Multispectral Semantic Segmentation for UAVs: A Benchmark Dataset and Baseline

Qiusheng Li, Hang Yuan, Tianning Fu, Zhibin Yu*, *Member, IEEE,* Bing Zheng, *Member, IEEE,* Shuguo Chen *Member, IEEE*

***Abstract*—Solidago canadensis L. is a typical invasive plant that has become a significant threat worldwide and profoundly impacts local ecosystems. An unmanned aerial vehicle (UAV)-based semantic segmentation system can help in monitoring the spread and location of Solidago canadensis L. To identify the growth range of this species with greater efficiency, we employ a high-speed multispectral camera, which provides richer color information and features with limited resolution, in conjunction with a high-quality RGB camera to construct a segmentation dataset. We construct a validated UAV multispectral (UAVM) dataset comprising 3260 pairs of calibrated RGB and multispectral images. All the images in the dataset underwent semantic annotation at a fine-grained pixel level, with 12 categories being covered. In addition, other plant categories can be employed in precision agriculture and ecological conservation. Moreover, we propose a benchmark model, UAVMNet. With the aid of the feature alignment module and the UAVMFuse module, UAVMNet efficiently integrates multispectral and high-quality RGB image information, enhancing its ability to perform semantic segmentation tasks effectively. To the best of our knowledge, this is the first model to colearn semantic representations via high-quality RGB and paired multispectral information on a UAV platform. We conduct comprehensive experiments on the proposed UAVM dataset.**

***Index Terms*—Semantic segmentation; UAV; Multispectral images; Solidago canadensis L.; UAVM dataset; UAVMNet.**

I. Introduction

With technological advances and innovations, remote sensing platforms such as unmanned aerial vehicles (UAVs) and satellites can easily acquire high-resolution images over a large geographic area. These advances have further increased the progress of fields such as agricultural production [2] and urban planning [3] and triggered the need for an intelligent understanding (e.g., scenewise classification, object detection, segmentation, data enhancement, etc. [4]) of remote sensing images. Unlike object detection tasks with bounding boxes, semantic segmentation tasks need to distinguish the class of the object at the pixel level [5]–[13]. Compared with object detection or scenewise classification tasks, semantic segmentation tasks can describe the boundary and spread of the targets more precisely with pixelwise classification outputs. Thus, semantic segmentation tasks based on remote sensing images can be more suitable for precise monitoring and investigation tasks (e.g., invasive species localization) than identifying only bounding boxes.

Unlike satellites, UAVs can provide smaller but more precise observations in a specified area. Owing to their outstanding local observation capability, flexibility, and mobility, UAVs are widely used in smart agriculture [14], precision agriculture [15], etc. Although numerous datasets are available in the field of UAV-based segmentation [16]–[21], there are only a few datasets for invasive species [22]. Solidago canadensis L. is native to North America and has become an invasive plant worldwide. It has now spread to most European countries, Asia, Australia, New Zealand, and elsewhere, posing a severe threat to local biodiversity and ecosystem function [23], [24]. Therefore, monitoring its growth range is crucial for preventing and controlling this invasive species. UAVs provide a convenient way to measure the distribution of Solidago canadensis L. on a small scale. Unfortunately, no public UAV-based datasets correlate with Solidago canadensis L. thus far. These factors drive us to construct a UAV-based multicategory semantic segmentation dataset for Solidago canadensis L. and vegetation monitoring. As a supplement to Solidago canadensis L., we also label 11 other categories, such as pines and bushes, which can be essential for precision agriculture.

Although many deep learning-based segmentation methods can localize and characterize the distribution of vegetation with RGB images [1], [6], [25]–[29], the RGB camera may sometimes be insufficient to characterize the extent of the vegetation due to shadows and the environmental complexity of the vegetation itself. For example, the segmentation results of the objects in the orange box are poor if only the RGB camera is used (Fig. 1). In the early stages of growth, Solidago canadensis L. is typically relatively small and often intermixed with bushes or other similar vegetation. While it is relatively easy to distinguish on the ground, it becomes visually indistinguishable from the surrounding environment when viewed from a UAV's perspective. Additionally, the variation in light and shadow on Solidago canadensis L. in aerial imagery further complicates the segmentation of this

This work was supported by the Natural Science Foundation of Shandong Province of China under Grant No. ZR2021LZH005 and the National Natural Science Foundation of China under Grant No. 62171419.

Qiusheng Li, Hang Yuan, and Tianning Fu are with the College of Electronic Engineering, Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266520, China.

Zhibin Yu and Bing Zheng are with the College of Electronic Engineering, Faculty of Information Science and Engineering, Ocean University of China, Qingdao 266520, China, and with the Key Laboratory of Ocean Observation and Information of Hainan Province, Sanya Oceano graphic Institution, Ocean University of China, Sanya 572025, China (e-mail: yuzhibin@ouc.edu.cn).

Shuguo Chen is with the Key Laboratory of Ocean Observation and Information of Hainan Province, Sanya Ocean Institute/College of Marine Technology, Ocean University of China, Qingdao 266100 / Sanya 572024, China.

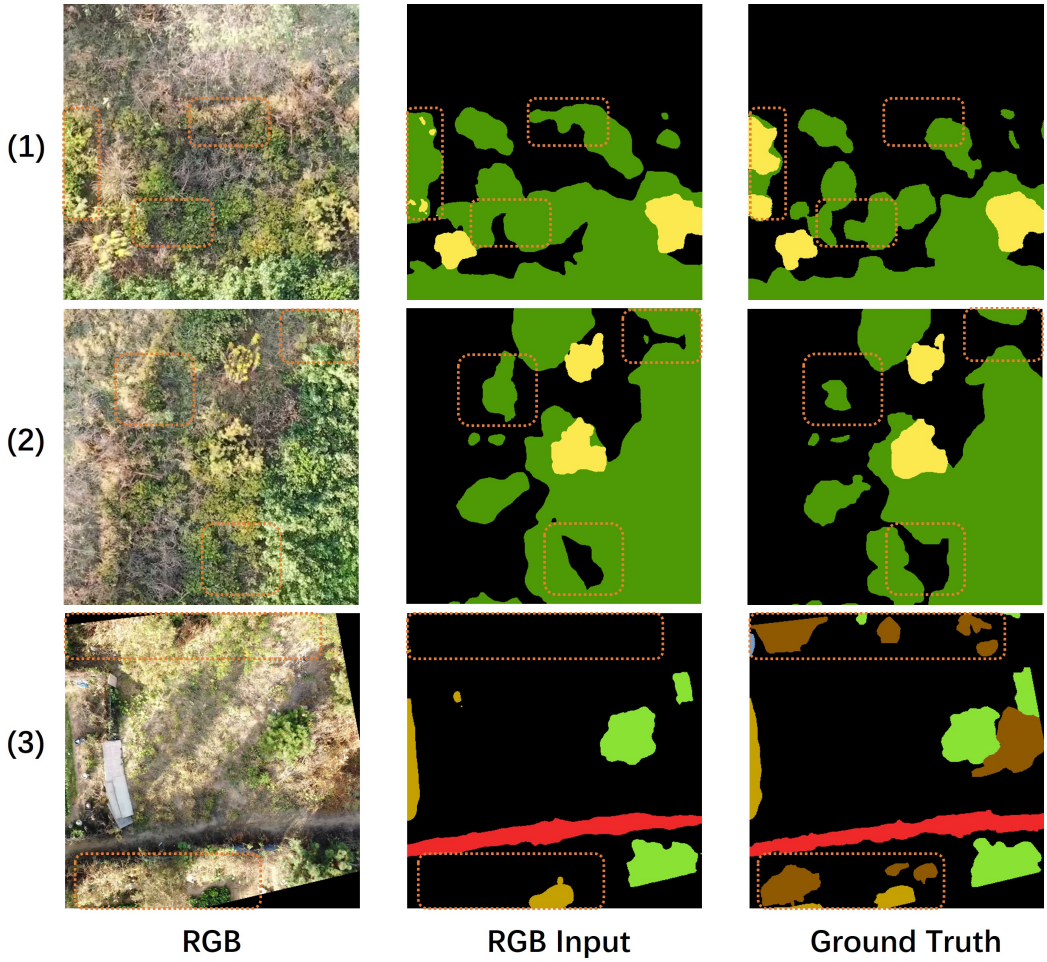

Fig. 1. Segmentation results with RGB inputs and the DeepLabv3plus [1] model. The first row from left to right shows the input RGB image, the visualized output of the RGB model, and the ground truth. The remaining rows show other images.

plant in RGB images using the current methods. In the first row, Solidago canadensis L. grows mixed with the bush, and their colors are similar, so the model fails to identify Solidago canadensis L. correctly. In the bush areas of the first and second rows, owing to the similarity in color between the weed and the bush, as well as the influence of shadows, the segmentation results are incorrect. In the third row, the distribution of reeds is more scattered and similar in color to the surrounding environment, which makes the model unable to correctly recognize reeds as vegetation types, resulting in poor segmentation results. Previous studies have shown that multispectral cameras with additional spectral bands can provide more detailed spectral information from different plant classes [30]. Although most multispectral cameras cannot offer high-quality imaging results with a high sampling rate simultaneously, multispectral cameras can collect more detailed spectral information in the visible range for complex vegetation distributions. These multispectral channels can help differentiate objects with similar colors. Thus, we curate a synchronous UAV multispectral semantic segmentation dataset called the UAVM dataset, which includes 3260 synchronized and calibrated high-quality RGB and multispectral image pairs. All the images are densely labeled with fine-grained semantic segmentation tags, constituting a rich collection of 12 object classes (Fig.2).

To simultaneously utilize the information provided by RGB images and multispectral images, we develop a specialized baseline model named the UAV multispectral semantic segmentation network (UAVMNet). The key to our design of UAVMNet is effectively synthesizing high-quality RGB and multispectral information. Considering the diversity of multispectral and high-quality RGB details, we design a two-branch model and develop a novel UAVMFuse module with a cross-spectrum attention mechanism to handle rich multispectral information with a moderate memory footprint. Moreover, to align the spectral information and promote the model to utilize both RGB and multispectral information, we propose a feature alignment module and further introduce a new UAVM alignment loss. We conduct comprehensive experiments on various state-of-the-art semantic segmentation models on the UAVM dataset. The experimental results validate the effectiveness of our UAVMNet model and shed light on the importance of the composite utilization of RGB and multispectral information in semantic segmentation.

Overall, the main contribution of this paper consists of three aspects:

- We present a synchronous UAV multispectral semantic segmentation dataset containing 3260 synchronized high-quality RGB and multispectral image pairs, which can be used to monitor the spread of the invasive species Solidago canadensis L. and can be applied to precision agriculture.
- We propose a new two-branch network architecture to establish correlations across spectra. To effectively learn to use high-quality RGB and multispectral information, we introduce the UAVMFuse module and the feature alignment module.
- We conduct comprehensive experiments on various state-of-the-art semantic segmentation models on the proposed UAVM dataset. The results show that our model is ahead of all other models in terms of performance.

II. Related Work

In this section, we review the most relevant literature on semantic segmentation datasets on remote sensing platforms, RGB semantic segmentation, multispectral semantic segmentation, and remote sensing hyperspectral image classification.

