Hyperspectral Image Instance Segmentation Using Spectral-Spatial Feature Pyramid Network

Leyuan Fang, *Senior Member, IEEE,* Yifan Jiang, Yinglong Yan, Jun Yue, and Yue Deng

***Abstract*—In recent years, hyperspectral image (HSI) classification and detection techniques based on deep learning have been widely applied to various aspects, such as environmental monitoring, urban planning, and energy surveys. As an important image content analysis method, instance segmentation can provide important support for the extraction of ground object information and monomeric application of HSI. This paper introduces instance segmentation into HSI interpretation for the first time. In this paper, we create the hyperspectral Instance Segmentation dataset (HS-ISD), which contains a total of 56 images, each with a size of 298 × 301 and a number of channels of 48. More than 1000 architectural examples are annotated to apply to the research of HSI instance segmentation. In addition, considering that HSI contains rich spectral and spatial information, and the traditional instance segmentation network model cannot well utilize both types of information effectively, we propose the spectral-spatial feature pyramid network (Spectral-Spatial FPN). The Spectral-Spatial FPN can integrate multi-scale Spectral information and multi-scale Spatial information in the feature extraction stage through attention mechanism and bidirectional feature pyramid structure, so as to better improve the performance of the network model by Spectral information and Spatial information, and realize end-to-end instance segmentation of HSI. The experimental results conducted on the HS-ISD show that the proposed Spectral-Spatial FPN can achieve state-of-the-art results.**

***Index Terms*—Hyperspectral image (HSI) instance segmentation, spectral-spatial feature pyramid network (Spectral-Spatial FPN), deep learning, spectral and spatial information, feature fusion.**

I. Introduction

INTELLIGENT interpretation of hyperspectral image (HSI) is of great value in scientific research and engineering applications. With the continuous advancement of deep learning in the field of remote sensing image interpretation, good results have been achieved in the classification [1, 2], object detection [3–5] and semantic segmentation [6, 7] of HSI. Meanwhile, with the development of remote sensing technology, the quality and resolution of HSI are significantly improved. Many important fields, such as the correction of civil map and the three-dimensional reconstruction of cities, need the support of more refined classification of remote sensing images [8]. Instance segmentation is not only to classify each pixel, but also to distinguish all individuals in the same category, which can make more precise use of HSI information. The implementation of end-to-end HSI instance segmentation can eliminate the redundant operation of traditional classification methods, and provide new opportunities for HSI processing. [9, 10].

Semantic segmentation can assign a corresponding category to each pixel in an image, but it does not distinguish objects in the same category [11]. As an important direction in computer vision, instance segmentation combines the two tasks of object detection and semantic segmentation to distinguish different instances on the basis of specific categories and achieve different categories of pixel-level segmentation on the results of instance-level object location [12]. With the application of deep convolution neural network, many instance segmentation frameworks have been proposed, such as InstanceFCN [13], Mask R-CNN [14], MS R-CNN [15], PointRend [16], etc. The process of instance segmentation can be divided into single stage and two stages. In 2017, Mask R-CNN model proposed by He Kaiming [14] and others has become the baseline model of many instance segmentation algorithms, and it is also the most frequently used instance segmentation algorithm today. BlendMask [17] integrates high-level and low-level semantic information through blender module, thus surpassing Mask R-CNN in speed and accuracy. Recently, transformer model has also aroused great interest in the field of computer vision. ISTR [18] is the first end-to-end instance segmentation framework based on transformer. Compared with CNN, visual transformer is also very competitive in the field of instance segmentation.

Compared with natural image, remote sensing image contains more objects and more complex backgrounds. The remote sensing image has a large range of variations in the size of similar targets and a large difference in color texture, which makes the instance segmentation task of the remote sensing image more challenging. Fig. 1 shows the difference between the HSI instance segmentation task and related task. It can be seen that instance segmentation of HSI is very challenging: (a) how to utilize the complex spatial and spectral information of HSI simultaneously, and (b) how to simply and effectively

This work was supported in part by the National Natural Science Foundation of China under Grant U22B2014 and Grant 62101072, in part by the Science and Technology Plan Project Fund of Hunan Province under Grant 2022RSC3064, in part by the Key Research and Development Program of Hunan Province of China under Grant 2021SK2039, in part by the Hunan Provincial Natural Science Foundation of China under Grant 2021JJ30003 and Grant 2021JJ40570, and in part by the Scientific Research Foundation of Hunan Education Department under Grant 20B022 and Grant 20B157. *(Corresponding author: Jun Yue.)*

Leyuan Fang is with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China, and also with the Peng Cheng Laboratory, Shenzhen 518000, China (e-mail: fangleyuan@gmail.com).

Yifan Jiang and Yinglong Yan are with the College of Electrical and Information Engineering, Hunan University, Changsha 410082, China (e-mail: yfjiang@hnu.edu.cn; yanyl@hnu.edu.cn).

Jun Yue is with the Department of Geomatics Engineering, Changsha University of Science and Technology, Changsha 410114, China (e-mail: jyue@pku.edu.cn).

Yue Deng is with the School of Astronautics, Beihang University, Beijing 100191, China (e-mail: ydeng@buaa.edu.cn).

distinguish different instances of objects of the same class. These are the main difficulties that the instance segmentation of HSI often cannot reflect the node and contour features of the instance objects well, which makes it difficult to achieve good segmentation results.

In this paper, we propose the spectral-spatial feature pyramid network (Spectral-Spatial FPN) for end-to-end instance segmentation of HSI. Firstly, due to the lack of datasets that can be used for HSI instance segmentation research, we construct a new hyperspectral instance segmentation dataset (HS-ISD) for the first time. HS-ISD contains 56 images of size 298×301, each with 48 channels, and a total of 1085 building instances are annotated. Secondly, our proposed Spectral-Spatial FPN is simple and effective without excessive computational overhead. It can be added to feature extraction in some state-of-the-art instance segmentation models to achieve the fusion of multi-scale spectral features with spatial features through the attention mechanism and the bidirectional fusion feature pyramid structure. The features extracted by the common model are mainly the spatial information of the image. Facing the HSI with more complex image details, to make full use of the rich spectral dimensional information of HSI, our proposed Spectral-Spatial FPN makes reasonable use of the spectral information to exploit the complementarity of Spatial information and Spectral information, which can better realize the fine-grained instance segmentation of remote sensing images. Based on the proposed dataset HS-ISD, numerous experiments on the popular mainstream network models in recent years show that the proposed Spectral-Spatial FPN can significantly promote the instance segmentation of HSI and achieve new state-of-the-art results.

Our contributions are summarized below.

- We introduce the idea of instance segmentation into HSI processing for the first time, and publish the HSI instance segmentation benchmark dataset named HS-ISD. The dataset will be made publicly available at https://github.com/sweetener08/Spectral-Spatial-FPN.
- We propose the Spectral-Spatial FPN for instance segmentation of HSI, which can realize the fusion of spatial and spectral information of HSI to achieve end-to-end instance segmentation of HSI. The source code will be available at https://github.com/sweetener08/Spectral-Spatial-FPN.
- We conduct experiments on HS-ISD using different instance segmentation network models, and the experimental results demonstrate the superiority of Spectral-Spatial FPN. The network model incorporating Spectral-Spatial FPN can achieve the state-of-the-art HSI instance segmentation results.

The remainder of this article is structured as follows. Section II summarizes the related works of instance segmentation and intelligent interpretation of HSI. Section III describes the details of the proposed method Spectral-Spatial FPN and the process of HSI instance segmentation. The introduction of the proposed dataset HS-ISD and the results and discussion of all experiments are given in Section IV. Finally, Section V concludes the paper and provides some ideas for future research work.

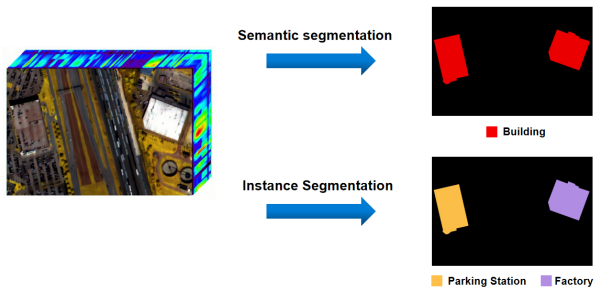


Fig. 1. The difference between the instance segmentation and semantic segmentation of HSI.

II. Related Works

A. General Instance Segmentation

Instance segmentation is one of the important tasks in computer vision, which can distinguish different instances based on semantic segmentation [12]. According to the number of stages involved in instance segmentation, fully supervised instance segmentation methods can be divided into two main categories: two-stage and single-stage methods [19].

