DBCG-Net : Dual Branch Calibration Guided Deep Network for UAV Images Semantic Segmentation

Chaoyun Mai, Yibo Wu, Yikui Zhai*, *Senior Member, IEEE,* Hao Quan, Jianhong Zhou, Angelo Genovese, *Senior Member, IEEE,* Vincenzo Piuri, *Fellow, IEEE,* and Fabio Scotti, *Senior Member, IEEE*

***Abstract*—Unmanned aerial vehicle (UAV) remote sensing images used for semantic segmentation possess distinct features compared to urban street scene images, including high resolution and a complex background. Spatial information plays a pivotal role in enhancing the performance of semantic segmentation for high-resolution images. The dual-branch architecture for semantic segmentation incorporates supplementary branches to capture spatial information. However, prior research on dual-branch semantic segmentation neglected the interaction between the contextual and spatial branches, leading to suboptimal model performance. In this discourse, the paper introduces a dual-branch semantic segmentation framework. This design advances the system's understanding of spatial information while facilitating inter-branch learning through two key modules. Initially, the spatial calibration feature extraction module employs frequency domain processing and learning tactics distinct from the contextual approach to generate image features under varied noise conditions. Calibration is achieved by generating features from diverse angles. Subsequently, the spatially-guided loss function directs the acquisition of spatial information for the spatial branch by condensing the deep image characteristics for the context branch. To assess the generalization capacity of the proposed method, experiments will be conducted on three different datasets. The proposed method's modules will be integrated into three representative dual-branch networks, allowing assessment of the generalization capacity of the key DBCG components. Empirical evidence demonstrates that this approach is highly effective, significantly surpassing the performance of the baseline network. Code is available at https://github.com/yikuizhai/DBCG-Net-master.**

***Index Terms*—CNN, deep learning, semantic segmentation, dual-branch calibration guided network, UAVs.**

This study was funded by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011576), Guangdong Science and Technology Planning Project (No. 2021A0505030080), Guangdong Science and Technology Planning Project (No. 2021A0505060011), Guangdong Higher Education Innovation and Strengthening School Project (No. 2022ZDZX1032), Guangdong Higher Education Innovation and Strengthening School Project (No. 2020ZDZX3031), Guangdong Jiangmen Science and Technology Project (No. 2220002000246), Guangdong Higher Education Innovation and Strengthening School Project (No. 2023ZDZX1029), Wuyi University Hong Kong and Macao Joint Research and Development Fund (No. 2022WGALH19), Guangdong Jiangmen Science and Technology special Correspondent Research Project (No. 2023760300070008390), 2022 Educational Science Planning Project Higher Education Special Project (No. 2022GXJK350).

Chaoyun Mai, Yibo Wu, Yikui Zhai, Jianhong Zhou are with the School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, Guangdong, China (e-mail: maichaoyun@foxmail.com; wuyibo0308@foxmail.com; yikuizhai@163.com; jackychou_lab126.com). *(Corresponding author: Yikui Zhai.)*

Hao Quan is with Dip. di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Ponzio 34, 20133 Milano (MI), Italy (e-mail: hao.quan@polimi.it).

Angelo Genovese, Vincenzo Piuri and Fabio Scotti are with Dipartimento di Informatica, Università Degli Studi di Milano, Via Celoria 18, 20133 Milano (MI), Italy (e-mail: angelo.genovese@unimi.it; vincenzo.piuri@unimi.it; fabio.scotti@unimi.it).

I. Introduction

Compared to satellite and aerial remote sensing, UAVs provide high-resolution imagery, detailed information on specific areas, and flexible imaging characteristics. Consequently, UAVs have emerged as a field of considerable interest [1], [2]. UAV imagery has found extensive applicability in research disciplines, notably environmental surveillance [3], precision farming [4], [5], accurate vehicle segmentation [6], and urban scene analysis [7]. UAVs frequently serve in damage assessment caused by natural calamities due to their rapid deployment prowess and adaptable information gathering capacity. For example, emergency detection of building collapses, floods, and fires [8], [9]. Within the realm of precision agriculture, UAVs prove to be instrumental in detecting crop lodging conditions [10]. Furthermore, in response to the scarcity of UAV datasets, a multitude of UAV datasets have emerged in recent years. For instance, ManipalUAVid [11] offers a dataset for UAV-based video semantic segmentation. Additionally, UAVPal [12] presents complex urban scenes from the UAV's perspective in Bhopal, Madhya Pradesh. Furthermore, there has been burgeoning interest in broad-scale models. SpectralGPT [13] introduced a comprehensive large-scale model tailored for the processing of remote sensing imagery. Currently, UAVs are utilized in both civilian and strategic capacities with an emphasis on information acquisition via visual sensors. Deep learning methods play a crucial role in enhancing UAVs' understanding of image data. Currently, UAVs have witnessed substantial progress in tasks, including object detection and recognition, within the domain of image applications. In recent years, research on UAV semantic segmentation has proliferated, indicating considerable potential for advancements in this area.

Semantic segmentation, pivotal for scene comprehension, is a deep learning-based image processing technique designed for dense prediction through pixel-by-pixel classification. This approach has found widespread application in distinct sectors, such as autonomous driving [14] and medical imaging [15]. Particularly, for applications requiring both location and boundary information in images, UAV semantic segmentation proves indispensable. This information can be captured at any moment, laying the groundwork for subsequent tasks. Research on semantic segmentation for UAV imagery is crucial for advancing applications utilizing UAV technology. While semantic segmentation algorithms have achieved significant results in the realm of autonomous driving, UAV scenarios present unique challenges, including higher image resolution,

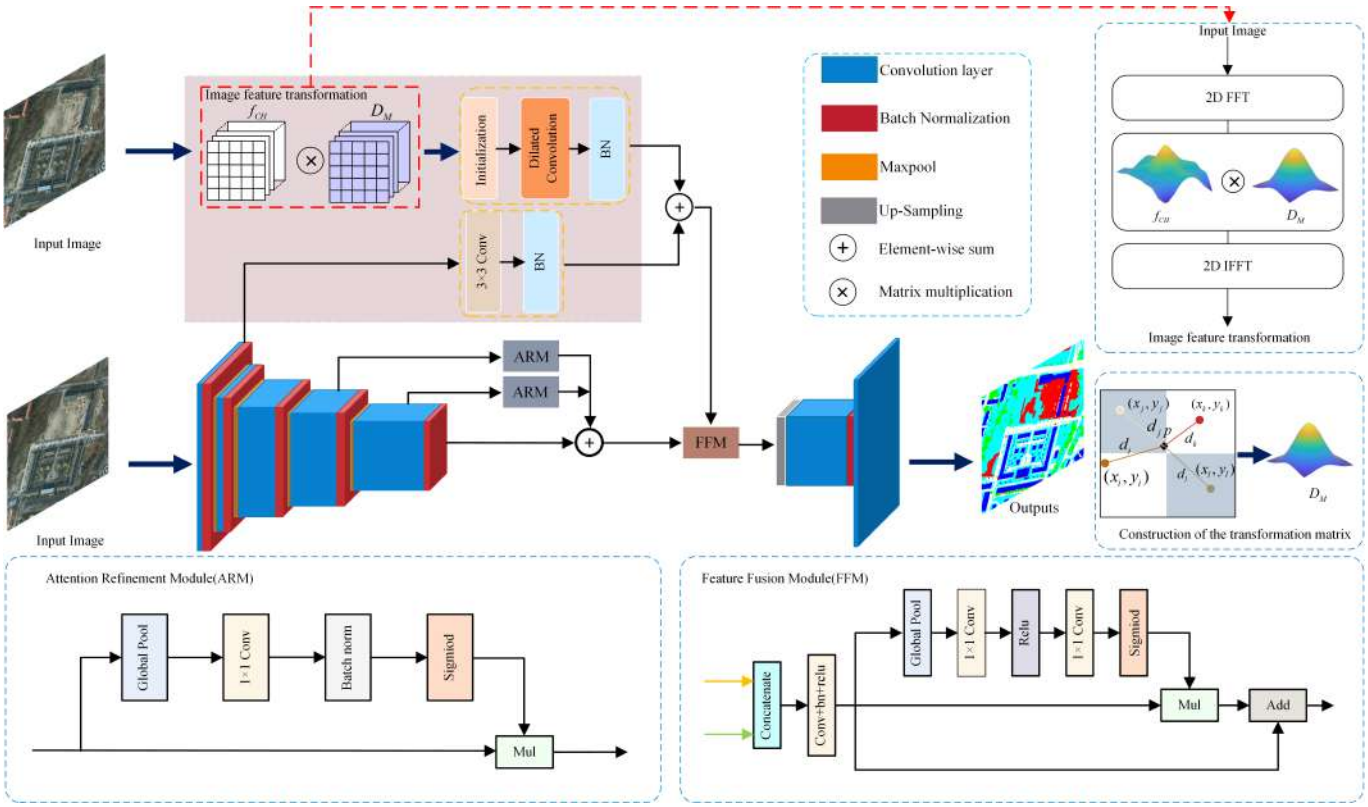

Fig. 1. A Dual-branch Calibration Guided Deep Learning Framework. ARM denotes Attention Refinement module, and FFM denotes Feature Fusion Module in [21]. The SCFEM is depicted in the pink background box, while the blue dashed box serves to delineate the Filtering process.

a larger field of vision, and more complex background data compared to images from autonomous driving contexts. For example, in UAV-related scenarios, object size variations are more pronounced compared to autonomous driving scenes, primarily due to differences in altitude. To address this issue, [16] introduces a solution in the form of a multi-scale convolutional neural network architecture, complemented by the utilization of superpixels. To tackle large-scale variations in UAV images, [17] introduces a multi-attention network that incorporates bidirectional fusion of multi-scale features, leading to enhanced feature extraction across multiple scales. The efficacy of this approach was demonstrated through promising results obtained on the UAVid dataset [18]. Despite significant progress made by existing semantic segmentation networks across various scene segmentations, UAV remote sensing semantic segmentation still faces numerous challenges, such as UAV images frequently encompassing a multitude of objects and intricate environments, making the detection of smaller objects within the images arduous. Moreover, challenges are compounded in UAV semantic segmentation due to factors like indistinct segmentation boundaries and ambiguity in segmenting the surrounding environments. To address challenges related to small objects and segmentation boundaries, a V-shaped encoder-decoder architecture is proposed [1]. This architecture incorporates upsampling, downsampling, and skip connections to enhance segmentation effectiveness for small targets and boundaries by refining the spatial detail feature maps extracted from the backbone. Cross-city scenarios could potentially impact the model's performance. C2Seg [19] offers a cross-city multimodal remote sensing dataset and introduces HighDAN (High-Resolution Domain Adaptive Network) to improve the model's generalization across diverse urban environments. As a result of the aforementioned challenges and difficulties, conventional semantic segmentation algorithms, originally intended for autonomous driving, tend to exhibit inadequate performance when applied to UAV images. This underscores the significance of devising robust semantic segmentation algorithms specifically tailored for UAV images.

