Information Fusion 114 (2025) 102689

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

STSNet: A cross-spatial resolution multi-modal remote sensing deep fusion network for high resolution land-cover segmentation

Beibei Yu [a], Jiayi Li [b,*], Xin Huang [a,b]

[a] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China
[b] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China

ARTICLE INFO

Keywords:
Land-cover segmentation
Multi-modal
High-resolution
Spatio-temporal-spectral
Cross-spatial resolution
Deep learning

ABSTRACT

Recently, deep learning models have found extensive application in high-resolution land-cover segmentation research. However, the most current research still suffers from issues such as insufficient utilization of multi-modal information, which limits further improvement in high-resolution land-cover segmentation accuracy. Moreover, differences in the size and spatial resolution of multi-modal datasets collectively pose challenges to multi-modal land-cover segmentation. Therefore, we propose a high-resolution land-cover segmentation network (STSNet) with cross-spatial resolution spatio-temporal-spectral deep fusion. This network effectively utilizes spatio-temporal-spectral features to achieve information complementary among multi-modal data. Specifically, STSNet consists of four components: (1) A high resolution and multi-scale spatial-spectral encoder to jointly extract subtle spatial-spectral features in hyperspectral and high spatial resolution images. (2) A long-term spatio-temporal encoder formulated by spectral convolution and spatio-temporal transformer block to simultaneously delineates the spatial, temporal and spectral information in dense time series Sentinel-2 imagery. (3) A cross-resolution fusion module to alleviate the spatial resolution differences between multi-modal data and effectively leverages complementary spatio-temporal-spectral information. (4) A multi-scale decoder integrates multi-scale information from multi-modal data. We utilized airborne hyperspectral remote sensing imagery from the Shenyang region of China in 2020, with a spatial resolution of 1authors declare that they have no known competm, a spectral number of 249, and a spectral resolution $\leq$ 5 nm, and its Sentinel dense time-series images acquired in the same period with a spatial resolution of 10 m, a spectral number of 10, and a time-series number of 31. These datasets were combined to generate a multi-modal dataset called WHU-H^{2}SR-MT, which is the first open accessed large-scale high spatio-temporal-spectral satellite remote sensing dataset (*i.e.*, with >2500 image pairs sized 300 m $\times$ 300 m for each). Additionally, we employed two open-source datasets to validate the effectiveness of the proposed modules. Extensive experiments show that our multi-scale spatial-spectral encoder, spatio-temporal encoder, and cross-resolution fusion module outperform existing state-of-the-art (SOTA) algorithms in terms of overall performance on high-resolution land-cover segmentation. The new multi-modal dataset will be made available at http://irsip.whu.edu.cn/resources/resources_en_v2.php, along with the corresponding code for accessing and utilizing the dataset at https://github.com/RS-Mage/STSNet.

1. Introduction

Land cover refers to the physical coverage of materials on the Earth's surface [1,2]. High spatial resolution land use and cover information is essential for decision-making in several areas. This information provides a vital foundation for conserving and sustainably utilizing natural resources by accurately monitoring changes in forest cover, water resource distribution, and soil erosion [3]. In the monitoring of geographic national conditions, high-resolution remote sensing data provide timely and accurate support for governments and relevant organizations in urban expansion, infrastructure construction, and natural disaster monitoring, ensuring scientific and effective decision-making [4,5]. Additionally, high-resolution data play a crucial role in achieving carbon neutrality by helping assess carbon stocks and the dynamics of the carbon cycle. These data facilitate the formulation of effective carbon management strategies that support carbon emission reduction and sustainable development goals [6–8]. In these applications, remote sensing technology, with its advantages of wide coverage, high

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China.
E-mail address: zjjerica@whu.edu.cn (J. Li).

https://doi.org/10.1016/j.inffus.2024.102689
Received 7 June 2024; Received in revised form 25 August 2024; Accepted 7 September 2024
Available online 8 September 2024

spatio-temporal-spectral resolution, and ability to access information without being constrained by ground conditions, has become an essential tool for obtaining high-resolution land-cover information, promoting technological progress and innovation across various fields [9,10]. By its very nature, land-cover segmentation assigns a land class label to each pixel of remote sensing imagery by means of semantic segmentation techniques [11–13]. Recently, deep learning has opened up the possibility of automated, fine-grained land-cover segmentation at high spatio-temporal-spectral resolution through end-to-end network training on large-scale samples [14–16].

High spatial resolution (*i.e.*, sub-meter to meter-level spatial resolution) imagery provides rich information on shape, texture, and structure for land-cover segmentation [17–19]. Hyperspectral imagery (*i.e.*, having a spectral resolution of $\leq$20 nm and covering the visible to near-infrared range) can collect information across the entire electromagnetic spectrum [20], and its fine spectral diagnostic capability is an effective means of distinguishing between land-cover materials. As sensor technology advances, some datasets with hyperspectral and high spatial resolution (H^2SR) are now available [21,22], which can comprehensively utilize rich high spatial-spectral information for land-cover segmentation. However, the above single-date remote sensing data are susceptible to the effects of weather changes, resulting in missing data and significant radiation differences [23]. In contrast, multi-temporal remote sensing data can dynamically monitor land-cover changes and provide information on land phenology and seasonal changes, which will effectively complement high-resolution observations. The fusion of low-temporal and medium-spatial resolution observations (*e.g.*, Landsat imagery) with high-temporal and low-spatial resolution ones (*e.g.*, MODIS imagery) can provide medium-spatial resolution and high-temporal records. However, the medium-spatial resolution imagery it provides lack the detailed texture information in high-spatial-resolution images [24,25], leading to inferior accuracy for land-cover segmentation. While PlanetScope is capable of obtaining high spatio-temporal resolution observations from small satellites, the commercial satellite data incurs significant data acquisition costs, particularly with increasing time series data. When there is a certain degree of reduction in the spatial resolution requirement (*i.e.*, a spatial resolution of 10–20 m), open-access Sentinel and Landsat satellites can provide more denser temporal sequences, richer spectral information, and more stable observations. In general, there is currently no data with high spatio-temporal-spectral resolution acquired from the same observation platform simultaneously. Therefore, to fully leverage the advantages of spatio-temporal-spectral information for land-cover segmentation, it is of great value to develop modal fusion technology to integrate remote sensing images from multiple observation platforms [26–28]. This study releases an open-source spatio-temporal-spectral multi-modal remote sensing dataset to explore this issue. This dataset consists of (1) airborne H^2SR imagery with a 1 m spatial resolution, 249 spectral bands, and a spectral resolution of $\leq$5 nm, and (2) multi-temporal Sentinel-2 imagery from the same year with a 10 m spatial resolution, 10 spectral bands, and temporal numbers of 31.

Meanwhile, from a technical perspective, despite substantial advancements in the field of high-resolution land-cover segmentation [29,30], certain limitations persist. These limitations primarily revolve around two aspects: (1) The existing fusion mainly refers to spatio-temporal and spatial-spectral fusion [31], and most of the fusion strategies are overly concerned with the spatial information; (2) Current fusion and interpretation strategies are mainly for data with small spatial resolution differences (*e.g.*, 2–4 times differences) [32–34], and few fusion strategies focus on data with large scale differences (*e.g.*, no $<$10 times differences, see Fig. 1) [35]. When dealing with the latter cases, it often struggles to integrate global contextual information from medium spatial resolution images and edge information from high-spatial-resolution images [32,36,37].

Recently, deep learning has made significant progress and emerged with UNet [38], and other networks. These data-driven deep learning

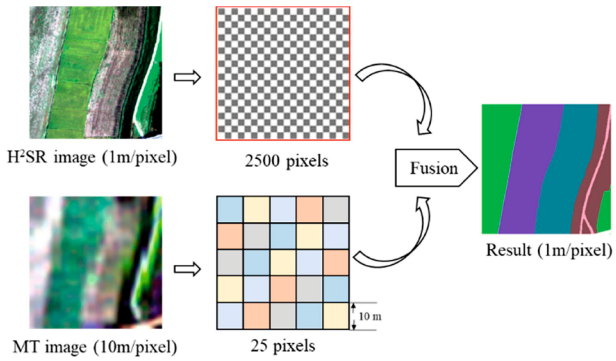

Fig. 1. Illustration of the fusion process for remote sensing images with different spatial resolutions. The figure shows how H^2SR images and MT images are fused to match spatial sizes. For example, if the 1-meter resolution image is 300 $\times$ 300 pixels, the corresponding 10-meter resolution image at the same location would be 30 $\times$ 30 pixels. The H^2SR image maps to a feature size of 50 $\times$ 50 pixels (2500 pixels), while the corresponding MT image maps to a feature size of 5 $\times$ 5 pixels (25 pixels). The pixel colors in the figure are for illustrative purposes only and have no practical significance.

methods provide new ideas for improving the accuracy of high-resolution land-cover segmentation. Due to the characteristics of H^2SR imagery, rich in spectral and spatial features, is typically captured by deep spatial-spectral network structures and multi-scale pyramidal bottleneck residual structures [34,39–41]. While deep structures help extract sematic information, they gradually degrade resolution and are prone to vanishing gradients or explosions. Multi-scale network structures perform feature fusion at each stage, facilitating better gradient flow between high-resolution and low-resolution branches. This mitigates to some extent the common problem of gradient vanishing in deep networks [42]. Nevertheless, frequent cross-scale connections in multi-scale networks can lead to multiple compressions and reconstructions of features, potentially degrading feature details and consistency. As an alternative for CNNs, transformer structures have emerged as an option for improving hyperspectral imagery segmentation by modeling long-range dependencies [43]. However, transformer models typically require substantial computational resources and extensive training data to learn global structural information. In contrast, depth-wise separable convolutions [44] (DSCs) effectively capture correlations, textures, shapes, and other information among individual hyperspectral bands, thereby improving the accuracy of spatial-spectral networks [45]. The core idea behind the attention mechanism is to emphasize important information through the allocation of weights while suppressing irrelevant data. In the field of remote sensing interpretation, convolutional attention [46], self-attention [47], and multi-head attention [48] have been introduced to enhance the model's ability to focus on key regions within image. Convolutional attention is designed to handle 2D data, more effectively, thereby improving the model's local feature extraction capability. In self-attention, query, key and value are all derived from the same input sequence, allowing global dependencies within the sequence to be captured. This mechanism is a core component of Transformer model. On basis of self-attention, multi-head attention enables model to compute attention in parallel across different subspaces, thereby capturing more feature information and enhancing the model's expressive power.

Single-date images only provide information at a specific time, cannot capture seasonal information, and are unsuitable for classifying land cover with temporal variations (*e.g.*, Cropland). In contrast, multi-temporal imagery performs better in land-cover segmentation tasks. Long Short-Term Memory (LSTM) models can mine time series changes in multi-temporal imagery [49]. Moreover, ConvLSTM integrates the strengths of both LSTM and CNN, primarily used to capture local spatiotemporal information in multi-temporal imagery [50,51].

Conversely, visual transformers excel at mining global spatiotemporal information through self-attention mechanisms [52]. Thus, it is meaningful to simultaneously consider local and global spatiotemporal information by ConvLSTM and visual transformer modules in a suitable way.

Currently, multi-modal fusion typically involves spatial-spectral fusion and spatio-temporal fusion. Spatial-spectral fusion focuses on combining the H^2SR features of land covers [53], commonly expressed through either direct fusion or attention-based fusion. In this context, the direct fusion strategy is primarily achieved by concatenating (or element-wise adding) spatial and spectral features [54]. Thus, it struggles to fully consider the complementary features of spatial and spectral characteristics. On the other hand, the attention-based works can assign weights to spatial and spectral features, thereby achieving the beneficial complementarity of local spatial and spectral characteristics [55]. Nonetheless, attention-based methods encounter challenges in effectively integrating multiscale spatial and spectral information.

Spatio-temporal information refers to remote sensing data capturing the Earth's surface across spatial and temporal domains. It includes geospatial distribution (*e.g.*, location, shape, size, spatial relationships) and temporal dynamics (*e.g.*, phenological changes, semantic changes over time) [56]. Spatio-temporal fusion seeks to explore the spatio-temporal information to strengthen the understanding of land cover features. Conventional spatio-temporal fusion methods typically start by dimensionality reduction of multi-temporal images, followed by stacking or weighted combination with high-spatial-resolution images [57]. However, the dimensionality reduction operation often leads to the loss of temporal dynamic information. Moreover, deep learning-based spatio-temporal fusion often employs a dual-stream network structure, fusing spatial-stream features with temporal-stream features in a similar manner like the direct fusion, to capture spatio-temporal information [58,59]. Similarly, these methods struggle to simultaneously consider both global and local spatio-temporal information. Local spatio-temporal information refers to the "shallow" spatial and temporal dynamics of adjacent time images, such as the texture of land parcels and spectral differences between adjacent period, whereas global spatio-temporal information involves the entire time series and spatial information. Taking farmland as an example, before the sowing season it may appear as bare ground, whereas presents spectral variations according to the growing states of crops (*e.g.*, sowing, tasseling, plucking, flowering, and fruiting). Global spatio-temporal information describes the spectral curve of the growth cycle of farmland, whereas local spatio-temporal information aids to accurately locate crop positions and understand the spectral differences between adjacent moments. Additionally, since multispectral temporal images typically comprising a series of multispectral images (*e.g.*, multispectral Sentinel-2 images), existing methods overlook the exploration of spectral information in multispectral temporal images [60].

Existing research on multi-modal fusion has not fully addressed the challenges of STSF, especially for fusing H^2SR images with multi-temporal data. Current strategies mainly focus on fusing images with similar resolutions (*e.g.*, 2–4 times difference), overlooking techniques for large-scale resolution differences (*e.g.*, 10 times) [61–64]. Direct fusion methods (*e.g.*, concatenation) often fail to integrate rich temporal, spectral, and spatial details effectively, while attention fusion methods, which rely on spatial and channel attention, struggle with resolution inconsistencies across sensors, limiting their ability to leverage complementary information in cross-resolution fusion.

To address the challenges associated with multi-modal fusion in high resolution land-cover segmentation, this paper proposes a spatio-temporal-spectral deep fusion network (called STSNet), which aims to fully exploit the advantages of high spatio-temporal-spectral resolution observations provided by multi-modal remote sensing imagery in land-cover segmentation. Specifically, STSNet contributes in the following aspects:

(1) To capture spatial-spectral information from the H^2SR image, we design a multi-scale encoder that performs cross-scale connections only at the beginning and the end stages of the network. This design avoids frequent cross-scale connections with compressions and reconstructions of features while enabling the fusion of high-level semantics with low-level detailed features. In addition, the spectral gated module is integrated into the multi-scale spatial-spectral encoder to enhance the diagnostic capability of hyperspectral data.
(2) To capture spatio-temporal information from the dense time series Sentinel-2 images, we propose an encoder that combines spectral convolution with a spatio-temporal transformer block. The block employs ConvLSTM to extract local spatio-temporal information and 3D self-attention to mine global spatio-temporal information.
(3) From the perspective of spatio-temporal-spectral fusion (STSF), the newly proposed cross-resolution module aims to mitigate the information loss caused by large spatial disparities in multi-modal remote sensing images. It adjusts the importance of features from each modality branch (*e.g.*, spatial-spectral and spatio-temporal features) at different resolutions, adaptively regulating their contribution to land-cover interpretation.

The remainder of the paper is organized as follows. Section 2 presents the proposed method. Section 3 describes the study area and datasets used. Section 4 evaluates performance of the proposed work, using the released multi-modal dataset and two separate open-source datasets (a H^2SR image, and a set of multi-temporal Sentinel-2 images). Finally, Section 5 concludes the paper. For reader convenience, commonly used acronyms in this paper are summarized in Table 1.