Since Hariharan et al. proposed the first instance segmentation method based on object detection results [20], two-stage instance segmentation has always been dominant in the task of instance segmentation [14, 21]. In addition, according to the sequence of detection and segmentation, the two-stage instance segmentation can be further divided into top-down method and bottom-up method. The top-down detection-based method can share the good performance of the detector, but the upper bound of this method is also limited by the detector performance. Pinheiro et al. proposed the DeepMask method to predict candidate masks on different spatial regions through sliding windows on the feature map [21], but this method only uses the results of the full connection layer as the feature representation, and this process inevitably loses accurate low-level location information. In order to overcome this problem, researchers began to extract features from multiple network layers. Liu et al. proposed a simple path aggregation network (PANet) [22] to realize the mutual supplement of low-level location information and high-level semantic information. In addition, inspired by the success of the object detection method Faster R-CNN [23], He et al. extended the mask prediction Branch [14] on the existing boundary box regression branch, and proposed an instance segmentation method of Mask R-CNN. Mask R-CNN had shown unprecedented excellent performance in the task of instance segmentation. At the same time, because of its simplicity and robustness, it has gradually become the benchmark of the task of instance segmentation.

The bottom-up segmentation-based method is based on the great progress in semantic segmentation [24–26], and realizes the instance mapping of each pixel through post-processing such as clustering or metric learning on the existing semantic segmentation methods [27, 28]. For the results of semantic

segmentation, it is easy to think of a method to directly use the bounding box for clipping [29]. On this basis, shape branches and global branches to solve the disconnection and occlusion problems are also derived [30]. Uhrig et al. also designed a network that can output semantic labels, depth order and direction information at the same time [31]. The direction information and depth order are used to determine whether different pixels and disconnected regions belong to the same instance. It is worth mentioning that the bounding box information is not important for this processing mode, which is similar to the human recognition mode. However, due to the need for high-precision segmentation methods and post-processing methods with strong generalization ability, this mode is more difficult to achieve than the detection-based instance segmentation methods.

In order to further improve the segmentation results and make better use of the relationship between detection and segmentation, researchers also proposed some multi-stage instance segmentation methods [32, 33]. Cai et al. proposed Cascade R-CNN [33] under the idea of multi-task cascade, which surpassed Mask R-CNN [14] on the COCO dataset. In addition, Fang et al. [34] proposed a method that makes full use of the one-to-one correspondence between query objects and instances at different stages to construct effective information flow at different levels in a continuous cascade. Due to the good performance in object detection based on self-attention, the instance segmentation method based on self-attention [35] has also received more attention. ISTR is the first to realize the end-to-end instance segmentation model based on Transformer [18]. It extracts image features through FPN, then uses image features and learnable location information as the input of Transformer, and finally decodes the output using different heads to achieve classification, positioning and segmentation.

Although two-stage instance segmentation methods have achieved promising results, they still have the same problems as object detection or semantic segmentation methods, such as cumbersome post-processing methods, high computational cost, and slow speed, which make it difficult to be applied to tasks requiring high real-time performance [36]. At the same time, the step-by-step process of executing instance segmentation cannot fully correlate the mutual information between object detection and instance segmentation [37]. However, the single-stage method achieves a better trade-off between speed and accuracy [38, 39].

The starting point of the single-stage instance segmentation method is to realize the parallel operation of segmentation and detection, which can greatly reduce the reasoning time [40]. Single-stage instance segmentation can be further divided into anchor-based method and anchor-free method [19].

The single-stage anchor-based instance segmentation method draws on the idea of candidate regions in most advanced object detection algorithms, such as RetinaNet [41], YOLOv3 [38] and Faster R-CNN [23]. They use the network to generate a group of class-independent candidate score maps or masks on the candidate regions, and employ the parallel semantic branches to extract instances. Li et al. integrated the region proposal network (RPN) into the whole architecture to achieve full convolution instance segmentation (FCIS) [42]. Bolya et al. proposed the YOLACT method, which uses two parallel substructures to generate the prototype mask and mask type and position information, and then completes the instance segmentation by linearly combining the two [40].

Single-stage anchor-free instance segmentation methods mostly borrow from the idea of directly predicting the length of each position to four bounding boxes in FCOS [43]. Based on the FCOS architecture, EmbedMask extended the bounding box regression branch, directly predicts a learnable margin by comparing the distance between pixel embedding and proposal embedding, and assigns each pixel to a different instance [44]. On the basis of FCOS, Xie et al. treated instance segmentation as an extended task of object detection in polar coordinates by modifying the number of polar diameters, and changed the border regression branch into the mask regression Branch [45].

B. Intelligent Interpretation of HSI

With the development of remote sensing technology, HSI has recently become the focus of most researchers. HSI as an important information source in remote sensing, is a multidimensional data cube and provides rich spectral-spatial information. These images come from the wavelength range between visible light and short-wave infrared light. Usually, two dimensions are used to represent the spatial coordinates, and the third dimension is used to describe the spectral information of each pixel [46]. Therefore, each pixel contains representative characteristics of the captured material, which makes them become an excellent data source for remote sensing information interpretation [47]. In most real-life situations, the analysis of HSI is very important. As a result, it is commonly used in many real-life applications such as crop monitoring, object tracking, mining, land cover analysis, agriculture, military surveillance, land fire monitoring and astronomy [48–51].

HSI benefits from hundreds of contiguous spectral bands with richer image feature information. The upper limit of HSI in image interpretation is higher, but the high dimension of HSI and the information redundancy between adjacent spectral bands also lead to higher computational costs [52, 53]. At the same time, another challenge of HSI is the lack of input samples, and the lack of relevant datasets is a great obstacle to its development [54].

Limited by the dataset, the research on intelligent interpretation of HSI is mainly concentrated in the fields of image classification and semantic segmentation, with less research in the field of object detection, while no relevant research has been seen in the field of instance segmentation. In the image classification task, Chen et al. [55] first applied the concept of deep learning to the HSI field. At present, the methods of HSI classification mainly focus on the combination of 3D-CNN, 2D-CNN and self-attention mechanism. Roy et al. [56] proposed an end-to-end spectral-spatial feature compression and excitation method for HSI. This method creates a feature transformation form and deletes useless feature layers to facilitate classification. Firat et al. [48] proposed a method combining 3D-CNN and ResNet50 methods to deal with HSI.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3240481

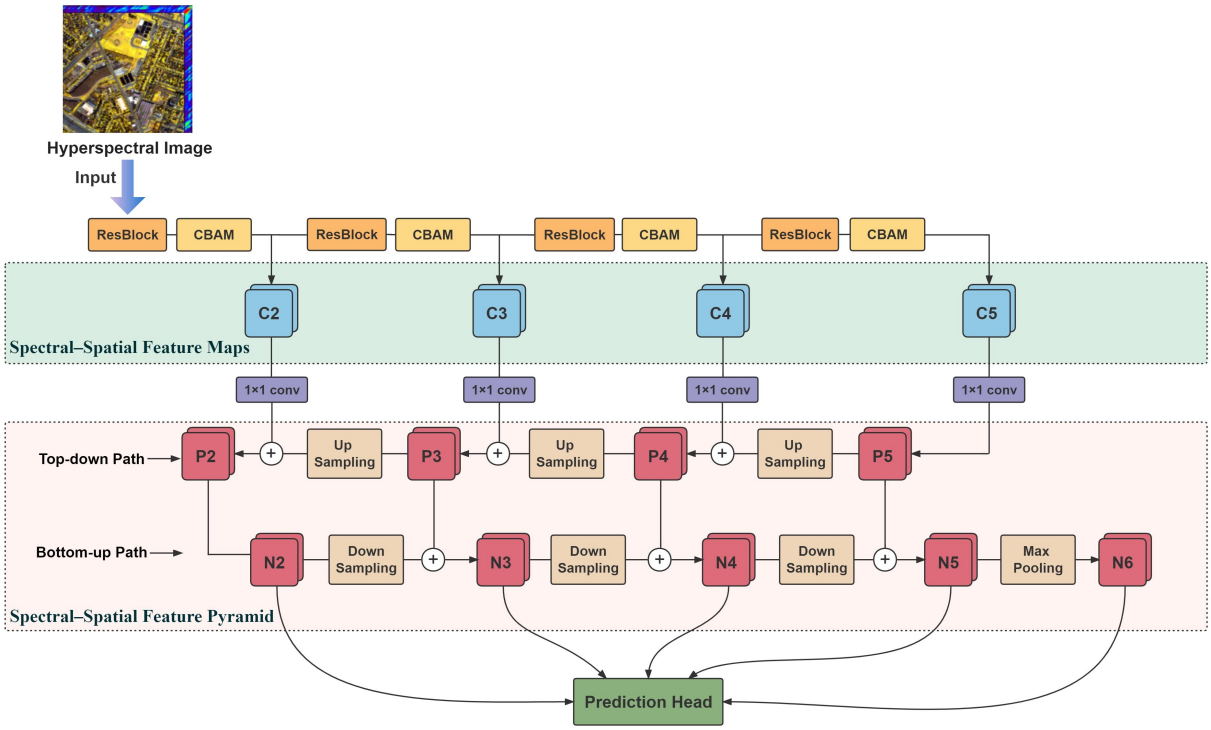


Fig. 2. The overview of the Spectral-Spatial FPN. It uses the spectral-spatial feature maps extracted by features in backbone to construct the spectral-spatial feature pyramid, which includes the top-down and bottom-up bidirectional feature fusion paths. The output features of the spectral-spatial feature pyramid are fed to the subsequent network for prediction.

