Estuaries and Coasts (2025) 48:72
https://doi.org/10.1007/s12237-025-01498-0

A Modified Swin-UNet Model for Coastal Wetland Classification Using Multi-Temporal Sentinel-2 Images

Binyu Wang[1] · Yuanheng Sun[1] · Xueyuan Zhu[1] · Senlin Teng[1] · Ying Li[1]

Received: 7 October 2024 / Revised: 14 January 2025 / Accepted: 27 January 2025
© The Author(s), under exclusive licence to Coastal and Estuarine Research Federation 2025

Abstract

Coastal wetlands are of great importance in protecting biodiversity, mitigating climate change, and providing natural resources. Using deep learning methods for the classification and mapping of coastal wetlands with optical remote sensing data can effectively monitor changes in wetlands, playing a crucial role in their protection. However, most current wetland classification methods focus on single-temporal data, with relatively few studies addressing multi-temporal data. Therefore, for the wetland classification task in the Bohai Rim region of China, an improved Swin-MTNet model based on the state-of-the-art deep learning model Swin-UNet is proposed in this study to better capture temporal feature variations with multi-temporal Sentinel-2 imagery. The Swin-MTNet is compared with Swin-UNet and DeepLabV3+, and the results indicate that Swin-MTNet achieves overall accuracy improvements of 5.12% and 2.85% and Kappa coefficient improvements of 6.85% and 3.86% over Swin-UNet and DeepLabV3+, respectively, when utilizing multi-temporal data. The classification improvement for *Spartina alterniflora* is the most significant, with F1 scores increasing by 0.45 and 0.47 compared to Swin-UNet and DeepLabV3+, respectively. These results demonstrate that the proposed Swin-MTNet model can effectively leverage the temporal features of multi-temporal data, significantly improving the accuracy of coastal wetland classification.

Keywords Wetland classification · Deep learning · Multi-temporal · Swin-UNet · Sentinel-2

Introduction

Coastal wetlands, as the intersection of terrestrial and marine ecosystems, harbor valuable ecological resources. Coastal wetlands provide natural habitats for numerous wildlife species (Barbier et al., 2011; Jamali and Mahdianpari, 2022a; Mao et al., 2020) and play important roles in regulating climate, conserving water resources and preventing soil erosion (Jamali et al., 2021a; Mahdianpari et al., 2018; Mao et al., 2021). However, due to the rise in human activities, the area of coastal wetlands continues to diminish. (Asselen et al., 2013; Mahdianpari et al., 2017; McCarthy et al., 2018; Mohammadimanesh et al., 2018). Globally, the phenomenon of coastal wetland loss is worsening year by year, posing significant challenges to ecosystems and human society (Chen et al., 2019; Mao et al., 2021).

High-resolution wetland classification products can effectively protect wetlands and help in land resource planning by providing precise and detailed information on wetland coverage (Zhu and Gong, 2014). Nevertheless, producing high-resolution wetland classification products is not an easy task (Mahdianpari et al., 2021). The temporal variability of wetland ecosystems and the remote geographical locations make traditional methods such as field surveys highly resource-intensive (Evans and Costa, 2013; Mahdianpari et al., 2017). Remote sensing technology has the characteristic of covering large areas and can quickly acquire data over extensive regions. Compared to traditional ground survey methods, remote sensing monitoring offers significant advantages in terms of time and cost (Amani et al., 2021; Mahdianpari et al., 2018). Multispectral remote sensing data is the most widely used data for wetland classification due to its high spatial resolution and ease of access, making it highly favored by many researchers. Sentinel-2 is a high-resolution multispectral imaging satellite equipped with a multispectral imager (MSI) at an altitude of 786 km,

Communicated by Paul A. Montagna

✉ Yuanheng Sun
yhsun@dlmu.edu.cn

[1] Environmental Information Institute, Navigation College, Dalian Maritime University, Dalian 116026, China

Published online: 24 February 2025

covering 13 spectral bands with a swath width of 290 km (Chen et al., 2024c). Sentinel-2 observation data features high ground resolution and, compared to traditional satellite data, additionally provides red-edge and shortwave infrared bands. These extra bands better capture vegetation characteristics, facilitating the classification of wetland vegetation (Saad El Imanni et al., 2022). This makes it an ideal satellite image to produce high-resolution wetland classification products and monitor changes in coastal wetland categories.

In traditional coastal wetland classification methods, machine (Bennett, 1994; Berhane et al., 2018; Chen et al., 2024a; Cortes and Vapnik, 1995) are widely used due to their advantages of low computational resource requirements and good performance with small datasets. However, as the data scale increases, traditional methods may struggle to handle high-dimensional data and large sample sizes. Compared to shallow models in machine learning, deep learning algorithms can typically process more complex data (Jamali et al., 2021b; Mohammadimanesh et al., 2019), making deep learning methods, especially Convolutional Neural Networks (CNNs), increasingly popular for large-scale wetland classification tasks (Hosseiny et al., 2022; Jamali et al., 2022; Lou et al., 2024).

However, this structured network reaches a bottleneck. The development of network architectures in the field of Natural Language Processing (NLP) has led researchers to see another direction. Liu et al. proposed a universal Transformer backbone called Swin Transformer (Z. Liu et al., 2021c). It has achieved great success as a backbone feature extraction network and has shown its advantages in the field of wetland classification in remote sensing images (Bazi et al., 2021; He et al., 2020; Hong et al., 2022; Jamali et al., 2023). Jamali et al. discovered in their study on wetland classification in Canada that the Swin Transformer achieved higher average accuracy compared to CNN models like AlexNet and VGG-16 (Jamali and Mahdianpari, 2022b). Additionally, by integrating VGG-16, 3D CNN, and Swin Transformer models, the classification performance was further enhanced. Cao et al. leveraged Swin Transformer to develop a U-shaped network structure called Swin-UNet (Cao et al., 2022). The studies mentioned above have already indicated the effectiveness of the Swin-UNet model in classification tasks (Hao et al., 2024; Xiao et al., 2023; Yao and Jin, 2022). However, these studies have focused on limited regions and scenarios, warranting further validation of its classification performance.

Moreover, most researches involving remote sensing wetland classification only explore the performance of models based on single-temporal image, no matter with CNN models or Transformer models, with few utilizing multi-temporal datasets. Chen et al. use multi-temporal data to extract remote sensing information of coastal aquaculture ponds in the Zhoushan Archipelago from 1984 to 2022 (Chen et al., 2024c), while Yang et al. apply multi-temporal data for large-scale land cover classification tasks (Yang et al., 2022). Wetland land cover may exhibit different characteristics in different seasons; for example, seasonal variations in wetland vegetation spectra and water levels may affect wetland classification results. Therefore, multiple temporal data not only help to timely detect changes in wetland ecosystems, providing important information for ecological environment protection and management but also assist models in better understanding such temporal changes, thereby improving classification accuracy. However, most studies based on multi-temporal data simply input the multi-temporal data into existing mature models to perform tasks such as classification. For example, Piaser and Villa (2023) only used spectral indices (SI) synthesized from multi-temporal Sentinel-2 data as input to established machine learning models to compare classification accuracy; Bill et al. (2024) utilized multi-temporal Landsat imagery, feeding each year's image separately into a support vector machine model for analysis. Current studies (Alam and Hossain, 2024; Chen et al., 2023; Moharrami et al., 2024) have mainly focused on applying existing models to multi-temporal data without considering the potential to modify these models to better capture the temporal dynamics and interactions in the data. Specifically, they have not considered how to adjust the model architecture to integrate multi-temporal features more effectively, which is critical to improving classification accuracy in wetland classification. How to incorporate this temporal information into the network model becomes an essential problem to be solved.

To address the current gap in research on utilizing multi-temporal information for wetland classification, this paper aims to propose an improved multi-temporal network model, naming Swin-MTNet, based on the Swin-UNet model, for wetland classification tasks. The technological innovation of this model lies in modifying the Swin-UNet architecture and introducing modules for learning temporal feature information and enhancing feature fusion, allowing the model to effectively capture information from the temporal dimension and improve wetland classification accuracy.

Material

Study Area

Bo Hai Coastal Region in Northern China, especially the Liao River estuary coastal wetland and the Yellow River delta coastal wetland which compress most coastal wetland land cover types, are chosen as our study areas. The geographical coordinates of the Liao River estuary wetland range from 121°28′ to 121°58′ east longitude and from 40°45′ to 41°05′ north latitude. Located in the jurisdiction of Panjin City, Liaoning Province, in the northeastern part of Liaodong Bay, and the central part of the Liao River delta.

The geographical coordinates of the Yellow River Delta wetland range from 118°32′ to 119°20′ east longitude and from 37°34′ to 38°12′ north latitude. It is located in the northeastern part of Dongying City, Shandong Province. Positioned on the south coast of Bohai Bay and the western part of Laizhou Bay, the wetland area spans 450,000 hectares in a fan-shaped distribution. The wetland types above all being coastal wetlands, and the wetland vegetation categories in the ecosystems are basically the same, including *Suaeda salsa*, *Spartina alterniflora*, and *Phragmites australis*, making mapping possible for the three regions together (Fig. 1).

Datasets

The remote sensing imagery data used in this study consists of Sentinel-2 satellite imagery downloaded from the Google Earth Engine (GEE) platform and Digital Elevation Model (DEM) data downloaded from the Earthdata platform. Sentinel-2 Level-2A surface reflectance products are employed in this study. The Level-2A data undergo spatial and atmospheric corrections, providing more accurate surface information suitable for land cover classification research. Sentinel-2 satellite imagery for the coastal wetlands of the Liao River estuary and the Yellow River delta in May, June, September, and November 2022 is selected. The specific band information used is presented in Table 1. These four months encompass a complete cycle of vegetation growth, maturity, and senescence, providing more comprehensive vegetation feature information. The DEM data utilized is the ASTER Global Digital Elevation Model V003 with a global spatial resolution of 30 meters, released in June 2019.

Methodology

Data Preprocessing

Sentinel-2 satellite imagery undergoes cloud and cloud shadow removal using the Quality Assessment (QA) band. Images for specified months are composited using a

