Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(6), 272-287
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

LRTACF-NET: Lowest-Resolution Temporal Attention and Cross Feedback for Multi-Temporal Crop Classification

Hanfen Zang[1], Xiangfeng Wei[2], Xiongyong Sun[3]

[1]Institute of Quantitative & Technological Economics, Chinese Academy of Social Sciences, Beijing, China
[2]Institute of Acoustics, Chinese Academy of Sciences, Beijing, China
[3]Lanyun Technology Group Co., Ltd., Beijing, China
Email: zanghf@163.com, wxf@mail.ioa.ac.cn, xiongyong97@163.com

How to cite this paper: Zang, H.F., Wei, X.F. and Sun, X.Y. (2025) LRTACF-NET: Lowest-Resolution Temporal Attention and Cross Feedback for Multi-Temporal Crop Classification. *Journal of Computer and Communications*, **13**, 272-287.
https://doi.org/10.4236/jcc.2025.136018

Received: May 18, 2025
Accepted: June 27, 2025
Published: June 30, 2025

Abstract

Accurate land cover and land use (LCLU) classification is critical for environmental monitoring, agricultural planning, and sustainable development. However, distinguishing spectrally similar land categories, such as crop types, remains challenging due to the limited ability of traditional methods to extract discriminative features. To address this, we propose a multi-feedback mechanism with a lightweight self-attention model, where multi-scale feature maps are progressively enhanced through deep supervision for robust feature extraction and fusion. Leveraging the high-resolution satellite time series data, LRTACF-NET in this paper demonstrates significant improvements over state-of-the-art approaches, achieving +10% mIoU and +9% mF1 in quantitative metrics. Notably, while maintaining high accuracy, our Model-7 reduces computational costs by 28% in FLOPs compared to UNet++. Although the best-performing model incurs higher computational cost in terms of FLOPs compared to the baseline, it achieves superior classification accuracy over existing LCLU approaches—including state-of-the-art foundation models—especially in mitigating misclassification among spectrally similar crops. Extensive experimental results demonstrate that LRTACF-NET achieves the highest scores in mIoU, mF1, and overall accuracy, thereby offering a scalable solution for precision LCLU mapping, particularly in crop classification.

Keywords

Temporal Attention, Satellite Images Temporal Series (SITS), Land Cover and Land Use (LCLU), Crop Mapping, Deep Learning

1. Introduction

Remote Sensing (RS) technology allows the observation and analysis of Earth's surface featured by detecting reflected and emitted radiation from a distance. Recently, RS image processing has undergone a technological evolution from traditional machine learning to deep learning, and more recently to vision Transformers. These advancements have played a significant role in applications such as crop classification, land cover and land use (LCLU) analysis, contributing greatly to precision agriculture and the supervision and management of natural resources.

Early crop classification in satellite RSIs (Remote Sensing Images) applied spectrum features such as NDVI and GLCM [1]-[5], and traditional machine learning (e.g. Random Forest [6]-[8], SVM [9]), but the precision is not high and the feature extraction is difficult. CNN-based models have made significant progress in crop classification due to its excellent ability for feature expression. The most classical model is U-Net [10], and Jia proposed an enhanced U-Net with spatial attention [11]. Other models such as ResNet-10 [12], Stacked Spectral Feature Space Patch (SSFSP) [13], were also applied for crop classification in RSIs. However, most of the CNN-based models are trained on RGB images (e.g., ImageNet). Recently researchers considered other deep learning models like Transformer, especially for time series satellite images. Transformer model was first proposed in NLP based on word embedding. Some researchers in the field of satellite remote sensing suggest that neighboring patches or tiles can be treated like word embeddings in NLP. Accordingly, patch embedding [14] and tile embedding techniques were introduced into RSIs [15]. Similarly, the self-attention mechanism in Transformer has been introduced into Satellite Images Time Series (SITS) due to its ability to handle the inherent heterogeneity and temporal dynamics of geophysical processes such as crop growth. Several Transformer-based models have been proposed [16]-[19], but self-attention alone cannot effectively address the continuous down-sampling of spatial information.

Reference [20] introduced Prithvi, a transformer-based geospatial foundation model pre-trained on more than 1TB of multispectral satellite data collected from the Harmonized Landsat–Sentinel-2 (HLS) dataset. Xie *et al.* [21] evaluated large-scale foundation models such as Prithvi, ViT, and Segformer, traditional machine learning models (e.g., RF, XGB) and conventional deep learning models (e.g., U-Net) using the IBM-NASA-HLS dataset. They found that traditional models still exhibited comparable or even superior performance to foundation models, especially in tasks where texture information is less informative for classification. Hsu *et al.* raised the question of whether the pre-trained Prithvi model, which was trained on six bands and a geospatial dataset, outperforms other pre-trained models [22]. The answer lies in the lack of joint learning of spectral, spatial, and temporal features, which has hindered the resolution of inter-class similarities among crops.

To bridge this gap, we propose LRTACF-Net, a novel deep learning framework that enhances the U-Net baseline by integrating temporal self-attention at the en-

coder's lowest-resolution level and a Cross Feedback with Mutual Attention (CFMA) mechanism in the decoder. It is a brand-new model based on U-Net for better performance in crop classification in MTCC (Multi-Temporal Crop Classification) dataset[1]. The proposed method advances the state-of-the-art in three key aspects:

1) Multi-scale temporal modeling: By embedding multi-head self-attention in the encoder, LRTACF-Net captures long-range temporal dependencies across crop growth stages, suppresses irrelevant spectral bands, and amplifies discriminative pixel-level features.

2) Cross-feature refinement: The CFMA module enables iterative feedback between decoder layers, fusing multi-level temporal attention features to refine the classification head and improve separability of spectrally similar crops.

3) Deeply supervised optimization: A hybrid loss function aggregates multi-layer predictions, enhancing training stability and segmentation metrics (e.g., mIoU, F1-score).

This paper addresses the challenge of adaptive temporal feature extraction in LCLU applications, offering a scalable solution for precision agriculture. Experimental validation on benchmark datasets demonstrates superior performance over existing methods, particularly in fine-grained crop classification scenarios.

2. Methodology

The overview of LRTACF-NET model is shown in **Figure 1**, and can be expressed as the following equations (1)-(5). The input image data was split into T1, T2 and T3, as shown in **Figure 2**, each has 6 bands and the image size is 224 × 224. The output includes six decoders, one final cross-feedback result and one refined result.