Mohan et al. [57] proposed a hybrid CNN method for HSI, which consists of multi-scale spatial-spectral features extracted based on 3D-2D CNN. Ahmad [58] proposed a fast 3D-CNN method, which can use spatial and spectral features to improve the classification accuracy in HSI. Roy et al. [59] introduced a new multimodal fusion transformer network for HSI, which also utilizes other multimodal data sources in addition to HSI.

In the field of semantic segmentation of HSI, its low spatial resolution has always affected the effect of the segmentation model. In order to overcome this problem, some works have introduced expanded convolution [60], which will significantly increase the size of the filter and prevent parameter explosion by padding with zero. However, these models are susceptible to data distortion and mixed pixels [61], and data processing in the spectral band will also lead to high computational costs. In order to solve this problem, García et al. [62] proposed his segmentation model based on adaptive rectangular convolution (ARC). The model learns the size and offset of the kernel of the adaptive layer through convolution, and uses average operation to achieve smaller parameter quantities.

In addition, change detection is also an important application field of HSI. Change detection refers to the process of extracting the change region from multiple remote sensing images of the same region acquired at different times [63]. Li et al. [64] proposed an iterative method for detecting changes in multispectral images at different spatial scales. On this basis, Wang et al. also [65] proposed a change detection framework based on sub-pixel convolution, which can utilize HSI with different spatial resolutions, and the low resolution feature map is merged into the high-resolution feature map, solving the problem of resolution matching and making better use of the context of pixels.

III. Methods

A. Spectral-Spatial FPN

Since ordinary network models usually focus more on the spatial information of images during feature extraction, for HSI, the attention to their spectral dimension is missing and the spectral characteristics are lost. In Feature extraction of images, Spatial information and Spectral information can be regarded as complementary features. Therefore, we comprehensively consider Spatial multi-scale information and Spectral information and propose the Spectral-Spatial FPN.

The structure of the Spectral-Spatial FPN is shown in Fig. 2. First we perform dual feature extraction of spectral dimension and spatial dimension on the input HSI by the role of the convolution module and attention module for feature extraction, and obtain multi-scale spectral and spatial fusion feature maps. We introduce the convolutional block attention module (CBAM) [66] after each Resblock in Resnet. The CBAM includes two parts, the channel attention module and the spatial attention module, to perform the attention fusion of spectral and spatial features, which is a very lightweight module that does not creat excessive memory and computational overhead. We take out the multi-scale feature maps of spectral information and spatial information fusion for feature

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3240481

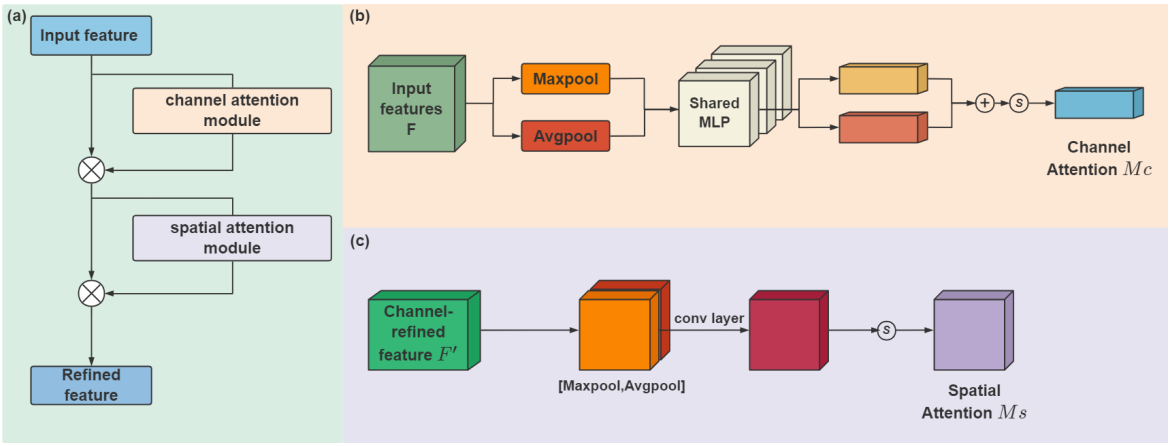


Fig. 3. Overall structure of CBAM module: (a) Overall process of CBAM module, (b) Overview of channel attention module, (c) Overview of spatial attention module.

pyramid structure construction for the next step of feature multi-scale fusion. In addition, the shallow features of the network are crucial for pixel-level classification tasks such as instance segmentation. In constructing the feature pyramid structure, we borrow the idea of PANet [22] and add a bottom-up structure by down sampling and max pooling operations to the top-down structure to build a bidirectional fusion pyramid structure, which can effectively shorten the information path and enhance the feature pyramid with the precise localization signals present in the lower levels.

CBAM is a simple and effective module, and its structure is shown in Fig. 3. The input features are sequentially passed through the channel attention module and the spatial attention module to enhance the useful information extracted in different dimensions. Channel attention module focuses on what is meaningful in the input image and which features on which channel are meaningful. The spatial attention module focuses on where is the most informative part, which complements the channel attention.

The channel attention mechanism can be expressed as follows:

$$M_c(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \tag{1}$$

where σ represents the sigmoid function. During the computation of the channel attention module, we use both average pooling and max pooling operations to aggregate spatial information. The shared network is composed of a weight sharing multilayer perception (MLP), and the core part of the Shared MLP is completed with 1×1 convolution to extract information.

Similarly, during the computation of the spatial attention module, the average pooling and max pooling operations are applied along the channel axis to obtain two feature maps, which are then concatenated based on the channel. Then, a 7×7 convolution operation is performed to reduce the number of channels to 1, and finally a sigmoid function is used to generate the spatial attention feature map ($M_s(F)$). The above process can be expressed as follows:

$$M_s\left(F'\right) = \sigma\left(f^{(7\times7)}\left(\mathrm{AvgPool}\left(F'\right); \mathrm{MaxPool}\left(F'\right)\right)\right) \tag{2}$$

After passing through the channel attention module and spatial attention module, the final feature map F'' with more fusion characteristics will be obtained. The entire CBAM process can be expressed as follows:

$$\begin{cases} F' = M_c(F) \otimes F \\ F'' = M_s\left(F'\right) \otimes F' \end{cases} \tag{3}$$

where $M_c(F)$ represents the channel attention map, $M_s(F)$ is the spatial attention map and $\otimes$ stands for element-wise multiplication.

The input HSI size is assumed to be $X^{H\times W\times B}$, where H and W are the height and width in the spatial dimension of the image, and B represents the number of channels in the spectral dimension of the image. In the backbone feature extraction network, multi-scale spectral-spatial feature maps can be obtained by Resblock and CBAM. We take out the results of the aspect compression twice ($C_2\frac{H}{4}\times\frac{W}{4}\times256$), three times ($C_3\frac{H}{8}\times\frac{W}{8}\times512$), four times ($C_4\frac{H}{16}\times\frac{W}{16}\times1024$), and five times ($C_5\frac{H}{32}\times\frac{W}{32}\times2048$) for the construction of the spectral-spatial feature pyramid architecture.

In the top-down structure of the feature pyramid architecture, the feature map obtained by upsampling operation on P_{i+1} is added to the feature map obtained by 1×1 convolution operation on C_i to obtain P_i. In the bottom-up structure, a shallower layer of N_i is fused with a deeper layer of P_{i+1} to get its next layer of N_{i+1}. Specifically, N_i is first downsampled by a 3×3 convolution operation with stride 2, and then added to each element of P_{i+1} through lateral connection to achieve an effective fusion of spectral and spatial features. Finally, a 3×3 convolution layer is used to operate on the fused feature map to obtain N_{i+1} and increase the characterization ability of the fused features. Among all feature maps fed to the prediction head, N_2 is obtained by copying the values of P_2, N_3, N_4, and N_5 are all obtained by the above fusion operation, and N_6 is obtained by max pooling operation of N_5.