The suboptimal performance observed when traditional semantic segmentation methods are applied to high-resolution UAV imagery can be attributed, in part, to the resolution itself. Effective semantic segmentation algorithms excel for several reasons, as indicated by prior research [20], with the vital role of spatial information being one of them, contributing to improved segmentation performance. A deeper understanding of such spatial elements can enhance the processing of high-resolution images. Consequently, developing methods to more effectively learn and utilize spatial details could significantly advance UAV scene semantic segmentation techniques. Existing strategies, such as dilated convolution, aim to uphold accuracy by preserving high resolution. However, the application of this method incurs significantly high computational costs in an effort to maintain superior resolution. In reaction to this, the dual-branch network [21] incorporates an added branch to learn spatial detail information, ensuring lower resolution while compensating for spatial information loss through an

auxiliary branch. Considering the effectiveness of these dual-branch networks within the domain of semantic segmentation, adjustments were made to streamline the original structure and corresponding adjustments to the semantic branch. Regarding the augmented computational demands imposed by [22], the integration of an additional spatial branch necessitated the subsequent recalibration of this specific branch, conducted by [23].

The fusion of information between the two branches enhances the model's learning efficacy, supported by numerous studies [24], [25], [26]. Prior methodologies, however, primarily facilitated learning across branches through feature fusion. In the study by [24], the integration of low-level features with high-level features is explored, focusing on promoting information fusion between the dual branches. This integration aims to facilitate the fusion of high-level features with low-level features. Nevertheless, these blending techniques share a common drawback: the discrepancy between deep and shallow features requires either up-sampling of low-resolution elements or down-sampling of high-resolution ones. In this process, an inevitable loss of information may occur, leading to suboptimal outcomes. Drawing from the concept of knowledge distillation [27], we view this as an information compression and refinement process. The deep semantic information approximates what we describe as the teacher model, while the student model is likened to shallow spatial information. By condensing the deep semantic information and extracting its knowledge, we facilitate the learning of shallow information, thereby addressing the aforementioned issues.

When reviewing the dual-branch network [21], [22], [23], we identified several issues. Firstly, the spatial branch appears overly simplistic in its configuration, hindering effective extraction of spatial feature information. Both branches operate independently without any interaction of information. Secondly, the direct fusion of information between the two distinct branches may encounter potential issues due to the significant difference between semantic and spatial information, potentially compromising the model's learning efficiency. Lastly, using identical images as inputs for both branches leads to a lack of diversity.

In this work, we introduce an Dual-branch Caibration Guided Neural Network architecture(DBCG-Net). Fig. 1 illustrates the main framework diagram. The novelty of this study lies in the introduction of the concepts of "calibration" and "guidance" when compared to existing literature. Firstly, we propose the notion of "calibration" within a dual-branch network. Knowledge calibration is a process aimed at improving accuracy through calibration, which can be achieved using two methods. The first method, known as repetitive calibration, involves evaluating the problem repeatedly from the same perspective and using the same method. Conversely, the second method, called differential calibration, entails examining the problem from distinct perspectives or adopting diverse approaches. It is evident that employing different methods or viewpoints to tackle the same problem can yield higher accuracy in comparison to repeatedly checking from the same perspective or using the same method. Therefore, based on this notion, we propose that "differential calibration" is superior to repetitive calibration. Building upon this idea, we introduce the SCFEM module. Diverging from previous dual-branch networks, which employ identical image features as inputs for both branches, we advocate for introducing some variation in the input images between the different branches. The SCFEM module initially transforms the spatial branch's image through frequency domain methods, enabling the capture of image features while accommodating diverse noise conditions. Finally, calibration is accomplished by merging the outputs of the dual branches.

Our proposal focuses on the concept of "guidance" in dual-branch learning. Earlier research on the interaction between dual branches mainly used feature fusion methods, involving merging the context branch's features into the spatial branch or vice versa. In contrast, our paper proposes the use of knowledge distillation as a technique for feature extraction. We facilitate the information exchange between the dual branches by guiding the learning of the spatial branch using deep knowledge from the context branch. We validate the effectiveness of the proposed method by conducting experiments to introduce these two concepts. Notable contributions of this research are summarized below:

1) We presented a dual-branch semantic segmentation framework designed to improve image semantic segmentation in UAV scenes. The approach focuses on enhancing information exchange between the two branches and increasing the spatial feature information extraction capacity.
2) The proposed architecture includes a spatial calibration extraction module. Calibration is an effective method for ensuring accuracy, and differential calibration surpasses repetitive calibration. Building upon this concept, the spatial calibration module introduces variations using frequency domain processing and accomplishes differential learning through two branches having distinct structures.
3) A spatially-guided loss function is proposed, intended to deliberate the learning of the spatial branch by leveraging the compression knowledge derived from deep features of images in the context branch. This approach promotes harmonious learning between the two branches of the dual-branch network during the training phase.
4) This study verified the generalization of the method on three datasets by integrating the core module of DBCGNet into three typical dual-branch networks. The proposed method demonstrated improvements compared to the baseline networks. Through comparative experiments, this method demonstrates certain competitiveness over other methods.

The remainder of this paper is structured as follows: Section II discusses existing methods pertinent to the research. Section III provides a comprehensive explanation of the methodology proposed in this study. Section IV delves into detailed experimental results and corresponding analyses. Finally, conclusions drawn from this research are concisely summarized in Section V.

II. RELATED WORK

This section provides a concise discussion of methodologies proposed by researchers in the field of spatial feature extraction, exploring the evolutionary development of dual-branch networks.

A. Spatial Feature Extraction

Previous research [15] has indicated that an effective semantic segmentation network should exhibit the following key attributes: 1. A robust backbone network with profound extraction capabilities. 2. The integration of multiscale features. 3. Implementation of a spatial attention mechanism. 4. Minimal computational complexity. Previous networks have consistently aimed to extract spatial information. For instance, [27] introduced a Context Propagation Network (CPN) guided by spatial details. This network employs strategies focused on propagating spatially detailed context, using superficial spatial details to guide the dispersion of low-resolution contextual data, thereby reducing computational costs. [28] proposed an improved dual-branch network, accompanied by a multi-scale context fusion module (DAPPM), which operates effectively at low resolution. One branch concentrates on extracting contextual features, while another emphasizes the retrieval of spatial information. Subsequently, these contextual and spatial features undergo fusion at various scales through a multi-scale context fusion module. The study [29] proposes a model named HR-ASPP, incorporating precise extraction of spatial features as well as enhanced extraction of shape features. It suggests utilizing parallel multi-resolution images and a fusion structure at multiple scales for effective extraction of spatial information, resulting in obtaining more accurate positional information. the SSFA module is proposed [30] to efficiently map semantic information onto spatial features, thereby improving the saliency of the target region of interest. To obtain rich spatial features of different scales, [31] applied multi-scale 2-D Singular Spectrum Analysis (SSA) after dimensionality reduction with SPCA on the image.

Various methodologies outlined above utilize direct spatial feature extraction on the original image. Nevertheless, owing to the substantial redundancy inherent in spatial visual information, it may lead to heightened computational demands and the acquisition of superfluous redundant data. Consequently, parallel subsampling is utilized on the original image to enable a faster capture of spatial information whilst simultaneously filtering out superfluous details. Subsequent learning of the spatial data is guided by context information to foster better interaction between the two branches, thereby bolstering the model's learning capability.

B. Dual Branch Network

Presently, two fundamental framework structures dominate mainstream semantic segmentation networks: the encoder-decoder structure and the dual-branch structure. Within the encoder-decoder architecture, the encoder typically operates as a robust feature extraction network, deployed to extrapolate semantic features. It condenses spatial information from images through repeated convolutional processes, consequently yielding high-level semantic data. The decoder utilizes an upsampling technique to reinstate the feature map to the original dimensions of the image, enabling dense image predictions. Standard encoder-decoder architectures, exemplified by UNet [32], incorporate ancillary lightweight lateral connections above the fundamental processes of encoding and decoding, facilitating upsampling to offset the deficiency in spatial data. Many subsequent methods are researched based on the foundation of encoding and decoding. Example models such as ERFNet [33] and ESPNet [34] utilize an encoder-decoder structure. While popular, this architecture has two principal limitations: Firstly, its full U-shaped configuration exacerbates computational requirements and notably hinders operational speed when employed with high-resolution feature maps. Secondly, the compression effect triggered by multiple encoding convolutions substantially diminishes the spatial information captured from an image, leading to significant information loss. Nonetheless, the skip connections employed in U-shaped structures merely serve as a palliative for spatial information loss, rather than providing a fundamental solution to the problem. To address the challenges inherent in the encoding-decoding framework, [21] proposed a dual-branch structure. Presently, numerous efficient methods utilize this structure, including Fast-SCNN [24]. This notable example incorporates a "learning downsampling" module to compute the low-level features across multiple resolution branches based on the dual-branch structure. BiSeNetV2 [22] advances the usage of global average pooling for context embedding and additionally introduces an attention-based fusion module. The CABiNet [35] utilizes a similar framework to that of the Fast-SCNN network and employs a more efficient encoding architecture, specifically, MobileNetV3 [36], during its encoding phase. In order to enhance the comprehensive integration of global and spatial information, [37] proposed DBFusion, a dual-branch network that combines convolutional neural networks and Swin transformers. The main advantage of this network lies in its ability to foster complementary interaction between local and global information through its unique dual-branch structure. The study by [38] encodes information regarding small objects by employing a dual-branch network along with a dual-mask branch to extract features using the dual-path masking technique. In previous work, each branch of a dual-branch network previously processed identical input images independently. These methodologies utilize a dual-pronged approach for feature extraction from an identical image. In applied scenarios, the optimal strategy involves scrutinizing the prevailing problem from multifaceted perspectives, subsequently endorsing and evaluating to diminish error rates. Hence, the dual-branch input images undergo processing with distinct forms of noise. Specifically, the context branch utilizes the pristine image for feature extraction, while the spatial branch employs a filter on the same image to yield an alternative noisy output. This strategy facilitates feature extraction from multiple viewpoints by leveraging the anomaly between the two branches. The dual-branch network consists of two branches that operate independently. This is evident in models such as BiSeNetV1 [21] and BiSeNetV2 [22]. Recent studies have initiated an exploration into information sharing

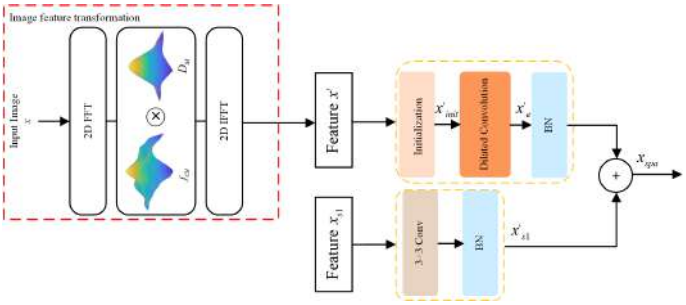

Fig. 2. The workflow diagram of the Spatial Calibration Feature Extraction Module comprises two main parts. The first part, known as "Filter," is primarily employed for extracting features under different noise conditions. The second part, referred to as "Fusion," assumes the main responsibility of integrating contextual and spatial branch information.

between these two distinct branches of the dual-branch network. Nonetheless, a noteworthy discrepancy exists between deep features and shallow features. Their hasty integration could potentially yield counter-productive results. As indicated by the Gate-SCNN model [25], the suggested context branch has attained an elevated comprehension of semantic scene understanding. The approach employed here diverges from a conventional process where two branches of features are reciprocally merged; rather, high-level semantic informational features direct the shallow features' development. This paper aims to enhance the spatial branch's feature extraction ability, guiding it towards learning spatial features through the lens of the context branch's features.