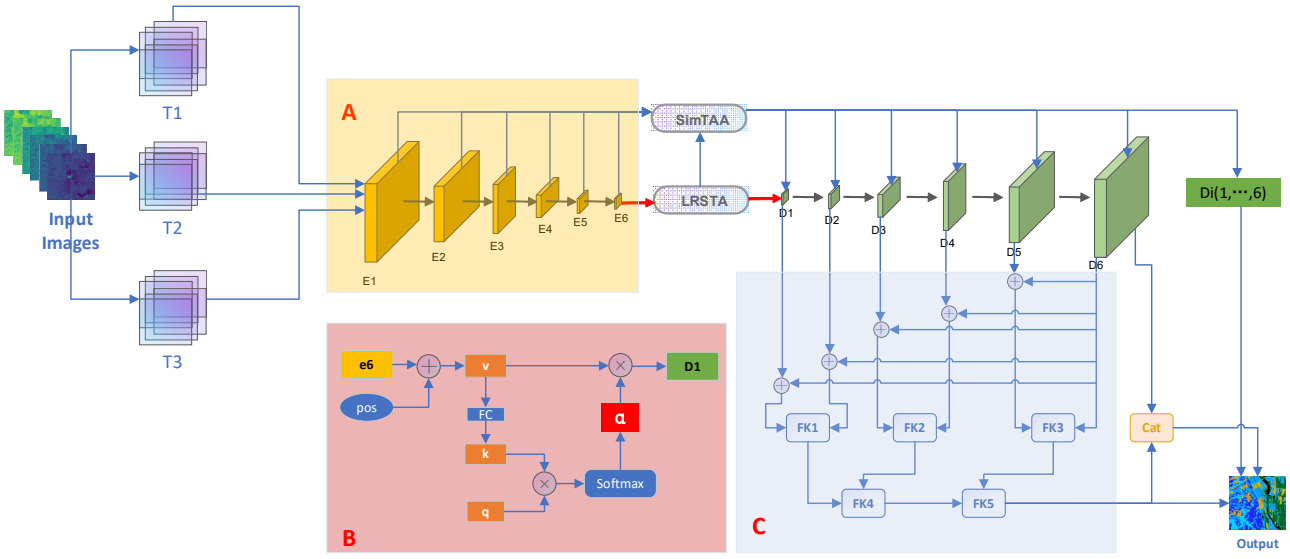


Figure 1. Overall framework of LRTACF-NET.

[1]https://huggingface.co/datasets/ibm-nasa-geospatial/multi-temporal-crop-classification.

Figure 2. Three temporal satellite images time series.

$$E_i^T = UNET\text{-}Encoder(T1, T2, T3)\ i = 1,\ldots,6 \quad (1)$$

$$D_i = SkipConn(Attention(E_i^T))\ i = 1,\ldots,6 \quad (2)$$

$$O_0 = CrossFeedback(D_i)\ i = 1,\ldots,6 \quad (3)$$

$$Refine = Concat(O_0, D_6) \quad (4)$$

$$Output = DeepSup(O_0, D_i, Refine)\ i = 1,\ldots,6 \quad (5)$$

2.1. UNET-Encoder Network

This module mainly extracts feature maps, including low-level and high-level features. The original input image is a 224 × 224 image with 18 channels (MTCC dataset). These 18 channels can be divided again into three temporal sets (each contains 6 channels). The three Temporal sets are satellite image data of the early, middle and late stages of crop growth as shown in **Figure 2**, defined as T1, T2 and T3.

The classic U-Net network encoder with 6 layers of depth is shown in **Figure 1** (Region A). Each layer processes T1, T2 and T3 simultaneously, by stacks of double 3 × 3 convolution, BatchNorm2D and ReLU Activation Function, and then followed by the pooling operation, which halves the image size. After a serial operation in 6 layers, the image size of the deepest layer reaches the size of 7 × 7, and the number of channels turns into 1024. The encoder results from the input image at time t is $E_i^t \in R^{(6\times H\times W)}$, where 6 is the number of channels at one temporal image (e.g., T1), and H and W both are 224. When putting E_1^1 feature map into UNET-Encoder layer 1, it will output 64 × H × W. After six layers' processing, get $E_6^1 \in 1024\times\frac{H}{32}\times\frac{W}{32}$ (channels of layer6 is also 1024).

2.2. LRTSA Module

After parallel processing in UNET-Encoder network, three 2-dimension feature maps in 6 encoder layers can be gotten, and were stacked in each layer to form 3-dimension image data as the input of LRTSA (Lowest Resolution Temporal Self-Attention) network. Similar to Transformer, the lowest resolution feature maps,

temporal self-attention is used. While in other 5 layers' feature maps, temporal self-attention aggregation is applied by upsampling the temporal self-attention matrix from the lowest resolution feature maps.

In order to apply self-attention mechanism in image segmentation network, the pixels in temporal images (T1, T2, and T3) are viewed as the words, and calculated PE (Positional Encoding) of each pixel. By summing up encoded position and feature map in the lowest resolution layer (the sixth layer in UNET-Encoder), three vectors are calculated: q(query), k(key), and v(value) from each pixel, as shown in Region B in **Figure 1**.

In order to calculate the positional features of feature map at t, [0, 1, 2] is encoded as the positional value of T1, T2 and T3 respectively, and then add these values to the feature map in one head pixel by pixel. According to Transformer model, q*k*v can be taken as the output classification head for the output sequence. In other words, D_1 (namely, decoder1) will be set as the decoder feature map, see (6) and Region B in **Figure 1**.

$$D_1 = \sum_{t=1}^{3} a_h^{(t)} \cdot v_{h,pos}^{(t)} \tag{6}$$

For reducing computing complexity, in other 5 layers in UNET-Encoder, attention is simply up-sampled and multiplied with feature maps (**Figure 1**). Here the lowest resolution's self-attention matrix $a_h^{(t)}$ must be up-sampled, then it will be multiplied by the corresponding Encoder e_l^t to get the higher decoder D_i (i = 1, 2, …, 6).

However, agricultural crops have difficulties distinguishing similar crops due to their similar shapes, colors and texture. In order to improve the performance of crop image segmentation, CFMA (Cross-Feedback with Mutual Attention) was designed in this paper (section 2.3.).

2.3. Cross-Feedback with Mutual Attention Module

After getting the Decoders (D_1 - D_6), for better obtaining complete features in each Decoder level, and outputting more precise class mapping, mutual attention cross feedback module was applied as shown in Region C in **Figure 1**.

Decoder1 (D_1) is defined as the lowest resolution from LRTSA module, whose size is 7 × 7 and channel is 1024. The highest resolution is D_6, whose size is the same as the original image size. Since D_6 contains global features from the final skip connection in LRTSA, it was down-sampled to match the size of the other five Decoders (denoted FB_i), with its channel dimension adjusted to 64. These global features (referred to as FB_i') are then be added to enable cross-feedback between the two decoders.

The first step of CFMA module is mutual attention between the two decoders, where each pixel at different layers and positions can receive attention from others. After a 3 × 3 convolution followed by batch normalization (BN) and ReLU activation, another convolution, BN, and a Sigmoid activation generate the attention weights.