A. Semantic Segmentation Datasets on Remote Sensing Platforms

Semantic segmentation datasets on remote sensing platforms can be roughly categorized into three types: those based on RGB images, those based on multispectral images, and those based on hyperspectral images. The semantic segmentation datasets based on RGB images include Aeroscapes [16], UAVid [17], and the ICG Drone Dataset [18]. For example, the Aeroscapes [16] aerial semantic segmentation benchmark consists of images captured via a UAV in the altitude range of 5 to 50 meters. The dataset provides 3269 720p images and 12 categories of true masks. The datasets based on multispectral imagery include ISPRS 2D [31] and VddNet [19]. The ISPRS 2D [31] dataset contains urban aerial imagery with image sizes of 2000×2000 or 6000×6000 and provides pixel-level annotations with resolutions of 9 cm or 5 cm. To identify land cover classes, this semantic annotation benchmark defines six categories. The VddNet [19] dataset is designed to efficiently and accurately detect early grapevine diseases to inhibit disease spread in a timely manner, so only four categories are defined in the dataset. There are many datasets based on hyperspectral images, such as Houston [32] and WHU-Hi-LongKo [33]. The Houston [32] dataset provides hyperspectral imagery with a spatial resolution of 2.5 m and an image size of 349×1905, which contains 144 bands (with a range of bands from 364 to 1046 nm). The study area of this dataset covers 15 feature types, such as roads, land, trees, highways, etc. The WHU-Hi-LongKo [33] dataset uses a UAV platform, and its study area is a simple agricultural scenario containing six crops: maize, cotton, sesame, broad-leaf soybean, narrow-leaf soybean, and rice. The UAV flies at an altitude of 500 m, and the image size is 550×400, containing 270 bands in the 400 to 1000 nm range, with a spatial resolution of approximately 0.463 meters. However, none of these works consider invasive species. Unlike these datasets, the proposed UAV multispectral semantic segmentation dataset is a synchronous UAV multispectral semantic segmentation dataset that can be used to monitor the spread of the invasive species Solidago canadensis L.

B. RGB-based Semantic Segmentation

Since the release of public datasets such as Cityscape [7], significant progress has been made in RGB-based semantic segmentation. FCN [8] is a groundbreaking work that transforms fully connected layers into convolutional layers, which allows the network to predict dense, high-resolution outputs directly from inputs of any size. Subsequently, the idea of using only fully convolutional layers has been widely adopted to develop a series of CNN-based semantic segmentation methods [1], [6], [25]–[29]. Recently, transformer models that use the self-attention mechanism have become popular in the field of semantic segmentation [34]–[37]. While traditional CNN models can only gradually expand the field of view by increasing the receptive field layer by layer, transformer models are able to focus on all parts of an image in a single computation through the self-attention mechanism, which can better capture the global contextual information. For remote sensing platforms, MCAFNet [38] handles image details at low resolution by introducing a global-local transformer block (GLTB). A channel attention optimization module and a fusion module are used in the decoder to strengthen the model's ability to acquire small-scale semantic information. DenseU-Net [39] uses cascading operations to connect features extracted by the network and fuses shallow and deep features via a symmetric structure to improve small object classes and the overall accuracy. The ERN [40] uses two edge-loss enhancement modules to preserve spatial information, which in turn reduces semantic ambiguity and reduces shadow interference. UnetFormer [41] develops a global-local attention mechanism to model local and global information. The model runs faster and has better segmentation. The clusterformer [42] contains a spatial channel feed-forward network (SC-FFN) and a cluster token mixer at the encoding stage to extract multiscale information and reduce interference information. This approach is effective in identifying pine wilt disease and improves accuracy. To intra-class variance in semantic segmentation, Genze [43] et al. propose a UAV weed segmentation dataset encompassing two classes and conduct comprehensive experiments on this dataset, with the standard deviation of the Dice Coefficient ranging from 0.0018 to 0.18. Li [44] et al. introduce a novel method for field cotton detection, enhancing the network's robustness and accuracy through unsupervised region generation and supervised semantic labeling prediction. In their experiments, the minimum standard deviation of Intersection over Union (IoU) is 3.8. PFT [45] utilizes cross-scale inter-query attention to exchange complementary information, with the maximum standard deviation of mean IoU (mIoU) in the ablation experiments reaching 0.7. PGNet [46] transmits long-range dependence and global context information to each pyramid-level feature, generating high-resolution feature maps with high-level semantic information to segment small and varying-sized objects. On the ISPRS Vaihingen dataset,

the standard deviation of mIoU is 0.3. Although significant progress has been made in image semantic segmentation, most of these models only consider RGB images as inputs, which restricts their ability to utilize multispectral data for segmentation tasks. Nevertheless, our proposed method is capable of utilizing both RGB and multispectral information for semantic segmentation tasks.

C. Multispectral Semantic Segmentation

The RGB and thermal (T) imaging multispectral semantic segmentation technique improves the robustness of the model by integrating multimodal visible and thermal image information. MFNet [47] and PST900 [48] are commonly used datasets with pairs of visible and infrared light images. On the basis of these two datasets, researchers have designed various RGBT models via different strategies for semantic segmentation of RGBT datasets. For example, Ha et al. [47] proposed a multispectral fusion network (MFNet) with two identical thermal and RGB image encoders and a decoder block. Sun [49] et al. proposed an RGB-thermal fusion network (RTFNet), which also uses two encoders and a decoder, but the outputs of the encoders are fused into the RGB encoder via elementwise summation of the thermal feature maps. PSTNet [48], on the other hand, uses a serial structure to train the RGB streaming network before the results are fed into the fusion stream along with thermal images. In addition to the models mentioned above, various other fusion strategies exist for integrating multispectral information. These include the domain adaptive approach [50], multilevel feature fusion strategy [51], and multimodal multistage fusion [52]. On the remote sensing platform BLASeNet [53], to obtain effective information from multispectral images, a 3D residual block is designed to encode spectral-spatial features in the encoding stage. Two matrices are used in the decoding stage to adaptively adjust the fusion ratio of high-level semantic and low-level detail features, improving the segmentation accuracy. The DSCNN [54] uses a two-branch network to combine the information from high-resolution panchromatic images and multispectral images, which are used to segment the built-up areas. PerceiverIO Conv3D [55] encodes multimodal features via 3D convolution and reduces the number of model parameters via PerceiverIO's cross-attention mechanism. VddNet [19] fuses RGB, near-infrared (NIR), and depth maps to monitor grape diseases. WeedMap [21] uses a sliding window to fuse RGB, red edge (RE), and NIR information to identify weeds. However, the number of channels (4 or 5) in most of these multispectral semantic segmentation datasets is limited. In contrast, our multispectral camera is able to collect more detailed spectral information, which significantly contributes to improving the segmentation results. Our method extracts advantageous features from more detailed spectral information and integrates them with RGB information, thereby significantly enhancing segmentation accuracy.

D. Remote Sensing Hyperspectral Image Classification

Remote sensing hyperspectral image classification (HSIC) is similar to the semantic segmentation task, where the goal is to assign unique labels to each pixel vector on the basis of the spectral or spatial properties of the HSI cube [56]. This technique has been widely used in environmental monitoring [57] and land use [58]. At present, the primary methods on the remote sensing platform for HSIC are based on deep networks, including methods based on stacked autoencoders (SAEs), convolutional neural networks (CNNs), and recurrent neural networks (RNNs). Among these methods, classifiers based on CNNs have a significant advantage in terms of accuracy. For example, Zhong [59] et al. proposed a 3D CNN with residual connectivity to improve the trainability and classification accuracy of the network. Mei [60] et al. designed two attention-based branches to extract discriminative feature vectors from both spatial and spectral aspects. The overall accuracy (OA) and average accuracy (AA) of both methods on the Indian Pines dataset exceed 98%. In comparison, the OA and AA of the SAE-based method are 85% and 86%, respectively, whereas the OA and AA of the RNN-based method are 91% and 82%, respectively. Obviously, CNN-based classifiers are significantly better than other classifiers. In recent years, significant research has been conducted on transformer-based frameworks. Zheng [61] et al. proposed a fast patchless global learning framework that enables faster patchless inference and learning from global spatial information, thus improving accuracy. Although hyperspectral images have a relatively high spectral resolution, they usually contain many redundant bands and are not available in real time. In contrast, our high-speed multispectral camera is capable of video acquisition with a high frame rate (50 Hz), which can be more efficient and flexible for invasive species monitoring. In addition, previous methods fail to integrate RGB and multispectral information effectively; hence, they are not suitable for multispectral semantic segmentation tasks.

III. UAVM Benchmark Dataset

In this section, we focus on describing the construction of the UAVM dataset and analyzing the statistical results.

A. Dataset Construction

Data collection. Our goal is to create a high-quality UAV multispectral semantic segmentation dataset. The dataset comprises RGB and multispectral image pairs with high-quality, dense annotations. We use a DJI UAV M300 for data acquisition, which is equipped with an RGB camera and a Silios Toucan T4 multispectral camera. The latter has ten channels, with wavelengths ranging from 400 nm to 900 nm. Qingdao West Coast Central Park is a pristine park that is in the early stage of being affected by Solidago canadensis L., which is crucial for prevention. The UAV captures all images at an altitude of 15 to 40 meters. The spatial resolutions of the RGB and multispectral images are 0.5167 cm to 1.378 cm and 0.5156 cm to 1.375 cm, respectively, at altitudes ranging from 15 to 40 meters, and the ratio of their spatial resolutions is approximately 1:1. We manually select high-quality images for the construction of our UAVM dataset. The dataset covers a complex suburban scene during the daytime and dusk, and we focus on the invasive species Solidago canadensis L. In

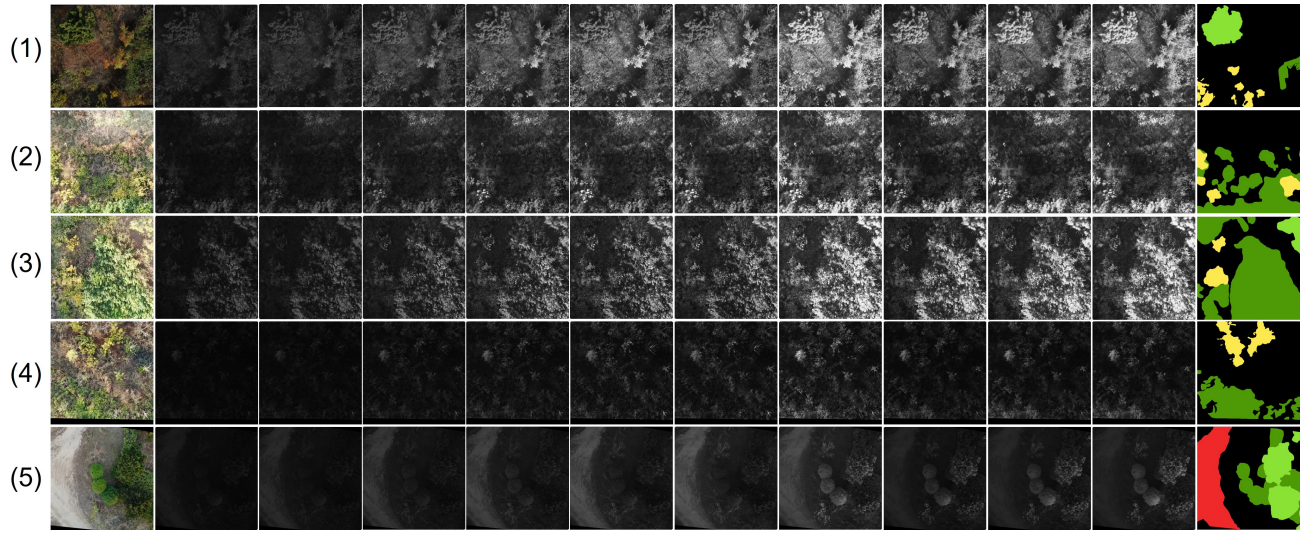

Fig. 2. Example images and labels from the UAVM dataset. From left to right are RGB images and multispectral 0-channel to 9-channel images, with corresponding ground truth labels.