Our proposed network module is loosely coupled and has strong adaptation and migration capabilities. The Spectral-Spatial FPN, which integrates spatial multiscale information

and spectral information, is capable of dual feature extraction for HSI, which is conducive to improving the learning ability of the network model and better meets the need for instance segmentation of HSI.

B. HSI Instance Segmentation

From the current research timeline, instance segmentation techniques can currently be grouped into two main categories based on the processing: two-stage instance segmentation model and single-stage instance segmentation model. Two-stage instance segmentation model generates candidate regions through common region selection methods such as selective search and bounding boxes, followed by feature extraction and regression classification in the candidate regions in combination with convolutional neural networks. The single-stage instance segmentation model directly distinguishes different instances for the image to be measured without adding the assistance of suggestion box, and saves the location and semantic information contained in different instances pixel by pixel to directly predict the location and region of different objects.

The framework of HSI segmentation using Spectral-Spatial FPN is shown in Fig. 4. In the two-stage instance segmentation task, the extracted [N_2, N_3, N_4, N_5, N_6] can be used as the effective feature layers of region proposal network (RPN), and the suggestion boxes are obtained using RPN. The extracted [N_2, N_3, N_4, N_5] can be used as the effective feature layers of the classification head and the mask head. For the head architecture of the network, we closely follow the architecture proposed in mainstream instance segmentation work, which has been verified to have good performance. The classification head is used to perform the next step on the effective feature layers to decode the proposed boxes to obtain the final prediction boxes; the mask head is used to perform the next step on the effective feature layers to obtain the semantic segmentation results inside each prediction box. We define f_{class} to represent the prediction result of the category, f_{box} to represent the prediction result of the location of each instance object in the HSI, and f_{mask} to represent the prediction result of the region division of each instance. The specific operation process is as follows:

$$\begin{cases} y_{pro} = F_{RPN}\left[N_2, N_3, N_4, N_5, N_6\right] \\ y_{feat} = F_{ROI}\left[y_{pro}, N_2, N_3, N_4, N_5\right] \\ f_{class}, f_{box} = F_{ODH}\left(y_{feat}\right) \\ f_{mask} = F_{MGH}\left(y_{feat}\right) \end{cases} \tag{4}$$

where y_{pro} represents the suggestion boxes obtained from the RPN network (F_{RPN}), y_{feat} represents the features used by ROI Align (F_{ROI}) output for result prediction, and finally the result of the instance segmentation of the input HSI is obtained after y_{feat} is applied to the Object Detection Head (F_{ODH}) and Mask Generation Head (F_{MGH}).

The process of the single-stage instance segmentation model is similar to that of the two-stage HSI instance segmentation model, but it is much simpler. Although the two-stage instance segmentation method has high accuracy, it is difficult to achieve the speed of real-time segmentation. Since the single-stage instance segmentation framework has been proposed, its advantages of simplicity, flexibility and speed have greatly influenced the research of instance segmentation. Part of the single-stage instance segmentation does not include Object Detection Head, and f_{class} and f_{mask} are directly output by mask. The main process of single-stage instance segmentation is as follows:

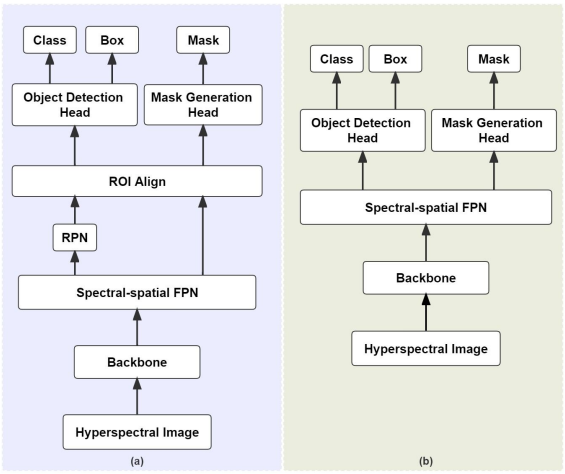


Fig. 4. Overview of HSI instance segmentation: (a) two-stage network model, (b) single-stage network model.

$$\begin{cases} f_{class}, f_{box} = F_{ODH}\left[N_2, N_3, N_4, N_5, N_6\right] \\ f_{mask} = F_{MGH}\left[N_2, N_3, N_4, N_5, N_6\right] \end{cases} \tag{5}$$

For the task of HSI instance segmentation, the multi-task loss function is defined as follows during training:

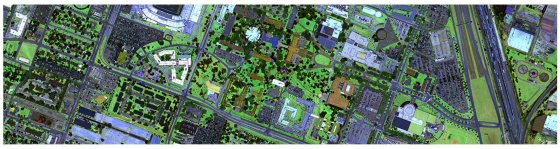
$$L = L_{cls} + L_{box} + L_{mask} \tag{6}$$

where L_{cls} represents classification loss, L_{box} represents bounding-box loss, and L_{mask} represents segmentation mask loss (part of single-stage instance segmentation methods do not include L_{box}). The definition of L_{mask} allows the network to generate masks for every class without competition among classes. This decouples mask and class prediction and is different from common practice when applying to semantic segmentation. Several experimental results from previous work have proved that loss function is the key for good instance segmentation results [14]. Our subsequent experiments show that this loss function is also applicable to HSI instance segmentation.

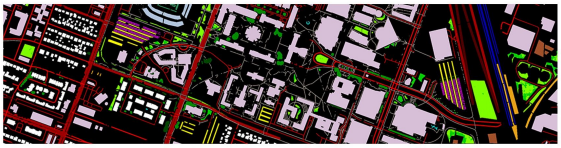
IV. EXPERIMENT AND RESULTS

A. HS-ISD Dataset

The work on HSI classification based on deep learning [67] is in a popular stage, but there is no existing research works on HSI instance segmentation, which limits the development of HSI research to some extent. Buildings are an important part of urban functional zoning and the main carrier of human social production and life. Building instances extraction from remote sensing images has become a popular research



(a)



(b)

Fig. 5. DFC2018 Houston dateaset. (a) False-color map. (b) The ground truth for the training region released for the contest.

direction, which is of great significance for urban planning and disaster assessment [68, 69]. In addition, compared with other ground objects, the structure of buildings is simpler and easier to label in the production of dataset. Based on the above, the research object in this paper is the building. The construction of the building-based HSI instance segmentation benchmark dataset contributes to the intelligent interpretation of HSI more directly and effectively applied to urban planning and application. In order to effectively promote the research of HSI instance segmentation, we construct the benchmark dataset HS-ISD.

DFC2018 Houston is the dataset used for the 2018 IEEE GRSS Data Fusion competition [70]. The data in this dataset was acquired by NCALM on February 16, 2017 between 16:31 and 18:18 GMT from the National Center for Airborne Laser Mapping, with data collected on and around the University of Houston campus. The sensors used for data collection in this competition include the OpTech TITAM M (14SEN/CON340), a lidar sensor with 3 different bands, the Dimac ULTRA-LIGHT+, a high resolution color imager with 70 mm focal length, and the ITRES CASI 1500, a hyperspectral imager. Multispectral lidar point cloud data in 1550 nm, 1064 nm, and 532 nm bands. The hyperspectral data cover a range of 380-1050 nm, with 48 bands and a spatial resolution of 1 m (in addition to the 48 bands containing hyperspectral data, there are also NADIR Channel and DEM Channel). The high-resolution RGB remote sensing images have a spatial resolution of 5 cm and are segmented into several individual images. The false-color and ground-truth maps are shown in Fig. 5 (a) and (b), respectively.

We choose the DFC2018 Houston used in the 2018 IEEE GRSS Data Fusion competition for the hyperspectral instance segmentation dataset, which has a dimension size of 4172×1202 and contains 48 channels in total. In the annotation, we take Google Map as an important reference basis and use ENVI software to annotate the remote sensing building instances provided in DFC2018 Houston that are segmented into 14 individual high-resolution RGB remote sensing image blocks. For the convenience of annotation and the feasibility of subsequent experiments, we cut each RGB remote sensing image into four pieces on average from the center, and the number of images is increased to 56 pieces accordingly. The corresponding HSIs are also cut into 56 hyperspectral mini-images of the same size. In the annotation, each building is marked as an individual structure, and a total of 1085 building examples are marked, including teaching buildings, residential buildings, commercial buildings, factories, etc. We generate a serial number label for each building instance, such as building No. 1, building No. 2, etc. All annotations have been repeatedly verified and corrected by multiple people to ensure the quality of annotations.