2. Methodology

To fully explore the spatio-temporal-spectral features of multi-modal data, this paper proposes a cross-spatial resolution spatio-temporal-spectral deep fusion network (STSNet) for high-resolution land-cover segmentation. The overall framework comprises the following components (see Fig. 2):

① the multi-scale spatial-spectral encoder module (see Section 2.1 for details) aims to effectively mine rich hyperspectral features and high spatial features at different scales;
② the spatio-temporal encoder module (see Section 2.2 for details) introduces a spatio-temporal transformer block to parallelly extract global and local spatio-temporal after a spectral convolution;
③ the cross-resolution fusion module (see Section 2.3 for details) fuses the spatio-temporal-spectral complementary information from the multi-modal images with different spatial resolutions;
④ the multi-scale decoder module merges the output features of ③ and the multi-scale spatial-spectral encoder features and then yields

Table 1
Acronyms and definitions.

Acronyms	Description
SOTA	State-of-the-art
H^2SR	Hyperspectral and high spatial resolution
STSF	Spatio-temporal-spectral fusion
CNN	Convolutional neural networks
DSC	Depth-wise separable convolution
LSTM	Long Short-Term Memory
MT	Multi-temporal
MLP	Multilayer Perceptron
ConvLSTM	Convolutional Long Short-Term Memory
MP	Max Pooling layer
AP	Average Pooling layer
3D W-MSA	Multi-head self attention modules with 3D regular windowing
3D SW-MSA	Multi-head self attention modules with 3D shifted windowing

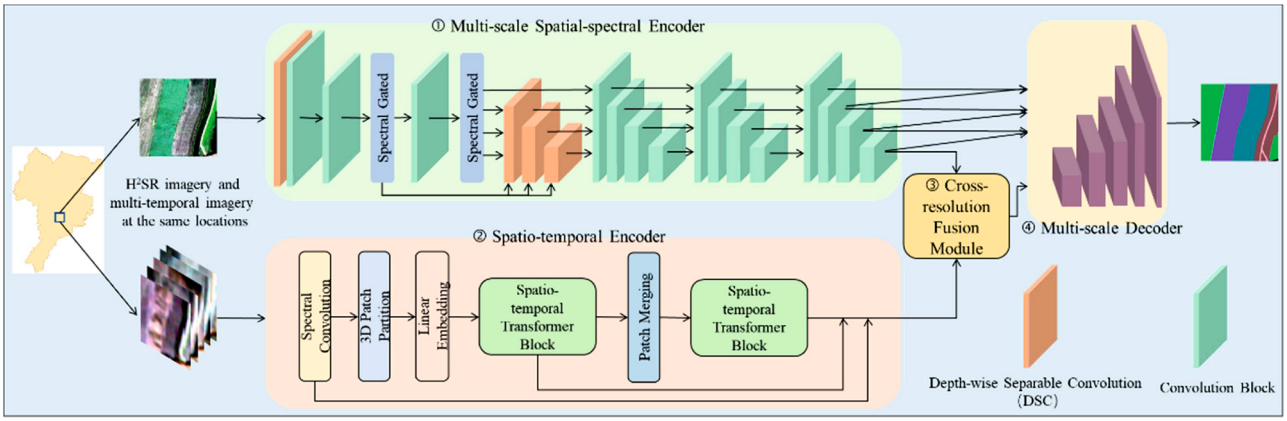

Fig. 2. Illustrates the overall framework of STSNet. It consists of ① multi-scale spatial-spectral module (light green area), including DSCs, spectral gated module, and convolutional blocks. Each convolutional block is composed of three layers of 3 × 3 convolutions, BatchNorm layers, and ReLU layers; ② spatio-temporal encoder module (refer to Fig. 5 below for details): This module includes spectral convolutions, 3D patch partition, linear embedding, and spatio-temporal transformer blocks; ③ cross-resolution fusion module (refer to Fig. 6 below for details); and ④ multi-scale decoder (light yellow area).

the final segmentation result. Firstly, 1 × 1 convolution is employed to ensure that the number of channels for features on each scale is the same. Then, fractional convolution [65] is applied to interpolating the multi-scale features to the highest resolution. Fractional convolution is a type of convolution operation that allows the convolution kernel to slide in fractional steps on the input feature. It is used to scale up the input feature map and is a variant of regular convolution. Last, the unsampled features are element-wise added, and the channel number of the final feature is adjusted to the target number of semantic class using a 1 × 1 convolution layer.

2.1. Multi-scale spatial-spectral encoder

Traditional spatial-spectral segmentation networks still employ spectral dimension reduction techniques to balance the contributions of the hyperspectral feature and the spatial feature [54,66] This often leads to insufficient exploration of the subtle spectral information. On the other hand, existing networks that focus on multi-scale feature extraction [67–69] lack exploration of the correlation and complementarity between multi-scale features. To address these issues, a multi-scale spatial-spectral encoder module is proposed to explore the multi-scale high spatial and hyperspectral characteristics of the land-covers in a tighter way. The encoder module comprises depth-wise separable convolutions (DSCs), spectral attention, and multiple convolutional blocks.

It adopts a four-branch parallel structure, each branch generates features down-sampled by a factor of 2 based on the size of the features from the previous branch. Each branch initially employs DSCs to process information between and within spectral bands, effectively capturing local spatial and spectral information. In the highest resolution branch spectral gated module is applied to weight different bands, thereby aggregating high spatial-spectral features. Specifically, adaptive average pooling layer and adaptive max pooling layer automatically calculate the size and stride of pooling layers based on the input and output (I/O) feature map sizes, thereby improving the model flexibility. The adaptive average pooling layer helps capture global spectral features, while the adaptive max pooling layer highlights local important features. Combining these two pooling layers leverages the advantages complementarity of global and local features. The formula for spectral gated module is as follows.

$$F_{SA\_AA} = \delta(MLP(AAP(F_x))) \tag{1}$$

$$F_{SA\_AM} = \delta(MLP(AMP(F_x))) \tag{2}$$

$$Spectral\_weights = F_{SA\_AA} \oplus F_{SA\_AM} \tag{3}$$

$$F_{SA} = F_x \otimes Spectral\_weights \tag{4}$$

where δ represents sigmoid function, *AAP* represents adaptive average pooling layer, *AMP* represents adaptive max pool layer, F_X represents input features, F_{SA_AA} represents the features after adaptive average pooling layer, F_{SA_AM} represents the features after adaptive max pooling layer, F_{SA} represents the final output features weighted by spectral gated module, $\otimes$ represents element-wise multiplication, and $\oplus$ represents element-wise addition.

Cross-scale connections involve up-sampling the feature maps of low-resolution branches to the same resolution as the high-resolution branches at each stage of the network, and then fusing them by pixel-wise addition. We observed that in existing high-resolution networks, frequent cross-scale connections lead to multiple compressions and reconstructions of features, resulting in information loss [70]. This is particularly disadvantageous for H^2SR imagery. To address this, performing two cross-scale connections at the beginning and end stages of network helps achieve a better balance in the fusion of spatial-spectral information. In the first cross-scale connection, the network can capture diverse spatial-spectral information at different scales, enhancing the richness of feature representation. In the second cross-scale connection, the network can integrate local and global spatial-spectral information more effectively. As shown in Fig. 6, this encoder produces spatio-spectral features at four scales. The highest semantic level feature (with lowest spatial resolution) being down-sampled 32 times on the original image is denoted as $F_{ST} \in R^{B\times C\times H\times W}$, where B, C, H, W denotes the batch-size, number of channels, height and width.

2.2. Spatio-temporal encoder

Existing spatio-temporal segmentation networks can be broadly categorized into two primary types: (1) those focusing on local spatio-temporal information [50,71,72] and (2) those emphasizing global spatio-temporal information [52,73,74]. However, these two approaches struggle to simultaneously consider global and local spatio-temporal information, leading to a certain loss of temporal or spatial information. Therefore, this paper proposes a novel spatio-temporal encoder module (Fig. 3), which consists of spectral convolution, 3D patch partition, linear embedding, and spatio-temporal transformer block.

Meanwhile, it is also noted that existing spatio-temporal fusion networks presents relatively less consideration to spectral information extraction. To address this, we employ a 3D spectral convolution to

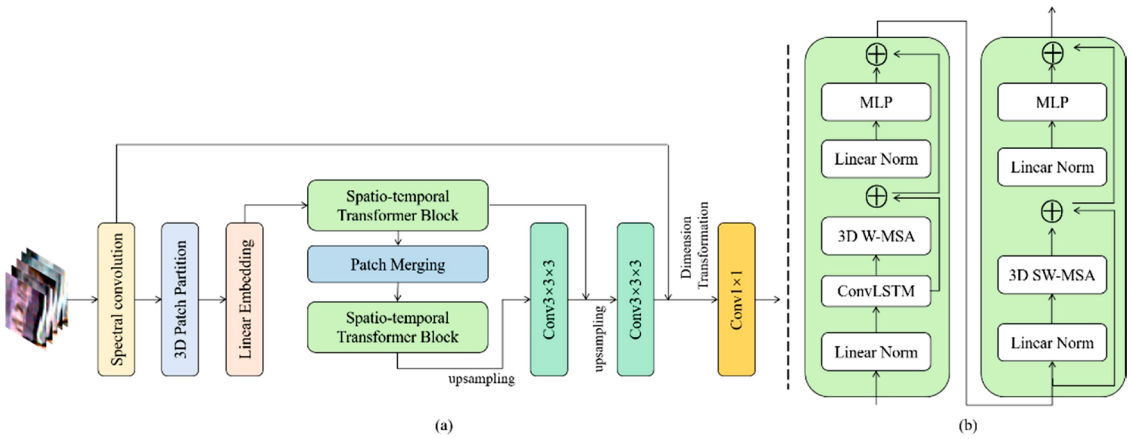

Fig. 3. Spatio-temporal encoder module. (a) Overall structure. (b) Spatio-temporal transformer block.

extract spectral features from multi-spectral Sentinel-2 images (see the yellow box in Fig. 3(a)). The initial feature shape of the input image is (B,T_1,D_1,H_1,W_1), where B, T_1, D_1 H_1, W_1 denote batch-size, temporal dimension, spectral dimension, height and width, respectively. After spectral convolution, the feature shape resizes into (B,T_1,D_2,H_1,W_1), where D_2 represents the spectral dimension after spectral convolution. Spectral convolution refers to a convolution layer performed for the spectral dimension in multispectral Sentinel-2 images. We utilize skip connections to fuse shallow with deep details, enhancing the utilization of spectral information. Meanwhile, to alleviate the issue of spatio-temporal information loss, a spatio-temporal block is designed. Firstly, the 3D patch partition divides the spatio-temporal feature into distinct blocks, where each block encapsulates temporal, spatial, and spectral information (see the light blue box in Fig. 3(a)). The shape of the partitioned feature is $(B,D_2,\frac{T_1}{2},\frac{H_1}{2},\frac{W_1}{2})$. Subsequently, the linear embedding conducts a linear transformation on the output features of the 3D patch partition to extract more sophisticated feature representations (see the light pink box in Fig. 3(a)). The feature shape resizes into $(B,D_3,\frac{T_1}{2},\frac{H_1}{2},\frac{W_1}{2})$, where D_3 represents the spectral dimension after linear embedding.

After the linear embedding, two spatio-temporal transformer blocks (see the light turquoise box in Fig. 3(a)) are used to mine global and local spatio-temporal information. The size of the output features of the first spatio-temporal transformer block become $(B,D_4,\frac{T_1}{2},\frac{H_1}{2},\frac{W_1}{2})$, where D_4 denotes the number of channels. The Patch Merging layer is in the middle of the two spatio-temporal transformer blocks (see the light sky blue box in Fig. 3(a)), which is used to down-scale the spatial and temporal dimensions, and the feature shape is resized into $(B,D_4,\frac{T_1}{4},\frac{H_1}{4},\frac{W_1}{4})$. After that, through the convolutional layer (the first light green box in Fig. 3(a)), the output features of the second spatio-temporal transformer block are further adjusted, as their size is converted to $(B,2D_4,\frac{T_1}{2},\frac{H_1}{2},\frac{W_1}{2})$. The second spatio-temporal transformer block is connected with the output features of the first transformer block in the form of short cut. The output features are passed through the second convolutional layer (the second light green box in Fig. 3(a)) in a similar manner, and the feature shape is adjusted to (B,D_2,T_1,H_1,W_1). After dimension transformation, the feature shape is adjusted to $(B,D_2 \times T_1,H_1,W_1)$. Finally, the features are fed into a 2D convolutional layer with a convolutional kernel of 1×1 (the light orange box in Fig. 3(a)) to generate the segmentation result, and the feature shape is adjusted to (B, C,H_1,W_1), where C denotes the dimension of the output features.

To mine multi-scale local and global spatio-temporal information, the spatio-temporal transformer block (Fig. 3(b)) is designed, including Layer Normalization, ConvLSTM, 3D W-MSA, 3D SW-MSA, and MLP. Layer Normalization is used to normalize the inputs in each feature dimension, which helps accelerate model convergence and improve the generalization ability of the model. 3D W-MSA and 3D SW-MSA effectively reduce the computational complexity by introducing the concepts of window and shifted window, respectively, thereby more thoroughly accounting for the spatial and temporal correlation in the data. ConvLSTM aims to mine the local spatio-temporal features, while the MLP primarily facilitates nonlinear mapping and augmentation of the input features to improve model representation.

ConvLSTM utilizes convolutional operations and internal memory cells to extract local spatio-temporal information from multi-temporal images. The formula for ConvLSTM is as follows:

$$i_t = \sigma(Z_{ii} * x_t + Z_{hi} * h_{t-1}) \tag{5}$$

$$f_t = \sigma(Z_{if} * x_t + Z_{hf} * h_{t-1}) \tag{6}$$

$$C_t = f_t \circ C_{t-1} + i_t * tanh(Z_{ig} * x_t + Z_{hg} * h_{t-1}) \tag{7}$$

$$o_t = \sigma(Z_{io} * x_t + Z_{ho} * h_{t-1}) \tag{8}$$

$$h_t = o_t \circ tanh(C_t) \tag{9}$$

where i_t, f_t, o_t represent outputs of input gate, forget gate, and output gate at time t, $t \in \left[1, \frac{T_1}{2}\right]$, respectively. Z_{ii} and Z_{hi} are weights for input gate. $Z_{if}$and Z_{hf} are weights for forget gate. Z_{ig} and Z_{hg} are weights for memory cell. Z_{io} and Z_{ho} are weights for output gate. σ denotes sigmoid function and $\circ$ denote the Hadamard product. $x_t \in R^{B \times D_3 \times \frac{H_1}{2} \times \frac{W_1}{2}}$, $h_t \in R^{B \times D_4 \times \frac{H_1}{2} \times \frac{W_1}{2}}$, $C_t \in R^{B \times D_4 \times \frac{H_1}{2} \times \frac{W_1}{2}}$ represents input feature, hidden state, memory cell at time t, respectively, where D_4 is the number of channels in the hidden state. When the channel numbers of x_t and h_{t-1} do not match (*i.e.*, $D_4 \neq D_4$), employ a 1×1 convolutional layer to adjust the channel numbers of x_t to match that of h_{t-1}. the convolution operation * processes x_t to capture spatio-spectral information. C_t responsible for storing and conveying spatio-temporal information. Through the coordinated interplay of the input gate, forget gate, memory cell, and output gate, the hidden state h_t can adeptly capture spatio-temporal information. The output $(h_1, h_2, h_3 \cdots h_{\frac{T_1}{2}}) \in R^{B \times \frac{T_1}{2} \times D_4 \times \frac{H_1}{2} \times \frac{W_1}{2}}$ of ConvLSTM is composed of hidden states.