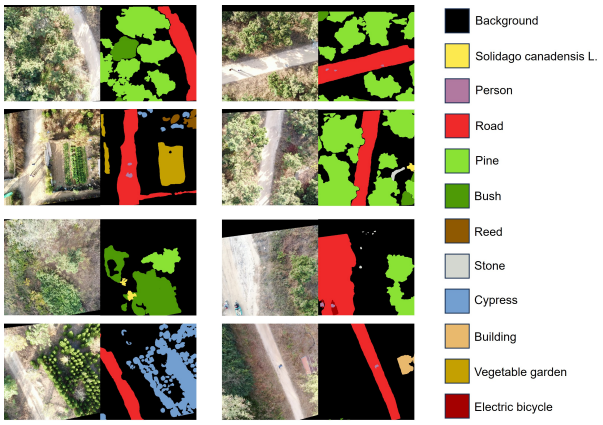

Fig. 3. On the left are examples of all categories in the UAVM dataset, and on the right are the visualization colors corresponding to each category.

Fig. 2, we show examples of RGB and multispectral images; in Fig. 3, we show examples of all the categories in the UAVM dataset and the colors corresponding to the labeling of each category.

Classes and annotations. To identify the categories to be labeled, we follow the guidance of [16] and consider the relevance of each category to the application, the frequency of occurrence, and the difficulty of the actual labeling work. We carefully checked the RGB and multispectral videos. We selected 12 categories that are worthy of annotation, including Solidago canadensis L., person, road, pine, bush, reed, stone, cypress, building, vegetable garden, electric bicycle, and background (unlabeled pixels). Since RGB images and multispectral images have different resolutions (the native resolution of RGB images is 1920×1080 and the native resolution of multispectral images is 512×512), we must image align them. As shown in Fig. 4, we use the scale-invariant feature transform (SIFT) algorithm [62] for feature point extraction for RGB and multispectral images. Feature point matching is performed via a brute-force matcher (BF). We compute the transformation matrix between the images on the basis of the matched feature point pairs. Finally, we apply the transformation matrix to the image to be matched and crop the image to a resolution of 512×512 to align it with the reference image.

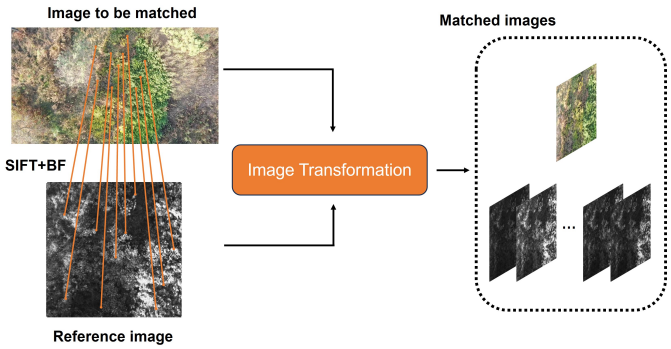

Fig. 4. Schematic of image matching via the SIFT algorithm.

The segment anything model (SAM) [63] has recently attracted much attention because of its strong zero-shot generalization capability. Therefore, to reduce the labeling effort, we adopt the interactive segmentation tool [64] based on the SAM [63] as an auxiliary tool in the annotation process. First, we roughly label the UAVM dataset via the interactive segmentation tool [64]. Subsequently, we manually adjust the boundaries for regions with inaccurate boundaries. Moreover, we adopt the adjacent image annotation approach, which ensures the consistency of annotations between neighboring images. Although the adjacent frame annotation step increases the annotation time, this approach makes it easier for the annotator to detect different annotations and unlabeled objects. Moreover, when labeling a specific frame, we have to refer to the video surrounding this frame, which can help the inspector determine the category of the labeled object by judging the height of the labeled object. For these reasons,

TABLE I
TOTAL PIXEL VOLUME (TPV), IMAGE SIZE, AND NUMBER OF IMAGE CHANNELS IN 8 DATASETS. B STANDS FOR BILLION, M STANDS FOR MILLION, AND PX MEANS PIXELS IN SIZE.

Dataset	Number of image channels	Image size	Number of classes	Volume	TPV
Aeroscapes [16]	3	1280×720Px	12	3269	9.0B
UAVid [17]	3	4096×2160 or 3840×2160Px	8	300	7.8B
ICG Drone Dataset [18]	3	6000×4000Px	20	600	43.2B
VddNet [19]	5	4626×3904Px	4	1	90M
MFNet [47]	4	640×480Px	9	1569	1.9B
PST900 [48]	4	1280×720Px	5	894	3.2B
Houston [32]	144	1905×349Px	15	1	95M
Xiongan New Area Dataset [65]	250	3750×1580Px	19	1	1.4B
UAVM (ours)	13	512×512Px	12	3260	11B

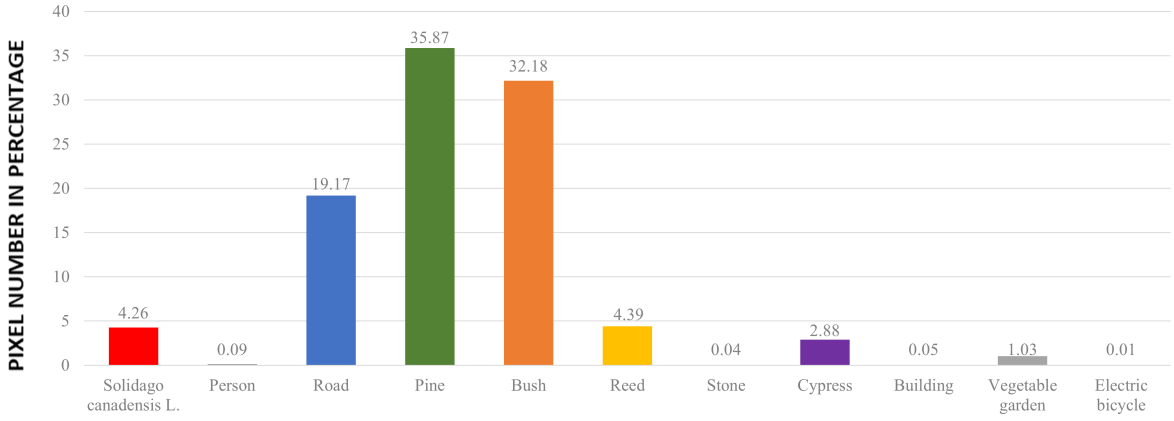

Fig. 5. Percentage of the number of pixels in each category (y-axis). The background category is not shown.

coupled with the complexity of challenging suburban scenes, annotating the UAVM dataset and quality control are very time-consuming. An average of thirty minutes per image is required for annotation.

Dataset splits. The UAVM dataset consists of 3260 synchronized and calibrated RGB and multispectral image pairs. To ensure that the training and validation sets have the same data distribution, we manually divided the dataset into training and validation sets, each consisting of 2704/556 annotated image pairs.

B. Statistical Analysis

Table I shows the number of images and the total pixel volume (TPV) for our UAVM dataset and several public remote sensing datasets (we consider only the source images and do not consider the data augmentation cases). Our UAVM dataset contains 3260 annotated image pairs of 12 categories. The hyperspectral dataset has a high number of image channels but a low number of TPVs due to its sampling rate. With respect to the UAV dataset, the number of TPVs in the ICG dataset [18] is greater than that in our UAVM dataset. This is because the multispectral camera we used can provide only 512×512 resolution for each channel, which is much less than the 6000×4000 resolution of the ICG dataset. Notably, these UAV datasets have fewer channels, which means that they contain less spectral information than our dataset does. The RGBT multispectral datasets [19], [47], [48] are similar to ours in terms of dataset type but are much smaller than our UAVM dataset in terms of dataset size. Fig. 5 shows the percentage of the number of pixels in each category of the UAVM dataset (excluding the background category). The proportions of the different categories in the UAVM dataset are imbalanced, similar to those in any other semantic segmentation dataset. Vegetation pixels account for a large portion of the UAVM dataset, and most of the pixels come from pines, bushes, and roads, with fewer pixels coming from people, buildings, and electric bicycles. The invasive species Solidago canadensis L. also has a high occupancy, which is consistent with its distributional characteristics.

IV. METHODOLOGY

A. Motivation

Various network models have been constructed for multispectral semantic segmentation [21], [47]–[49], [53], [54], [66]. These models focus on feature fusion techniques for

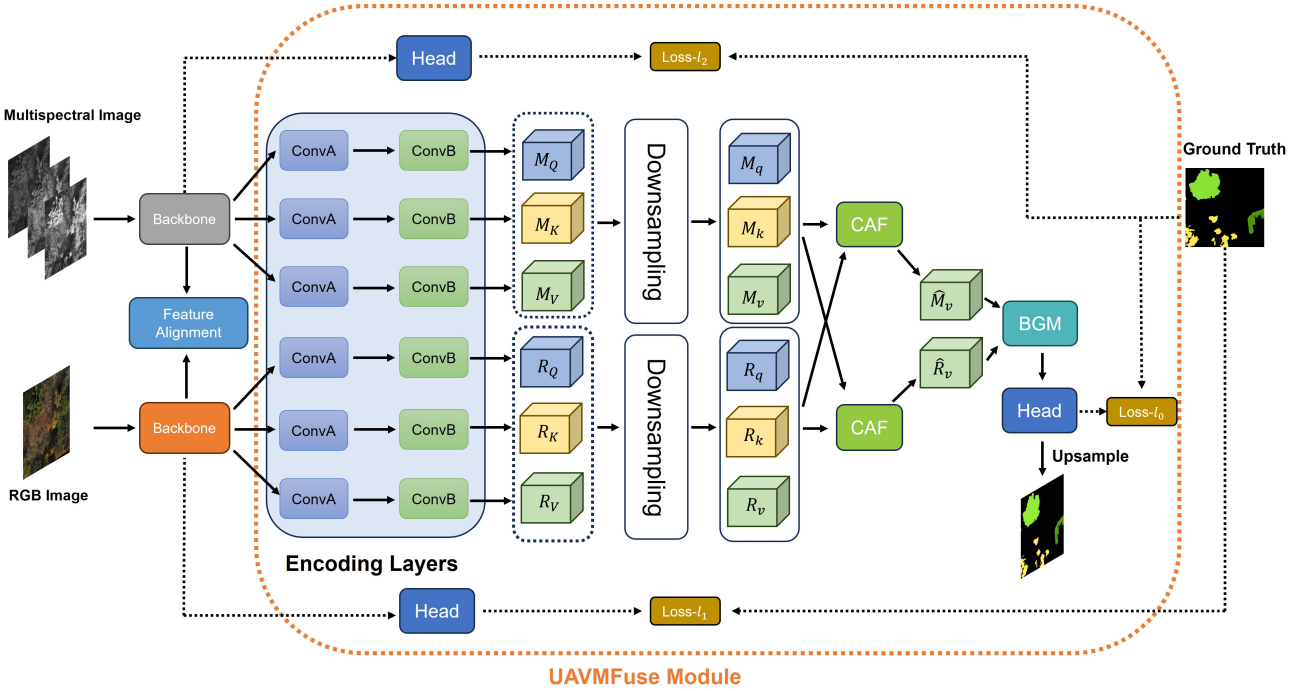

Fig. 6. The architecture of the proposed UAVMNet. The inputs consist of RGB and multispectral image pairs. UAVMNet contains a feature extraction module, a UAVMFuse module, and a feature alignment module. ConvA and ConvB are 1×1 convolution and 3×3 convolution, respectively.