The second step of CFMA module is to generate a 4-dilation convolution (rate = 1, 2, 3, 4). The four dilation convolutions can be used for classification head, and each convolution was combined with its last dilation convolution. This can enlarge the vision field and obtain more features in different layers without extra parameters. After the four dilation convolutions, the feature map can be used for final output, which is the classification of each pixel in the input image.

As shown in region C in **Figure 1**, five cross-feedback modules (FK1-FK5) are designed for the fusion of all levels of feature maps. After five times of cross-feedback modules (FK1-FK5), the result DC_fuse contains high-level, low-level features, and the expanded fields by dilation convolution.

2.4. Deep Supervision and Loss Function

To integrate the outputs of the decoder network and the final cross-feedback result, the combined result of D_6 and DC_fuse is used as a refined fusion feature. The refined fusion can be outputted as the final 14 classes result (including background).

All the results, including refined result, the final result of cross-feedback result, six results of decoder (D_1 - D_6), can serve as classified result for deep supervision. Since the output of each layer of the network (the image size can be made consistent with the original image size through up-sampling) has certain features, deep supervision takes advantage of this characteristics to calculate the loss function, for obtaining a unified loss, and then for adjusting the model's training parameters through backpropagation. To balance computational cost and model recognition performance, we calculate the total loss based on the final output of the cross-feedback block, the refined fusion features, and the outputs of all layers in the LRTSA module, with a hybrid loss function. The objective is to integrate the respective advantages of the two employed loss functions to improve classification performance. To address the class imbalance problem, a combination of weighted Binary Cross-Entropy loss and IoU loss is employed, as shown in (7).

$$\mathfrak{L}\text{oss}_{JOINT} = BCE_{loss} * Weight_{BCE} + IOU_{loss} * Weight_{IOU} \tag{7}$$

3. Experiments and Results

The dataset is MTCC (see footnote 1), and there are 13 classes in it, without background class. The proposed LRTACF-NET and the comparative methods were implemented using PyTorch on a system equipped with an Intel Core i7-13700K processor (3.4 GHz), 32 GB of RAM, and an NVIDIA GeForce RTX 3090 graphics processing unit with 24 GB of memory. There are 9 methods for comparison in **Table 1**.

In training phrase, batch size is 6, initial learning rate is 0.011, learning decay strategy is polynomial decay policy, max epoch is 100, and the optimizer is SAM. Loss function is a jointed loss with BCE and IoU in section 2.

In inference phrase, F1 and IoU are adopted to evaluate the performance of the predicted results on the test set, including class-wise F1-Score, Precision and Recall Scores across all classes, overall accuracy (OA), mean F1 (mF1) and mean intersection over union (mIoU). All experimental results of 9 models and their

comparisons with other papers are shown in **Table 2** and **Table 3** (without background class).

Table 1. Models settings.

Model	Encoder	DS	LOSS	LRTSA	CFMA
1	U-Net		Tversky		
2	U-Net		wIoU + wBCE		
3	U-Net	√	wIoU + wBCE		
4	UNet++		Tversky		
5	UNet++		wIoU + wBCE		
6	UNet++	√	wIoU + wBCE		
7	U-Net	√	wIoU + wBCE	√	
8	U-Net	√	wIoU + wBCE	√	3 times
9	U-Net	√	wIoU + wBCE	√	5 times

DS = Deep Supervision; wIoU = weighted IoU (Intersection over Union); wBCE = weighted Binary Cross Entropy.

Table 2. IoU comparison for multi-temporary crop classification (%).

Model	Natural vegetation	Forest	Corn	Soybeans	Wetlands	Developed /Barren	Open water	Winter wheat	Alfalfa	Fallow/Idle cropland	Cotton	Sorghum	Other	mIoU
Model-9	56.72	55.91	70.29	67.48	49.72	58.08	77.34	59.90	45.80	42.86	49.06	45.76	48.52	55.96
Model-8	56.87	55.27	70.45	67.17	49.70	57.87	77.18	59.87	45.51	42.69	47.90	45.92	48.78	55.78
Model-7	56.92	55.74	69.98	67.01	49.24	57.58	77.37	59.23	45.29	42.53	48.37	45.31	48.10	55.59
Model-6	55.72	55.31	66.36	63.99	49.38	58.75	76.40	56.95	43.32	40.14	42.86	42.33	43.70	53.48
Model-5	55.39	55.07	65.76	63.02	48.33	58.29	76.61	56.32	42.64	38.70	40.90	42.08	42.99	52.78
Model-4	47.26	51.06	56.19	55.00	44.83	51.58	76.57	51.59	40.97	31.85	34.35	38.55	35.08	47.30
Model-3	55.34	54.75	64.71	61.94	48.32	56.77	76.86	56.32	42.22	39.26	40.72	40.97	42.02	52.32
Model-2	54.22	54.44	63.47	61.03	45.68	56.61	75.98	54.10	40.37	38.00	37.11	39.11	38.45	50.66
Model-1	46.30	48.98	53.80	51.79	44.18	48.92	76.64	49.61	39.49	33.88	33.16	37.56	33.18	45.96
Prithvi[b]	*40.38*	*47.47*	*54.91*	*52.97*	*40.20*	*36.11*	*68.04*	*49.67*	*30.84*	*34.93*	*32.37*	*32.83*	*34.27*	*42.69*
U-Net[b]	*55.21*	*54.79*	*64.29*	*63.08*	*40.99*	*54.65*	*76.86*	*58.17*	*39.12*	*42.64*	*42.26*	*41.76*	*42.20*	*52.00*
RFaug[b]	*50.25*	*53.76*	*59.56*	*56.51*	*44.20*	*42.00*	*76.10*	*53.24*	*30.64*	*38.68*	*35.60*	*39.80*	*39.72*	*47.70*
ViT[b]	*44.39*	*41.63*	*50.07*	*49.22*	*37.69*	*28.33*	*68.34*	*41.44*	*24.74*	*32.22*	*29.47*	*32.68*	*29.78*	*39.23*
Segformer[b]	*47.14*	*45.10*	*52.95*	*51.30*	*37.14*	*32.05*	*69.34*	*46.69*	*29.00*	*34.58*	*31.86*	*35.20*	*33.73*	*42.01*

[a]Red, Green, Blue and orange means the best, second best, third best and other paper's best results based on different models. [b]See Reference Paper, Yiqun Xie, etc. 2024 [21].

Table 3. F1 comparison for multi-temporary crop classification (%).