The shape file obtained by ENVI software after labeling holds the location and related properties of the building instances, but it cannot be directly applied to the subsequent experiments and operations, and needs to be transformed into the JSON file in the form of coco through geographic coordinate transformation and file format conversion. The markup file mainly includes the image ID number, category ID number and the coordinate position of the positioning border of each building instance.

Our final dataset consists of HSIs cut into 56 small images, each with a size of 298×301 and a number of 48 channels, as well as remote sensing building example annotation files corresponding to each image. According to the ratio of 7:3 we divide the dataset into a training set and a test set, i.e., 40 images for training and 16 images for testing, and HS-ISD will be used for subsequent experiments and applications. Several example diagrams of HS-ISD are shown in Fig. 6.

We classify HS-ISD in different dimensions, namely building area and building complexity in Fig. 7. In the division of building area, we refer to the definition of object in COCO dataset [71], and consider the building pixel area less than 32×32 as small building, the building pixel area more than 96×96 as large building, and the building pixel area between them as medium building. In the division of building complexity, we can consider buildings with more number of edges as more complex structures according to the labeling method of dotting and tracing edges in the labeling process. We consider those with less than 15 edges as simple buildings, and the rest as complex buildings.

In terms of the division schemes mentioned above, the vast majority of HS-ISD are small buildings (78%), followed by medium buildings (20%) and large buildings (2%). In terms of building complexity, most of them are complex buildings (60%) and relatively few are simple buildings (40%). HSI instance segmentation task is very challenging in these cases.

B. Evaluation Metrics

In order to evaluate the performance of the proposed method, we use the typical index mAP of instance segmentation to reflect the accuracy of HSI instance segmentation, including four indexes: Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50}. The larger the value of each index is, the more accurate the result is and the better the segmentation effect is. In addition, we choose the typical metrics Parameters and GFLOPs to evaluate the model structure. Larger values of both indicate a larger number of model parameters and a more complex model, respectively.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3240481

Fig. 6. Several example diagrams of the HS-ISD, where each building is covered by different color masks to represent individual building instances.

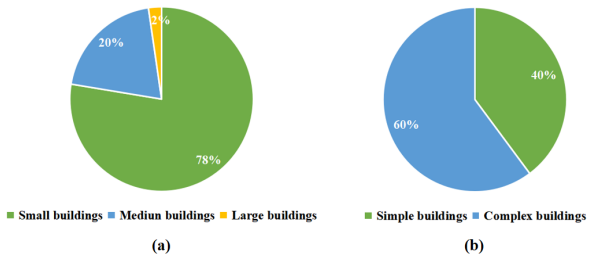


Fig. 7. The HS-ISD is divided according to different types of dimensions: (a) building area, and (b) building complexity.

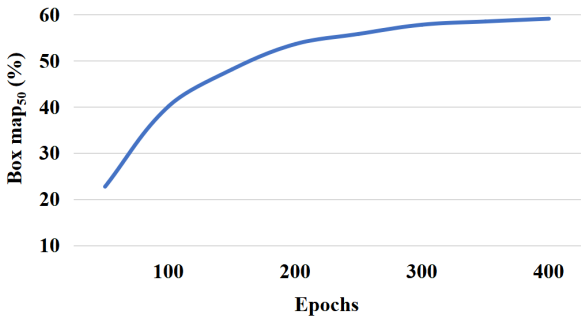


Fig. 8. The influence of different training epochs on the accuracy of test results.

C. Comparative Methods and Implementation Details

In order to implement the instance segmentation of HSI and verify the good performance of the proposed Spectral-Spatial FPN, we combine the existing typical instance segmentation methods on the brand-new HSI instance segmentation benchmark dataset HS-ISD. We choose the mainstream instance segmentation network models in recent years for our experiments, mainly including single-stage network models: SOLO [72], SOLOv2 [73] and PointRend [16], and two-stage network models: Mask R-CNN [14], Mask Scoring R-CNN [15], Cascade Mask R-CNN [33] and Queryinst [34]. All our experiments are done based on MMDetection [74], an open source object detection toolbox based on PyTorch provided by OpenMMLab platform.

Fig. 8 shows the effect of different training epochs on the

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3240481

TABLE I
RESULTS (%) OF DIFFERENT SINGLE-STAGE INSTANCE SEGMENTATION NETWORK MODELS ON THE HS-ISD

Method	Backbone	Box mAP	Box mAP_{50}	Mask mAP	Mask mAP_{50}	Parameters	GFLOPs
SOLO	Resnet50	-	-	19.5	43.1	35.89 M	54.61
SOLOv2	Resnet50	-	-	21.5	46.8	46.00 M	89.70
Pointrend	Resnet50	26.5	49.5	23.2	49.4	55.48 M	39.00
SOLO + Spectral-Spatial FPN	Resnet50	-	-	20.5	48.2	39.43 M	57.09
SOLOv2 + Spectral-Spatial FPN	Resnet50	-	-	22.1	49.6	49.54 M	92.18
Pointrend + Spectral-Spatial FPN	Resnet50	**27.6**	**56.3**	**24.5**	**55.1**	59.02 M	41.47
SOLOv2 + Spectral-Spatial FPN	Resnet101	-	-	26.9	56.5	68.54 M	99.79
Pointrend + Spectral-Spatial FPN	Resnet101	**30.1**	**59.3**	**27.7**	**57.1**	77.96 M	49.08

TABLE II
RESULTS (%) OF DIFFERENT TWO-STAGE INSTANCE SEGMENTATION NETWORK MODELS ON THE HS-ISD

Method	Backbone	Box mAP	Box mAP_{50}	Mask mAP	Mask mAP_{50}	Parameters	GFLOPs
Mask R-CNN	Resnet50	29.0	54.7	25.2	53.3	43.75 M	88.27
Mask Scoring R-CNN	Resnet50	31.1	56.4	24.7	50.8	60.01 M	88.27
Cascade Mask R-CNN	Resnet50	31.6	58.5	27.5	57.1	76.80 M	219.04
Queryinst	Resnet50	33.6	57.5	28.9	56.9	172.27 M	983.62
Mask R-CNN + Spectral-Spatial FPN	Resnet50	30.8	57.8	27.3	56.6	47.29 M	90.75
Mask Scoring R-CNN + Spectral-Spatial FPN	Resnet50	31.7	60.5	25.2	52.8	63.55 M	90.75
Cascade Mask R-CNN + Spectral-Spatial FPN	Resnet50	33.6	60.4	28.1	58.2	80.34 M	221.51
Queryinst + Spectral-Spatial FPN	Resnet50	**35.8**	**62.3**	**31.3**	**62.1**	175.81 M	986.09
Mask R-CNN + Spectral-Spatial FPN	Resnet101	31.3	58.5	27.5	56.3	66.28 M	98.36
Mask Scoring R-CNN + Spectral-Spatial FPN	Resnet101	32.4	60.7	28.4	58.6	82.54 M	98.36
Cascade Mask R-CNN + Spectral-Spatial FPN	Resnet101	34.4	62.4	29.7	62.2	99.33 M	229.12
Queryinst + Spectral-Spatial FPN	Resnet101	**37.4**	**64.3**	**31.9**	**62.5**	194.80 M	993.70

final performance. It is clear to see that setting the number of training epochs to 300 achieves the best balance between time and performance. In the training process, the batch size of each GPU and the number of data reading threads of each GPU are set to 4 in the training stage. We use SGD as the optimizer and set the base learning rate to 0.01, weight decay to 0.0001 and SGD momentum to 0.9. The total number of training epochs is set as 300 epochs, and the learning rate is decayed by step. The learning rate is reduced in the 240th and 270th epochs, respectively (by default, it is reduced to 0.1 times of the original each time). The warmup strategy is adopted to set a very small initial learning rate (0.001 times of the base learning rate), and it increases linearly in the initial iteration rounds of training. We use the two-stage and single-stage network models based on resnet50 as the baseline on the HS-ISD, and add the proposed Spectral-Spatial FPN to each model separately for a comprehensive comparison of the results with the baseline. The mmdetection version and PyTorch version used are 2.25.0 and 1.7.0, respectively. In order to accelerate the training of the network model, our experiments are conducted on NVIDIA GeForce RTX 3090.

D. Main Results

We report our experimental results with different instance segmentation network models on HS-ISD to prove that the proposed Spectral-Spatial FPN can perform more effective feature extraction and has better performance when performing instance segmentation on HSI. The experimental results of the two-stage network models and single-stage network models are shown in Table I and Table II, respectively.

