Online ISSN: 3104-9338
Print ISSN: 3104-932X

A Review of Applications of Deep Learning-Driven Multi-Source Remote Sensing Spatiotemporal Fusion in Wetland Monitoring

Zhiming Chen, Can Qin, Hao Zeng*

Yiyang Normal College

Received: December 1, 2025

Revised: December 12, 2025

Accepted: December 14, 2025

Published online: December 15, 2025

To appear in: *International Journal of Advanced AI Applications,* Vol. 2, No. 1 (January 2025)

* Corresponding Author: Hao Zeng(crawler 2015@163.com)

Abstract. Wetlands are vital ecosystems on Earth, and the dynamic monitoring of their changes is crucial for ecological conservation, biodiversity maintenance, and water resource management. Multi-source remote sensing spatiotemporal fusion technology can integrate the advantages of datasets with different spatiotemporal resolutions to generate image sequences with high spatiotemporal resolution, providing core data support for dynamic wetland monitoring. By virtue of its powerful feature extraction and nonlinear fitting capabilities, deep learning has greatly improved the accuracy and adaptability of spatiotemporal fusion, driving wetland monitoring toward refinement and dynamization. This paper systematically sorts out the research background and significance of deep learning-driven multi-source remote sensing spatiotemporal fusion technology, categorically elaborates on mainstream fusion models, summarizes their application achievements in four core scenarios (wetland boundary extraction, type classification, ecological parameter inversion, and dynamic change monitoring), analyzes the key problems faced by current technologies, and prospects future research directions, thereby providing a reference for the in-depth application of this technology in wetland monitoring.

Keywords: *Deep Learning; Multi-source Remote Sensing; Spatiotemporal Fusion*; *Wetland Monitoring*

1. Introduction

Wetlands are transitional zones between terrestrial and aquatic ecosystems, boasting crucial ecological functions such as water purification, flood regulation and storage, biodiversity conservation, and carbon sequestration and emission reduction, hence earning the title of the kidneys of the Earth [1-4]. However, affected by global climate change, urbanization expansion, and irrational human development, the global wetland area has shown a shrinking trend with severe degradation of ecological functions [5]. Wetland protection and dynamic monitoring have thus become one of the core tasks in global ecological and environmental governance. Remote sensing technology, leveraging its advantages of large-scale, non-contact, and periodic observation, has emerged as the mainstream technical means for wetland monitoring.

Current commonly used remote sensing sensors face a trade-off bottleneck between spatial and temporal resolution [6]. Specifically, high-spatial-resolution sensors (e.g., Landsat series, Sentinel-2) can capture the fine structure of wetlands, such as the distribution of vegetation communities and microtopographic features [7], but their long revisit periods make it difficult to capture short-term scale dynamic changes in wetlands (e.g., flood season water level fluctuations, abrupt changes in vegetation phenology) [8,9]. In contrast, high-temporal-resolution sensors (e.g., MODIS) enable daily high-frequency observations to monitor wetland dynamic processes, yet their low spatial resolution prevents accurate identification of small-scale wetland types and boundaries [10]. Additionally, wetland ecosystems exhibit high heterogeneity, encompassing various landscape types such as water bodies, herbaceous vegetation, emergent vegetation, and swamps. They are also susceptible to noise interference from clouds, shadows, water vapor, etc. Consequently, data from a single sensor is insufficient to meet the requirements of refined wetland monitoring.

Multi-source remote sensing spatiotemporal fusion technology, by establishing correlation models between data with different spatial and temporal resolutions, generates continuous image sequences with both high spatial and high temporal resolution, providing an effective approach to break through the "spatiotemporal resolution trade-off" bottleneck [11-13]. Traditional spatiotemporal fusion methods perform adequately in homogeneous areas, but in wetlands with complex and heterogeneous scenarios, their fusion accuracy is limited due to the difficulty in depicting complex spatiotemporal correlation characteristics.

In recent years, deep learning technology has significantly improved fusion performance in complex scenarios by virtue of its ability to automatically extract deep spatiotemporal features, offering a new technical paradigm for wetland monitoring [14-16]. Therefore, this paper systematically reviews the applications of deep learning-driven multi-source remote sensing spatiotemporal fusion technology in wetland monitoring. Focusing on wetlands as a special and complex ecosystem, it classifies and summarizes the adaptive improvements and innovations of deep learning-based spatiotemporal fusion models, clarifies the advantages and limitations of different models in wetland scenarios, enriches the application theoretical system of multi-source remote sensing spatiotemporal fusion, and promotes the deepening of interdisciplinary research (integrating remote sensing science, ecology, and artificial intelligence). Furthermore, this paper identifies the core value of this technology in wetland boundary extraction, type classification, ecological parameter inversion, and dynamic change monitoring, providing precise and efficient technical support for wetland protection and management. Meanwhile, it analyzes the key problems faced by the current technical applications, proposes targeted solutions, and facilitates the transformation of this technology from laboratory research to operational application, thereby offering a scientific basis for the protection, restoration, and sustainable management of wetland ecosystems.

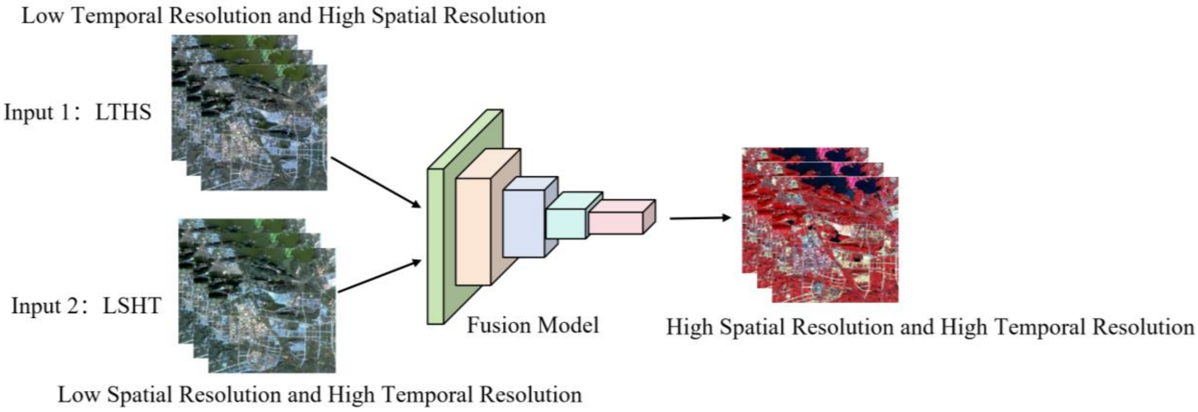

Figure 1. Spatio-temporal Fusion Structure Diagram of Multi-source Remote Sensing.