3D W-MSA that calculate self-attention [75,76] (see formula 10) within a 3D window (The three dimensions of temporal, height, and width) is employed to address the inability of interaction between various windows, thereby extracting global spatio-temporal information.

$$\text{Attention}(Q,K,V) = \text{SoftMax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{10}$$

where Q, K, and V represent the query, key, and value matrices, respectively. d denotes the dimension of query/key. The output features $(h_1, h_2, h_3 \cdots h_{\frac{T_1}{2}})$ after ConvLSTM undergoes three independent linear transformations to obtain Q, K, and V, respectively. Q and K are combined to calculate the distribution of attention weights (see formula 10), where Q reflects the significance of the current time and K acts as a reference for Q. This weight distribution is then utilized for the weighted combination of V, extracting global information by computing weights across the entire scope.

2.3. Cross resolution fusion of spatial-spectral modal features and spatio-temporal modal features

There is significant noise and resolution disparity between remote sensing data from different sensors, making direct fusion methods (*e.g.*, concatenation or addition operations) challenging [77]. To amplify valuable features and suppress irrelevant ones, a cross-resolution fusion module (see Fig. 4) is formulated to alleviate spatial resolution differences among multi-modal data.

Considering the substantial resolution disparity between H^2SR image and dense time series Sentinel-2 images, we achieve spatio-temporal-spectral cross-modal integration in the form of channel attention using the encoding features at the highest semantic levels of both modalities. First, we spatially resample the encoder features of the spatio-temporal modality to the same spatial dimensions as the semantic features of the spatial-spectral modality, denoted as $F_{ST} \in R^{B\times C\times H\times W}$. The specific calculation for spatio-temporal-spectral integrated fusion is shown in formula (11).

$$F_{STS} = F_{SS} \oplus \left(\alpha \times \left(\left(F_{SS} \otimes W_{MLP_{SS}}\right) \oplus \left(F_{ST} \otimes W_{MLP_{ST}}\right)\right)\right), \text{ where } \alpha = SR_{TT}/SR_{ST} \tag{11}$$

Here, $F_{STS} \in R^{B\times C\times H\times W}$ represents the fused features. The adjustment coefficient α balances the importance of the spatial-spectral features and the spatio-temporal features, determined by the ratio of spatial resolution pixels of spatial-spectral image (SR_{TT}) to that of spatio-temporal images (SR_{ST}), with smaller value indicating greater spatial resolution differences. $W_{MLP\_SS} \in R^{B\times C\times 1\times 1}$ and $W_{MLP\_ST} \in R^{B\times C\times 1\times 1}$ are the weights of the two modalities.

The cross-modal attention in this paper consists of Sigmoid function (σ), Max Pooling layer (*MP*), Average Pooling layer (*AP*), shared convolution [78] (φ), and MLP. Specifically, we employ *MP* and *AP* separately in the spatial dimension to capture channel attention on each modality. Considering that features from each modality may contribute differently to land-cover segmentation, shared convolution is used to simultaneously process the attention of each modality, and Concatenation operation (*Concat*) conducted for preliminary fusion of spatio-temporal-spectral features, as formulated in Eq. (12):

$$W_{SC\_STSF} = Concat(\varphi(AP(\sigma(F_{SS}))), \varphi(MP(\sigma(F_{SS}))), \varphi(AP(\sigma(F_{ST}))), \varphi(AP(\sigma(F_{ST})))) \tag{12}$$

Then, with the characteristic that capture complex relationships and feature interactions between channels [79,80], MLP is employed to effectively compressing and integrating information across different modalities. Subsequently, the weights through MLP are split into equal-sized W_{MLP_SS} and W_{MLP_ST}.

3. Study area and data

3.1. Study area

The study area is located in the southern part of Shenyang City, Liaoning Province, northeast China (122°33′ - 122°52′, 41°12′ - 41°25′). With a total area of 227.79 km^2, the research area falls within the temperate humid continental climatic zone (see Fig. 5). Characterized by windy springs, rainy summers, it has an average frost-free period of 171 days.

3.2. Data

We have released an open-source dataset combining hyperspectral and high spatial resolution data with multi-temporal data (named WHU-H^2SR-MT). To our knowledge, this is currently the largest dataset available for spatio-temporal-spectral interpretation. This dataset serves not only to validate the effectiveness of the algorithms presented in this paper but also to contribute to the research community in this field. Additionally, we use two separate open-source datasets (a H^2SR image, and a set of multi-temporal images) to verify the effectiveness of the proposed multi-scale spatial-spectral encoder (see Section 2.1 for details) and spatio-temporal encoder (see Section 2.2 for details), respectively.

The H^2SR images were acquired in September 2020, with a spectral range between 391 nm and 984 nm, comprising 249 bands. The image was preprocessed by relative radiometric calibration and atmospheric correction. The spectral resolution of each band is ≤5 nm.

Multi-temporal Sentinel-2 images for 2020 were acquired for the experimental area. First, Fmask [55], a widely used cloud removal technique, was applied to each Sentinel-2 image. This operation assigns a null value to pixels affected by clouds or cloud shadows, ensuring clear and reliable spectral data for further analysis. Then, monthly averaging has been performed for each pixel, thereby obtaining a multi-temporal multi-spectral Sentinel-2 images, resulting in a total of 31 Sentinel-2

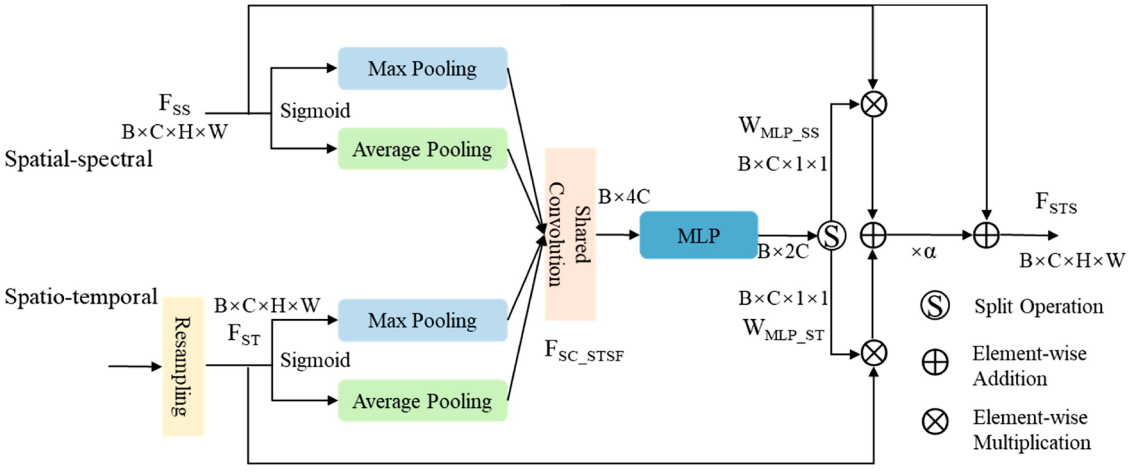

Fig. 4. Cross-resolution fusion module. B, C, H, W denotes the batch-size, number of channels, height and width for the input spatial-spectral features.

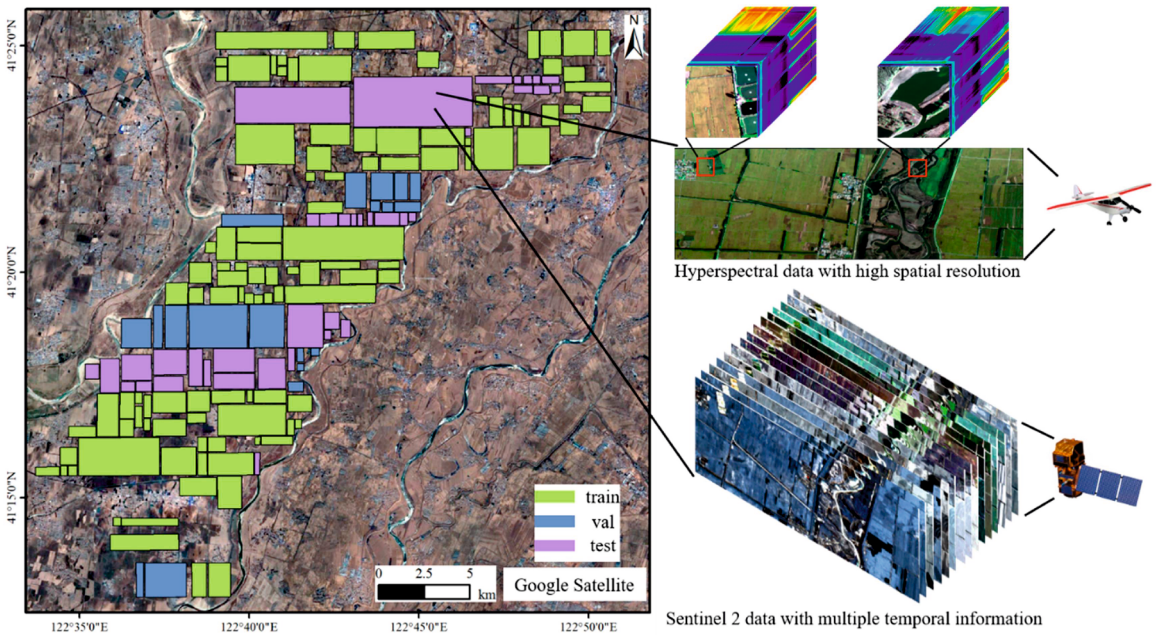

Fig. 5. The geographical location and distribution of the study area.

images (see Fig. 6) after cloud removal operations. Ten land-cover relevant bands (B2, B3, B4, B5, B6, B7, B8, B8a, B11, B12) [81] were selected, and resampled the bands with coarser spatial resolution (B5, B6, B7, B8a, B11, B12) to 10 m.

The spatio-temporal-spectral dataset consists of the H^{2}SR images and the multi-temporal Sentinel-2 images. The land cover reference label was derived from the land cover project completed by the Ministry of Natural Resources of China in 2020. All labels were manually collected and validated through field survey. First, we aligned the H^{2}SR images using the coordinate system and projection: WGS_1984_UTM_Zone_51 N. Later, the multi-temporal Sentinel-2 images were downloaded using the vector range of the H^{2}SR images. Firstly, each image is spatially aligned with the manually collected land-cover labels and uniformly cropped into patches of size 300 m × 300 m, with an 80 m spatial overlap. A total of 2531 pairs of samples were generated, with each pair consisting of one H^{2}SR patch, a set of 31 Sentinel-2 patches, and the corresponding land-cover label patch. This dataset comprises eight land-cover categories (Fig. 7). The dataset has eight land-cover categories: paddy field, dry farmland, forest land, grassland, building, highway, greenhouse, and water body. The dataset was divided into training, validation, and test sets with a ratio of 6:1:3. We named the entire multi-modal dataset WHU-H^{2}SR-MT, with the H^{2}SR subset called WHU-H^{2}SR, and the multi-temporal subset called WHU-MT.

4. Results and discussions

4.1. Experimental settings

To fully validate the proposed network, in addition to the dataset in the experiment areas, we utilized two open-source datasets: one with spatial-spectral information and the other with spatio-temporal information.

(1) AeroRIT dataset [21] consists of a H^{2}SR image captured by the Headwall Micro E sensor with a spatial resolution of 0.4 m and 372 spectral bands. Currently, it is open-source and includes 51 spectral bands, with a size of 1973 × 3975 pixels and wavelengths ranging from 397 to 1003 nm. This dataset comprises five land-cover categories (Fig. 8). This dataset is divided into training, validation, and test sets with a ratio of 6:2:2.
(2) Sen4AgriNet dataset [82] consists of 180,000 pairs of Sentinel-2 patches. Each pair comprises image patches from 12 time-series, with each patch including 13 spectral bands. The patch size is 61 × 61 pixels, and the spatial resolution is 10 m. From this dataset, we selected seven land cover categories (Fig. 9). The dataset is divided into training, validation, and test sets with a 6:2:2 ratio.

Model training: All experiments were conducted using the PyTorch framework and on a server equipped with two NVIDIA GeForce RTX 3090 GPUs. During the experimental training, the initial learning rate was set to 0.001, the Adam optimizer [83] was used, the cross-entropy loss function was applied, and the batch size was set to 8. The model was trained for 45 epochs on WHU-H^{2}SR-MT dataset and 20 epochs on AeroRIT and Sen4AgriNet datasets.

Performance assessment: To quantitatively evaluate the segmentation accuracy, we selected the following metrics: intersection over union (IoU), F1-Score (F1), overall accuracy (OA), and mean IoU (MIoU) [84–86]. These metrics have values ranging between 0 and 100 %, with

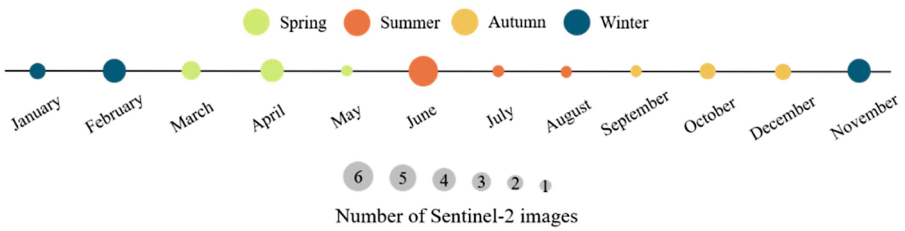

Fig. 6. Number of multi-temporal images available in each month of 2020.

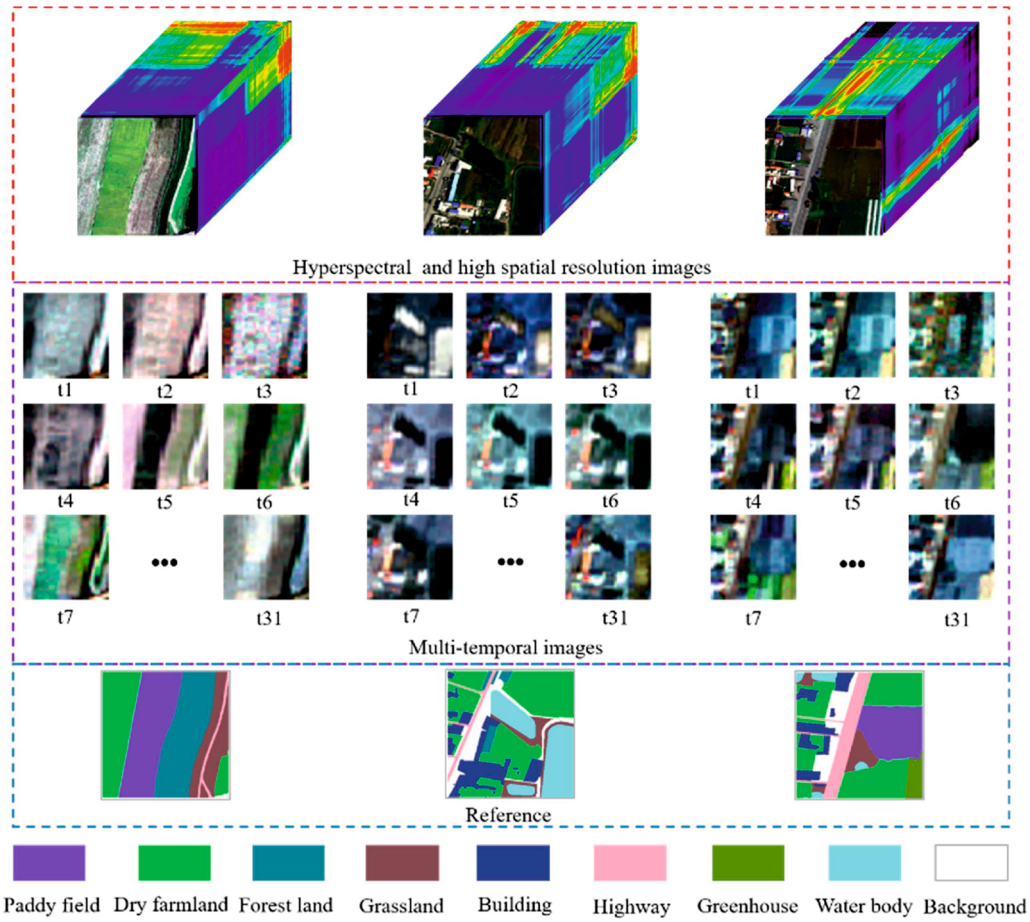

Fig. 7. WHU-H^{2}SR-MT. The first line represents WHU-H^{2}SR, the second line represents WHU-MT, the third line represents the associated semantic annotations.

higher values indicating better performance. In addition, we added the number of parameters and training time per epoch.

$$IoU = \frac{TP}{FN + FP + TP} \tag{13}$$

$$Recall = \frac{TP}{FN + TP} \tag{14}$$

$$Precision = \frac{TP}{FP + TP} \tag{15}$$

$$F1 - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{16}$$

$$OA = \frac{TP + TN}{FN + TN + FP + TP} \tag{17}$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{TP}{FN + FP + TP} \tag{18}$$

Where k is the number of land cover category, and TP, FP, TN, and FN denote the number of true-positive, false-positive, true-negative, and false-negative pixels in each category.

