PSDA: Pyramid Spatial Deformable Aggregation for Building Segmentation in Multiview Remote Sensing Images

Xuejun Huang, Yi Wan, *Member, IEEE*, Yongjun Zhang, *Member, IEEE*, Xinyi Liu, Bin Zhang, Yameng Wang, Haoyu Guo, Yingying Pei, and Zhonghua Hu

***Abstract*—As increasingly more deep learning models are designed and implemented, the performance of single-view image semantic segmentation is approaching its upper limit. With the increasing availability of multiview satellite images, using multiview information is gaining attention as it can address occlusion problems in single-view images and achieve cross-validation to reduce inappropriate segmentation. However, current multiview semantic segmentation methods often rely on multiview voting or require complex preprocessing steps, which may not fully leverage the advantages of multiview images. We analyzed the complementarity and constraints of multiview information and introduced the pyramid spatial deformable aggregation (PSDA) module, a plug-and-play module designed to enhance multiview feature fusion. PSDA is the core component of our early multiview segmentation framework, which facilitates early-stage information fusion by directly extracting features from multiview images, avoiding the complex and time-consuming production of true orthoimages. In this article, we first show how we created the multiview segmentation dataset (MVSeg dataset) using orthoimages generated from different-view images. Then, the results are shown to prove that our method outperformed the corresponding single-view segmentation method, namely by increasing the intersection over union (IoU) metric by approximately 1.23% –3.68% on both datasets. Due to the fusion of multiview images at an early stage, the computational complexity is 0.29–0.74 times that of the state-of-the-art method, and the IoU metric improved by approximately 2.20% –7.52% on both datasets.**

***Index Terms*—Deformable convolutional network (DCN), multiview image fusion, orthoimage, semantic segmentation.**

Received 26 October 2024; revised 25 December 2024 and 27 February 2025; accepted 15 March 2025. Date of publication 20 March 2025; date of current version 8 April 2025. This work was supported in part by the China Railway Group Laboratory Basic Research Project under Grant L2023G014, in part by the National Natural Science Foundation of China under Grant 42030102 and Grant 42201474, in part by the Major special projects of Guizhou [2022]001, in part by the Key Laboratory of Spatial-temporal Big Data Analysis and Application of Natural Resources in Megacities, MNR under Grant KFKT-2024-04, in part by the Postdoctoral Research Project for China Railway Siyuan Survey and Design Group Co., LTD. under Grant KY2023127S, and in part by the China Postdoctoral Science Foundation under Grant 2024M753810. *(Corresponding authors: Yi Wan; Yongjun Zhang.)*

Xuejun Huang, Yameng Wang, Yingying Pei, and Zhonghua Hu are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China (e-mail: huangxuejun@whu.edu.cn; ymw@whu.edu.cn; painyy@whu.edu.cn; yqchlsl@whu.edu.cn).

Yi Wan is with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities, MNR, Shanghai 200063, China (e-mail: yi.wan@whu.edu.cn).

Yongjun Zhang and Xinyi Liu are with the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China, and also with the Technology Innovation Center for Collaborative Applications of Natural Resources Data in GBA, Ministry of Natural Resources, Guangzhou 510075, China (e-mail: zhangyj@whu.edu.cn; liuxy0319@whu.edu.cn).

Bin Zhang is with the China Railway Siyuan Survey and Design Group Co., Ltd., Wuhan 430063, China (e-mail: bin.zhang@whu.edu.cn).

Haoyu Guo is with the Institute of Water Engineering Sciences, Wuhan University, Wuhan 430079, China (e-mail: haoyu.guo@whu.edu.cn).

Our MVSeg dataset is available at https://github.com/NanCheng2001/MVSeg-Dataset.

Digital Object Identifier 10.1109/JSTARS.2025.3553030

I. Introduction

BUILDING segmentation plays a pivotal role in remote sensing image analysis and is critical to applications such as urban planning [1] and geographic data updating [2]. However, current methods primarily focus on single-view orthoimages, which can limit the accuracy of segmentation due to the challenges posed by the complex urban landscape [3], [4], [5], [6]. Multiview imaging, a fundamental capability of satellites [7], may provide complementary information that could potentially enhance segmentation accuracy.

Despite their potential, previous studies have primarily focused on applying multiview imagery to 3-D reconstruction [8], [9], while the integration of multiview information for building segmentation remains an underexplored area. Existing multiview semantic segmentation methods often rely on multiview voting at the decision-making stage, which is simple and straightforward, but fails to fully explore the explicit integration and utilization of multiview semantics [10], [11], [12]. To date, only a few deep learning-based methods have fused multiview information [13], [14]. Although these methods outperform multiview voting approaches, they often depend on complex preprocessing steps, such as digital surface model (DSM) generation and true ortho rectification, which significantly hinder their scalability.

Building upon the motivation to effectively fuse multiview image features for improved building segmentation, this study now addresses two critical challenges: 1) How can multiview satellite images be efficiently fused to capture complementary information? 2) What fusion strategy is most effective for accurate building segmentation in this context?

To address these challenges, we propose the pyramid spatial deformable aggregation (PSDA) module, which fully exploits the complementarity and constraints of multiview images. PSDA is a core component of our early multiview segmentation

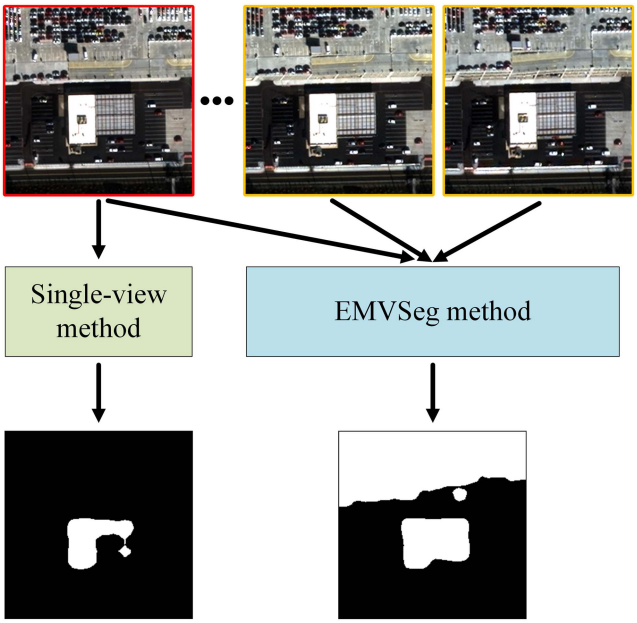

Fig. 1. Example of how multiview fusion can help differentiate between the parking lot and the building, where the reference image is surrounded by a red frame and other images are its neighboring images. It is difficult for humans as well to distinguish the building only by a reference image.