2.Deep Learning-Based Spatiotemporal Fusion Models

The high heterogeneity and susceptibility to noise interference of wetlands impose higher requirements on the feature extraction capability, anti-interference performance, and adaptability of deep learning-based spatiotemporal fusion models. As illustrated in Figure 1, the model takes two complementary types of remote sensing data as inputs: low-temporal-resolution, high-spatial-resolution data (LTHS, e.g., Landsat), which provides fine spatial details of wetlands but lacks temporal continuity; and low-spatial-resolution, high-temporal-resolution data (LSHT, e.g., MODIS), which captures dynamic changes but has coarse spatial details. By leveraging deep learning fusion models including Convolutional Neural Network (CNN) [17], Recurrent Neural Network (RNN) [18], Generative Adversarial Network (GAN) [19]and Transformer [20], the advantages of the two data types are integrated—compensating for the blurriness of low-spatial-resolution data with high-spatial details, and filling the temporal gaps of low-temporal-resolution data with high-temporal dynamics. The resulting fused data, characterized by both high spatial and high temporal resolution, retains fine spatial features of wetlands while enabling high-frequency temporal observation, thus enabling adaptation to complex wetland monitoring scenarios. These four categories of models have optimized fusion accuracy and stability through targeted improvements, and their principles and performance characteristics will be elaborated in subsequent sections. The mathematical formula for pixel calculation in the process of multi-source remote sensing spatiotemporal fusion can be expressed as follows:

$$Out_{fusion}(x,y,t) = \alpha \cdot I_{HS}(x,y) + \beta \cdot I_{HT}(x,y,t)$$

$$\alpha + \beta = 1$$

Among these parameters, I_{HS} denotes high-spatial-resolution data, I_{HT} denotes high-temporal-resolution data, and α and βrepresent weight coefficients.

2.1. Convolutional Neural Network based Models

Due to its local connectivity, weight sharing, and pooling operations, the Convolutional Neural Network (CNN) possesses robust spatial feature extraction capabilities, enabling it to effectively capture the spatial texture details of wetlands (e.g., vegetation community distribution, water body boundary contours) [21,22]. It serves as a fundamental model for spatiotemporal fusion in wetland monitoring. Its core principle lies in extracting temporal features from high-temporal-resolution images and spatial features from high-spatial-resolution images via an encoder, followed by up sampling through a decoder to reconstruct high spatiotemporal resolution images, thereby achieving the fusion goal of spatial detail enhancement temporal sequence continuity. The mathematical formula of the Convolutional Neural Network (CNN) can be expressed as follows:

$$X_l^{out} = \sigma\left(\sum_{k=1}^{K} W_k * X_{l-1}^{in} + b_k\right)$$

Among these parameters, $*$ denotes Convolution operation, X_{l-1}^{in} denotes input feature of the previous layer, and W_k denotes weight of the k convolution kernel and b_k represent bias value, X_l^{out} denotes output of the current layer, σ represent Sigmoid function.

For wetland scenarios, the main improvement directions of CNN-based models focus on enhancing edge detail preservation and anti-noise ability [23]. As illustrated in Figure 2, the classic U-Net model fuses feature from the encoder and decoder through skip connections, effectively preserving the boundary details of wetland water bodies and vegetation [24]. However, when dealing with large-scale spatiotemporal changes in wetlands, it exhibits insufficient global feature capture. To address this, as illustrated in Figure 3, researchers have proposed the U-Net++ model, which constructs a deeper feature fusion structure through dense skip connections [25]. In wetland boundary extraction scenarios, the boundary clarity of fused images has been improved by 12%–18% compared to the traditional U-Net. Targeting the problem of frequent cloud and fog in wetlands, Zhang et al. [26] introduced an attention gating module into the CNN model to focus on core target regions such as water bodies and vegetation, suppressing cloud and fog noise interference. In the fusion experiment conducted on the Poyang Lake Wetland, the root mean square error (RMSE) of the model for cloud and fog-affected areas was reduced by 23% compared to the traditional CNN.

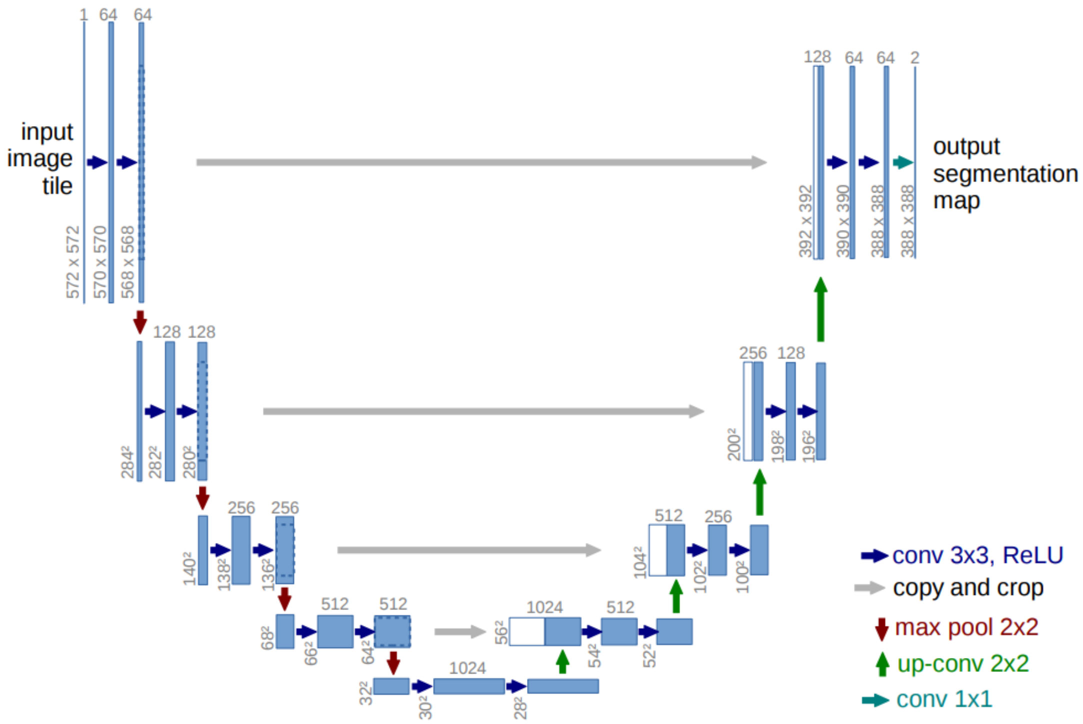

Figure 2. U-Net model architecture.

In addition, to adapt to the multi-scale features of wetlands (e.g., macro wetland landscapes and micro vegetation individuals), multi-scale CNN models have emerged as the times require. Li et al. [27] constructed a pyramid CNN fusion model, which extracts coarse-scale dynamics and fine-scale structural features of wetlands through convolution kernels of different sizes. In the fusion experiment conducted on the Wetland, the model's discrimination accuracy between emergent vegetation and herbaceous vegetation was improved by 15% compared with single-scale CNN models. The advantages of CNN-based models lie in their excellent spatial detail restoration effect and high computational efficiency, making them suitable for scenarios such as wetland boundary extraction and small-scale type classification. However, when used alone, they are insufficient in capturing the long-term sequential dynamic changes of wetlands.

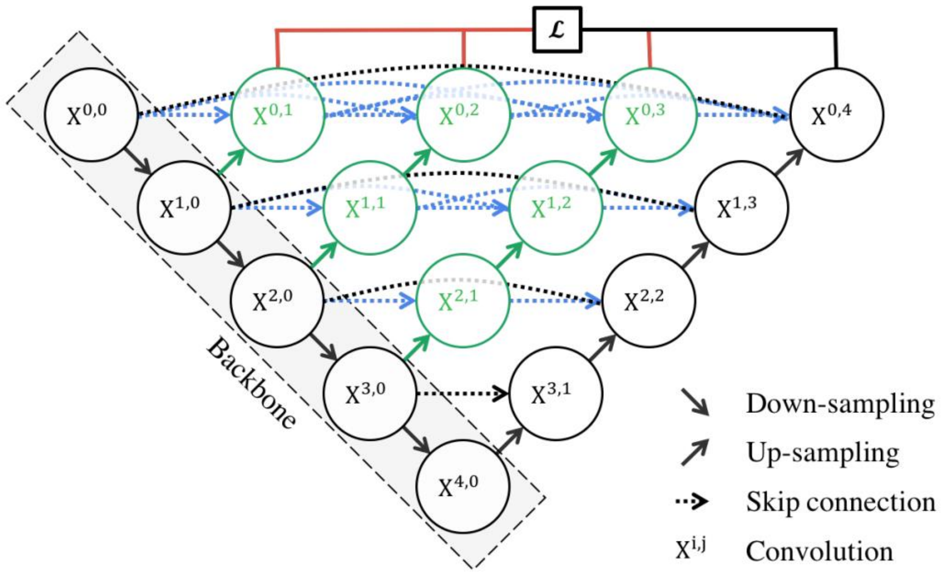

Figure 3. U-Net++ model architecture.

