2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

HRS-UNet: A Semantic Segmentation Model for Precise Crop Classification in Hyperspectral Remote Sensing Image

Zhiyu Yang[1] , Lei Zou[1] and Yuhuai Lin[1,*]

[1] College of Information and Electrical Engineering, China Agricultural University, Beijing, China
{yangzhiyu, 2022308250126, 2022308250214}@cau.edu.cn

Abstract. Precise crop classification, as a pivotal technology underpinning precision agriculture, has attracted considerable attention in recent years. Hyperspectral imaging systems mounted on Unmanned Aerial Vehicles (UAVs) are capable of producing high spatial resolution hyperspectral imagery, offering distinct advantages including low operational costs, high operational flexibility, and real-time data acquisition. As a result, these systems have emerged as an optimal tool for precise crop classification within precision agriculture monitoring. Nevertheless, existing methods for crop classification using UAV hyperspectral imagery encounter a trade-off between global feature perception and computational complexity, frequently leading to the loss of spatial features. To tackle this issue, this study introduces a hyperspectral segmentation network, HRS-UNet, designed to achieve precise crop classification from hyperspectral samples. And we propose a Multiscale Spectral Aggregation (MSA) module, which greatly reduces the computational burden of the backbone network through feature enhancement and dimensionality reduction. Evaluation results on the UAV-HSI-Crop dataset reveals that our model attains state-of-the-art performance, achieving an overall classification accuracy of 89.96% and a Kappa coefficient of 0.8814, outperforming existing approaches. Our model offers a novel technical pathway for efficient monitoring in precision agriculture.

Keywords: Precision Agriculture, Hyperspectral Image, Semantic Segmentation, Remote Sensing.

1 Introduction

Driven by the dual imperatives of global food security challenges and the United Nations Sustainable Development Goals (SDGs), precision agriculture has emerged as a pivotal paradigm propelling modern agriculture towards sustainability and intelligence[1]. This paradigm, deeply rooted in the integration of information and engineering technologies, aims to achieve synergistic improvements in crop yield, resource use efficiency, and environmental protection by precisely sensing and responding to the spatiotemporal variability of agroecosystems[2]. The realization of this objective critically hinges on the capability to monitor crop life cycle status with precision, real-time

* Corresponding author

responsiveness, and non-destructiveness. However, traditional coarse-grained land cover classification methods fall short of meeting the demands for fine-grained management in precision agriculture[3]. Consequently, the development of classification techniques capable of accurately distinguishing between crop varieties and even growth stages at the field level has become an urgent scientific imperative[4].

To address this challenge, hyperspectral remote sensing provides robust data and technical support for the precise crop classification. Unlike multispectral imaging systems that capture information only in a few discrete and broad spectral bands, hyperspectral sensors acquire data across hundreds of continuous and extremely narrow spectral channels, enabling the detailed characterization of unique spectral reflectance curves of ground objects[5]. This exceptional spectral resolution allows hyperspectral data to reveal subtle spectral features arising from differences in the physical structure and chemical composition of ground objects—features often indiscernible in conventional multispectral data—thereby laying a solid spectral information foundation for the precise identification and quantitative inversion of surface targets.

In recent years, the rapid advancements in computer vision, machine learning, and particularly deep learning, have significantly propelled the progress of hyperspectral image classification techniques. Early studies predominantly relied on traditional machine learning algorithms such as Support Vector Machines (SVM) or Random Forests (RF) for pixel-wise independent classification[6, 7]. However, these methods overlooked the spatial contextual information within images, failing to leverage the spatial correlations between pixels. With the advent of deep learning, Convolutional Neural Networks (CNNs)[8, 9] and Transformer-based model[10] have become mainstream technological pathways for hyperspectral classification due to their powerful nonlinear feature representation capabilities, giving rise to general segmentation frameworks like SegNet, SETR, UNet, and TransUNet. Nonetheless, these models still grapple with inherent limitations: CNNs are constrained by their local receptive fields, struggling to capture long-range spatial dependencies and global contextual features; while Transformers excel at global modeling, their self-attention mechanisms incur a computational complexity that scales quadratically with the number of pixels, posing significant challenges for dense prediction tasks like hyperspectral image classification, which involves high-dimensional feature redundancy.