(EMVSeg) framework, which enables direct multiview feature fusion in an end-to-end manner without the complex DSM generation and true orthorectification process. For instance, as shown in Fig. 1, when vehicles are parked on a building's roof—a challenging case for human interpretation using a single-view image—our method accurately identifies the building by effectively fusing multiview image information. This demonstrates the critical role of multiview data in improving building segmentation in complex scenarios.

Compared to previous research in this area, our method makes the following contributions.

1) Our primary and novel PSDA module, which offers strong portability, fully harnesses the complementarity and constraints of multiview images. By incorporating joint offset prediction and enhanced spatial deformable convolution (ESDC), our method with PSDA improves the IoU metric by 1.23% –3.68% compared to corresponding single-view approaches.
2) A new framework, EMVSeg, has been developed for building segmentation that integrates multiview information. To the best of our knowledge, this is the first end-to-end framework for building segmentation that eliminates the need for complex and time-consuming preprocessing, greatly simplifying the multiview segmentation process.
3) Two new datasets: i) SpaceNet4-MVSeg and ii) DFC19-MVSeg, which collectively we named the multiview segmentation dataset (MVSeg dataset). The MVSeg dataset is the first dedicated benchmark dataset specifically curated for the quantitative evaluation of building segmentation using multiview orthoimages.

The rest of this article is structured as follows. Section II contains a brief review of the related literature; Section III provides a detailed description of our proposed framework EMVSeg and the PSDA module; Section IV describes the new MVSeg dataset; Section V presents our experimental results; Section VI provides a discussion of the generalizability of our approach; and Section VII concludes this article.

II. Related Work

In this section, we briefly review related previous works, including the multiview remote sensing image fusion, and deformable convolutional network (DCN).

A. Multiview Remote Sensing Image Fusion

Data fusion strategies are commonly classified into three main approaches: early fusion, middle fusion, and late fusion [15]. Early-fusion strategies combine features at the input stage and process the fused features with a single network. In contrast, middle-fusion strategies integrate features encoded by independent encoders in the decoder part, thereby facilitating feature fusion by using a shared decoder. Late-fusion strategies, also known as decision-level fusion, combine segmentation results from multiple networks at the decision-making stage with the main goal of tackling the problem of inconsistent segmentation outcomes for the same region obtained from various inputs.

In remote sensing, multiview information fusion has been successful in tasks such as classification, crop nutrition estimation, and façade parsing [16], [17], [18]. However, most image segmentation methods often do not fuse multiview information. They only take multiview images as inputs, rely on dense matching methods to generate DSM, and combine them with orthophotos to enhance segmentation accuracy [19], [20], [21]. These methods do not deeply explore the fusion mechanism of multiview information.

Currently, most multiview segmentation methods are based on multiview voting, where semantic segmentation is performed independently for each viewpoint, and the results are merged using time-consuming voting techniques [10], [11], [12]. While this approach is simple and straightforward, it overlooks the complementary feature-level information across perspectives, thereby limiting the full potential of multiview data.

In recent years, advancements in deep learning have led to the emergence of multiview segmentation methods based on multiview fusion. Comandur and Kak [13] proposed a deep learning-based multiview segmentation strategy, achieving pixel-level alignment through DSM generation and the derivation of true orthophotos. Subsequently, Chen et al. [14] leveraged stereo labels to enable efficient feature fusion in dual space. However, they rely on complex preprocessing steps, such as dense matching and true ortho rectification, which significantly increase computational costs.

In summary, the recent studies most relevant to our work can be categorized into two main types: multiview voting methods and multiview fusion methods, as summarized in Table II. Unlike previous methods, our approach eliminates the need for complex pre-processing and employs an early fusion strategy that achieves high computational efficiency.

B. Deformable Convolutional Network

The DCN was introduced in [22], using a learnable offset to modify the shape of the convolutional kernel. This adaptable convolutional structure can adjust more effectively to changes in the target shape and enhance the model's perception ability. DCNs have been widely used in various tasks, such as video compression [23], semantic segmentation [24], and video super-resolution [25].

In the field of semantic segmentation, DHCNet was proposed in [26], a new method for hyperspectral images (HSIs) classification that uses a DCN. DHCNet utilizes deformable convolution's ability to adjust sampling positions dynamically based on the spatial context of the HSI, which significantly enhanced the accuracy of HSI classification. In addition, Liu et al. [27] made a notable contribution in this field by integrating deformable convolutional blocks into the encoder, allowing for more comprehensive contextual feature extraction. The model's adaptability to cloud-induced variations was significantly improved by this approach, surpassing multiple state-of-the-art (SOTA) methods. Furthermore, Deng et al. [28] introduced restricted deformation convolution for semantic segmentation, which effectively models geometric transformations. Their method successfully addressed significant distortions in fisheye images, thereby enhancing semantic segmentation accuracy.

Based on these advancements, we carefully designed our PSDA module to address the issue of incorrect fusion of multiview feature information resulting from conventional convolution fusion methods by using sampling position offsets.

III. Theory and Methodology

In this section, we first discuss the guiding principles of our approach. Then, we present our framework EMVSeg and the PSDA module, and the loss function used to train the framework in detail.

A. Multiview Information Complementarity and Constraints

Our study explored multiview image information aggregation methods that aim to make use of multiview information complementarity and constraints. The evaluation criteria for semantic segmentation are mainly influenced by two factors: the correctness of pixel classification and the accuracy of boundary segmentation [29].

In terms of multiview information complementarity, segmentation based on single-view orthoimages is often disrupted by occlusion and shadows, particularly in semantically ambiguous areas where incorrect segmentation results are common. Because multiview images can provide richer feature information [14], the correct information fusion method can achieve complementary multiview information and ensure the correctness of pixel classification.