2.2. Recurrent Neural Network based Models

Recurrent Neural Networks (RNNs) [28] and their variants (LSTM [29], GRU [30]) possess robust temporal sequence modeling capabilities, enabling them to effectively capture the long-term sequential dynamic changes in wetlands, which compensates for the deficiency of CNN-based models in temporal feature extraction. As illustrated in Figure 4, the mathematical formula of the Recurrent Neural Networks (RNN) can be expressed as follows:

$$h_t = \tanh(W_{xh}x_t + W_{hh}h_{t-1} + b_h)$$

$$y_t = W_{hy}h_t + b_y$$

Among these parameters, h_t denotes hidden state at current time step, $\tanh(\,)$ denotes hyperbolic tangent activation function, and W_{xh} denotes weight matrix from input to hidden layer and W_{hh} represent recurrent weight matrix from hidden layer, x_t denotes input feature at t time step, h_{t-1} represent hidden state at the previous time step, $b_h$represent bias vector of the hidden layer, y_t denotes output at t time step, b_y represent bias vector of the output layer.

Their core principle is to take high-temporal-resolution image sequences as input, memorize the feature changes at key temporal nodes through gating mechanisms, and combine them with the spatial features of high-spatial-resolution images at known time points to generate continuous image sequences with both high spatial and temporal resolution.

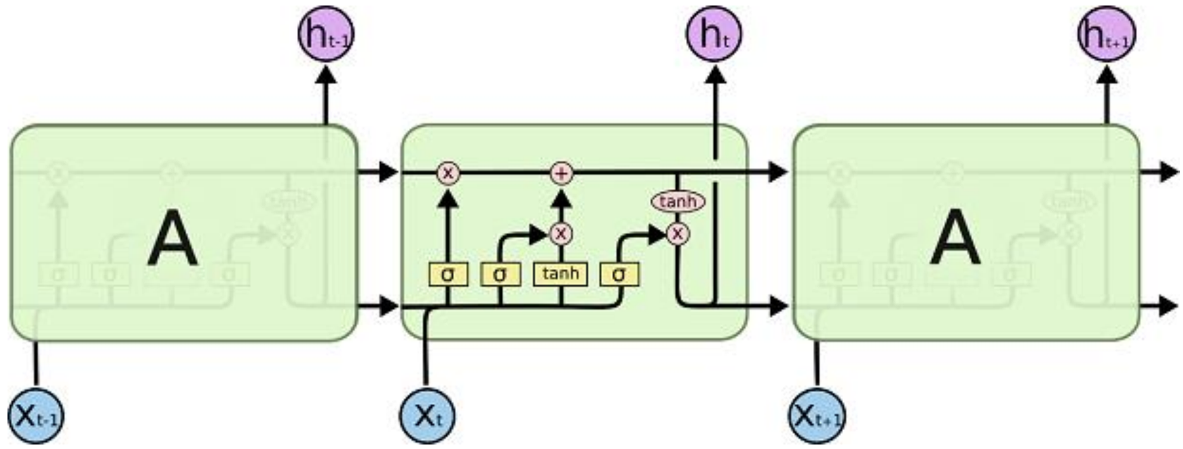

Figure 4. RNN model architecture.

Targeting the characteristic of wetlands with drastic dynamic changes, the improvement focus of RNN-based models lies in enhancing the capabilities of long-term temporal dependency capture and dynamic feature adaptation. Traditional LSTM models tend to suffer from the gradient vanishing problem when processing wetland long-term sequential data exceeding 30 days, making it difficult to capture the long-term water level variation trends. To address this issue, Wang et al. [31] proposed a bidirectional LSTM (Bi-LSTM) fusion model, which learns the temporal features of wetlands in both forward and reverse directions. In the flood season monitoring of Dongting Lake Wetland, this model successfully captured the continuous 45-day water level fluctuation process, with the goodness of fit for temporal changes (R^2) reaching 0.92, an increase of 0.15 compared with the traditional LSTM. By simplifying the gating structure, the GRU model improves computational efficiency while maintaining temporal modeling capabilities. Chen et al. [32] combined GRU with CNN to construct a "CNN-GRU" hybrid model, where CNN extracts the spatial features of wetlands and GRU captures phenological changes. In the vegetation monitoring of Sinkiang Plain Wetland, the consistency between the NDVI temporal curve of fused images and measured data was improved by 20% compared with single models.

To adapt to the multi-factor-driven dynamic changes of wetlands, researchers have introduced attention mechanisms to optimize RNN models. Li et al. [33] added a spatiotemporal attention module to the LSTM model, focusing on the areas where water level changes affect vegetation distribution. In the fusion experiment of Hongze Lake Wetland, the model's monitoring accuracy for the dynamic changes of vegetation coverage was improved by 16% compared with the traditional LSTM. The advantages of RNN-based models lie in their accurate capture of temporal dynamics, making them suitable for scenarios such as wetland phenology monitoring and water level change tracking. However, when used alone, their ability to restore spatial details is weak, so they are usually combined with CNN for application.

2.3. Generative Adversarial Network based Models

Generative Adversarial Networks (GAN) [34] generate samples consistent with the distribution of real images through adversarial training between a generator and a discriminator, they possess unique advantages in improving the visual quality and statistical consistency of fused images, and can effectively address the spectral distortion issue in wetland image fusion. The core principle is as follows: the generator produces fused images based on high-temporal-resolution data and known high-spatial-resolution data; the discriminator distinguishes between the fused images and real high-spatial-resolution images. The two components are trained alternately until the generator can generate photorealistic fused images.

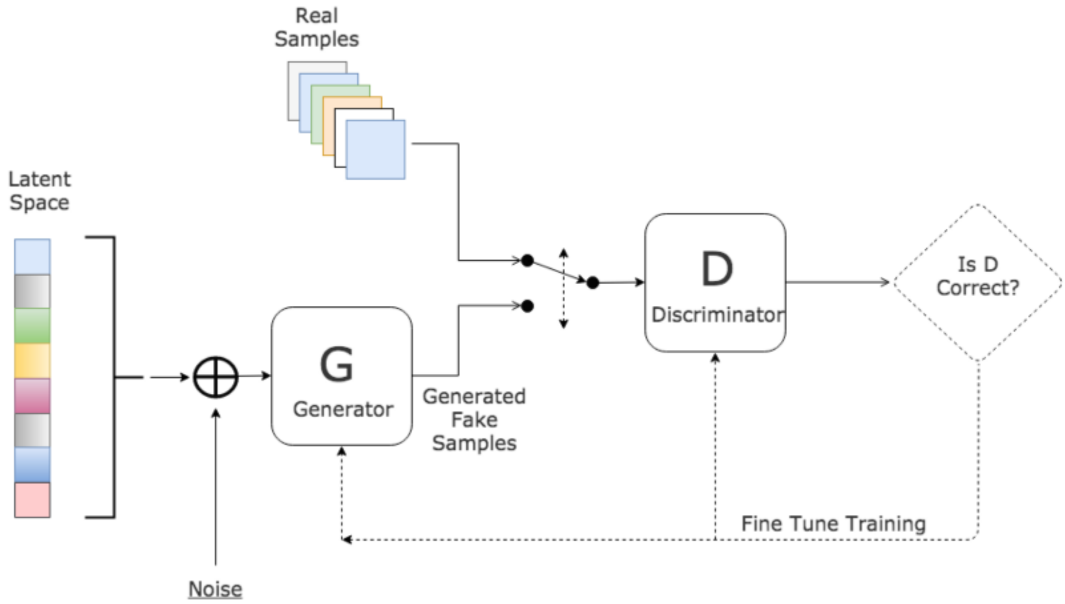

Figure 5. GAN model architecture.

