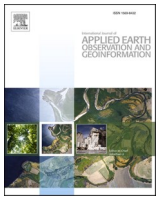
International Journal of Applied Earth Observation and Geoinformation 138 (2025) 104457

Contents lists available at ScienceDirect

International Journal of Applied Earth Observation and Geoinformation

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



PatchOut: A novel patch-free approach based on a transformer-CNN hybrid framework for fine-grained land-cover classification on large-scale airborne hyperspectral images

Renjie Ji [a,b,c], Kun Tan [a,b,c,d,*], Xue Wang [a,b,c], Shuwei Tang [a,b,c], Jin Sun [a,b,c], Chao Niu [a,b,c], Chen Pan [e]

[a] Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China
[b] Key Laboratory of Geographic Information Science (Ministry of Education), East China Normal University, Shanghai 200241, China
[c] School of Geographic Sciences, East China Normal University, Shanghai 200241, China
[d] School of Geospatial Artificial Intelligence, East China Normal University, Shanghai 200241, China
[e] Shanghai Municipal Institute of Surveying and Mapping, Shanghai 200063, China

ARTICLE INFO

Keywords:
Hyperspectral image classification
Patch-free
Semantic segmentation
Multi-scale feature fusion

ABSTRACT

Airborne hyperspectral systems can provide high-resolution hyperspectral images (HSIs) covering large scenes, enabling fine-grained land-cover classification. However, the most popular patch-based methods are limited by low computational efficiency and broken classification results, which hinders the full utilization of this powerful technology in Earth observation applications. Therefore, in this paper, considering the efficiency requirements for large-scale land-cover classification, a novel **p**atch-free **a**pproach based on a **T**ransformer-**C**NN **h**ybrid (PatchOut) framework is proposed. The proposed PatchOut framework adopts an encoder-decoder architecture, enabling rapid semantic segmentation for HSI classification. For the encoder module, we introduce a computationally efficient reduced Transformer module integrated with convolutional neural network (CNN), to leverage their complementary strengths for long-range and local feature extraction, respectively. A multi-scale spatial-spectral feature fusion (MSSSFF) module is also proposed to amalgamate the characteristics of different levels from the encoder, which enhances the overall feature representation. Then, to address the loss of semantic detail and resolution inherent in multi-level feature extraction, a novel feature reconstruction module (FRM) is applied to recover high-quality semantic features. Finally, a large-scale benchmark dataset, Qingpu-HSI, is presented, comprising airborne HSIs covering 33.91 km^2 with 20 land-cover classes. Experiments on the Qingpu-HSI and another public dataset demonstrate the superior accuracy and efficiency of our proposed PatchOut framework, outperforming several well-known patch-free and patch-based methods. The Qingpu HSI dataset, along with the PatchOut framework code will be released at https://github.com/busbyjrj/PatchOut.

1. Introduction

Accurately classifying and mapping land use through remote sensing represents a persistent research focus in Earth observation (Anderson et al., 2017; Yao et al., 2023). Hyperspectral images (HSIs) can reflect more detailed information of ground objects due to their unprecedented hundreds to thousands of continuous narrow bands and the fine-grained spatial distribution information (Paoletti et al., 2019). Consequently, HSIs play an essential role in facilitating land-cover classification and monitoring, such as wetland classification (Su et al., 2021), agricultural crop classification (Zhong et al., 2020) and tree species identification (Fu et al., 2023).

HSI classification seeks to assign a unique semantic label to each pixel vector. (Bioucas-Dias et al., 2013). In recent years, deep learning has demonstrated its powerful capabilities for image recognition, leading to its increased application in other related fields, including HSI classification (Li et al., 2019; Paoletti et al., 2019). As a classical deep learning task, classification models based on a patch input were first

* Corresponding author at: Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities (Ministry of Natural Resources), East China Normal University, Shanghai 200241, China.
E-mail address: tankuncu@gmail.com (K. Tan).

https://doi.org/10.1016/j.jag.2025.104457
Received 31 October 2024; Received in revised form 25 February 2025; Accepted 1 March 2025

used in HSI classification. Convolutional neural network (CNN)-based models such as SSRN (Zhong et al., 2018), A^2S^2K-ResNet (Roy et al., 2021), etc., have shown great performances in HSI classification. These models employ 2D- or 3D-CNNs to effectively learn hierarchical spatial-spectral representations. Recently, Transformer-based models have become a hot research topic. Researchers are exploring the combined advantages of CNNs and Transformers. Hence, CNN and Transformer hybrid networks, e.g. SSFTT (Sun et al., 2022), and the spatial-spectral feature extraction network with patch attention module (PASSNet) (Ji et al., 2023), have recently been designed, and have shown powerful performances in feature extraction.

However, the aforementioned methods, based on CNNs and Transformers, are implemented using HSI patches. The patch-based methods take an HSI patch as input, but only one pixel in its center can be predicted at a time. Although the patch-based methods can achieve satisfactory classification maps, it is difficult to avoid the problem of calculation redundancy and high time consumption. Recently, large HSIs are becoming more common due to advanced high-resolution hyperspectral sensors for airborne and unmanned aerial vehicle (UAV) platforms. The limitations of patch-based models prevent effective HSI classification in large scenes. With the fast patch-free global learning (FPGA) framework, a fully end-to-end classification framework was introduced, providing a more efficient option for HSI classification, which is actually a semantic segmentation task (Zheng et al., 2020). Benefiting from the end-to-end pixel-level prediction capability of the semantic segmentation models, patch-free models, e.g., unified multi-scale learning (UML) (Wang et al., 2022b), and the lightweight Transformer (LiT) network (Zhang et al., 2023a), can efficiently classify and annotate every pixel in the input image with high precision.

For the semantic segmentation models, a larger labeled dataset is required. It has been pointed out that the limited availability of HSI classification datasets has constrained the development of deep learning models (Schmitt et al., 2023). Some models utilizing whole-image inputs face inherent data leakage challenges between training and evaluation sets, even with label mask strategies, thus affecting the reliability of accuracy assessments (Zhang et al., 2023b). On the other hand, the input feature size in some models, such as the LiT network, is set very small, e.g., 32×32 or 64×64, which affects the ability of acquiring and associating long-range dependencies.

Deep learning approaches have demonstrated impressive performance in HSI classification, but training highly accurate deep learning models is still faced with many challenges. There are three primary issues that warrant attention:

(1) As the spatial resolution of HSIs improves, a larger input image size is required to capture complete ground object boundaries. Subsequently, fusing the spatial and spectral features is essential for achieving comprehensive scene understanding and accurate segmentation.

(2) The Transformer architecture possesses the capability to capture long-range dependencies; however, it suffers from a substantial computational overhead and is prone to a suboptimal generalization performance when trained on limited samples.

(3) The existing HSI datasets struggle to meet the needs of the patch-free classification methods. In order to properly utilize HSI data for semantic segmentation and reasonably evaluate the performance, a large-scale finely labeled HSI classification dataset is required.

In order to alleviate the above problems, a patch-free approach based on a Transformer-CNN hybrid (PatchOut) framework is proposed for large-scale HSI classification tasks in this paper. Specifically, employing the encoder-decoder architecture enables PatchOut to extract spatial-spectral features at multi levels, adapting to the fine-grained features of various land-cover types. The PatchOut framework is aimed at harnessing the strength of Transformers in capturing long-range dependencies and the prowess of CNNs in modeling local features and incorporating inductive bias. Moreover, the multi-scale spatial-spectral feature fusion (MSSSFF) module facilitates multi-level feature fusion, enabling the framework to extract intricate HSI features and enhance classification performance. Furthermore, a feature reconstruction module (FRM) based on a lightweight Transformer structure is proposed to bridge disparity between encoder and decoder features with different sizes. The primary contributions of this work are summarized below.

(1) To mitigate the computational complexity inherent in Transformer structures, a reduced Transformer block (RTB) mechanism is introduced. In the encoder blocks, this mechanism efficiently captures deep and long-range associative information with a small computational overhead. In the decoder blocks, the proposed FRM is able to fuse and reconstruct low-resolution features, which not only enhances the feature quality but also mitigates the potential semantic discrepancies in encoder-decoder skip connections.

(2) To further amalgamate the characteristics of the different levels from the encoder module, the MSSSFF module is proposed. Features spanning diverse scales are encoded into a unified dimensional space, and the Transformer is leveraged to aggregate the contextual semantic information, which enables the module to capture the local cross-channel interaction and inter-channel dependencies, enhancing the overall feature representation.

(3) We built a large-scale manually annotated HSI classification dataset—the Qingpu HSI dataset—dedicated to fine-grained classification of vegetation species and land-cover categories. This dataset is able to effectively avoid the issue of training and test data leakage, and provides a novel benchmark for the fine-grained classification of vegetation and land-cover types, while being particularly suitable for patch-free semantic segmentation models.

2. Related work

2.1. CNNs for HSI classification

CNNs have proven highly effective at extracting spatial-spectral features from HSI data. In the early days, one-dimensional (1D) CNNs were first used for HSI classification; however, these methods solely utilize spectral information, while neglecting spatial features (Yue et al., 2015). The advent of 2D CNNs has made up for this deficiency, resulting in a significant improvement in classification accuracy (Makantasis et al., 2015). To better extract spatial-spectral features simultaneously, 3D CNNs have been introduced. SSRN achieves higher accuracy with the 3D CNN architecture and consecutive spectral-spatial residual blocks, but requires significantly more computation (Zhong et al., 2018). The hybrid spectral convolutional neural network concatenates 3D and 2D convolutions, balancing the complexity and classification accuracy (Roy et al., 2020). Recent advancements in group convolution and attention mechanisms have enabled lightweight CNN architectures to achieve competitive classification results, e.g., the lightweight spectral-spatial attention network (Cui et al., 2022), etc.

On the other hand, fully convolutional networks (FCNs) eliminate the final fully connected layers, enabling pixel-to-pixel classification on input images. The FPGA framework adopts the encoder and decoder structure, and can obtain spatial and spectral features of different scales (Zheng et al., 2020). UML introduces channel shuffle and channel attention mechanisms into an FCN, resulting in a lightweight and efficient model (Wang et al., 2022b). The spectral patching network integrates a residual architecture and atrous spatial pyramid pooling modules to effectively capture multi-scale semantic information (Hu et al., 2022). 3D-HRNet uses an attention-based 3D CNN module to capture global–local spectral features, thereby optimizing performance (Xu et al., 2023).

However, the CNN-based architectures face inherent limitations in modeling long-range semantic relations due to their finite receptive fields. In the process of multi-scale feature fusion, particular attention should be paid to the disparities between local and long-range features across different spatial resolutions.

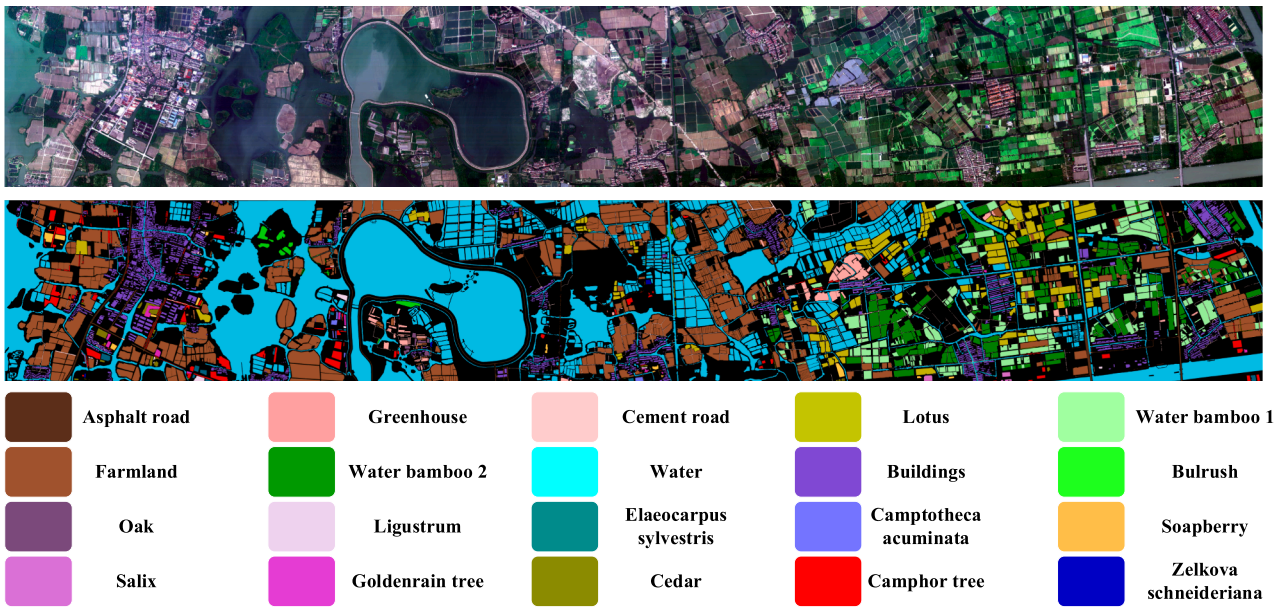


Fig. 1. The true-color images and the corresponding ground truths of the Qingpu HSI dataset.