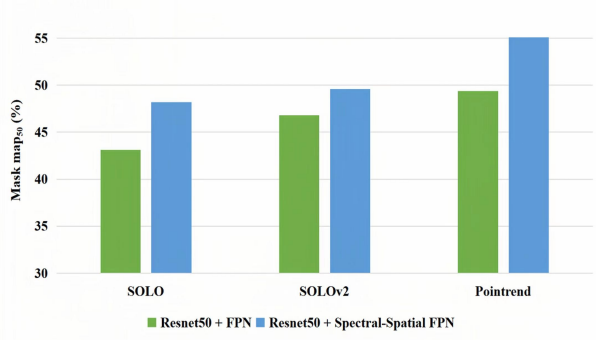


Fig. 9. Effect of Spectral-Spatial FPN on the results of single-stage instance segmentation methods.

For the single-stage instance segmentation network models, it can be seen from Table I that compared with the baseline after addition of Spectral-Spatial FPN, the Mask mAP and Mask mAP_{50} scores of SOLO increase by 1% and 5.1%, respectively; the Mask mAP and Mask mAP_{50} scores of SOLOv2 increase by 0.6% and 2.8%, respectively; and the Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} scores of Pointrend increase by 1.1%, 6.8%, 1.3% and 5.7%, respectively (the evaluation metrics for SOLO and SOLOv2 do not include Box mAP and Box mAP_{50}). Fig. 9 intuitively shows the effect of spectral-Spatial FPN on improving the result accuracy of the single-stage models. In addition, we modify

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3240481

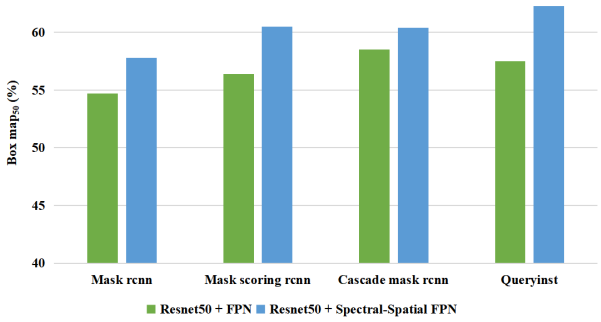


Fig. 10. Effect of Spectral-Spatial FPN on the results of two-stage instance segmentation methods.

the network models after adding Spectral-Spatial FPN and then conduct experiments, including adjusting the backbone of SOLOv2 and Pointrend to resnet101, the evaluation scores can be significantly improved. Pointrend + Spectral-spatial FPN with backbone of resnet101 achieves the single-stage optimal performance with evaluation scores Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} of 30.1%, 59.3%, 27.7% and 57.1%, respectively.

Table II shows the results of the different two-stage instance segmentation network models on the HS-ISD. Overall the accuracy of the two-stage network model results is better than the single-stage model results, but the model structure is more complex. When the backbone is all set to Resnet50, after adding Spectral-Spatial FPN, the Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} scores of Mask R-CNN increase by 1.8%, 3.1%, 2.1% and 3.3%,respectively; the Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} scores of Mask Scoring R-CNN increase by 0.6%, 4.1%, 0.5%, and 2%, respectively; and the Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} scores of Cascade Mask R-CNN increase by 2%, 1.9%, 0.6%, and 1.1%, respectively. The Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} scores of Queryinst + Spectral-Spatial FPN increased by 2.2%, 4.8%, 2.4% and 5.2%, respectively. Fig. 10 clearly shows that the result accuracy of different two-stage models with the addition of Spectral Spatial FPN is significantly improved. After modifying backbone to resnet101, the performance of the network model can basically be effectively improved, where Queryinst + Spectral-Spatial FPN with backbone of resnet101 achieves the two-stage optimal performance with evaluation scores Box mAP, Box mAP_{50}, Mask mAP and Mask mAP_{50} of 37.4%, 64.3%, 31.9%, and 62.5%, respectively.

It can be seen from the above quantitative experimental results that the integrated use of spectral information and spatial information can more effectively reflect the ground object information of HSI, distinguish buildings from other objects, and achieve higher precision segmentation results. The Spectral-Spatial FPN can significantly improve the instance segmentation accuracy of existing models on HSI without creating excessive memory and computing overhead.

The qualitative comparison results of the proposed Spectral-Spatial FPN are shown in Fig. 11. Spectral-Spatial FPN can reduce the omission of small buildings and achieve a more complete area segmentation in most cases. In addition, Spectral-Spatial FPN can achieve more detailed segmentation of building edge details (Row 2 of Fig. 11), and for building examples in close proximity, Spectral-Spatial FPN can also distinguish them well (Row 3 of Fig. 11). These visualization results validate the superiority of our Spectral-Spatial FPN in the task of HSI instance segmentation.

In summary, the Spectral-Spatial FPN has good adaptability and migration capability. Both the single-stage instance segmentation model and the two-stage instance segmentation model perform well on the HS-ISD with the addition of Spectral-Spatial FPN, which can lead to effective improvement of performance results.

E. Ablation Study

In the Spectral-Spatial FPN module, the CBAM module and PAN module are used to extract multi-scale spectral-spatial feature maps and fuse the utilization of multi-scale features for more accurate HSI instance segmentation, respectively. To verify the effectiveness of these two modules, we design several ablation experiments on Spectral-Spatial FPN to verify the effectiveness of CBAM module and PAN module. In the following experiment, the "Base" represents the basic model without CBAM and PAN. "Base + CBAM" and "Base + PAN" represent models with CBAM and PAN modules, respectively.

As shown in Table III, the combined use of the CBAM module and PAN module can significantly improve the model performance. More specifically, the addition of the CBAM module improves the Box mAP_{50} and Mask mAP_{50} scores on the HS-ISD by 2.6% and 3.3%, respectively, and the PAN module improves the Box mAP_{50} and Mask mAP_{50} scores on the HS-ISD by 2.8% and 2.5%, respectively. In addition, after combining the CBAM and PAN, we get 4.8% and 5.2% improvement in Box mAP_{50} and Mask mAP_{50} scores on the HS-ISD, respectively. These results prove the effectiveness of the CBAM module and PAN module, and show that the combination of CBAM module and PAN module can bring significant gain effect to HSI instance segmentation.

In addition, since the CBAM module includes two submodules, SAM and CAM, we also conduct several ablation experiments on the CBAM module. "Base + SAM" and "Base + CAM" represent the models with SAM and CAM modules, respectively. As can be seen from Table IV, the addition of SAM module increased Box mAP_{50} and Mask mAP_{50} scores by 0.6% and 1.8%, respectively, and the addition of CAM module increased Box mAP_{50} and Mask mAP_{50} scores by 1.4% and 1.7%, respectively on the HS-ISD. This indicates that both submodules of CBAM can enhance the ability of the model to extract spectral-spatial feature maps. In Fig. 12, we can clearly see that the model mask with CBAM module can better cover the target building area, indicating that the CBAM module can well utilize the information in the target building area and extract the spectral and spatial features.

V. CONCLUSION

In this paper, we introduced instance segmentation into the intelligent interpretation of HSI for the first time, and

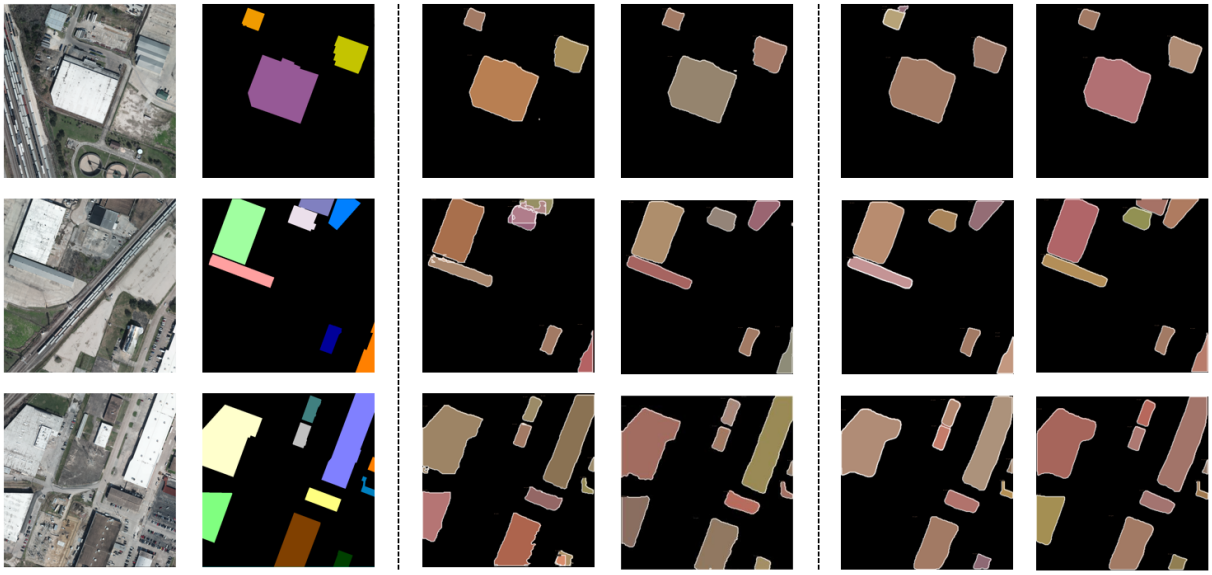


Fig. 11. Comparison of the visualization results of different methods on the HS-ISD, setting the fraction threshold for visualization to 0.3, and using different mask colors in each image to indicate different building instances. (a) RGB image. (b) Ground truth. (c) Pointrend. (d) Pointrend + Spectral-Spatial FPN. (e) Queryinst. (f) Queryinst + Spectral-Spatial FPN.