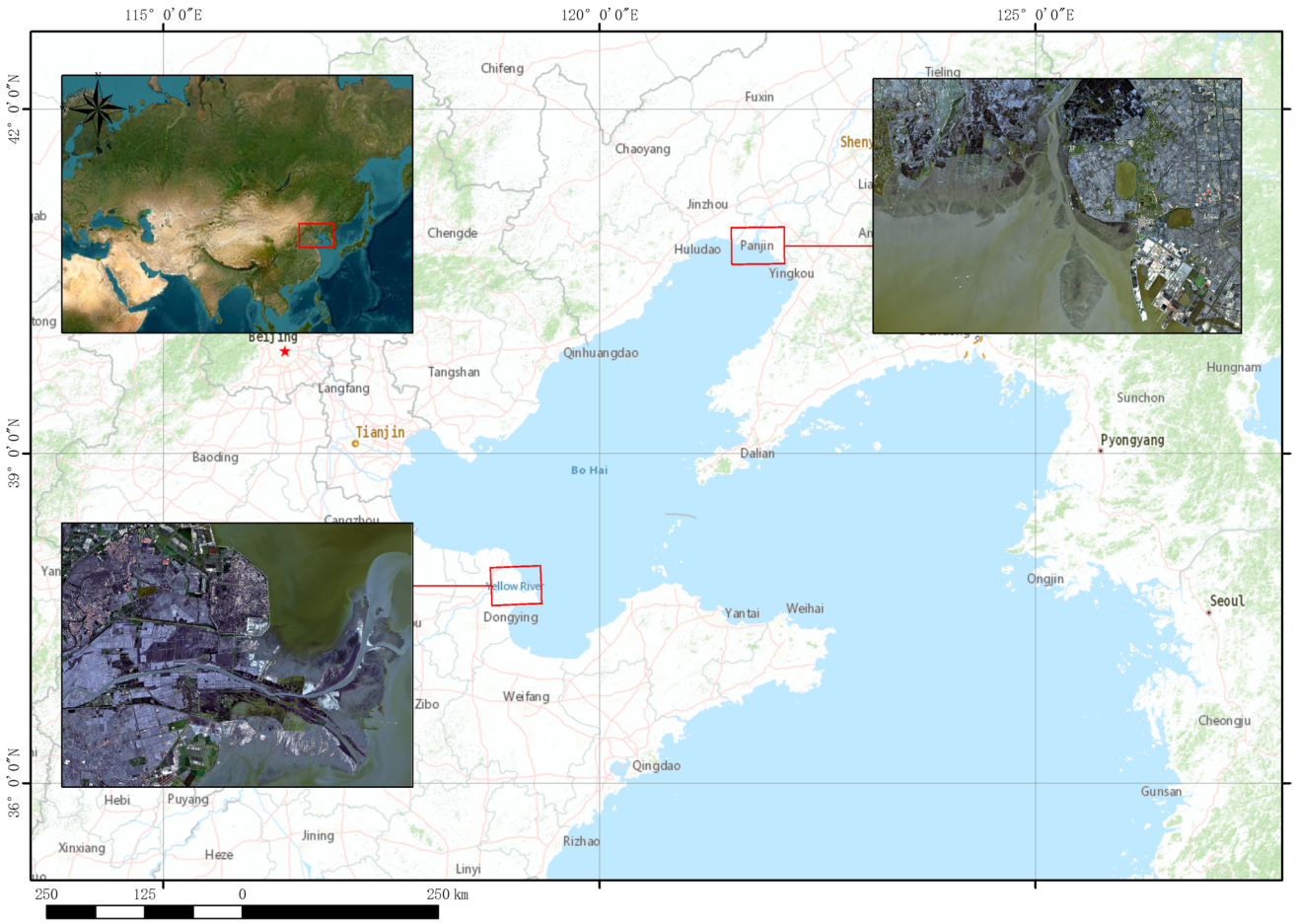


Fig. 1 Schematic map of the Bohai Rim study area (the top left corner is an overview of the East Asia region, the top right corner is a true-color composite remote sensing image of the Liao River estuary coastal wetland, and the bottom left corner is a true-color composite remote sensing image of the Yellow River delta wetland)

Tabel 1 Sentinel-2 band information

Bands	Wavelength (nm)	Resolution (m)
B2 (blue)	458–523	10
B3 (green)	543–578	10
B4 (red)	650–680	10
B5 (red-edge 1)	698–713	20
B6 (red-edge 2)	733–748	20
B7 (red-edge 3)	773–793	20
B8 (NIR)	785–900	10
B11 (SWIR 1)	1565–1655	20
B12 (SWIR 2)	2100–2280	20

median operation. Both the spectral bands listed in Table 1 and the DEM data are resampled to a 20-m resolution to maintain consistent resolution across all bands. Finally, the imagery data and DEM data are clipped to the size of the study area.

The label is generated by manual visual interpretation with Sentinel-2 images in 2022. The land cover across the Bo Hai Coastal Region is classified into nine categories: *Suaeda salsa*, *Spartina alterniflora*, *Phragmites australis*, mudflat, natural water bodies, farmland, reservoirs, aquaculture farm, and impervious surfaces. Table 2 shows the image features for each category in different months. For areas where the category is difficult to determine, high-resolution images in Google Earth are used as references to ensure the labels are as close to reality as possible.

All Sentinel-2 images and the corresponding label are cropped into 128 × 128 using a sliding window. The cropped dataset is augmented using horizontal and vertical flips, increasing the total number of images to 5000. After shuffling the dataset, it is split into training, validation, and testing sets in a 6:2:2 ratio. Additionally, manual selection is performed to ensure that each dataset contains all classification categories. As mainstream network models typically accept single-temporal inputs, whereas multi-temporal inputs are accepted by our proposed model, four different datasets are constructed to explore the model's ability to learn temporal information during the classification. A detailed description of these datasets is provided in Table 3.

Proposed Multi-temporal Network for Classification

Structure of Proposed Network

The Swin-UNet (Cao et al., 2022) has demonstrated tremendous potential in remote sensing image segmentation, thus is adopted as the baseline network in this study. To enable the model to learn information from different time phases, the encoder of Swin-UNet is designed as four independent branches. The four branches operate independently, with each pathway composed of four Swin Transformer blocks, resulting in a network depth of 4. The input data passes through the Patch Extract layer and the Patch Embedding layer for encoding window size and position. Subsequently, it enters the Swin Transformer block for feature extraction. The features extracted from the four branches are concatenated and input into the CSAM for learning across the temporal dimension channels. The features learned through time dimension are downsampled using the Patch Merging layer, and then proceed through Swin Transformer blocks and CSAM. This process is repeated three times to form the backbone feature extraction network. The SAFF module is used to implement skip connections in the network. It is used to fuse the features outputted by each layer of the Swin Transformer blocks with the features outputted by the CSAM at the same layer of the feature extraction network. The Patch Expanding layer is then utilized to upsample the features, resulting in features that are consistent in size with the input image. This network structure is named as Swin-MTNet in this work, and its structure is shown in Fig. 2. To mitigate overfitting, we incorporate regularization and dropout techniques into the Swin-MTNet. Specifically, we apply L2 regularization during training to reduce the model's over-dependence on the training data, and introduce Dropout layers to further enhance the model's robustness. As a result, this approach effectively reduces the number of training parameters, with the total number of parameters in the model being 2,302,465.

Swin-MTNet can be represented as:

$$f_{SMTN} = f_{dec}(f_{CSAM}(f_{enc}(x_1^N), \ldots, f_{enc}(x_4^N))) \quad (1)$$

where the input data is x_1, x_2, x_3, x_4, each image has N bands, f_{enc} represents the encoder, f_{dec} represents the decoder. SAFF is used for feature fusion in the decoder part.

Channel Self-Attention Model

The Swin Transformer, by design, is not capable of capturing temporal information as it mainly focuses on extracting features from individual time steps. The CSAM, however, plays a crucial role in learning and incorporating temporal information. CSAM is a weight feature matrix with the same size as the original feature matrix, which uses a global self-attention mechanism to allocate weights to combination feature images concatenated at different times and to local feature extraction. It is responsible for assigning weights to the features extracted by the Swin Transformer at different

Table 2 Typical false-color composite image of Sentinel-2 for each category in different months of 2022

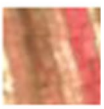
Categories	May	June	September	November
Suaeda salsa (SS)				
Spartina alterniflora (SA)				
Phragmites australis (PA)				
mudflat (MU)				
natural water bodies (NWB)				
farmland (FA)				
reservoirs (RE)				
aquaculture farm (AF)				
impervious surfaces (IS)				

Table 3 Detailed description of four constructed datasets

No.	Name	Input block dimension	Input band description
Dataset I	Single-temporal dataset	(128, 128, 10)	Sentinel-2 imagery with 9 bands and DEM data for September.
Dataset II	Single-temporal augmented dataset	(128, 128, 10)	Sentinel-2 imagery with 9 bands and DEM data for May, June, September, and November.
Dataset III	Band-stacked dataset	(128, 128, 37)	The Sentinel-2 imagery with 9 bands for May, June, September, and November stacked together, along with DEM data, forming a 37-band image.
Dataset IV	Multi-temporal dataset	(128, 128, 10, 4)	Adding the temporal dimension, the Sentinel-2 imagery with 9 bands and DEM data for May, June, September, and November are respectively placed in the last dimension.

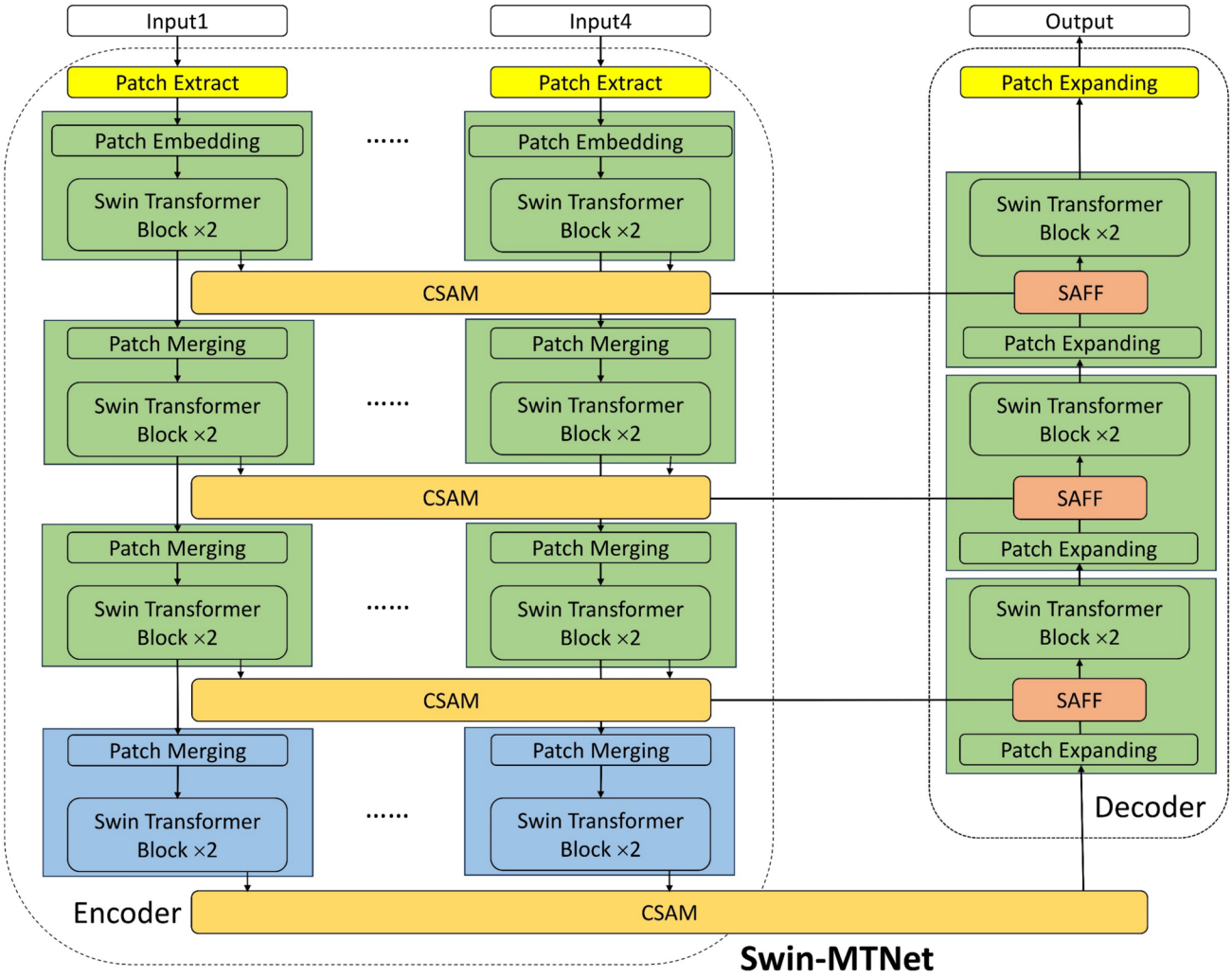


Fig. 2 Structure of Swin-MTNet network

time and selecting the most valuable temporal information. Channels with significant temporal changes are assigned higher weights, while those with minor changes are assigned lower weights.