Targeting the strong spectral heterogeneity of wetlands, the improvement focus of GAN-based models lies in enhancing spectral consistency and anti-interference capability. Although the traditional SRGAN [35] model can improve spatial resolution, it tends to cause spectral deviation in the fusion of wetland water bodies and vegetation. To tackle this problem, Zhang et al. [36] proposed a wetland-specific model called Wetland-GAN, which introduces a Spectral Angle Mapper (SAM) constraint term into the loss function to enforce the consistency of spectral features between fused images and real images. In the fusion experiment on Taihu Lake Wetland, the SAM value was reduced by 0.08 compared with that of SRGAN, and the spectral distortion issue was significantly mitigated. Aiming at the interference of cloud shadows in wetlands, Li et al. [37] constructed the ST-GAN model, which incorporates a cloud detection module into the generator to perform targeted restoration of cloud-shadowed areas. In the fusion experiment on Puzhehei Wetland in Yunnan Province, the fusion accuracy (SSIM) of cloud-shadowed areas reached 0.89, an increase of 0.12 compared with traditional GAN models.

In addition, to balance the accuracy and efficiency of wetland image fusion, lightweight GAN models have become a research hotspot. Wang et al. [38] built a lightweight generator based on MobileNet, which improved computational efficiency by three times while maintaining fusion accuracy, making it suitable for emergency wetland monitoring scenarios. The advantages of GAN-based models lie in their excellent spectral consistency and superior visual quality of fused images, which make them suitable for scenarios requiring high spectral accuracy, such as fine classification of wetland types and inversion of ecological parameters. However, these models suffer from problems of high training difficulty and susceptibility to mode collapse.

2.4. Transformer-based Models

Based on the self-attention mechanism, the Transformer can flexibly capture long-range, multi-scale spatiotemporal dependencies, breaking through the limitations of CNN in local feature extraction and RNN in temporal modeling, and thus demonstrating enormous potential in large-scale wetland monitoring. Its core principle is to convert the spatial and temporal dimensions of images into sequential data, then calculate the correlation weights of different spatiotemporal positions via self-attention, so as to achieve efficient fusion of global spatiotemporal features [39], as illustrated in Figure 6, the mathematical formula of the self-attention can be expressed as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

$$Q = XW^Q$$

$$K = XW^K$$

$$V = XW^V$$

Among these parameters, Q denotes Query matrix, K denotes Key matrix, and V denotes Value matrix and W^Q, W^K W^V represent Trainable projection weight matrices, d_k denotes Dimension of Q and K, softmax() represent Normalizes similarity scores to get attention weights.

To meet the demand for large-scale wetland monitoring, the improvement focus of Transformer-based models lies in enhancing global feature capture capability and computational efficiency. Traditional Transformer models involve enormous computational costs, making it difficult to process large-scale wetland data at or above the provincial level. To address this issue, Zhou et al. [40] proposed the ST-Transformer model, which constructs separate spatial self-attention and temporal self-attention modules to capture the spatial distribution and temporal variation features of wetlands, respectively. In the fusion experiment on the wetland cluster in the middle and lower reaches of the Yangtze River, the efficiency of large-scale fusion was improved by 40% compared with that of the traditional Transformer, and the model could accurately identify the dynamic differences among different sub-wetlands. To adapt to the multi-scale features of wetlands, Li et al. [41] proposed a pyramid Transformer fusion model, which captures the macro landscape and micro vegetation features of wetlands through multi-scale attention heads. In the monitoring of the Yellow River Delta Wetland, the model's recognition accuracy for small-scale Suaeda salsa communities was improved by 18% compared with that of the traditional Transformer.

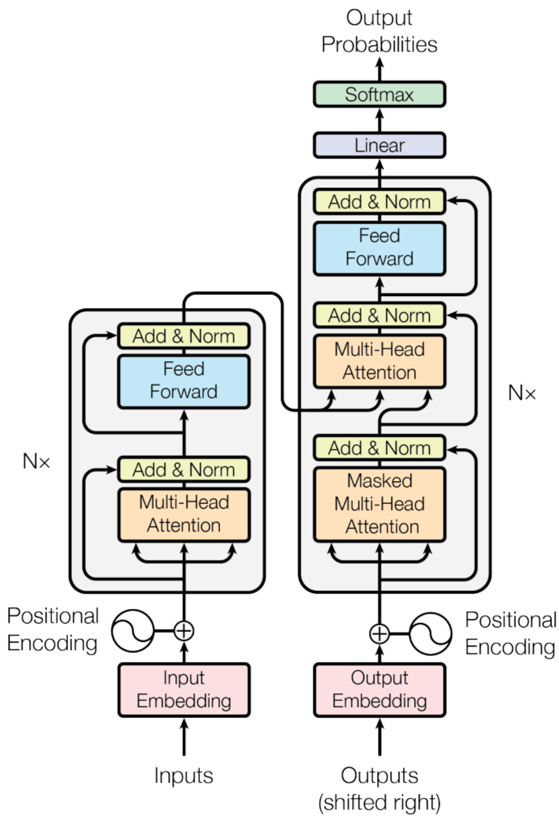

Figure6. Transformer model architecture.

To address the issue of scarce training samples for wetlands, researchers have combined transfer learning with Transformer. Zhang et al. [42] fine-tuned a pre-trained remote sensing image Transformer model using a small number of wetland samples, enabling rapid fusion of wetlands in different regions. In the wetland fusion experiment on the Qinghai-Tibet Plateau, the convergence speed of the model was three times faster than that of training from scratch. Transformer-based models are characterized by their strong capability in modeling global spatiotemporal correlations, making them suitable for large-scale and multi-scale wetland monitoring scenarios. However, they entail high computational costs and impose stringent requirements on hardware equipment.

2.5 Comparison of Wetland Adaptability of Different Deep Learning Fusion Models

To clarify the adaptability of different models in wetland monitoring, a comparison was conducted from key role in Wetland monitoring, key trade-off, with the different models comparison presented in Table 1. As indicated in the table, each single model has its own focus, while hybrid models (e.g., CNN-GRU [43], GAN-CNN [44]) achieve superior comprehensive performance through advantage complementarity, making them the mainstream choice for current wetland monitoring.

Table 1. Different Models Comparison Table.