III. METHODOLOGY

This study introduces a dual-branch calibration guided deep learning architecture designed for semantic segmentation of high-resolution remote sensing images gathered from UAVs. This section delineates each module within the proposed deep learning structure in detail.

A. Overall Architecture

The DBCG-Net deep learning framework primarily comprises two core components: the Spatial Calibration Feature Extraction Module (SCFEM) and the spatially-guided loss function(SGL). In the initial phase of the framework, the system acquires knowledge primarily via two branches that utilize different noise and learning strategies, processing the same image from distinct perspectives. Subsequently, in the second phase, integration of information from both branches allows for successful calibration and validation functions. Guided by a deep comprehension of image features contextualized by the teacher model, the spatially-guided loss function directs the acquisition of shallow spatial details in the spatial branch. This methodology fosters information sharing between both branches. while subsequent sections introduce two key proposed modules: the Spatial Calibration Feature Extraction Module(SCFEM) and the spatially-guided loss function(SGL).

B. Spatial Calibration Feature Extraction Module

This study proposes the adoption of distinct learning strategies for spatial branches, enabling the transformation of the supplied input image $x' \subset R^{H\times W\times C}$. Such a method allows both the spatial and context branches to acquire images under a diverse range of noise conditions. Spatial features are extracted via the spatial branches of disparate learning frameworks. The context branch serves to guide spatial branch learning, thereby supplanting the reciprocal integration previously occurring between both branches. The flowchart as a whole of the Module for Spatial Calibration Feature Extraction can be seen in Fig. 2

Primarily, achieve $x' \subset R^{H\times W\times C}$ by transmuting the provided input image, denoted as $x' = F_T(x)$. F_T denotes the transformation function that converts x into x'. The Fast Fourier Transform (FFT) algorithm is applied to the image, resulting in f_{CH}. Construct a transformation matrix, $D_M \subset R^{H\times W}$, after multiplying the acquired f_{CH} with the transformation matrix $D_M \subset R^{H\times W}$, take the absolute value of the fast Fourier inverse transform to obtain a new image $x' \subset R^{H\times W\times C}$. The transformation function is defined as:

$$X' = |IFFT(FFT(X)D_M)| \tag{1}$$

Construct the transformation matrix, $D_M \subset R^{H\times W}$. The primary purpose of this matrix is to filter out noise. Once filtered, the features of various noise types can be obtained for use as inputs. To construct $D_M \subset R^{H\times W}$, it is essential to determine the distance from each point in the image to the central point. Here, d and n serve as hyperparameters that aid in adjusting the extent of noise removal. Let L_{dis} denote the Euclidean distance from each point to the central point. In this context, C denotes the value of the central point, while $x_i, x_j \subset R^{H\times W}$ represents the values of all other points.

$$D_M = \frac{1}{(1+(L_{dis}/d))^{2n}} \tag{2}$$

$$L_{dis} = \sqrt{(C_i - x_i)^2 + (C_j - x_j)^2} \tag{3}$$

The learning of space information characteristics is steered by context branching, which in turn informs spatial branching. Given the substantial redundancy inherent in the original image data, it becomes vital to compress the transformed image x', into a more efficient representation. Doing so allows for enhanced capture of spatial features. Consequently, parallel initialization operations are utilized to condense the original data. Specifically, downsampling and max-pooling operations are executed independently on the original dataset. Following the procurement of the preliminary feature map, termed as $x'_{init} \subset R^{c\times\frac{R}{4}\times\frac{R}{4}}$, the secondary feature map (denoted as $x'_e \subset R^{c\times\frac{R}{4}\times\frac{R}{4}}$) is acquired via the implementation of dilated convolution. The objective of this phase is to enhance the spatial receptive field. By directing the learning process of the spatial branch, the context branch can circumvent the issue of guidance difficulty that arises due to a substantial divergence between the two branches. At this stage, the feature map $x_{s1} \subset R^{C\times\frac{R}{4}\times\frac{R}{4}}$, originating from the initial convolutional layer of the context branch, is combined with x'_e to form the feature map x_{spa} This resultant feature map is subsequently processed through an activation function, thereby deriving spatial information features. The entire process can be illustrated as follows:

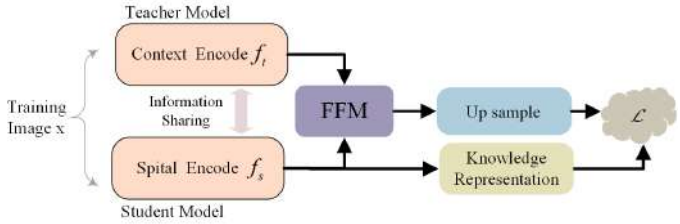

Fig. 3. Diagram illustrating the framework of the spatially-guided loss function (SGL).The proposed architecture employs deep semantic features to guide spatial branch learning. In this study, we denote the SGL method proposed in this paper as $\mathcal{L}$.

$$x_{spa} = \sigma(cat(conv(F_{init}(x'), x_{s1}))) \tag{4}$$

C. Spatially-Guided Loss Function

Neural networks predominantly generate class probabilities for each category utilizing a "softmax" output layer. The outputs from this layer are transformed from variable z_i to probability p^S.

$$p_i = \frac{z_i/\tau}{\sum_j exp(z_j/\tau)} \tag{5}$$

Herein, τ represents temperature, typically assigned a value of 1. An increase in the value of τ results in a more diluted class distribution. Knowledge Distillation (KD) predominantly employs Kullback-Leibler divergence to define the loss function. It posits the spatial branch as the student model, denoting the context branch as the teacher model. Accordingly, it leverages the context branch for guiding the learning process of the spatial branch. Hence, there is a modification in the original loss function, which is illustrated in Equation (6).

$$\mathcal{L}_{KD} = KL(P^C \parallel p^S) = \sum_{i=1}^{C} p_i^C log\left(\frac{p_i^C}{p_i^S}\right) \tag{6}$$

In this context, p^C symbolizes the class probability resulting from the transformation of context branch's output layer z_i. Concurrently, p^S represents the class probability derived from the transformation of the spatial branch's output layer z_i.

An additional dice loss function has been implemented to address the issue of class imbalance arising from the rough outcomes produced by the cross-entropy loss function. Because this function is not sensitive to either foreground or background information, it can effectively mitigate class imbalance.

The overall loss function is characterized as follows:

$$\mathcal{L}_{dice}(p_d, g_d) = 1 - \frac{2\sum_i^{H\times W} p_d^i g_d^i + \epsilon}{\sum_i^{H\times W}(p_d^i)^2 + \sum_i HW(g_d^i)^2 + \epsilon} \tag{7}$$

$\mathcal{L}_{oral}(\cdot)$ denotes the original loss function of the model, Fig. 3 illustrates the schematic diagram of the Spatially-guided Loss (SGL) method, p_d signifies the final prediction result, and g_d stands for Ground Truth. λ is a key hyperparameter; empirical evidence suggests optimal outcomes when this parameter is set to 0.1. Hence, in the experimental section, the default value for λ is established as 0.1. $\mathcal{L}_{dice}(p_d, g_d)$ denotes the dice loss function, which is given as follows:

$$\mathcal{L} = \mathcal{L}_{oral}(p_d, g_d) + \lambda\mathcal{L}_{KD}(P^C, P^S) + \beta\mathcal{L}_{dice}(p_d, g_d) \tag{8}$$

IV. Experiments

A. Dataset and Evaluation Metrics

1) Evaluation Metrics: A set of experiments will be carried out using three distinct datasets: UDD6 [39], UAVid [40], and Potsdam. These datasets will primarily serve for both ablation studies and comparative analyses. The network optimization will employ the Stochastic Gradient Descent (SGD) optimizer, configured with a learning rate of 0.01 and a momentum set to 0.9. Ubuntu 18.04 served as the operating system for all experiments. The code was adapted, and experiments were conducted using mmsegmentation. The system specifications included CUDA 11.1, CuDNN 8.0.5, OpenCV 4.6.0, MMCV 1.3.16, and Python 3.7. All experiments were performed on an NVIDIA GeForce RTX 3080 GPU with a batch size of 1 to assess the performance of different methods. The training cycle comprises 300 epochs, with the minimum learning rate identified as 1e-4. Details regarding the specific pre-processing of images are thoroughly addressed in the introductory segment of the dataset.

The metric employed for performance evaluation is the Mean Intersection over Union (mIoU), representing the average IoU across all categories within the dataset. The mathematical formulation of mIoU can be defined as follows:

$$mIoU = \frac{1}{k}\sum_{i=1}^{k}\frac{TP}{FN + FP + TP} \tag{9}$$

In this study, "TP" denotes True Positives, "FP" represents False Positives, "TN" indicates True Negatives, and "FN" refers to False Negatives.

2) Dataset: The UDD dataset is a collaborative effort involving Peking University, Huludao City, Henan University, and Cangzhou City, amalgamating various unmanned aerial vehicle image datasets. This dataset encompasses five primary categories: vegetation, buildings, roads, vehicles, and backgrounds. The images in this dataset were segmented into dimensions of 512×512. Post-segmentation, 3880 images were assigned to the training set, while 1280 images were allocated to the test set.