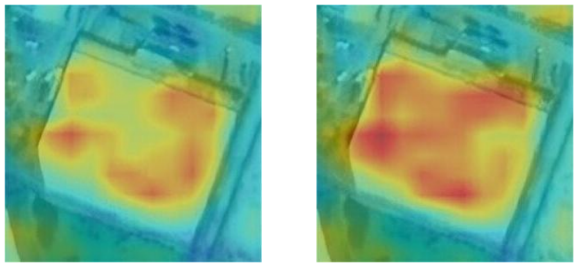


Fig. 12. Visual comparison of intermediate feature map between the CBAM module and the baseline: (a) Baseline, (b) Baseline + CBAM.

TABLE III
ABLATION STUDY (%) FOR THE PROPOSED SPECTRAL-SPATIAL FPN ON THE HS-ISD

Method	Box mAP_{50}	Mask mAP_{50}
Base	57.5	56.9
Base + CBAM	60.1	60.2
Base + PAN	60.3	59.4
Base + Spectral-Spatial FPN	**62.3**	**62.1**

TABLE IV
ABLATION STUDY (%) FOR THE CBAM MODULE ON THE HS-ISD

Method	Box mAP_{50}	Mask mAP_{50}
Base	57.5	56.9
Base + SAM	58.1	58.7
Base + CAM	58.9	58.6
Base + CBAM	**60.1**	**60.2**

implemented end-to-end HSI instance segmentation on the new benchmark dataset HS-ISD by using Spectral-Spatial FPN. HS-ISD takes different building examples as annotation objects, which can provide important support for the intelligent interpretation of HSI and urban planning. In addition, we proposed the Spectral-Spatial FPN, which first extracted the multi-scale spectral-spatial feature maps using the attention mechanism, and then performed a more effective feature fusion of spectral and spatial information of HSI through a bidirectional fused feature pyramid structure to achieve end-to-end instance segmentation of HSI. We conducted experiments on HS-ISD based on the mainstream two-stage and single-stage instance segmentation network models, and the experimental results demonstrate that the proposed Spectral-Spatial FPN can achieve state-of-the-art HSI instance segmentation results. The Spectral-Spatial FPN can effectively distinguish individual instances with similar structural information and achieve finer segmentation of edge details. In the future, we will further explore the field of instance segmentation of HSI and develop more effective HSI instance segmentation network models to meet the needs of more complex and diverse practical scenes.

REFERENCES

[1] S. K. Roy, G. Krishna, S. R. Dubey, and B. B. Chaudhuri, "Hybridsn: Exploring 3-d–2-d cnn feature hierarchy for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 2, pp. 277–281, 2019.

[2] N. Audebert, B. Le Saux, and S. Lefèvre, "Deep learning for classification of hyperspectral data: A comparative review," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 159–173, 2019.

[3] Z. Dong and B. Lin, "Bmf-cnn: an object detection method based on multi-scale feature fusion in vhr remote

sensing images," *Remote Sens. Lett.*, vol. 11, no. 3, pp. 215–224, 2020.

[4] X. Yang, H. Sun, X. Sun, M. Yan, Z. Guo, and K. Fu, "Position detection and direction prediction for arbitrary-oriented ships via multitask rotation region convolutional neural network," *IEEE Access*, vol. 6, pp. 50 839–50 849, 2018.

[5] S. Zhang, G. He, H.-B. Chen, N. Jing, and Q. Wang, "Scale adaptive proposal network for object detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 864–868, 2019.

[6] R. Kemker, C. Salvaggio, and C. Kanan, "Algorithms for semantic segmentation of multispectral remote sensing imagery using deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 60–77, 2018.

[7] X. Yuan, J. Shi, and L. Gu, "A review of deep learning methods for semantic segmentation of remote sensing imagery," *Expert Syst. Appl.*, vol. 169, p. 114417, 2021.

[8] J. Yue, L. Fang, P. Ghamisi, W. Xie, J. Li, J. Chanussot, and A. Plaza, "Optical remote sensing image understanding with weak supervision: Concepts, methods, and perspectives," *IEEE Geosci. Remote Sens. Mag.*, vol. 10, no. 2, pp. 250–269, 2022.

[9] S. Xia, S. Xu, R. Wang, J. Li, and G. Wang, "Building instance mapping from als point clouds aided by polygonal maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[10] S. Xia and R. Wang, "Extraction of residential building instances in suburban areas from mobile lidar data," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 453–468, 2018.

[11] S. Minaee, Y. Y. Boykov, F. Porikli, A. J. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2021.

[12] A. M. Hafiz and G. M. Bhat, "A survey on instance segmentation: state of the art," *Int. J. Multimedia Inf. Retrieval*, vol. 9, no. 3, pp. 171–189, 2020.

[13] J. Dai, K. He, Y. Li, S. Ren, and J. Sun, "Instance-sensitive fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 534–549.

[14] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2961–2969.

[15] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 6409–6418.

[16] A. Kirillov, Y. Wu, K. He, and R. Girshick, "Pointrend: Image segmentation as rendering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9799–9808.

[17] H. Chen, K. Sun, Z. Tian, C. Shen, Y. Huang, and Y. Yan, "Blendmask: Top-down meets bottom-up for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 8573–8581.

[18] J. Hu, L. Cao, Y. Lu, S. Zhang, Y. Wang, K. Li, F. Huang, L. Shao, and R. Ji, "Istr: End-to-end instance segmentation with transformers," *arXiv preprint arXiv:2105.00637*, 2021.

[19] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12 193–12 202.

[20] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous detection and segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 297–312.

[21] P. O. O Pinheiro, R. Collobert, and P. Dollár, "Learning to segment object candidates," *Proc. NIPS*, vol. 28, 2015.

[22] S. Liu, L. Qi, H. Qin, J. Shi, and J. Jia, "Path aggregatwork for instance segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2018, pp. 8759–8768.

[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Proc. NIPS*, vol. 28, 2015.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[25] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Appl. Soft Comput.*, vol. 70, pp. 41–65, 2018.

[26] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2015, pp. 3431–3440.

[27] A. Newell, Z. Huang, and J. Deng, "Associative embedding: End-to-end learning for joint detection and grouping," *Proc. NIPS*, vol. 30, 2017.

[28] X. Liang, L. Lin, Y. Wei, X. Shen, J. Yang, and S. Yan, "Proposal-free network for instance-level object segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 12, pp. 2978–2991, 2017.

[29] A. Arnab and P. H. Torr, "Bottom-up instance segmentation using deep higher-order crfs," *arXiv preprint arXiv:1609.02583*, 2016.

[30] Arnab, Anurag and Torr, Philip HS, "Pixelwise instance segmentation with a dynamically instantiated network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 441–450.

[31] J. Uhrig, M. Cordts, U. Franke, and T. Brox, "Pixel-level encoding and depth layering for instance-level semantic labeling," in *Proc. German Conf. Pattern Recog. (GCPR)*. Springer, 2016, pp. 14–25.

[32] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 4974–4983.

[33] Z. Cai and N. Vasconcelos, "Cascade r-cnn: high quality object detection and instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 5, pp. 1483–1498,

2019.

[34] Y. Fang, S. Yang, X. Wang, Y. Li, C. Fang, Y. Shan, B. Feng, and W. Liu, "Instances as queries," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 6910–6919.

[35] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 213–229.

[36] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact++: Better real-time instance segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2020.

[37] S. Wang, Y. Gong, J. Xing, L. Huang, C. Huang, and W. Hu, "Rdsnet: A new deep architecture forreciprocal object detection and instance segmentation," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, vol. 34, no. 07, 2020, pp. 12 208–12 215.

[38] J. Redmon and A. Farhadi, "Yolov3: An incremental improvement," *arXiv preprint arXiv:1804.02767*, 2018.

[39] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 21–37.

[40] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "Yolact: Real-time instance segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 9157–9166.

[41] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017, pp. 2980–2988.