2.2. Transformers for HSI classification

Recently, vision Transformer (ViT) models have offered an alternative architectural paradigm for image processing tasks (Dosovitskiy et al., 2021). The ViT-based architectures, which process images as sequences of flattened patches, have gained prominence in various visual tasks, including the HSI classification. Their efficacy in capturing long-range dependencies and extracting hierarchical features has led to improved performance across spatial and spectral dimensions. SpectralFormer (SF) rethinks the HSI classification task from a sequential perspective with the Transformer model, and can extract and leverage the local spectral sequence information from the adjacent spectral bands (Hong et al., 2022). Recently, some hybrid models have combined Transformers and CNNs to utilize their respective advantages and realize comprehensive mining of local and long-range features. PASSNet uses hybrid CNN-Transformer architecture with patch attention mechanism to extract spatial-spectral features from local to long range scales (Ji et al., 2023).

For semantic segmentation tasks with limited data, integrating CNNs and Transformers has demonstrated enhanced feature extraction capabilities (Dai et al., 2021), notably in applications like medical image segmentation (Gao et al., 2021), building extraction (Fu et al., 2024), etc. For HSI data with a small sample size, there has still been little research on hybrid semantic segmentation models. HSI-TransUNet leverages residual-connect Transformers to extract lobal contextual features, and obtain superior performance in the UAV HSI crop classification (Niu et al., 2022). The LiT network integrates lightweight convolutional modules and self-attentions structures, employing a controlled multiclass stratified sampling strategy to avoid classification overfitting problems (Zhang et al., 2023a). For multi-scale HSI feature extraction, the S^2HM^2 framework utilizes a self-supervised learning approach based on a 3D masking strategy and a 3D SwinTransformer, demonstrating effective performance (Tu et al., 2024).

In general, the Transformer-based networks for HSI classification tasks generally demonstrate a higher computational complexity than their CNN-based counterparts. Furthermore, in semantic segmentation tasks, the Transformer architecture requires larger training datasets. Consequently, the development of hybrid CNN-Transformer models merits further investigation, to optimize the performance and efficiency in HSI classification applications.

2.3. Other HSI classification methods

Some other deep learning models have also been applied to HSI classification. For example, generative adversarial networks are utilized to automatically augment training datasets to mitigate the small sample size issue in HSI classification (Roy et al., 2022). The combination of capsule network with CNN and Transformer is also widely used for hyperspectral classification (Wang et al., 2023). Recently, a novel network architecture based on state space models, named Mamba, has also been introduced for HSI classification tasks (Wang et al., 2024). Graph Convolutional Networks (GCNs), when integrated with superpixels, offer an effective way to improve the consistency of classification results, providing another alternative to patch-based methods (Li et al., 2023).

3. Materials

3.1. Qingpu-HSI dataset

The limited size of HSI classification datasets limits further development of deep learning in this field (Schmitt et al., 2023), and also leads to the problem of training and test data leakage (Zhang et al., 2023a). For the fine land-cover classification needs of large-scale aerial HSIs, we built the Qingpu-HSI dataset. The Qingpu dataset was finely annotated manually, according to the properties of the ground features, rather than generated labels from geographic information data. To our knowledge, Qingpu-HSI is currently the largest manually annotated aerial HSI classification dataset.

The airborne HSIs were captured in Qingpu District, which is an outer suburb of Shanghai, China on June 16, 2022, using the Airborne Multi-Modality Imaging Spectrometer (AMMIS) developed by the Shanghai Institute of Technical Physics at the Chinese Academy of Sciences. The visible and near-infrared (VNIR) module of AMMIS utilizes a 256-band sensor (400–1000 nm). After removing five bad bands, 251 bands were ultimately used for analysis. A fixed-wing aircraft equipped with the AMMIS sensor captured HSIs at 3000 m altitude with a of 0.75 m spatial resolution. We conducted dark current correction, radiometric calibration, and geometric calibration on the HSIs. For the specific steps, please refer to Niu et al. (2024).

A representative area (20480 × 2944 pixels) was selected for the precise land-cover classification (Fig. 1). Ground truth was established

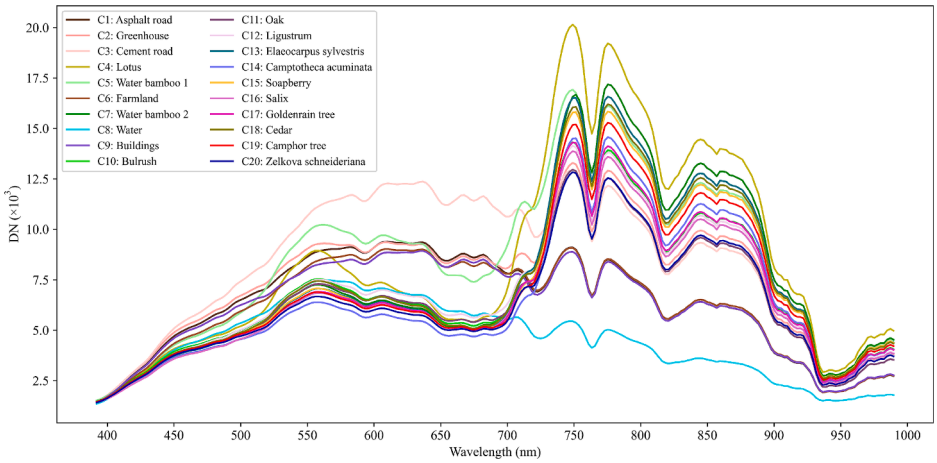


Fig. 2. Mean digital number values of the spectral signatures for the 20 land-cover types.

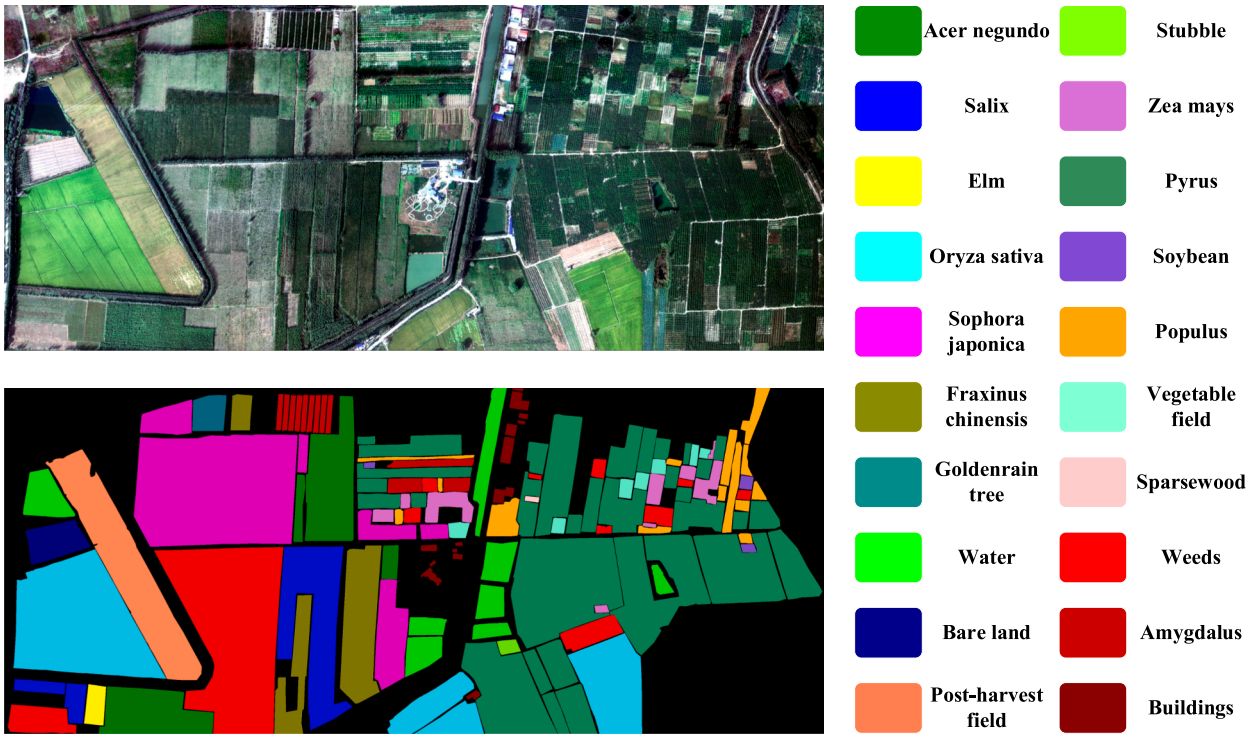


Fig. 3. The true-color images and the corresponding ground truths of the Matiwan HSI dataset.

through comprehensive field investigation using the Xingse mobile application (version 3.16), identifying 20 distinct land-cover types and their spatial distributions. To delineate ground object boundaries, historical images from ESRI and Google Earth, supplemented by 0.1-m aerial RGB imagery acquired in 2022 by the Shanghai Institute of Surveying and Mapping, was consulted. The resulting Qingpu HSI datasets comprises 20 land-cover categories, including four artificial surfaces, five crop types, ten tree species, and water bodies. Fig. 2 presents the mean digital number (DN) values of the spectral signatures for the 20 land-cover types. While exhaustive labeling of land-cover types and pixels is prioritized, the resulting dataset exhibits a highly imbalanced class distribution with a pronounced long tail, posing a significant challenge for semantic segmentation tasks.

3.2. *Matiwan HSI dataset*

The Matiwan HSI dataset serves as another benchmark for evaluating fine-grained vegetation classification performance, which was acquired in Xiong'an New Area in China, by the AMMIS VNIR module in 2017, with 256 spectral bands (Jia et al., 2022). The Matiwan dataset comprises 3750 × 1580 pixels at 0.5 m spatial resolution. It has 20 categories, among which crops are the main categories. Fig. 3 shows the RGB true-color image and corresponding labels for the Matiwan dataset. The Matiwan dataset exhibits significant intra-class variation, primarily due to varying vegetation growth stage and shadowing from street trees, posing a challenge for accurate HSI classification (Jia et al., 2022).