RGB and NIR or IR, and a variety of excellent feature fusion methods have been devised. These models consider a two-branch structure that has shown significant advantages in improving segmentation accuracy and robustness by resorting to near-infrared or infrared imaging information. However, there are two specific problems with this generic scheme:

i) There is a lack of alignment between independent encoders. Using independent encoders may lead to disjointed and inconsistent extracted features when applied to dozens of pieces of spectral information, decreasing accuracy due to insufficient information fusion ability.
ii) There is an inherent difference between high-quality RGB and multispectral modalities. High-quality RGB data provide detailed visible appearance information and high-quality imaging. In contrast, multispectral methods can provide richer spectral information. This difference in modality results in feature embeddings of RGB and multispectral images being distributed in different embedding spaces, and suboptimal cross-spectral features affect the full utilization of cross-spectral complementarity. Therefore, addressing the modality difference problem is essential for effectively utilizing the complementary information of RGB and multispectral images.

In Section IV-B, we present a well-designed baseline called UAVMNet to address these problems. It solves the two problems mentioned above and thus enables better exploitation of RGB and multispectral information for semantic segmentation tasks.

B. Overall Architecture

We first illustrate the list of symbols and abbreviations mentioned in our paper in Table II to provide better readability.

TABLE II
LIST OF SYMBOLS AND ABBREVIATION.

Notation	Description
CE loss	Cross entropy loss
Q	Query maps generated by encoding layers
K	Key maps generated by encoding layers
V	Value maps generated by encoding layers
R	Feature map of RGB stream
M	Feature map of multi-spectra stream
CAF	Cross-spectrum attention fusion module
BGM	Balance guidance module
Add	Element-wise summation operation

RGB images are high quality, whereas multispectral images contain rich spectral information. Both sides have their strengths and weaknesses. To achieve complementary advantages, we first employ a feature alignment module that allows the RGB branch to learn additional spectral information from the multispectral branch. During the decoding stage, the RGB branch queries required information from the rich spectral information of the multispectral branch, while the multispectral branch queries required information from the high-quality RGB branch. Therefore, we construct comprehensive feature representations through this bidirectional querying process, achieving complementary advantages between the two modalities.

The network structure of the proposed UAVMNet is shown in Fig. 6. The structure contains three parts: the feature

extraction module; the feature alignment module for problem-solving i); and the UAVMFuse module for problem-solving ii).

Given the input RGB and multispectral image pairs, we feed them into their respective backbone networks for feature extraction. Moreover, we employ the feature alignment module to ensure consistency between the features and further enhance the fusion effect by calculating the UAVM alignment loss. Then, we feed the feature maps extracted from the two-branch backbone into the proposed UAVMFuse module for decoding. The UAVMFuse module fuses RGB and multispectral features via a cross-spectral attention mechanism and is able to generate the final segmentation maps.

The training process of our UAVMNet for multispectral semantic segmentation (MSS) is shown in Algorithm 1.

Algorithm 1 Training of UAVMNet for MSS

Input: $\{(\mathbf{I}_r^n, \mathbf{I}_m^n, \mathbf{Y})\}_{n=1}^{N_{data}}$, a dataset of sequence pairs.
Input: $\mathbf{\Theta}$, the initialized parameters.
Output: $\hat{\mathbf{\Theta}}$, the trained parameters.
for $i = 1, 2, ..., N_{epoches}$ **do**
  for $n = 1, 2, ..., N_{data}$ **do**
    Compute UAVMNet($\mathbf{I}_r^n, \mathbf{I}_m^n, \mathbf{\Theta}$) via forward propagation
    Compute l_{total} via loss function
    Update $\mathbf{\Theta}$ via backward propagation
  end
end
return $\hat{\mathbf{\Theta}} = \mathbf{\Theta}$

C. Feature Alignment

We propose a simple but effective alignment module for feature learning during training. Since we use two independent encoders with the same structure, both encoders include the same scale features. Thus, the feature alignment loss can easily match the features in the two encoders. As shown in Fig. 8, the feature alignment module convolves stage 3 and stage 4 features in the two backbones separately to reduce the number of channels. Then, UAVM alignment loss is applied to the features to align the channel information.

UAVM alignment loss. To align the spectral information, we need an alignment loss function. Inspired by channelwise distillation loss (CWD loss) [67], we further design a tailored UAVM alignment loss for the UAVMNet model. Formally, the UAVM alignment loss is defined as:

$$\Phi\left(\mathbf{X}_{i,j}\right) = \frac{\exp\left(\frac{\mathbf{X}_{i,j}}{T}\right)}{\sum_{j=1}^{C}\exp\left(\frac{\mathbf{X}_{i,j}}{T}\right)} \tag{1}$$

$$l_{Ali} = \frac{T^2}{S}\sum_{i=1}^{S}\sum_{j=1}^{C}\Phi\left(\mathbf{X}_{i,j}^m\right)\cdot\log\left[\frac{\Phi\left(\mathbf{X}_{i,j}^m\right)}{\Phi\left(\mathbf{X}_{i,j}^r\right)}\right] \tag{2}$$

where $i = 1, 2, ..., S$ represents the spatial location index and $j = 1, 2, ..., C$ denotes the channel index. T is a hyperparameter (temperature). The larger T is, the softer the probability distribution is. A larger T means that we are concerned with a wider range of channels for each spatial location. Following [67], we set the temperature $T = 4$. $\mathbf{X}^m$ and $\mathbf{X}^r \in \mathbb{R}^{S\times C}$ are the feature maps of the multispectral and RGB branches, respectively. $\Phi(\cdot)$ transforms feature activation into a probability distribution of spatial location directions. According to the above equation, if $\Phi(\mathbf{X}_{i,j}^m)$ is large, $\Phi(\mathbf{X}_{i,j}^r)$ should also be as large as $\Phi(\mathbf{X}_{i,j}^m)$ to minimize the loss function. However, when $\Phi(\mathbf{X}_{i,j}^m)$ is small, the loss function is less concerned with minimizing $\Phi(\mathbf{X}_{i,j}^r)$. Since the RGB channel has higher image quality and the multispectral channel contains richer color information, we design UAVM alignment loss to take advantage of multispectral information to support the RGB branch. This alignment mechanism forces the RGB branch to learn additional channel and global context information for the multispectral branch.

D. UAVMFuse Module

We have developed a UAVMFuse module in Fig. 6, which consists of encoding layers, downsampling, a cross-spectrum attention fusion module, and a balance guidance module. Each component is described below.

Encoding layers. Considering the backbone structure (e.g., ResNet-50), the dimension of the feature maps obtained after two independent backbone extractions can reach 2048, which is a large value. Therefore, feeding them directly into the decoder is both time-consuming and expensive. A suitable way to reduce the computational effort is to restrict the number of channels. Generally, a 1×1 convolution is a common solution for reducing the number of channels. However, a 1×1 convolution cannot capture the related features among neighboring pixels, which are crucial for segmentation tasks. Although a larger convolution kernel can capture spatial information among neighboring pixels, it increases the number of parameters and computational cost. Considering the above factors, inspired by the bottleneck in ResNet [68], we suggest an approach that balances performance and computational cost. We first reduce the channels via a 1×1 convolution. Then, a 3×3 convolution captures spatial information for spatial information coding. After encoding, to establish correlations across spectra, we generate query (Q), key (K), and value (V) maps corresponding to the RGB stream and the multispectral stream. The features of the RGB stream are denoted as $\mathbf{R}$, and the features of the multispectral stream are denoted as $\mathbf{M}$. We denote these features as $\mathbf{I} = \{\mathbf{R}, \mathbf{M}\}$.

Downsampling. The two branches are encoded to obtain the corresponding query, key, and value maps, and we use a simple but effective strategy to downsample the data. As shown in Fig. 6, we apply the maximum pooling operation $\psi_n(\cdot)$ with stride size n to the query, key, and value maps of the two branches to obtain the downsampled features $\psi_n(\mathbf{I}_Q), \psi_n(\mathbf{I}_K)$ and $\psi_n(\mathbf{I}_V)$:

$$\mathbf{I}_q = \psi_n(\mathbf{I}_Q), \quad \mathbf{I}_k = \psi_n(\mathbf{I}_K), \quad \mathbf{I}_v = \psi_n(\mathbf{I}_V) \tag{3}$$

Using the downsampling module, we can effectively reduce the computational complexity while preserving the key features. Our experiments have shown that $n = 2$ is the

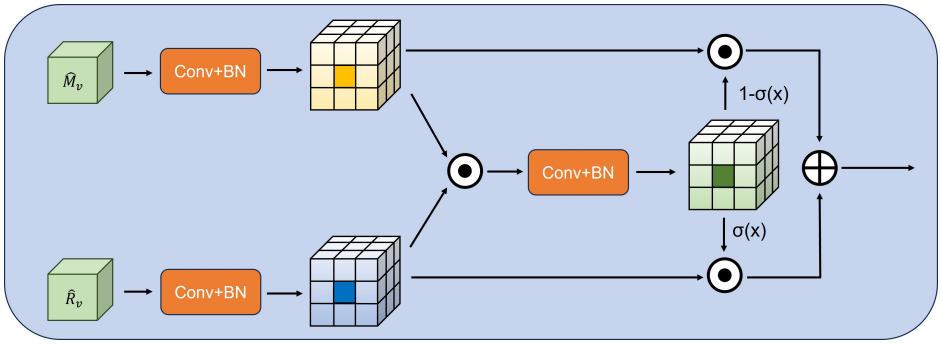

Fig. 7. Illustration of the balance guidance module. $\sigma(x)$ denotes the sigmoid function. Conv stands for a 1×1 convolution. BN stands for batch norm.

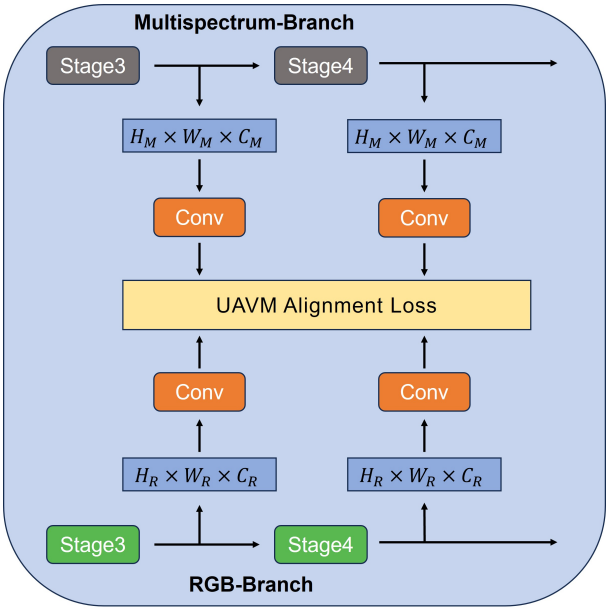

Fig. 8. Illustration of the feature alignment module. Conv stands for a 1×1 convolution.

most appropriate. Moreover, the spatial information can be well maintained while reducing the computational amount and improving the model's inference speed.