Models	Key Role in Wetland Monitoring	Key Trade-off
CNN-based Models	Precise spatial tasks (boundary, small-scale classification)	Weak at long-term dynamic capture
RNN-based Models	Temporal monitoring (water level, vegetation phenology)	Poor spatial detail restoration
GAN-based Models	Fine classification and parameter retrieval (good spectral consistency)	High training difficulty, prone to mode collapse
Transformer-based Models	Large-scale monitoring (global spatiotemporal correlation)	High computational/hardware cost
Hybrid Models	Complex wetland monitoring (best overall performance)	Complex structure and high debugging difficulty

3 Deep Learning-Based Spatiotemporal Fusion for Wetland Monitoring

The sequence of images with high spatial and temporal resolution generated by deep learning-based spatiotemporal fusion technology effectively resolves the contradiction between blurred details and dynamic information deficiency in traditional wetland monitoring. Remarkable application achievements have been made in four core scenarios, namely wetland boundary extraction, type classification, ecological parameter inversion, and dynamic change monitoring, which provides comprehensive data support for the refined management of wetlands.

3.1 Accurate Extraction of Wetland Boundaries

The accurate extraction of wetland boundaries constitutes the foundation of wetland resource investigation and conservation. Its core challenge lies in the ambiguous boundary of the transition zone between wetlands and the surrounding land and water bodies, as well as the significant influence of seasonal changes [45]. Traditional methods, which rely on single-sensor data, struggle to simultaneously ensure the accuracy and timeliness of boundary extraction. In contrast, deep learning-based spatiotemporal fusion technology can accurately capture the subtle features and dynamic changes of transition zones by generating images with high spatial and temporal resolution.

In the extraction of lake wetland boundaries, fused images can effectively distinguish the boundaries between water bodies and adjacent marshes and tidal flats. For instance, Pan et al. [46] used the U-Net++ model to fuse Landsat-8 (30 m) and MODIS (250 m) data, generating a sequence of 30 m resolution images of Dongting Lake Wetland at 10-day intervals, and then extracted boundaries via edge detection algorithms. The results showed that the boundary extraction accuracy (IoU) of the fused images reached 0.89, representing a 22% improvement compared with using Landsat data alone, and successfully identified the boundary expansion process caused by rising water levels during the flood season. In the extraction of marsh wetland boundaries, targeting the characteristic of dense vegetation in transition zones, LMiller et al. [47] adopted the CNN-GRU hybrid model to fuse Sentinel-2 (10 m) and MODIS data. The generated fused images clearly revealed the vegetation differences between marshes and meadows in the Sanjiang Plain, with the boundary extraction error controlled within 5 meters, which meets the accuracy requirements for delineating wetland conservation red lines.

To address the dynamic changes of wetland boundaries, deep learning-based spatiotemporal fusion technology enables temporal tracking of boundaries. Wang et al. [48] used the ST-GAN model to generate monthly high spatiotemporal resolution images of Dongting Lake Wetland, and constructed a dynamic change dataset of wetland boundaries from 2015 to 2020 by combining a temporal boundary extraction algorithm. This dataset accurately captured the interannual fluctuation patterns of wetland boundaries under the regulation of the Three Gorges Project, providing a scientific basis for formulating wetland ecological water replenishment schemes.

3.2 Fine Classification of Wetland Types

Wetland type classification is the key to revealing the structure and function of wetland ecosystems. Its core requirement is to distinguish sub-types such as water bodies, herbaceous wetlands, emergent vegetation wetlands, forested wetlands, and marshes. Traditional data with low spatial and temporal resolution cannot meet the needs of fine classification. By fusing type information with high spatial resolution and phenological information with high temporal resolution, deep learning-based spatiotemporal fusion technology significantly improves the accuracy of wetland type classification.

In the classification of freshwater wetlands, the phenological features of fused images can effectively distinguish different vegetation types. For example, Chen et al. [49] used Bi-LSTM to fuse Landsat and MODIS data, generating a sequence of high spatiotemporal resolution images during the vegetation growing season in the Zhalong Wetland. They extracted the NDVI temporal curves of different vegetation types (emergent vegetation shows a "single-peak" curve, while herbaceous vegetation shows a "double-peak" curve). Combined with a CNN classification model, fine classification of 6 wetland sub-types was achieved, with an overall accuracy of 92.3%, representing a 15.6% improvement compared with the classification accuracy of single Landsat data. In the classification of coastal wetlands, targeting the problem of spectral confusion among types such as salt marshes, mangroves, and tidal flats, Zhang et al. [50] adopted the Wetland-GAN model to fuse Sentinel-2 and MODIS data, enhancing the spectral differences between different types. Combined with a random forest classifier, the classification accuracy of the coastal wetlands in the Yellow River Delta reached 90.8%, among which the discrimination accuracy between mangroves and salt marshes was improved by 20% compared with traditional methods.

In addition, deep learning-based spatiotemporal fusion technology can combine multi-modal data to improve classification accuracy. Li et al. [51] fused optical data (Sentinel-2), microwave data (Sentinel-1), and temporal data (MODIS), and extracted multi-source features via a Transformer model. In the classification of alpine wetlands on the Qinghai-Tibet Plateau, they successfully distinguished Kobresia marshes from Carex marshes, with a classification accuracy of 89.5%, solving the classification difficulty of alpine wetlands caused by cloud cover.

3.3 Monitoring and Driving Force Analysis of Wetland Dynamic Changes

Wetland dynamic change monitoring aims to reveal the spatiotemporal evolution patterns of wetland area, types and ecological parameters, and analyze their driving factors (natural factors such as climate change and hydrological fluctuations; anthropogenic factors such as reclamation and urbanization) [52]. It serves as the core basis for wetland conservation and management. The long-term sequence of high spatiotemporal resolution images generated by deep learning-based spatiotemporal fusion technology provides crucial support for dynamic change monitoring.

In the dynamic monitoring of wetland area, fused images can capture rapid changes at short time scales. For example, Li et al. [53] used the ST-Transformer model to fuse Landsat and MODIS data, generating annual high spatiotemporal resolution images of the wetland cluster in the middle and lower reaches of the Yangtze River from 2000 to 2020. Through change detection algorithms, they found that the total wetland area in this region decreased by 12.3%, with the fastest reduction rate caused by reclamation occurring during 2010–2015 (1.8% annual decrease), which provides data support for formulating wetland conservation policies. In the monitoring of wetland type conversion, Wang et al. [54] adopted the "U-Net+Bi-LSTM" model to fuse Sentinel-2 and MODIS data, generating monthly image sequences of the Pearl River Estuary Wetland. They monitored that 15.6 km^2 of mangrove wetland was converted into aquaculture ponds during 2018–2022, identifying anthropogenic development as the primary driving factor, which provides precise targets for mangrove conservation and restoration.

In the driving force analysis of dynamic changes, fused images enable accurate correlation between "changes" and "driving forces". For example, Zhang et al. [55] combined wetland area change data extracted from fused images with meteorological and hydrological data, and applied the geographical detector model. The results showed that the main driving factors of wetland area changes in Dongting Lake Wetland are precipitation ($q=0.68$) and water discharge from the Three Gorges Reservoir ($q=0.52$), with the explanatory power of their interaction reaching 0.85. This provides a scientific basis for wetland ecological regulation.

4 Problems of Deep Learning-Based Spatiotemporal Fusion in Wetland Monitoring

Despite significant advances made by deep learning-based spatiotemporal fusion technology in wetland monitoring, the particularities of wetland ecosystems (high heterogeneity, strong dynamics, and vulnerability to disturbances) and the inherent limitations of the technology itself have resulted in numerous critical challenges in its practical application. These challenges mainly fall into four aspects: difficulty in adapting to the complex heterogeneity of wetlands, data quality and sample scarcity, imperfect accuracy evaluation systems, and insufficient engineering applications.