[42] Y. Li, H. Qi, J. Dai, X. Ji, and Y. Wei, "Fully convolutional instance-aware semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 2359–2367.

[43] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, 2017, pp. 2117–2125.

[44] H. Ying, Z. Huang, S. Liu, T. Shao, and K. Zhou, "Embedmask: Embedding coupling for one-stage instance segmentation," *arXiv preprint arXiv:1912.01954*, 2019.

[45] E. Xie, P. Sun, X. Song, W. Wang, X. Liu, D. Liang, C. Shen, and P. Luo, "Polarmask: Single shot instance segmentation with polar representation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 12 193–12 202.

[46] A. F. Goetz, G. Vane, J. E. Solomon, and B. N. Rock, "Imaging spectrometry for earth remote sensing," *science*, vol. 228, no. 4704, pp. 1147–1153, 1985.

[47] S. Wang, J. Yue, J. Liu, Q. Tian, and M. Wang, "Large-scale few-shot learning via multi-modal knowledge discovery," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 718–734.

[48] H. Firat and D. Hanbay, "3b esa tabanlı resnet50 kullanılarak hiperspektral görüntülerin sınıflandırılması classification of hyperspectral images using 3d cnn based resnet50," in *Proc. 29th Signal Process. Commun. Appl. Conf. (SIU)*, 2021, pp. 6–9.

[49] J. Yue, L. Fang, and M. He, "Spectral-spatial latent reconstruction for open-set hyperspectral image classification," *IEEE Trans. Image Process.*, 2022.

[50] H. Firat, M. Uçan, and D. Hanbay, "Classification of hyperspectral remote sensing images using hybrid 3d-2d cnn architecture," *J. Comput. Sci., vol. IDAP-2021, no. Special*, pp. 132–140, 2021.

[51] Firat, H and Uçan, M and Hanbay, D, "Hyperspectral image classification using minivggnet," *J. Comput. Sci., vol. IDAP-2021, no. Special*, pp. 295–303, 2021.

[52] J. Yue, L. Fang, H. Rahmani, and P. Ghamisi, "Self-supervised learning with adaptive distillation for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[53] Y. Li, H. Zhang, and Q. Shen, "Spectral–spatial classification of hyperspectral imagery with 3d convolutional neural network," *Remote Sens.*, vol. 9, no. 1, p. 67, 2017.

[54] J. Yue, D. Zhu, L. Fang, P. Ghamisi, and Y. Wang, "Adaptive spatial pyramid constraint for hyperspectral image classification with limited training samples," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.

[55] Y. Chen, Z. Lin, X. Zhao, G. Wang, and Y. Gu, "Deep learning-based classification of hyperspectral data," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 7, no. 6, pp. 2094–2107, 2014.

[56] S. K. Roy, S. Chatterjee, S. Bhattacharyya, B. B. Chaudhuri, and J. Platoš, "Lightweight spectral–spatial squeeze-and-excitation residual bag-of-features learning for hyperspectral classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5277–5290, 2020.

[57] A. Mohan and M. Venkatesan, "Hybridcnn based hyperspectral image classification using multiscale spatiospectral features," *Infrared Phys. Technol.*, vol. 108, p. 103326, 2020.

[58] M. Ahmad, A. M. Khan, M. Mazzara, S. Distefano, M. Ali, and M. S. Sarfraz, "A fast and compact 3-d cnn for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, 2020.

[59] S. K. Roy, A. Deria, D. Hong, B. Rasti, A. Plaza, and J. Chanussot, "Multimodal fusion transformer for remote sensing image classification," *arXiv preprint arXiv:2203.16952*, 2022.

[60] B. Pan, X. Xu, Z. Shi, N. Zhang, H. Luo, and X. Lan, "Dssnet: A simple dilated semantic segmentation network for hyperspectral imagery classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 11, pp. 1968–1972, 2020.

[61] J. M. Bioucas-Dias, A. Plaza, N. Dobigeon, M. Parente, Q. Du, P. Gader, and J. Chanussot, "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 5, no. 2, pp. 354–379, 2012.

[62] J. L. García, M. E. Paoletti, L. I. Jiménez, J. M. Haut, and A. Plaza, "Efficient semantic segmentation of hyperspectral images using adaptable rectangular convolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[63] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, "A

review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 2, pp. 140–158, 2019.

[64] X. Li, F. Ling, G. M. Foody, and Y. Du, "A superresolution land-cover change detection method using remotely sensed images with different spatial resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 7, pp. 3822–3841, 2016.

[65] L. Wang, L. Wang, H. Wang, X. Wang, and L. Bruzzone, "Spcnet: A subpixel convolution-based change detection network for hyperspectral images with different spatial resolutions," *IEEE Trans. Geosci. Remote Sens.*, 2022.

[66] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 3–19.

[67] J. Yue, W. Zhao, S. Mao, and H. Liu, "Spectral–spatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sens. Lett.*, vol. 6, no. 6, pp. 468–477, 2015.

[68] J. Kang, M. Körner, Y. Wang, H. Taubenböck, and X. X. Zhu, "Building instance classification using street view images," *ISPRS J. Photogramm. Remote Sens.*, vol. 145, pp. 44–59, 2018.

[69] S. Ji, Y. Shen, M. Lu, and Y. Zhang, "Building instance change detection from large-scale aerial images using convolutional neural networks and simulated samples," *Remote Sens.*, vol. 11, no. 11, p. 1343, 2019.

[70] Y. Xu, B. Du, L. Zhang, D. Cerra, M. Pato, E. Carmona, S. Prasad, N. Yokoya, R. Hänsch, and B. Le Saux, "Advanced multi-sensor optical remote sensing for urban land use and land cover classification: Outcome of the 2018 ieee grss data fusion contest," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 12, no. 6, pp. 1709–1724, 2019.

[71] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2014, pp. 740–755.

[72] X. Wang, T. Kong, C. Shen, Y. Jiang, and L. Li, "Solo: Segmenting objects by locations," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 649–665.

[73] X. Wang, R. Zhang, T. Kong, L. Li, and C. Shen, "Solov2: Dynamic and fast instance segmentation," *Proc. NIPS*, vol. 33, pp. 17 721–17 732, 2020.

[74] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C. C. Loy, and D. Lin, "MMDetection: Open mmlab detection toolbox and benchmark," *arXiv preprint arXiv:1906.07155*, 2019.

Leyuan Fang (S'10-M'14-SM'17) received the Ph.D. degree from the College of Electrical and Information Engineering, Hunan University, Changsha, China, in 2015.

From August 2016 to September 2017, he was a Postdoc Researcher with the Department of Biomedical Engineering, Duke University, Durham, NC, USA. He is currently a Professor with the College of Electrical and Information Engineering, Hunan University. His research interests include sparse representation and multi-resolution analysis in remote sensing and medical image processing. He is the associate editors of IEEE Transactions on Image Processing, IEEE Transactions on Geoscience and Remote Sensing, IEEE Transactions on Neural Networks and Learning Systems, and Neurocomputing. He was a recipient of one 2nd-Grade National Award at the Nature and Science Progress of China in 2019.

Yifan Jiang received the B.S. degree from Xiangtan University, Xiangtan, China, in 2021. He is pursuing the M.S. degree with the College of Electrical and Information Engineering, Hunan University, Changsha, China.

His research interests include hyperspectral image processing, instance segmentation, and deep learning.

Yinglong Yan received the B.S. degree from China University of Petroleum (East China), Qingdao, China, in 2022. He is pursuing the M.S. degree with the College of Electrical and Information Engineering, Hunan University, Changsha, China.

His research interests include remote sensing image processing, semantic segmentation, and instance segmentation.

Jun Yue received the B.Eng. degree in geodesy from Wuhan University, Wuhan, China, in 2013 and the Ph.D. degree in GIS from Peking University, Beijing, China, in 2018.

He is currently an Assistant Professor with the Department of Geomatics Engineering, Changsha University of Science and Technology. His research interests include satellite image understanding, pattern recognition, and few-shot learning. Dr. Yue serves as a reviewer for IEEE Transactions on Image Processing, IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Geoscience and Remote Sensing, ISPRS Journal of Photogrammetry and Remote Sensing, IEEE Geoscience and Remote Sensing Letters, IEEE Transactions on Biomedical Engineering, Information Fusion, Information Sciences, etc.

Yue Deng received the B.E. degree (Hons.) in automatic control from Southeast University, Nanjing, China, in 2008, and the Ph.D. degree (Hons.) in control science and engineering from the Department of Automation, Tsinghua University, Beijing, China, in 2013.

He is currently a Faculty with the School of Astronautics, Beihang University, Beijing, China. His current research interests include machine learning, signal processing, and computational biology.