4.2. Performance of multi-scale spatial-spectral encoder

To evaluate the effectiveness of the multi-scale spatial-spectral encoder, we combined it with a multi-scale decoder for segmentation. Seven SOTA spatial-spectral segmentation networks were selected for comparison: RSSAN [34], DFFN [39], DPRN [40], HRNet [42], FreeNet [41], FTUNetFormer [87], H2Former [88], applied to the segmentation of H^{2}SR images. A brief description of the characteristics of the compared methods is provided below:

RSSAN: It is a popular spectral-spatial feature learning for hyperspectral image segmentation, and it imported spectral and spatial attention modules into a standard CNN with five 2D convolutional block.

DFFN: On the basis of standard CNN, it introduces residual learning to construct a deeper network, which is composed of three residual blocks.

DPRN: On the basis of DFFN, it employs a pyramidal bottleneck residual structure to comprehensively extract spatial-spectral features. It is noted that, although the spatial detailed information obtained by the pyramidal bottleneck residual structure is better than the above two networks, it is still at risk of over-smoothing in dealing with pixel-wise segmentation.

HRNet: It emphasizes multi-scale feature fusion through multi-resolution network architecture to ensure the high spatial resolution interpretation. However, its frequent cross-scale connections often lead to multiple compressions and reconstructions of features, and it ignores hyperspectral characteristics. Thus, our proposed work reduces the number of cross-scale connections and utilizes depth-separable convolution to further capture spatial and spectral features in an efficient manner.

FreeNet: Inspired by auto-encoder structure of Unet [38], it consists of an encoder equipped with on spectral attention and a decoder module for pixel-wise segmentation. Whereas, we design a multi-branch structure that begins with depth separable convolution to simultaneously extract multiscale spectral and spatial features, and incorporate spectral attention to enhance the capability to explore spectral information.

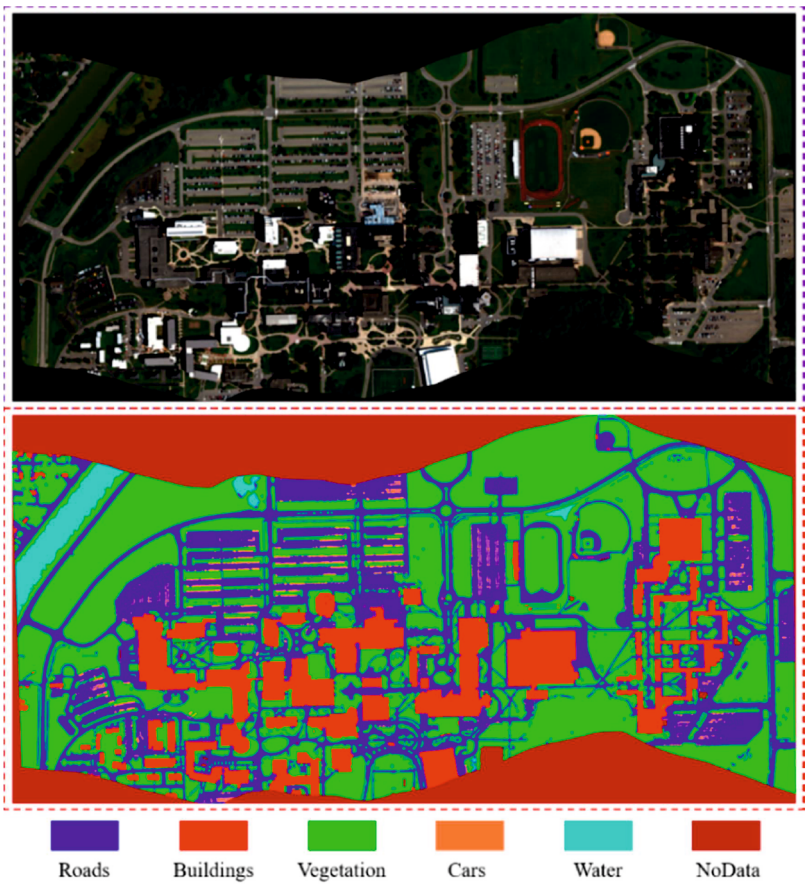

Fig. 8. AeroRIT dataset. The purple dashed box contains the true color (R:7, G:15, B:25) map and the red dashed box contains the corresponding semantic labels.

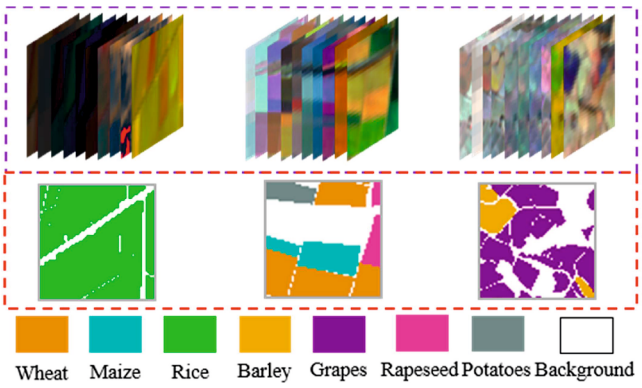

Fig. 9. Sen4AgriNet dataset. The purple dashed box contains the multi-temporal Sentinel-2 patches, and the red dashed box contains the corresponding semantic annotations.

FTUNetFormer: This model utilizes Swin Transformer structure as encoder and employs a UNet-like decoder, enabling the modeling of both global and local spatial-spectral information.

H2Former: This model integrates convolution, multiscale channel attention, and Transformer components through hybrid strategies to enhance its ability to capture multiscale long-distance dependencies and local spatial information.

(1) Experiment 1: WHU-H^{2}SR

Table 2 presents the quantitative segmentation results of the spectral-spatial fusion networks. It is noted that the proposed multi-scale spectral-spatial encoder-based work achieves the highest OA and mIoU. The increments in OA and mIoU relative to the compared algorithms range from 0.53 % ~ 9.40 % and 0.65 % ~ 16.76 %, respectively. The results indicate that the proposed multi-scale spectral-spatial encoder module has a significant advantage over other spectral-spatial fusion methods. Although the parameters of the proposed encoder are the fourth-largest, its training cost is the second most efficient.

As seen in the cases of Fig. 10, RSSAN, which ignores multi-scale spatial information, leads to noticeable noise for all categories. From Cases 4 and 5 of Fig. 10, HRNet demonstrates good accuracy in the forest land and grassland but exhibits misclassifications for dry farmland and greenhouse. It is obvious from Case 1, 3 and 4 that due to multiple down-sampling operations that cause spatial information loss, DFFN and DPRN tend to misclassify the greenhouse land as buildings, while all cases of Fig. 10 suggest that FreeNet exhibits misclassifications in the dry farmland, highway and water bodies. From Cases 1 and 3 in Fig. 10, both FTUNetFormer and H2Former exhibit poor performance in local feature extraction, leading to misclassifications of building and greenhouse boundaries.

(2) Experiment 2: AeroRIT

To verify the generality of the multi-scale spatial-spectral encoder, we conducted tests on the AeroRIT open dataset. As depicted in Table 3, our proposed method achieved the highest OA and mIoU compared to other spatial-spectral fusion networks. Our method achieved performance improvements ranging from 0.40 % ~ 3.24 % in OA and 3.14 % ~ 10.10 % in mIoU.

In Case 3 of Fig. 11, DFFN, DPRN, HRNet and FTUNetFormer have apparent omissions in cars. While in Case 1 and Case 5, RSSAN, DFFN, DPRN, FreeNet, FTUNetFormer and H2Former exhibit clear misclassifications in roads. In Case 4 of Fig. 11,

Table 2
Comparison between the spatial-spectral fusion networks on WHU-H^{2}SR.

Method		RSSAN	DFFN	DPRN	HRNet	FreeNet	FTUNetFormer	H2Former	Ours
Paddy field	IoU	81.94	90.21	89.15	89.58	90.53	87.06	89.80	**91.64**
	F1	90.01	94.85	94.26	94.50	95.03	93.08	94.63	95.63
Dry farmland	IoU	66.31	82.04	79.26	79.12	80.48	77.05	78.53	**81.01**
	F1	79.67	90.14	88.43	88.35	89.19	87.04	87.97	89.50
Forest land	IoU	26.45	**54.82**	53.36	51.69	50.22	48.47	50.76	53.50
	F1	40.42	70.81	69.57	68.15	66.86	65.29	67.34	69.70
Grassland	IoU	4.26	25.99	25.35	**27.32**	25.94	25.91	23.60	24.71
	F1	8.17	41.22	40.45	42.91	41.20	41.15	38.18	39.62
Building	IoU	54.25	69.10	61.08	**69.17**	68.87	66.31	66.83	68.44
	F1	70.30	81.73	75.69	81.78	81.57	79.74	80.12	81.27
Highway	IoU	26.74	38.18	33.24	30.62	38.26	**39.11**	16.06	37.30
	F1	41.69	55.26	49.89	46.88	55.31	56.23	27.68	54.33
Greenhouse	IoU	30.13	39.21	44.40	57.00	**60.16**	49.27	54.64	59.35
	F1	46.26	55.85	61.41	72.61	75.10	66.02	70.67	75.40
Water body	IoU	45.90	53.54	52.10	52.82	49.73	**54.11**	44.59	53.38
	F1	62.82	69.74	68.49	69.13	66.41	70.22	61.68	70.05
#Parameters(M)		0.13	0.53	8.66	29.69	2.69	33.41	33.93	22.81
Training time per epoch(min)		10.25	9.16	9.17	8.88	9.25	9.51	9.82	8.92
OA		77.50	86.25	85.05	86.09	86.37	84.33	85.00	**86.90**
mIoU		41.91	56.63	54.75	57.17	58.02	55.91	53.11	**58.67**

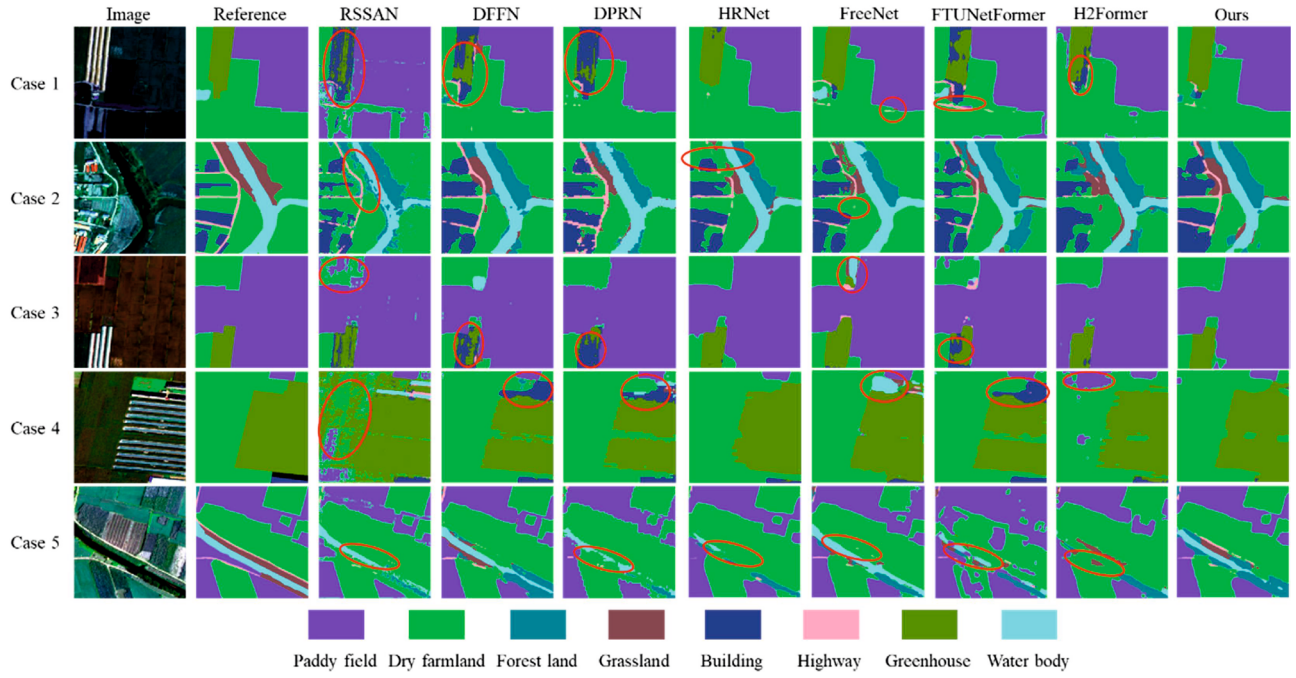

Fig. 10. Visualization of segmentation results of different the spectral-spatial fusion networks on WHU-H^{2}SR. Misclassifications are highlighted with red circles.

Table 3
Comparison between the spatial-spectral fusion networks on AeroRIT.