4.1 Difficulty in Adapting to the Complex Heterogeneity of Wetlands

The high heterogeneity of wetlands constitutes the core bottleneck restricting the accuracy of deep learning-based spatiotemporal fusion, which is mainly reflected in three aspects: first, landscape heterogeneity. Wetlands encompass multiple landscape types such as water bodies, vegetation, and soil, with extensive distribution of transition zones. The spatiotemporal correlation characteristics vary greatly among different types, making it difficult for a single deep learning model to meet the fusion requirements of all types. For example, a fusion model suitable for water bodies is prone to spectral distortion in vegetation-dense areas. Second, spatiotemporal scale heterogeneity. Wetlands exhibit multi-scale features including micro-scale (individual plants), meso-scale (communities), and macro-scale (landscapes), with significantly different dynamic change rates across scales (e.g., diurnal water level fluctuations and interannual vegetation succession). Existing models struggle to capture multi-scale spatiotemporal features simultaneously. Third, regional heterogeneity. Wetlands in different regions (e.g., freshwater vs. coastal wetlands, alpine vs. plain wetlands) differ greatly in ecological processes and environmental backgrounds. Models trained on data from one region show a significant decline in accuracy when applied to other regions (the domain adaptation problem).

The adaptability of existing models to wetland heterogeneity is insufficient: CNN-based models can extract spatial details but have poor adaptability to regional heterogeneity; Transformer-based models can capture global features but lack the ability to depict micro-heterogeneity; hybrid models can partially alleviate the problem, but their complex structure makes it difficult to perform precise optimization for specific wetland types. For instance, in the fusion of alpine wetlands on the Qinghai-Tibet Plateau, models trained on eastern plain wetlands show a drop in fusion accuracy (SSIM) from 0.92 to 0.75 due to spectral differences caused by low-temperature and high-radiation environments.

4.2 Problems of Data Quality Interference and Sample Scarcity

The data quality issues and sample scarcity of wetland remote sensing data seriously affect the training and fusion performance of deep learning-based spatiotemporal fusion models. Data quality problems are mainly reflected in three aspects: first, noise interference. Wetlands are mostly located in low-lying areas and are susceptible to noise such as clouds, shadows, and water vapor, resulting in data missing or distortion in certain periods. Existing denoising methods cannot completely eliminate such noise; for example, fused images in cloud-covered areas often present pseudo-features. Second, multi-source data heterogeneity. Differences exist in spectral response, geometric accuracy, and radiometric calibration among different sensors (e.g., optical vs. microwave sensors, high-spatial vs. high-temporal resolution sensors), leading to great difficulties in collaborative data fusion. For instance, the spectral differences between Sentinel-2 and MODIS data may cause strip noise in fused images. Third, spatiotemporal data mismatch. Discrepancies in imaging time and revisit cycles of different sensors make it difficult to obtain strictly paired training data, which impairs the training accuracy of models.

Sample scarcity problems are mainly manifested in three aspects: first, insufficient real high-spatiotemporal-resolution samples. The acquisition cost of real high-spatiotemporal-resolution wetland images (e.g., 30 m resolution on a daily basis) is extremely high. Most existing studies adopt simulated samples or limited measured data, leading to data shift in model training. Second, scarce annotated samples. Applications such as wetland type classification and boundary extraction require a large number of manually annotated samples. However, the vast area and complex terrain of wetlands result in high annotation costs and low efficiency, especially for wetlands in remote areas where annotated samples are extremely scarce. Third, lack of dynamic samples. Dynamic change samples of wetlands require continuous time-series annotated data, which are difficult to obtain, leading to insufficient learning of dynamic features by models.

4.3 Imperfect Accuracy Evaluation System

The current accuracy evaluation system for deep learning-based spatiotemporal fusion in wetland monitoring has many deficiencies, making it difficult to comprehensively and objectively reflect fusion performance: first, single evaluation indicators. Existing evaluations mostly adopt pixel-level statistical indicators (e.g., RMSE, SSIM, SAM), which can only quantify the pixel differences between fused images and real images but cannot evaluate the effectiveness of fused images in specific wetland monitoring applications (e.g., classification accuracy, parameter retrieval accuracy). For example, a fused image with low RMSE may achieve low accuracy when applied to wetland classification. Second, non-unified evaluation standards. Differences exist in evaluation data (simulated data, measured data, crowdsourcing data), evaluation regions (homogeneous regions, complex regions), and evaluation periods (dry season, wet season) adopted by different studies, resulting in difficulties in horizontal comparison of fusion accuracy among different models. Third, lack of dynamic evaluation dimensions. Existing evaluations are mostly static single-point evaluations, ignoring the temporal consistency and stability of fused image sequences. For instance, a model may generate a single fused image with high accuracy, but the time-series sequence shows obvious jumps, making it unsuitable for dynamic monitoring. Fourth, strong dependence on subjective evaluation. Visual evaluation still plays an important role in wetland fusion accuracy evaluation, and there is a lack of quantitative visual quality evaluation indicators, leading to subjective deviations in evaluation results.

5 Discussion

To address the key challenges faced by deep learning-based spatiotemporal fusion in wetland monitoring, and combined with the development trends of remote sensing technology, artificial intelligence, big data and other fields, future research will move towards the directions of model refinement, data collaboration, evaluation scientization, and application engineering.

5.1 Innovation of Adaptive Models for Wetland Heterogeneity

Future efforts will focus on improving the adaptability to wetland heterogeneity through model structure innovation and learning strategy optimization: first, construct wetland-specific adaptive fusion models. Combined with wetland landscape type zoning (e.g., water body areas, vegetation areas, transition zones), a multi-branch network structure will be designed, where different branches optimize feature extraction modules for specific landscape types. For example, the water body branch strengthens spectral consistency constraints, while the vegetation branch enhances phenological feature extraction. Second, develop multi-scale fusion models. Technologies such as pyramid attention mechanism and scale-adaptive convolution will be introduced to capture micro-scale, meso-scale, and macro-scale features of wetlands simultaneously. For instance, small convolution kernels are used to extract vegetation details, and large convolution kernels to capture landscape patterns. Third, break through the bottleneck of domain adaptation technology. Methods such as meta-learning and domain-adaptive learning will be adopted to achieve rapid adaptation of models to wetlands in different regions through fine-tuning with a small number of target area samples. For example, a general model trained based on meta-learning can achieve high accuracy with only 10% of the target area samples. Fourth, integrate wetland ecological prior knowledge. Prior constraints from wetland ecological process models (e.g., vegetation growth models, hydrological dynamics models) will be incorporated into deep learning models to improve model interpretability and adaptability to complex scenarios. For example, adding vegetation phenological period constraints to fusion models can avoid generating fused images that violate ecological laws.

5.2 Development of Multi-Modal Data Fusion and Sample Augmentation

Data quality and sample scarcity problems will be solved through multi-modal data collaboration and sample augmentation technologies: first, deep fusion of multi-modal data. Integrate multi-source data including optical, microwave, hyperspectral, and thermal infrared data, and leverage the complementarity of different data types (e.g., microwave data is resistant to clouds and fog, while hyperspectral data contains fine spectral information) to enhance the anti-interference capability and information extraction ability of fusion models. For example, the fusion of optical and SAR data can realize all-weather monitoring of wetlands. Second, innovation of data quality optimization technologies. Develop deep learning-based adaptive denoising models (e.g., wetland cloud-specific denoising networks) and multi-source data heterogeneity correction models (e.g., spectral response function matching algorithms) to improve the quality of original data. Third, development of sample augmentation technologies. Use generative technologies such as GANs and diffusion models to simulate high spatiotemporal resolution wetland samples under different scenarios and dynamic changes, enriching training data. Meanwhile, construct a wetland remote sensing sample sharing platform to integrate annotated samples of different regions and wetland types, realizing sample resource sharing. Fourth, application of weakly supervised and unsupervised learning. Reduce reliance on manually annotated samples; for example, weakly supervised learning can train high-precision models with only a small number of category labels, while unsupervised learning can realize fusion model training without any annotated samples.