Model	Natural vegetation	Forest	Corn	Soybeans	Wetlands	Developed/ Barren	Open water	Winter wheat	Alfalfa	Fallow/Idle cropland	Cotton	Sorghum	Other	mF1
Model-9	**72.38**	**71.72**	**82.55**	**80.58**	**66.42**	**73.48**	**87.22**	**74.92**	**62.83**	**60.00**	**65.83**	**62.79**	**65.34**	**71.24**
Model-8	**72.51**	**71.19**	**82.66**	**80.36**	**66.40**	**73.31**	**87.12**	**74.90**	**62.55**	**59.84**	**64.78**	**62.94**	**65.57**	**71.09**
Model-7	**72.54**	**71.58**	**82.34**	**80.25**	**65.99**	**73.08**	**87.24**	**74.40**	**62.35**	**59.68**	**65.20**	**62.36**	**64.96**	**70.92**
Model-6	**71.56**	**71.23**	**79.78**	**78.04**	**66.11**	**74.01**	**86.62**	**72.57**	**60.45**	**57.29**	**60.01**	**59.48**	**60.82**	**69.07**
Model-5	**71.29**	**71.03**	**79.34**	**77.32**	**65.17**	**73.65**	**86.75**	**72.06**	**59.79**	**55.80**	**58.05**	**59.24**	**60.13**	**68.43**
Model-4	**64.19**	**67.60**	**71.95**	**70.96**	**61.91**	**68.06**	**86.73**	**68.06**	**58.13**	**48.31**	**51.13**	**55.65**	**51.94**	**63.43**
Model-3	**71.25**	**70.76**	**78.57**	**76.50**	**65.16**	**72.43**	**86.92**	**72.06**	**59.38**	**56.38**	**57.88**	**58.12**	**59.17**	**68.04**
Model-2	**70.31**	**70.50**	**77.66**	**75.80**	**62.71**	**72.29**	**86.35**	**70.21**	**57.52**	**55.07**	**54.13**	**56.23**	**55.54**	**66.49**
Model-1	**63.29**	**65.75**	**69.96**	**68.24**	**61.28**	**65.70**	**86.78**	**66.32**	**56.62**	**50.61**	**49.80**	**54.61**	**49.83**	**62.21**
Prithvi[b]	*57.53*	*64.38*	*70.89*	*69.25*	*57.34*	*53.06*	*80.98*	*66.37*	*47.14*	*51.77*	*48.91*	*49.43*	*51.05*	*59.09*
U-Net[b]	*71.14*	*70.79*	*78.27*	*77.36*	*58.14*	*70.67*	*86.92*	*73.56*	*56.24*	*59.78*	*59.41*	*58.92*	*59.36*	*67.74*
RFaug[b]	*66.88*	*69.92*	*74.65*	*72.21*	*61.30*	*59.16*	*86.42*	*69.48*	*46.91*	*55.78*	*52.51*	*56.94*	*56.86*	*63.77*
ViT[b]	*44.39*	*41.63*	*50.07*	*49.22*	*37.69*	*28.33*	*68.34*	*41.44*	*24.74*	*32.22*	*29.47*	*32.68*	*29.78*	*39.23*
Segformer[b]	*47.14*	*45.10*	*52.95*	*51.30*	*37.14*	*32.05*	*69.34*	*46.69*	*29.00*	*34.58*	*31.86*	*35.20*	*33.73*	*42.01*

[a]Red, Green, Blue and orange means the best, second best, third best and other paper's best results based on different models. [b]See Reference Paper, Yiqun Xie, etc. 2024 [21].

The best mIoU (Model-9) is 55.96%, which is higher than the results in [21], as shown in **Table 2**. The best mF1 (Model-9) is 71.24%, which is also higher than the results in [21], as shown in **Table 3**. The results show that U-Net based models are better than other models like ViT, Segformer, even the Prithvi model of IBM and NASA, which was based on big foundation model. From Model-4 to Model-6, UNet++ was applied with different configurations, shown in **Table 1**. Model-4 is UNet++ based on Tversky loss. The results are lower than U-Net. After replacing mixed-loss (Model-5) and adding Deep Supervision (Model-6), the results improved over 8%.

Table 4 shows the performance about parameter scale, time consuming, Overall Accuracy (OA) score, Average Accuracy, Average Precision and mIoU of all models. In Model-2, just changing loss function from Tversky (Model-1) to weighted BCE_IoU loss function (Model-2), all indexes have improved. After adding deep supervision to Model-3, the time-consuming jump up from nearly 6 hours to 7 hours and 24 minutes, but all indexes also increased higher. When temporal attention and cross-feedback module were added, the results of four evaluation indexes (OA, Avg. Accuracy, Avg. Precision, and mIoU) were also higher and higher, while the time-consuming is increased over 2 times.

Table 4. Parameters, FLOPs, OA, and other metrics.

Model	Parameters (M)	FLOPs (G)	Time	OA	Avg. accuracy	Avg. precision	mIoU
Model-9	165.8	886.8	15h16m	**0.7332**	**0.7095** (61.9[1] 60.7[2] 66.8[3] 68.8[4])	0.6654	**0.5596** (42.6[1] 42.7[2] 48.6[3] 50.7[4])
Model-8	165.6	748.1	13h7m	0.7320	0.7063	**0.6660**	0.5578
Model-7	**164.7**	***605.4***	10h9m	**0.7311**	**0.7042**	**0.6650**	**0.5559**
Model-6	95.9	849.8	11h56m	0.7169	0.6876	0.6459	0.5348
Model-5	95.9	849.8	11h17m	0.7114	0.68	0.6409	0.5278
Model-4	95.9	849.8	11h23m	0.6524	0.6523	0.5816	0.4730
Model-3	157.9	466.2	**7h24m**	0.7070	0.6779	0.6361	0.5232
Model-2	157.9	**462.1**	**5h54m**	0.6956	0.6604	0.6247	0.5066
Model-1	157.9	**462.1**	5h56m	0.6372	0.6442	0.5672	0.4596

[a]Other values of Avg Accuracy and mIoU noted 1,2,3,4 are from https://doi.org/10.48550/arXiv.2412.02732. 1, U-Net; 2, Prithvi-EO-1.0-100M; 3, Prithvi-EO-2.0-300M; 4, Prithvi-EO-2.0-600M. the red means the best, and underlined Figures means the second best, and **the bold** means **the third**.

The parameters in our model are slightly increased from 157.9M to 165.8M. However, despite the increased training parameters, LRTACF-NET (Model-9) demonstrates superior mapping performance compared to the other models, highlighting its unique merits. The performance of each class has increased and overall accuracy (mIoU) shows a 10% improvement compared to the foundation model and baseline model (Model-1)[2].