In the task of precise crop classification, hyperspectral addresses the limitations of traditional remote sensing methods, particularly the inadequate spectral resolution for distinguishing between similar crops. UAVs equipped with hyperspectral systems are capable of generating high spatial resolution hyperspectral imagery. With advantages such as low operational costs, high flexibility, and real-time data acquisition, UAV-based hyperspectral imagery has emerged as a vital data source for precise crop classification. Recent research on precise crop classification leveraging UAV-based hyperspectral imagery has achieved notable progress. For example, Zhong et al.[11] proposed a framework that integrates a deep convolutional neural network with a conditional random field classifier (CNNCRF), demonstrating commendable classification performance on their publicly available WHU-Hi dataset. Guo et al.[12] developed a hybrid model combining a convolutional neural network (CNN) and a Transformer, which employs a spectral-spatial feature extraction module to capture shallow features

and a dual-branch architecture to simultaneously extract local and global features. This model exhibited strong potential for hyperspectral classification of agricultural fields on the WHU-Hi dataset. Furthermore, Tang et al.[3] designed a spatial-spectral attention network that overcomes the shortcomings of existing spatial-spectral attention mechanisms, which focus solely on single features, thereby enhancing classification performance.

Despite these advancements, current methods are still constrained by the trade-off between global perception capability and computational complexity. Predominantly, existing models rely on extracting local features from pixel neighborhoods, lacking the capacity to perceive features across larger spatial extents. In UAV-based hyperspectral classification tasks, where image spatial resolution often reaches the centimeter level, crop distributions may extend beyond predefined neighborhood ranges. This limited spatial perception can result in the omission of critical features. To address this issue, Niu et al.[13] introduced a more complex dataset for precise crop classification, UAV-HSI-Crop, and proposed the HSI-TransUNet model to achieve holistic classification at larger scales. Nevertheless, the model's performance and computational efficiency still warrant further optimization. Consequently, the development of an advanced model capable of efficiently integrating multi-scale spatial features, while balancing local details with global context and maintaining manageable computational costs, remains a pressing scientific challenge in the domain of precise hyperspectral crop classification.

In light of this, this study proposes a hyperspectral classification model tailored for precise crop classification, named HRS-UNet. This model achieves an effective balance between global perception capability and computational complexity, while utilizing the Multiscale Spectral Aggregation (MSA) module for spectral feature extraction and dimensionality reduction. This approach provides a novel technical pathway for hyperspectral crop classification, poised to advance the further development of precision agriculture.

The main contributions of this study can be summarized as follows:

(1) We propose a novel U-Net-based model for precise crop classification, effectively balancing global perception capability with computational complexity.

(2) We design a Multiscale Spectral Aggregation module within the model to extract the spectral feature and dimensionality reduction. Additionally, we integrate a spatial-transformer and global attention mechanisms to strengthen the model's global perception capability.

(3) Experimental results demonstrate that our model achieves state-of-the-art performance on the UAV-HSI-Crop dataset.

2 Methods

The different bands in hyperspectral images contain rich feature information of each crop, which is beneficial for the model to learn. Our proposed method aggregates multi-scale spectral channel features and utilizes an improved UNet architecture for spatial feature perception.