Method		RSSAN	DFFN	DPRN	HRNet	FreeNet	FTUNetFormer	H2Former	Ours
Water	IoU	75.69	63.60	73.97	71.08	77.68	**77.82**	55.62	68.40
	F1	86.16	79.60	85.03	75.10	87.44	87.53	71.48	81.17
Cars	IoU	19.72	27.88	24.34	32.76	22.95	29.27	20.85	**48.37**
	F1	32.89	43.19	38.57	47.43	37.34	45.28	34.51	65.18
Vegetation	IoU	97.22	95.41	95.42	97.04	**97.61**	96.84	97.39	97.55
	F1	98.59	97.65	97.66	98.66	98.79	98.40	98.68	98.76
Roads	IoU	75.64	74.42	78.05	84.57	80.59	79.39	81.76	**86.16**
	F1	96.11	85.34	87.67	92.57	89.25	88.51	89.96	92.57
Buildings	IoU	82.65	73.45	83.05	86.70	86.62	80.16	**89.03**	87.39
	F1	90.47	84.70	90.74	93.57	92.83	88.99	94.20	93.27
#Parameters(M)		0.13	0.53	8.66	29.69	2.69	33.41	33.93	22.81
Training time per epoch(min)		0.91	5.34	1.02	4.42	1.11	3.01	3.35	1.16
OA		93.81	93.20	94.19	96.04	95.05	94.69	95.31	**96.44**
mIoU		70.19	67.47	70.96	74.43	73.09	72.70	68.93	**77.57**

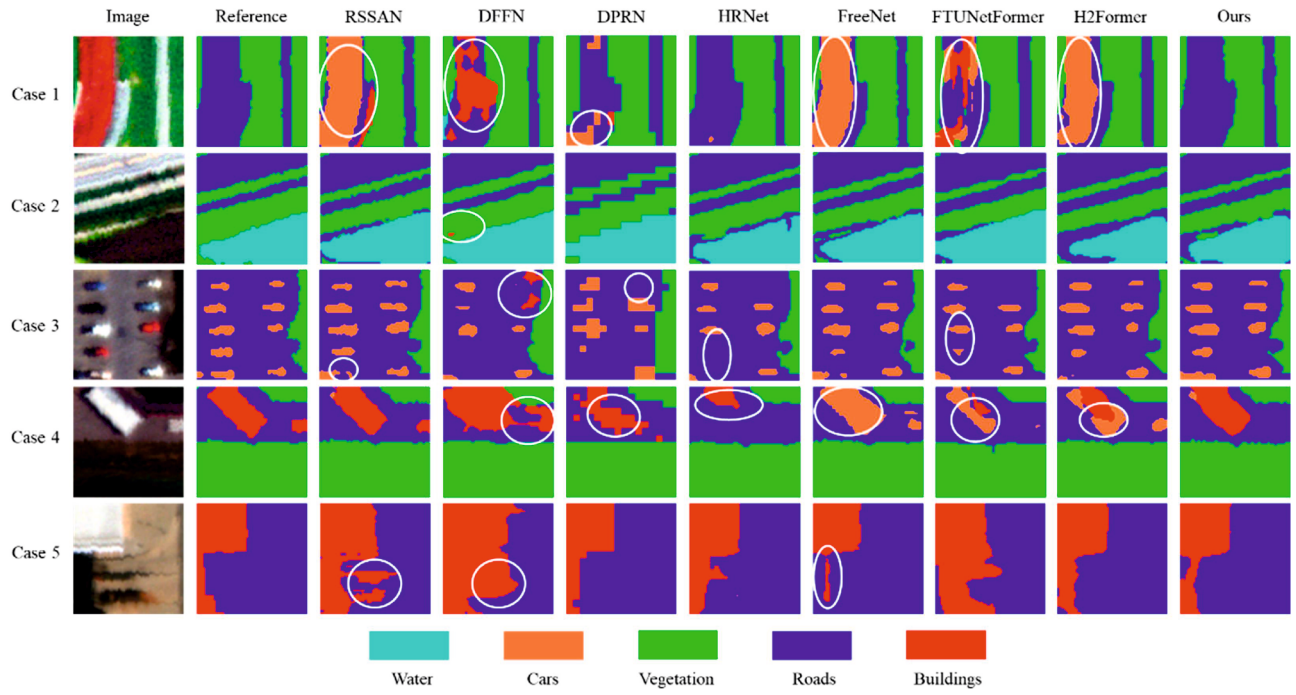

Fig. 11. Visualization of segmentation results of different spectral-spatial fusion networks on AeroRIT. Misclassifications are highlighted with white circle.

RSSAN and DFFN also present commissions in buildings, while HRNet, FreeNet, FTUNetFormer and H2Former omit buildings. Compared to WHU-H^{2}SR, AeroRIT has a smaller size, therefore the required time for this experiment is shorter.

(3) Ablation experiment of multi-scale spatial-spectral encoder

Our proposed multi-scale spatial-spectral encoder performs two cross-scale connections at the beginning and end stages of the network, which avoids frequent cross-scale connections that lead to multiple compression and reconstruction problems of the features. Meanwhile, we integrate a spectral gated module into the encoder to enhance the diagnostic capability of hyperspectral data. Overall, experimental results demonstrate that our multi-scale spatial-spectral encoder achieves higher performance with a lower computational cost (mIoU: 58.67 %, Parameters: 22.81 M) compared to the mainstream multi-scale network HRNet (mIoU: 57.17 %, Parameters: 29.69 M).

Further, we conducted thorough ablation experiments to assess the impact of these key components. We employed a multi-branch structure composed of the convolutional blocks as the baseline. When using the baseline for segmentation, the mIoU is the lowest (see Table 4). The incorporation of DSCs and spectral gated module separately resulted in a noticeable improvement in mIoU. When both are used in combination, the mIoU reaches its highest value. It is indicated that introducing DSCs and spectral gated module into multiscale spatial-spectral encoder is essential. DSCs effectively captures spectral and spatial information, while spectral gated module explores global and local significant features in spectral bands.

The key hyperparameter in STSNet is the number of channels (C). When the number of channels is set to (32,32,64,128), the mIoU of the spatial-spectral fusion network is 55.91 %, and the number of parameters is 0.72 M. As the value of C increases, both the number of parameters and the overall accuracy increase (see Table 5). Considering both the accuracy and efficiency of the model, C is set as (249, 300, 512, 512) for the following experiments in this study.

Table 4
Multi-scale spatial-spectral encoder module ablation experiments on different datasets.

Dataset	Spectral gated module	DSCs	mIoU
WHU-H^{2}SR			53.17
	✓		55.17
		✓	57.82
	✓	✓	**58.67**
AeroRIT			71.08
	✓		71.59
		✓	74.33
	✓	✓	**77.57**

4.3. Performance of spatio-temporal encoder

To assess the effectiveness of the proposed spatio-temporal encoder, we connected it with a 1 × 1 convolution layer, which adjusts the number of channels to the target number of semantic categories. We selected five SOTA spatio-temporal fusion networks as comparative algorithms: 3DUnet [73], ConvLSTM [51,89], FPN-ConvLSTM [50], ConvGRU [90], and TSViT [52], and, for the segmentation of the multi-temporal images.

3DUnet: The model follows the Unet structure, with the encoding branch using 3D convolutions to process both spatio-temporal features.

ConvLSTM: This model is a LSTM based recursive neural network, and it replaces the linear layers with spatial convolutions.

FPN-ConvLSTM: This model utilizes a pyramid structure to learn spatial information and then employs ConvLSTM to extract time-series information.

ConvGRU: On the basis of ConvLSTM, this model introduces a variant of Gated Recurrent Unit (GRU) to replace the LSTM block.

Table 5
Segmentation accuracy and the number of channels C.

Hyperparameter C	mIoU	#Parameters(M)
(32,32,64,128)	55.91	0.72
(64,64,128,256)	57.22	2.84
(128,128,256,512)	58.06	11.04
(249,300,512,512)	58.67	22.81
(249,512,512,1024)	58.75	43.71

TSViT: This model is based on the visual transformer, enhancing segmentation capabilities by utilizing temporal positional encoding and class tokens.

(1) Experiment 1: WHU-MT

Table 6 presents the quantitative segmentation results of various spatio-temporal fusion networks. Our proposed method, which utilizes the spatio-temporal encoder, achieves optimal segmentation performance, with OA and mIoU reaching 85.93 % and 53.88 %, respectively. Compared to other comparative algorithms, our method obtains an increase of 1.46 %–11.83 % and 2.73 %–13.22 % in terms of OA and mIoU, respectively. These experimental results highlight the remarkable superiority of the proposed spatio-temporal encoder module.

From Case1, 3 and 4 of Fig. 12, it is seen that all compared networks misclassify the greenhouse to other classes. In Case2 of Fig. 12, 3DUnet and TSViT misclassify dry farmland as buildings. And in Case5, water bodies are misclassified as forest land by ConvLSTM and TSViT. The improved segmentation results achieved by our spatio-temporal encoder. The main reasons for this superiority can be primarily attributed to the following factors: the spectral convolution, focusing on the spectral information of multi-temporal data, and the utilization of ConvLSTM and 3D self-attention within the spatio-temporal transformer block extracts both local and global information in the multi-temporal Setineal-2 images. This approach can to some extent reduce the loss of spectral, temporal, and spatial information, thereby improving segmentation accuracy.

(2) Experiment 2: Sen4AgriNet

To furtherly validate the performance of the spatio-temporal encoder, we selected the Sen4AgriNet open-source dataset for testing. Table 7 presents the quantitative segmentation results of the spatio-temporal fusion networks on Sen4AgriNet. As seen from this table, our proposed work demonstrates superior performance. Compared to the comparative algorithms, our work achieved an OA gains ranging from 0.53 % – 2.73 % and 0.39 % – 5.07 % in mIoU. The core component of TsViT is the Vision Transformer (ViT) [91], a standard transformer module that learns global features and prefers large-scale data training. In contrast, our proposed multi-scale hierarchical spatio-temporal transformer can simultaneously learn local-global spatio-temporal features with reliable performance even if the training sample is limited. Sen4AgriNet is a large-scale open-source dataset with over 180,000 Sentinel-2 images, which is much larger in scale compared to WHU-MT. Thus, the overall accuracy of TSViT was more evident on Sen4AgriNet compared to WHU-MT.

In Case 2 and 4 of Fig. 13, 3DUnet presents good accuracy for wheat, barley and potatoes, but poor performance for other categories (especially maize). In Case1, FPN-ConvLSTM, ConvGRU, and TSViT also omit maize. ConvLSTM and FPN-ConvLSTM misclassified maize as potatoes in Case 2, 3DUnet and ConvGRU presents errors for barley and maize in Case 3, while ConvLSTM and TSViT omit rice in Case 5. The primary reason for these misclassifications lies in the insufficient integration of temporal and spatial information by 3DUnet, ConvLSTM, ConvGRU, and TSViT.

(3) Spatio-temporal-spectral characteristics in multi-temporal multispectral Sentinel-2 images

Considering the spatio-temporal-spectral characteristics inherent in multi-temporal multispectral Sentinel-2 images, we propose a spatio-temporal encoder that combines spectral convolution and spatio-temporal transformer block. The architecture of the spatio-temporal encoder adopts 3DSwinT [76], which is a hierarchical multi-scale structure. In each layer of the structure, we originally propose the spatio-temporal transformer block, which combines ConvLSTM with 3D self-attention to construct a transformer structure to exploit both global and local spatio-temporal information. Table 8 compares the proposed transformer with the SOTA 3DSwinT (*i.e.*, baseline listed in the first line of the table). It compares and analyzes the performance of each component, demonstrating the superiority of our proposed approach.

When use only the baseline for segmentation, mIoU shows the lowest value. Subsequently, with the introduction of the spectral convolution and the spatio-temporal transformer block, the mIoU significantly improves. When combined, mIoU reaches its highest level. On the one hand, as Sentinel-2 images contain rich spectral information, spectral convolution can deeply extract the spectral features. On the other hand, compared to existing spatio-temporal fusion networks, the proposed spatio-temporal transformer block explores spatio-temporal information from both local and global perspectives, ensuring the improvement of segmentation accuracy.

4.4. Performance of cross-resolution fusion module

To assess the effectiveness of our proposed spatio-temporal-spectral cross-resolution fusion module, we selected five SOTA fusion strategies for comparison: Direct module [92], MSE module [37], AFFB module [93], FFM module [94], and DPFN module [95] for accuracy comparison. The test data used is WHU-H^{2}SR-MT.

Direct module: Directly concatenates different modal features.

MSE module: Concatenates different modal features initially, obtains squeezed features through a global average pooling layer, acquires new feature mappings through channel-wise multiplication, adds the feature mappings to the spatial branch, obtains fused feature maps, and finally achieves multimodal fusion.

AFFB module: Performs element-wise summation, multiplication, and maximization operations on different modal features and then

Table 6
Comparison between the spatio-temporal fusion networks on WHU-MT.

Methods	3DUnet		ConvLSTM		FPN—ConvLSTM		ConvGRU		TSViT		Ours	
Class	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Paddy field	70.27	82.54	90.18	94.83	69.45	81.97	90.71	95.13	68.73	81.47	**91.54**	95.58
Dry farmland	67.04	80.27	79.31	88.46	66.43	79.83	79.25	88.43	63.30	77.52	**80.65**	89.29
Forest land	44.99	62.06	47.59	64.49	40.32	57.47	45.79	62.81	36.80	53.80	**48.93**	65.71
Grassland	22.11	36.20	21.94	35.99	19.07	32.04	20.36	33.82	16.51	28.35	**25.30**	40.37
Building	57.41	72.94	58.48	73.80	58.82	74.08	59.45	74.57	52.95	69.24	**62.01**	76.55
Highway	**15.98**	27.55	8.66	15.94	14.47	25.28	7.62	14.13	11.23	20.19	15.88	27.40
Greenhouse	47.41	64.32	54.73	70.75	43.25	60.38	55.61	71.47	40.24	57.39	**55.66**	71.48
Water body	43.90	61.01	48.30	65.14	41.56	58.72	48.84	65.63	35.55	52.46	**51.12**	67.65
#Parameters(M)	1.55		0.02		0.76		0.02		8.76		6.33	
Training time per epoch(min)	0.91		0.93		22.61		0.72		1.18		1.35	
OA	76.74		84.47		76.10		84.39		74.10		**85.93**	
mIoU	46.13		51.15		44.17		50.95		40.66		**53.88**	

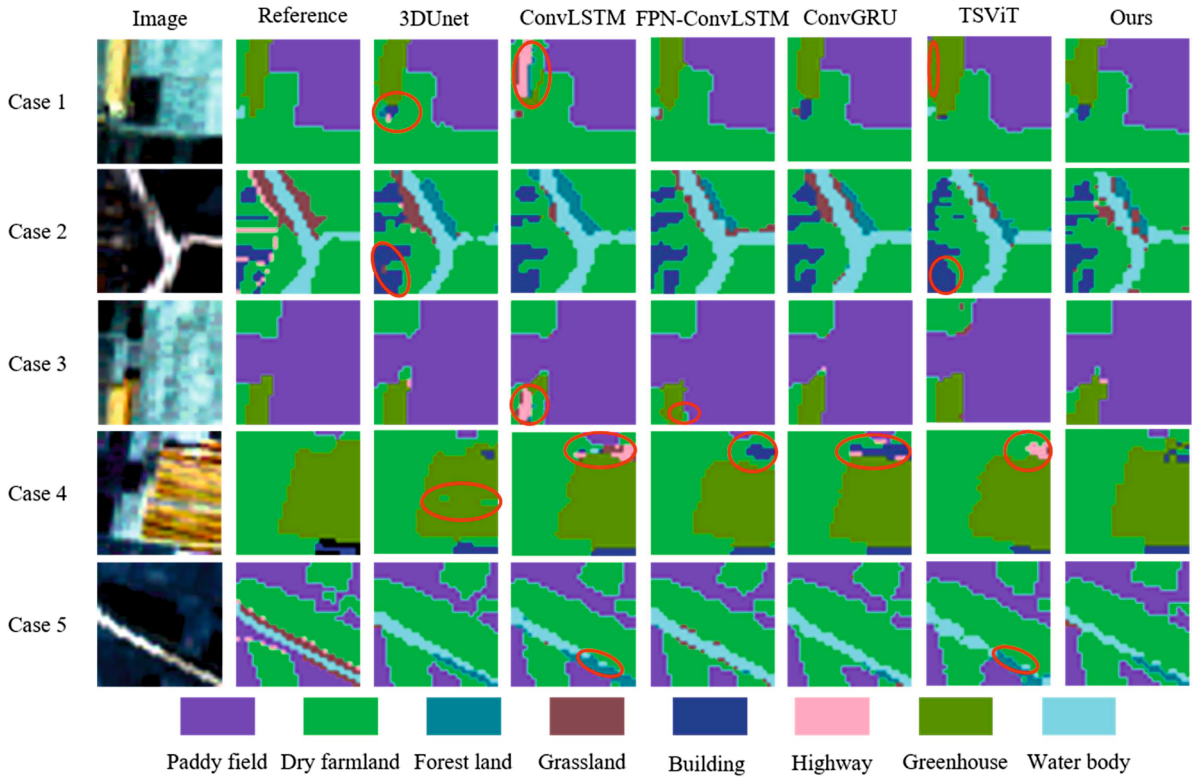

Fig. 12. Visualization of segmentation results of different spatio-temporal fusion networks on WHU-MT. Misclassifications are highlighted with red circle.