The construction of the CSAM is illustrated in Fig. 3. The features from multiple encoder branches are concatenated and then enter two branches:

(1) Branch for extracting global attention. First, the concatenated features from different time steps are fed into a global average pooling layer, which reduces the dimensions to one-dimensional to capture the global information. Then, the features are passed through a 1×1 convolutional layer to extract global features, reducing the number of channels by a factor of r. Next, the features undergo Batch Normalization (Ioffe and Szegedy, 2015) and ReLU activation for non-linear activation. Another 1×1 convolution operation followed by Batch Normalization is applied to deepen the feature extraction and restore the number of channels to their original count. Finally, the global feature map is fed into the self-attention mechanism to enable each channel to participate in the computation, obtaining global features. The self-attention mechanism SA(x) is represented as:

$$\mathrm{SA}(x) = softmax\left(\frac{\mathrm{Q}(x)\mathrm{K}(x)^{\mathrm{T}}}{\sqrt{dk}}\right)\mathrm{V}(x) \tag{2}$$

where Q, K, and V matrices are all obtained from x. First, we compute the dot product between Q and K, and to prevent the result from being too large, we divide by the scale standard $\sqrt{dk}$. Then, we use

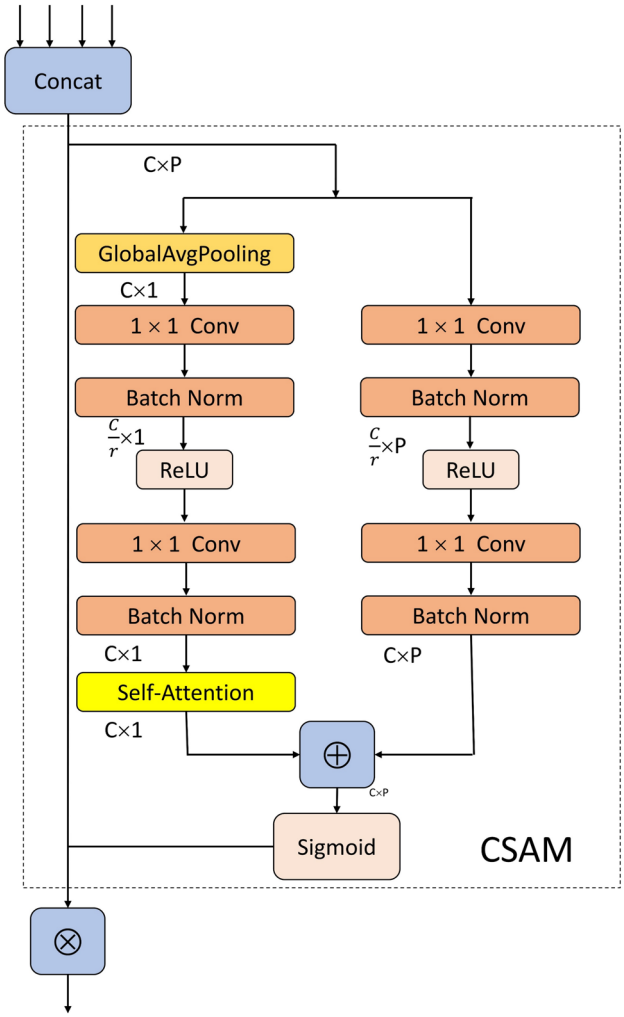


Fig. 3 Channel self-attention model

softmax to normalize the result into a probability distribution, and multiply by matrix V to get the weighted sum representation.

(2) Branch for extracting local attention. The global average pooling layer and the self-attention module in the branch responsible for extracting global attention were removed, limiting it to learning only local features.

The features from the two branches are then added together, fusing the global and local information. Additionally, the locally extracted features have the same shape as the input features, allowing the preservation and enhancement of fine details in the lower-level features. The fused features are transformed using a sigmoid function to obtain the final feature weights. These feature weights are applied to the combined features, resulting in the final weighted fused features.

The CSAM module can be represented as:

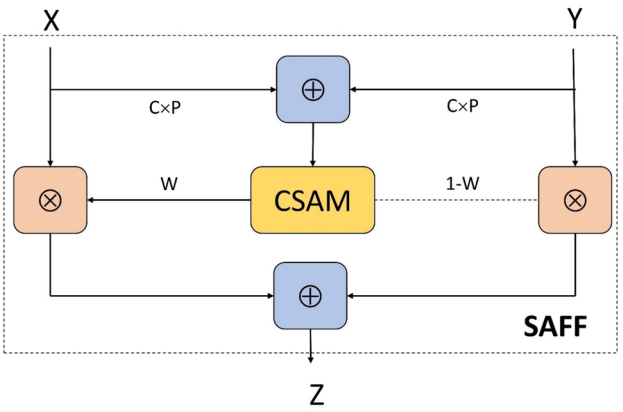


Fig. 4 Self-attention feature fusion

$$f_{CSAB}(x) = f_{sigmoid}(\mathrm{SA}(\mathrm{W}_2^1(f_{ReLU}(\mathrm{W}_1^1(f_{GAP}(x_{concat}))))) + \mathrm{W}_2^1(f_{ReLU}(\mathrm{W}_1^1(x_{concat})))) \quad (3)$$

Self-Attention Feature Fusion

The Swin-MTNet model is based on the U-shaped encoder-decoder network, where skip connections are an essential part. Traditional skip connections typically concatenate low-level and high-level features followed by a convolution operation. This approach only captures local information, whereas the CSAM module can extract both global and local features. Therefore, we improved CSAM into two feature fusion modules called SAFF to enhance the model's feature learning capability and improve classification accuracy.

As shown in Fig. 4, the SAFF structure works by adding two input features, X and Y, to obtain fused features, which are then passed through the CSAM for feature weight allocation, resulting in the weight matrix W. Since the feature weights in CSAM are output by a sigmoid function, the W matrix contains values between 0 and 1. This ensures that both W and 1−W are positive, allowing the network to perform a soft selection or weighted averaging between X and Y (Dai et al., 2021). Through SAFF, the network can better explore the connections between shallow and deep network information, selectively extracting more valuable information from both shallow and deep layers.

The SAFF module can be represented as:

$$f_{SAFF}(\mathrm{X}, \mathrm{Y}) = (X \otimes f_{CSAM}(X \oplus Y)) \oplus (Y \otimes (1 - f_{CSAM}(X \oplus Y))) \quad (4)$$

Model Training Settings

The TensorFlow 2.0 framework is adopted for building and training the Swin-MTNet model and other image segmentation models. The model training is operated on a 64-bit

Windows 11 system, with an AMD Ryzen 7 5800H processor (CPU), a graphics processing unit (GPU) with 6 GB of memory from NVIDIA GeForce RTX 3060, and 16 GB of random-access memory (RAM). We utilize the Adam optimizer with an initial learning rate of 1×10^{-5}. During training, a warm-up strategy was employed to dynamically adjust the learning rate. The model with the lowest validation set loss during training is saved, and the batch size is set to 16. The loss function combines focal loss and dice loss.

Comparison and Evaluation

Evaluation Metrics

To evaluate the performance of the Swin-MTNet model and its classification effectiveness for wetland classification, precision, overall accuracy, recall rate, F1 score, and Kappa coefficient are adopted to evaluate the model. Accuracy, recall rate, and F1 score reflect the model's ability to correctly classify each coastal wetland category, while overall accuracy reflects the model's ability to correctly classify all categories. The Kappa coefficient can well reflect the accuracy of the model in classifying small sample categories. The formulas of abovementioned metrics are as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Overall Accuracy} = \frac{\text{TP} + \text{TN}}{\text{Total number of pixels}} \times 100 \tag{6}$$

$$Recall = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{F1} - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

$$\text{Kappa} = \frac{P_o - P_e}{1 - P_e}, P_o = \frac{\sum x_{ii}}{N}, P_e = \frac{\sum x_{i+} x_{i+}}{N^2} \tag{9}$$

In Eqs. (5) to (7), where TP, FP, and FN represent the number of pixels correctly identified as the foreground, incorrectly identified as the foreground, and wrongly identified as the background, respectively. In Eq. (9), N represents the sum of all elements in the confusion matrix; $\sum x_{ii}$ represents the sum of diagonal elements in the confusion matrix, which is the correct elements; x_{i+} represents the sum of elements in the row i of the matrix.

Comparing Schemes

To investigate whether multi-temporal information can improve the classification accuracy of coastal wetlands and whether the Swin-MTNet model can better utilize temporal information compared to mainstream models with a large number of citations, Swin-UNet and DeepLabV3+ (Chen et al., 2018) are selected as comparative models. The comparative scheme is designed as shown in Table 4. Swin-UNet, as the optimized model, has been shown in a few studies to be effective for (Hao et al., 2024; Xiao et al., 2023), and the comparative results can better reflect the optimization effect. DeepLabV3+, as a popular deep learning model, has demonstrated excellent classification results in many fields, and many researchers use it as the basis for their studies, providing a powerful benchmark for evaluating the performance of the models (Gonzalez-Perez et al., 2022; M. Liu et al., 2021a, 2021b).

Scheme 1 uses mainstream models to classify coastal wetlands with the single-temporal dataset introduced in the "Data Preprocessing" section, providing only the imagery information from September when the vegetation is at its growing peak. This scheme investigates the classification capability of the models on single-temporal data. Schemes 2 and 3 examine the classification capabilities of mainstream models with different forms of multi-temporal information. Scheme 2 uses the single-temporal expanded dataset described in the "Data Preprocessing" section, where each sample lacks multi-temporal information, but the dataset contains images from four different months. Each input image is from a single temporal instance. Scheme 3 uses the band-stacked multi-temporal dataset described in the "Data Preprocessing" section. Each sample contains all band information from four different temporal instances, forming a multi-temporal dataset. These schemes help explore the mainstream models' ability to learn from different forms of multi-temporal information. Scheme 4 is an ablation study for Swin-MTNet. By comparing the classification results with and without the CSAM and SAFF modules, we investigate the role of these modules in learning temporal information and their contribution to classification accuracy.

Results

Ablation Study

Our proposed Swin-MTNet model introduced the CSAM module and SAFF module to learn temporal features and enhance feature fusion. To better evaluate the impact of these modules and the structure on improving the classification accuracy, we conducted ablation experiments first. The comparison results of Swin-MTNet, i.e., Schemes 4, on the test dataset are shown in Table 5. The overall accuracy (OA) for model using only the Swin-MTNet four-branch model, as in Scheme 4's Swin-MTNet-ECS, reached 92.53%, and the Kappa coefficient is 90.22%. As the CSAM is added in

Table 4 Comparison schemes

Scheme	Input dataset	Model
Scheme 1	Single-temporal dataset	Swin-UNet DeepLabV3+
Scheme 2	Single-temporal augmented dataset	Swin-UNet DeepLabV3+
Scheme 3	Band-stacked dataset	Swin-UNet DeepLabV3+
Schemes 4 (Ablation Study)	Multi-temporal dataset	Swin-MTNet Swin-MTNet-ES Swin-MTNet-ECS

Swin-MTNet represents Swin-MTNet with both the CSAM and SAFF modules. Swin-MTNet-ES represents the Swin-MTNet model excluding the SAFF module, and Swin-MTNet-ECS represents the Swin-MTNet model excluding both the CSAM and SAFF modules

the branch fusion part, as in Scheme 4's Swin-MTNet-ES, the model is enabling to better learn temporal features. By filtering temporal features, the accuracy of most categories improved compared to Swin-MTNet-ECS, with the OA and Kappa coefficient increasing by 0.75% and 0.97%, respectively. Especially for the classification of *Spartina alterniflora*, the F1 score increased by 8%. Furthermore, the SAFF module is added to the skip connections, as in Scheme 4's Swin-MTNet, to better capture the relationship between shallow and deep information. This resulted in further improved accuracy compared to Swin-MTNet-ES without the SAFF module, achieving an OA of 93.83% and a Kappa coefficient of 91.91%.