In multiview images, whether they are aerial images based on center projection or satellite images based on oblique parallel projection, the roof offset follows the principle of epipolar line translation based on multiview imaging geometry theory. The segmentation boundaries of multiview images are constrained through this principle. Fig. 2 illustrates how powerful geometric constraints ensure that each perspective's image has precise architectural boundaries. However, the classification of ground information is usually performed on orthoimages, which can lead to the loss of strict geometric imaging relationships. To this end, we designed the EMVSeg framework, which pioneered the use of multiview orthoimages for information fusion without time-consuming preprocessing.

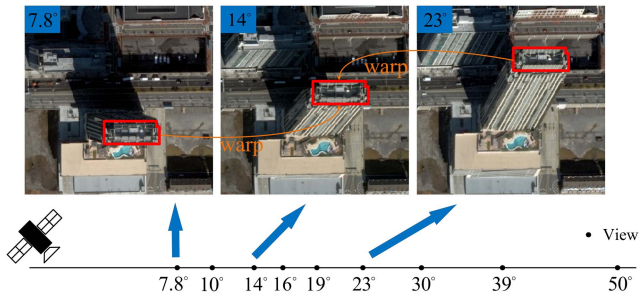

Fig. 2. Schematic diagram of constraints between building contours in multiview images based on strict imaging models.

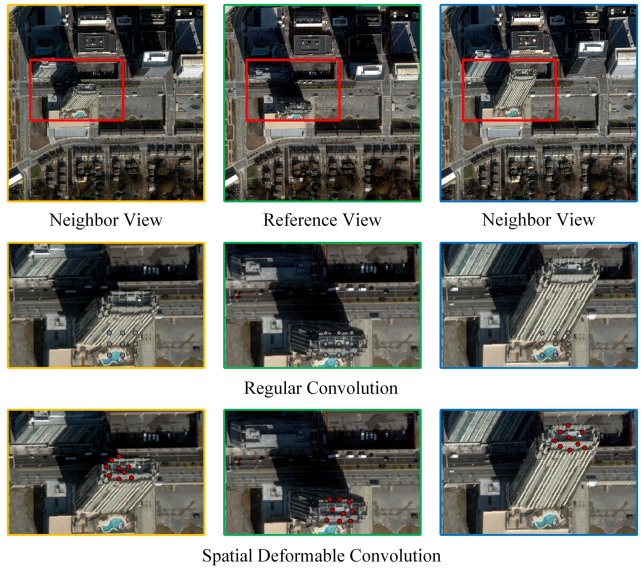

Fig. 3. Fixed receptive field of traditional convolution (second row) and adaptive receptive field of spatial deformable convolution (third row).

Furthermore, we carefully designed the PSDA module to take full advantage of the complementarity and constraints contained in multiview orthoimages and its core is ESDC. We used intelligent learning to obtain the convolution kernel offset, which enables the fusion of multiview image features at the correct location, ensuring the complementarity of multiview image information aggregation (as shown in Fig. 3). In addition, our PSDA is capable of assigning learnable confidence scores to the boundaries of each image, thereby integrating the boundary information of buildings in the reference image and its neighboring images. In other words, our module can suppress erroneous boundary fusion information from neighboring images, similar to the operation of back-end weighted voting fusion based on strict geometric imaging models, which achieved constraints on multiperspective information. Thus, our

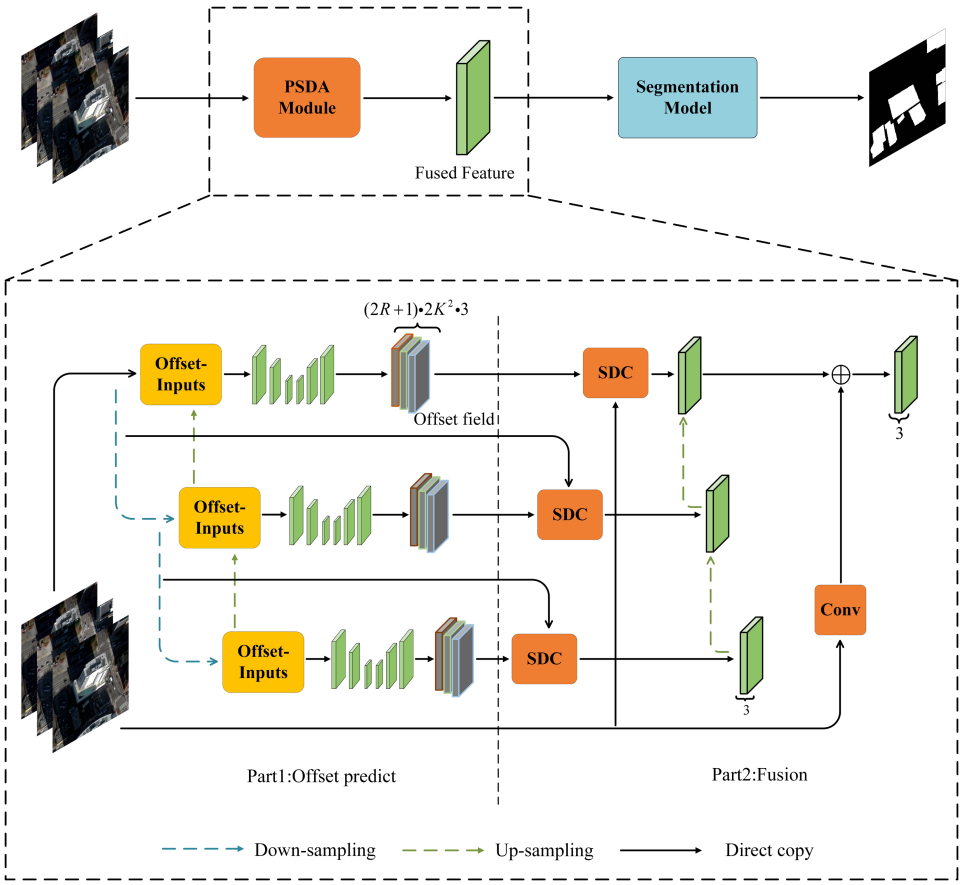

Fig. 4. Architecture of the EMVSeg.

method improves both the correctness of pixel classification and the accuracy of boundary segmentation, leading to better segmentation results.