4. Visualization and Discussion

In **Figure 3**, it is obvious that Model-9 (this paper proposed) outperforms Model-1 (baseline) in classifying Corn and Developed/Barren. In the first and second images in **Figure 3**, Model-1 generated false positives by misidentifying 'Corn' pixels as 'Other' or 'Developed/Barren', while Model-9 correctly predicted 'Corn' with higher spectral discrimination accuracy. In order to evaluate the average predicted visualized effect for all models (from Model-1 to Model-9), the correct predicted class was colored with the black, and the incorrect predicted class was colored with the white (as shown in **Figure 4**). Therefore, the blacker means the classification result is better. It is clear that the models in this paper (Model-7~9) are the best, because most of them are black. There are 9 models, which are explained in **Table 1**. In **Figure 4**, the black prediction result denotes completely right classes, while the white means misclassified classes.

4.1. Analysis of Predicted Single Class

To evaluate the accuracy and error in a single crop class for some models, True

[2]https://github.com/ClarkCGA/multi-temporal-crop-classification-baseline.

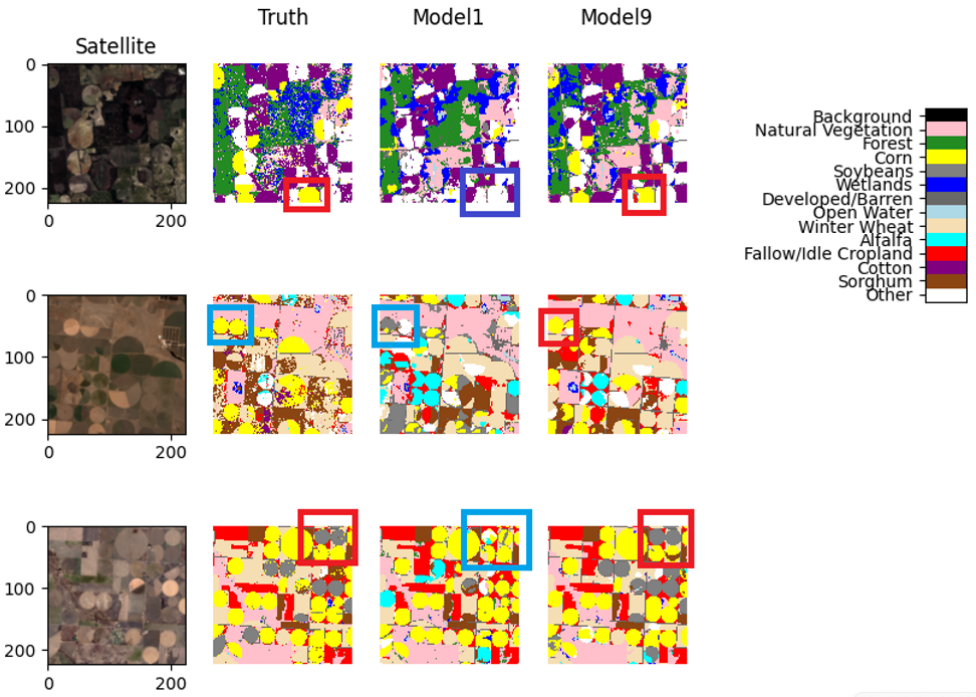


Figure 3. Three 224 * 224 images' prediction.

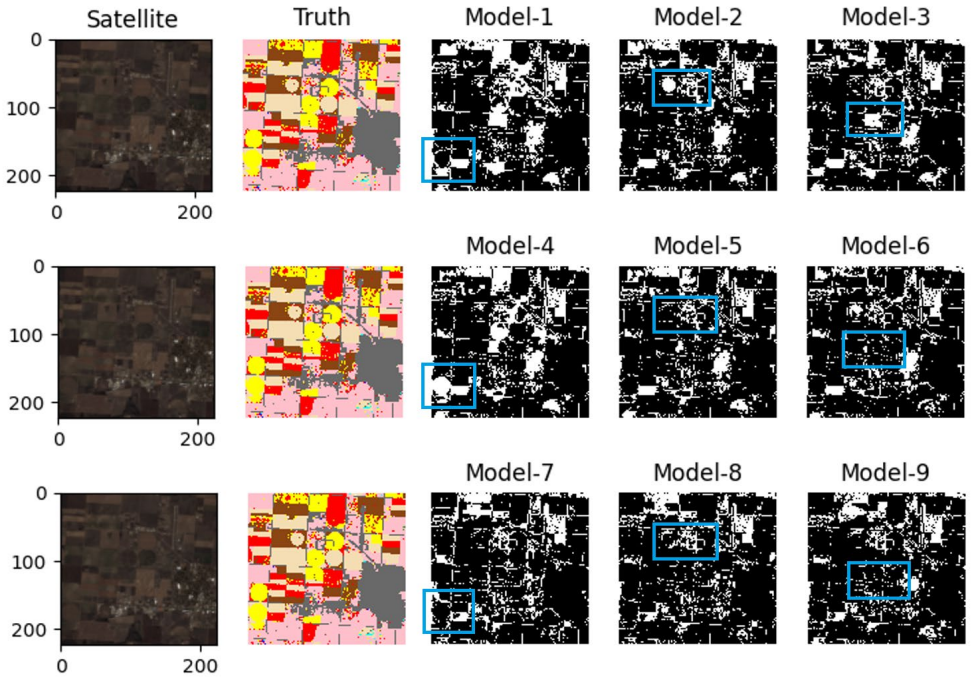
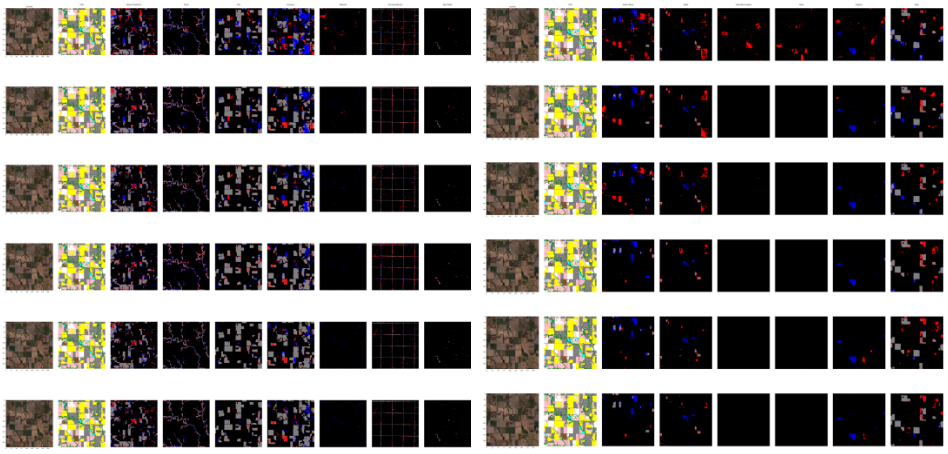


Figure 4. The same 224 * 224 image's prediction of 9 models.

Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN) were considered and visualized. For example, in **Figure 5**, TP is colored with the gray color; TN is colored with the black color; FP is colored with the red color; FN is colored with the blue color.