2.1 Pipeline

The architecture of our model is shown in Fig. 1. The input hyperspectral image $I \in \mathbb{R}^{N \times C \times H \times W}$ first undergoes a dimension reduction operation through the Multiscale Spectral Aggregation (MSA) module, obtained the Aggregated hyperspectral image $\hat{I} \in \mathbb{R}^{N \times C' \times H \times W}$. The structure of our backbone network is similar to the UNet[14]. The encoder of our network consists of four important modules: ResBlock, Spatial-Transformer, Global Attention, and Downsample Blocks. The detail of these modules is shown in Fig. 2. Each Resblock changes the number of channels of the feature map, Spatial-Transformer extracts local features on the feature map using the self-attention mechanism, followed by the Global Attention perceiving the image features in channel and spatial dimensions. Finally, the Downsample Block will perform a two-times downsampling of the feature map. The structure of the decoder is similar to the encoder, where the class prediction probability of each pixel $P \in \mathbb{R}^{N \times Classes \times H \times W}$ is obtained through the Fully Connected (FC) layer after the feature map size is recovered using the transposed convolution.

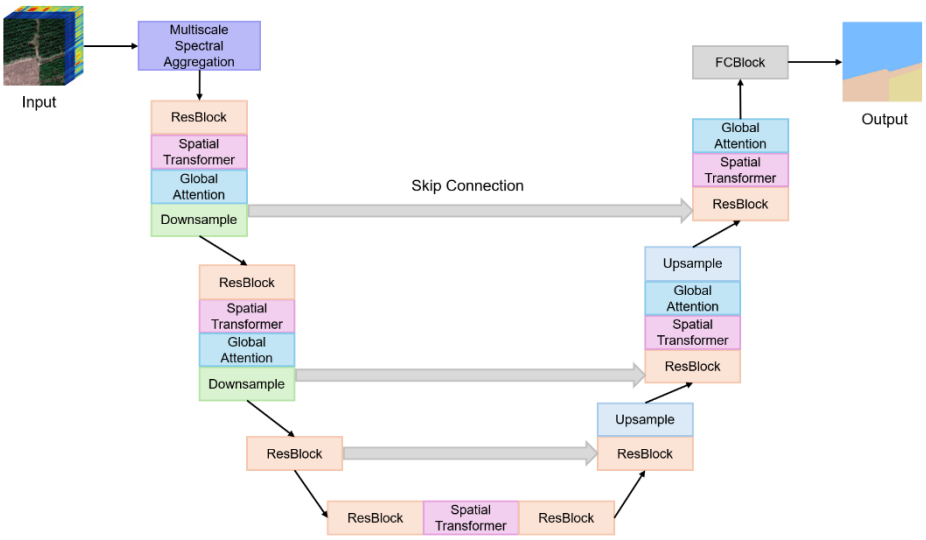

Fig. 1. The overall pipeline of our proposed method.

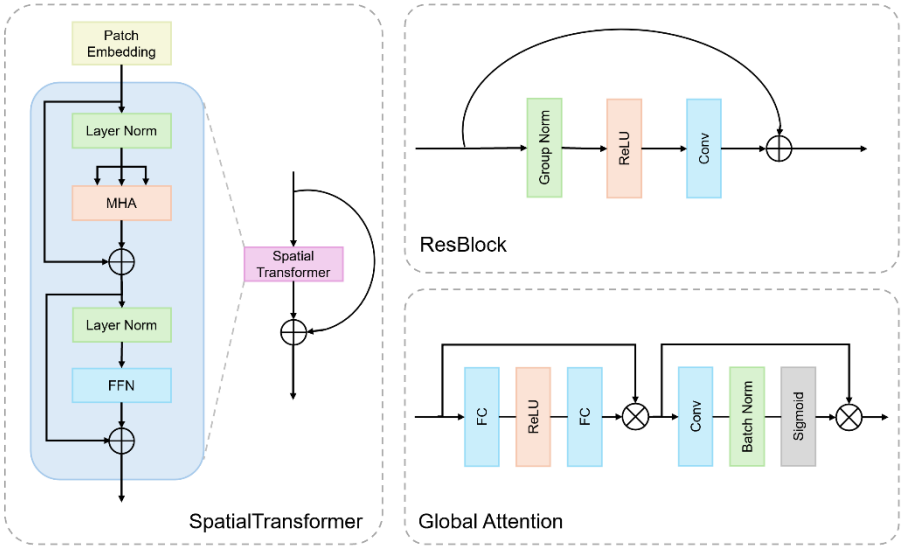

Fig. 2. The detail of modules: ResBlock, Spatial-Transformer and Global Attention.