Table 7
Comparison between the spatio-temporal fusion networks on Sen4AgriNet.

Methods	3DUnet		ConvLSTM		FPN—ConvLSTM		ConvGRU		TSViT		Ours	
Class	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Wheat	94.94	97.41	94.10	96.96	93.17	96.47	94.53	97.19	92.20	95.94	**95.78**	97.84
Maize	93.92	96.87	93.36	96.56	91.75	95.77	91.39	95.50	88.06	93.65	**94.54**	97.19
Rice	99.95	99.97	99.86	99.92	99.82	99.91	99.89	99.95	**99.93**	99.96	99.90	99.95
Barley	89.21	94.30	87.71	93.45	85.52	92.20	86.74	92.90	83.62	91.08	**91.23**	95.41
Grapes	**99.49**	99.74	98.50	99.24	98.87	99.44	98.25	99.12	97.98	98.98	99.08	99.54
Rapeseed	95.84	97.87	94.45	97.15	90.67	95.10	93.31	96.54	88.19	93.72	**96.39**	98.16
Potatoes	**96.55**	98.25	92.56	96.13	93.09	96.42	93.18	96.47	87.16	93.14	95.74	97.82
#Parameters(M)	1.55		0.02		0.76		0.02		8.76		6.33	
Training time per epoch(min)	280		156		793		249		366		537	
OA	96.86		96.24		95.51		96.14		94.66		**97.39**	
mIoU	95.70		94.36		93.27		93.90		91.02		**96.09**	

concatenates the obtained feature maps through stacking.

FFM module: Utilizes a cross-attention to globally exchange information between two modalities, and upscales output feature to highest spatial resolution through channel mixing.

DPFN module: Utilizes separate convolutional layers to extract information between two modalities and then concatenates the generated feature maps through stacking.

Table 9 displays the segmentation results of different fusion strategies, demonstrating that our fusion strategy outperforms others. Particularly, mIoU is improved by 8.08 % ৰ্ 9.69 %. As seen from the cases in Fig. 14, the segmentation results of our proposed method are closer to the reference label. MSE, FFM and DPFN perform poorly on greenhouse in Case 1 and 3, MSE, AFFB and FFM noticeably misclassify dry farmland in Case 4, Direct misclassifies dry farmland as grassland in Case 2, and Direct, AFFB, FFM, and DPFN present inferior performance on highway. The main reasons are as follows: (1) WHU-H^{2}SR-MT has inconsistent spatial resolutions, with WHU-H^{2}SR having a spatial resolution of 1 m and WHU-MT having a spatial resolution of 10 m, resulting in a significant difference in data resolution. The use of direct fusion strategies (*e.g.*, concatenation) is difficult to form the complementary advantages of multimodal features. (2) Existing fusion strategies are not particularly designed for spatio-temporal-spectral feature fusion. In contrast, the newly proposed cross-resolution module aims to mitigate the information loss caused by large spatial disparities in multi-modal remote sensing images. It adjusts the importance of features from each modality branch (*e.g.*, spatial-spectral and spatio-temporal features) at different resolutions, adaptively regulating their contribution to land-cover interpretation. This module improves accuracy by over 8 % in mIoU compared to other fusion strategies.

4.5. Modalities performance analysis

Table 10 presents the segmentation results for different modalities. Specifically, H^{2}SR-VNIR represents the segmentation result of the image obtained by merging the spectral bands of WHU-H^{2}SR based on the band configuration of Sentinel-2 imagery, H^{2}SR denotes the segmentation result using WHU-H^{2}SR (*i.e.*, the high-spatial and hyperspectral image), H^{2}SR-10 denotes the result of the data by downscaling the spatial

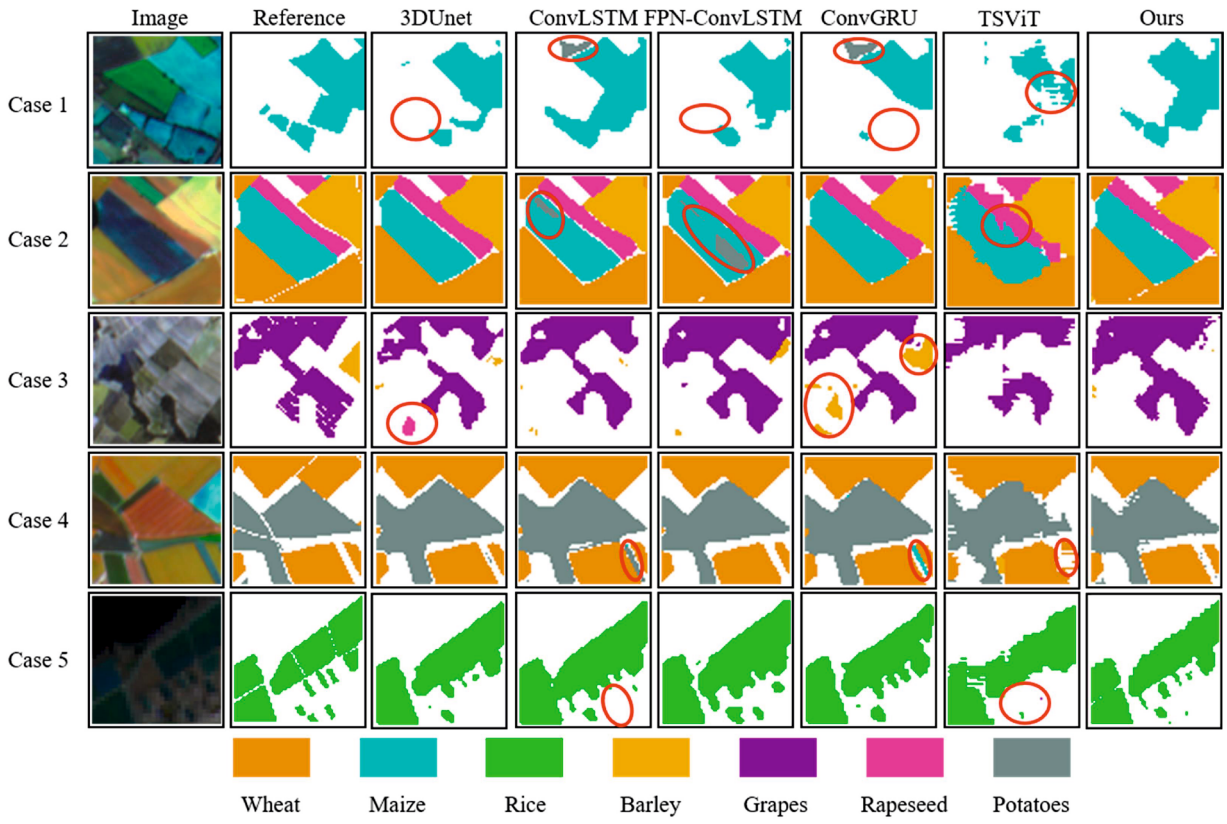

Fig. 13. Visualization of segmentation results of different spatio-temporal fusion networks on Sen4AgriNet. Misclassifications are highlighted with red circle.

Table 8
Spatio-temporal encoder module ablation experiments on two datasets.

Dataset	Spectral convolution	ConvLSTM	3D self-attention	mIoU
WHU-MT				44.78
	✓			52.97
	✓	✓		53.29
	✓		✓	53.42
	✓	✓	✓	**53.88**
Sen4AgriNet				95.20
	✓			95.53
	✓	✓		95.61
	✓		✓	95.67
	✓	✓	✓	**96.09**

resolution of WHU-H^{2}SR from 1 m to 10 m (*i.e.*, the spatial resolution of Sentinel-2 imagery), MT denotes the result using WHU-MT (*i.e.*, the multi-temporal Sentinel-2 images), H^{2}SR-VNIR-MT denotes the result using the fusion of WHU-VNIR and WHU-MT, and H^{2}SR-MT denotes the results using the fusion of WHU-H^{2}SR and WHU-MT. We replaced the original multi-temporal Sentinel-2 images (spatial resolution: 10 m) with multi-temporal Landsat 8 images (spatial resolution: 30 m) to explore the adaptability of the cross-resolution fusion module to a broader range of resolution changes. H^{2}SR-Landsat denotes the result using the fusion of WHU-H^{2}SR and WHU-Landsat (*i.e.*, the multi-temporal Landsat 8 images). Through comparing the above results, the following phenomena can be observed:

1) In the case of single modality, WHU-H^{2}SR (with both high spatial-spectral resolution) performs best, and the high spatial resolution H^{2}SR-VNIR and the multi-temporal WHU-MT have their own

Table 9
Comparison between the fusion modules on WHU-H^{2}SR-MT.

Methods	Direct		MSE		AFFB		FFM		DPFN		Ours	
Class	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1	IoU	F1
Paddy field	70.46	82.67	70.30	82.56	69.80	82.21	70.24	82.52	70.06	82.39	**92.78**	96.25
Dry farmland	67.98	80.94	67.74	80.77	67.48	80.58	68.45	81.27	67.92	80.90	**82.46**	90.39
Forest land	49.98	66.65	49.65	66.35	49.86	66.54	49.02	65.79	46.95	63.90	**54.97**	70.94
Grassland	25.28	40.36	25.89	41.14	23.09	37.51	26.20	41.52	23.13	37.57	**26.29**	41.58
Building	63.61	77.76	61.97	76.52	64.81	78.65	64.61	78.50	64.75	78.60	**68.64**	81.41
Highway	36.39	53.36	40.29	57.43	35.31	52.19	**40.31**	57.46	35.40	52.29	38.79	55.89
Greenhouse	47.91	64.78	48.59	65.40	41.09	58.25	46.18	63.19	46.13	63.14	**59.33**	74.41
Water body	50.17	66.82	44.06	61.17	48.89	65.66	48.20	65.05	47.24	64.17	**54.53**	70.58
#Parameters(M)	21.89		22.02		22.19		25.84		87.75		28.71	
#FLOPs(G)	56.10		56.12		56.22		56.51		81.92		56.13	
Training time per epoch(min)	10.15		10.21		10.58		11.13		12.61		10.30	
OA	78.13		77.87		77.89		78.18		77.95		**87.52**	
mIoU	51.47		51.06		50.04		51.65		50.20		**59.73**	

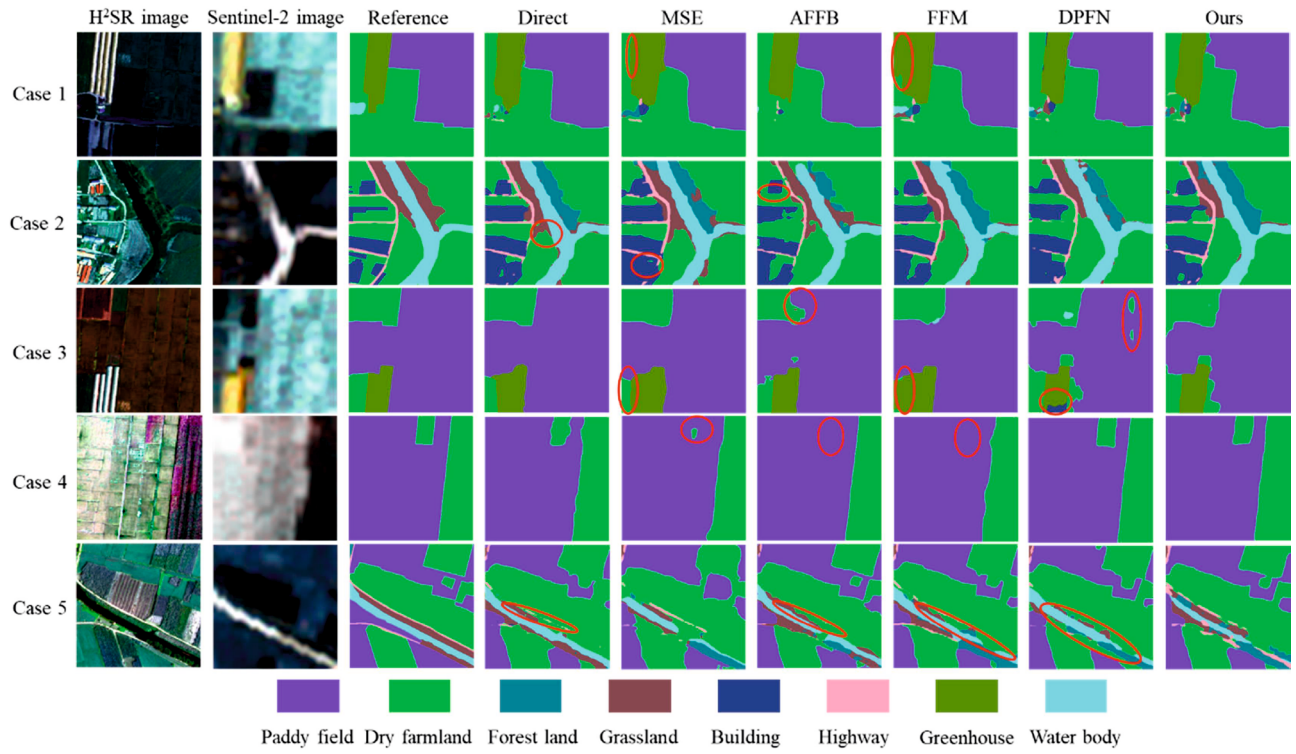

Fig. 14. Visualization of segmentation results of different spatio-temporal-spectral fusion networks on WHU-H^{2}SR-MT. Misclassifications are highlighted with red circle.

Table 10
Comparison of different modalities segmentation results.

		Paddy field	Dry farmland	Forest land	Grassland	Building	Highway	Greenhouse	Water body	mIoU	OA
Single-modal data	H^{2}SR-VNIR	88.45	78.57	54.29	22.97	63.45	29.25	52.79	47.79	54.70	84.76
	H^{2}SR	91.64	81.01	53.50	24.71	68.44	37.30	**59.35**	53.38	58.67	86.90
	H^{2}SR-10	86.65	73.67	44.97	19.29	58.69	12.35	42.99	44.23	47.83	82.15
	MT	91.54	80.65	48.93	25.30	62.01	15.88	55.66	51.12	53.88	85.93
Multi-modal data	H^{2}SR-VNIR-MT	90.88	79.48	53.20	23.70	65.79	32.78	49.65	46.72	55.30	85.94
	H^{2}SR-Landsat	92.14	81.82	**55.44**	24.45	68.24	38.32	57.47	52.80	58.84	87.28
	H^{2}SR-MT	**92.78**	**82.46**	54.97	**26.29**	**68.64**	**38.79**	59.33	**54.53**	**59.73**	**87.52**

strengths. For example, agricultural crops (including paddy field, dry farmland, greenhouse) prefer time-series information, while buildings and highways prefer spatial characteristics. Comparing multi-spectral with hyperspectral data, *i.e.*, H^{2}SR-VNIR *vs* H^{2}SR, hyperspectral characteristics provides stable accuracy gains for both artificial surfaces and agricultural crops. It is noteworthy that when downscaling the spatial resolution from 1 m to 10 m (*i.e.*, H^{2}SR *vs* H^{2}SR-10), the loss of spatial information leads to a sharp decrease in the segmentation accuracy of each category, demonstrating the importance of meter-level spatial characteristics for land cover interpretation. When multi-modal data are available, with the aid of the newly proposed cross-resolution fusion module, both the fusion of the H^{2}SR with multi-temporal Sentinel-2 (H^{2}SR-MT) and the fusion of H^{2}SR with multi-temporal Landsat-8 (H^{2}SR-Landsat) outperform H^{2}SR. Specifically, H^{2}SR-MT with higher spatial resolution is slightly superior to H^{2}SR-Landsat. This demonstrates that the proposed cross-resolution fusion module has significant advantages in fusing images with large spatial resolution differences.