The UAVid dataset represents a comprehensive, high-resolution semantic segmentation dataset derived from UAV imagery, particularly focusing on urban street scenes at a resolution of 3840×2160. This dataset includes 420 images divided into three distinct sets: 200 images for the training set, 70 images for the validation set, and the remaining 150 images for the testing set. Given its intricate scenes and high resolution, it serves as a challenging benchmark. The image size maintained during the training process is 512×1024 pixels. Importantly, no data augmentation techniques were applied at any stage of the process.

The Potsdam dataset consists of high-resolution images (6000×6000 pixels) with a Ground Sampling Distance (GSD)

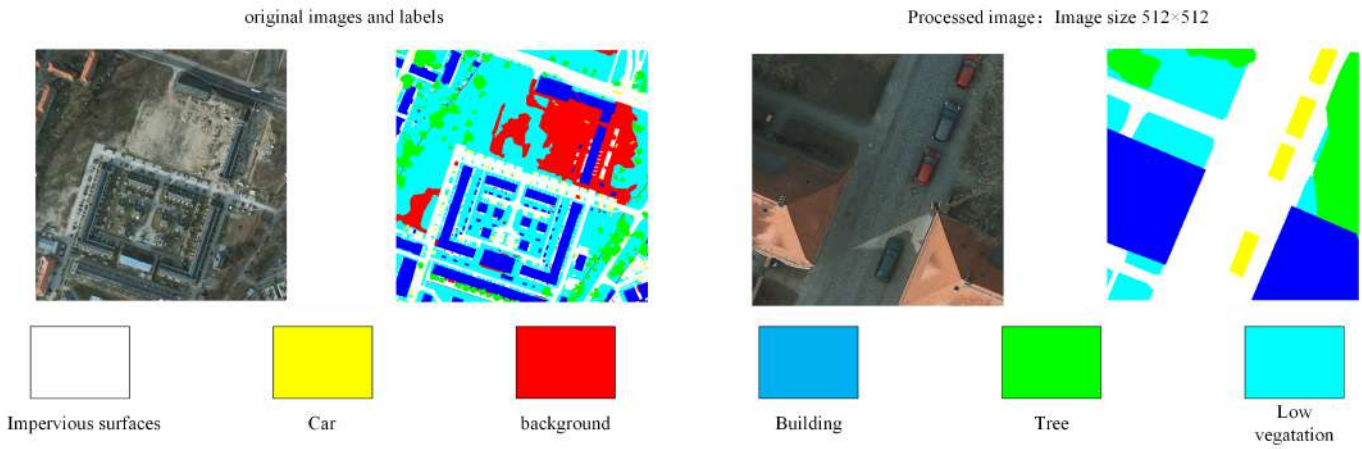

Fig. 4. Example images and labels from ISPRS Potsdam dataset and Processed Example images, The figure comprises color-coded labels, original images along with their corresponding labels, and processed images employed in the training process.

TABLE I
SPATIAL CALIBRATION FEATURE EXTRACTION MODULE ABLATION EXPERIMENT

Method	mIoU(UDD)	mIoU(UAVid)	mIoU(Potsdam)
BiSeNetV1	67.4	61.1	69.7
BiSeNetV1+SCFEM	67.9	64.6	70.3
BiSeNetV2	64.2	64.5	67.8
BiSeNetV2+SCFEM	65.6	65.3	68.3
STDC	67.2	65.4	70.1
STDC+SCFEM	**69.8**	**66.1**	**71.1**

TABLE II
SPACE SHARING LOSS FUNCTION ABLATION EXPERIMENT

Method	mIoU (UDD)	mIoU (UAVid)	mIoU (Potsdam)
BiSeNetV1+SCFEM	67.9	64.6	70.3
BiSeNetV1+SCFEM+SGL	68.2	66.1	71.0
BiSeNetV2+SCFEM	65.6	65.3	68.3
BiSeNetV2+SCFEM+SGL	66.8	68.8	68.4
STDC+SCFEM	69.8	66.1	71.1
STDC+SCFEM+SGL	**72.8**	**67.6**	**71.4**

of 5 cm, ensuring exceptional clarity. During the training phase, 3465 images from this dataset were used. Due to the resource-intensive nature of handling 6000×6000 pixel images during training, they were resized to 512×512 pixels. The remaining images constituted the validation set and underwent identical processing as the training ones. Refer to Fig. 4 for examples of the cropped dataset's images and annotations.

B. Ablation Experiments

The ablation study conducted in this article seeks to validate the efficacy of the Spatial Calibration Feature Extraction Module (SCFEM) and the Spatially-guided Loss Function (SGL) proposed in our methodology. Separate ablation studies will be performed on BiSeNetV1, BiSeNetV2, and STDC. The detailed testing results are presented in Table I.

TABLE III
ABLATING EXPERIMENT ON THE SCFEM MODULE

Filter	Fusion	mIoU (UDD)	mIoU (UAVid)	mIoU (Potsdam)
×	×	58.9	55.0	57.8
×	✓	59.2	59.5	66.1
✓	✓	**64.6**	**67.9**	**70.3**

Table I shows the improved model. It achieved an accuracy of 69.8% on the UDD dataset, 66.1% on the UAVid dataset, and 69.8% on the Potsdam dataset, which reflects an enhancement compared to the baseline. Initially, the top two rows of the table indicate ablation studies conducted on BiSeNet. With the advent of the UAVid dataset, an improvement of 3.5% is observed relative to the original methodology. In the Potsdam dataset, we observed a 0.6% improvement over the original method, whereas the UDD dataset displayed a 0.5% enhancement when compared to the initial approach.The central two rows of the table pertain to ablation experiments conducted on BiSeNetV2, resulting in a performance improvement of 1.4% on the UDD dataset. The model further exhibited an advancement of 0.8% on the UAVid dataset and a 0.5% enhancement on the Potsdam dataset. The final two lines highlight the enhancements achieved on STDC. Consequently, there is a recorded increase of 2.6% on the UDD dataset, an improvement of 0.7% on the UAVid dataset, and progress of 1.0% on the Potsdam dataset. As evidenced by the integrated analysis, the proposed Spatial Calibration Feature Extraction Module (SCFEM) asserts its superiority over the existing baseline, thereby affirming its effectiveness.

The outcomes of the ablation experiment for SGL are depicted in Table II. This study is performed for comparison purposes using SCFEM as a basis, given that SGL is an enhanced version of SCFEM. As such, there have been no ablation experiments to contrast the initial method with SGL.

In Table II, the enhanced methodology yields a score of 72.8% on the UDD dataset, 67.6% on the UAVid dataset, and 71.4% on the Potsdam dataset. This table's initial two

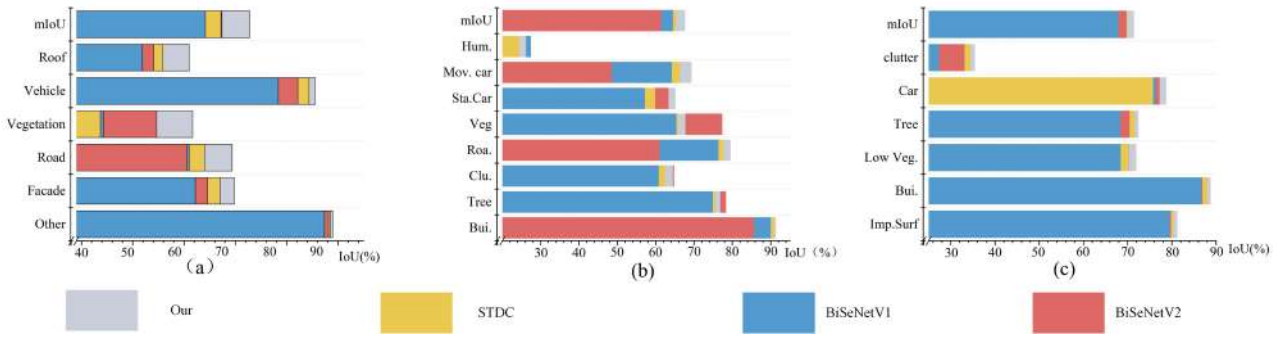

Fig. 5. Ablation experiments on each category, This study presents a comparative analysis between three dual-branch networks and the proposed method, utilizing incremental bar charts. (a) The performance of the proposed method was evaluated on the UDD dataset. (b) The performance of the proposed method was evaluated on the UAVid dataset. (c) Performance analysis was carried out on the Potsdam dataset. The proposed method is depicted in gray on the graph.

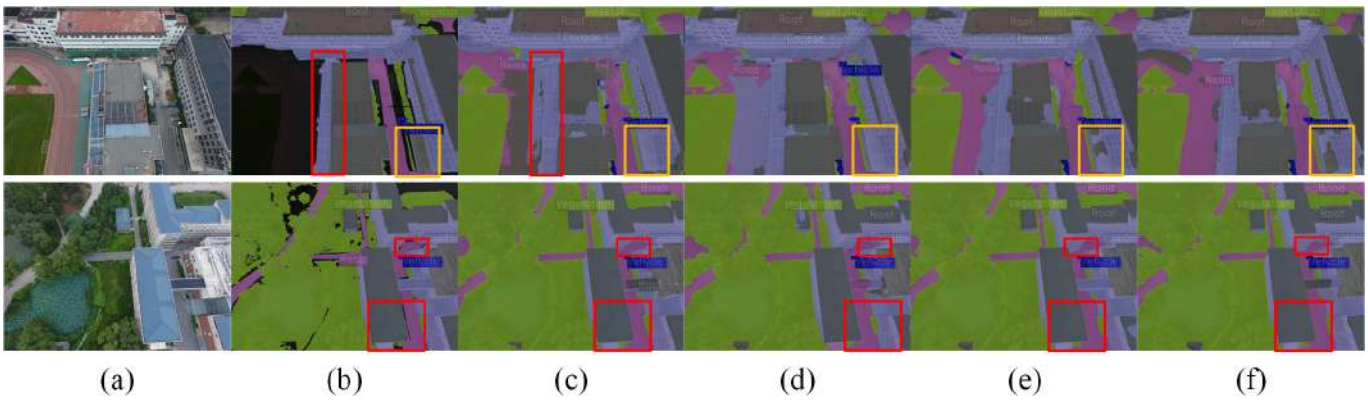

Fig. 6. Comparative visualization between the model inference and ground truth on the UDD test dataset. (a) Image. (b) Ground truth. (c) BiSeNetV1. (d) BiSeNetV2. (e) STDC. (f) DBCG.