4. Proposed Methodology

This study aims to design a novel hybrid convolution and

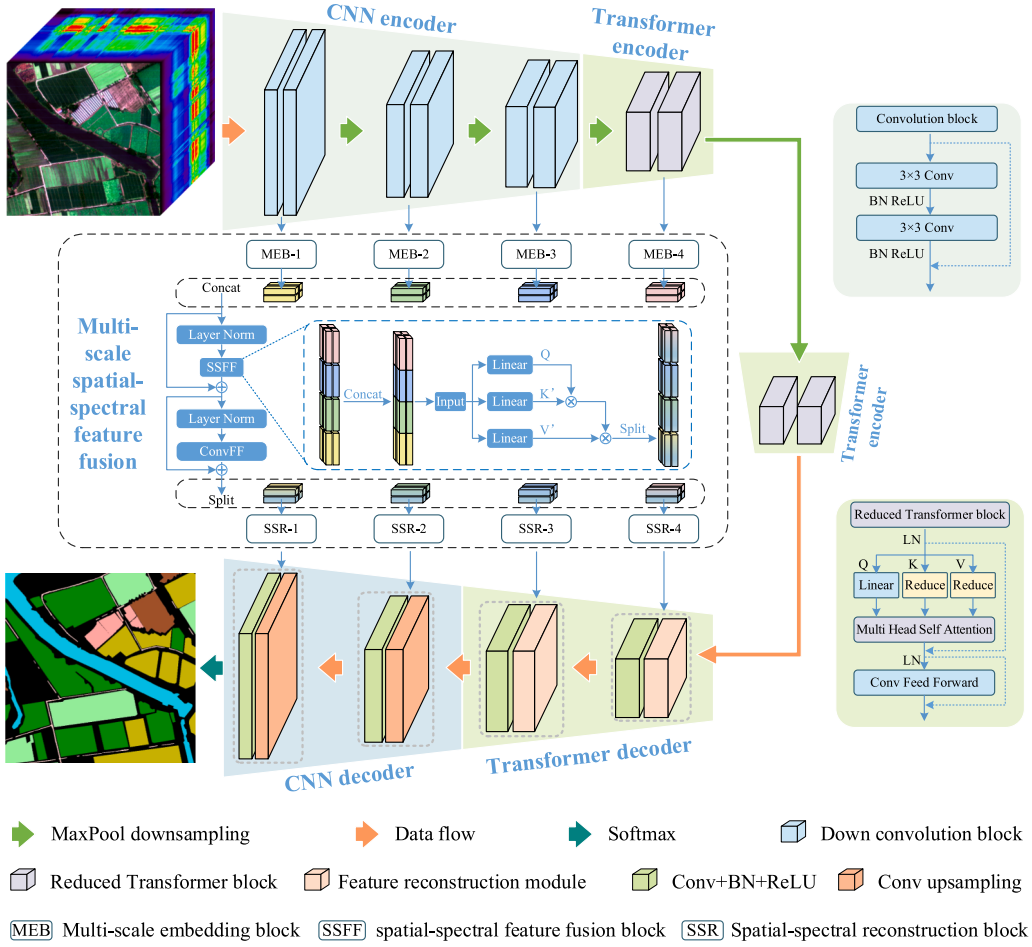


Fig. 4. The proposed PatchOut architecture.

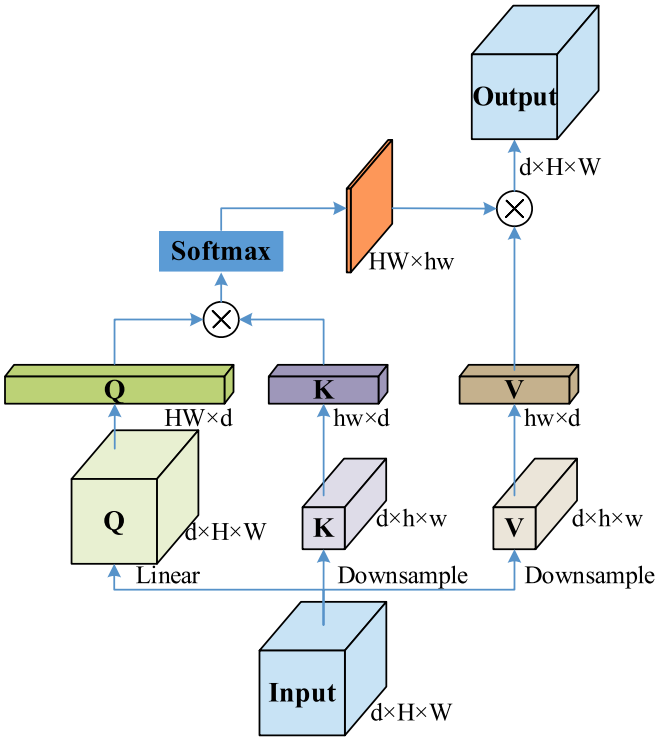


Fig. 5. The proposed reduced Transformer block.

Transformer method, with an encoder-decoder structure, to enhance local and long-range feature fusion in HSIs. Specifically, a reduced Transformer block (RTB) is used to build a long-range feature acquisition module in the encoder and an FRM in the decoder. Besides, the MSSSFF module is also proposed between the encoder and the decoder modules. In the following, detailed descriptions of the proposed PatchOut framework are given, and the overview of the architecture is shown in Fig. 4.

4.1. Transformer blocks

4.1.1. Reduced Transformer block

The Transformer blocks, which are composed of a multi-head self-attention (MHSA) layer and a feed-forward layer, are capable of capturing the long-range dependencies and contextual features from the input. However, the computational demands of Transformers are considerable, especially for a larger input scale. Considering that remote sensing images are structured data, in high-resolution remote sensing images, a single ground object often consists of multiple pixels with similar spectral features. In other words, we can compress the input features while extracting features to reduce redundant and inefficient recompilation. Thus, we designed a reduced Transformer block, as shown in Fig. 5.

Similar to MHSA, consider an input feature map $X \in \mathscr{R}^{C\times H\times W}$, where C, H, and W are the number of channels, spatial height, and width, respectively. Firstly, X is projected through depthwise convolution (*DWConv*) layers to obtain a query $Q \in \mathscr{R}^{C\times H\times W}$, a key $K \in \mathscr{R}^{C\times H\times W}$, and a value $V \in \mathscr{R}^{C\times H\times W}$. The difference is that the proposed RTB also uses

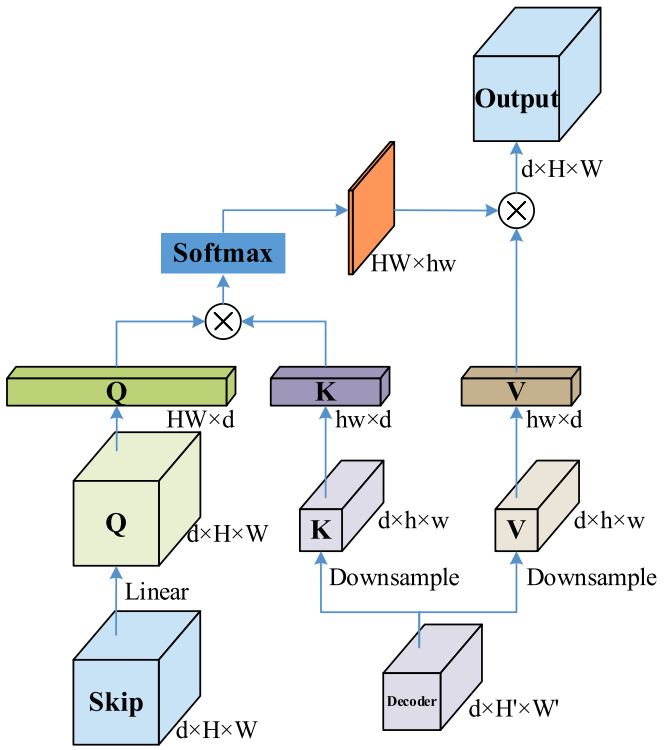


Fig. 6. The proposed feature reconstruction module.

bilinear downsampling to compress the spatial size of K and V into low-resolution features $K^{'}, V^{'} \in \mathscr{R}^{C\times h\times w}$, where h and w are the reduced spatial size of the feature map after downsampling ($h < H$ and $w < W$). The details of RTB can be formulated as follows, to simplify the presentation, and the multi-head structure is not presented:

$$Q = DWConv(X) \tag{1}$$

$$K^{'} = SR(K) = SR(DWConv(X)) \tag{2}$$

$$V^{'} = SR(V) = SR(DWConv(X)) \tag{3}$$

where $DWConv$ has a kernel size of 3. SR represents the downsampling bilinear interpolation operation to reduce the spatial size of K or V. The proposed reduced self-attention is then calculated as:

$$Attention(Q, K^{'}, V^{'}) = Softmax\left(\frac{QK^{'T}}{\sqrt{d_{head}}}\right)V^{'} \tag{4}$$

$$\widehat{X} = Attention(Q, K^{'}, V^{'})W^{O} \tag{5}$$

where d_{head} is the dimension of the heads. $W^{O} \in \mathbb{R}^{C\times C}$ refers to the output linear projection matrix. With the above formula, it can be found that the compilation amount will decrease significantly, and thus the proposed RTB has the ability to handle larger input feature maps, without image patch embedding.

Since HSI features are rich in spatial characteristics, a convolution operation replaces linear projection in the feed-forward layer. Thus, the output of the convolutional feed-forward (ConvFF) layers is computed as:

$$\widehat{X} = GELU(PWConv(X)) \tag{6}$$

$$\widehat{X} = GELU(DWConv(\widehat{X})) + X \tag{7}$$

where $PWConv$ is the pointwise convolution, and $DWConv$ has a kernel size of 3.

As depicted in Fig. 4, the RTB incorporate layer normalization before MHSA and ConvFF layers, along with residual connections. These operations contribute to improved model stability and accelerated learning. Considering that the convolution operation is applied instead of a linear operation, no position encoding methods are adopted.

4.1.2. Feature reconstruction module

The decoder blocks reconstruct high-resolution representations by fusing contextual information from low-resolution features of high-level encoder with spatial details from high-resolution features derived from skip connections. Since HSIs have rich deep features, simple fusion methods such as addition or concatenation cannot deal with the details of spatial-spectral fusion. In order to better construct the fusion decoder and reconstruct the high-resolution feature map, we designed the FRM by using the RTB blocks.

As shown in Fig. 6, firstly, a Transformer block is used to fuse the low-level features L and high-level features H. The MHSA mechanism used in the FRM has a similar structure to the RTB, but it takes two inputs. The query $Q \in \mathscr{R}^{C\times H\times W}$ with high-resolution features comes from the skip connections of the MSSSFF module H, and the key $K \in \mathscr{R}^{C\times H^{'}\times W^{'}}$ and value $V \in \mathscr{R}^{C\times H^{'}\times W^{'}}$ with low-resolution features come from the previous decoder layer L.

$$Q = DWConv(H) \tag{8}$$

$$K^{'} = SR(K) = SR(DWConv(L)) \tag{9}$$

$$V^{'} = SR(V) = SR(DWConv(L)) \tag{10}$$

Bilinear interpolation is then used to enhance the spatial resolution of L and form the residual structure, followed by the ConvFF structure.

$$\widehat{X} = Attention(Q, K^{'}, V^{'})W^{O} + Up(L) \tag{11}$$

where Up denotes bilinear upsampling to restore the spatial size of L.

The fused result $\widehat{X}$ and the high-level features H are then concatenated, and a convolutional layer is used to obtain U, followed by batch normalization and the rectified linear unit ($ReLU$) activation function, thereby ensuring that the scales of the features are balanced.

$$U = ReLU(BN(Conv(Concat(\widehat{X}, H)))) \tag{12}$$

where $Concat$ denotes the concatenation operation, $Conv$ denotes the convolution operation with the kernel size of 3, BN refers to the batch normalization, and U refers to the final outputs.

4.2. Encoder-decoder structure

The integration of CNN and Transformer within a hybrid structure can fuse the advantages of translation equivariance and a global receptive field and balance the generalization ability and model capacity (Dai et al., 2021; Ji et al., 2023). This work explores the integration of convolution and self-attention mechanisms to capitalize on their respective strengths.

4.2.1. Encoder blocks

The encoder (Fig. 4) consists of three CNN layers succeeded by two reduced Transformer layers. Considering the high dimensionality of HSIs, it is not sufficient for convolution kernels to use a small number of channels, so the stem layer employs 64 channels. At the beginning of each encoder layer, a max-pooling operation of size 2×2 is employed to halves the spatial resolution. The convolutional layers comprise two convolutional modules with the kernel size of 3. Then, to mitigate the computational demands of the Transformer layers, the reduced ratio in RTB is set to 4 and 2 for the Qingpu and Matiwan datasets, respectively, resulting in a reduced feature map width and height of $K^{'}, V^{'}$ as 16, thereby lessening the computation burden. The last four stages employ 128, 256, 512, and 512 channels, respectively, to promote efficient capture of spatial-spectral information.