Figure 5 shows the predicted results with different Models in the same image, that is Model-1 to Model-3 (the first row to third row in **Figure 5**) and Model-7 to Model-9 (the 4th row to 6th row in **Figure 5**), from Class1 to Class7 (Natural Vegetation, Forest, Corn, Soybeans, Wetlands, Developed/Barren, Open Water) in columns of **Figure 5** (a), and from Class8 to Class13 (Winter Wheat, Alfalfa, Fallow/Idle Cropland, Cotton, Sorghum and Other) in **Figure 5** (b).

We can see clearly that there is less blue and less red in our models (last three rows) compared with first three models (first three rows), which means there are fewer FP and FN in all classes in our models (Model-7 to Model-9), especially in Model-9.



(a) (b)

Figure 5. (a) TP, TN, FP, FN results from Class 1 to 7 of the same image, the first column is satellite image, the second column is ground truth; (b) TP, TN, FP, FN results from Class 8 to 13 of the same image, the first column is satellite image, the second column is ground truth.

4.2. Discussion

Compared to Model-5 and Model-2 in **Table 4**, Model-7 demonstrates a significantly higher mIoU, owing to the integration of the Lowest Resolution Temporal Self-Attention (LRTSA) module. This indicates the advantages of Multi-Temporal Self-Attention and the necessity of using SITS as input for crop classification tasks. It also demonstrates the efficacy of SITS in uncovering the phenological cycle of crops. However, three phenological temporal data are far from enough because the vegetation growth cycle comprises seven critical phenological stages: the DOYs (Day Of Years) of greening onset, greening midpoint, maturity, peak, senescence onset, senescence midpoint, and dormancy. In contrast, spectral-temporal features of crops extracted from remote sensing time-series data can significantly improve crop classification accuracy [6].

4.2.1. Confusion Matrix

There exist many challenges in particular in a similar crop even though they have different growth cycle. They present similar spectral signatures or maybe similar

temporal circular, which make models cannot distinguish them from each other well. Taking Alfalfa as an example (in **Figure 6**, row Alfalfa means the ground truth is Alfalfa), Alfalfa can be confused with all other objects, except there is no confusion with Cotton. However, Alfalfa was easily confused with Natural Vegetation (9% in **Figure 6** and 13.9% in **Figure 7**), which means that Model-9 is easier confused between Alfalfa and Natural Vegetation. While Model-9 have fewer confused rate in other classes (except "Other" class) than Model-1, the F1 value in diagonal of Alfalfa in **Figure 7** is higher than **Figure 6**, which means Model-9 has higher performance in recognizing Alfalfa, but prefer misclassifying Alfalfa with Natural Vegetation.

4.2.2. Network Complexity Analysis

Compared with the other models, the proposed LRTACF-NET (Model-9 in **Table 1**) achieves significant improvements in performance. When replacing the

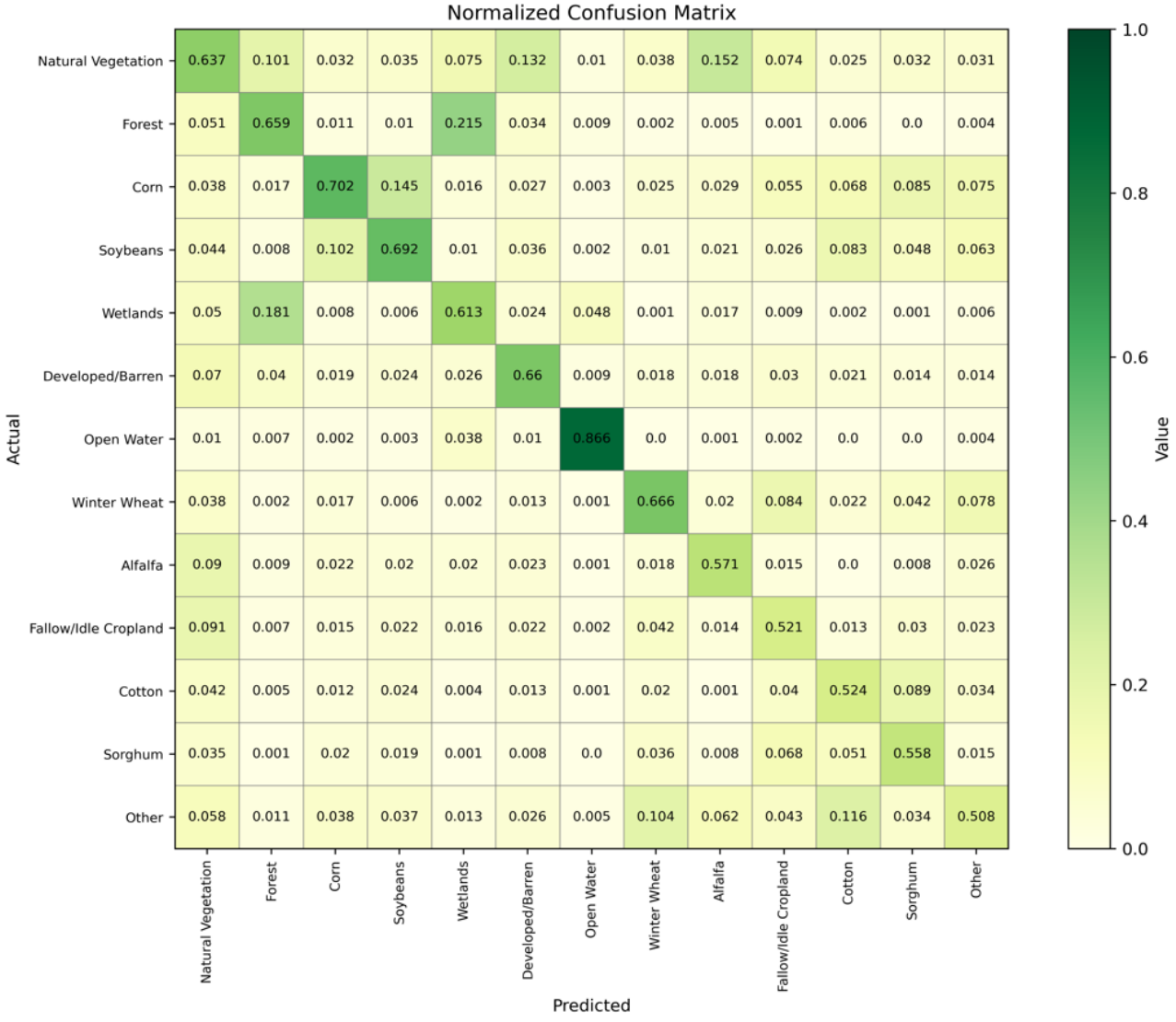


Figure 6. The confusion matrix result (F1 value) of Model-1.