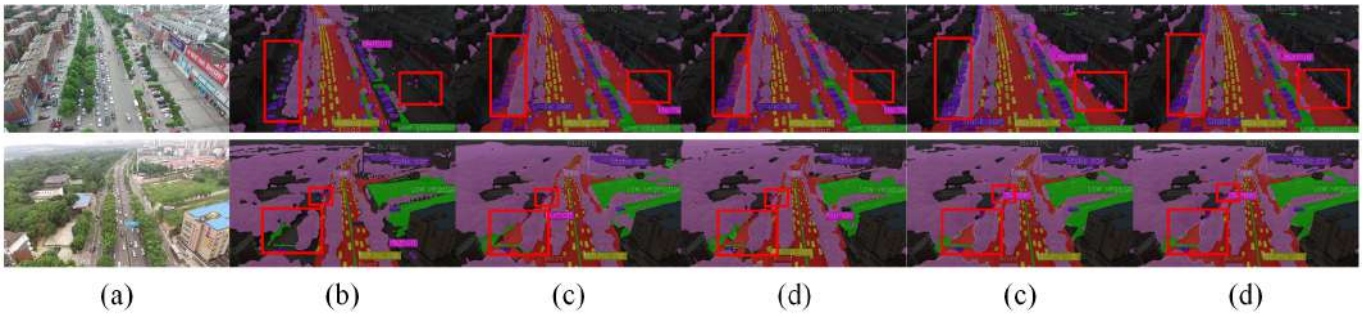

Fig. 7. Comparative visualization between the model inference and ground truth on the UAVid test dataset. (a) Image. (b) Ground truth. (c) BiSeNetV1. (d) BiSeNetV2. (e) STDC. (f) DBCG.

rows present a comparison of their effects on BiSeNet. The results showed a 0.3% increase in the UDD dataset, a 1.5% increase in the UAVid dataset, and a 0.7% increase in the Potsdam dataset. The center rows of the table illustrate an enhancement in BiSeNetV2's performance, indicated by a surge of 1.2% on the UDD dataset, 3.5% on the UAVid dataset, and a marginal increase of 0.1% on the Potsdam dataset. The final two sentences present comparative results from the STDC study. An improvement of 3% was observed in the UDD dataset, with a 1.5% and a slight 0.3% increase recorded for the UAVid and Potsdam datasets respectively. Through in-depth analysis, we found that the spatial sharing loss function has been optimized. However, the results on several datasets did not show statistically significant effects. This indicates that there are still differences between the teacher model and the student model, thereby affecting the student's ability to understand the teacher model. Exploring this issue will be one of the key focuses of future research.

In Table III We performed ablative experiments on two modules of SCFEM, utilizing the backbone is Resnet18. Firstly, upon comparing the initial two rows of the comparative table, it becomes evident that the spatial branch significantly

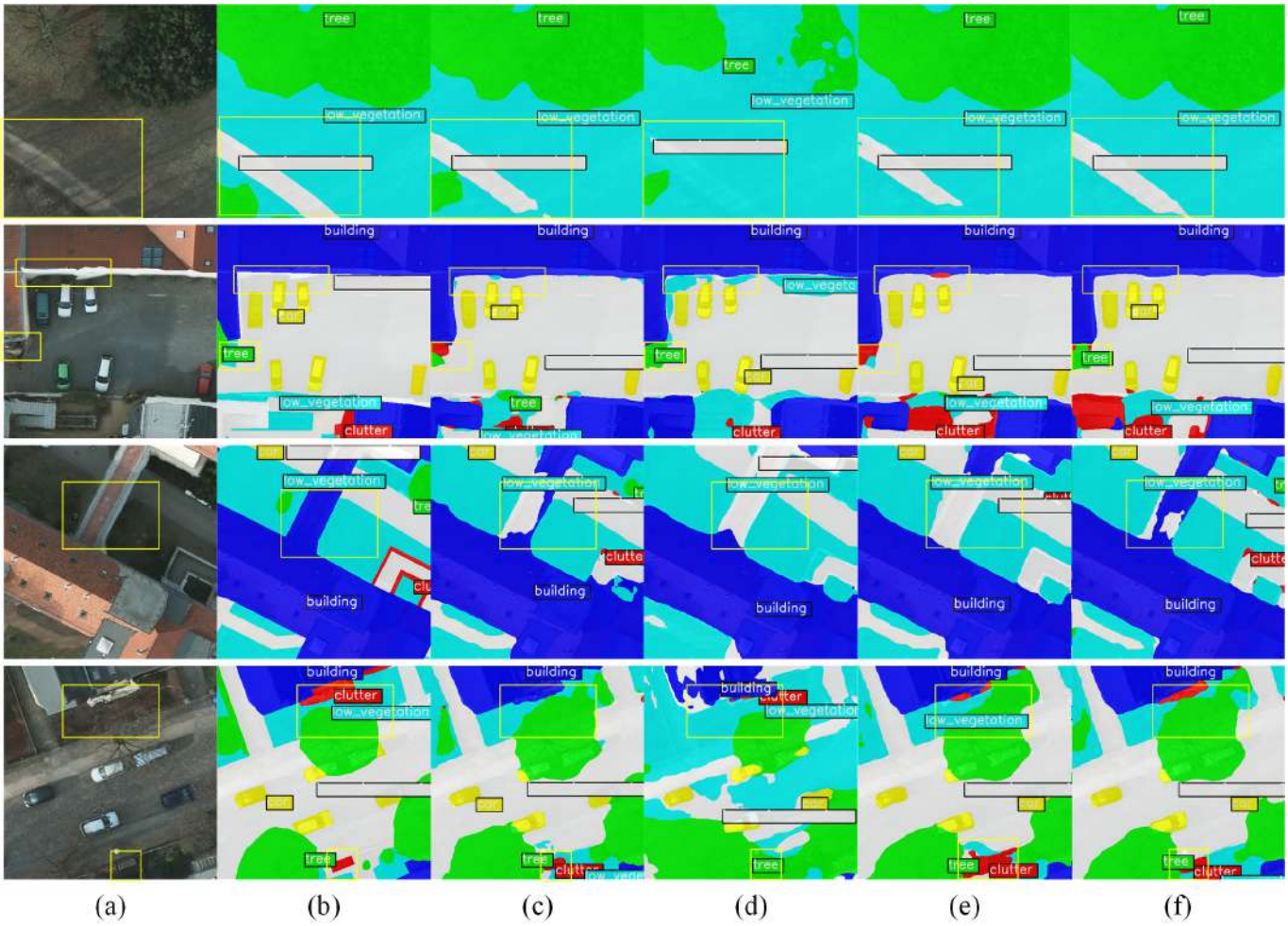

Fig. 8. Comparative visualization between the model inference and ground truth on the Potsdam test dataset. (a) Image. (b) Ground truth. (c) BiSeNetV1. (d) BiSeNetV2. (e) STDC. (f) DBCG.

TABLE IV
EXPERIMENTAL INVESTIGATION OF LAYER ABLATION BETWEEN DUAL BRANCHES

Layer	mIoU (UDD)	mIoU (UAVid)	mIoU (Potsdam)
None	69.5	62.4	67.7
Layer1	**72.8**	**67.6**	**71.4**
Layer2	72.0	63.1	69.6
Layer3	72.3	63.3	70.4
Layer4	71.5	63.2	70.1

influences the performance of the dual-branch network. Not including the spatial branch adversely affects the model's performance. Secondly, upon comparing the final two rows in the subsequent tables, the importance of filters for the overall SCFEM module becomes apparent. Additionally, this finding supports the notion that obtaining features under diverse noise conditions enhances the differential detection method.

In Table IV an ablation experiment was conducted to explore the optimal layer for sharing by comparing the contributions of different layers between the contextual branch and the spatial branch. Comparisons with no sharing and other layers indicate that layer sharing between the dual branches improves model performance. When comparing Layer 1, Layer 2, Layer 3, and Layer 4, we conclude that the sharing effect is optimal when applied to Layer 1. This is because the first layer retains a significant amount of detailed information, which is inevitably lost during the learning process of the spatial branch. Sharing the first layer helps compensate for this loss of information. Fig. 9 displays the visualization results of the shared layers for the UDD6 dataset. Refer to Fig. 11 for the visualization results of the shared layers for the UAVid dataset. Fig. 11 illustrates the visualization results for the Potsdam dataset. The visual analysis confirms that Layer1 performs optimally, displaying precise detection coverage and finer edges than other layers.The second figure, as shown in Fig. 11, demonstrates that the red low vegetation other layer is classified as "clu.", whereas Layer1 accurately identifies it.

C. Ablation Experiments on Each Category

This section examines the effects of the proposed method on each category. The method will be evaluated for its impact on each category. Ablation experiments for each category

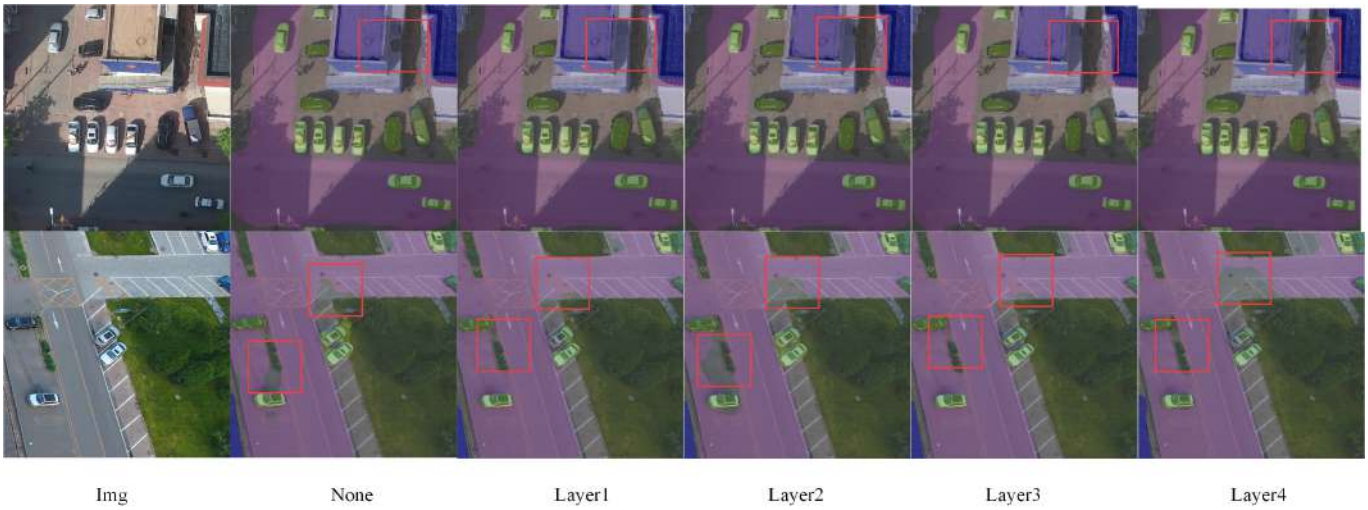

Fig. 9. Comparing the visualized results obtained from different shared layers using the UDD6 test dataset.