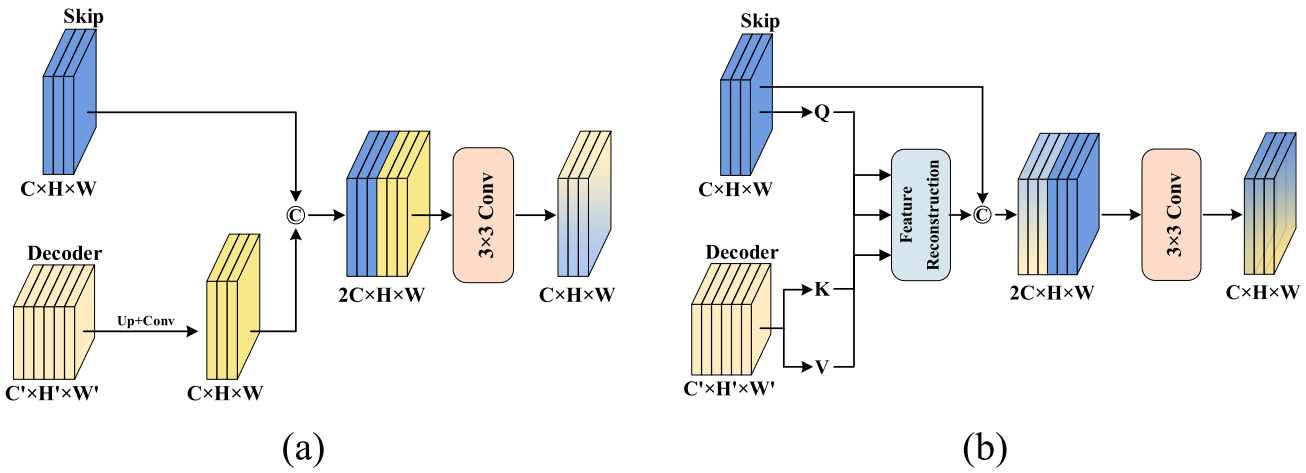


Fig. 7. The structure of (a) the CNN-based decoder, and (b) the Transformer-based decoder.

4.2.2. Decoder blocks

Similar to the encoder blocks, the decoder blocks also comprise two CNN layers and two Transformer layers. During the deep feature fusion stage, as shown in Fig. 7, two FRMs are employed. Due to the low spatial resolution yet high channel characteristics of deep features, the use of a Transformer-based FRM can better fuse long-distance and global features. During the shallow feature fusion stage, the input features are characterized by a high spatial resolution, which enables the effective extraction of local features via convolutional modules, resulting in optimized spatial details for the task of land-cover classification.

4.3. Multi-level spatial-spectral feature fusion

In the proposed PatchOut framework, the encoder extracts spatial-spectral features at multiple scales. Since the different channels may capture distinct semantic patterns, it is crucial to effectively fuse these features, bridging potential semantic gaps, rather than relying on concatenation or addition. To solve the feature fusion problem between the encoder and decoder module, inspired by UCTransNet (Wang et al., 2022a), we propose the MSSSFF module, which incorporates Transformer mechanisms to effectively model long-range dependencies across different encoder stages. Specifically, in Fig. 4, the MSSSFF module comprises three steps: a multi-scale feature embedding block (MEB), a spatial-spectral feature fusion (SSFF) block, and a spatial-spectral reconstruction (SSR) block.

During the encoder stages, four different-level skip connection features are obtained as $F_i \in \mathscr{R}^{C_i \times \frac{H}{2^i} \times \frac{W}{2^i}}, (i = 1, 2, 3, 4)$. Firstly, for multi-scale feature embedding, DWConv is used to tokenize the features at four different levels to have the same spatial size. In this study, we set the feature size as 16×16 after tokenization. Subsequently, PWConv unifies the channel dimension of the four features to 128 channels each. The tokens of the four layers are then concatenated to form a multi-level fusion feature:

$$E_i = PWConv(DWConv(F_i)), E_i \in \mathscr{R}^{c \times h \times w} \tag{13}$$

$$E = Concat(E_1, E_2, E_3, E_4) \tag{14}$$

For the multi-level SSFF, MHSA and ConvFF modules that are essentially consistent with those utilized in the RTB are utilized. Regarding the Q, K, and V components, they are projected from the same

Fig. 8. Distribution map of training samples (within the red boxes). (a) Qingpu HSI dataset. (b) Matiwan HSI dataset.

Table 1
Sample distribution across the land-cover classes in the Qingpu and Matiwan datasets.

Qingpu dataset				Matiwan dataset			
No.	Classes	Train + Val	Testing	No.	Classes	Train + Val	Testing
C1	Asphalt road	36,545	363,112	**C1**	Acer negundo	67,250	158,397
C2	Greenhouse	62,687	297,340	**C2**	Salix	65,277	115,489
C3	Cement road	34,864	368,230	**C3**	Elm	7688	7665
C4	Lotus	114,804	1,363,364	**C4**	Oryza sativa	67,988	384,156
C5	Water bamboo 1	192,083	1,570,698	**C5**	Sophora japonica	133,010	342,581
C6	Farmland	322,359	7,033,623	**C6**	Fraxinus chinensis	59,808	109,534
C7	Water bamboo 2	172,562	1,907,426	**C7**	Goldenrain tree	6653	16,651
C8	Water	524,298	14,465,019	**C8**	Water	60,584	105,063
C9	Buildings	244,680	1,615,848	**C9**	Bare land	15,336	23,073
C10	Bulrush	23,963	78,620	**C10**	Post-harvest field	71,257	122,573
C11	Oak	19,399	9240	**C11**	Stubble	4224	1388
C12	Ligustrum	15,942	18,298	**C12**	Zea mays	41,078	18,087
C13	Elaeocarpus sylvestris	41,730	19,546	**C13**	Pyrus	332,121	694,392
C14	Camptotheca acuminata	23,470	2296	**C14**	Soybean	3975	3176
C15	Soapberry	49,966	33,202	**C15**	Populus	45,917	45,155
C16	Salix	7649	21,300	**C16**	Vegetable field	19,071	10,077
C17	Goldenrain tree	34,385	20,218	**C17**	Sparsewood	588	908
C18	Cedar	39,395	83,672	**C18**	Weeds	119,845	301,945
C19	Camphor tree	124,765	381,818	**C19**	Amygdalus	38,317	27,197
C20	Zelkova schneideriana	2315	8340	**C20**	Buildings	9584	20,032

source E. The difference is that the tokenization embedding has been performed, rendering the reduced key and value components redundant. Consequently, the reduction ratio is set to 1. The multi-level SSFF is repeated four times in order to enforce the deep fusion of multi-level features.

After this, the fused feature $E^{'}$ is split into four groups according to the concatenation order. For each fused feature $F_i^{'} \in \mathscr{R}^{c\times h\times w}$, it is restored to its original size using nearest neighbor interpolation, followed by a DWConv layer, and adaptively added with the original skip connections:

$$F_i^{'} = DWConv(split(E^{'})) \tag{15}$$

$$O_i = \alpha F_i^{'} + (1-\alpha)F_i \tag{16}$$

where O_i is the final skip connection after the multi-level SSFF, and $\alpha \in (0,1)$ denotes the adaptive weight to balance the original and fused skip connections.

5. Experiments and analysis

5.1. Training data preprocessing

As shown in Fig. 8 and Table 1, to mitigate the impact of sparse and imbalanced labeled samples on the accuracy of the semantic segmentation model, and to prevent test set leakage, firstly, regions with more concentrated categories were cropped for training and validation, while the remaining regions formed the test set. For the Qingpu HSI dataset, sixteen images of size 512 × 512 were cropped for the training and validation, representing 6.96 % of the total area. For the Matiwan HSI dataset, a total of 22 images of size 256 × 256 were applied, representing 24.33 % of the total area.

Then, a global stochastic stratified sampling strategy was adopted in this experiment (Zheng et al., 2020), which is a form of data augmentation that also helps to address the issue of class imbalance (Wang et al., 2022b; Zhu et al., 2022). In detail, to avoid the long-tail distribution of ground types and ensure balanced training, 5000 and 1000 samples per land-cover class were extracted from cropped images for the Qingpu-HSI and Matiwan datasets, respectively. The remaining labeled pixels within the selected cropped regions served as the validation set. Besides, a minibatch sampler was also applied during the training process. When evaluating model performance, the cropped areas for training and validation are excluded to ensure that pixels used for training and validation do not participate in accuracy assessment. Besides, for the patch-based models used in the comparative experiments, HSI patches of size 9 × 9, centered on each training and validation pixel, were extracted to construct the respective datasets. These HSI patches were sampled from 16 or 22 distinct cropped regions, ensuring independence from the test set.

5.2. Experimental setup

Given the limited available HSI training data, data augmentation procedures, including random rotation, cropping, and flipping, were implemented to augment the training set and improve model generalization performance. The number of training epochs was set to 100, and the batch size was set to 4 for the Qingpu dataset and 8 for the Matiwan dataset. For the proposed PatchOut model, the learning rate was set to 0.001, and a stochastic gradient descent (SGD) optimizer was employed to update the training parameters. Cross-entropy served as the loss function. To facilitate a fair comparison, the experimental setups for the other comparison models were replicated as reported in the original publications. During the inference, we employed a sliding window approach with 50 % overlap between adjacent tiles to mitigate the edge deterioration prevalent in semantic segmentation tasks (Sun et al., 2019).

All the experiments were performed entirely on an NVIDIA GeForce RTX 4090 GPU. The overall accuracy (OA), Kappa, mean intersection over union (mIoU), and frequency weighted intersection over union (FWIoU) were selected for the accuracy evaluation of all the classification models.

5.3. Comparison algorithms

To evaluate the HSI classification performance of the proposed PatchOut framework, the following two aspects of comparison deep learning methods were considered. The patch-based methods were SSRN (Zhong et al., 2018), SpectralFormer (Hong et al., 2022), and PASSNet (Ji et al., 2023), for which the input patch size was set to 9. The patch-free methods were FPGA (Zheng et al., 2020), ABCNet (Li et al., 2021), Swin-Unet (Cao et al., 2023), and ConvNext-V2 (Woo et al., 2023). The parameters of the comparison methods were as consistent as possible with the original articles.

Table 2
Classification accuracies of the different methods on the Qingpu HSI dataset.