It is evident from Table 5 that the accuracy for *Spartina alterniflora*, *Phragmites australis*, impervious surfaces, and aquaculture significantly improved with the addition of the CSAM and SAFF modules, with F1 scores increasing by 15%, 3%, 4%, and 3% respectively. The accuracy for *Suaeda salsa*, natural water bodies, mudflat, farmland, and reservoirs showed little change, with improvements less than 3%. Without using CSAM to learn temporal features, the Swin-MTNet-ECS model tends to misclassify *Spartina alterniflora* as mudflat. This is because *Spartina alterniflora* does not grow during certain months, and the model fails to focus on the characteristics of the growing months, thus predicting them as mudflat. The Swin-MTNet model with CSAM reduced the misclassification of *Spartina alterniflora* as mudflat by more than half. *Phragmites australis* have features similar to farmland, leading to many *Phragmites australis* pixels being misclassified as farmland. However, the slightly different growing cycles between the two allow the Swin-MTNet model to better distinguish them, improving classification accuracy. *Suaeda salsa* has the lowest F1 score due to being a small sample with few pixels and scattered distribution, resulting in lower classification accuracy. Impervious surfaces experience the most significant misclassification and omission errors. This is due to the large sample size and the presence of *Phragmites australis* and other vegetation around buildings or roads, making it difficult to distinguish them in 20-m resolution remote sensing images.

Comparison with Other Methods

The best-performing model from Scheme 4 in Table 4, Swin-MTNet, is selected as the representative to compare against the contrast models Swin-UNet and DeepLabV3+ on different types of datasets, as shown in Schemes 1–3 in Table 4. The comparison results are presented in Table 6. It can be seen that Swin-MTNet achieved the highest OA and Kappa coefficients among all the models compared, with an OA of 93.83% and a Kappa coefficient of 91.91%. The contrast models performed the worst on the single-temporal dataset (Scheme 1). For the same contrast models, the classification accuracy for coastal wetlands is higher when using either the single-temporal augmented dataset (Scheme 2) or the temporal-stacked dataset (Scheme 3) compared to the single-temporal dataset. Compared to the single-temporal dataset, the Swin-UNet model's OA increased by

Table 5 Comparison results of ablation experiments

Class	Swin-MTNet			Swin-MTNet-ES			Swin-MTNet-ECS		
	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
SS	0.79	0.70	0.74	0.80	0.66	0.73	0.79	0.71	0.75
SA	0.85	0.83	0.84	0.77	0.78	0.78	0.76	0.65	0.70
NWB	0.98	0.98	0.98	0.98	0.98	0.98	0.99	0.97	0.98
MU	0.93	0.97	0.95	0.93	0.97	0.9475	0.90	0.97	0.93
AF	0.94	0.92	0.93	0.93	0.91	0.9241	0.92	0.92	0.92
IS	0.83	0.80	0.82	0.82	0.78	0.7973	0.80	0.76	0.78
PA	0.91	0.90	0.91	0.90	0.89	0.8920	0.90	0.87	0.88
FA	0.93	0.95	0.94	0.93	0.94	0.9346	0.91	0.94	0.93
RE	0.93	0.96	0.94	0.95	0.95	0.9484	0.97	0.90	0.93
OA (%)	93.83			93.28			92.53		
Kappa (%)	91.91			91.19			90.22		

Table 6 Classification results of Swin-MTNet and contrast models on different schemes

Model	Scheme	Evaluation metrics	SS	SA	NWB	MU	AF	IS	PA	FA	RE	OA (%)	Kappa (%)
Swin-UNet	Scheme 1	Precision	0.10	0.61	0.89	0.74	0.81	0.67	0.81	0.86	0.39	83.49	77.93
		Recall	0.07	0.30	0.95	0.60	0.76	0.65	0.73	0.90	0.25		
		F1-score	0.08	0.40	0.92	0.67	0.79	0.66	0.77	0.88	0.30		
	Scheme 2	Precision	0.26	0.28	0.95	0.81	0.89	0.76	0.86	0.88	0.96	88.71	85.06
		Recall	0.33	0.15	0.97	0.78	0.85	0.73	0.80	0.92	0.90		
		F1-score	0.29	0.19	0.96	0.80	0.87	0.75	0.83	0.90	0.92		
	Scheme 3	Precision	0.47	0.44	0.97	0.87	0.86	0.71	0.79	0.88	0.85	88.14	84.41
		Recall	0.37	0.36	0.96	0.89	0.86	0.67	0.75	0.91	0.65		
		F1-score	0.41	0.39	0.96	0.88	0.86	0.69	0.77	0.89	0.74		
DeepLabV3+	Scheme 1	Precision	0.13	0.80	0.88	0.74	0.80	0.70	0.81	0.86	0.37	83.44	77.82
		Recall	0.03	0.49	0.95	0.58	0.73	0.67	0.75	0.91	0.29		
		F1-score	0.04	0.60	0.92	0.65	0.76	0.68	0.78	0.89	0.33		
	Scheme 2	Precision	0.67	0.73	0.96	0.86	0.91	0.82	0.90	0.89	0.97	90.98	88.05
		Recall	0.52	0.25	0.98	0.79	0.89	0.76	0.86	0.94	0.95		
		F1-score	0.58	0.37	0.97	0.82	0.90	0.79	0.88	0.92	0.96		
	Scheme 3	Precision	0.79	0.66	0.96	0.89	0.88	0.73	0.80	0.89	0.80	88.98	85.49
		Recall	0.54	0.55	0.97	0.94	0.84	0.68	0.80	0.91	0.20		
		F1-score	0.64	0.60	0.96	0.91	0.86	0.70	0.80	0.90	0.32		
Swin-MTNet	Scheme 4	Precision	0.79	0.85	0.98	0.93	0.94	0.83	0.91	0.93	0.93	93.83	91.91
		Recall	0.70	0.83	0.98	0.97	0.92	0.80	0.90	0.95	0.96		
		F1-score	0.74	0.84	0.98	0.95	0.95	0.82	0.91	0.94	0.94		

approximately 5.22% and the DeepLabV3+ model's OA increased by approximately 7.54% when using the single-temporal augmented dataset. When using the temporal-stacked dataset, the Swin-UNet model's OA increased by approximately 4.65% and the DeepLabV3+ model's OA increased by approximately 5.54%. The result indicates that multi-temporal data provide more information to the models compared to single-temporal data, during the classification task. The models perform better especially for categories with significant temporal variations, such as *Suaeda salsa*, *Phragmites australis*, and mudflats, with multi-temporal data compared to single-temporal data. The classification results of Swin-UNet and DeepLabV3+ models have similar performance on the band-stacked dataset and the single-temporal expanded dataset. Moreover, the OA and Kappa coefficients indicate that models using the single-temporal expanded dataset outperform those using the band-stacked dataset—the OA of the Swin-UNet model improves by 0.57%, and the DeepLabV3+ model improves by 2%. However, for small sample categories with fewer pixels, such as *Suaeda salsa* and *Spartina alterniflora*, using the band-stacked dataset results in better classification performance.

Figure 5 illustrates the classification results of several representative scenes in the Liao River estuary region using different datasets with Swin-MTNet and contrast models under different schemes. The Swin-MTNet model effectively segments the edges between wetlands and natural water bodies compared with other results (first and second rows in Fig. 5). While contrast models using the band-stacked multi-temporal dataset (Scheme 3) can also segment the edges, they misclassify some central wetland areas as water bodies. Contrast models using the single-temporal expanded dataset (Scheme 2) and the single-temporal dataset (Scheme 1) fail to classify the wetlands into complete shapes. This is because the mudflat areas that labeled are the regions most exposed during the four temporal instances. If only single-temporal information is used, complete classification is almost impossible. The single-temporal expanded dataset may contain mismatches between labels and image information for categories like mudflats with significant temporal variations, leading to incomplete segmentation. The Swin-MTNet model performs exceptionally well in classifying wetland vegetation, such as *Suaeda salsa* and *Phragmites australis*, with clear contours. The contrast models using the single-temporal augmented dataset failed to classify *Suaeda salsa* (first row in Fig. 5), and models from other schemes also showed blurred edges in *Suaeda salsa* classification (second row in Fig. 5). Similarly, except for the Swin-MTNet model, the contrast models from other schemes misclassified *Phragmites australis* as farmland or impervious surfaces (upper right corner of the second row in Fig. 5). For the classification of aquaculture areas, the Swin-MTNet model showed a significant reduction in misclassification.

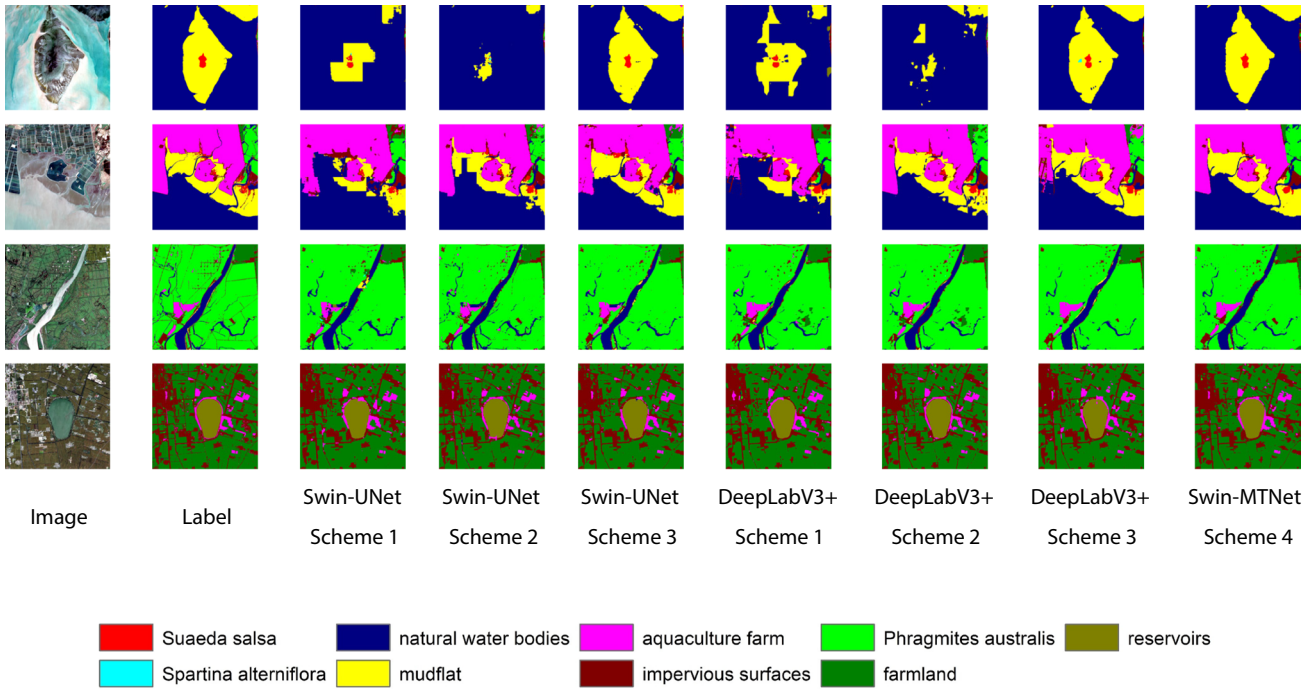


Fig. 5 Visualization of classification results using different schemes with Swin-MTNet and contrast models in several representative scenes of the Liao River estuary region

Contrast models using the temporal-stacked dataset and the single-temporal dataset misclassified aquaculture areas as natural water bodies (second row in Fig. 5), and other schemes also showed similar misclassification to varying degrees (third row in Fig. 5). The Swin-MTNet model provided a more complete classification of reservoirs. Contrast models using the temporal-stacked dataset exhibited some misclassification of small parts of the reservoir, with surrounding aquaculture areas being misclassified as impervious surfaces.