2.2 Multiscale Spectral Aggregation

To address the issue of increased computational complexity caused by multi-channel features in hyperspectral images, we propose a Multiscale Spectral Aggregation (MSA) module. This module achieves end-to-end joint feature learning through parallelized spectral feature compression and multi-scale spatial feature extraction. As shown in Fig. 3, this module includes convolutional layers, Batch Normalization (BN), ReLU, and self-attention mechanism, where "Conv_k" represents a convolutional layer with kernel size of k.

The MSA module utilizes multi-branch parallelism to construct multi-scale convolutional layer groups to capture spatial features of different scales. In addition, the model uses convolution for feature channel dimension transformation to compress the spectral channels. The calculation process of each branch is as follows:

$$F_i = SA\left(ReLU\left(BN\left(W^k \times \hat{I}\right)\right)\right) \tag{1}$$

where F_i denotes the output of i-th branch, and W^k represents the parameters of a convolutional layer with kernel size of k. The calculation of Self-Attention is carried out in the spectral dimension. After flattening the spatial dimension of the feature map, we use a linear layer to calculate the Query, Key and Value vectors of the embedding. Finally, we concatenate the outputs of each branch in the spectral channel dimension. And the final output is obtained by fusing convolutional layer.

In this study, we used MSA with four branches, each of which compressed the spectral channels to 32 dimensions. The convolution kernel sizes for the four branches are set to 1, 3, 5, and 7. The selection of branch number and scale size will be validated in

ablation experiments. After concatenation and fusion, we finally obtained the output of 128 channels.

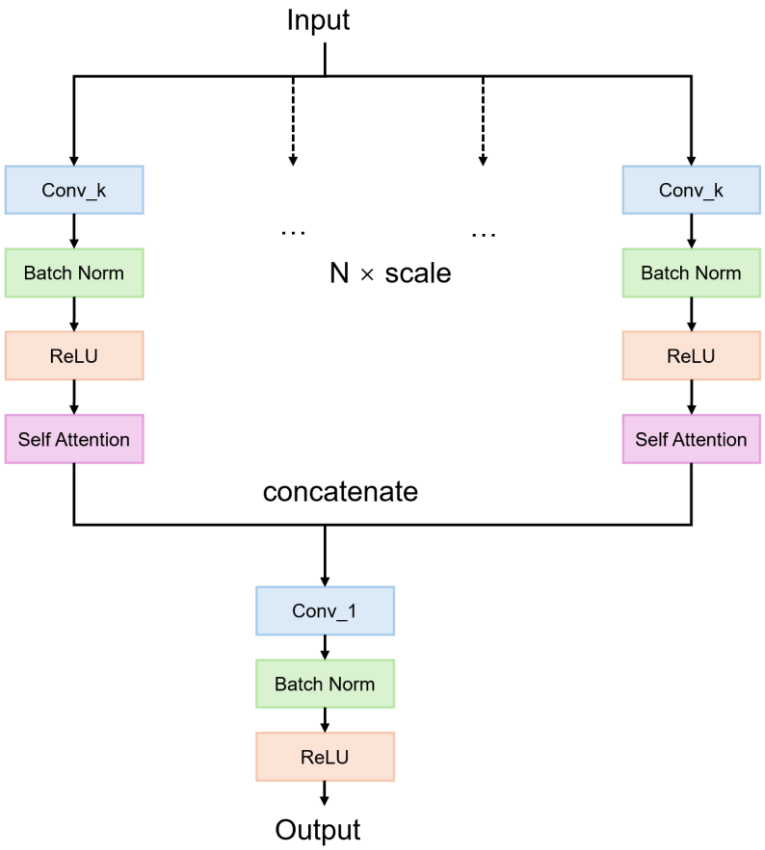

Fig. 3. The architecture of Multiscale Spectral Aggregation module.