B. Overall Architecture

As mentioned above, previous methods either required time-consuming voting operations or true orthorectification of the image to achieve pixel-level alignment. The existing multiview segmentation framework is very complex. We designed a novel framework based on multiview orthoimages to simplify the segmentation process, named EMVSeg, which is shown in Fig. 4 (some convolutional structures are omitted in the figure).

First, we use a multiview stereo problem setup in which one image in the multiview orthoimages is used as the reference image and the rest of the images are referred to as neighboring images. Then, the PSDA module is used to aggregate multiview information and select effective features from neighboring images to enhance the feature representation of the reference image. Finally, the enhanced features are fed into various semantic segmentation models, such as FastFCN [30] and DeepLab [31]. EMVSeg works directly from the orthoimages, eliminating the need for true orthorectification or other complex projection operations.

The PSDA module is the core of EMVSeg and is designed to achieve correct feature fusion. The module is comprised of two key stages: joint offset prediction and fusion with ESDC. In the first stage, PSDA uses the reference image and its neighboring images as input to predict the offset field and confidence. This process enables deformable convolution to dynamically select the best feature extraction positions in multiview images. Unlike traditional fixed convolutional kernels, PSDA can flexibly cope with feature shifts caused by differences in viewpoints, thus capturing more representative features. In the second stage, the jointly predicted offset field is fed into the deformable convolutional layer to complete multiview information aggregation. In addition, to promote the flow of information, we added a skip connection outside the structure. The PSDA then produces an enhanced feature map using multiview information aggregation, which is fed into the subsequent network for semantic segmentation.

As previously discussed, the only step required is to insert the PSDA module at the front end of the segmentation model, specifically before the backbone to achieve an early fusion of features. This approach significantly simplifies the feature aggregation process.

C. Joint Offset Prediction

As shown in Fig. 5, the PSDA module process begins by calculating the offset for the ESDC network. To incorporate more contextual information into the network and improve its

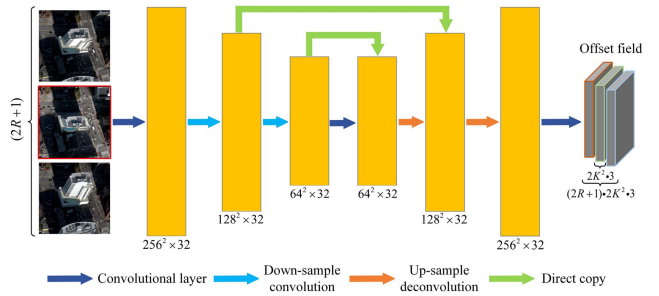

Fig. 5. Structure diagram of offset joint prediction module (U-shaped network), taking the top layer of the pyramid as an example.

precision in predicting offsets, we implemented a joint estimation technique instead of the traditional estimation methods that process a pair of images at a time [25], [32].

Moreover, we used a U-shaped network to predict offsets and confidences, which involves three rounds of convolutional down-sampling and three rounds of up-sampling. Skip connections are used in the middle two layers to improve the information flow. Ultimately, this process yields offset predictions with a channel number of $(2R+1)\cdot 2K^2\cdot 3\ (2R+1)$, which represents the number of images in the multiview image set (e.g., when $R=2$, there are five images in the multiview image set).

The offset prediction module enabled joint prediction for offset and confidence, which integrates richer feature information, resulting in more accurate predictions of offsets and confidences for multiview images.

We define I_v as the reference image, where H and W are the height and width of the reference image respectively. The front and rear R view images are regarded as neighboring images and sent to the PSDA module together with the target slice I_v. The formula for predicting offset $\Delta p$and confidence $\Delta m$is as follows:

$$\Delta p, \Delta m = F_{\text{offset}}\left(\left[I_{v-R}, \ldots, I_v, \ldots, I_{v+R}\right]\right). \tag{1}$$

Among them, $[I_{v-R}, \ldots, I_v, \ldots, I_{v+R}]$ represents the input set of multiview images and F_{offset} is the U-shaped network used to estimate the offset and confidence.

D. Fusion With ESDC

We developed a novel ESDC, which uses deformable convolution to aggregate feature information from multiview images. In contrast to previous methods, we utilized coarse-to-fine fusion to adapt to the low spatial resolution of remote sensing images, resulting in more precise fusion. Our method utilizes a pyramid structure to generate offset-inputs and fused features. The offset-inputs and fused features are then propagated to higher scales to achieve more accurate offset estimation and feature fusion.

Here, the up-sampling method and down-sample convolution are used to transfer information between the low-scale and high-scale layers. In addition, each layer in ESDC uses a fusion module called spatial deformable convolution (SDC), as shown in Fig. 6, which can accommodate changes in the spatial location of buildings in multiview orthoimages.

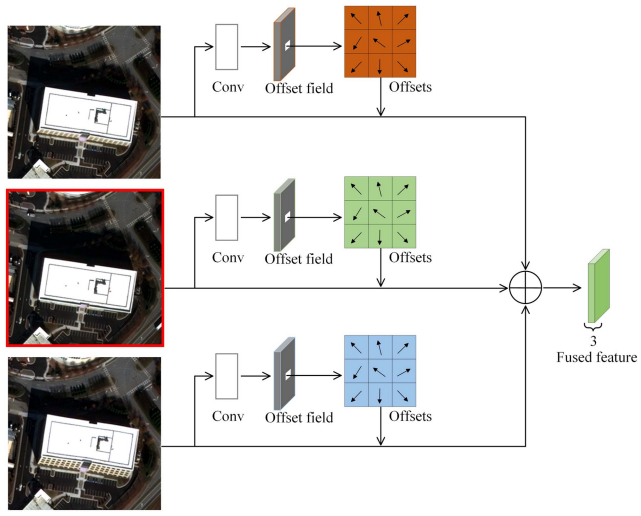

Fig. 6. Structure diagram of SDC with offsets prediction.