Figure 6 displays the visualization results of Swin-MTNet and contrast models for classifying the Yellow River delta

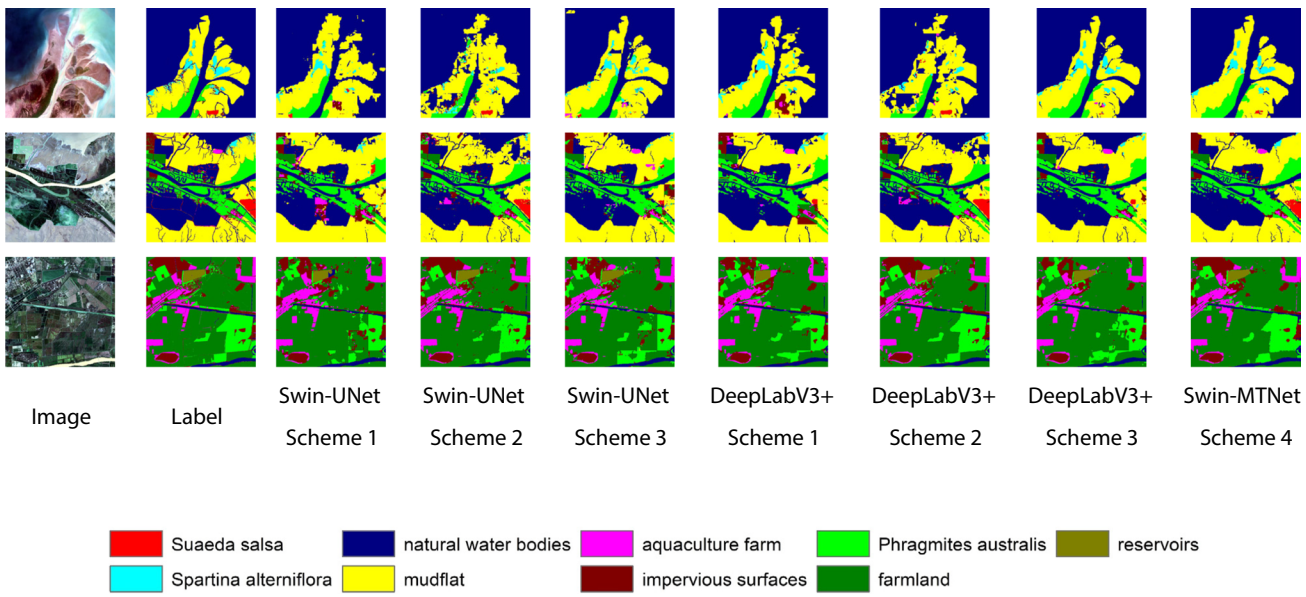


Fig. 6 Visual results of wetland classification using different schemes with Swin-MTNet and contrast models in several representative scenes of the Yellow River delta region

wetland area using different schemes. For the classification of wetland vegetation such as *Spartina alterniflora*, *Suaeda salsa*, and *Phragmites australis*, the Swin-MTNet model still outperforms the contrast models from other schemes. The contrast models using the single-temporal augmented dataset (Scheme 2) made extensive errors in classifying *Spartina alterniflora* (first row in Fig. 6), and the *Spartina alterniflora* classified by contrast models from other schemes had blurry and fragmented contours. The classification results for *Suaeda salsa* and *Phragmites australis* are similar to those for *Spartina alterniflora*. Models using the temporal-stacked dataset (Scheme 3) and the single-temporal dataset (Scheme 1) failed to classify *Suaeda salsa* (second row in Fig. 6). Except for the Swin-MTNet model, other models misclassified parts of the *Phragmites australis* as farmland and mudflat (second and third rows in Fig. 6). Compared to other results, the Swin-MTNet model provided accurate classifications for reservoirs and farmland, while other models had serious misclassifications for farmland. Models using single-temporal data (Scheme 1) severely misclassified reservoirs as natural water bodies or aquaculture areas (third row in Fig. 6). For the classification of impervious surfaces, the Swin-MTNet model showed clear edges and better performance compared to the models from other schemes (third row in Fig. 6).

As seen from Table 6 and Figs. 5 and 6, the Swin-MTNet model outperforms the contrast models from other schemes in both the segmentation of mudflat edges and the classification of wetland vegetation. The F1 scores for mudflat, *Suaeda salsa*, *Spartina alterniflora* and are also the highest, at 0.95, 0.74, and 0.84, respectively. Compared to Scheme 2's DeepLabV3+ model, which has the second-highest overall accuracy after Swin-MTNet, these three categories show improvements of 0.13, 0.16, and 0.47, respectively. Moreover, the DeepLabV3+ model in Scheme 2 does not perform well in classifying mudflat, as the segmented images are incomplete and exhibit large areas of misclassification. The Swin-UNet model, in both Scheme 2 and Scheme 3, does not perform well in classifying *Suaeda salsa* and *Spartina alterniflora*, with frequent mutual misclassifications. Contrast models using multi-temporal data (Schemes 2 and 3) generally perform better than those using single-temporal data (Scheme 1). This indicates that multi-temporal data provides more information to the models, thereby aiding in the improvement of classification accuracy.

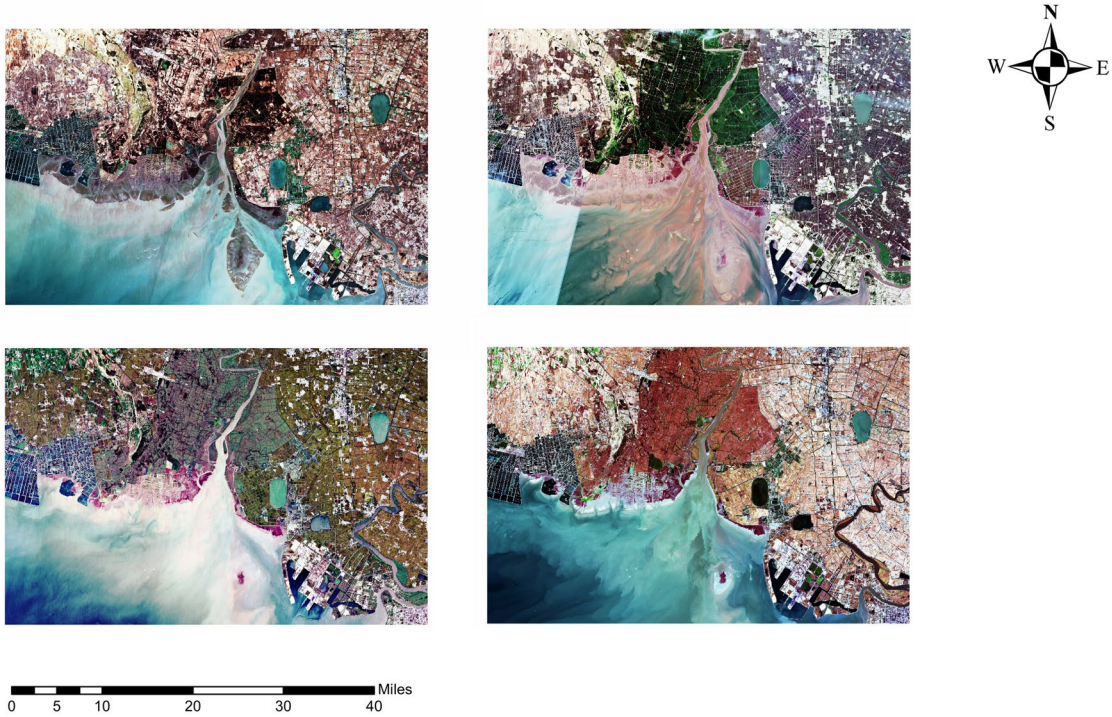
Wetland Maps of the Study Area

The models with the highest overall accuracy from Swin-MTNet, Swin-UNet, and DeepLabV3+ are selected to perform wetland mapping of the study area in 2022. The results are shown in Figs. 7 and 8. As can be seen from the results, the Swin-MTNet model significantly outperforms the Swin-UNet and DeepLabV3+ models in the classification of mudflats and wetland vegetation.

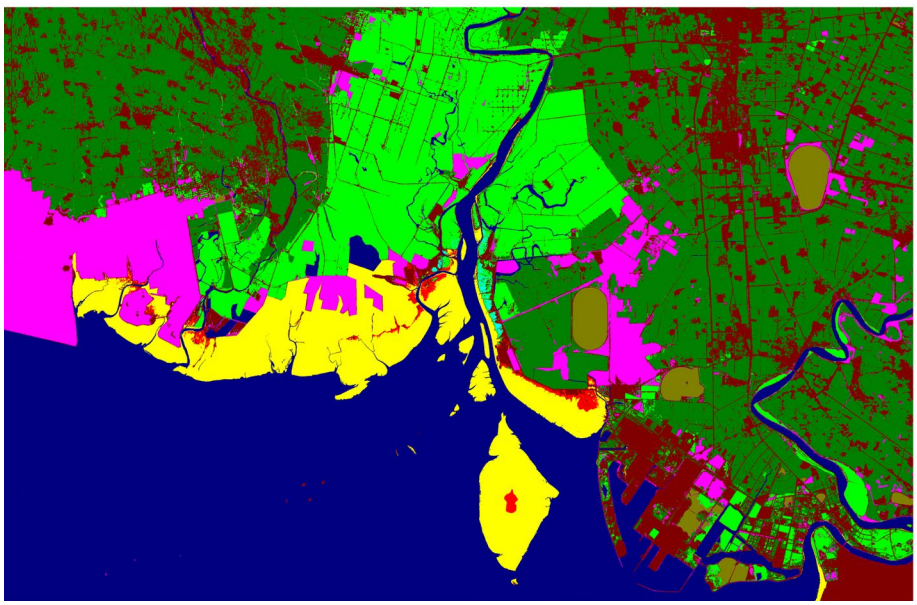
Discussion

Accurate mapping of coastal wetlands is a crucial and challenging task, and multi-temporal data can provide more information and is expected to improve accuracy. In this study, we propose the Swin-MTNet network model, which can learn multi-temporal information. The classification results of Swin-MTNet are compared and analyzed against the results of contrast models using both multi-temporal and single-temporal data. The results demonstrate that the proposed Swin-MTNet network model can better learn temporal features, outperforming contrast models on both temporally-augmented and temporally-stacked datasets, with significantly improved accuracy. Furthermore, it is concluded that multi-temporal information is more beneficial for the classification of vegetation or other categories with significant temporal variations, particularly for categories such as *Suaeda salsa*, *Spartina alterniflora*, and *Phragmites australis*. Models trained on single-temporal dataset perform poorly due to the lack of complete temporal information, making it difficult to distinguish between different growth stages. Multi-temporal data, however, provides the model with more features, resulting in improved accuracy in coastal wetland classification. Current research (Chen et al., 2024b, 2022) also indicates that multi-temporal information can enhance classification accuracy. Guo et al. used a random forest model to classify the coastal wetlands of the Liaohe Estuary and found that methods utilizing temporal features achieved higher overall accuracy and Kappa coefficient compared to methods lacking temporal features (Guo et al., 2024).

The proposed time learning module, CSAM, and feature fusion module, SAFF, are applicable to other models as well. CSAM is essentially a weight matrix, making it adaptable to other information filtering learning processes, enabling the selection of more important features through local and global feature learning. The SAFF module is another implementation of the CSAM module applied to the model's skip connections. Hence, the CSAM module exhibits broad scalability and application scenarios, warranting further exploration of its performance in other scenarios. Xiao et al. design a parallel branch with a context aggregation module in the Swin-UNet encoder to enhance contextual information extraction. The results showed that compared to Swin-UNet, the mIoU, mF1, and OA values increased by 2.83%, 2.47%, and 2.05%, respectively (Xiao et al., 2023).