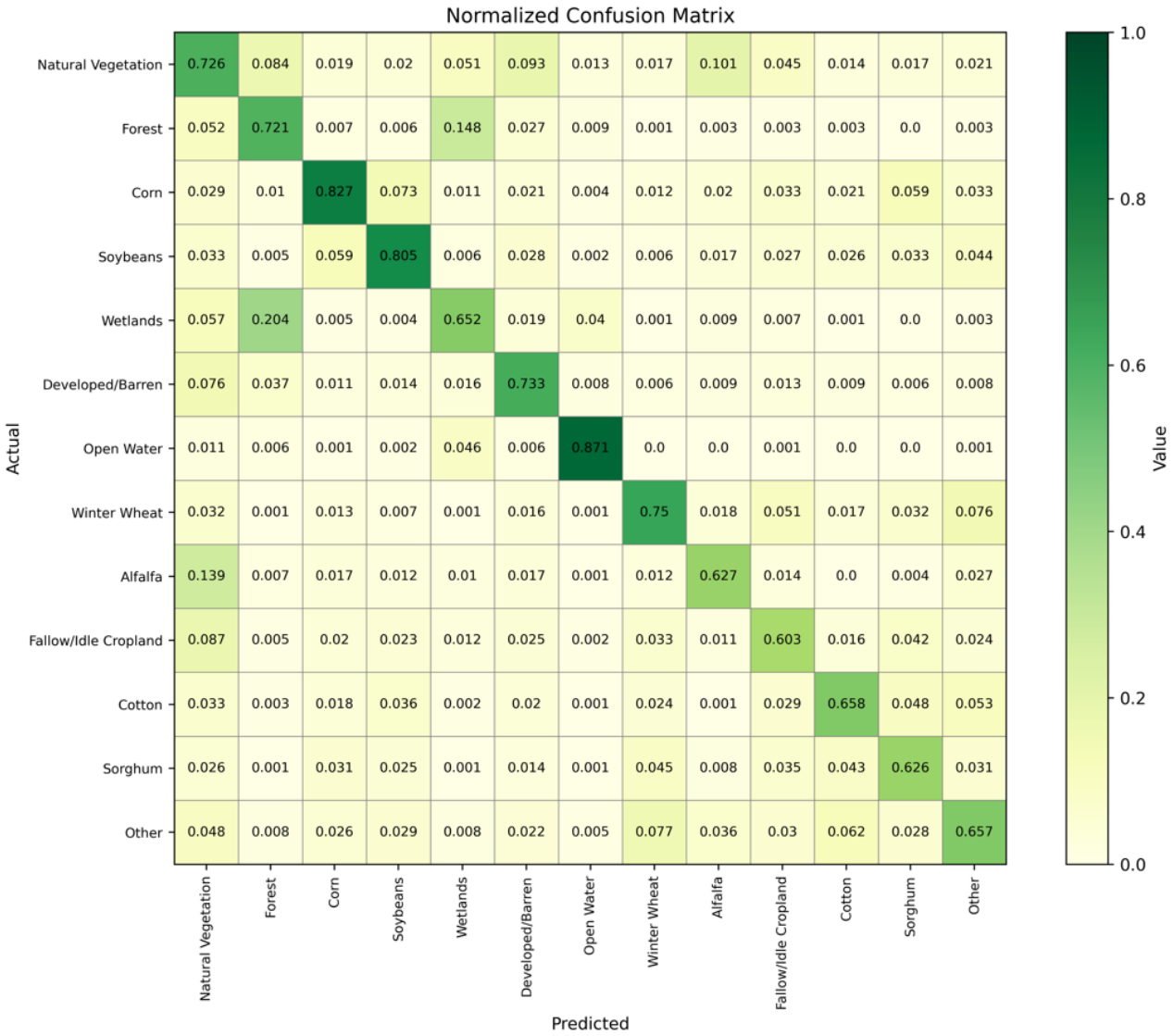


Figure 7. The confusion matrix result (F1 value) of Model-9.

Tversky loss with Weighted-BCE-mIoU, the FLOPs remain unchanged (462.10, see **Table 4**), but the mIoU increases from 45.96 to 50.66. Further incorporating Deep Supervision leads to a slight increase in FLOPs (from 462.10 to 466.15) and a minor rise in inference time (from nearly 6 hours to a little over 7 hours). However, this modification also boosts the mIoU further, reaching 52.32 (see **Table 4**).

In **Table 4**, the proposed method achieves an optimal balance between computational cost and performance (Model-2), delivering higher mIoU with competitive FLOPs. Notably, Model-8 (with 3 Feed-Backs) maintains lower FLOPs than Models 4-6, yet achieves superior mIoU. Although Model-9 (with 5 Feed-Backs) incurs the highest computational cost (FLOPs), it also attains the highest mIoU. Future work should focus on designing a more lightweight model to reduce operational overhead while preserving accuracy.

5. Conclusions

This paper presents LRTACF-NET, a deep supervised U-Net architecture, with substitutional Lowest Resolution Temporal Attention module in skip connection and five Cross-Feedbacks in decoder. It also details the architectures and mathematical formulation for components in this deep learning network model. Through the comparative experiments with some other papers and the ablation experiment of the proposed model, three principal conclusions emerge: 1) The joint loss function combining weighted BCE and IoU demonstrates superior performance improvement (e.g. mIoU 4.7%) over Tversky loss, particularly in addressing class imbalance in crop classification. 2) The LRTSA module effectively utilizes temporal feature even under limited multi-temporal inputs (three periods), achieving 82.66% classification accuracy in corn compared to 69.96% in baseline models. 3) CFMA mechanism elevates crop classification precision by 0.37 percent in mIoU and 0.32 percent in mF1 through multi-times cross-feedback mutual attention and dilation convolution operations. 4) This paper further confirms that temporal or phenological features are essential for distinguishing between spectrally similar crops. However, the resulting computational burden suggests a critical trade-off between high accuracy and computational efficiency.

Future research should focus on the development of lightweight cross-feedback mutual attention architectures capable of reducing false positives and false negatives, enhancing the interpretability of deep learning models, and achieving higher IoU and accuracy through improved edge detection and advanced mathematical approaches. Moreover, to enhance the generalization capability of the proposed model, it is essential to apply it to more diverse datasets, thereby ensuring greater scalability.