2) For each land-cover category, the importance of features decreases in the order of meter-level spatial information, multi-temporal information to nanometer-level hyperspectral information. Moreover, it can be seen that the two artificial land types, highway and building, rely more on spatial characteristic, and highway cannot be effectively recognized with a spatial resolution of 10 m.

4.6. Applicability of different modal images

In terms of modality, we compare the segmentation result of each single-modal image (as shown in Table 10), including H^{2}SR, a spectral-synthetic image H^{2}SR-VNIR (merging the hyperspectral bands of H^{2}SR according to the spectral configuration of Sentinel-2 imagery), a spatial-synthetic data H^{2}SR-10 (downscaling H^{2}SR to the spatial resolution of Sentinel-2), and the multi-temporal Sentinel-2 images (called MT). Based on the experimental results, the following observations can be obtained:

When only a single modality data is available, H^{2}SR shows the most outstanding performance for all classes, which indicates that the joint spatial-spectral information has a significant advantage in land cover interpretation task. When only high-spatial multi-spectral modality (H^{2}SR-VNIR) is available, the interpretation results for artificial surfaces (including buildings, highways) are still good. When only the hyperspectral modality (H^{2}SR-10) is available, the accuracy for all categories shows a significant drop, indicating the importance of high spatial information. On the basis of multi-spectral information, the inclusion of temporal information (*i.e.*, MT) is beneficial for agricultural land

(including paddy fields, dry farmlands, and greenhouses) as well as natural surfaces (including grassland and water bodies). Considering interpretation ability and data accessibility, it can be said that the multi-temporal Sentinel-2 images have more potential for large scale tasks, while high-spatial and hyperspectral image is more beneficial for the tasks requiring high accuracy.

Comprehensive agricultural area monitoring and management information is provided by the spatio-temporal-spectral fusion method in the field of agriculture. Fine spatial information is delivered by high spatial resolution data, which accurately identifies detailed features in agricultural fields. Rich spectral information is provided by hyperspectral data, allowing detailed spectral features for each pixel to be identified, thus enabling precise identification and quantification of the physiological characteristics, nutritional status, and diseases of different crops. In addition, multi-temporal data facilitate the monitoring of the entire growth cycle of crops from sowing to harvesting, enabling the identification of crop characteristics at different growth periods and the analysis of crop growth through multi-temporal data. Comprehensive and refined farmland monitoring and management are enabled by STSNet, which enhances the efficiency and effectiveness of agricultural production and offers powerful data support and decision-making assistance for modern agriculture.

5. Conclusion

In this paper, we propose a land cover segmentation network, named STSNet, which performs cross-resolution spatio-temporal-spectral deep fusion. First, in the multi-scale spatial-spectral encoder, STSNet employs DSCs to capture spatial and spectral information separately and uses spectral gated module to enhance the analysis and discrimination performance of hyperspectral images. The encoder employs a multi-scale structure and cross-scale connection approach, ensuring effective interaction among multi-scale features. Two cross-scale connections are performed at the beginning and the end of the network, avoiding frequent cross-scale connections that involve compressions and reconstructions of the features. This design preserves the interrelated and complementary information of the features, facilitating the balanced fusion of spatial-spectral information and enhancing the accuracy of spatial-spectral fusion segmentation. Second, STSNet considers the spatial, temporal, and spectral characteristics of multi-temporal Sentinel-2 images in the spatio-temporal encoder module. To this end, a spatio-temporal transformer block is developed, which extracts both global and local spatio-temporal information from Sentinel-2 images. Additionally, spectral convolution is employed to extract spectral information. Lastly, to achieve comprehensive fusion of spatio-temporal-spectral multi-modal images with different spatial resolutions, the newly proposed cross-resolution module effectively mitigates challenges such as information loss in the process of multi-modal data fusion. Specifically, an adjustment coefficient (α) is designed to regulate the importance of features extracted by each modal branch at different spatial resolutions. This mechanism adaptively adjusts the contribution of each modality's features to land cover interpretation.

To better evaluate STSNet, we have open-sourced a multi-modal dataset called the WHU-H^{2}SR-MT, which consists of H^{2}SR images with multi-temporal images. To our knowledge, this is currently the largest spatio-temporal-spectral multi-modal interpretation dataset. Experimental results demonstrate that our multi-scale spatial-spectral encoder module, spatio-temporal encoder module, and cross-resolution fusion module outperform existing mainstream algorithms overall.

In future research, we will explore methods for fusing spatio-temporal-spectral-angle remote sensing images with high spatial resolution. By introducing multi-view images, the spatial structure of features can be captured more comprehensively, effectively overcoming occlusion and distortion issues that may occur in a single view.

CRediT authorship contribution statement

Beibei Yu: Writing – original draft, Software, Methodology, Conceptualization. **Jiayi Li:** Writing – original draft, Validation, Methodology, Investigation. **Xin Huang:** Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Our datasets will be made available at http://irsip.whu.edu.cn/resources/resources_en_v2.php

Acknowledgments

The research was supported by the National Natural Science Foundation of China (under Grants 42471391, 42271328, and 42071311)

References

[1] HankuiK. Zhang, D.P. Roy, D. Luo, Demonstration of large area land cover classification with a one dimensional convolutional neural network applied to single pixel temporal metric percentiles, Remote Sens. Environ. 295 (2023) 113653, https://doi.org/10.1016/j.rse.2023.113653.
[2] J. Yang, X. Huang, The 30 m annual land cover dataset and its dynamics in China from 1990 to 2019, Earth. Syst. Sci. Data 13 (2021) 3907–3925, https://doi.org/10.5194/essd-13-3907-2021.
[3] S.W. Myint, P. Gober, A. Brazel, S. Grossman-Clarke, Q. Weng, Per-pixel *vs*. object-based classification of urban land cover extraction using high spatial resolution imagery, Remote Sens. Environ. 115 (2011) 1145–1161, https://doi.org/10.1016/j.rse.2010.12.017.
[4] F. Zhang, X. Yang, Improving land cover classification in an urbanized coastal area by random forests: the role of variable selection, Remote Sens. Environ. 251 (2020) 112105, https://doi.org/10.1016/j.rse.2020.112105.
[5] W.Y. Yan, A. Shaker, N. El-Ashmawy, Urban land cover classification using airborne LiDAR data: a review, Remote Sens. Environ. 158 (2015) 295–310, https://doi.org/10.1016/j.rse.2014.11.001.
[6] T. Pei, J. Xu, Y. Liu, X. Huang, L. Zhang, W. Dong, C. Qin, C. Song, J. Gong, C. Zhou, GIScience and remote sensing in natural resource and environmental research: status quo and future perspectives, Geogr. Sustain. 2 (2021) 207–215, https://doi.org/10.1016/j.geosus.2021.08.004.
[7] M. Najafzadeh, F. Homaei, H. Farhadi, Reliability assessment of water quality index based on guidelines of national sanitation foundation in natural streams: integration of remote sensing and data-driven models, Artif. Intell. Rev. 54 (2021) 4619–4651, https://doi.org/10.1007/s10462-021-10007-1.
[8] M. Dalponte, T. Jucker, S. Liu, L. Frizzera, D. Gianelle, Characterizing forest carbon dynamics using multi-temporal lidar data, Remote Sens. Environ. 224 (2019) 412–420, https://doi.org/10.1016/j.rse.2019.02.018.
[9] X. Liu, Y. Huang, X. Xu, X. Li, X. Li, P. Ciais, P. Lin, K. Gong, A.D. Ziegler, A. Chen, P. Gong, J. Chen, G. Hu, Y. Chen, S. Wang, Q. Wu, K. Huang, L. Estes, Z. Zeng, High-spatiotemporal-resolution mapping of global urban change from 1985 to 2015, Nat. Sustain. 3 (2020) 564–570, https://doi.org/10.1038/s41893-020-0521-x.
[10] P. Gong, X. Li, J. Wang, Y. Bai, B. Chen, T. Hu, X. Liu, B. Xu, J. Yang, W. Zhang, Y. Zhou, Annual maps of global artificial impervious area (GAIA) between 1985 and 2018, Remote Sens. Environ. 236 (2020) 111510, https://doi.org/10.1016/j.rse.2019.111510.
[11] B. Huang, B. Zhao, Y. Song, Urban land-use mapping using a deep convolutional neural network with high spatial resolution multispectral remote sensing imagery, Remote Sens. Environ. 214 (2018) 73–86, https://doi.org/10.1016/j.rse.2018.04.050.
[12] Y. Li, T. Shi, Y. Zhang, W. Chen, Z. Wang, H. Li, Learning deep semantic segmentation network under multiple weakly-supervised constraints for cross-domain remote sensing image semantic segmentation, ISPRS J. Photogramm. Remote Sens. 175 (2021) 20–33, https://doi.org/10.1016/j.isprsjprs.2021.02.009.
[13] P. Du, X. Bai, K. Tan, Z. Xue, A. Samat, J. Xia, E. Li, H. Su, W. Liu, Advances of four machine learning methods for spatial data handling: a review, J. Geovisual. Spat. Anal. 4 (2020) 13, https://doi.org/10.1007/s41651-020-00048-5.
[14] H. Huang, J. Wang, C. Liu, L. Liang, C. Li, P. Gong, The migration of training samples towards dynamic global land cover mapping, ISPRS J. Photogramm. Remote Sens. 161 (2020) 27–36, https://doi.org/10.1016/j.isprsjprs.2020.01.010.