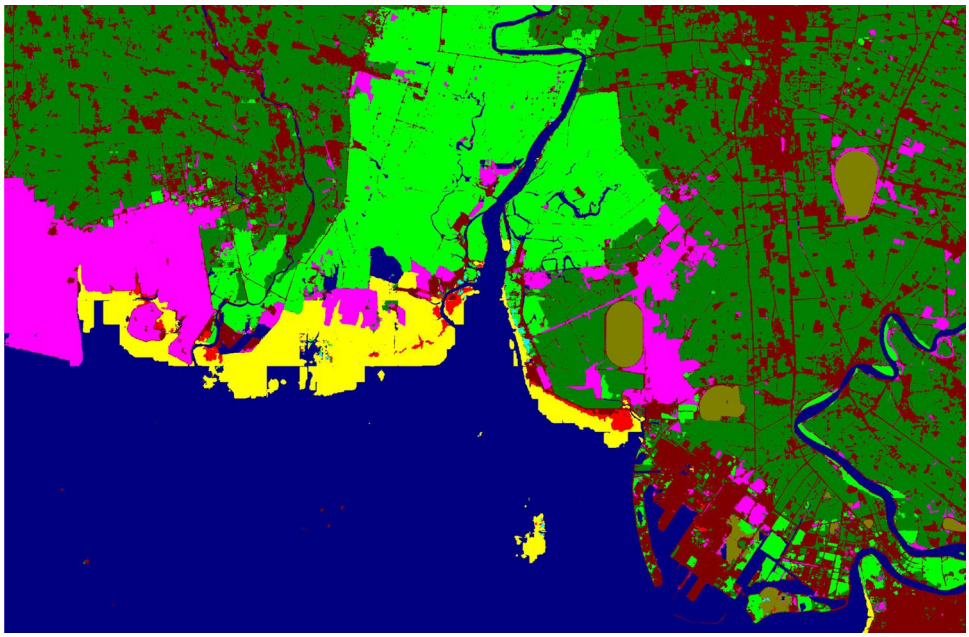
(a) Image



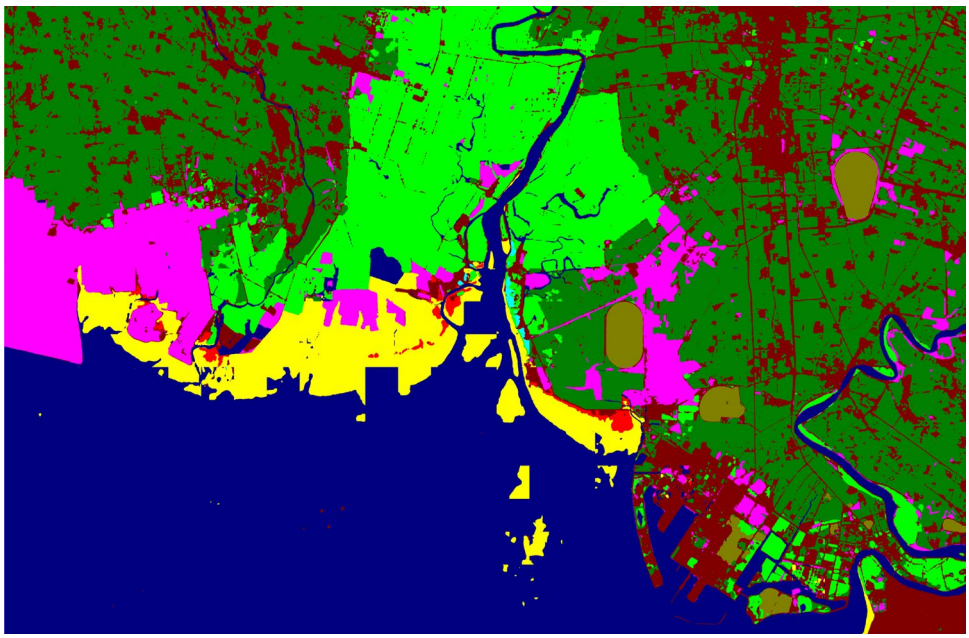
(b) label

Fig. 7 (a-e) Wetland map of the Liao River estuary region

Despite the excellent utilization of multi-temporal information by Swin-MTNet, there are still some limitations. Image data must be complete and cannot contain missing values. If some pixel points in the samples are null values, training cannot be performed. However, after cloud removal and cloud shadow masking of the downloaded Sentinel-2 imagery, some areas with clouds may result in empty values. Therefore, for the parts of the image with



(c) Swin-UNet



(d) DeepLabV3+

Fig. 7 (continued)

missing values after cloud removal, the nearest available image is used for filling, meeting Swin-MTNet's requirement for high image quality, in this work. Nevertheless, since the filled parts are not from the same day as the original image, there may be some numerical deviations, which may affect the improvement of image classification accuracy. Experiments can be conducted using data generated by GANs to enhance the model's robustness against this challenge in future research. Second, the utilized DEM elevation data is released in June 2019, with a global spatial resolution of 30 m. Visual inspection revealed that in recent years, there have been small-scale land reclamation

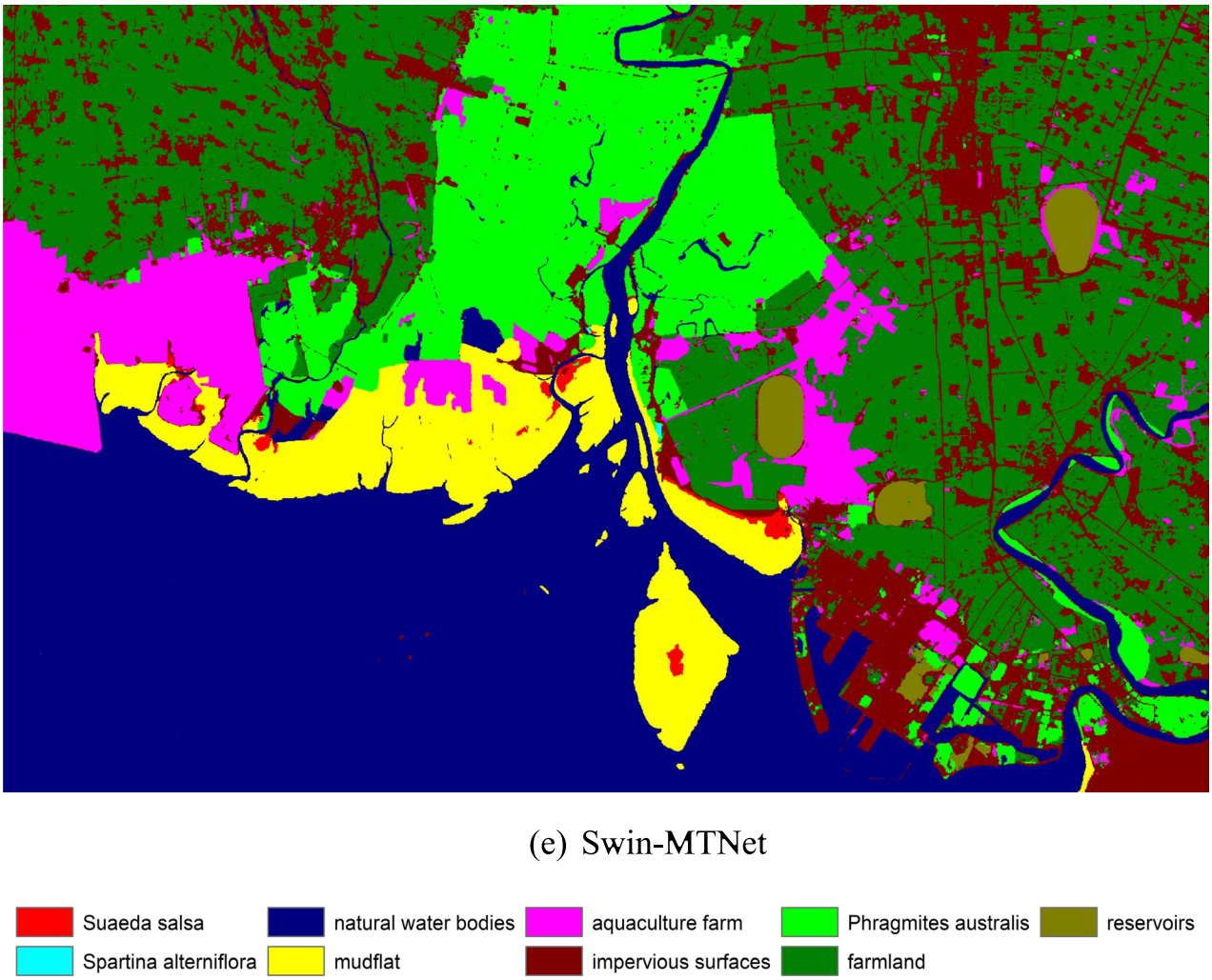


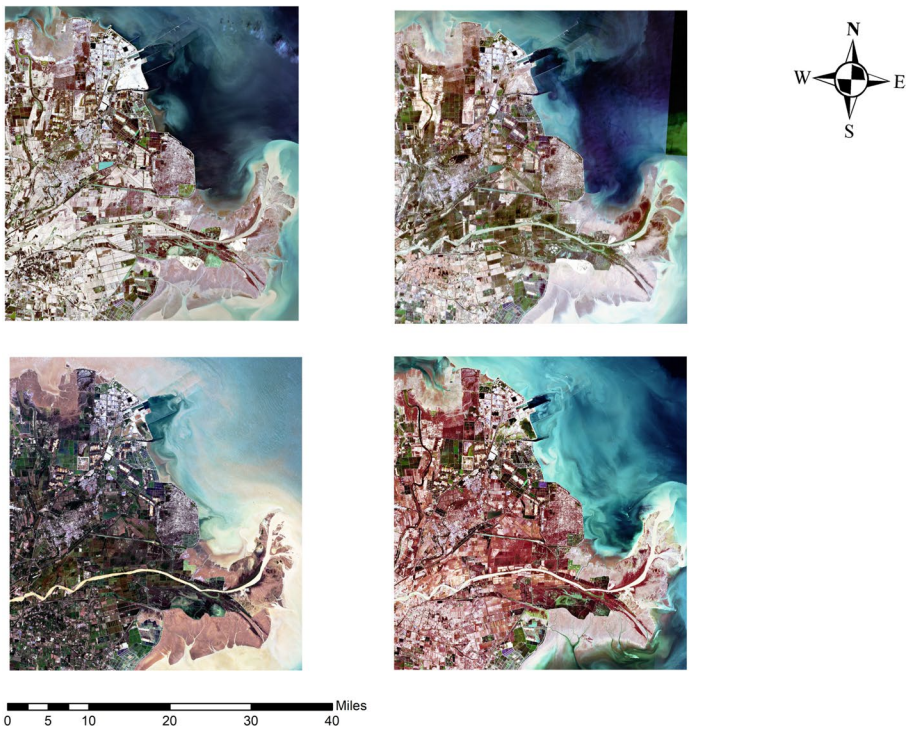
Fig. 7 (continued)

projects in the Bohai Bay area. Additionally, the DEM data is resampled to a 20-m resolution, causing some discrepancies between the DEM data and the 2022 Sentinel-2 imagery of the Liao River estuary and Yellow River delta regions.

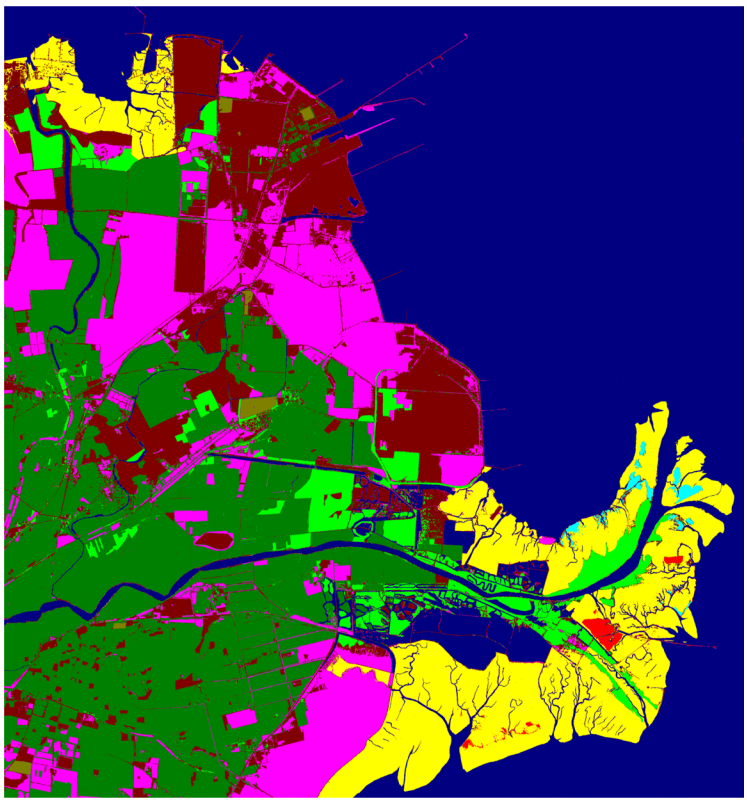
Due to the model's use of temporal information across different seasons for learning, vegetation feature that are sensitive to seasonal variations significantly influence its performance. In regions with vegetation feature that differ markedly from the study area, such as coastal wetlands in southern China, the model's classification accuracy may be affected by discrepancies in vegetation growth. This hypothesis was validated in preliminary experiments conducted in the Yancheng area of Jiangsu Province. Results showed a significant reduction in classification accuracy for *Suaeda salsa* and *Spartina alterniflora*, with accuracies of 0.36 and 0.69, respectively, while the accuracy for farmland and *Phragmites australis* decreased slightly to 0.88 and 0.83, respectively. Future research should focus on optimizing and improving the model for wetland types with significant regional differences and incorporating their feature to enhance its generalization ability.