Acknowledgements

We would like to thank Lanyun Technology Group Co., Ltd. for providing computational resources and technical support for the model-related works.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Orynbaikyzy, A., Gessner, U., Mack, B. and Conrad, C. (2020) Crop Type Classification Using Fusion of Sentinel-1 and Sentinel-2 Data: Assessing the Impact of Feature Selection, Optical Data Availability, and Parcel Sizes on the Accuracies. *Remote Sensing*, **12**, Article No. 2779. https://doi.org/10.3390/rs12172779

[2] Chen, X., Zhan, Y., Liu, Y., Gu, X., Yu, T., Wang, D., *et al.* (2020) Improving the Classification Accuracy of Annual Crops Using Time Series of Temperature and Vegetation Indices. *Remote Sensing*, **12**, Article No. 3202. https://doi.org/10.3390/rs12193202

[3] Ashourloo, D., Shahrabi, H.S., Azadbakht, M., Rad, A.M., Aghighi, H. and Radiom, S. (2020) A Novel Method for Automatic Potato Mapping Using Time Series of Sen-

tinel-2 Images. *Computers and Electronics in Agriculture*, **175**, Article ID: 105583. https://doi.org/10.1016/j.compag.2020.105583

[4] Wang, H., Chen, X., Zhang, T., Xu, Z. and Li, J. (2022) CCTNet: Coupled CNN and Transformer Network for Crop Segmentation of Remote Sensing Images. *Remote Sensing*, **14**, Article No. 1956. https://doi.org/10.3390/rs14091956

[5] Pech-May, F., Aquino-Santos, R., Rios-Toledo, G. and Posadas-Durán, J.P.F. (2022) Mapping of Land Cover with Optical Images, Supervised Algorithms, and Google Earth Engine. *Sensors*, **22**, Article No. 4729. https://doi.org/10.3390/s22134729

[6] Liu, X., Xie, S., Yang, J., Sun, L., Liu, L., Zhang, Q., *et al.* (2023) Comparisons between Temporal Statistical Metrics, Time Series Stacks and Phenological Features Derived from NASA Harmonized Landsat Sentinel-2 Data for Crop Type Mapping. *Computers and Electronics in Agriculture*, **211**, Article ID: 108015. https://doi.org/10.1016/j.compag.2023.108015

[7] You, N., Dong, J., Huang, J., Du, G., Zhang, G., He, Y., *et al.* (2021) The 10-M Crop Type Maps in Northeast China during 2017-2019. *Scientific Data*, **8**, Article No. 41. https://doi.org/10.1038/s41597-021-00827-9

[8] Tariq, A., Yan, J., Gagnon, A.S., Riaz Khan, M. and Mumtaz, F. (2022) Mapping of Cropland, Cropping Patterns and Crop Types by Combining Optical Remote Sensing Images with Decision Tree Classifier and Random Forest. *Geo-Spatial Information Science*, **26**, 302-320. https://doi.org/10.1080/10095020.2022.2100287

[9] Yan, S., Yao, X., Zhu, D., Liu, D., Zhang, L., Yu, G., *et al.* (2021) Large-Scale Crop Mapping from Multi-Source Optical Satellite Imageries Using Machine Learning with Discrete Grids. *International Journal of Applied Earth Observation and Geoinformation*, **103**, Article ID: 102485. https://doi.org/10.1016/j.jag.2021.102485

[10] Ronneberger, O., Fischer, P. and Brox, T. (2015) U-net: Convolutional Networks for Biomedical Image Segmentation. In: Navab, N., Hornegger, J., Wells, W. and Frangi, A., Eds., *Medical Image Computing and Computer-Assisted Intervention—MICCAI* 2015: 18*th International Conference*, Munich, 5-9 October 2015, 234-241.

[11] Jia, X., Wang, W., Zhang, M. and Zhao, B. (2025) Atten-Nonlocal Unet: Attention and Non-Local Unet for Medical Image Segmentation. *Computers in Biology and Medicine*, **191**, Article ID: 110129. https://doi.org/10.1016/j.compbiomed.2025.110129

[12] Cai, J., Shi, J., Leau, Y., Meng, S., Zheng, X. and Zhou, J. (2024) Res50-SimAM-ASPP-Unet: A Semantic Segmentation Model for High-Resolution Remote Sensing Images. *IEEE Access*, 12, 192301-192316. https://doi.org/10.1109/access.2024.3519260

[13] Chen, H., Qiu, Y., Yin, D., Chen, J., Chen, X., Liu, S., *et al.* (2022) Stacked Spectral Feature Space Patch: An Advanced Spectral Representation for Precise Crop Classification Based on Convolutional Neural Network. *The Crop Journal*, **10**, 1460-1469. https://doi.org/10.1016/j.cj.2021.12.011

[14] Fried, O., Avidan, S. and Cohen-Or, D. (2017) Patch2vec: Globally Consistent Image Patch Representation. *Computer Graphics Forum*, **36**, 183-194. https://doi.org/10.1111/cgf.13284

[15] Jean, N., Wang, S., Samar, A., Azzari, G., Lobell, D. and Ermon, S. (2019) Tile2vec: Unsupervised Representation Learning for Spatially Distributed Data. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 3967-3974. https://doi.org/10.1609/aaai.v33i01.33013967

[16] Garnot, V.S.F., Landrieu, L., Giordano, S. and Chehata, N. (2020) Satellite Image Time Series Classification with Pixel-Set Encoders and Temporal Self-Attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*,

Seattle, 13-19 June 2020, 12325-12334.

[17] Garnot, V.S.F. and Landrieu, L. (2020) Lightweight Temporal Self-Attention for Classifying Satellite Images Time Series. In: *Advanced Analytics and Learning on Temporal Data*: 5*th ECML PKDD Workshop*, Springer Nature, Vol. 12588, 171-181.

[18] Stergioulas, A., Dimitropoulos, K. and Grammalidis, N. (2022) Crop Classification from Satellite Image Sequences Using a Two-Stream Network with Temporal Self-attention. 2022 *IEEE International Conference on Imaging Systems and Techniques* (*IST*), Kaohsiung, 21-23 June 2022, 1-6. https://doi.org/10.1109/ist55454.2022.9827752

[19] MacDonald, E., Jacoby, D. and Coady, Y. (2024) VistaFormer: Scalable Vision Transformers for Satellite Image Time Series Segmentation. https://arxiv.org/abs/2409.08461

[20] Jakubik J., Roy S., Phillips C.E., Fraccaro P., Godwin D., Zadrozny B., *et al.* (2023) Foundation Models for Generalist Geospatial Artificial Intelligence. https://arxiv.org/abs/2310.18660

[21] Xie Y.Q., Wang Z.H., Chen W.Y., Li Z.L., Jia X.W., Wang R.C., *et al.* (2024) When are Foundation Models Effective? Understanding the Suitability for Pixel-Level Classification Using Multispectral Imagery. https://arxiv.org/abs/2404.11797

[22] Hsu, C., Li, W. and Wang, S. (2024) Geospatial Foundation Models for Image Analysis: Evaluating and Enhancing NASA-IBM Prithvi's Domain Adaptability. *International Journal of Geographical Information Science*. https://doi.org/10.1080/13658816.2024.2397441