms other SOTA methods and 2) our PAN framework noticeably improves the classification results without using any label information in the target domain. The visualization of deep features proves that the advancement of indices is achieved by aligning hierarchical phenology features effectively.

References

[1] Mesay Belete Bejiga, Farid Melgani, and Pietro Beraldini. Domain adversarial neural networks for large-scale land cover classification. *Remote Sensing*, 11(10):1153, 2019. 3

[2] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006. 2

[3] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001. 6, 7

[4] Mark Broich, Matthew C Hansen, Peter Potapov, Bernard Adusei, Erik Lindquist, and Stephen V Stehman. Time-series analysis of multi-resolution optical imagery for quantifying forest cover loss in sumatra and kalimantan, indonesia. *International Journal of Applied Earth Observation and Geoinformation*, 13(2):277–291, 2011. 3

[5] Yaping Cai, Kaiyu Guan, Jian Peng, Shaowen Wang, Christopher Seifert, Brian Wardlow, and Zhan Li. A high-performance and in-season classification system of field-level crop types using time-series landsat data and a machine learning approach. *Remote sensing of environment*, 210:35–47, 2018. 1

[6] RS De Fries, M Hansen, JRG Townshend, and R Sohlberg. Global land cover classifications at 8 km spatial resolution: the use of training data derived from landsat imagery in decision tree classifiers. *International Journal of Remote Sensing*, 19(16):3141–3168, 1998. 3

[7] Dennis C Duro, Steven E Franklin, and Monique G Dubé. A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using spot-5 hrg imagery. *Remote sensing of environment*, 118:259–272, 2012. 3

[8] Steven E Franklin, Oumer S Ahmed, Michael A Wulder, Joanne C White, Txomin Hermosilla, and Nicholas C Coops. Large area mapping of annual land cover dynamics using multitemporal change detection and classification of landsat time series data. *Canadian Journal of Remote Sensing*, 41(4):293–314, 2015. 3

[9] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3146–3154, 2019. 5

[10] Cristina Gómez, Joanne C White, and Michael A Wulder. Optical remotely sensed time series data for land cover classification: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 116:55–72, 2016. 3

[11] Patrick Griffiths, Sebastian van der Linden, Tobias Kuemmerle, and Patrick Hostert. A pixel-based landsat compositing algorithm for large area land cover mapping. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 6(5):2088–2101, 2013. 2

[12] Nancy B Grimm, Stanley H Faeth, Nancy E Golubiewski, Charles L Redman, Jianguo Wu, Xuemei Bai, and John M Briggs. Global change and the ecology of cities. *science*, 319(5864):756–760, 2008. 1

[13] Yousra Hamrouni, Eric Paillassa, Véronique Chéret, Claude Monteil, and David Sheeren. From local to global: A transfer learning-based approach for mapping poplar plantations at national scale using sentinel-2. *ISPRS Journal of Photogrammetry and Remote Sensing*, 171:76–100, 2021. 3

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 3

[15] Reza Khatami, Giorgos Mountrakis, and Stephen V Stehman. A meta-analysis of remote sensing research on supervised pixel-based land-cover image classification processes: General guidelines for practitioners and future research. *Remote Sensing of Environment*, 177:89–100, 2016. 2

[16] Nataliia Kussul, Mykola Lavreniuk, Sergii Skakun, and Andrii Shelestov. Deep learning classification of land cover and crop types using remote sensing data. *IEEE Geoscience and Remote Sensing Letters*, 14(5):778–782, 2017. 1

[17] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 3

[18] Zhaokui Li, Xiangyi Tang, Wei Li, Chuanyun Wang, Cuiwei Liu, and Jinrong He. A two-stage deep domain adaptation method for hyperspectral image classification. *Remote Sensing*, 12(7):1054, 2020. 3

[19] Ajay Mathur and Giles M Foody. Crop classification by support vector machine with intelligently selected training data for an operational application. *International Journal of Remote Sensing*, 29(8):2227–2240, 2008. 3

[20] Lichao Mou, Lorenzo Bruzzone, and Xiao Xiang Zhu. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. *IEEE Transactions on Geoscience and Remote Sensing*, 57(2):924–935, 2018. 3

[21] Giorgos Mountrakis, Jungho Im, and Caesar Ogole. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 66(3):247–259, 2011. 1

[22] Charlotte Pelletier, Geoffrey I Webb, and François Petitjean. Temporal convolutional neural network for the classification of satellite image time series. *Remote Sensing*, 11(5):523, 2019. 3, 6, 7

[23] Kyle Pittman, Matthew C Hansen, Inbal Becker-Reshef, Peter V Potapov, and Christopher O Justice. Estimating global cropland extent with multi-year modis data. *Remote Sensing*, 2(7):1844–1863, 2010. 3

[24] Marc Rußwurm and Marco Körner. Multi-temporal land cover classification with sequential recurrent encoders. *ISPRS International Journal of Geo-Information*, 7(4):129, 2018. 3

[25] Marc Rußwurm and Marco Körner. Self-attention for raw optical satellite time series classification. *ISPRS Journal of Photogrammetry and Remote Sensing*, 169:421–435, 2020. 3

[26] Marc Rußwurm, Sébastien Lefèvre, and Marco Körner. Breizhcrops: A satellite time series dataset for crop type identification. In *Proceedings of the International Conference on Machine Learning Time Series Workshop*, 2019. 3

[27] Xin-Yi Tong, Gui-Song Xia, Qikai Lu, Huanfeng Shen, Shengyang Li, Shucheng You, and Liangpei Zhang. Land-cover classification with high-resolution remote sensing images using transferable deep models. *Remote Sensing of Environment*, 237:111322, 2020. 3

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017. 3, 6, 7

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 5

[30] Jinfan Xu, Yue Zhu, Renhai Zhong, Zhixian Lin, Jialu Xu, Hao Jiang, Jingfeng Huang, Haifeng Li, and Tao Lin. Deepcropmapping: A multi-temporal deep learning approach with improved spatial generalizability for dynamic corn and soybean mapping. *Remote Sensing of Environment*, 247:111946, 2020. 3, 6, 7

[31] Dujuan Zhang, Yaozhong Pan, Jinshui Zhang, Tangao Hu, Jianhua Zhao, Nan Li, and Qiong Chen. A generalized approach based on convolutional neural networks for large area cropland mapping at very high resolution. *Remote Sensing of Environment*, 247:111912, 2020. 3

[32] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 5

[33] Liangpei Zhang, Lefei Zhang, and Bo Du. Deep learning for remote sensing data: A technical tutorial on the state of the art. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):22–40, 2016. 3

[34] Liheng Zhong, Lina Hu, and Hang Zhou. Deep learning based multi-temporal crop classification. *Remote sensing of environment*, 221:430–443, 2019. 3

[35] Ruixi Zhu, Li Yan, Nan Mo, and Yi Liu. Semi-supervised center-based discriminative adversarial learning for cross-domain scene-level land-cover classification of aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:72–89, 2019. 3

[36] Xiao Xiang Zhu, Devis Tuia, Lichao Mou, Gui-Song Xia, Liangpei Zhang, Feng Xu, and Friedrich Fraundorfer. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, 5(4):8–36, 2017. 3

[37] Zhe Zhu and Curtis E Woodcock. Automated cloud, cloud shadow, and snow detection in multitemporal landsat data: An algorithm designed specifically for monitoring land cover change. *Remote Sensing of Environment*, 152:217–234, 2014. 2