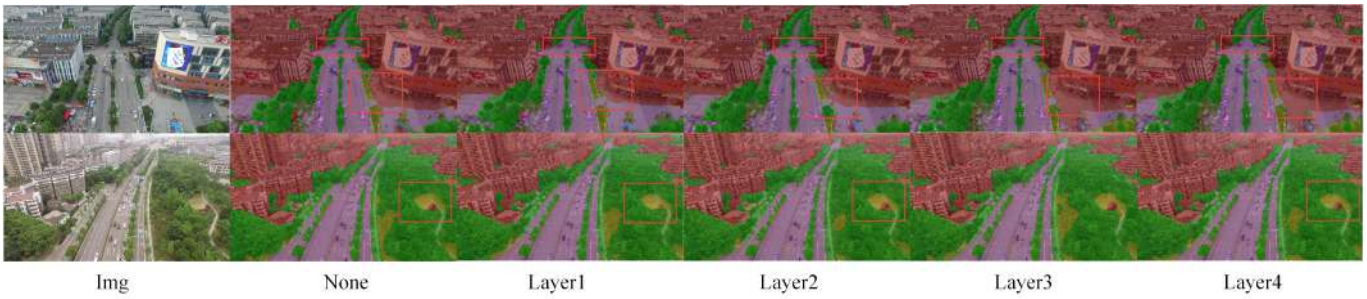

Fig. 10. Comparing the visualized results obtained from different shared layers using the UAVid test dataset.

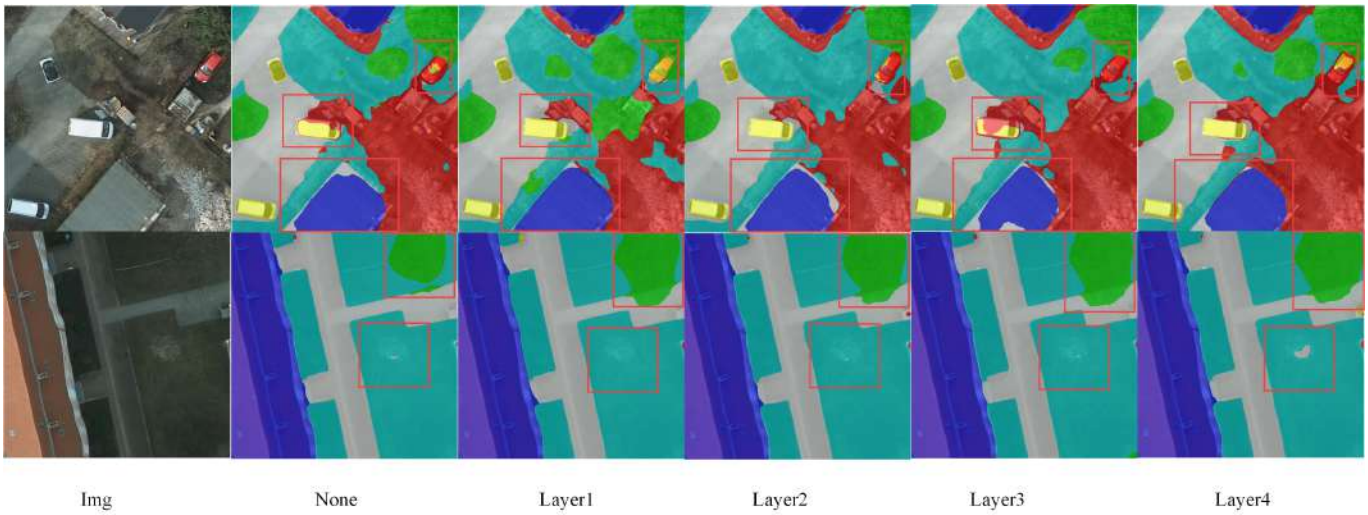

Fig. 11. Comparative visualization between the model inference and ground truth on the Potsdam test dataset.

will be depicted using incremental bar charts as a visual representation, as illustrated in Fig. 5.

Among them, (a) indicates the performance on the UDD dataset. The graph clearly indicates that the gray color corresponds to the proposed method in this paper. The prevalence of gray color at the top clearly demonstrates that the proposed method achieves the highest accuracy in all categories. Moreover, notably in the roof, vegetation, and Rood categories, there is a substantial increase in the gray color area, indicating significant enhancements achieved by the proposed method across these specific categories. Additionally, it is evident that the categories experiencing the most notable changes are the ones with initially low accuracy. Consequently, the proposed method effectively improves the accuracy for these previously

TABLE V
COMPARATIVE EXPERIMENTAL RESULTS ON THE UAVID DATASET

Method	Bui.	Tree	clu.	Roa.	Veg.	Sta.Car.	Mov.Car	Hum.	mIoU
SexNext [41]	86.0	66.4	53.0	66.4	59.3	29.6	35.5	3.1	50.2
Topformer [42]	85.9	67.3	52.7	68.4	42.2	67.3	57.5	2.2	53.4
Seaformer [43]	83.7	67.1	49.1	67.0	56.9	41.1	50.1	3.7	52.3
Fast-SCNN [24]	87.7	72.6	56.9	73.9	63.8	47.1	59.6	15.5	59.6
BiSeNetV2 [22]	90.1	75.0	60.9	76.5	65.5	57.3	64.3	**27.4**	64.6
APCNet [26]	90.6	76.2	61.9	76.2	66.7	61.0	65.7	24.9	65.3
STDC [23]	91.0	75.6	77.7	62.5	65.9	59.8	66.4	24.5	65.4
ICNet [44]	90.3	75.3	61.8	66.5	**77.3**	60.5	68.0	25.4	65.6
BiSeNetV1 [21]	85.7	78.3	64.7	61.1	**77.3**	63.4	48.6	17.5	61.5
ShelfNet [45]	76.9	73.2	44.1	61.4	43.4	21.0	52.6	3.6	47.0
BANet [46]	85.4	**78.9**	66.6	**80.7**	62.1	52.8	69.3	21.0	64.6
SwiftNet [52]	85.3	78.2	64.1	61.5	76.4	62.1	51.1	15.7	61.1
Ours	**91.3**	76.8	**79.5**	64.4	67.7	65.1	**69.3**	26.1	**67.6**

TABLE VI
COMPARATIVE EXPERIMENTAL RESULTS ON THE POTSDAM DATASET

Method	imp.	Bui.	Low. Veg.	Tree	Car	clutter	mIoU
SexNext [41]	79.8	87.3	71.2	**73.2**	77.5	34.0	70.5
Topformer [42]	66.1	65.2	**77.3**	71.5	60.0	**55.3**	65.9
Seaformer [43]	59.9	63.2	64.6	64.9	57.4	48.8	59.8
Fast-SCNN [24]	72.5	80.2	60.3	52.8	60.1	9.7	55.9
BiSeNetV2 [22]	79.5	86.7	68.5	68.3	76.3	27.3	67.8
STDC [23]	80.4	88.1	70.1	71.7	75.7	34.5	70.0
ICNet [44]	78.2	84.9	64.4	65.3	78.5	25.5	65.4
BiSeNetV1 [21]	79.9	87.0	70.3	70.5	77.3	33.2	69.8
APCNet [26]	77.8	84.5	67.0	71.2	78.9	25.0	67.6
ANNNet [47]	79.8	87.4	69.8	71.2	**79.1**	35.9	70.5
PSANet [48]	77.7	83.6	66.2	68.2	76.7	27.8	66.7
PSPNet [49]	79.5	86.8	68.6	69.0	76.7	29.8	68.4
Ours	**81.2**	**88.7**	71.9	72.4	78.7	35.4	**71.4**

TABLE VII
EXPERIMENTAL RESULTS ON THE UDD6 DATASET COMPARED TO THE PROPOSED METHOD

Method	Other	Fac.	Roa.	Veg.	Veh.	Roof	mIoU
SexNext [41]	88.9	**71.1**	66.6	60.5	**87.1**	59.3	72.3
Topformer [42]	84.9	63.5	54.6	45.6	80.7	47.8	63.2
Seaformer [43]	82.9	49.0	47.5	34.4	71.2	30.1	52.5
APCNet [26]	84.3	65.9	56.8	42.7	76.5	50.7	62.3
BiSeNetV2 [22]	87.4	62.2	61.0	44.3	78.4	51.9	64.2
ICNet [44]	87.5	63.2	57.3	47.2	81.1	52.7	64.8
STDC [23]	88.5	67.0	64.0	43.7	84.4	55.9	67.2
BiSeNetV1 [21]	88.2	64.6	60.6	54.7	82.2	54.1	67.4
ANNNet [47]	87.7	67.0	63.7	53.0	79.6	54.1	67.6
DANet [50]	88.0	62.6	65.4	58.2	80.6	56.4	68.6
Fast-SCNN [24]	**89.2**	68.4	66.7	49.2	84.6	59.0	69.5
CCNet [51]	88.9	68.9	65.8	**63.9**	85.3	60.0	72.2
Ours	89.0	69.8	**69.4**	61.7	85.6	**61.0**	**72.8**

low-accuracy categories in the UDD dataset. Notably, BeSeNetV2 demonstrates the lowest performance on the UDD dataset, exhibiting lower accuracy compared to BeSeNetV1. The detection results of STDC in the Vegetation category fail to attain satisfactory performance. Fig. 6 displays the visualization diagram. In the visualization of the first figure, inside the red box, other categories often detect a broader region along the road edge and classify it as Road, whereas our method closely matches the ground truth (GT) range. Within the orange box, roofs have a tendency to be misclassified as roads, however, our method accurately recognizes them as roofs in contrast to other methods. The second visualization diagram showcases that our method exhibits finer edge details within the red box in comparison to other methods. (b) showcases the performance on the UAVid dataset. Although there is a minor decrease in accuracy compared to BiSeNetV1 in specific categories like Hum., clu., and Tree, it consistently outperforms BiSeNetV1 in all other categories. Nevertheless, there is only a marginal difference in overall performance area when compared to BiSeNetV1, while its accuracy significantly exceeds that of BiSeNetV1 in other categories. Fig. 7 displays the visualization diagram. The first visualization diagram depicts that small vehicles may be missed or incorrectly detected in alternative methods, however, our approach ensures more precise segmentation. In the second figure, it illustrates how our method attains a distinct boundary between stationary cars and roads. Our method exhibits higher accuracy in detecting human edges. Alternatively, other methods display both missed detections and coarse linear edges in human subjects. (c) illustrates the effectiveness of our proposed method on the Potsdam dataset. The graph clearly shows that each category is positioned on the far right side, indicating the excellent performance attained by the approach presented in this paper. Notably, BiSeNetV2 displays the poorest performance, whereas STDC underperforms specifically in the car category.Despite surpassing the remaining three dual-branch networks, the extent of improvement indicated by the gray color area suggests that the achieved enhancement with our method is relatively insignificant. Hence, there is still ample scope for improvement. Fig. 8 displays the visualization diagram. The first and third figures reveal that the proposed method achieves a more comprehensive detection of roads and buildings in comparison to the other three methods. The second figure demonstrates lower rates of misjudgment compared to the other three methods. For instance, BiSeNetV1 and STDC misclassified tree objects as clutter. Some parts of the road were misclassified as low vegetation by BiSeNetV2. The fourth figure demonstrates that the proposed method produces

more detailed results for smaller clutter when compared to the aforementioned three methods.