Formally, the fusion process using SDC can be described as

$$F(p) = \sum_{c=3\cdot(v-R)}^{3\cdot(v+R)} \sum_{k=1}^{K^2} w_{c,k}\cdot I_v\left(p+p_k+\Delta p_{c,k}\right)\cdot \Delta m_{c,k} \tag{2}$$

where F represents the fused feature map, $w_c \in R^{K^2}$ represents the convolution kernel of the cth channel, p represents any spatial position, and p_k represents the regular convolution offset. For example, when $K=3$, $p_k \in \{(-1,-1),(-1,0),\ldots,(1,1)\}$. $\Delta p_{c,k}$ is our additional learnable offset. Note that the deformable offset $\Delta p_{c,k}$ is simultaneously affected by its spatial location and channel position. Therefore, the spatial deformation and changes in observation angles of multiview images can be represented simultaneously. Since the learnable offset can be of fractional order, we follow [22] and apply differentiable bilinear interpolation to sample subpixels $I_v(p+p_k)$. In addition, since we are using the early-fusion strategy, the number of channels in the fused feature map is 3. For simplicity, we only consider the learnable offsets and ignore the modulation confidences in the description and figures.

In addition, we employ a pyramid processing approach to fuse multiview features progressively from coarse to fine. Specifically, after generating the features W^l of the l layer, we use double up-sampling at $l+1$ layer to obtain the offset and fusion features $F_v^l$of the $l$layer. Note that no upsampling inputs are required to obtain the offset and fusion feature of the third layer. The formula for this process is as follows:

$$W^l = \text{Conv}\left(\left[I_{v-R}, \ldots, I_v, \ldots, I_{v+R}\right]\right) \tag{3}$$

$$\Delta P_{v\_in}^l = \begin{cases} g_1\left(g_2\left(W^l\right), \left(\Delta P_{v\_in}^{l+1}\right)^{\uparrow 2}\right), l<3 \\ g_1\left(g_2\left(W^3\right)\right), l=3 \end{cases} \tag{4}$$

$$F_v^l = \begin{cases} g_3\left(\text{SDC}\left(W^l, g_4\left(\Delta P_{v\_in}^l\right)\right), \left(F_v^{l+1}\right)^{\uparrow 2}\right), l<3 \\ \text{ReLU}\left(\text{SDC}\left(W^3, g_4\left(\Delta P_{v\_in}^3\right)\right)\right), l=3 \end{cases}. \tag{5}$$

TABLE I
COMPARISON OF OUR METHOD WITH PREVIOUS MULTIVIEW SEGMENTATION METHODS

Methods	Fusion Strategies	Computational Efficiency	Preprocessing-Free?
Multi-view voting [10-12]	Late fusion	Low	√
Multi-view fusion [13, 14]	Late fusion or middle fusion	Low	×
Ours	Early fusion	High	√

TABLE II
DETAILS OF SPACENET4-MVSEG AND DFC19-MVSEG DATASETS

	SpaceNet4-MVSeg	DFC19-MVSeg
GSD (m/pixel)	0.5	0.3
Number of viewpoints	5	3
Number of images	16242	10454
Image size (px)	256 × 256	256 × 256
City coverage	Atlanta	Jacksonville & Omaha
Type of data source	WorldView-2	WorldView-3

In the above-mentioned formula, ΔP_{v_in} is the input of the offset joint prediction module, ReLU$(\cdot)$ represents the rectified linear unit activation function, $(\cdot)^{\uparrow 2}$ represents the double upsampling, SDC is the spatial deformable convolution described in (2), and g_1, g_2, g_3, g_4 all represent functions composed of multiple convolutional layers. Furthermore, we use a three-level pyramid in this module, i.e., $l \in (1, 2, 3)$.

Our coarse-to-fine multiview fusion method guarantees the accuracy and semantic richness of the fused feature map, thereby enhancing the segmentation accuracy.

E. Loss Function

To train the proposed model, we utilized a standard cross-entropy loss function to evaluate its predictions. The loss function is defined as follows:

$$L = -\frac{1}{M}\frac{1}{N}\sum_{c}^{M}\sum_{i}^{N} y_{c,i}\log\left(P_{c,i}\right) \tag{6}$$

where M is the total number of categories, N is the total number of pixels in the output label map, c is the category index, and i is the pixel index. $y_{c,i}$ and $P_{c,i}$ are the true label and the probability predicted by the model that the ith pixel belongs to the cth class respectively.

IV. MULTIVIEW SEGMENTATION DATASET

Although multiview remote sensing images covering the same geographical areas are becoming increasingly available, there is a lack of carefully curated multiview orthoimage datasets specifically designed for segmentation tasks. To address this gap, we present two datasets: SpaceNet4-MVSeg and DFC19-MVSeg, which were carefully curated and processed to meet the requirements of semantic segmentation tasks. We first provide the details of these two datasets in Table I, followed by a comprehensive description in Section IV-A and IV-B.

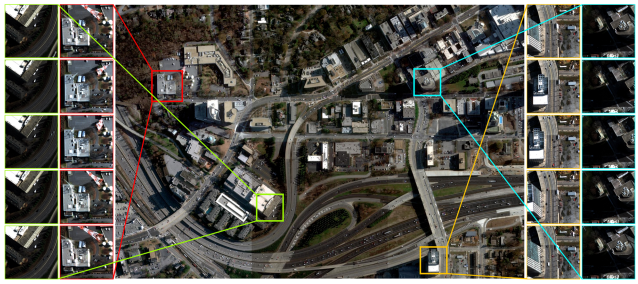

Fig. 7. Several samples in the SpaceNet4-MVSeg dataset, which form a standardized set of five-view orthoimages dataset from top to bottom.

A. SpaceNet4-MVSeg Dataset

The SpaceNet4 dataset is a specialized collection of multiview remote sensing orthoimages [33]. It was created to investigate enhanced algorithmic capabilities in handling off-nadir imagery. The dataset is comprised of 27 WorldView-2 satellite images with 0.5 m ground sampling distance (GSD) captured at off-nadir angles (denoting the angular distance between the satellite's nadir directly below it and the scene's center) ranging from 7.8° to 54°. All of these images were acquired within a 5-min timeframe, covering an expansive area of 665 km^2 in downtown Atlanta. The dataset encompasses approximately 126 747 building footprints [34]. Notably, the original dataset includes roof labels for the image with the smallest off-nadir angle (7.8°), where only visible building portions are labeled in cases of partial occlusion.