Conclusion

This paper proposes a multi-temporal deep learning network model, Swin-MTNet, for classifying complex coastal wetlands using multi-temporal Sentinel-2 image data. Swin-UNet model is improved into a four-branch input model, with each branch receiving input from different temporal data, and a time learning module, CSAM, is constructed to filter important temporal information from different branches. Additionally, based on the CSAM module, a feature fusion module, SAFF, is designed to fuse shallow and deep feature information at the model's



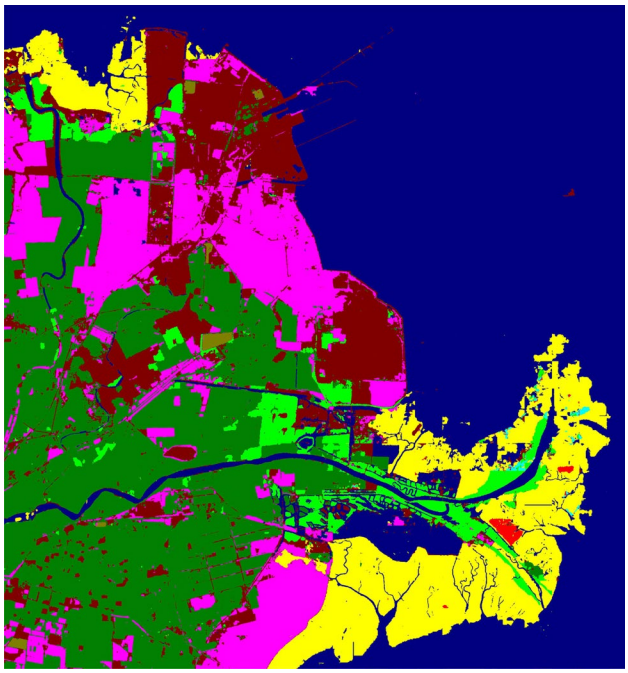
(a) Image



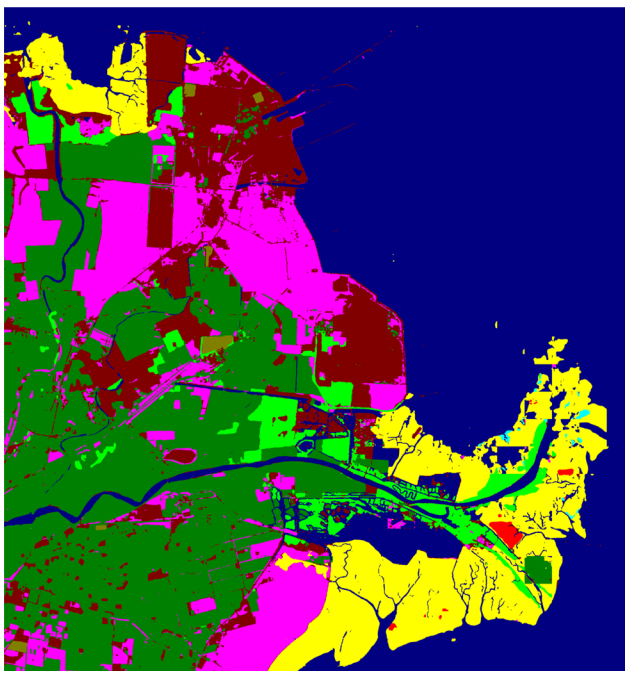
(b) label

Fig. 8 (a-e) Wetland map of the Yellow River delta region

Fig. 8 (continued)

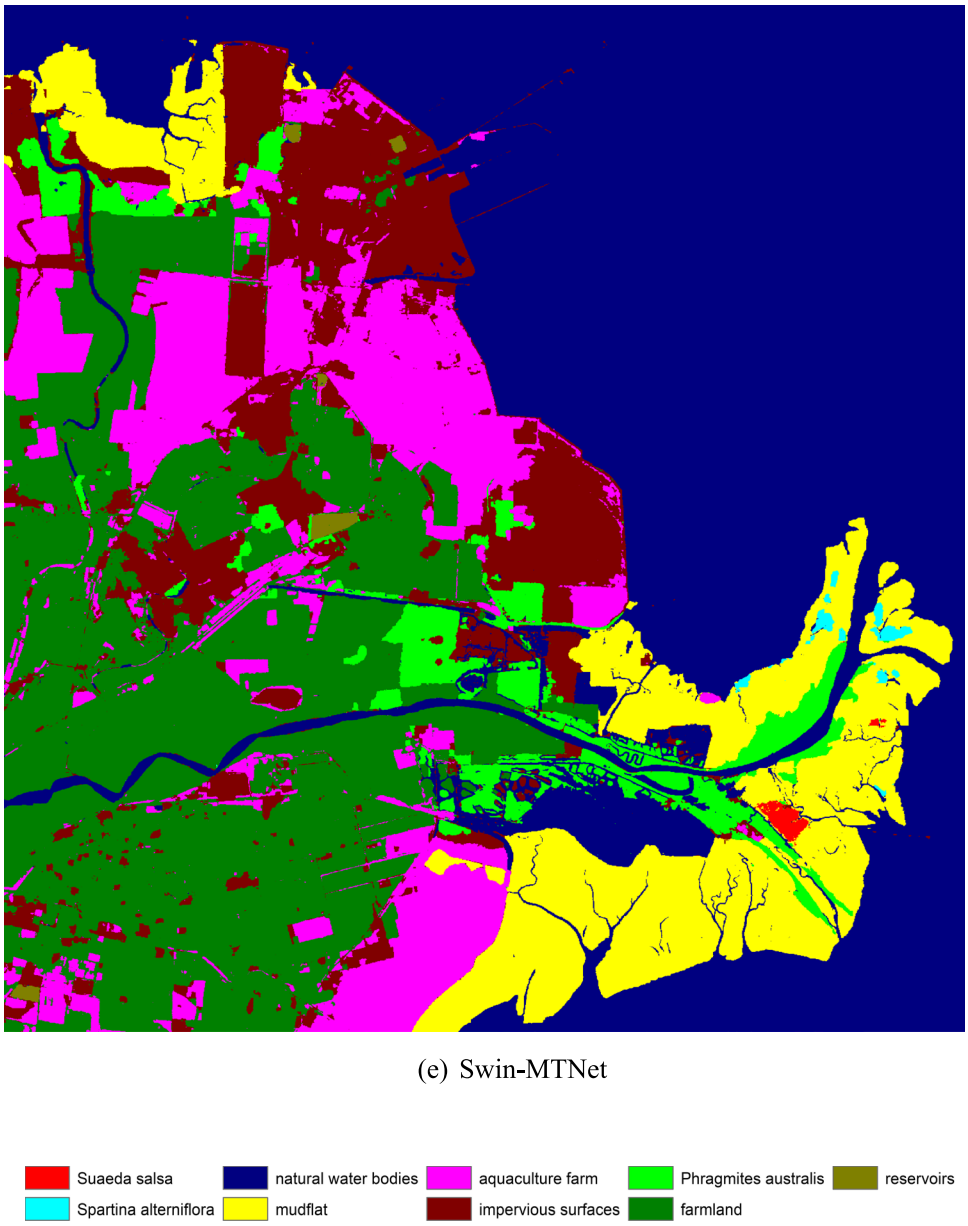


(c) Swin-UNet



(d) DeepLabV3+

Fig. 8 (continued)



(e) Swin-MTNet

skip connections. Different multi-temporal datasets are constructed to explore the differences in coastal wetland classification abilities between contrast models using multi-temporal and single-temporal information, and the ability of Swin-MTNet model to learn temporal features compared to other contrast models. The results indicate that after utilizing multi-temporal information, the average accuracy of Swin-UNet model for coastal wetland classification increases by 5.22%, and the Kappa coefficient increases by 7.13%. The average accuracy of DeepLabV3+ model increases by 7.54%, and the Kappa coefficient increases by 10.23%. Swin-MTNet model achieves a 5.12% increase in average accuracy and a 6.85% increase in the Kappa coefficient compared to Swin-UNet model using multi-temporal information, and a 2.85% increase in average accuracy and a 3.86% increase in the Kappa coefficient compared to DeepLabV3+ model using multi-temporal information. The experimental results demonstrate that utilizing multi-temporal information from remote sensing can improve classification performance of coastal wetlands compared to utilizing single-temporal information, and our proposed Swin-MTNet model exhibits better ability to learn multi-temporal information than contrast models.

Author Contribution All listed authors make significant contributions to the manuscript. Binyu Wang is responsible for data collection, program development, and drafting the first version of the manuscript. Yuanheng Sun defines the overall objectives of the project, conducts data analysis, and revises the manuscript. All authors contribute to the initial data review and quality control and approve the manuscript for submission.

Funding This work is supported by the National Natural Science Foundation of China (42301391) and Fundamental Research Funds for the Central Universities of China (3132024120).

Data Availability Data will be made available on request.

Declarations

Competing Interests The authors declare no competing interests.

References

Alam, S. M. R., & Hossain, M. S. (2024). Using a water index approach to mapping periodically inundated saltmarsh land-cover vegetation and eco-zonation using multi-temporal Landsat 8 imagery. *Journal of Coastal Conservation, 28*, 19.

Amani, M., Brisco, B., Mahdavi, S., Ghorbanian, A., Moghimi, A., DeLancey, E. R., Merchant, M., Jahncke, R., Fedorchuk, L., Mui, A., Fisette, T., Kakooei, M., Ahmadi, S. A., Leblon, B., & LaRocque, A. (2021). Evaluation of the Landsat-based Canadian wetland inventory Map using multiple sources: Challenges of large-scale wetland classification using remote sensing. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 32–52. https://doi.org/10.1109/JSTARS.2020.3036802

Barbier, E. B., Hacker, S. D., Kennedy, C., Koch, E. W., Stier, A. C., & Silliman, B. R. (2011). The value of estuarine and coastal ecosystem services. *Ecological Monographs, 81*, 169–193. https://doi.org/10.1890/10-1510.1

Bazi, Y., Bashmal, L., Rahhal, M. M. A., Dayil, R. A., & Ajlan, N. A. (2021). Vision transformers for remote sensing image classification. *Remote Sensing, 13*, 516.

Bennett, K. P. (1994). Global tree optimization: A non-greedy decision tree algorithm. *Computing Science and Statistics, 26*, 156–156.

Berhane, T. M., Lane, C. R., Wu, Q., Autrey, B. C., Anenkhonov, O. A., Chepinoga, V. V., & Liu, H. (2018). Decision-tree, rule-based, and random forest classification of high-resolution multispectral imagery for wetland mapping and inventory. *Remote sensing, 10*, 580.