Cross-spectrum attention fusion module. Next, we propose a new approach to attentional fusion. In this approach, we combine the RGB and multispectral streams to construct a complete feature representation of the respective streams. Inspired by spatiotemporal attention [69], we compute the cross-spectrum attention information for the RGB and multispectral streams separately. For example, for the RGB stream, the attention is fused in the following way:

$$\hat{\mathbf{R}}_v = \boldsymbol{Softmax}\left(\frac{\mathbf{R}_q\mathbf{M}_k^T}{\sqrt{d_k}}\right)\mathbf{M}_v + \mathbf{R}_v \tag{4}$$

where $\mathbf{R}_q, \mathbf{R}_v, \mathbf{M}_k, \mathbf{M}_v$ represents the downsampling mapping features, as shown in Equation 3, and d_k represents the dimensions of the query and key. Similarly, the fusion of multispectral streaming attention can be obtained as:

$$\hat{\mathbf{M}}_v = \boldsymbol{Softmax}\left(\frac{\mathbf{M}_q\mathbf{R}_k^T}{\sqrt{d_k}}\right)\mathbf{R}_v + \mathbf{M}_v \tag{5}$$

where $\mathbf{M}_q, \mathbf{M}_v, \mathbf{R}_k, \mathbf{R}_v$ represents the downsampling mapping features, as expressed in Equation 3.

Next, we obtain the complete feature representations of the RGB and multispectral streams, named $\hat{\mathbf{R}}_v$ and $\hat{\mathbf{M}}_v$, respectively. The RGB stream can retrieve critical information from the multispectral stream through this elaborate attention mechanism. Moreover, the multispectral stream is able to retrieve the desired information from the RGB stream. This interactive attention mechanism allows the two streams to complement each other in the information exchange process and can effectively capture the nonlocal correlations between cross-spectral pixels.

Balance guidance module. We introduce the balance guidance module (BGM) to fuse the value maps obtained from the two branches more efficiently, inspired by the attention mechanism [70]. As shown in Fig. 7, this module allows the RGB branch to selectively fuse valuable features from the multispectral branch. We define the RGB branch and multispectral branch value maps as $\hat{\mathbf{R}}_v$ and $\hat{\mathbf{M}}_v$, respectively; then, the output of the sigmoid function can be expressed as:

$$\mathbf{P} = \boldsymbol{Sigmoid}\left(\varphi\left(\varphi\left(\hat{\mathbf{R}}_v\right)\cdot\varphi\left(\hat{\mathbf{M}}_v\right)\right)\right) \tag{6}$$

where $\varphi(\cdot)$ denotes the convolution and batch norm (BN) computation. $\mathbf{P}$ denotes the probability that these two pixels belong to the same object. A larger $\mathbf{P}$ means that the model would trust the RGB branch rather than the multispectral branch. Therefore, the output of the BGM can be written as:

$$\mathbf{O}_{BGM} = \mathbf{P}\hat{\mathbf{R}}_v + (1-\mathbf{P})\hat{\mathbf{M}}_v \tag{7}$$

E. Loss Function

We inherit ResNet-18/50 [68] to build our backbones, which include 4 stages, for feature extraction. Following [71], we place a segmentation head at the output of stage 3 for each of the two backbones. The additional segmentation head can generate additional semantic losses l_1, l_2, which can better optimize the whole network, where l_1 and l_2 are calculated using cross-entropy (CE) loss. The CE loss is defined as follows:

$$l_{CE} = -\sum_{i=0}^{N} y_i \ln(p_i) \tag{8}$$

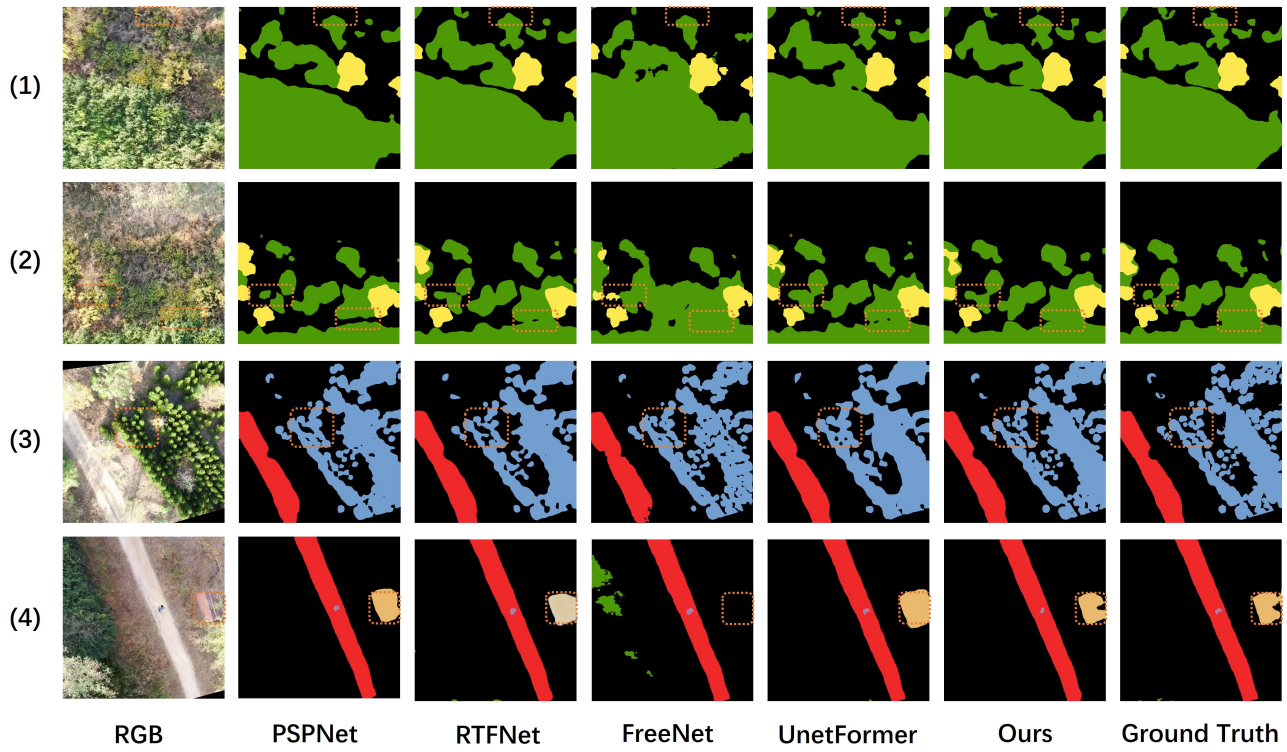

Fig. 9. Qualitative results on the UAVM dataset. We highlight the details with orange boxes.

where p is the model prediction, y is the label, and N is the number of categories.

Finally, the final loss of the whole UAVMNet is:

$$l_{total} = \lambda_0 l_0 + \lambda_1 l_1 + \lambda_2 l_2 + \sum_{i=0}^{1} \lambda_{Ali}^{i} l_{Ali}^{i} \tag{9}$$

where i represents stage 3 and stage 4 of the backbone when it is 0 and 1, respectively. Empirically, we refer to PIDNet [72] and set the hyperparameters of UAVMNet to $\lambda_0 = 1, \lambda_1 = 0.4, \lambda_2 = 0.4$. Additionally, we apply a greedy search strategy to choose $\lambda_{Ali} = [5, 5]$ for stage 3 and stage 4.

V. Experiments

In this section, we build a benchmark for the UAV multi-spectral semantic segmentation task and evaluate the proposed UAVMNet.

A. Implementation Details and Evaluation Metrics

Our code is implemented on the PyTorch platform. The model parameters are initialized via the ImageNet [74] pre-training model, which is then trained on a single GeForce RTX 3090 GPU and converged for optimal performance. The batch size is set to 4. The data augmentation strategies include random resizing, random horizontal flipping, random cropping, and padding to avoid potential overfitting. We use the Adam optimizer with an initial learning rate of 6e-5 and adaptive scheduling on the basis of training losses. During testing, data augmentation strategies are not applied, and images are kept at their original resolution for evaluation. We test on a single GeForce RTX 3090 GPU and set the batch size to 1 to measure the inference speed. Our implementation is based on MMSegmentation [75].

We use the mean intersection over union (mIoU), the mean pixel accuracy (mPA), and the mF1 score for evaluation. Pixel accuracy (PA) denotes the ratio of the number of correctly predicted pixels for a category to the total number of pixels, and mPA denotes the average percentage of correctly categorized pixels in the sample used for accuracy assessment. Intersection over union (IoU) denotes the ratio of the intersection of the model's predictions and true values for a category to the concatenation. MIoU denotes the average intersection over union result of the samples used for accuracy assessment. The F1 score is the harmonic mean of precision and recall, whereas the mF1 score is the average of the F1 scores of all categories. The definitions of the mIoU, mPA, and mF1 score are as follows:

$$mIoU = \frac{1}{k} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}} \tag{10}$$

$$mPA = \frac{1}{k} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}} \tag{11}$$

$$mF1score = \frac{1}{k} \sum_{i=0}^{k} \frac{2p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji}} \tag{12}$$

where p_{ij} refers to the value in the i^{th} row and j^{th} column of the confusion matrix; p_{ji} refers to the value in the j^{th} row and i^{th} column of the confusion matrix; the predicted category is j and the true category is i; and k is the number of categories.

TABLE III
QUANTITATIVE EVALUATION ON THE VALIDATION SET OF THE UAVM DATASET.

Type	Model	Backbone	mIoU(%)	mPA(%)	mF1-score(%)
Generic SS Model	UperNet [29]	ResNet-18	57.24	63.71	64.32
	UperNet [29]	ResNet-50	67.25	73.01	73.77
	Deeplabv3plus [1]	ResNet-18	62.33	68.99	67.49
	Deeplabv3plus [1]	ResNet-50	68.00	76.72	74.28
	UperNet [29]	Swin-T	54.54	59.29	62.57
	UperNet [29]	Swin-S	63.72	66.62	67.19
	FCN [8]	ResNet-18	50.03	55.21	56.87
	FCN [8]	ResNet-50	60.04	66.76	66.37
	PSPNet [28]	ResNet-18	52.17	58.45	57.41
	PSPNet [28]	ResNet-50	70.26	75.72	78.42
RGBT Image-based MSS Model	MFNet [47]	Mini-Inception	63.32	68.54	71.69
	EGFNet [66]	ResNet-152	54.56	64.10	66.55
	RTFNet [49]	ResNet-152	67.32	74.38	74.62
	PSTNet [48]	No	57.40	63.79	64.22
HSIC Model	HyLITE [73]	Vit	38.98	44.99	43.85
	FreeNet [61]	Swin-S	56.62	61.70	60.74
Remote Sensing Image-based SS Model	UnetFormer [41]	SWSL_ResNet-18	67.52	74.21	71.49
	Clusterformer [42]	No	62.60	68.43	70.71
Remote Sensing Image-based MSS Model	UAVMNet(ours)	ResNet-18	69.31	77.23	80.61
	UAVMNet(ours)	ResNet-50	**79.50**	**85.86**	**87.95**

TABLE IV
RESULTS ON THE PST900 DATASET.

Model	Background	Fire-Extinguisher	Backpack	Hand-Drill	Survivor	mIoU(%)
CCNet [76]	99.05	51.84	66.42	32.27	57.50	61.42
RTFNet-50 [49]	98.84	43.49	70.58	1.00	28.00	48.40
RTFNet-152 [49]	98.92	52.03	75.30	25.37	36.43	57.61
MFNet [47]	98.63	60.35	64.27	41.13	20.70	57.02
ERFNet [77]	98.73	58.79	68.08	52.76	34.38	62.55
MAVNet [78]	97.89	22.58	51.52	31.94	34.73	47.74
UNet [27]	97.95	42.96	52.89	38.27	31.64	52.74
Fast-SCNN [79]	98.51	35.48	64.60	15.50	21.68	47.15
ACNet [80]	99.25	51.46	83.19	59.95	65.19	71.81
PSTNet [48]	98.85	53.60	69.20	70.12	50.03	68.36
EGFNet [66]	99.20	**74.30**	83.00	71.20	64.60	78.50
UAVMNet(ours)	**99.45**	72.33	**84.50**	**74.15**	**74.84**	**81.06**

B. Comparison

UAVM dataset. Owing to the lack of prior work directly applicable to the new UAV multispectral semantic segmentation task, we first select the segmentation model relevant to this task from various segmentation methods. We replicate these methods via published code and default settings and benchmark them on the UAVM dataset. These models include semantic segmentation (SS) models that are based on RGB images (UperNet [29], Deeplabv3plus [1], FCN [8], and PSPNet [28]), multispectral semantic segmentation (MSS) models that are based on RGBT images (MFNet [47], RTFNet [49], EGFNet [66], and PSTNet [48]), SS models that are based on remote sensing images (UnetFormer [41], Clusterformer [42]), HSIC models (HyLITE [73], FreeNet [61]), and our proposed UAVMNet model. For the single-branch model, we concatenate the RGB and multispectral images in the channel direction and then feed them into the model for training. In our UAVMNet model, we expect our model to make full use of RGB and multispectral information to increase the segmentation accuracy.