2.3 Module design

ResBlock. The architecture of ResBlock is shown in Fig. 2. The key part of this module consists of convolutional layers, Group Normalization (GN) and ReLU activation functions, and the final output forms a residual connection with the input. GN normalizes data by grouping channels, preserving the correlation of channels within each group, making it more suitable for capturing local features. GN is not affected by batch size and can maintain a stable distribution of data even when the batch size is small, with robustness. ReLU combines the smoothness of Sigmoid and the nonlinear characteristics of ReLU, and has received widespread attention for its excellent performance in deep networks. The calculation process of ResBlock can be defined as:

$$Z_{out} = conv\left(ReLU\left(GN\left(Z_{in}\right)\right)\right) + Z_{in} \tag{2}$$

Spatial-Transformer. The architecture of Spatial-Transformer is shown in Fig. 2. Unlike ViT [15], our Spatial-Transformer uses convolutional layers with kernel size of 1 to patch the image. Namely, we only change the dimension of feature embedding and use self-attention mechanism to capture the relationships in the overall space of the feature map. After patch embedding, the basic transformer[16] block, including Layer Normalization (LN), Multi-Head Attention (MHA), and Feed Forward Network (FFN), is used to process feature maps:

$$Z_{out} = FFN\left(Z'_{in} + LN\left(Z'_{in}\right)\right) + Z'_{in} \tag{3}$$

$$Z'_{in} = MHA\left(Z_{in} + LN\left(Z_{in}\right)\right) + Z_{in} \tag{4}$$

Global Attention. Global Attention replaces traditional local operations with lightweight global attention computation, aiming to enhance the global modeling capability of the model by capturing long-range dependencies. And Global Attention dynamically integrates multidimensional information through dual-path attention to jointly model spectral channels and spatial features. Firstly, perform attention calculation on the spectral channel:

$$Z'_{in} = FC\left(ReLU\left(FC\left(Z_{in}\right)\right)\right) \odot Z_{in} \tag{5}$$

Next, we use convolutional layers and Batch Normalization (BN) to model spatial information and generate attention weights using Sigmoid, ultimately obtaining the output of Global Attention:

$$Z_{out} = Sigmoid\left(BN\left(conv\left(Z'_{in}\right)\right)\right) \odot Z'_{in} \tag{6}$$

2.4 Loss Function Design

In this study, we use a hybrid loss function that effectively combines Cross-Entropy Loss[17] and Log-Cosh Dice Loss[13] to address challenges in segmentation tasks. The loss function is defined as:

$$\mathcal{L} = \mathcal{L}_{CE} + \mathcal{L}_{LC-Dice} \tag{7}$$

The Cross-Entropy Loss $\mathcal{L}_{CE}$ is widely used in classification tasks and is particularly effective in penalizing pixel-level misclassifications. The Cross-Entropy Loss is computed as follows:

$$\mathcal{L}_{CE} = -\sum\sum y_{i,c} \log\left(p_{i,c}\right) \tag{8}$$

where $y$denotes the one-hot encoding of the ground truth label, and $p_{i,c}$ stands for the probability distribution of the predicted result, i represents the i-th pixel, c refers to the c-th category.

To better address inherent class imbalance issues in segmentation tasks and improve boundary delineation, we use the Log-Cosh Dice Loss. This loss function combines the advantages of Dice loss and the logarithmic hyperbolic cosine function, providing stable gradients and robustness to outliers. The Log-Cosh Dice Loss is defined as:

$$\mathcal{L}_{LC-Dice} = \log\left(\cosh\left(1 - Dice\left(p, y\right)\right)\right) \tag{9}$$

where $Dice\left(\cdot,\cdot\right)$ represents the Dice Loss:

$$Dice\left(p, y\right) = \sum_{i} \left(1 - \frac{2|P_i \cap GT_i| + \epsilon}{|P_i| + |GT_i| + \epsilon}\right) \tag{10}$$

where P_i and GT_i stand for the model's prediction and the GT label, respectively, while index i denotes a class index. The addition of $\epsilon = 1e-5$ is to avoid zero in the denominator when $P_i = GT_i = 0$.