Class	SSRN	SpectralFormer	PASSNet	FPGA	ABCNet	Swin-Unet	ConvNext-V2	PatchOut
C1	71.03 ± 4.57	71.76 ± 0.71	71.38 ± 7.73	68.54 ± 9.94	70.09 ± 6.95	56.43 ± 6.00	67.06 ± 2.43	**72.18 ± 4.79**
C2	89.89 ± 1.86	89.05 ± 0.48	90.48 ± 3.52	93.59 ± 1.14	91.51 ± 1.77	84.70 ± 2.36	87.10 ± 2.25	**95.54 ± 0.86**
C3	95.33 ± 1.08	95.58 ± 0.57	**95.60 ± 0.83**	94.59 ± 0.66	89.06 ± 4.56	85.22 ± 2.31	91.68 ± 1.01	94.17 ± 1.58
C4	95.55 ± 0.78	96.29 ± 0.02	97.53 ± 0.67	85.23 ± 4.36	96.95 ± 0.57	83.37 ± 3.59	97.96 ± 0.69	**98.54 ± 0.58**
C5	87.28 ± 2.10	88.91 ± 0.88	92.17 ± 2.05	87.79 ± 1.59	**95.49 ± 0.55**	91.79 ± 0.97	94.63 ± 0.53	95.22 ± 0.47
C6	87.93 ± 2.32	93.87 ± 0.12	95.15 ± 0.95	88.68 ± 2.78	90.36 ± 2.91	96.48 ± 0.76	96.48 ± 0.86	**97.86 ± 0.22**
C7	94.69 ± 1.23	92.16 ± 1.35	92.30 ± 0.76	91.62 ± 1.79	89.37 ± 3.44	84.40 ± 2.32	82.28 ± 1.60	**96.54 ± 0.25**
C8	97.56 ± 0.84	98.16 ± 0.11	98.28 ± 0.31	94.73 ± 0.93	91.39 ± 6.14	98.31 ± 0.40	98.58 ± 0.29	**98.58 ± 0.15**
C9	93.15 ± 0.75	89.66 ± 0.42	**95.90 ± 0.62**	90.52 ± 2.14	88.33 ± 1.62	88.30 ± 3.21	92.49 ± 1.71	94.30 ± 0.78
C10	64.53 ± 1.19	68.86 ± 0.01	72.32 ± 2.86	58.95 ± 0.59	**77.62 ± 10.85**	63.71 ± 1.87	74.00 ± 3.70	64.97 ± 1.92
C11	97.00 ± 0.58	95.64 ± 0.36	98.19 ± 0.60	**100.00 ± 0.00**	100.00 ± 0.01	93.92 ± 4.19	93.71 ± 6.41	99.26 ± 0.87
C12	22.29 ± 1.62	18.35 ± 0.33	22.96 ± 1.36	16.77 ± 1.23	**26.41 ± 1.20**	17.49 ± 1.30	17.62 ± 3.48	23.85 ± 3.93
C13	87.29 ± 1.99	78.31 ± 1.55	81.67 ± 0.17	76.27 ± 6.49	80.78 ± 0.84	40.25 ± 8.12	61.27 ± 10.38	**87.65 ± 4.73**
C14	**96.33 ± 2.29**	86.73 ± 8.24	79.55 ± 6.64	79.49 ± 20.19	88.01 ± 5.57	84.31 ± 7.44	67.74 ± 9.98	88.12 ± 9.26
C15	64.42 ± 3.04	60.76 ± 0.20	**64.75 ± 3.01**	48.85 ± 7.85	42.20 ± 14.17	52.76 ± 3.80	54.15 ± 4.72	46.84 ± 11.78
C16	83.95 ± 6.03	82.04 ± 3.50	71.11 ± 7.76	53.47 ± 9.84	49.15 ± 21.23	**84.39 ± 9.16**	33.94 ± 7.20	36.99 ± 6.43
C17	82.72 ± 4.98	81.44 ± 1.56	**86.82 ± 1.35**	48.13 ± 0.85	48.23 ± 4.56	76.80 ± 5.74	74.27 ± 13.37	51.39 ± 4.98
C18	56.32 ± 2.85	53.07 ± 0.71	**69.68 ± 7.21**	41.45 ± 3.29	40.95 ± 5.26	43.66 ± 1.65	45.42 ± 5.38	54.55 ± 4.35
C19	**86.43 ± 0.55**	73.16 ± 1.28	85.11 ± 2.09	57.17 ± 6.77	73.30 ± 7.21	63.25 ± 1.69	73.84 ± 5.35	81.38 ± 2.66
C20	6.75 ± 10.82	0.64 ± 0.53	0.94 ± 0.54	59.41 ± 12.69	38.27 ± 24.25	20.21 ± 18.39	22.29 ± 8.71	**62.79 ± 6.52**
OA	93.30 ± 0.83	94.59 ± 0.02	95.76 ± 0.30	90.82 ± 0.43	90.45 ± 2.78	93.71 ± 0.24	95.13 ± 0.31	**96.82 ± 0.08**
Kappa	0.905 ± 0.011	0.923 ± 0.000	0.939 ± 0.004	0.871 ± 0.006	0.867 ± 0.036	0.910 ± 0.003	0.930 ± 0.004	**0.954 ± 0.001**
mIoU	0.538 ± 0.003	0.515 ± 0.005	0.583 ± 0.009	0.466 ± 0.008	0.483 ± 0.024	0.448 ± 0.007	0.511 ± 0.005	**0.596 ± 0.007**
FWIoU	0.890 ± 0.013	0.910 ± 0.001	0.926 ± 0.005	0.860 ± 0.006	0.850 ± 0.036	0.904 ± 0.004	0.919 ± 0.004	**0.945 ± 0.001**

5.4. Classification results

5.4.1. Results for the Qingpu HSI dataset

For the Qingpu HSI dataset, the quantitative evaluation of the PatchOut framework and other comparison algorithms are presented in Table 2, wherein the highest accuracy and second-highest accuracy is indicated in bold and underline, respectively. The OA measures the global pixelwise classification, while the mIoU assesses the segmentation quality across all classes. The experimental results indicate that our PatchOut framework attained superior performance, attaining an OA of 96.82 %, Kappa of 0.954, mIoU of 0.596, and FWIoU of 0.945. In detail, compared with PASSNet, which was the best performing patch-based method, the OA of PatchOut is improved by 1.11 %, while the mIoU and FWIoU of PatchOut are improved by 2.23 % and 2.05 %, respectively. Compared with the other patch-free models, the proposed PatchOut framework exhibits a significant performance improvement. Quantitatively, it demonstrates improvements of 1.78 %, 16.63 %, and 2.83 % in OA, mIoU and FWIoU, respectively, when compared to ConvNext-V2. Overall, the proposed PatchOut framework performs well in the fine-grained land-cover classification task, especially for artificial surfaces and crop types.

Fig. 9 and Fig. 10 illustrate the overall and local classification outcomes generated by the different methods for several representative areas in the Qingpu dataset, enabling a visual comparison of the model performance across diverse land-cover types. As shown in Fig. 9, the proposed PatchOut framework improved visual quality, exhibiting minimal salt-and-pepper noise, preserving the most complete internal structures, and delineating clear boundaries, particularly for cropland and water classifications. From the classification results of the local magnification (Fig. 10), for instance, from the first and second rows, PatchOut could accurately distinguish the two different types of water bamboo, which are in different growth cycles. From the third row, some models (e.g., Swin-Unet and ConvNext-V2) fail to accurately identify camphor tree, which is the main tree species type in the Qingpu area. In addition, the proposed PatchOut framework can extract the farmland type most completely. From the fourth and fifth rows, when faced with fine tree species classification, despite the relatively comparable classification accuracies between the patch-based and patch-free methods, the visual quality of the patch-based results is notably inferior, demonstrating poor spatial coherence and continuity in the classification output. The final row depicts the classification map of residential area and its vicinity, with PatchOut exhibiting better integrity rather than fragmentation.

5.4.2. Results for the Matiwan HSI dataset

Table 3 reports the classification results of all compared methods for the Matiwan HSI dataset. Comprehensively, the PatchOut framework again achieves a superior performance, outperforming the other approaches. The quantitative evaluation demonstrates that the proposed PatchOut framework obtains a superior performance, yielding an OA of 89.96 %, Kappa of 0.883, mIoU of 0.704, and FWIoU of 0.834. PASSNet and SSRN, which belong to the patch-based methods, achieve the second-highest classification accuracy. The precision for the "Sparse-wood" (C17) category in most of the patch-free methods is lower than that for the patch-based methods, which can be attributed to the extreme scarcity of samples and their confinement to a single training image. These data limitations significantly constrain the performance of the patch-free semantic segmentation models.

Fig. 11 presents the visualized classification maps generated by the various models on the complete Matiwan HSI dataset. Clearly, the patch-based models exhibit salt-and-pepper noise in their classification maps, particularly noticeable in the higher-resolution Matiwan dataset due to their pixelwise prediction mode. In contrast, the patch-free models, especially the proposed PatchOut, generate smoother and visually more appealing results. Furthermore, due to the inherent sparsity and varied growth conditions of certain classes within the Matiwan dataset, FPGA, which is the most lightweight method among the patch-free architectures, exhibits notable limitations in its feature extraction and generalization capabilities. ABCNet, incorporating a bilateral structure, is noted for exhibiting a serrated appearance at the edges of objects. Swin-Unet, featuring a sliding Transformer window, tends to manifest a checkerboard pattern in its classification results. The ConvNext-V2 model demonstrates a generally favorable classification performance; however, it encounters occasional misclassification characterized by sparse voids in specific regions, e.g., *Sophora japonica*, due to the limited receptive field inherent to CNNs. The proposed PatchOut framework, leveraging the combined strengths of the CNN's local feature extraction capability and the Transformer's ability to capture long-range dependencies, achieves the best visual performance. It effectively mitigates issues such as voids and checkerboard effects, thereby enhancing the overall classification accuracy and visual coherence.

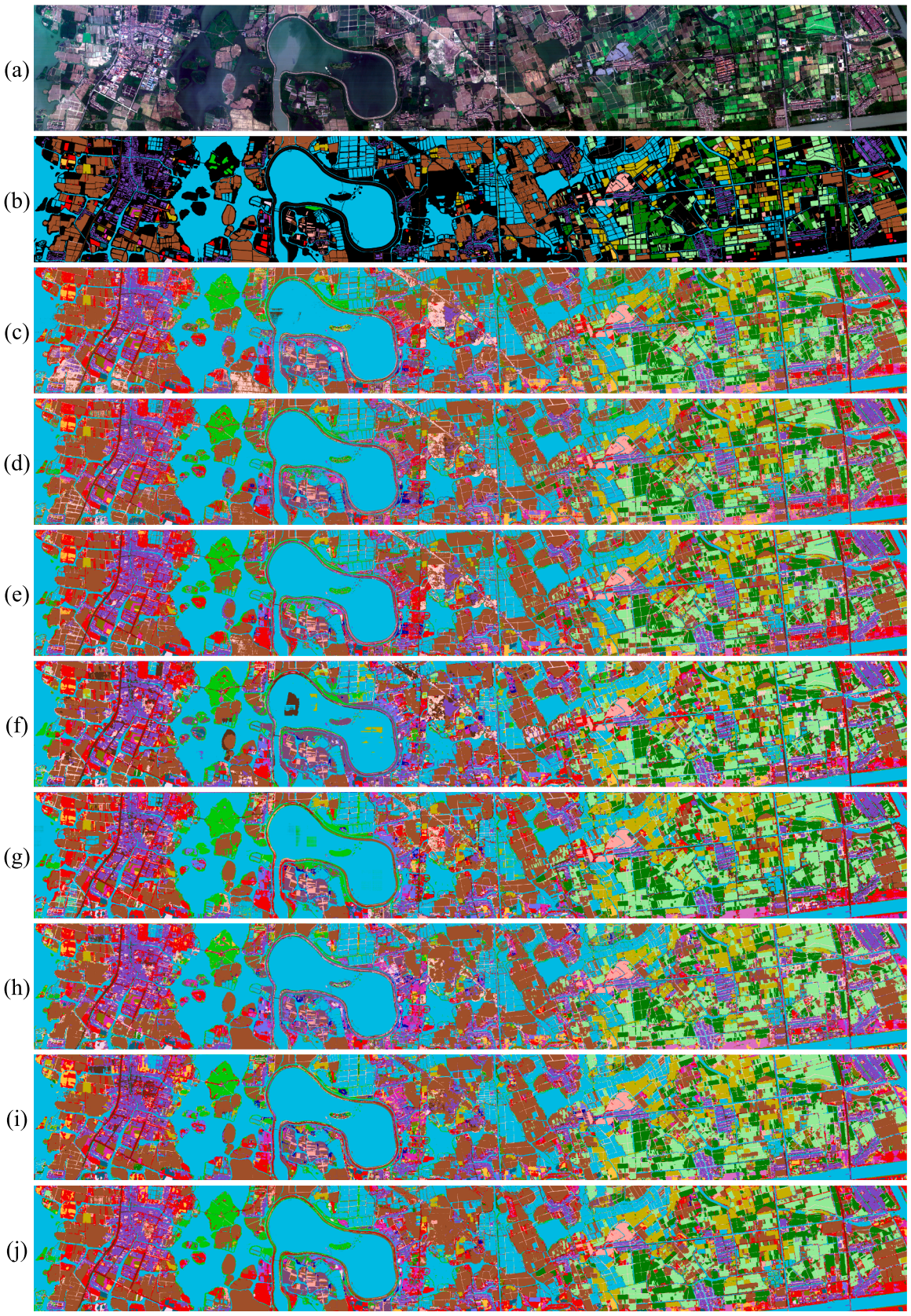


Fig. 9. Classification maps obtained by the different methods on the Qingpu HSI dataset. (a) True-color images. (b) Labels. Patch-based methods: (c) SSRN. (d) SpectralFormer. (e) PASSNet. Patch-free methods: (f) FPGA. (g) ABCNet. (h) Swin-Unet. (i) ConvNext-V2. (j) PatchOut.