D. Comparative Experiments

In this section, a comparison will be made between the network proposed in this paper, and existing semantic segmentation algorithms like Fast-SCNN and ICNet. The goal is to obtain an objective statistical evaluation that demonstrates the effectiveness of the proposed architecture. Quantitative analysis of the proposed method will be conducted based on Intersection over Union (IOU) and mean Intersection over Union (mIoU) metrics for each category.

The performance of the proposed method on the UAVid dataset is presented in Table V. The table clearly demonstrates that the proposed method surpasses other approaches, achieving an accuracy of 67.6%. Our approach achieves the highest accuracy among all methods for the categories Bui., Clu., and Mov. Car, obtaining percentages of 91.3%, 79.5%, and 69.3%. It is worth noting SexNeXt [41], Topformer [42], and Seaformer [43] have demonstrated promising performance in urban street scenes. However, experimental results indicate that these models do not perform well on datasets involving unmanned aerial vehicles (UAVs). In particular, the detection of human-related categories in the UAVid dataset exhibits generally low accuracy, leading to an overall decline in performance. Conversely, the method presented in this paper exhibits exceptional performance on the dataset collected by drones, highlighting the merits of our approach in regard to drone data. After conducting an analysis, it is evident that the proposed method still has certain limitations. Since the method relies on convolutional neural networks (CNNs), it inevitably encounters a limitation commonly associated with CNNs - their limited capacity to integrate global information. Despite the overall advantages of the proposed method, it fails to fully integrate both global and local information. Moreover, the proposed method exhibits suboptimal accuracy when distinguishing between vegetation and trees. This discrepancy arises due to the top-down or side-view perspective typically captured by unmanned aerial vehicles (UAVs), which deviates from the conventional viewpoint observed in urban scenes. Analysis of the collected dataset reveals a diminished distinction between vegetation and trees within this particular viewing angle, consequently considerably augmenting the segmentation challenge for both categories. Future investigations aim to develop attention mechanisms applicable to unmanned aerial vehicle scenarios in order to tackle these challenges.

Table VI presents the results of the comparative experiment on the Potsdam dataset, with our proposed method demonstrating superiority over other network models. While some categories have not achieved the optimal level, the difference compared to the optimal category is relatively small. Consequently, an overarching conclusion can be drawn that our method exhibits superior overall performance.

Experiments comparing the selected networks were executed using the UDD dataset. The segmentation test results, as shown in Table VII, provide comparative insights into the performance of different networks. Upon evaluation, it becomes evident that our method outperforms other network models. Notably, the mIoU achieved was 72.8%. However, the accuracy of the "Veg." category is relatively lower when compared to SegNeXt. Upon analyzing these categories' accuracy, the variance appears insignificant. Notwithstanding, the improved method exhibits superior accuracy in other categories, translating to an overall advantage. By analyzing three datasets, the proposed method demonstrates several advantages. However, it also exhibits significant shortcomings, particularly in accurately categorizing vegetation, failing to meet the desired outcome. Recognizing this category becomes challenging due to the unique perspective offered by unmanned aerial vehicles (UAVs). To address and overcome this issue, further research will be undertaken.

V. Discussion

In this section, we will discuss the inspiration and background that our research contributes to future studies, as well as its impact on potential areas of interest.

Dual-branch calibration: We propose the introduction of the concept of calibration in a dual-branch network. Multiple solutions and perspectives are possible when solving a problem or understanding a phenomenon. If multiple perspectives or methods arrive at the same result, it is considered correct. Differential testing is a method of verification that enhances accuracy. Building upon this concept, we present SCFEM as our proposition. Further exploration opportunities exist regarding this concept, including the investigation of calibration between different models and the implementation of more refined differential testing techniques.

Dual-branch guidance: we introduce the concept of dual-branch guidance as the basis for our research. We propose a novel technique called SGL that aims to facilitate information interaction between the two branches. SGL not only considers knowledge distillation as a means of extracting knowledge from one model to another but also promotes the learning process within each branch of the model itself. The implications of this approach are significant and can inspire future research endeavors.

Looking Ahead: This study offers novel insights into the aspect of information interaction within the dual-branch architecture for semantic segmentation in drone technology, thereby serving as a source of inspiration. The ideas presented in this study expand the avenues for future research.

VI. Conclusions

This study introduces DBCG-Net to address UAV-based semantic segmentation in remote sensing images. We conduct an analysis of the inherent limitations of current dual-branch architectures, exemplified by BiSeNet, and manage these deficiencies using two key architectural modifications. The Spatial Calibration Feature Extraction Module efficiently tackles the problem of insufficient feature extraction in spatial branches when processing high-resolution images. By utilizing different types of noise, the reciprocity between the two branches of the dual-branch network facilitates mutual error correction. Additionally, the spatially-guided loss function conceptualizes

the contextual branch as a teacher model and the spatial branch as a student model. This framework encourages learning transfer from the contextual to the spatial branch. The efficacy of this proposed method has been validated through empirical testing. The methods presented here have been substantiated to exhibit enhancements over the established baseline. Moreover, they achieved an accuracy of 67.6% on the UAVid dataset, 71.4% on the Potsdam dataset, and 72.8% on the UDD dataset.

Our future work will address the following two aspects. First, the experimental results indicate suboptimal performance in the tree and vegetation categories. This can be attributed to the unique perspective of UAVs, which results in nuanced variations in category characteristics. Second, the differing structures between the two branches prevent the student model from fully comprehending the knowledge transferred from the teacher model. Consequently, the improvement effect of SGL has failed to meet the expected outcome. Based on these analyses, we aim to employ more sensitive feature extraction methods, enhance the learning approach, and improve information transfer efficiency between the teacher and student models to resolve these concerns.

REFERENCES

[1] S. Yi, X. Liu, J. Li and L. Chen," Uavformer: a composite transformer network for urban scene segmentation of uav images,"*Pattern Recognit.*, vol. 133, pp. 109019, 2023.

[2] E. Alvarez-Vanhard, T. Corpetti and T. Houet, "Uav & satellite synergies for optical remote sensing applications: A literature review,"*Sci. Remote Sens.*, vol. 4, pp. 100019, 2021.

[3] A. Singh, H. Kalke, M. Loewen and N. Ray," Riverice segmentation with deep learning,"*IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 7570–7579, 2020.

[4] T. Xie, J. Li, C. Yang, Z. Jiang, Y. Chen, L. Guo, and J. Zhang, " Crop height estimation based on uav images: Methods, errors, and strategies,"*Comput. Electron. Agric.*, vol. 185, pp. 106155, 2021.

[5] D. Ogawa, T. Sakamoto, H. Tsunematsu, T. Yamamoto, N. Kammo, Y. Nonoue, and J. Yonemaru, " Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8929–5238, 2019.

[6] M. Fayyaz, M. H. Saffar, M. Sabokrou, M. Fathy, F. Huang, and R. Klette."Stfcn: spatio-temporal fully convolutional neural network for semantic segmentation of street scenes," in *Computer Vision–ACCV 2016 Workshops: ACCV 2016 International Workshops*, Taipei, Taiwan, China, 2017, pp. 493–509.

[7] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using cnn-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 19, pp. 303–312, 1992.

[8] C. Kyrkou, and T. Thecharides "Emergencynet: Efficient aerial image classification for drone-based emergency monitoring using atrous convolutional feature fusion," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* , vol. 13, pp. 1687–1699, 2020.

[9] M. Rahnemoonfar et al., "Floodnet: A high resolution aerial imagery dataset for post flood scene understanding," *IEEE Access*, vol. 9, pp. 89644–89654, 2021.

[10] M. Rahnemoonfar et al., "Identifying sunflower lodging based on image fusion and deep semantic segmentation with uav remote sensing imaging," *Comput. Electron. Agric.*, vol. 179, pp. 105812, 2020.

[11] A. Maiti, S. O. Elberink, and G. Vosselman, "Uavpal: A new dataset for semantic segmentation in complex urban landscape with efficient multiscale segmentation," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 17, pp. 464–475, 2023.

[12] S. Girisha, U. Verma, M. M. Pai, and R. M. Pai, "Uvidnet: Enhanced semantic segmentation of uav aerial videos by embedding temporal information," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 4115–4127, 2021.

[13] D. Hong, B. Zhang, X. Li, et al., "SpectralGPT: Spectral foundation model," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024. DOI:10.1109/TPAMI.2024.3362475.

[14] I. Papadeas, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "Real-time semantic image segmentation with deep learning for autonomous driving: A survey," *Appl. Sci.-Basel*, vol. 11, no. 19, pp. 8802, 2021.

[15] N. Siddique, S. Paheding, C. P Elkin, and V. Devabhaktuni., "U-net and its variants for medical image segmentation: A review of theory and applications," *IEEE Access*, vol. 9, pp. 82031–82057, 2021.

[16] T. K. Behera, S. Bakshi, M. Nappi, and P. K. Sa, "Superpixel-based multiscale cnn approach toward multiclass object segmentation from uav-captured aerial images," *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, vol. 16, pp. 1771–1784, 2023.

[17] Y. Lyu, G. Vosselman, G. Xia, and M. Y. Yang, "Bidirectional multi-scale attention networks for semantic segmentation of oblique UAV imagery,"2021, *arXiv:2102.03099*.

[18] Y. Lyu, G. Vosselman, G. S. Xia, A. Yilmaz, and M. Y. Yang, "Uavid: A semantic segmentation dataset for uav imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 108–119, 2020.

[19] D. Hong, B. Zhang, H. Li, et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, pp. 113856, 2023.

[20] M. Guo, C. Lu, Q. Hou, Z. Liu, M. Cheng, and S. Hu, " Segnext: Rethinking convolutional attention design for semantic segmentation,"in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, pp. 1140–1156, 2022.