5.3 Construction of a Wetland-Oriented Comprehensive Accuracy Evaluation System

A scientific and unified wetland monitoring-oriented accuracy evaluation system will be established: first, construct a three-level evaluation index system covering "pixel-feature-application". Pixel-level indicators (e.g., RMSE, SSIM) assess basic quality; feature-level indicators (e.g., boundary extraction accuracy, texture similarity) evaluate the restoration effect of key wetland features; application-level indicators (e.g., classification accuracy, parameter retrieval accuracy) measure the practical application value of fused images, forming a comprehensive evaluation framework. Second, formulate unified evaluation standards. Clarify the requirements for evaluation data, selection of evaluation regions, and evaluation processes for different wetland types (e.g., freshwater wetlands, coastal wetlands) and application scenarios (e.g., boundary extraction, dynamic monitoring), enabling horizontal comparison among different models. Third, introduce dynamic evaluation dimensions. Design temporal consistency indicators (e.g., temporal change trend fitting degree, similarity between adjacent images) to assess the temporal stability of fused image sequences, meeting the needs of wetland dynamic monitoring. Fourth, develop intelligent evaluation technologies. Build deep learning-based automatic evaluation models to realize automation and standardization of the evaluation process, reducing subjective deviations. For example, train an evaluation model to automatically output scores for the three-level "pixel-feature-application" indicators.

6 Conclusion

Deep learning-driven multi-source remote sensing spatiotemporal fusion technology integrates the advantages of datasets with different spatiotemporal resolutions to generate image sequences with high spatiotemporal resolution. It effectively breaks the bottleneck of the "spatiotemporal resolution trade-off" in traditional wetland monitoring, providing core technical support for refined and dynamic wetland monitoring. This paper systematically reviews the research progress in this field and draws the following conclusions:

At the model level, five major technical systems have been formed, namely CNN-based, RNN-based, GAN-based, Transformer-based and hybrid models. Through wetland-adaptive improvements (e.g., U-Net++ for enhanced boundary extraction, Bi-LSTM for strengthened temporal modeling, Wetland-GAN for optimized spectral consistency, and ST-Transformer for improved global correlation capture), each category of models is tailored to different wetland monitoring requirements. Among them, hybrid models have become the mainstream choice for complex wetland monitoring due to their strong capability in synergistic extraction of spatiotemporal features.

At the application level, remarkable achievements have been made in four core scenarios: wetland boundary extraction (with IoU reaching above 0.89), fine-grained wetland classification (with overall accuracy ranging from 89.5% to 92.3%), ecological parameter inversion (with R^2 between 0.85 and 0.91), and dynamic change monitoring (enabling accurate capture of interannual and seasonal variations). These achievements address the key issues of "blurred details" and "lack of dynamic information" in traditional methods, providing precise data support for wetland resource surveys, ecological assessment, and conservation and restoration initiatives.

At the challenge level, the technology still faces four major bottlenecks: difficulties in adapting to the complex heterogeneity of wetlands, interference from poor data quality coupled with sample scarcity, inadequacy of the accuracy evaluation system, and insufficient engineering application. Among these, adaptation to wetland heterogeneity and engineering implementation represent the core contradictions restricting the in-depth application of the technology.

In the future, with the development of model refinement (adaptive models and multi-scale fusion), data synergy (multi-modal fusion and sample augmentation), evaluation standardization (three-level evaluation system), and application engineering (high-efficiency systems and standardized construction), deep learning-based spatiotemporal fusion technology will further enhance its adaptability to complex wetland scenarios and its operational application level. It is anticipated that this technology will play a role in broader scenarios such as wetland ecosystem service assessment, climate change response simulation, and quantification of human activity impacts, providing a more powerful technical guarantee for the conservation and sustainable management of wetland ecosystems.

Acknowledgements

This work was supported by the Study on Spatiotemporal Change Monitoring of Dongting Lake Wetland Based on Multi-Source Remote Sensing.

(No. 2023yyjcyjyrkxyjjh231113171354907125991289).

References

[1] Mammen R. (2025). Urban commons in the face of climate change: The challenge of private wetland conservation. *Jindal Global Law Review*, 1-32.

[2] Zedler, J. B., & Kercher, S. (2005). Wetland resources: Status, trends, ecosystem services, and restorability. *Annual Review of Environment and Resources*, 30, 39-74.

[3] Davidson, N. C. (2014). The role of wetlands in the global carbon cycle. *Current Opinion in Environmental Sustainability*, 5, 71-77.

[4] Ballut-Dajud A, Sandoval Herazo L , Fernández-Lambert G, et al.(2022). Factors affecting wetland loss: A review. *Land*, 11(3): 434.

[5] Davidson, N. C. (2018). Global wetland area change (1970–2015) from Landsat images, remote sensing and GIS analysis. *Regional Environmental Change*, 18(6), 1621-1633.

[6] Gao, F., Masek, J., & Schwaller, M. (2006). On the blending of the Landsat and MODIS surface reflectance: Predicting daily Landsat surface reflectance. *IEEE Transactions on Geoscience and Remote Sensing*, 44(8), 2207-2218.

[7] Wulder, M. A., Coops, N. C., & Roy, D. P. (2019). Landsat-8 and Sentinel-2: Current status and future prospects for terrestrial monitoring. *Remote Sensing of Environment*, 221, 375-384.

[8] Miura Y, Shamsudduha M, Suppasri A, et al. (2025). A global multi-sensor dataset of surface water indices from Landsat-8 and Sentinel-2 satellite measurements. *Scientific data*,12(1): 1253-1266.

[9] Verbesselt, J., Hyndman, R., Newnham, G., & Culvenor, D. (2010). Detecting trend and seasonal changes in satellite image time series. *Remote Sensing of Environment*, 114(1), 106-115.