Table III shows the segmentation results on the UAVM validation set. As shown in the table, our UAVMNet vastly outperforms the other methods and achieves the highest segmentation accuracy. For example, in the SS model, our ResNet-18-based model achieves a 69.31% mIoU, which is comparable in accuracy to those of the ResNet-50-based PSPNet [28] and Deeplabv3plus [1] models. For the MSS model, ResNet-152-based RTFNet achieves the highest mIoU of 67.32%, which is still not as high as that of our ResNet-18-based model. Our ResNet-50-based model achieves state-of-the-art results on the UAVM dataset (mIoU of 79.50%, mPA of 85.86% and mF1 scores of 87.95%).

Fig. 9 shows the visualization of the segmentation results in challenging scenarios. The plants shade each other in the bush regions in the first and second rows, and the plant distribution

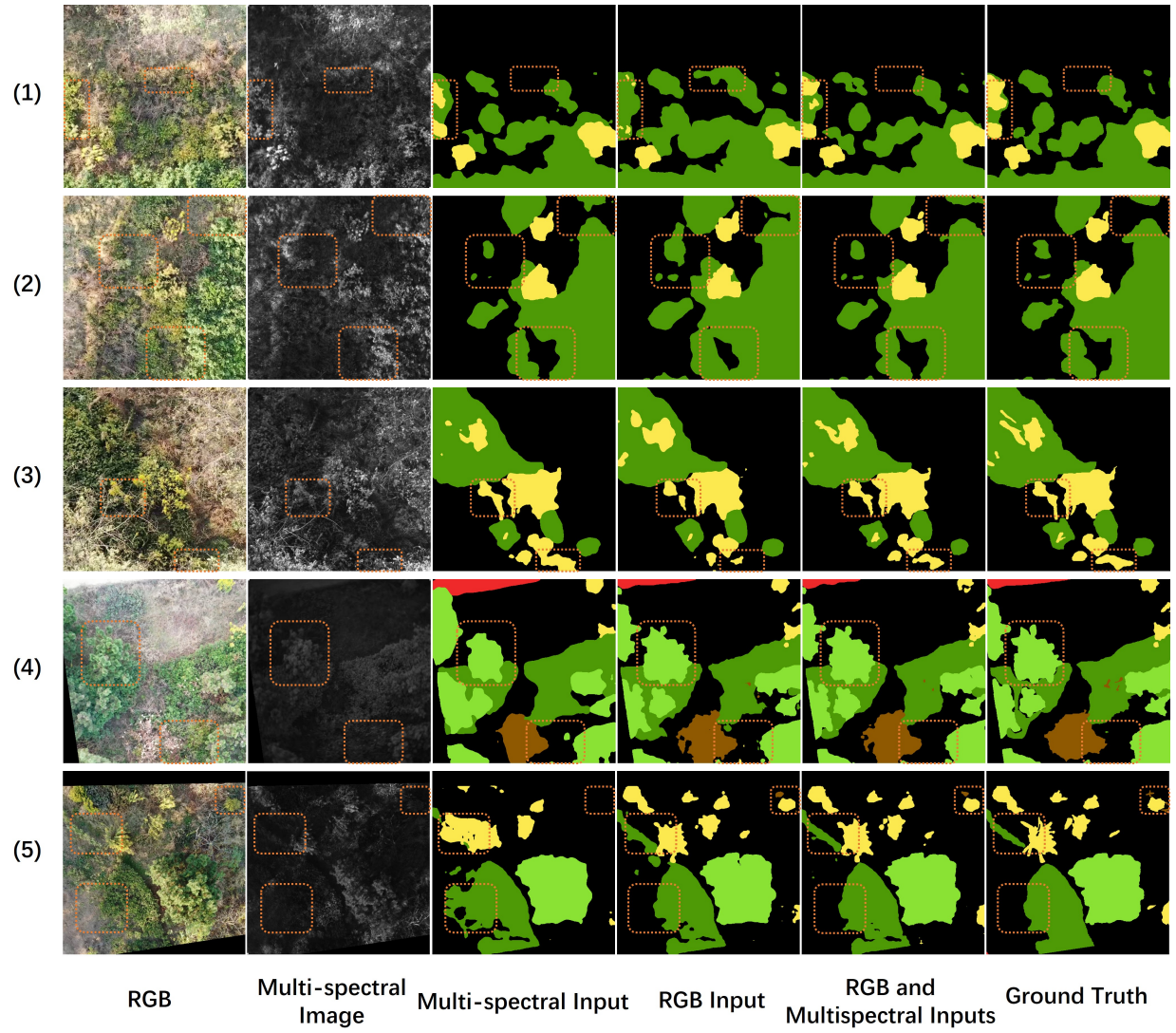

Fig. 10. Qualitative results of Deeplabv3plus [1] on the UAVM dataset. We present one of the channels in the multispectral images. From left to right are the original RGB images, the original multispectral images, the segmentation results trained with multispectral images, the segmentation results trained with only RGB images, the segmentation results trained with direct fusion of RGB images and multispectral images, and the ground truth. We highlight the details with orange boxes.

is more complex, making segmentation very difficult. For example, FreeNet wrongly segmented the bush into Solidago canadensis L. in the second row. Nevertheless, compared with other models, our model can segment plant boundaries more accurately and with higher precision. In the third row, the color of the cypress tree changes due to shadows, resulting in less accurate segmentation results. However, by fully fusing RGB and multispectral information, our UAVMNet can still perform relatively accurate inference in this case. Our model segmentation result is also much better for nonplant categories, such as the building in the last row of images. These results show that our UAVMNet can effectively fuse RGB and multispectral information to achieve more accurate segmentation results in challenging scenarios.

PST900. To validate the effectiveness of our model, we conducted experiments on the PST900 dataset, which is similar in type to the UAVM dataset. The PST900 dataset contains 894 aligned pairs of RGB and thermal images. The dataset has five categories of semantic annotations: background, fire extinguisher, backpack, hand drill, and survivor.

Table IV presents the results of our model on the PST900 test set. The results demonstrate that our model performs excellently, achieving the highest overall segmentation accuracy with an mIoU of 81.06. This indicates that our model can be effectively applied to various types of datasets, showcasing its robust generalization capability.

C. Ablation Analysis

We conduct ablation experiments on the UAVMNet model on the validation set. The results are shown in Tables V to IX, and the experimental results fully validate the effectiveness of our core design module. Unless otherwise stated, all the ablation experiments use ResNet-50 as the backbone.

RGB and multispectral information. We first investigate the advantages of fusing RGB and multispectral information in Table V. As shown in the table, for example, for the UperNet

TABLE V
QUANTITATIVE RESULTS FROM RGB AND MULTISPECTRAL INFORMATION FUSION ABLATION STUDIES.

Model	Backbone	Input Image		mIoU(%)
		RGB	Multi-spectrum	
Deeplabv3plus [1]	ResNet50	✓		65.26
			✓	51.34
		✓	✓	68.00
UperNet [29]	ResNet50	✓		64.59
			✓	50.75
		✓	✓	67.25

[29] model trained with only RGB or multispectral images, the mIoUs are 64.59% and 50.75%, respectively. Using the simplest direct fusion strategy (RGB and multispectral images are concatenated), the segmentation accuracy can also be improved, with an mIoU of 67.25%. The experimental results on the Deeplabv3plus [1] model are also similar.

Fig. 10 illustrates the qualitative tests performed on the Deeplabv3plus [1] model. For the area of the bush in the first and second rows, the RGB colors of the bush are similar to those of the nearby weed, and segmentation using only the RGB image results in error. However, weeds can be more easily distinguished in multispectral images, so we can obtain better segmentation results using only multispectral images. In the first row, the RGB colors of Solidago canadensis L. are similar to those of the bush, and the model mistakes Solidago canadensis L. for the bush with only RGB image segmentation. However, the model can clearly distinguish Solidago canadensis L. from the bush with the help of multispectral images. In the third row, due to the interference of tree branches, segmentation using the multispectral image is able to obtain more accurate results than using the RGB image. However, opposite results are obtained for the area of pine in the fourth row and for the area of the bush and Solidago canadensis L. in the fifth row, which may be because the multispectral image contains some redundant bands that affect the segmentation results. Another factor is that the RGB channel can provide high-quality information, which is more suitable for describing the contour details of pines. We obtain better results when we use both RGB images and multispectral images as inputs than when we use multispectral images or RGB images separately. This experiment demonstrates the benefits of using both RGB and multispectral information to improve semantic segmentation.

Effectiveness of extra losses. To facilitate the optimization of the whole network and enhance the functionality of each component, UAVMNet introduces three additional losses. According to Table VI, the use of the segmentation header to compute the additional semantic losses l_1, l_2 is crucial to improve the segmentation accuracy. In particular, the joint use of l_1 and l_2 (+2.06% mIoU) strongly justifies the introduction of additional auxiliary segmentation heads. We evaluated the UAVM alignment loss l_{Ali} in Table VI. Integrating our UAVM alignment loss l_{Ali} improves the mIoU score by 3.35% with additional semantic losses l_1 and l_2. This improvement may be because UAVM alignment loss converts feature activation into probability distributions in the direction of spatial locations, thus focusing on the overall information alignment of each spatial location. These spatial locations may contain rich global contextual information. This experimental result demonstrates the effectiveness of introducing a feature alignment module.

TABLE VI
ABLATION STUDY OF EXTRA LOSSES FOR UAVMNET.

Extra Loss			mIoU(%)
l_1	l_2	l_{Ali}	
			76.15
✓			76.92(+0.77)
	✓		76.68(+0.53)
✓	✓		78.21(+2.06)
✓	✓	✓	**79.50**(+3.35)

The weight of UAVM alignment loss. When additional semantic losses l_1 and l_2 are used, the weights of the alignment losses are shown in Table VII. Theoretically, the weighted vector should not be too large or too small. If a larger weight is assigned, it may dominate the training process while omitting the other losses. A model with a smaller weight may not effectively utilize the high-quality spectral information of the multispectral stream. For the weights of the feature maps, a weight of [5, 5] is the most appropriate and can significantly improve the performance (i.e., 78.21% $\rightarrow$ 79.50%). Boundary gains occur when the weights are increased beyond [5, 5]. Conversely, when the weights are too large, the model accuracy decreases.

TABLE VII
ABLATION STUDY ON THE WEIGHT OF UAVM ALIGNMENT LOSS.