To adapt this dataset for multiview segmentation, we selected four additional orthoimages with off-nadir angles (10°, 14°, 19°, 23°) and combined them with the (7.8°) image to create a standardized five-view dataset. The dataset is illustrated in Fig. 7. Each image in the five-view orthoimages dataset was cropped to a size of 256×256 pixels, resulting in a total of 13 048 sets in the training dataset, 1422 sets in the validation dataset, and 1772 sets in the test dataset. We named this dataset the

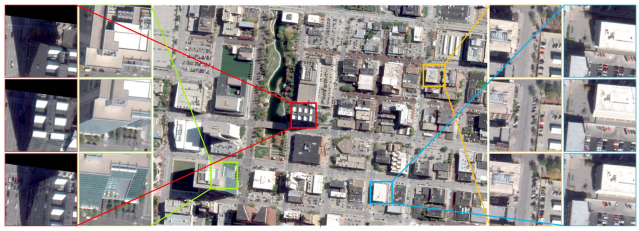

Fig. 8. Several samples in the DFC19-MVSeg dataset, which form a standardized set of three-view orthoimages datasets from top to bottom.

SpaceNet4-MVSeg dataset. It is essential to highlight that within each group of images, only the image with an off-nadir angle of 7.8° possesses an authentic building mask.

B. DFC19-MVSeg Dataset

The DFC19 dataset is comprised of multiview satellite images of two large US cities, Jacksonville, Florida (JAX) and Omaha, Nebraska (OMA), along with ground truth geometric and semantic labels [35]. It includes WorldView-3's visible imagery with approximately 0.3 m GSD. The dataset contains 26 images collected in JAX, from 2014 to 2016 and 43 images collected in OMA from 2014 to 2015. Notably, the DFC19 dataset provides semantic labels for buildings from various viewpoints [36], [37], [38].

We selected original images with different off-nadir angles from the JAX and OMA city images. The corresponding index values are 1, 14, 20 for JAX and 4, 6, 39 for OMA. Then, we used DEM with 10 m GSD for orthorectification and combined the orthoimages from three different views to create a set of standardized three-view orthoimages datasets. We named this composite dataset the DFC19-MVSeg dataset, which is illustrated in Fig. 8. Each image in the DFC19-MVSeg dataset is cropped to a size of 256×256 pixels. The training dataset is comprised of 8364 sets of images, the validation dataset has 1046, and the test dataset has 1044. Unlike SpaceNet4-MVSeg, DFC19-MVSeg includes semantic labels for all perspectives.

V. Experiments

In this section, we first elaborate on the experimental setting and evaluation metrics we utilized. Next, the EMVSeg is compared with SOTA methods on both datasets. The results are further visualized for detailed analysis. We then discuss the ablation studies conducted to verify the rationality of the proposed PSDA module, including an evaluation of the early fusion strategy, as well as portability experiments to demonstrate its generality.

A. Experimental Setting and Evaluation Metrics

Experimental Setting. To evaluate the efficiency of our approach, we implemented a consistent experimental setup for training both single-view and multiview methods. All of the experiments were conducted on a Linux PC equipped with an NVIDIA GeForce RTX 2080 Ti 11G GPU. The implementation of our architecture and the code we reproduced were both developed using the PyTorch deep learning framework.

During training, we maintained uniform configurations across all the dataset experiments. The batch size was set to 2, the maximum number of epochs was limited to 20, and we utilized the stochastic gradient descent optimizer with a learning rate of 0.0001, weight decay of 0.0001, and momentum of 0.9. The training process for the SpaceNet4-MVSeg dataset exclusively utilized the image with the smallest off-nadir angle (7.8°) as the reference image. The images from the other four perspectives were used as neighboring images to enhance the semantic information. In the DFC19-MVSeg dataset, each image served as both a reference and a neighboring image.

Evaluation Metrics. To evaluate the segmentation results, we used the overall accuracy (OA), intersection over union (IoU) and $F1$ Score, which are the commonly used indicators for evaluating semantic segmentation models [39]. OA is defined as the ratio of the number of correctly classified pixels to the total number of pixels

$$\mathrm{OA} = \frac{P_{\mathrm{correct}}}{P_{\mathrm{all}}}. \tag{7}$$

Given that there is only one category in this task, we denote TP as true positives, FP as false positives, TN as true negatives, and FN as false negatives of the building class in the confusion matrix. IoU is defined as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{8}$$

The $F1$ Score is the harmonic mean of Precision and Recall. The $F1$ Score for building category is calculated as follows:

$$F1\ \mathrm{Score} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{9}$$

where Precision and Recall are defined as follows:

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \quad \mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}. \tag{10}$$

B. Comparison With SOTA Methods

1) Performance Assessment of the Proposed Framework: As shown in Table III, we compared our proposed EMVSeg method (the semantic segmentation method with the PSDA module) to the baseline single-view methods (DeepLab and FastFCN), the multiview semantic segmentation benchmark methods (multiview voting [10], [11], [12] and CFA [18]), and the SOTA multiview semantic segmentation networks (MV-train [13] and MVMapper [14]). All the methods shared the same experimental hyperparameters.

For each method in Table III, not all the approaches utilizing multiview images as input outperformed the single-view baseline methods. This discrepancy might be attributed to the reliance of the multiview comparative algorithms on pixel-level aligned multiview images. In contrast, the EMVSeg outperformed the current SOTA multiview segmentation methods in the remote sensing domain across all metrics.