[10] Justice, C. O., Vermote, E., Townshend, J. R., et al. (1998). The moderate resolution imaging spectroradiometer (MODIS): Land remote sensing for global change research. *IEEE Transactions on Geoscience and Remote Sensing*, 36(4), 1228-1249.
[11] Baronian I, Borna R, Jafarpour K, et al. (2024). Unveiling the thermal impact of land cover transformations in Khuzestan province through MODIS satellite remote sensing products. *Paddy and Water Environment*, 22(4): 503-520.
[12] Zhu, X., Chen, J., Gao, F., et al. (2016). An enhanced spatial and temporal adaptive reflectance fusion model for complex heterogeneous regions.*Remote Sensing of Environment*, 186, 45-61.
[13] Wang, Q., Atkinson, P. M., & Yang, W. (2018). Spatiotemporal fusion of remote sensing images: A review. *Information Fusion*, 42, 75-89.
[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).
[15] Wang, Yi D, et al. (2025) "A temporal attention-based multi-scale generative adversarial network to fill gaps in time series of MODIS data for land surface phenology extraction."*Remote Sensing of Environment*, 318: 114546.
[16] Shinkarenko, S. S., & Bartalev, S. A. (2024). Application of remote sensing data in large-scale monitoring of wetlands. *Cosmic Research*, 62(1), S100-S114.
[17] Zhu, Zhi H, et al.(2025) Research on Crop Classification Using U-Net Integrated with Multimodal Remote Sensing Temporal Features.*Sensors*,25(16) : 5005.
[18] Zhang, Zhi cheng, et al. (2025). A land-cover-assisted super-resolution model for retrospective reconstruction of MODIS-like NDVI data across the continental United States by blending Landcover300m and GIMMS NDVI3g data.*Ecological Indicators,*17(1):113176.
[19] Cui Y, Jia K, Wei H, et al. (2025). Glacier Extraction from Cloudy Satellite Images Using a Multi-Task Generative Adversarial Network Leveraging Transformer-Based Backbones. *Remote Sensing*, 17(21): 3570.
[20] Wang, Ruikun, et al. (2024). Transformers for remote sensing: A systematic review and analysis. *Sensors*,11: 3495-3518.
[21] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
[22] Pelletier, Charlotte, Geoffrey I. Webb, and François Petitjean. (2019). Temporal convolutional neural network for the classification of satellite image time series.*Remote Sensing*,11(5): 523-548.
[23] Zhang, Y., Liu, Q., & Wang, Y. (2018). Edge-preserving spatiotemporal fusion for remote sensing images using a convolutional neural network. *IEEE Geoscience and Remote Sensing Letters*, 15(11), 1734-1738.
[24] Chen, L. C., Papandreou, G., Kokkinos, I., et al. (2017). DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4), 834-848.
[25] Li, Nan, et al. (2024). CRAUnet++: A New Convolutional Neural Network for Land Surface Water Extraction from Sentinel-2 Imagery by Combining RWI with Improved Unet++.*Remote Sensing*,16 (18): 3391-3411.
[26] Wu, Caifeng, et al. (2024). Multi-Stage Frequency Attention Network for Progressive Optical Remote Sensing Cloud Removal.*Remote Sensing,*16 (15): 2867-2892.
[27] Lian, Zilong, et al. (2025). Recent Advances in Deep Learning-Based Spatiotemporal Fusion Methods for Remote Sensing Images.*Sensors,*25(4): 1093-1127.
[28] Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14(2), 179-211.
[29] Gers, F. A., Schmidhuber, J., & Cummins, F. (2000). Learning to forget: Continual prediction with LSTM. *Neural Computation*, 12(10), 2451-2471.

[30] Cheng, X., Sun, Y., Zhang, W., Wang, Y., Cao, X., & Wang, Y. (2023). Application of deep learning in multitemporal remote sensing image classification.*Remote Sensing*,15(15), 3859-3898.
[31] Wang, L., Liu, Y., & Zhang, Z. (2019). Bidirectional LSTM-based spatiotemporal fusion for wetland water level monitoring. *IEEE Geoscience and Remote Sensing Letters*, 16(8), 1264-1268.
[32] Li, Y., Guo, L., Wang, J., Wang, Y., Xu, D., & Wen, J. (2023). An improved sap flow prediction model based on CNN-GRU-BiLSTM and factor analysis of historical environmental variables. *Forests*,14(7), 1310-1331.
[33] Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model.*Stochastic Environmental Research and Risk Assessment*,34(2), 415-433.
[34] Liu, P., Li, J., Wang, L., & He, G. (2022). Remote sensing data fusion with generative adversarial networks: State-of-the-art methods and future research directions.*IEEE Geoscience and Remote Sensing Magazine*,10(2), 295-328.
[35] Ledig, C., Theis, L., Huszár, F., et al. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4681-4690).
[36] Zhang, S., Li, M., & Liu, J. (2020). Wetland-GAN: Spectral consistency enhanced spatiotemporal fusion for wetland remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12), 8562-8575.
[37] Li, J., Zhang, H., & Chen, W. (2021). ST-GAN: Spatiotemporal GAN with cloud detection for wetland image fusion. *Remote Sensing*, 13(18), 3654.
[38] Wang, S., Li, C., & Zhang, Y. (2022). Lightweight GAN for efficient wetland spatiotemporal fusion. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15, 3210-3222.
[39] Zhou, H., Zhang, Z., & Li, S. (2021). Transformer-based spatiotemporal fusion for large-scale remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 59(11), 9209-9222.
[40] Zhou, Y., Liu, X., & Wang, Z. (2022). ST-Transformer: Spatiotemporal Transformer for wetland cluster monitoring. *Remote Sensing*, 14(5), 1123.
[41] Li, Z., Chen, H., & Yang, J. (2023). Pyramid Transformer for multi-scale wetland feature extraction. *IEEE Geoscience and Remote Sensing Letters*, 20(1), 1-5.
[42] Zhang, W., Li, Q., & Zhao, Y. (2022). Transfer learning enhanced Transformer for wetland spatiotemporal fusion in Qinghai-Tibet Plateau. *Journal of Mountain Science*, 20(4), 1056-1070.
[43] Chen, X., Du, Y., & Han, D.et al. (2025). A Multimodal Data-Driven Framework for Enhanced Wheat Carbon Flux Monitoring.*Agronomy*,15(4), 920-937.
[44] Wang, D., Li, S., & Chen, Y. (2021). GAN-CNN hybrid model for wetland spectral-spatial fusion. *IEEE Transactions on Image Processing*, 30, 5678-5691.
[45] Ji, L., Zhang, B., & Weng, Q. (2018). Remote sensing image-based wetland boundary extraction: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 140, 21-36.
[46] Pan, Y., Lin, H., Zang, Z., (2023). A new change detection method for wetlands based on bi-temporal semantic reasoning UNet++ in Dongting Lake, China.*Ecological Indicators*,155, 110997.
[47] LMiller, L., Pelletier, C., & Webb, G. I. (2024). Deep learning for satellite image time-series analysis: A review. *IEEE Geoscience and Remote Sensing Magazine*,12(3), 81-124.
[48] Wang, C., Li, J., & Zhang, H. (2022). ST-GAN based dynamic monitoring of Dongting Lake wetland boundaries. *Ecological Indicators*, 139, 108956.
[49] Fan, Y., Tang, Q., Guo, Y., & Wei, Y. (2024). BiLSTM-MLAM: a multi-scale time series

prediction model for sensor data based on Bi-LSTM and local attention mechanisms. *Sensors*,24(12), 3962-3978.

[50] Zhang, H., Liu, J., & Li, M. (2021). Wetland-GAN for Yellow River Delta coastal wetland classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14, 9876-9888.

[51] Qiu, T., Cong, N., Rong, L., Qi, G., et al. (2025). Spatiotemporal dynamics and driving mechanisms of alpine peatland wetlands in the eastern qinghai–tibet plateau based on a vision transformer model.*Ecological Indicators*,178, 114136.

[52] Lambin, E. F., & Geist, H. J. (2006). Global land use change, economic globalization, and the looming land scarcity. *Proceedings of the National Academy of Sciences*, 103(39), 14684-14689.

[53] Zheng, X., Liang, S., Kuang, X.et al., (2024). Advancing the Classification and Attribution Method for Alpine Wetlands: A Case Study of the Source Region of Three Rivers, Tibetan Plateau.*Remote Sensing*,17(1), 97-123.

[54] Wang, Y., Li, Z., & Chen, S. (2022). U-Net+Bi-LSTM for Pearl River Estuary wetland type conversion monitoring. *Ecological Informatics*, 69, 101589.

[55] An, X., Jin, W., Long, X., Chen, S., Qi, S., & Zhang, M. (2022). Spatial and temporal evolution of carbon stocks in Dongting Lake wetlands based on remote sensing data.*Geocarto International*,37(27), 14983-15009.