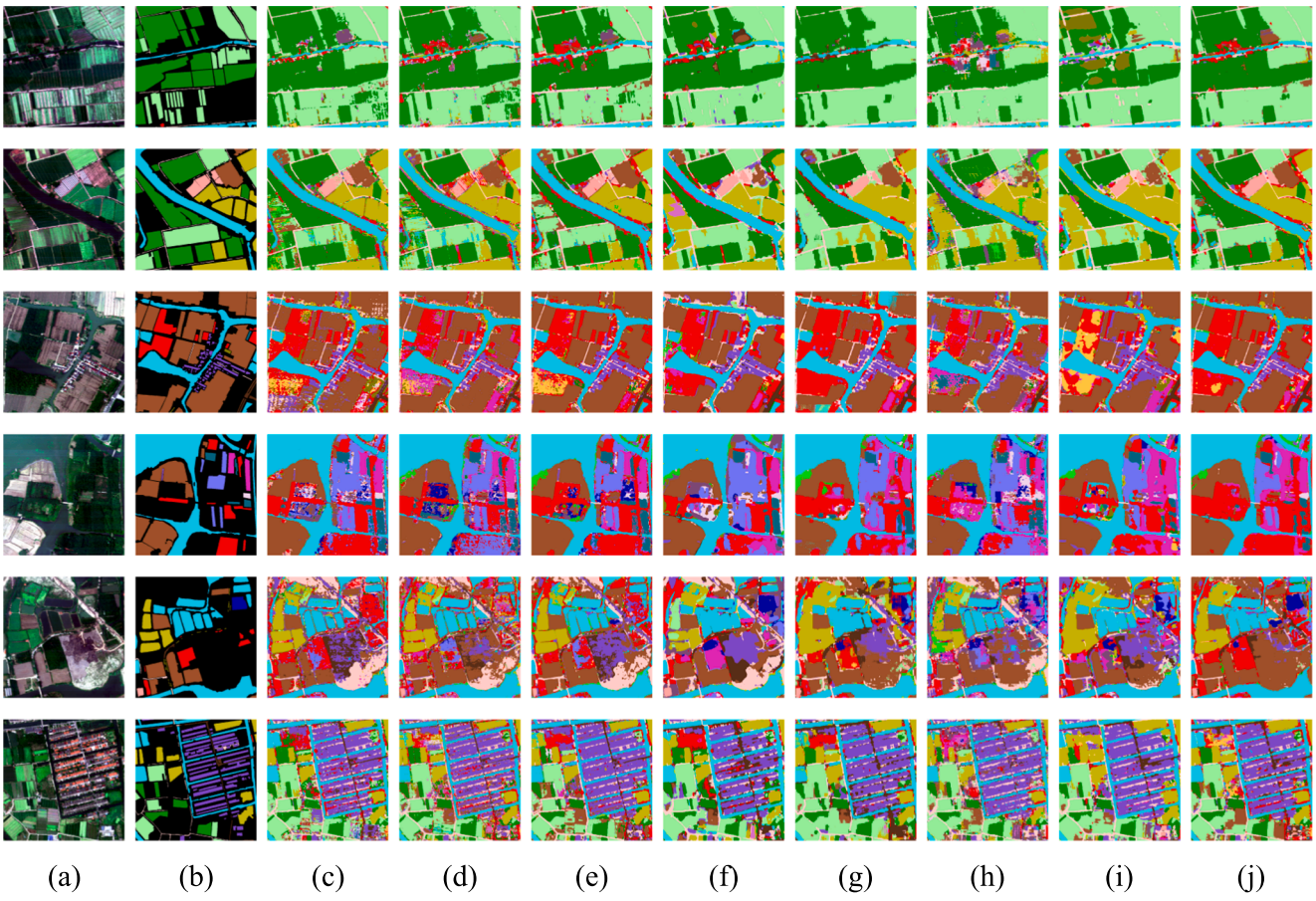


Fig. 10. Local classification maps obtained by the different methods on the Qingpu HSI dataset. (a) True-color images. (b) Labels. Patch-based methods: (c) SSRN. (d) SpectralFormer. (e) PASSNet. Patch-free methods: (f) FPGA. (g) ABCNet. (h) Swin-Unet. (i) ConvNext-V2. (j) PatchOut.

Table 3
Classification accuracies of the different methods on the Matiwan HSI dataset.

Class	SSRN	SpectralFormer	PASSNet	FPGA	ABCNet	Swin-Unet	ConvNext-V2	PatchOut
C1	83.59 ± 1.34	77.82 ± 1.58	85.90 ± 2.88	59.31 ± 6.47	71.01 ± 10.60	72.78 ± 7.39	80.03 ± 5.19	**87.05 ± 0.55**
C2	93.11 ± 0.81	90.09 ± 1.34	93.16 ± 1.10	89.20 ± 3.86	84.25 ± 0.42	93.88 ± 1.98	92.69 ± 6.12	**98.18 ± 0.50**
C3	95.63 ± 1.36	92.91 ± 0.67	**95.98 ± 3.05**	64.79 ± 6.02	18.88 ± 4.09	80.36 ± 30.19	95.80 ± 5.20	84.72 ± 12.54
C4	95.43 ± 0.34	95.27 ± 0.36	94.84 ± 0.52	88.56 ± 4.29	90.30 ± 7.67	96.35 ± 0.75	**96.59 ± 0.75**	96.23 ± 1.26
C5	84.58 ± 3.18	77.53 ± 2.57	85.00 ± 1.10	57.01 ± 6.37	80.01 ± 3.13	71.44 ± 3.06	72.15 ± 8.26	**88.03 ± 2.19**
C6	**94.65 ± 0.70**	93.82 ± 0.56	88.98 ± 4.48	74.59 ± 6.75	77.30 ± 3.81	91.94 ± 2.44	93.79 ± 3.03	91.94 ± 2.41
C7	92.96 ± 2.09	76.99 ± 4.30	79.25 ± 11.76	51.76 ± 11.16	80.39 ± 5.29	70.57 ± 14.35	90.48 ± 1.50	**93.75 ± 4.83**
C8	86.38 ± 1.81	86.85 ± 2.79	89.52 ± 1.98	73.54 ± 2.16	**93.66 ± 2.81**	84.84 ± 3.90	86.34 ± 4.51	92.83 ± 2.11
C9	89.09 ± 3.19	92.14 ± 0.97	91.90 ± 2.83	54.67 ± 21.48	89.70 ± 4.66	89.02 ± 3.39	80.10 ± 3.27	**99.15 ± 0.48**
C10	98.52 ± 0.43	97.94 ± 0.25	**99.57 ± 0.11**	93.29 ± 3.79	91.04 ± 3.94	82.41 ± 2.43	89.38 ± 7.14	99.48 ± 0.48
C11	82.74 ± 6.67	82.36 ± 2.34	81.54 ± 3.96	70.09 ± 7.03	64.93 ± 9.32	67.78 ± 9.07	89.12 ± 4.95	**89.51 ± 2.19**
C12	86.60 ± 2.11	91.36 ± 0.78	90.91 ± 1.29	84.89 ± 0.64	90.19 ± 1.45	88.04 ± 1.56	93.10 ± 1.97	**95.07 ± 0.45**
C13	87.18 ± 1.63	74.72 ± 1.12	85.77 ± 1.67	85.96 ± 2.44	82.05 ± 2.55	78.31 ± 2.61	**92.37 ± 1.44**	88.68 ± 0.73
C14	36.78 ± 5.58	36.12 ± 4.83	48.22 ± 7.04	58.43 ± 20.45	**62.40 ± 5.59**	20.08 ± 2.31	24.78 ± 10.29	51.33 ± 9.41
C15	79.60 ± 4.05	79.26 ± 2.79	87.19 ± 2.10	74.68 ± 2.23	**89.01 ± 2.45**	77.03 ± 3.39	80.33 ± 1.45	88.83 ± 1.81
C16	79.25 ± 2.91	**80.33 ± 2.13**	77.86 ± 3.02	77.40 ± 4.41	72.35 ± 15.86	61.20 ± 14.49	70.34 ± 9.88	77.86 ± 7.84
C17	57.56 ± 15.31	66.85 ± 3.85	**77.71 ± 6.53**	2.38 ± 2.44	0.00 ± 0.00	72.05 ± 30.99	47.22 ± 13.82	21.01 ± 10.78
C18	83.93 ± 2.14	70.49 ± 1.68	**85.25 ± 1.84**	66.81 ± 3.48	70.59 ± 5.67	68.65 ± 4.73	85.25 ± 5.25	79.52 ± 4.53
C19	88.98 ± 4.55	79.71 ± 1.64	89.92 ± 2.61	52.60 ± 17.42	81.43 ± 2.98	69.90 ± 2.47	74.63 ± 2.82	**98.15 ± 0.90**
C20	74.20 ± 8.16	74.54 ± 0.88	81.07 ± 3.95	76.23 ± 4.32	80.41 ± 7.62	67.55 ± 10.95	71.38 ± 10.37	**86.37 ± 2.00**
OA	88.34 ± 0.34	81.58 ± 0.38	88.34 ± 0.18	76.57 ± 1.18	81.70 ± 2.06	80.18 ± 1.14	87.42 ± 1.64	**89.96 ± 0.36**
Kappa	0.865 ± 0.004	0.788 ± 0.005	0.865 ± 0.002	0.726 ± 0.014	0.787 ± 0.024	0.771 ± 0.013	0.853 ± 0.019	**0.883 ± 0.004**
mIoU	0.620 ± 0.005	0.564 ± 0.009	0.645 ± 0.008	0.486 ± 0.007	0.569 ± 0.019	0.530 ± 0.032	0.669 ± 0.033	**0.704 ± 0.021**
FWIoU	0.814 ± 0.005	0.728 ± 0.006	0.810 ± 0.003	0.648 ± 0.016	0.712 ± 0.030	0.701 ± 0.013	0.789 ± 0.021	**0.834 ± 0.005**

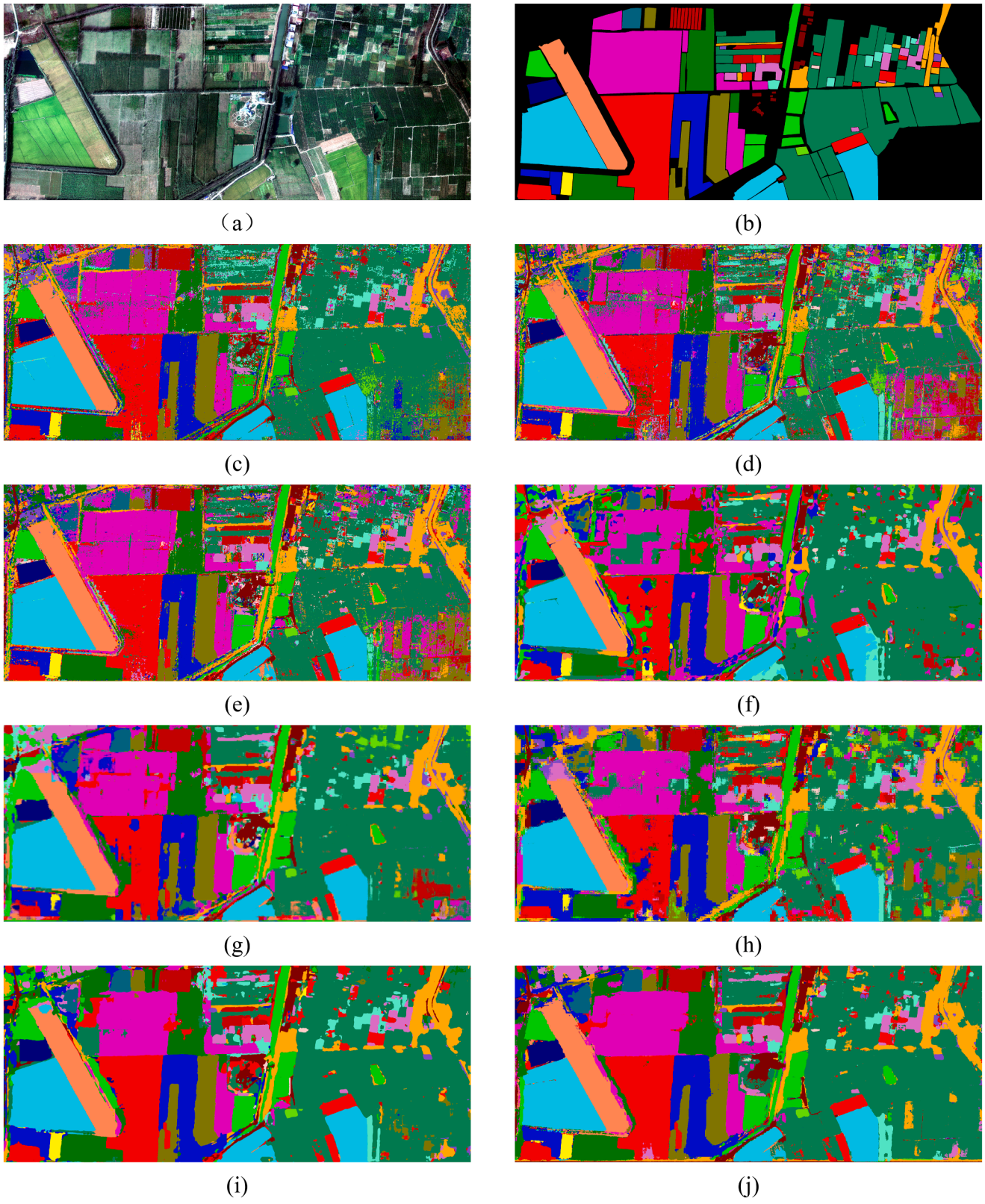


Fig. 11. Classification maps obtained by the different methods on the Matiwan HSI dataset. Patch-based methods: (a) True-color images. (b) Labels. Patch-based methods: (c) SSRN. (d) SpectralFormer. (e) PASSNet. Patch-free methods: (f) FPGA. (g) ABCNet. (h) Swin-Unet. (i) ConvNext-V2. (j) PatchOut.