Loss Weight	mIoU(%)
0, 0	78.21
1, 1	78.36
3, 3	79.37
5, 5	**79.50**
10, 10	79.44
15, 15	78.78
20, 20	78.44

Collaboration of CAF and BGM. We also tested the effectiveness of the proposed BGM and CAF modules. Add represents the elementwise summation operation. Under the same conditions, we evaluated the model's performance in four cases: Add+Add, CAF+Add, Add+BGM, and CAF+BGM. The experimental results in Table VIII demonstrate the superiority of the CAF and BGM working together. The joint use of CAF and BGM improves the mIoU score by 1.27% (i.e., 78.23% $\rightarrow$ 79.50%).

Fig. 11 shows the results of the corresponding visualizations, where we visualize the features separately for different categories. The first and fourth rows concern the visualization of the pine tree category. When Add+Add is used, the first row incorrectly identifies the pine tree as background. In the fourth row, bushes are incorrectly identified as pines, and

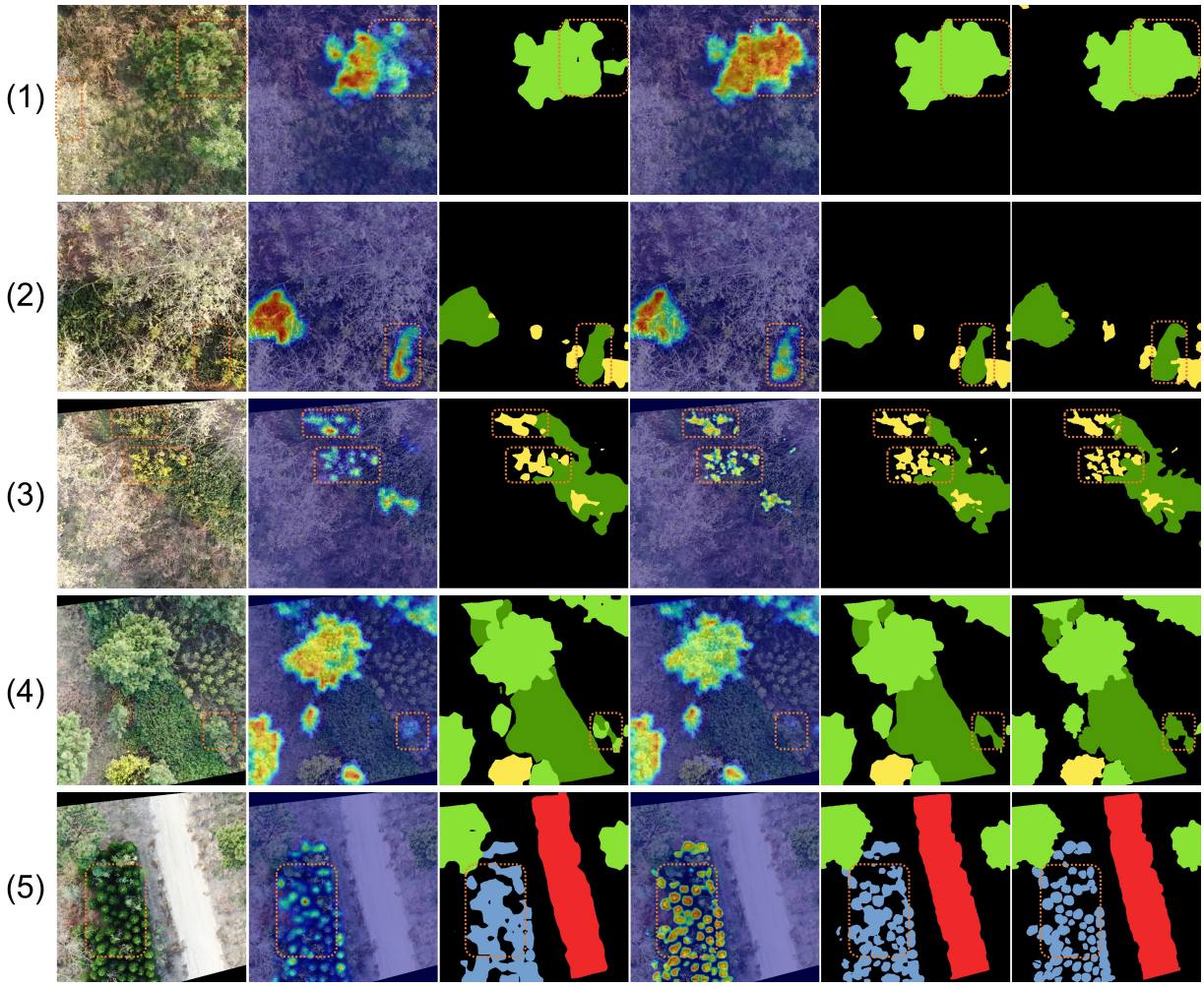

Fig. 11. Visualization results of the joint use of the CAF and BGM on the UAVM dataset. The first column indicates the RGB input images; the second and third columns are qualitative results obtained via Add+Add; the fourth and fifth columns are qualitative results obtained via CAF+BGM; the sixth column is the ground truth; the second and fourth columns are feature visualizations; and the third and fifth columns are the segmentation results. We highlight the details with orange boxes for comparison.

TABLE VIII
ABLATION STUDY OF CAF AND BGM ON UAVMNET.

Fusion1		Fusion2		mIoU(%)
ADD	CAF	ADD	BGM	
✓		✓		78.23
	✓	✓		78.44(+0.21)
✓			✓	78.67(+0.44)
	✓		✓	**79.50**(+1.27)

the orange box in the heatmap of the feature visualization also shows the discrimination error. However, both errors are avoided when the CAF and BGM are used jointly, suggesting that the two attentional mechanisms adequately fuse RGB and multispectral information. The second, third and fifth rows show the visualization results for the bush, Solidago canadensis L. and cypress categories, respectively. When both the CAF and BGM are used, the activation values in the heatmap are more concentrated. Thus, the boundaries are more precise for complex vegetation contour segmentation, which further demonstrates the effectiveness of our module in fusing RGB and multispectral information.

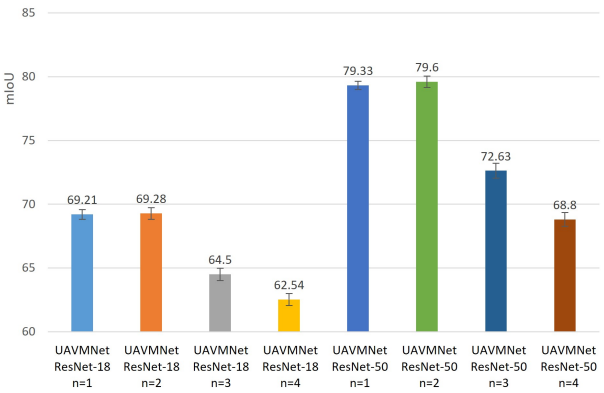

Fig. 12. Results of the standard deviation experiments.

Attention downsampling. To increase the computational efficiency, we employ an attentional downsampling operation. We conduct ablation experiments on the stride n. As shown

TABLE IX
EFFECT OF DIFFERENT DOWNSAMPLING STRIDE n ON THE UAVM DATASET.

Model	Backbone	Evaluation	$n=1$	$n=2$	$n=3$	$n=4$
UAVMNet	ResNet-18	mIoU(%)	69.23	69.31	64.53	62.49
		FPS	45.14	52.13	52.50	52.76
UAVMNet	ResNet-50	mIoU(%)	79.36	79.50	72.67	68.78
		FPS	20.10	23.81	23.97	24.66

in Table IX, when the stride size n is increased, the model accuracy is slightly improved, while the inference speed is also increased. Specifically, the model performs optimally when the stride n is 2. The results indicate that the maximum pooling downsampling operation is very effective in sparsely extracting spatial information. However, the model accuracy decreases rapidly as the information becomes too sparse when the stride n further increases. In this case, the inference speedup is not significant because the feature map resolution is already low.

To ensure the stability and repeatability of the experiments, we perform standard deviation experiments on the proposed UAVMNet network based on this ablation experiment. We maintain consistent experimental settings and retrain the network five times. The specific experimental results are in Fig. 12. The height of each bar represents the mean and the short line represents the standard deviation. As shown in Fig. 12, the maximum variance is only 0.59, indicating that our experiment exhibits high stability and repeatability.

VI. CONCLUSION

This paper presents a preliminary study of the UAV multispectral semantic segmentation problem. To monitor the spread of the invasive species Solidago canadensis L., we present the UAVM dataset, a collection of RGB and multispectral image pairs containing 3260 aligned and annotated pairs. We conduct comprehensive benchmark experiments on the UAVM dataset and propose a simple yet effective baseline framework (UAVMNet). RGB and multispectral information each have their own advantages and disadvantages. To effectively fuse RGB and multispectral information, we first employ a feature alignment module to ensure the coherence of features. We subsequently utilize the UAVMFuse module to construct a comprehensive feature representation and decode it, thereby maximizing the complementary nature of the cross-spectral data. Through this design, UAVMNet achieves state-of-the-art performance on the UAVM dataset. In the future, we plan to extend the UAVM dataset with richer forms of data annotation.

REFERENCES

[1] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[2] R. Sheikh, A. Milioto, P. Lottes, C. Stachniss, M. Bennewitz, and T. Schultz, "Gradient and log-based active learning for semantic segmentation of crop and weed for agricultural robots," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 1350–1356.

[3] H. Bi, L. Xu, X. Cao, Y. Xue, and Z. Xu, "Polarimetric sar image semantic segmentation with 3d discrete wavelet transform and markov random field," *IEEE transactions on image processing*, vol. 29, pp. 6601–6614, 2020.

[4] L. P. Osco, J. M. Junior, A. P. M. Ramos, L. A. de Castro Jorge, S. N. Fatholahi, J. de Andrade Silva, E. T. Matsubara, H. Pistori, W. N. Gonçalves, and J. Li, "A review on deep learning in uav remote sensing," *International Journal of Applied Earth Observation and Geoinformation*, vol. 102, p. 102456, 2021.

[5] Q. Bi, S. You, W. Ji, and T. Gevers, "Learning rotation equivalent scene representation from instance-level semantics: A novel top-down perspective," *Computer Vision and Image Understanding*, vol. 229, p. 103635, 2023.

[6] 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 transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[7] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[8] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[9] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang *et al.*, "Deep high-resolution representation learning for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3349–3364, 2020.

[10] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 408–14 419.

[11] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong *et al.*, "Swin transformer v2: Scaling up capacity and resolution," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 009–12 019.

[12] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 1290–1299.

[13] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 000–16 009.

[14] P. Lottes, R. Khanna, J. Pfeifer, R. Siegwart, and C. Stachniss, "Uav-based crop and weed classification for smart farming," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3024–3031.

[15] N. Chebrolu, T. Läbe, and C. Stachniss, "Robust long-term registration of uav images of crop fields for precision agriculture," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3097–3104, 2018.

[16] I. Nigam, C. Huang, and D. Ramanan, "Ensemble knowledge transfer for semantic segmentation," in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2018, pp. 1499–1508.

[17] Y. Lyu, G. Vosselman, G.-S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS journal of photogrammetry and remote sensing*, vol. 165, pp. 108–119, 2020.

[18] "Icg drone dataset," *http://dronedataset.icg.tugraz.at.*

[19] M. Kerkech, A. Hafiane, and R. Canals, "Vddnet: Vine disease detection network based on multispectral images and depth map," *Remote Sensing*, vol. 12, no. 20, p. 3305, 2020.

[20] ——, "Vine disease detection in uav multispectral images with deep learning segmentation approach." 2020.

[21] I. Sa, M. Popović, R. Khanna, Z. Chen, P. Lottes, F. Liebisch, J. Nieto, C. Stachniss, A. Walter, and R. Siegwart, "Weedmap: A large-scale semantic weed mapping framework using aerial multispectral imaging and deep neural network for precision farming," *Remote Sensing*, vol. 10, no. 9, p. 1423, 2018.