[21] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, " Bisenet: Bilateral segmentation network for real-time semantic segmentation,"in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 325–341.

[22] 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," *Int. J. Comput. Vision*, vol. 129, pp. 3051–3068, 2021.

[23] M. Fan, S. Lai, J. Huang, X. Wei, Z. Chai, J. Luo, and X. Wei, "Rethinking bisenet for real-time semantic segmentation," in *in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9716–9725. DOI: 10.1109/CVPR46437.2021.00959.

[24] R. P. K. Poudel, S. Liwicki, and R. Cipolla, "Fast-scnn: Fast semantic segmentation network," in *30th Proc. Brit. Mach. Vis. Conf.*, Cardiff, UK, 2019, pp. 289. [Online]. Available: https://doi.org/10.48550/arXiv.1902.04502

[25] T. Takikawa, D. Acuna, V. Jampani, and S. Fidler, "Gated-scnn: Gated shape cnns for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 5229–5238. DOI: 10.1109/ICCV.2019.00533.

[26] J. He, Z. Deng, L. Zhou, Y. Wang, and Y. Qiao, "Adaptive pyramid context network for semantic segmentation,"in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7519–7528. DOI: 10.1109/CVPR.2019.00770.

[27] S. Hao, Y. Zhou, Y. Guo, R. Hong, J. Cheng, and M. Wang, " Real-time semantic segmentation via spatial-detail guided context propagation," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–12, 2022.

[28] H. Pan, Y. Hong, W. Sun, and Y. Jia, "Deep dual-resolution networks for real-time and accurate semantic segmentation of traffic scenes," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 3, pp. 3448–3460, 2022.

[29] J. Zhang, H. Hu, T. Yang, et al., "HR-ASPP: An improved semantic segmentation model of cervical nucleus images with accurate spatial localization and better shape feature extraction based on Deeplabv3+," in *Proc. Int. Conf. Dig. Image Proc.*, 2023, pp.1–8. [Online]. Available: https://doi.org/10.1145/3604078.3604094

[30] Y. Zhang, W. Chen, X. Li, Z. Lai, H. Kong, "Adversarial Keyword Extraction and Semantic-Spatial Feature Aggregation for Clinical Report Guided Thyroid Nodule Segmentation,"in *Pattern Recognit. Comput Vis.*, 2023, pp.235–247. [Online]. Available: https://doi.org/10.1007/978-981-99-8558-6_20

[31] H. Fu, G. Sun, J. Ren, A. Zhang and X. Jia, "Fusion of PCA and Segmented-PCA Domain Multiscale 2-D-SSA for Effective Spectral-Spatial Feature Extraction and Data Classification in Hyperspectral Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1-14, 2022.

[32] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation,"in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, 2015, pp. 234–241. [Online]. Available: https://doi.org/10.1007/978-3-319-24574-4_28

[33] E. Romera, J. M. Alvarez, L. M. Bergasa, and R. Arroyo, "Efficient convnet for real-time semantic segmentation," in *IEEE Int. Veh. Sym.*, 2017, pp. 1789–1794. DOI: 10.1109/IVS.2017.7995966

[34] S. Mehta, M. Rastegari, A. Caspi, L. Shapiro, and H. Hajishirzi, "Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 552–568. [Online]. Available: https://doi.org/10.1007/978-3-030-01249-6_34

[35] S. Kumaar, Y. Lyu, F. Nex, and M. Y. Yang, "Cabinet: Efficient context aggregation network for low-latency semantic segmentation,"

in *IEEE Int. Conf. Robot. Autom.*, 2021, pp. 13517–13524. DOI: 10.1109/ICRA48506.2021.9560977
[36] A. Howard, M. Sandler, G. Chu, et al., "Searching for mobilenetv3," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, PP. 1314–1324.
[37] W. Jiang, D. Zhang, G. Hui,"A dual-branch fracture attribute fusion network based on prior knowledge," *Eng. Appl. Artif. Intell.*, vol. 127, pp. 14, 2024.
[38] Y. Sun, L. Su, S. Yuan and H. Meng, "DANet: Dual-Branch Activation Network for Small Object Instance Segmentation of Ship Images.," *Pattern Recognit. Comput. Vis.*, vol. 33, no. 11, pp. 6708–6720, 2023.
[39] Y. Chen, Y. Wang, P. Lu, et al., "Large-scale structure from motion with semantic constraints of aerial images," *Pattern Recognit. Comput. Vis.*, 2018, pp. 347-359. [Online]. Available: https://doi.org/10.1007/978-3-030-03398-9_30
[40] Y. Lyu, G. Vosselman, G. **a, et al., "UAVid: A semantic segmentation dataset for UAV imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 165, pp. 6708–6720, 2020.
[41] M. Guo, C. Lu, Q. Hou, et al., "Segnext: Rethinking convolutional attention design for semantic segmentation,"in *IEEE Trans. Circuits Syst. Video Technol.*, 2022, pp. 1140-1156. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2022/file/08050f40fff41616ccfc3080e60a301a-Paper-Conference.pdf
[42] W. Zhang, Z. Huang, G. Luo, et al., "TopFormer: Token pyramid transformer for mobile semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12083–12093 DOI: 10.1109/CVPR52688.2022.01177
[43] Q. Wan, Z. Huang, J. Lu, et al., "Seaformer: Squeeze-enhanced axial transformer for mobile semantic segmentation,", 2023, *arXiv:1811.11721*
[44] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 418–434. [Online]. Available: https://doi.org/10.1007/978-3-030-01219-9_25
[45] J. Zhuang, J. Yang, L. Gu, and N. Dvornek, "Shelfnet for fast semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2019, pp. 847–856, DOI: 10.1109/ICCVW.2019.00113.
[46] X. Chen, D. Qi, and J. Shen, "Boundary-aware network for fast and high-accuracy portrait segmentation," 2019, *arXiv:1901.03814*.
[47] Z. Zhu, M. Xu, S. Bai, T. Huang, and X. Bai, "Asymmetric non-local neural networks for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 593–602. DOI: 10.1109/ICCV.2019.00068.
[48] H. Zhao, Y. Zhang, S. Liu, J. Shi, C. C. Loy, D. Lin, and J. Jia, "Psanet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283. [Online]. Available: https://doi.org/10.1007/978-3-030-01240-3_17
[49] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network,"' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 267–283. DOI: 10.1109/CVPR.2017.660
[50] J. Fu, J. Liu, H. Tian, Y. Li, Y. Bao, Z. Fang, and H. Lu, "Dual attention network for scene segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2019, pp. 3146–3154. DOI: 10.1109/CVPR.2019.00326
[51] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, " Ccnet: Crisscross attention for semantic segmentation," 2019, *arXiv:1811.11721*
[52] H. Wang, X. Jiang, H. Ren, Y. Hu, and Song Bai, "SwiftNet: Real-Time Video Object Segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1296–1305. DOI: 10.1109/CVPR46437.2021.00135

Chaoyun Mai is an Associate Professor at Wuyi University, Guangdong, China. He received the Ph.D. degree in information and communication engineering from Beihang University, Beijing, China, in 2017. Currently, he holds the position of associate professor in the School of Electronics and Information Engineering at Wuyi University. His current research interests include image processing and signal processing.

Yibo Wu received the B.S. degree from Shangqiu Normal University. Currently, he is enrolled in the master's program at the School of Electronics and Information Engineering, Wuyi University, China. His research focuses on computer vision, panoramic semantic segmentation.

Yikui Zhai (Senior Member, IEEE) received the bachelor's degree in optical electronics information and communication engineering from Shantou University, Shantou, China, in 2004, the master's degree in signal and information processing from Shantou University, in 2007, and the Ph.D. degree in signal and information processing from Beihang University, Beijing, China, in June 2013. Since October 2007, he has been working with the School of Electronics and Information Engineering, Wuyi University, Jiangmen, China, where he is a Professor. He has been a Visiting Scholar with the Department of Computer Science, University of Milan, Milan, Italy, since 2016. His research interests include image processing, deep learning, and pattern recognition.

Hao Quan received the bachelor's degree in computer science from the Università degli Studi di Milano, in 2014, the master's degree in computer science from the Università degli Studi di Milano, in 2017, and the Ph.D. degree in computer science and engineering from Polytechnic University of Milan, in 2022. Since July 2023, he has been a postdoctoral researcher with the Polytechnic University of Milan, Milan, Italy. His current research interests include artificial intelligence, robotic technology, machine learning.

Jianhong Zhou is currently pursuing his Master in School of Electronics and Information Engineering, Wuyi University. He received his B.S. degree in Wuyi University of Communication Engineering in 2020. His research interests include: computer vision, representational contrastive learning, image processing and object detection.

Angelo Genovese (Senior Member, IEEE) received the Ph.D. degree in computer science from the Università degli Studi di Milano, Milan, Italy, in 2014.,He has been a Post-Doctoral Research Fellow in computer science with the Università degli Studi di Milano since 2014. He has been a Visiting Researcher with the University of Toronto, Toronto, ON, Canada. Original results have been published in over 30 articles in international journals, proceedings of international conferences, books, and book chapters, and patents. His current research interests include signal and image processing, 3-D reconstruction, computational intelligence technologies for biometric systems, industrial and environmental monitoring systems, and design methodologies and algorithms for self-adapting systems.,Dr. Genovese is an Associate Editor of the Journal of Ambient Intelligence and Humanized Computing (Springer).

Vincenzo Piuri (Fellow, IEEE) received the M.S. and Ph.D. degrees in computer engineering from Politecnico di Milan, Milan, Italy, in 1984 and 1988, respectively.

He was the Department Chair with the University of Milan, Milan, from 2007 to 2012, where he has been a Full Professor since 2000. He was an Associate Professor with Politecnico di Milan from 1992 to 2000, a Visiting Professor with The University of Texas at Austin, Austin, TX, USA from 1996 to 1999, and a Visiting Researcher with George Mason University, Fairfax, VA, USA, from 2012 to 2016. He founded a startup company, Sensure srl, Bergamo, Italy, in the area of intelligent systems for industrial applications (leading it from 2007 to 2010) and was active in industrial research projects with several companies.

Dr. Piuri is an ACM Fellow.

Fabio Scotti (Senior Member, IEEE) received the Ph.D. degree in computer engineering from the Politecnico di Milan, Milan, Italy, in 2003.

Dr. Scotti is an Associate Editor with the IEEE Transactions on Human-Machine Systems and Soft Computing (Springer). He has been an Associate Editor with the IEEE Transactions on Information Forensics and Security, and a Guest Coeditor for the IEEE Transactions on Instrumentation and Measurement.