In the SpaceNet4-MVSeg dataset, particularly noteworthy is the improvement in IoU for the building category, which

TABLE III
QUANTITATIVE COMPARISON OF THE PROPOSED EMVSEG WITH OTHER METHODS

BaseNet	Method	SpaceNet4-MVSeg			DFC19-MVSeg		
		IoU	OA	F1 Score	IoU	OA	F1 Score
DeepLab	Single-view	70.87	95.84	82.95	66.66	96.01	79.99
	Multi-view voting	70.17	95.72	82.47	61.67	95.35	76.29
	CFA	72.84	96.25	84.29	65.19	95.75	78.93
	MV-train	70.99	95.94	83.03	64.69	95.69	78.56
	MVMapper	72.35	96.20	83.96	60.79	95.05	75.62
	EMVSeg	**74.55**	**96.60**	**85.42**	**68.31**	**96.25**	**81.17**
FastFCN	Single-view	71.80	96.29	83.59	65.56	95.98	79.20
	Multi-view voting	70.93	96.14	83.00	60.42	95.31	75.33
	CFA	72.47	96.32	84.03	63.71	95.83	77.84
	MV-train	70.36	95.85	82.60	62.93	95.81	77.25
	MVMapper	70.53	96.27	82.72	60.03	95.06	75.03
	EMVSeg	**73.51**	**96.51**	**84.73**	**66.79**	**96.15**	**80.09**

TABLE IV
COMPARISON OF PARAMS AND FLOPS OF THE PROPOSED EMVSEG WITH OTHER METHODS

BaseNet	Method	SpaceNet4-MVSeg		DFC19-MVSeg	
		Params	FLOPs	Params	FLOPs
DeepLab	Single-view	39.6M	82.06G	39.6M	82.06G
	Multi-view voting	39.6M	410.30G	39.6M	246.18G
	CFA	39.6M	82.10G	39.6M	82.09G
	MV-train	39.6M	410.30G	39.6M	246.18G
	MVMapper	39.6M	282.45G	39.6M	182.26G
	EMVSeg	40.6M	120.03G	40.4M	111.46G
FastFCN	Single-view	66.4M	71.60G	66.4M	71.60G
	Multi-view voting	66.4M	358.01G	66.4M	214.81G
	CFA	66.4M	71.64G	66.4M	71.63G
	MV-train	66.4M	358.01G	66.4M	214.81G
	MVMapper	66.4M	202.90G	66.4M	137.25G
	EMVSeg	67.3M	109.57G	67.2M	101.00G

increased by 2.20% –3.56% compared to the SOTA methods, 1.04% –4.38% compared to the benchmark methods, and 1.71% –3.68% compared to the single-view baseline methods. These results demonstrate the superior performance of our method in semantic segmentation tasks leveraging multiview remote sensing images.

In the DFC19-MVSeg dataset, the IoU of the building category also improved, which increased by 3.62% –7.52% compared to the SOTA methods, 3.08% –6.64% compared to the benchmark methods, and 1.23% –1.65% compared to the single-view baseline methods.

In addition, Table IV shows that the number of model parameters (Params) and the floating point operations of the model (FLOPs) were also important indicators for evaluating the quality of our model. The former reflects the memory size occupied by the model and reflects the spatial complexity of the model, and the latter reflects the computational complexity of the model. As shown in Table IV, while our model has a slightly higher number of parameters compared to the SOTA methods, the increase is relatively small (only 0.8–1M). This slight increase is mainly attributed to the additional modules introduced for enhancing multiview feature interaction and semantic segmentation performance. Nevertheless, thanks to our early-fusion strategy, the computational complexity is 0.29–0.74 times that of the SOTA methods, which shows that our model performed well. Since the SOTA methods do not use an early fusion strategy, they require 5 times/3 times feature extraction of the base model, which increases computational complexity.

2) Visual Analysis: We further show the visualization results of the experiments and provide an in-depth analysis of the EMVSeg. Figs. 9 and 10 present a visual comparison of the qualitative results on the SpaceNet4-MVSeg and DFC19-MVSeg datasets, respectively, and demonstrate the effectiveness of the EMVSeg in improving semantic segmentation accuracy for multiview remote sensing images.

As can be seen from Figs. 9 and 10, the multiview semantic segmentation network integrated with the PSDA module effectively improved the segmentation effect. This improvement overcame semantic confusion and ambiguity to a certain extent, particularly for buildings that are highly reflective, buildings with ambiguous semantics, and buildings with similar spectral characteristics to the surrounding ground.

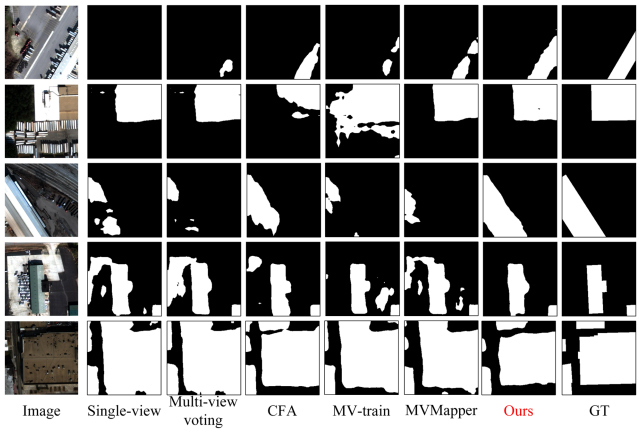

Fig. 9. Qualitative comparison of the EMVSeg with baseline competitors on the SpaceNet4-MVSeg dataset.

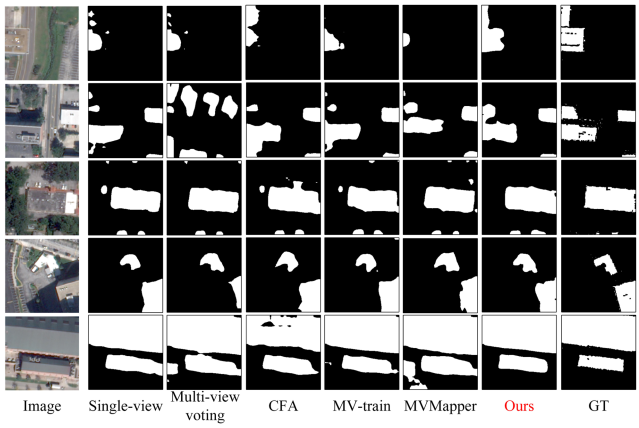

Fig. 10. Qualitative comparison of the EMVSeg with baseline competitors on the DFC19-MVSeg dataset.

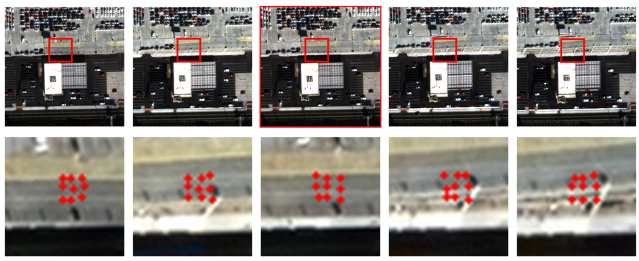

Fig. 11. Visualization of the feature sampling locations, where the reference image is surrounded by a red box.