The logarithmic hyperbolic cosine function, which can be defined as:

$$\log\left(\cosh\left(x\right)\right) = \log\left(\frac{e^x + e^{-x}}{2}\right) \tag{11}$$

2.5 Evaluation Metrics

We evaluated the model's semantic segmentation performance on the dataset using Overall Accuracy (OA) and the kappa coefficient. OA represents the proportion of correctly classified samples to the total number of samples, while the kappa coefficient quantifies the reduction in classification error compared to a completely random classification. The formulas are defined as:

$$OA = \frac{1}{N}\sum_{i=1}^{r} x_{ii} \tag{12}$$

$$Kappa = \frac{N\sum_{i=1}^{r} x_{ii} - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)}{N^2 - \sum_{i=1}^{r}\left(x_{i+} \times x_{+i}\right)} \tag{13}$$

where N denotes the total number of elements in the confusion matrix,x_{ii} represents the diagonal elements of the confusion matrix, x_{i+} is the sum of row x_{i*}, x_{+i} is the sum of column x_{*i}.

3 Experiments

3.1 Experiment Set-up

Dataset. We have conducted our method in the UAV-HSI-Crop dataset[13]. The UAV-HSI-Crop dataset from China Agricultural University was collected through hyperspectral image acquisition in farmland plots located in both Majiakou Village and Xijingmeng Village, Shenzhou City, Hebei Province, China. The data was captured using a Pika L hyperspectral imager manufactured by Resonon, covering a spectral range of 400–1000 nm with 200 spectral bands. The ground sampling distance is approximately 100 meters, and the spatial resolution is 0.1 meters per pixel. The dataset includes 433

hyperspectral images of size 96×96 pixels, covering 27 distinct vegetation categories such as bare soil and weed, Chinese cabbage, corn, millet, and others.

Implementation Details. All experiments are conducted with 8 × NVIDIA RTX 3090 GPUs. The training lasts for 200 epochs with the batch size of 8. The Adam optimizer is employed with an initial learning rate of 4e-5 and weight decay of 0.0005. Our model performs two downsampling operations, while ResBlock performs a double-dimensional operation on the channel. We only apply the Spatial-Transformer and Global Attention modules during downsampling. The depth of the Spatial-Transformer is set to 1.

3.2 Results

Comparison with other methods. Our proposed method demonstrates significant improvements over existing state-of-the-art approaches when evaluated on the UAV-HSI-Crop benchmark dataset. As is shown in Tab. 1, our method achieves a remarkable overall accuracy (OA) of 89.96% with a corresponding Kappa coefficient of 0.8814, representing a substantial performance gain of 4.54 percentage points in OA compared to the HSI-TransUNet baseline. This significant enhancement can be attributed to several key architectural innovations and methodological contributions.

The superior performance metrics demonstrate the efficacy of our integrated pipeline architecture, which synergistically combines the novel Multiscale Spectral Aggregation (MSA) module with advanced loss function optimization. The MSA module effectively captures multi-resolution spectral-spatial features across different scales, enabling more comprehensive representation learning from hyperspectral data. Furthermore, the implementation of Log-Cosh Dice Loss addresses the inherent class imbalance challenges prevalent in agricultural land cover classification tasks, particularly for minority classes representing small agricultural parcels.

Qualitative assessment through semantic segmentation visualization (Fig. 4) provides additional validation of our method's effectiveness in preserving fine-grained boundary details and accurately delineating complex agricultural landscape features. The enhanced boundary preservation capability is particularly crucial for precision agriculture applications where accurate field boundary detection is essential for yield estimation and crop management decisions.

Table 1. Results on the UAV-HSI-Crop dataset compared to the state-of-the-art methods.