Table 4
Total trainable parameters, FLOPs, and prediction times for the different methods on the Qingpu and Matiwan datasets.

Dataset		SSRN	SpectralFormer	PASSNet	FPGA	ABCNet	Swin-Unet	ConvNext-V2	PatchOut
Qingpu	Params	0.92 M	0.49 M	0.70 M	2.84 M	15.18 M	27.53 M	11.16 M	30.65 M
	FLOPs	269.06 M	50.87 M	56.84 M	165.67 G	118.51 G	37.62 G	58.032 G	275.87 G
	Inference time (s)	4284.07	9679.69	2010.76	157.92	147.88	151.42	92.4	135.22
Matiwan	Params	0.93 M	0.50 M	0.72 M	2.84 M	15.21 M	27.54 M	11.17 M	30.65 M
	FLOPs	273.41 M	52.37 M	58.35 M	41.70 G	30.14 G	9.44 G	14.52 G	69.17 G
	Inference time (s)	432.78	696.39	183.15	17.81	17.69	27.54	20.56	30.22

Table 5
Ablation experiments on the Qingpu and Matiwan datasets.

Method			Qingpu dataset				Matiwan dataset			
RTB	MSSSFF	FRM	OA	Kappa	mIoU	FWIoU	OA	Kappa	mIoU	FWIoU
ViT	×	×	95.93	0.942	0.554	0.930	87.60	0.856	0.647	0.803
√	×	×	96.11	0.944	0.545	0.935	87.95	0.860	0.657	0.812
√	√	×	96.57	0.951	0.569	0.941	89.46	0.878	0.676	0.831
√	×	√	96.53	0.950	0.578	0.939	89.19	0.875	0.680	0.823
√	√	√	**96.82**	**0.954**	**0.596**	**0.945**	**89.96**	**0.883**	**0.704**	**0.834**
ViT	√	ViT	96.47	0.949	0.571	0.939	89.44	0.877	0.687	0.826

6. Discussion

6.1. Time consumption comparison

The analysis of the computational efficiency is shown in Table 4, focusing on three key metrics: model parameter count, computational complexity (FLOPs), and inference time. Firstly, the patch-free methods, used for semantic segmentation, exhibit larger parameter counts and FLOPs compared to the patch-based approaches, which are typically used for classification tasks, primarily attributed to the substantial differences in input dimensions. The patch-free models can process and predict entire images of 512 × 512 or 256 × 256 pixels in a single forward pass, while the patch-based methods operate on smaller 9 × 9 patches, generating predictions for only the central pixel of each patch. Consequently, in terms of inference time, despite the implementation of GPU parallelization optimization to maximize the GPU utilization for the patch-based methods, the patch-free approaches demonstrate significantly faster performances. The disparity in inference time becomes even more pronounced when processing larger HSIs. For the Qingpu HSI dataset, the proposed framework requires only 135.22 s, including data loading. In contrast, the fastest patch-based method—PASSNet—exceeds 2000 s. While achieving comparable or superior accuracy to patch-based methods, the proposed patch-free model offers substantial and noteworthy improvements in computation efficiency.

6.2. Ablation study

Several ablation experiments were conducted to evaluate the performance brought by the proposed RTB, MSSSFF and FRM modules. The baseline model employed a hybrid encoder architecture comprising a standard Vision Transformer and CNN, while utilizing solely CNN upsampling in the decoder phase. The proposed modules were then incrementally incorporated into the baseline. Additionally, in order to evaluate the performance of proposed RTB, it was removed from the PatchOut framework, replacing it with a standard Vision Transformer block. As can be seen in Table 5, the ablation study results demonstrate that the inclusion of these modules effectively enhances the performance of the proposed PatchOut framework. Specifically, especially for the mIoU and FWIoU indicators, the integration of the three proposed modules leverages the Transformer's capacity for long-range feature extraction and the exploitation of adjacent pixels, resulting in more comprehensive classification outcomes and reduced noise.

Table 6
Classification accuracies of the different methods on the Qingpu and Matiwan dataset under different image size and batch size.

Dataset	Image Size	FPGA	ABCNet	Swin-Unet	ConvNext-V2	PatchOut
Qingpu	I256B4	91.89 ± 1.67	89.86 ± 1.80	93.92 ± 0.32	93.97 ± 0.67	95.15 ± 0.43
	I256B8	92.61 ± 0.36	90.55 ± 1.72	94.20 ± 0.41	94.28 ± 0.79	96.33 ± 0.16
	I256B16	92.86 ± 0.13	91.78 ± 1.22	93.94 ± 0.40	94.63 ± 0.17	96.76 ± 0.08
	I512B4	90.82 ± 0.43	90.45 ± 2.78	93.71 ± 0.24	95.13 ± 0.31	96.82 ± 0.08
Matiwan	I128B8	77.48 ± 0.56	80.90 ± 1.48	80.87 ± 0.69	85.99 ± 2.15	87.47 ± 0.89
	I128B16	77.12 ± 2.50	82.81 ± 1.12	81.24 ± 0.79	87.43 ± 0.71	87.53 ± 0.96
	I128B32	77.74 ± 0.59	84.58 ± 0.66	80.93 ± 0.95	86.68 ± 0.37	88.69 ± 0.91
	I256B8	76.57 ± 1.18	81.70 ± 2.06	80.18 ± 1.14	87.42 ± 1.64	89.96 ± 0.36

Notes: *I* means image size, *B* means batch size. I256B4 means the image size is set as 256 and the batch size is set as 4.

6.3. Sensitivity analysis on image size and batch size

To evaluate the potential impact of image size and batch size settings on model performance, further experiments were implemented. The original training datasets were divided into four subsets for training. Specifically, for the Qingpu dataset, 256 × 256 patches were used with batch sizes of 4, 8, and 16. For the Matiwan dataset, 128 × 128 patches were used with batch sizes of 8, 16, and 32. As shown in Table 6, under different combinations of image size and batch size, our proposed PatchOut model consistently maintains the best overall accuracy performance. Besides, the results indicate that smaller training sample sizes (implying fewer land cover types per image) generally require larger batch sizes for most models to maximize type diversity within each iteration.

6.4. Sensitivity analysis on overlap percentage

While overlapping inference is widely adopted in large-scale image semantic segmentation for its benefits, it inherently leads to redundant computations. Thus, we reduced the percentage of overlap to 33 % or 25 % for further experiments. As shown in Table 7, the model performance

Table 7
Classification accuracies of the patch-free methods on the Qingpu and Matiwan dataset using different overlap ratios.

Dataset	Overlap ratio	Accuracy	FPGA	ABCNet	Swin-Unet	ConvNext-V2	PatchOut
Qingpu	25 %	OA	91.03 ± 0.36	90.89 ± 2.74	93.73 ± 0.24	95.31 ± 0.28	96.82 ± 0.10
		mIoU	0.472 ± 0.009	0.491 ± 0.025	0.449 ± 0.007	0.515 ± 0.006	0.595 ± 0.014
	33 %	OA	90.90 ± 0.41	91.15 ± 2.78	94.56 ± 0.23	95.37 ± 0.31	96.83 ± 0.09
		mIoU	0.471 ± 0.007	0.499 ± 0.026	0.469 ± 0.008	0.515 ± 0.008	0.596 ± 0.013
	50 %	OA	90.82 ± 0.43	90.45 ± 2.78	93.71 ± 0.24	95.13 ± 0.31	96.82 ± 0.08
		mIoU	0.466 ± 0.008	0.483 ± 0.024	0.448 ± 0.007	0.511 ± 0.005	0.596 ± 0.007
Matiwan	25 %	OA	76.66 ± 1.04	82.02 ± 1.48	80.37 ± 1.11	88.44 ± 1.58	90.34 ± 0.32
		mIoU	0.490 ± 0.014	0.568 ± 0.016	0.535 ± 0.033	0.697 ± 0.034	0.718 ± 0.017
	33 %	OA	76.69 ± 1.03	83.03 ± 1.29	82.43 ± 1.13	87.96 ± 1.78	90.26 ± 0.30
		mIoU	0.482 ± 0.014	0.576 ± 0.014	0.570 ± 0.031	0.686 ± 0.039	0.710 ± 0.016
	50 %	OA	76.57 ± 1.18	81.70 ± 2.06	80.18 ± 1.14	87.42 ± 1.64	89.96 ± 0.36
		mIoU	0.486 ± 0.007	0.569 ± 0.019	0.530 ± 0.032	0.669 ± 0.033	0.704 ± 0.021

varied across different overlap ratios. We posit that this variation may be attributed to the extent of edge degradation inherent in each model. Our proposed PatchOut model achieved consistent accuracy across varying overlap ratios. This performance stability suggests that the hybrid Transformer-CNN architecture effectively integrates the local receptive fields of CNNs and the global receptive fields of Transformers, thus mitigating the impact of edge effects on accuracy.

7. Conclusion

In this paper, we propose a novel patch-free framework—PatchOut—specifically designed for the task of HSI semantic segmentation. This framework integrates Transformer and CNN architectures, leveraging their combined capacity to learn both global and local representations, thereby facilitating a more comprehensive understanding of the spectral-spatial characteristics of HSI. By synergistically leveraging the respective strengths of the CNN and Transformer, the unified framework facilitates the simultaneous modeling of high-resolution local features and low-resolution global representations, thus enabling the extraction of multi-level spatial and spectral characteristics. The proposed MSSSFF module leverages the Transformer architecture to aggregate and extract multi-scale features, thereby facilitating a more comprehensive understanding of the spectral-spatial characteristics inherent in HSIs. In addition, the proposed FRM enhances the low-resolution features in the decoder by integrating them with high-resolution features from the encoder, enabling effective feature restoration. Moreover, we have described how a large-scale HSI dataset covering Qingpu District, Shanghai, China, was constructed for fine land-cover classification. Empirical evaluations demonstrate that the proposed PatchOut framework, as a semantic segmentation approach, can effectively extract spectral-spatial features from HSIs. Compared to the existing patch-free methods, PatchOut achieves superior classification accuracy, while significantly outperforming the patch-based approaches in terms of computational efficiency. Given the challenges in obtaining labeled HSI data, feature research will explore the self-supervised learning paradigms to minimize the requirement for annotated samples.

CRediT authorship contribution statement

Renjie Ji: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Methodology, Data curation, Conceptualization. **Kun Tan:** Writing – review & editing, Validation, Methodology, Funding acquisition, Conceptualization. **Xue Wang:** Writing – review & editing, Visualization, Validation, Supervision, Software, Methodology, Conceptualization. **Shuwei Tang:** Visualization, Validation, Software, Methodology, Investigation. **Jin Sun:** Visualization, Validation, Software, Methodology, Investigation. **Chao Niu:** Visualization, Software, Methodology, Investigation. **Chen Pan:** Visualization, Methodology, Investigation.

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.

Acknowledgments

We would like to extend our sincere gratitude to the Shanghai Institute of Surveying and Mapping for their provision of the airborne hyperspectral imagery data. Furthermore, we deeply appreciate the invaluable assistance of our laboratory members in the labeling process for the Qingpu HSI dataset. This work was supported in part by Yangtze River Delta Science and Technology Innovation Community Joint Research (Basic Research) Project (No. 2024CSJZN1300), Shanghai Municipal Education Commission Science and Technology Project (2024AI02002), National Natural Science Foundation of China (No. 42171335) and National Civil Aerospace Project of China (No. D040102).