[15] Y. Zhao, D. Feng, L. Yu, X. Wang, Y. Chen, Y. Bai, H.J. Hernández, M. Galleguillos, C. Estades, G.S. Biging, J.D. Radke, P. Gong, Detailed dynamic land cover mapping of Chile: accuracy improvement by integrating multi-temporal data, Remote Sens. Environ. 183 (2016) 170–185, https://doi.org/10.1016/j.rse.2016.05.016.
[16] M.A. Friedl, D.K. McIver, J.C.F. Hodges, X.Y. Zhang, D. Muchoney, A.H. Strahler, C. E. Woodcock, S. Gopal, A. Schneider, A. Cooper, A. Baccini, F. Gao, C. Schaaf, Global land cover mapping from MODIS: algorithms and early results, Remote Sens. Environ. 83 (2002) 287–302, https://doi.org/10.1016/S0034-4257(02)00078-0.
[17] X.-Y. Tong, G.-S. Xia, X.X. Zhu, Enabling country-scale land cover mapping with meter-resolution satellite imagery, ISPRS J. Photogramm. Remote Sens. 196 (2023) 178–196, https://doi.org/10.1016/j.isprsjprs.2022.12.011.
[18] X. Zhang, S. Du, Learning selfhood scales for urban land cover mapping with very-high-resolution satellite images, Remote Sens. Environ. 178 (2016) 172–190, https://doi.org/10.1016/j.rse.2016.03.015.
[19] Y. Cao, X. Huang, Q. Weng, A multi-scale weakly supervised learning method with adaptive online noise correction for high-resolution change detection of built-up areas, Remote Sens. Environ. 297 (2023) 113779, https://doi.org/10.1016/j.rse.2023.113779.
[20] Y. Zhang, W. Li, R. Tao, J. Peng, Q. Du, Z. Cai, Cross-scene hyperspectral image classification with discriminative cooperative alignment, IEEE Trans. Geosci. Remote Sens. 59 (2021) 9646–9660, https://doi.org/10.1109/TGRS.2020.3046756.
[21] A. Rangnekar, N. Mokashi, E.J. Ientilucci, C. Kanan, M.J. Hoffman, AeroRIT: a new scene for hyperspectral image analysis, IEEE Trans. Geosci. Remote Sens. 58 (2020) 8116–8124, https://doi.org/10.1109/TGRS.2020.2987199.
[22] Y. Xu, J. Gong, X. Huang, X. Hu, J. Li, Q. Li, M. Peng, Luojia-HSSR: a high spatial-spectral resolution remote sensing dataset for land-cover classification with a new 3D-HRNet, Geo-Spatial Inform. Sci. (2022) 1–13, https://doi.org/10.1080/10095020.2022.2070555.
[23] J. Li, X. Huang, L. Tu, WHU-OHS: a benchmark dataset for large-scale Hersepctral Image classification, Int. J. Appl. Earth Obs. Geoinf. 113 (2022) 103022, https://doi.org/10.1016/j.jag.2022.103022.
[24] J. Ren, Y. Shao, H. Wan, Y. Xie, A. Campos, A two-step mapping of irrigated corn with multi-temporal MODIS and Landsat analysis ready data, ISPRS J. Photogramm. Remote Sens. 176 (2021) 69–82, https://doi.org/10.1016/j.isprsjprs.2021.04.007.
[25] F. Gao, M.C. Anderson, X. Zhang, Z. Yang, J.G. Alfieri, W.P. Kustas, R. Mueller, D. M. Johnson, J.H. Prueger, Toward mapping crop progress at field scales through fusion of Landsat and MODIS imagery, Remote Sens. Environ. 188 (2017) 9–25, https://doi.org/10.1016/j.rse.2016.11.004.
[26] Y. Wen, X. Li, H. Mu, L. Zhong, H. Chen, Y. Zeng, S. Miao, W. Su, P. Gong, B. Li, J. Huang, Mapping corn dynamics using limited but representative samples with adaptive strategies, ISPRS J. Photogramm. Remote Sens. 190 (2022) 252–266, https://doi.org/10.1016/j.isprsjprs.2022.06.012.
[27] J. Han, Z. Zhang, Y. Luo, J. Cao, L. Zhang, J. Zhang, Z. Li, The RapeseedMap10 database: annual maps of rapeseed at a spatial resolution of 10 m based on multi-source data, Earth. Syst. Sci. Data 13 (2021) 2857–2874, https://doi.org/10.5194/essd-13-2857-2021.
[28] D. Ienco, R. Interdonato, R. Gaetano, D. Ho Tong Minh, Combining Sentinel-1 and Sentinel-2 satellite image time series for land cover mapping via a multi-source deep learning architecture, ISPRS J. Photogramm. Remote Sens. 158 (2019) 11–22, https://doi.org/10.1016/j.isprsjprs.2019.09.016.
[29] D. He, Q. Shi, J. Xue, P.M. Atkinson, X. Liu, Very fine spatial resolution urban land cover mapping using an explicable sub-pixel mapping network based on learnable spatial correlation, Remote Sens. Environ. 299 (2023) 113884, https://doi.org/10.1016/j.rse.2023.113884.
[30] Y. Zhang, G. Chen, S.W. Myint, Y. Zhou, G.J. Hay, J. Vukomanovic, R. K. Meentemeyer, UrbanWatch: a 1-meter resolution land cover and land use database for 22 major cities in the United States, Remote Sens. Environ. 278 (2022) 113106, https://doi.org/10.1016/j.rse.2022.113106.
[31] J. Li, D. Hong, L. Gao, J. Yao, K. Zheng, B. Zhang, J. Chanussot, Deep learning in multimodal remote sensing data fusion: a comprehensive review, Int. J. Appl. Earth Obs. Geoinf. 112 (2022) 102926, https://doi.org/10.1016/j.jag.2022.102926.
[32] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, J. Chen, DKDFN: domain Knowledge-Guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification, ISPRS J. Photogramm. Remote Sens. 186 (2022) 170–189, https://doi.org/10.1016/j.isprsjprs.2022.02.013.
[33] V. Sainte Fare Garnot, L. Landrieu, N. Chehata, Multi-modal temporal attention models for crop mapping from satellite time series, ISPRS J. Photogramm. Remote Sens. 187 (2022) 294–305, https://doi.org/10.1016/j.isprsjprs.2022.03.012.
[34] M. Zhu, L. Jiao, F. Liu, S. Yang, J. Wang, Residual spectral–spatial attention network for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 59 (2021) 449–462, https://doi.org/10.1109/TGRS.2020.2994057.
[35] S. Liu, H. Zhao, Q. Du, L. Bruzzone, A. Samat, X. Tong, Novel cross-resolution feature-level fusion for joint classification of multispectral and panchromatic remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14, https://doi.org/10.1109/TGRS.2021.3127710.
[36] M. Maimaitijiang, V. Sagan, P. Sidike, S. Hartling, F. Esposito, F.B. Fritschi, Soybean yield prediction from UAV using multimodal data fusion and deep learning, Remote Sens. Environ. 237 (2020) 111599, https://doi.org/10.1016/j.rse.2019.111599.
[37] B. Ren, S. Ma, B. Hou, D. Hong, J. Chanussot, J. Wang, L. Jiao, A dual-stream high resolution network: deep fusion of GF-2 and GF-3 data for land cover classification, Int. J. Appl. Earth Obs. Geoinf. 112 (2022) 102896, https://doi.org/10.1016/j.jag.2022.102896.
[38] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: J, WWM, F.A.F. Navab Nassir, Hornegger (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Springer International Publishing, Cham, 2015, pp. 234–241.
[39] W. Song, S. Li, L. Fang, T. Lu, Hyperspectral image classification with deep feature fusion network, IEEE Trans. Geosci. Remote Sens. 56 (2018) 3173–3184, https://doi.org/10.1109/TGRS.2018.2794326.
[40] M.E. Paoletti, J.M. Haut, R. Fernandez-Beltran, J. Plaza, A.J. Plaza, F. Pla, Deep pyramidal residual networks for spectral–spatial hyperspectral image classification, IEEE Tran. Geosci. Remote Sens. 57 (2019) 740–754, https://doi.org/10.1109/TGRS.2018.2860125.
[41] Z. Zheng, Y. Zhong, A. Ma, L. Zhang, FPGA: fast patch-free global learning framework for fully end-to-end hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 58 (2020) 5612–5626, https://doi.org/10.1109/TGRS.2020.2967821.
[42] K. Sun, B. Xiao, D. Liu, J. Wang, Deep high-resolution representation learning for human pose estimation, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 5686–5696, https://doi.org/10.1109/CVPR.2019.00584.
[43] D. Liao, C. Shi, L. Wang, A spectral–spatial fusion transformer network for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–16, https://doi.org/10.1109/TGRS.2023.3286950.
[44] F. Chollet, Xception: deep learning with depthwise separable convolutions, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 1800–1807, https://doi.org/10.1109/CVPR.2017.195.
[45] B. Cui, X.-M. Dong, Q. Zhan, J. Peng, W. Sun, LiteDepthwiseNet: a lightweight network for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15, https://doi.org/10.1109/TGRS.2021.3062372.
[46] Q. Zhao, J. Liu, Y. Li, H. Zhang, Semantic segmentation with attention mechanism for remote sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13, https://doi.org/10.1109/TGRS.2021.3085889.
[47] K. Li, D. Wang, X. Wang, G. Liu, Z. Wu, Q. Wang, Mixing self-attention and convolution: a unified framework for multisource remote sensing data classification, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–16, https://doi.org/10.1109/TGRS.2023.3310521.
[48] X. Li, M. Xu, S. Liu, H. Sheng, J. Wan, Ultralightweight feature-compressed multihead self-attention learning networks for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–14, https://doi.org/10.1109/TGRS.2024.3404929.
[49] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Netw. 18 (2005) 602–610, https://doi.org/10.1016/j.neunet.2005.06.042.
[50] J.A. Chamorro Martinez, L.E. Cué La Rosa, R.Q. Feitosa, I.D. Sanches, P.N. Happ, Fully convolutional recurrent networks for multidate crop recognition from multitemporal image sequences, ISPRS J. Photogramm. Remote Sens. 171 (2021) 188–201, https://doi.org/10.1016/j.isprsjprs.2020.11.007.
[51] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W. Wong, W. WOO, Convolutional LSTM network: a machine learning approach for precipitation nowcasting, in: C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, R. Garnett (Eds.), Adv Neural Inf Process Syst, Curran Associates, Inc., 2015, in: https://proceedings.neurips.cc/paper/2015/file/07563a3fe3bbe7e3ba84431ad9d055af-Paper.pdf.
[52] M. Tarasiou, E. Chavez, S. Zafeiriou, ViTs for SITS: vision transformers for satellite image time series, (2023). http://arxiv.org/abs/2301.04944.
[53] Z. Qiu, J. Xu, J. Peng, W. Sun, Cross-channel dynamic spatial–spectral fusion transformer for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–12, https://doi.org/10.1109/TGRS.2023.3324730.
[54] U.A. Bhatti, Z. Yu, J. Chanussot, Z. Zeeshan, L. Yuan, W. Luo, S.A. Nawaz, M. A. Bhatti, Q.U. Ain, A. Mehmood, Local similarity-based spatial–spectral fusion hyperspectral image classification with deep CNN and gabor filtering, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15, https://doi.org/10.1109/TGRS.2021.3090410.
[55] S. Jia, Z. Min, X. Fu, Multiscale spatial–spectral transformer network for hyperspectral and multispectral image fusion, Inform. Fusion 96 (2023) 117–129, https://doi.org/10.1016/j.inffus.2023.03.011.
[56] J. Adrian, V. Sagan, M. Maimaitijiang, Sentinel SAR-optical fusion for crop type mapping using deep learning and Google Earth Engine, ISPRS J. Photogramm. Remote Sens. 175 (2021) 215–235, https://doi.org/10.1016/j.isprsjprs.2021.02.018.
[57] Z. Cai, Q. Hu, X. Zhang, J. Yang, H. Wei, J. Wang, Y. Zeng, G. Yin, W. Li, L. You, B. Xu, Z. Shi, Improving agricultural field parcel delineation with a dual branch spatiotemporal fusion network by integrating multimodal satellite data, ISPRS J. Photogramm. Remote Sens. 205 (2023) 34–49, https://doi.org/10.1016/j.isprsjprs.2023.09.021.
[58] Feng Gao, J. Masek, M. Schwaller, F. Hall, On the blending of the Landsat and MODIS surface reflectance: predicting daily Landsat surface reflectance, IEEE Trans. Geosci. Remote Sens. 44 (2006) 2207–2218, https://doi.org/10.1109/TGRS.2006.872081.
[59] X. Liu, C. Deng, J. Chanussot, D. Hong, B. Zhao, StfNet: a two-stream convolutional neural network for spatiotemporal image fusion, IEEE Trans. Geosci. Remote Sens. 57 (2019) 6552–6564, https://doi.org/10.1109/TGRS.2019.2907310.
[60] X. Meng, Q. Liu, F. Shao, S. Li, Spatio–temporal–spectral collaborative learning for spatio–temporal fusion with land cover changes, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16, https://doi.org/10.1109/TGRS.2022.3185459.
[61] X. Chen, X. Meng, F. Shao, W. Sun, PSSTFN: a progressive spatial-temporal-spectral fusion network for remote sensing images, IEEE Trans. Geosci. Remote Sens. (2023), https://doi.org/10.1109/TGRS.2023.3329531, 1–1.

[62] X. Li, G. Zhang, H. Cui, S. Hou, Y. Chen, Z. Li, H. Li, H. Wang, Progressive fusion learning: a multimodal joint segmentation framework for building extraction from optical and SAR images, ISPRS J. Photogramm. Remote Sens. 195 (2023) 178–191, https://doi.org/10.1016/j.isprsjprs.2022.11.015.
[63] J. Li, Q. Hu, Y. Zhang, Multimodal image matching: a scale-invariant algorithm and an open dataset, ISPRS J. Photogramm. Remote Sens. 204 (2023) 77–88, https://doi.org/10.1016/j.isprsjprs.2023.08.010.
[64] D. Hong, B. Zhang, H. Li, Y. Li, J. Yao, C. Li, M. Werner, J. Chanussot, A. Zipf, X. X. Zhu, Cross-city matters: a multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks, Remote Sens. Environ. 299 (2023) 113856, https://doi.org/10.1016/j.rse.2023.113856.
[65] M.D. Zeiler, G.W. Taylor, R. Fergus, Adaptive deconvolutional networks for mid and high level feature learning, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2018–2025, https://doi.org/10.1109/ICCV.2011.6126474.
[66] K. Wu, J. Fan, P. Ye, M. Zhu, Hyperspectral image classification using spectral–spatial token enhanced transformer with hash-based positional embedding, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–16, https://doi.org/10.1109/TGRS.2023.3258488.
[67] H. Fu, G. Sun, L. Zhang, A. Zhang, J. Ren, X. Jia, F. Li, Three-dimensional singular spectrum analysis for precise land cover classification from UAV-borne hyperspectral benchmark datasets, ISPRS J. Photogramm. Remote Sens. 203 (2023) 115–134, https://doi.org/10.1016/j.isprsjprs.2023.07.013.
[68] S. Pande, B. Banerjee, HyperLoopNet: hyperspectral image classification using multiscale self-looping convolutional networks, ISPRS J. Photogramm. Remote Sens. 183 (2022) 422–438, https://doi.org/10.1016/j.isprsjprs.2021.11.021.
[69] X. Qiao, S.K. Roy, W. Huang, Multi-scale neighborhood attention transformer with optimized spatial pattern for hyperspectral image classification, IEEE Trans. Geosci. Remote Sens. (2023), https://doi.org/10.1109/TGRS.2023.3314550, 1–1.
[70] D. Mehta, A. Skliar, H. Ben Yahia, S. Borse, F. Porikli, A. Habibian, T. Blankevoort, Simple and efficient architectures for semantic segmentation, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE, 2022, pp. 2627–2635, https://doi.org/10.1109/CVPRW56347.2022.00296.
[71] L. Yang, R. Huang, J. Huang, T. Lin, L. Wang, R. Mijiti, P. Wei, C. Tang, J. Shao, Q. Li, X. Du, Semantic segmentation based on temporal features: learning of temporal–spatial information from time-series SAR images for paddy rice mapping, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16, https://doi.org/10.1109/TGRS.2021.3099522.
[72] M. Tarasiou, R.A. Guler, S. Zafeiriou, Context-self contrastive pretraining for crop type semantic segmentation, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–17, https://doi.org/10.1109/TGRS.2022.3198187.
[73] V. Sainte Fare Garnot, L. Landrieu, Panoptic segmentation of satellite image time series with convolutional temporal attention networks, in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE, 2021, pp. 4852–4861, https://doi.org/10.1109/ICCV48922.2021.00483.
[74] Rose M Rustowicz, Robin Cheong, Lijing Wang, Stefano Ermon, Marshall Burke, David Lobell, Semantic segmentation of crop type in Africa: a novel dataset and analysis of deep learning methods, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2019, pp. 75–82.
[75] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L.u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, Curran Associates, Inc., 2017, pp. 6000–6010. Vol. 30, https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[76] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, H. Hu, Video swin transformer, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2022, pp. 3192–3201, https://doi.org/10.1109/CVPR52688.2022.00320.
[77] H. Hosseinpour, F. Samadzadegan, F.D. Javan, CMGFNet: a deep cross-modal gated fusion network for building extraction from very high-resolution remote sensing images, ISPRS J. Photogramm. Remote Sens. 184 (2022) 96–115, https://doi.org/10.1016/j.isprsjprs.2021.12.007.
[78] M. Sui, H. Li, Z. Zhu, F. Zhao, AFNet-M: adaptive fusion network with masks for 2D +3D facial expression recognition, in: 2023 IEEE International Conference on Image Processing (ICIP), IEEE, 2023, pp. 116–120, https://doi.org/10.1109/ICIP49359.2023.10222441.
[79] X. Tang, M. Li, J. Ma, X. Zhang, F. Liu, L. Jiao, EMTCAL: efficient multiscale transformer and cross-level attention learning for remote sensing scene classification, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–15, https://doi.org/10.1109/TGRS.2022.3194505.
[80] R. Zhao, Z. Shi, Z. Zou, High-resolution remote sensing image captioning based on structured attention, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14, https://doi.org/10.1109/TGRS.2021.3070383.
[81] P. Griffiths, C. Nendel, P. Hostert, Intra-annual reflectance composites from Sentinel-2 and Landsat for national-scale crop and land cover mapping, Remote Sens. Environ. 220 (2019) 135–151, https://doi.org/10.1016/j.rse.2018.10.031.
[82] D. Sykas, I. Papoutsis, D. Zografakis, Sen4AgriNet: a harmonized multi-country, multi-temporal benchmark dataset for agricultural earth observation machine learning applications, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, IEEE, 2021, pp. 5830–5833, https://doi.org/10.1109/IGARSS47720.2021.9553603.
[83] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, (2014), arXiv preprint arXiv:1412.6980.
[84] S. Xu, X. Zhu, J. Chen, X. Zhu, M. Duan, B. Qiu, L. Wan, X. Tan, Y.N. Xu, R. Cao, A robust index to extract paddy fields in cloudy regions from SAR time series, Remote Sens. Environ. 285 (2023) 113374, https://doi.org/10.1016/j.rse.2022.113374.
[85] F. Zhao, R. Sun, L. Zhong, R. Meng, C. Huang, X. Zeng, M. Wang, Y. Li, Z. Wang, Monthly mapping of forest harvesting using dense time series Sentinel-1 SAR imagery and deep learning, Remote Sens. Environ. 269 (2022) 112822, https://doi.org/10.1016/j.rse.2021.112822.
[86] L. Bai, S. Du, X. Zhang, H. Wang, B. Liu, S. Ouyang, Domain adaptation for remote sensing image semantic segmentation: an integrated approach of contrastive learning and adversarial learning, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13, https://doi.org/10.1109/TGRS.2022.3198972.
[87] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, P.M. Atkinson, UNetFormer: a UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery, ISPRS J. Photogramm. Remote Sens. 190 (2022) 196–214, https://doi.org/10.1016/j.isprsjprs.2022.06.008.
[88] A. He, K. Wang, T. Li, C. Du, S. Xia, H. Fu, H2Former: an efficient hierarchical hybrid transformer for medical image segmentation, IEEE Trans. Med. ImAging 42 (2023) 2763–2775, https://doi.org/10.1109/TMI.2023.3264513.
[89] S.K. Sønderby, C.K. Sønderby, H. Nielsen, O. Winther, Convolutional LSTM networks for subcellular localization of proteins, in: 2015: pp. 68–80. https://doi.org/10.1007/978-3-319-21233-3_6.
[90] N. Ballas, L. Yao, C. Pal, A. Courville, Delving deeper into convolutional networks for learning video representations, (2016), arXiv preprint arXiv:1511.06432.
[91] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16×16 words: transformers for image recognition at scale, (2020), arXiv preprint arXiv:2010.11929.
[92] X. Huang, L. Zhang, An SVM ensemble approach combining spectral, structural, and semantic features for the classification of high-resolution remotely sensed imagery, IEEE Trans. Geosci. Remote Sens. 51 (2013) 257–272, https://doi.org/10.1109/TGRS.2012.2202912.
[93] Y. Cao, Y. Wu, M. Li, W. Liang, X. Hu, DFAF-Net: a dual-frequency PolSAR image classification network based on frequency-aware attention and adaptive feature fusion, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–18, https://doi.org/10.1109/TGRS.2022.3152854.
[94] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, R. Stiefelhagen, CMX: cross-modal fusion for RGB-X semantic segmentation with transformers, (2023), arXiv preprint arXiv:2203.04838.
[95] J. Wang, Z. Shao, X. Huang, T. Lu, R. Zhang, A dual-path fusion network for pan-sharpening, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14, https://doi.org/10.1109/TGRS.2021.3090585.