Bill Donatien, L. M., Clobite, B. B., & MerisMidel, M. L. (2024). Land use land cover change detection using multi-temporal Landsat imagery in the North of Congo Republic: A case study in Sangha region. *Geocarto International, 39*, 2425184.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M. (2022). Swin-unet: Unet-like pure transformer for medical image segmentation. In *European Conference on Computer Vision*. Springer, pp. 205–218. https://doi.org/10.1007/978-3-031-25066-8_9

Chen, C., Fu, J., Zhang, S., & Zhao, X. (2019). Coastline information extraction based on the tasseled cap transformation of Landsat-8 OLI images. *Estuarine, Coastal and Shelf Science, 217*, 281–291.

Chen, C., Liang, J., Xie, F., Hu, Z., Sun, W., Yang, G., Yu, J., Chen, L., Wang, Lihua, & Wang, Liyan. (2022). Temporal and spatial variation of coastline using remote sensing images for Zhoushan archipelago, China. *International Journal of Applied Earth Observation and Geoinformation, 107*, 102711.

Chen, R., Yang, H., Yang, G., Liu, Y., Zhang, C., Long, H., Xu, H., Meng, Y., & Feng, H. (2023). Land-use mapping with multi-temporal sentinel images based on google earth engine in Southern Xinjiang Uygur Autonomous Region. *China. Remote Sensing, 15*, 3958.

Chen, C., Sun, W., Yang, Z., Yang, G., Jia, M., Zhang, Z., Liang, J., Chen, Y., Ren, T., & Hu, X. (2024b). Tracking dynamics characteristics of tidal flats using landsat time series and Google Earth Engine cloud platform. *Resources, Conservation and Recycling, 209*, 107751.

Chen, C., Zou, Z., Sun, W., Yang, G., Song, Y., & Liu, Z. (2024c). Mapping the distribution and dynamics of coastal aquaculture ponds using Landsat time series data based on U2-Net deep learning model. *International Journal of Digital Earth, 17*, 2346258.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H. (2018). Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*. pp. 801–818.

Chen, C., Liang, J., Sun, W., Yang, G., Meng, X. (2024a). An automatically recursive feature elimination method based on threshold decision in random forest classification. *Geo-spatial Information Science,* 1–26. https://doi.org/10.1080/10095020.2024.2387457

Cortes, C., & Vapnik, V. (1995). *Support-vector networks. Machine learning, 20*, 273–297.

Dai, Y., Gieseke, F., Oehmcke, S., Wu, Y., Barnard, K. (2021). Attentional feature fusion. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. pp. 3560–3569.

Evans, T. L., & Costa, M. (2013). Landcover classification of the Lower Nhecolândia subregion of the Brazilian Pantanal Wetlands using ALOS/PALSAR, RADARSAT-2 and ENVISAT/ASAR imagery. *Remote Sensing of Environment, 128*, 118–137.

Gonzalez-Perez, A., Abd-Elrahman, A., Wilkinson, B., Johnson, D. J., & Carthy, R. R. (2022). Deep and machine learning image classification of coastal wetlands using unpiloted aircraft system multispectral images and lidar datasets. *Remote Sensing, 14*, 3937.

Guo, S., Feng, Z., Wang, P., Chang, J., Han, H., Li, H., Chen, C., Du, W. (2024). Mapping and classification of the Liaohe Estuary Wetland based on the combination of object-oriented and temporal features. IEEE Access. https://doi.org/10.1109/ACCESS.2024.3389935

Hao, M., Dong, X., Jiang, D., Yu, X., Ding, F., & Zhuo, J. (2024). Land-use classification based on high-resolution remote sensing imagery and deep learning models. *Plos one, 19*, e0300473.

He, J., Zhao, L., Yang, H., Zhang, M., & Li, W. (2020). HSI-BERT: Hyperspectral image classification using the bidirectional encoder representation from transformers. *IEEE Transactions on Geoscience and Remote Sensing, 58*, 165–178. https://doi.org/10.1109/TGRS.2019.2934760

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., & Chanussot, J. (2022). SpectralFormer: Rethinking hyperspectral image classification with transformers. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–15. https://doi.org/10.1109/TGRS.2021.3130716

Hosseiny, B., Mahdianpari, M., Brisco, B., Mohammadimanesh, F., & Salehi, B. (2022). WetNet: A spatial–temporal ensemble deep learning model for wetland classification using Sentinel-1 and Sentinel-2. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1–14. https://doi.org/10.1109/TGRS.2021.3113856

Ioffe, S., Szegedy, C., 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of the 32nd International Conference on Machine Learning. Presented at the International Conference on Machine Learning*, PMLR, pp. 448–456.

Jamali, A., & Mahdianpari, M. (2022a). Swin transformer and deep convolutional neural networks for coastal wetland classification using Sentinel-1, Sentinel-2, and LiDAR data. *Remote Sensing, 14*, 359. https://doi.org/10.3390/rs14020359

Jamali, A., & Mahdianpari, M. (2022b). Swin transformer for complex coastal wetland classification using the integration of Sentinel-1 and Sentinel-2 imagery. *Water, 14*, 178. https://doi.org/10.3390/w14020178

Jamali, A., Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., & Salehi, B. (2021a). Deep forest classifier for wetland mapping using the combination of Sentinel-1 and Sentinel-2 data. *GIScience & Remote Sensing, 58*, 1072–1089. https://doi.org/10.1080/15481603.2021.1965399

Jamali, A., Mahdianpari, M., Mohammadimanesh, F., Brisco, B., & Salehi, B. (2021b). A synergic use of Sentinel-1 and Sentinel-2 imagery for complex wetland classification using generative adversarial network (GAN) scheme. *Water, 13*, 3601. https://doi.org/10.3390/w13243601

Jamali, A., Mahdianpari, M., Brisco, B., Mao, D., Salehi, B., & Mohammadimanesh, F. (2022). 3DUNetGSFormer: A deep learning pipeline for complex wetland mapping using generative adversarial networks and Swin transformer. *Ecological Informatics, 72*, 101904. https://doi.org/10.1016/j.ecoinf.2022.101904

Jamali, A., Roy, S. K., & Ghamisi, P. (2023). WetMapFormer: A unified deep CNN and vision transformer for complex wetland mapping. *International Journal of Applied Earth Observation and Geoinformation, 120*, 103333. https://doi.org/10.1016/j.jag.2023.103333

Liu, M., Fu, B., Fan, D., Zuo, P., Xie, S., He, H., Liu, L., Huang, L., Gao, E., & Zhao, M. (2021a). Study on transfer learning ability for classifying marsh vegetation with multi-sensor images using DeepLabV3+ and HRNet deep learning algorithms. *International Journal of Applied Earth Observation and Geoinformation, 103*, 102531. https://doi.org/10.1016/j.jag.2021.102531

Liu, M., Fu, B., Xie, S., He, H., Lan, F., Li, Y., Lou, P., & Fan, D. (2021b). Comparison of multi-source satellite images for classifying marsh vegetation using DeepLabV3 Plus deep learning algorithm. *Ecological Indicators, 125*, 107562. https://doi.org/10.1016/j.ecolind.2021.107562

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B. (2021c). Swin transformer: Hierarchical vision transformer using shifted windows. Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10012–10022.

Lou, A., He, Z., Zhou, C., & Lai, G. (2024). Long-term series wetland classification of Guangdong-Hong Kong-Macao Greater Bay Area based on APSMnet. *International Journal of Applied Earth Observation and Geoinformation, 128*, 103765. https://doi.org/10.1016/j.jag.2024.103765

Mahdianpari, M., Salehi, B., Mohammadimanesh, F., & Brisco, B. (2017). An assessment of simulated compact polarimetric SAR data for wetland classification using random forest algorithm. *Canadian Journal of Remote Sensing, 43*, 468–484.

Mahdianpari, M., Salehi, B., Rezaee, M., Mohammadimanesh, F., & Zhang, Y. (2018). Very deep convolutional neural networks for complex land cover mapping using multispectral remote sensing imagery. *Remote Sensing, 10*, 1119. https://doi.org/10.3390/rs10071119

Mahdianpari, M., Brisco, B., Granger, J., Mohammadimanesh, F., Salehi, B., Homayouni, S., & Bourgeau-Chavez, L. (2021). The third generation of Pan-Canadian wetland map at 10 m resolution using multisource earth observation data on cloud computing platform. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 14*, 8789–8803. https://doi.org/10.1109/JSTARS.2021.3105645

Mao, D., Wang, Z., Du, B., Li, L., Tian, Y., Jia, M., Zeng, Y., Song, K., Jiang, M., & Wang, Y. (2020). National wetland mapping in China: A new product resulting from object-based and hierarchical classification of Landsat 8 OLI images. *ISPRS Journal of Photogrammetry and Remote Sensing, 164*, 11–25. https://doi.org/10.1016/j.isprsjprs.2020.03.020

Mao, D., Wang, Z., Wang, Y., Choi, C.-Y., Jia, M., Jackson, M.V., Fuller, R.A. (2021). Remote observations in China's Ramsar sites: Wetland dynamics, anthropogenic threats, and implications for sustainable development goals. *Journal of Remote Sensing*. https://doi.org/10.34133/2021/9849343

McCarthy, M. J., Radabaugh, K. R., Moyer, R. P., & Muller-Karger, F. E. (2018). Enabling efficient, large-scale high-spatial resolution wetland mapping using satellites. *Remote sensing of environment, 208*, 189–201.

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Brisco, B., & Motagh, M. (2018). Multi-temporal, multi-frequency, and multi-polarization coherence and SAR backscatter analysis of wetlands. *ISPRS Journal of Photogrammetry and Remote Sensing, 142*, 78–93. https://doi.org/10.1016/j.isprsjprs.2018.05.009

Mohammadimanesh, F., Salehi, B., Mahdianpari, M., Gill, E., & Molinier, M. (2019). A new fully convolutional neural network for semantic segmentation of polarimetric SAR imagery in complex land cover ecosystem. *ISPRS Journal of Photogrammetry and Remote Sensing, 151*, 223–236. https://doi.org/10.1016/j.isprsjprs.2019.03.015

Moharrami, M., Attarchi, S., Gloaguen, R., & Alavipanah, S. K. (2024). Integration of Sentinel-1 and Sentinel-2 data for ground truth sample migration for multi-temporal land cover mapping. *Remote Sensing, 16*, 1566.

Piaser, E., & Villa, P. (2023). Evaluating capabilities of machine learning algorithms for aquatic vegetation classification in temperate wetlands using multi-temporal Sentinel-2 data. *International Journal of Applied Earth Observation and Geoinformation, 117*, 103202.

Saad El Imanni, H., El Harti, A., Hssaisoune, M., Velastegui-Montoya, A., Elbouzidi, A., Addi, M., El Iysaouy, L., & El Hachimi, J. (2022). Rapid and automated approach for early crop mapping using Sentinel-1 and Sentinel-2 on Google Earth Engine; A case of a highly heterogeneous and fragmented agricultural region. *Journal of Imaging, 8*, 316. https://doi.org/10.3390/jimaging8120316

van Asselen, S., Verburg, P. H., Vermaat, J. E., & Janse, J. H. (2013). Drivers of wetland conversion: A global meta-analysis. *PloS one, 8*, e81292.

Xiao, D., Kang, Z., Fu, Y., Li, Z., & Ran, M. (2023). Csswin-unet: A Swin-unet network for semantic segmentation of remote sensing images by aggregating contextual information and extracting spatial information. *International Journal of Remote Sensing, 44*, 7598.

Yang, X., Zhang, B., Chen, Z., Bai, Y., & Chen, P. (2022). A multi-temporal network for improving semantic segmentation of large-scale landsat imagery. *Remote Sensing, 14*, 5062. https://doi.org/10.3390/rs14195062

Yao, J., & Jin, S. (2022). Multi-category segmentation of Sentinel-2 images based on the Swin UNet Method. *Remote Sensing, 14*, 3382. https://doi.org/10.3390/rs14143382

Zhu, P., & Gong, P. (2014). Suitability mapping of global wetland areas and validation with remotely sensed data. *Sci. China Earth Sci., 57*, 2283–2292. https://doi.org/10.1007/s11430-014-4925-1

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.