[22] A. J. Abreu, L. A. Alexandre, J. A. Santos, and F. Basso, "Ludvision–remote detection of exotic invasive aquatic floral species using drone-mounted multispectral data," *arXiv preprint arXiv:2207.05620*, 2022.

[23] D. Mei, L. Jian-Zhong, Z. Wen-Ju, C. Jia-Kuan, and L. Bo, "Canada goldenrod (solidago canadensis): an invasive alien weed rapidly spread-

ing in china," *Journal of Systematics and Evolution*, vol. 44, no. 1, p. 72, 2006.

[24] C. Schittko and S. Wurst, "Above-and belowground effects of plant-soil feedback from exotic solidago canadensis on native tanacetum vulgare," *Biological Invasions*, vol. 16, pp. 1465–1479, 2014.

[25] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1520–1528.

[26] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[27] O. Ronneberger, P. Fischer, and T. Brox, "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*. Springer, 2015, pp. 234–241.

[28] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2881–2890.

[29] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 418–434.

[30] F. López-Granados, J. Torres-Sánchez, A. Serrano-Pérez, A. I. de Castro, F.-J. Mesas-Carrascosa, and J.-M. Pena, "Early season weed mapping in sunflower using uav technology: variability of herbicide treatment maps against weed thresholds," *Precision agriculture*, vol. 17, pp. 183–199, 2016.

[31] F. Rottensteiner, G. Sohn, M. Gerke, J. D. Wegner, U. Breitkopf, and J. Jung, "Results of the isprs benchmark on urban object detection and 3d building reconstruction," *ISPRS journal of photogrammetry and remote sensing*, vol. 93, pp. 256–271, 2014.

[32] C. Debes, A. Merentitis, R. Heremans, J. Hahn, N. Frangiadakis, T. van Kasteren, W. Liao, R. Bellens, A. Pižurica, S. Gautama *et al.*, "Hyperspectral and lidar data fusion: Outcome of the 2013 grss data fusion contest," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2405–2418, 2014.

[33] X. Hu, Y. Zhong, C. Luo, and X. Wang, "Whu-hi: Uav-borne hyperspectral with high spatial resolution (h2) benchmark datasets for hyperspectral image classification," *arXiv preprint arXiv:2012.13920*, 2020.

[34] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[36] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.

[37] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[38] M. Yuan, D. Ren, Q. Feng, Z. Wang, Y. Dong, F. Lu, and X. Wu, "Mcafnet: a multiscale channel attention fusion network for semantic segmentation of remote sensing images," *Remote Sensing*, vol. 15, no. 2, p. 361, 2023.

[39] R. Dong, X. Pan, and F. Li, "Denseu-net-based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65 347–65 356, 2019.

[40] S. Liu, W. Ding, C. Liu, Y. Liu, Y. Wang, and H. Li, "Ern: Edge loss reinforced semantic segmentation network for remote sensing images," *Remote Sensing*, vol. 10, no. 9, p. 1339, 2018.

[41] L. Wang, R. Li, C. Zhang, S. Fang, C. Duan, X. Meng, and P. M. Atkinson, "Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 190, pp. 196–214, 2022.

[42] H. Liu, W. Li, W. Jia, H. Sun, M. Zhang, L. Song, and Y. Gui, "Clusterformer for pine tree disease identification based on uav remote sensing image segmentation," *IEEE Transactions on Geoscience and Remote Sensing*, 2024.

[43] N. Genze, R. Ajekwe, Z. Güreli, F. Haselbeck, M. Grieb, and D. G. Grimm, "Deep learning-based early weed segmentation using motion blurred uav images of sorghum fields," *Computers and Electronics in Agriculture*, vol. 202, p. 107388, 2022.

[44] Y. Li, Z. Cao, H. Lu, Y. Xiao, Y. Zhu, and A. B. Cremers, "In-field cotton detection via region-based semantic image segmentation," *Computers and Electronics in Agriculture*, vol. 127, pp. 475–486, 2016.

[45] Z. Qin, J. Liu, X. Zhang, M. Tian, A. Zhou, S. Yi, and H. Li, "Pyramid fusion transformer for semantic segmentation," *IEEE Transactions on Multimedia*, 2024.

[46] B. Liu, J. Hu, X. Bi, W. Li, and X. Gao, "Pgnet: Positioning guidance network for semantic segmentation of very-high-resolution remote sensing images," *Remote Sensing*, vol. 14, no. 17, p. 4219, 2022.

[47] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 5108–5115.

[48] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "Pst900: Rgb-thermal calibration, dataset and segmentation network," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 9441–9447.

[49] Y. Sun, W. Zuo, and M. Liu, "Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[50] J. Vertens, J. Zürn, and W. Burgard, "Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images. in 2020 ieee," in *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 8461–8468.

[51] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. Hwang, "Gmnet: Graded-feature multilabel-learning network for rgb-thermal urban scene semantic segmentation," *IEEE Transactions on Image Processing*, vol. 30, pp. 7790–7802, 2021.

[52] X. Lan, X. Gu, and X. Gu, "Mmnet: Multi-modal multi-stage network for rgb-t image semantic segmentation," *Applied Intelligence*, vol. 52, no. 5, pp. 5817–5829, 2022.

[53] Z. Liang and X. Wang, "Semantic segmentation network with band-location adaptive selection mechanism for multispectral remote sensing images," in *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*. IEEE, 2022, pp. 3488–3491.

[54] Y. Tan, S. Xiong, and Y. Li, "Automatic extraction of built-up areas from panchromatic and multispectral remote sensing images using double-stream deep convolutional neural networks," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 11, no. 11, pp. 3988–4004, 2018.

[55] N. Kieu, K. Nguyen, S. Sridharan, and C. Fookes, "General-purpose multimodal transformer meets remote sensing semantic segmentation," *arXiv preprint arXiv:2307.03388*, 2023.

[56] M. Ahmad, S. Shabbir, S. K. Roy, D. Hong, X. Wu, J. Yao, A. M. Khan, M. Mazzara, S. Distefano, and J. Chanussot, "Hyperspectral image classification—traditional to deep models: A survey for future prospects," *IEEE journal of selected topics in applied earth observations and remote sensing*, vol. 15, pp. 968–999, 2021.

[57] T. R. Martha, N. Kerle, C. J. Van Westen, V. Jetten, and K. V. Kumar, "Segment optimization and data-driven thresholding for knowledge-based landslide detection by object-based image analysis," *IEEE transactions on geoscience and remote sensing*, vol. 49, no. 12, pp. 4928–4943, 2011.

[58] Q. Zhu, Y. Zhong, B. Zhao, G.-S. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, 2016.

[59] Z. Zhong, J. Li, Z. Luo, and M. Chapman, "Spectral–spatial residual network for hyperspectral image classification: A 3-d deep learning framework," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 2, pp. 847–858, 2017.

[60] X. Mei, E. Pan, Y. Ma, X. Dai, J. Huang, F. Fan, Q. Du, H. Zheng, and J. Ma, "Spectral-spatial attention networks for hyperspectral image classification," *Remote Sensing*, vol. 11, no. 8, p. 963, 2019.

[61] Z. Zheng, Y. Zhong, A. Ma, and L. Zhang, "Fpga: Fast patch-free global learning framework for fully end-to-end hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 8, pp. 5612–5626, 2020.

[62] A. Witkin, "Scale-space filtering: A new approach to multi-scale description," in *ICASSP'84. IEEE international conference on acoustics, speech, and signal processing*, vol. 9. IEEE, 1984, pp. 150–153.

[63] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[64] A.-z. yatengLG and horffmanwang, "ISAT with segment anything: A interactive semi-automatic annotation tool based segment anything," 2023, open source software available from https://github.com/yatengLG/ISAT with segment anything. [Online]. Available: https://github.com/yatengLG/ISAT_with_segment_anything
[65] "Xiong'an Aerial Hyperspectral Remote Sensing Image Dataset," *www.hrs-cas.com/a/share/shujuchanpin/2019/0501/1049.html*, 2019.
[66] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3571–3579.
[67] C. Shu, Y. Liu, J. Gao, Z. Yan, and C. Shen, "Channel-wise knowledge distillation for dense prediction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5311–5320.
[68] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.
[69] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9226–9235.
[70] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[71] C. Yu, C. Gao, J. Wang, G. Yu, C. Shen, and N. Sang, "Bisenet v2: Bilateral network with guided aggregation for real-time semantic segmentation," *International Journal of Computer Vision, 129(11):3051–3068, 2021*.
[72] J. Xu, Z. Xiong, and S. P. Bhattacharyya, "Pidnet: A real-time semantic segmentation network inspired by pid controllers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 19 529–19 539.
[73] F. Zhou, M. Kilickaya, and J. Vanschoren, "Locality-aware hyperspectral classification," *arXiv preprint arXiv:2309.01561*, 2023.
[74] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.
[75] M. Contributors, "MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark," https://github.com/open-mmlab/mmsegmentation, 2020.
[76] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 603–612.
[77] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Erfnet: Efficient residual factorized convnet for real-time semantic segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 1, pp. 263–272, 2017.
[78] T. Nguyen, S. S. Shivakumar, I. D. Miller, J. Keller, E. S. Lee, A. Zhou, T. Özaslan, G. Loianno, J. H. Harwood, J. Wozencraft *et al.*, "Mavnet: An effective semantic segmentation micro-network for mav-based tasks," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3908–3915, 2019.
[79] R. P. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," *arXiv preprint arXiv:1902.04502*, 2019.
[80] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *2019 IEEE international conference on image processing (ICIP)*. IEEE, 2019, pp. 1440–1444.

Qiusheng Li received a B.S. degree in communication engineering from Hangzhou Dianzi University in 2022. He is currently pursuing a graduate degree from the Department of Information Science and Engineering, Ocean University of China. His research interests include deep learning and computer vision.

Hang Yuan received a B.Eng. degree in information engineering from the Qingdao University of Science and Technology in 2022. He is currently pursuing a graduate degree from the Department of Information Science and Engineering, Ocean University of China. His research interests include deep learning and computer vision.

Tianning Fu received a B.Eng. degree in Internet of Things engineering from Hainan University in 2022. He is currently pursuing a graduate degree from the Department of Information Science and Engineering, Ocean University of China. His research interests include deep learning and computer vision.

Zhibin Yu (Member, IEEE) received a B.E. degree from the Harbin Institute of Technology, China, and an M.E. degree in computer engineering and a Ph.D. degree in electrical engineering from Kyungpook National University, South Korea, in 2009 and 2016, respectively. In 2016, he joined the Department of Electronic Engineering, Ocean University of China, where he is currently an Associate Professor in the College of Electronic Engineering. His current research interests include underwater image processing, image-to-image translation, and generative neural networks.

Bing Zheng (Member, IEEE) received a B.S. degree in electronics and information systems, an M.S. degree in marine physics, and a Ph.D. degree in computer application technology from the Ocean University of China, Qingdao, China, in 1991, 1995, and 2013, respectively. He is currently a Professor in the College of Electronic Engineering, Ocean University of China. His research interests include ocean optics, underwater imaging, and optical detection.

Shuguo Chen received a Ph.D. degree in marine information detection and processing from the Ocean University of China, Qingdao, in 2015. Currently, he is with the College of Marine Technology, Ocean University of China, and his research focuses primarily on ocean optics and ocean color remote sensing, particularly on the calibration and validation of ocean color satellite sensors and in situ measurements and atmospheric correction, as well as UAV remote sensing.