Method	OA%	Kappa
SegNet[18]	43.61	0.5415
SETR[19]	69.47	0.7267
UNet[14]	76.07	0.7131
TransUNet[20]	78.64	0.7456
HSI-TransUNet[13]	86.05	0.8347
HRS-UNet(Ours)	**89.96**	**0.8814**

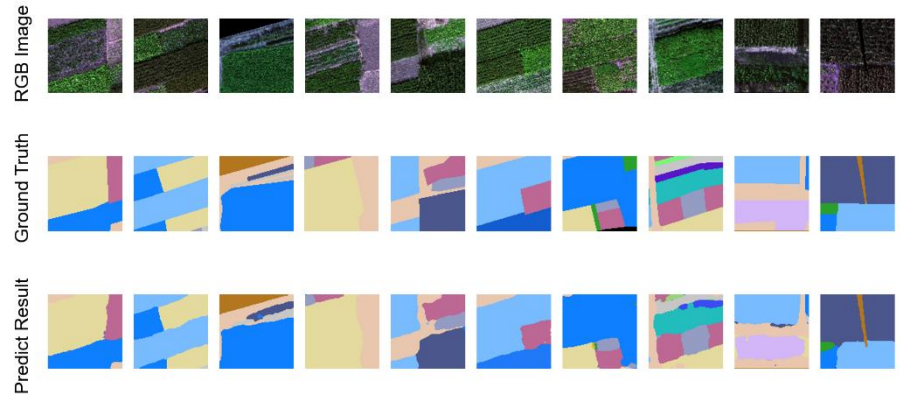

Fig. 4. The visualization results of semantic segmentation.

3.3 Ablation Study

Effectiveness of Multiscale Spectral Aggregation and other modules in Pipeline. To validate the contributions of key modules, ablation experiments were conducted on the UAV-HSI-Crop dataset, and the results are shown in Tab. 2. For each module, we remove it and replace it with a single convolutional layer with kernel size of 1, to verify its effectiveness. From the experimental results, it can be seen that after removing MSA, the overall accuracy decreased to 87.04%. This indicates that our MSA module plays an important role in the entire pipeline, effectively aggregating the features of spectral channels. After removing other modules, our model can still maintain high performance, which proves the robustness of our model.

Table 2. Ablation results on the pipeline, including Multiscale Spectral Aggregation, ResBlock, Spatial-Transformer and Global Attention modules.

MSA	ResBlock	ST	GA	OA(%)	Kappa
×	-	-	-	87.04	0.8462
-	×	-	-	89.46	0.8757
-	-	×	-	87.55	0.8531
-	-	-	×	87.78	0.8550
-	-	-	-	**89.96**	**0.8814**

Branch and scale selection for MSA. The ablation experiment results of branch and scale selection for MSA are shown in the Tab. 3. We set different numbers of branches and the size of convolution kernels in each branch separately. From the experimental results, it can be seen that our selection of branch numbers and scales achieved the best performance. When the number of branches decreases, the feature expression performance of the model will significantly decline. And as the number of branches increases, the model exhibits overfitting, which affects the accuracy of segmentation. The experimental results indicate that we have chosen the optimal branch number and scale.

Table 3. Ablation results on the different branches and kernel sizes of MSA.

Branches	Kernel Size	OA(%)	Kappa
3	1 3 5	88.14	0.8598
5	1 3 5 7 9	88.80	0.8677
4	3 5 7 9	88.50	0.8642
4	1 3 5 7	**89.96**	**0.8814**

4 Conclusion

This study proposes a semantic segmentation model, HRS-UNet, designed for precise crop classification from UAV-based hyperspectral imagery. HRS-UNet addresses the limitations of existing methods through two innovative components. The pipeline of our proposed model is based on the UNet framework, consisting of ResBlock, Spatial-Transformer and Global Attention mechanism. And we introduce the Multiscale Spectral Aggregation module, which enhances feature representation and lowers the computational complexity of the model through spectral channels aggregation. Experimental evaluations on the benchmark UAV-HSI-Crop dataset demonstrate that HRS-UNet achieves state-of-the-art performance, with an overall classification accuracy of 89.96% and a Kappa coefficient of 0.8814, surpassing existing methods. These findings indicate that HRS-UNet offers efficient and accurate crop classification support for precision agriculture monitoring, with substantial application potential.