Data availability

Data will be made available on request.

References

Anderson, K., Ryan, B., Sonntag, W., Kavvada, A., Friedl, L., 2017. Earth observation in service of the 2030 Agenda for Sustainable Development. Geo-Spat. Inf. Sci. 20, 77–96. https://doi.org/10.1080/10095020.2017.1333230.

Bioucas-Dias, J.M., Plaza, A., Camps-Valls, G., Scheunders, P., Nasrabadi, N., Chanussot, J., 2013. Hyperspectral Remote Sensing Data Analysis and Future Challenges. IEEE Geosci. Remote Sens. Mag. 1, 6–36. https://doi.org/10.1109/MGRS.2013.2244672.

Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M., 2023. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In: Karlinsky, L., Michaeli, T., Nishino, K. (Eds.), Computer Vision – ECCV 2022 Workshops. Springer Nature Switzerland, Cham, pp. 205–218.

Cui, Y., Xia, J., Wang, Z., Gao, S., Wang, L., 2022. Lightweight Spectral–Spatial Attention Network for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/TGRS.2021.3080394.

Dai, Z., Liu, H., Le, Q.V., Tan, M., 2021. CoAtNet: Marrying Convolution and Attention for All Data Sizes.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale.

Fu, H., Sun, G., Zhang, L., Zhang, A., Ren, J., Jia, X., Li, F., 2023. Three-dimensional singular spectrum analysis for precise land cover classification from UAV-borne hyperspectral benchmark datasets. ISPRS J. Photogramm. Remote Sens. 203, 115–134. https://doi.org/10.1016/j.isprsjprs.2023.07.013.

Fu, W., Xie, K., Fang, L., 2024. Complementarity-Aware Local–Global Feature Fusion Network for Building Extraction in Remote Sensing Images. IEEE Trans. Geosci. Remote Sens. 62, 1–13. https://doi.org/10.1109/TGRS.2024.3370714.

Gao, Y., Zhou, M., Metaxas, D., 2021. UTNet: A Hybrid Transformer Architecture for Medical Image Segmentation.

Hong, D., Han, Z., Yao, J., Gao, L., Zhang, B., Plaza, A., Chanussot, J., 2022. SpectralFormer: Rethinking Hyperspectral Image Classification With Transformers. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2021.3130716.

Hu, X., Zhong, Y., Wang, X., Luo, C., Zhao, J., Lei, L., Zhang, L., 2022. SPNet: Spectral Patching End-to-End Classification Network for UAV-Borne Hyperspectral Imagery With High Spatial and Spectral Resolutions. IEEE Trans. Geosci. Remote Sens. 60, 1–17. https://doi.org/10.1109/TGRS.2021.3049292.

Ji, R., Tan, K., Wang, X., Pan, C., Xin, L., 2023. PASSNet: A Spatial–Spectral Feature Extraction Network With Patch Attention Module for Hyperspectral Image Classification. IEEE Geosci. Remote Sens. Lett. 20, 1–5. https://doi.org/10.1109/LGRS.2023.3322422.

Jia, J., Chen, J., Zheng, X., Wang, Y., Guo, S., Sun, H., Jiang, C., Karjalainen, M., Karila, K., Duan, Z., Wang, T., Xu, C., Hyyppä, J., Chen, Y., 2022. Tradeoffs in the Spatial and Spectral Resolution of Airborne Hyperspectral Imaging Systems: A Crop Identification Case Study. IEEE Trans. Geosci. Remote Sens. 60, 1–18. https://doi.org/10.1109/TGRS.2021.3096999.

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P.M., 2021. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. ISPRS J. Photogramm. Remote Sens. 181, 84–98. https://doi.org/10.1016/j.isprsjprs.2021.09.005.

Li, S., Song, W., Fang, L., Chen, Y., Ghamisi, P., Benediktsson, J.A., 2019. Deep Learning for Hyperspectral Image Classification: An Overview. IEEE Trans. Geosci. Remote Sens. 57, 6690–6709. https://doi.org/10.1109/TGRS.2019.2907932.

Li, Z., Meng, Q., Guo, F., Wang, L., Huang, W., Hu, Y., Liang, J., 2023. Feature-guided dynamic graph convolutional network for wetland hyperspectral image classification. Int. J. Appl. Earth Obs. Geoinformation 123, 103485. https://doi.org/10.1016/j.jag.2023.103485.

Makantasis, K., Karantzalos, K., Doulamis, A., Doulamis, N., 2015. Deep supervised learning for hyperspectral data classification through convolutional neural networks. In: In: 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Presented at the 2015 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 4959–4962. https://doi.org/10.1109/IGARSS.2015.7326945.

Niu, B., Feng, Q., Chen, B., Ou, C., Liu, Y., Yang, J., 2022. HSI-TransUNet: A transformer based semantic segmentation model for crop mapping from UAV hyperspectral imagery. Comput. Electron. Agric. 201, 107297. https://doi.org/10.1016/j.compag.2022.107297.

Niu, C., Tan, K., Wang, X., Du, P., Pan, C., 2024. A semi-analytical approach for estimating inland water inherent optical properties and chlorophyll a using airborne hyperspectral imagery. Int. J. Appl. Earth Obs. Geoinformation 128, 103774. https://doi.org/10.1016/j.jag.2024.103774.

Paoletti, M.E., Haut, J.M., Plaza, J., Plaza, A., 2019. Deep learning classifiers for hyperspectral imaging: A review. ISPRS J. Photogramm. Remote Sens. 158, 279–317. https://doi.org/10.1016/j.isprsjprs.2019.09.006.

Roy, S.K., Haut, J.M., Paoletti, M.E., Dubey, S.R., Plaza, A., 2022. Generative Adversarial Minority Oversampling for Spectral–Spatial Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 60, 1–15. https://doi.org/10.1109/TGRS.2021.3052048.

Roy, S.K., Krishna, G., Dubey, S.R., Chaudhuri, B.B., 2020. HybridSN: Exploring 3-D–2-D CNN Feature Hierarchy for Hyperspectral Image Classification. IEEE Geosci. Remote Sens. Lett. 17, 277–281. https://doi.org/10.1109/LGRS.2019.2918719.

Roy, S.K., Manna, S., Song, T., Bruzzone, L., 2021. Attention-Based Adaptive Spectral–Spatial Kernel ResNet for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 59, 7831–7843. https://doi.org/10.1109/TGRS.2020.3043267.

Schmitt, M., Ahmadi, S.A., Xu, Y., Taşkin, G., Verma, U., Sica, F., Hänsch, R., 2023. There Are No Data Like More Data: Datasets for deep learning in Earth observation. IEEE Geosci. Remote Sens. Mag. 11, 63–97. https://doi.org/10.1109/MGRS.2023.3293459.

Su, H., Yao, W., Wu, Z., Zheng, P., Du, Q., 2021. Kernel low-rank representation with elastic net for China coastal wetland land cover classification using GF-5 hyperspectral imagery. ISPRS J. Photogramm. Remote Sens. 171, 238–252. https://doi.org/10.1016/j.isprsjprs.2020.11.018.

Sun, Y., Tian, Y., Xu, Y., 2019. Problems of encoder-decoder frameworks for high-resolution remote sensing image segmentation: Structural stereotype and insufficient learning. Neurocomputing 330, 297–304. https://doi.org/10.1016/j.neucom.2018.11.051.

Sun, L., Zhao, G., Zheng, Y., Wu, Z., 2022. Spectral–spatial feature tokenization transformer for hyperspectral image classification. IEEE Trans. Geosci. Remote Sens. 60, 1–14. https://doi.org/10.1109/TGRS.2022.3144158.

Tu, L., Li, J., Huang, X., Gong, J., Xie, X., Wang, L., 2024. S ^{2} HM ^{2} : A Spectral–Spatial Hierarchical Masked Modeling Framework for Self-Supervised Feature Learning and Classification of Large-Scale Hyperspectral Images. IEEE Trans. Geosci. Remote Sens. 62, 1–19. https://doi.org/10.1109/TGRS.2024.3392962.

Wang, C., Huang, J., Lv, M., Du, H., Wu, Y., Qin, R., 2024. A local enhanced mamba network for hyperspectral image classification. Int. J. Appl. Earth Obs. Geoinformation 133, 104092. https://doi.org/10.1016/j.jag.2024.104092.

Wang, H., Cao, P., Wang, J., Zaiane, O.R., 2022a. UCTransNet: Rethinking the Skip Connections in U-Net from a Channel-Wise Perspective with Transformer, in. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 2441–2449. https://doi.org/10.1609/aaai.v36i3.20144.

Wang, X., Tan, K., Du, P., Han, B., Ding, J., 2023. A capsule-vectored neural network for hyperspectral image classification. Knowl.-Based Syst. 268, 110482. https://doi.org/10.1016/j.knosys.2023.110482.

Wang, X., Tan, K., Du, P., Pan, C., Ding, J., 2022b. A Unified Multiscale Learning Framework for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 1–1. https://doi.org/10.1109/TGRS.2022.3147198.

Woo, S., Debnath, S., Hu, R., Chen, X., Liu, Z., Kweon, I.S., Xie, S., 2023. ConvNeXt V2: Co-designing and Scaling ConvNets with Masked Autoencoders.

Xu, Y., Gong, J., Huang, X., Hu, X., Li, J., Li, Q., Peng, M., 2023. Luojia-HSSR: A high spatial-spectral resolution remote sensing dataset for land-cover classification with a new 3D-HRNet. Geo-Spat. Inf. Sci. 26, 289–301. https://doi.org/10.1080/10095020.2022.2070555.

Yao, J., Zhang, B., Li, C., Hong, D., Chanussot, J., 2023. Extended Vision Transformer (ExViT) for Land Use and Land Cover Classification: A Multimodal Deep Learning Framework. IEEE Trans. Geosci. Remote Sens. 61, 1–15. https://doi.org/10.1109/TGRS.2023.3284671.

Yue, J., Zhao, W., Mao, S., Liu, H., 2015. Spectral–spatial classification of hyperspectral images using deep convolutional neural networks. Remote Sens. Lett. 6, 468–477. https://doi.org/10.1080/2150704X.2015.1047045.

Zhang, X., Su, Y., Gao, L., Bruzzone, L., Gu, X., Tian, Q., 2023a. A Lightweight Transformer Network for Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 61, 1–17. https://doi.org/10.1109/TGRS.2023.3297858.

Zhang, X., Yan, J., Tian, J., Li, W., Gu, X., Tian, Q., 2023b. Objective evaluation-based efficient learning framework for hyperspectral image classification. Giscience Remote Sens. 60, 2225273. https://doi.org/10.1080/15481603.2023.2225273.

Zheng, Z., Zhong, Y., Ma, A., Zhang, L., 2020. FPGA: Fast Patch-Free Global Learning Framework for Fully End-to-End Hyperspectral Image Classification. IEEE Trans. Geosci. Remote Sens. 58, 5612–5626. https://doi.org/10.1109/TGRS.2020.2967821.

Zhong, Y., Hu, X., Luo, C., Wang, X., Zhao, J., Zhang, L., 2020. WHU-Hi: UAV-borne hyperspectral with high spatial resolution (H2) benchmark datasets and classifier for precise crop identification based on deep convolutional neural network with CRF. Remote Sens. Environ. 250, 112012. https://doi.org/10.1016/j.rse.2020.112012.

Zhong, Z., Li, J., Luo, Z., Chapman, M., 2018. Spectral–Spatial Residual Network for Hyperspectral Image Classification: A 3-D Deep Learning Framework. IEEE Trans. Geosci. Remote Sens. 56, 847–858. https://doi.org/10.1109/TGRS.2017.2755542.

Zhu, Q., Deng, W., Zheng, Z., Zhong, Y., Guan, Q., Lin, W., Zhang, L., Li, D., 2022. A spectral-spatial-dependent global learning framework for insufficient and imbalanced hyperspectral image classification. IEEE Trans. Cybern. 52, 11709–11723. https://doi.org/10.1109/TCYB.2021.3070577.