References

1. Rao N.C., Bathla S., Kumar A., Jha G.K.: Agriculture and sustainable development goals: an overview and issues. Agricultural Economics Research Review. 31. 1 (2018).
2. van Noordwijk M., Duguma L.A., Dewi S., Leimona B., Catacutan D.C., Lusiana B., Öborn I., Hairiah K., Minang P.A.: SDG synergy between agriculture and forestry in the food, energy, water and income nexus: reinventing agroforestry? Current opinion in environmental sustainability. 34. 33 (2018).
3. Tang G., Wang X., Zhao H., Hu X., Jin G., Zhong Y.: Attention in attention for hyperspectral with high spatial resolution (H) image classification. IEEE Transactions on Geoscience and Remote Sensing. 62. 1 (2023).
4. Dong J., Wu W., Huang J., You N., He Y., Yan H.: State of the art and perspective of agricultural land use remote sensing information extraction. Journal of Geo-information Science. 22. 772 (2020).
5. Bioucas-Dias J.M., Plaza A., Camps-Valls G., Scheunders P., Nasrabadi N., Chanussot J.: Hyperspectral remote sensing data analysis and future challenges. IEEE Geoscience and remote sensing magazine. 1. 6 (2013).
6. Ham J., Chen Y., Crawford M.M., Ghosh J.: Investigation of the random forest framework for classification of hyperspectral data. IEEE Transactions on Geoscience and Remote Sensing. 43. 492 (2005).

7. Tu B., Wang J., Kang X., Zhang G., Ou X., Guo L.: KNN-based representation of superpixels for hyperspectral image classification. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 11. 4032 (2018).
8. Fu C., Du B., Zhang L.: Resc-net: Hyperspectral image classification based on attention-enhanced residual module and spatial-channel attention. IEEE Transactions on Geoscience and Remote Sensing. 2024).
9. Zhang Z., Huang L., Wang Q., Jiang L., Qi Y., Wang S., Shen T., Tang B., Gu Y.: UAV Hyperspectral Remote Sensing Image Classification: A Systematic Review. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing. 2024).
10. [10] Li K., Wan Y., Ma A., Zhong Y.: A lightweight multi-scale and multi-attention hyperspectral image classification network based on multi-stage search. IEEE Transactions on Geoscience and Remote Sensing. 2025).
11. Zhong Y., Hu X., Luo C., Wang X., Zhao J., Zhang L.: 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 Sensing of Environment. 250. 112012 (2020).
12. Guo X., Feng Q., Guo F.: CMTNet: a hybrid CNN-transformer network for UAV-based hyperspectral crop classification in precision agriculture. Scientific Reports. 15. 12383 (2025).
13. Niu B., Feng Q., Chen B., Ou C., Liu Y., Yang J.: HSI-TransUNet: A transformer based semantic segmentation model for crop mapping from UAV hyperspectral imagery. Computers and Electronics in Agriculture. 201. 107297 (2022).
14. Ronneberger O., Fischer P., Brox T.: U-net: Convolutional networks for biomedical image segmentation. In: Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18, vol. 234. Springer,2015)
15. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929. 2020).
16. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.: Attention is all you need. Advances in neural information processing systems. 30. 2017).
17. Cox D.R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society Series B: Statistical Methodology. 20. 215 (1958).
18. Badrinarayanan V., Kendall A., Cipolla R.: Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence. 39. 2481 (2017).
19. Zheng S., Lu J., Zhao H., Zhu X., Luo Z., Wang Y., Fu Y., Feng J., Xiang T., Torr P.H.: Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, vol. 6881. 2021)
20. Chen J., Lu Y., Yu Q., Luo X., Adeli E., Wang Y., Lu L., Yuille A.L., Zhou Y.: Transunet: Transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306. 2021).