In addition, to explore how the PSDA module achieves the fusion of multiview image features at the correct location, we visualize the convolutional kernel sampling location of the PSDA module. The experimental results are shown in Fig. 11. In challenging scenarios, such as when the top of the building is a parking lot, the PSDA module intelligently selects the appropriate feature sampling locations, successfully capturing the building façade information from neighboring images and achieving accurate segmentation (the segmentation results are shown in Fig. 1). We also visualize the boundary constraint process in Fig. 12. From the figure, it is evident that the PSDA module successfully implements multiview information constraints by intelligently assigning weights to building contours.

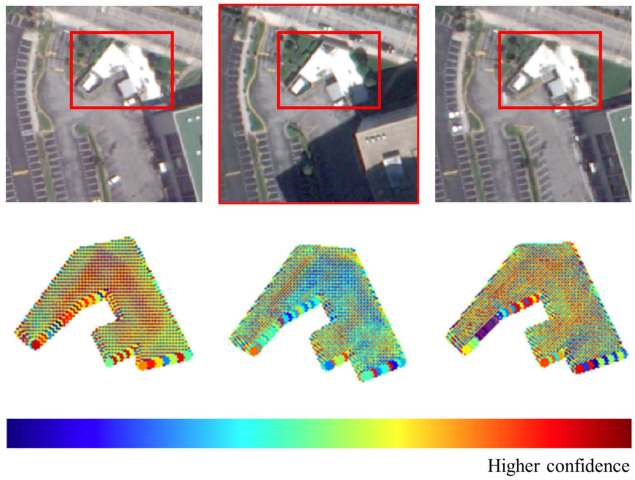

Fig. 12. Visualization of the confidence, where the reference image is surrounded by a red box.

C. Ablation Study

In this section, we present our ablation experiments on the proposed PSDA module and early fusion strategy, and explore the impact of the number of multiview images on the model as well as the transportability of the PSDA module. The above-mentioned experiments were performed on both datasets.

1) Effect of PSDA Module: To assess the performance of our fusion module, we conducted experiments with various configurations, the results of which are summarized in Table V. The PSDA method in Table V integrates our PSDA module into the base single-view semantic segmentation model. The PRA module adopts the pyramid design of the PSDA module but replaces SDC with regular convolution. The SDA module retains the SDC of PSDA while omitting the pyramid structure. The RA module uses a standard convolution only.

From Table V, we offer three main conclusions. First, our PSDA method effectively fused multi-view features (comparing the last line with the first line in each BaseNet). Compared to traditional regular sampling convolution, the flexible convolution kernel enhanced our PSDA's multiview feature extraction capabilities, which we demonstrated using the SpaceNet4-MVSeg dataset results as an example for analysis. In the DeepLab-based model, its performance improved by 0.97% on IoU, 0.26% on OA, and 0.64% on the F1 score compared to method RA. In the FastFCN-based model, its performance compared to method RA improved by 1.31% on IoU, 0.13% on OA, and 0.87% on the F1 score. Second, our ablation experiments demonstrated the effectiveness of the designed pyramid structure (comparing the fourth line with the second line in each BaseNet). Using the PSDA module with a pyramid structure, more accurate offsets were obtained to make the multiview fusion of SDC more accurate. Finally, the SDC we applied was

TABLE V
ABLATION STUDY OF PSDA MODULE (INCLUDING THE USE OF PYRAMID STRUCTURES AND SDC) ON THE SPACENET4-MVSEG AND DFC19-MVSEG DATASETS

BaseNet	Module	Pyramid ?	SDC?	SpaceNet4-MVSeg			DFC19-MVSeg		
				IoU	OA	F1 Score	IoU	OA	F1 Score
DeepLab	RA	--	--	73.58	96.34	84.78	65.36	95.83	79.05
	SDA	--	√	73.37	96.29	84.64	67.54	96.09	80.63
	PRA	√	--	73.36	96.32	84.63	67.03	96.05	80.26
	PSDA	√	√	**74.55**	**96.60**	**85.42**	**68.31**	**96.25**	**81.17**
FastFCN	RA	--	--	72.20	96.38	83.86	64.33	95.83	78.29
	SDA	--	√	73.45	**96.55**	84.69	66.42	96.10	79.82
	PRA	√	--	73.44	96.52	84.69	66.56	96.12	79.92
	PSDA	√	√	**73.51**	96.51	**84.73**	**66.79**	**96.15**	**80.09**

TABLE VI
ABLATION STUDY OF THE EFFECT OF EARLY FUSION STRATEGY ON THE SPACENET4-MVSEG AND DFC19-MVSEG DATASETS

BaseNet	Method	Early?	Late?	SpaceNet4-MVSeg			DFC19-MVSeg		
				IoU	OA	F1 Score	IoU	OA	F1 Score
DeepLab	SV	--	--	70.87	95.84	82.95	66.66	96.01	79.99
	L-PSDA	--	√	71.58	96.14	83.43	64.98	95.61	78.77
	L-MVtrain	--	√	70.99	95.94	83.03	64.69	95.69	78.56
	E-L-PSDA	√	√	57.98	92.30	73.40	67.01	95.88	80.25
	E-PSDA-L-MVtrain	√	√	44.99	87.04	62.06	64.44	95.75	78.38
	E-PSDA	√	--	**74.55**	**96.60**	**85.42**	**68.31**	**96.25**	**81.17**
FastFCN	SV	--	--	71.80	96.29	83.59	65.56	95.98	79.20
	L-PSDA	--	√	70.74	96.04	82.86	62.50	95.62	76.92
	L-MVtrain	--	√	70.36	95.85	82.60	62.93	95.81	77.25
	E-L-PSDA	√	√	50.30	89.73	66.94	54.75	93.49	70.76
	E-PSDA-L-MVtrain	√	√	53.88	91.50	70.03	55.14	94.15	71.09
	E-PSDA	√	--	**73.51**	**96.51**	**84.73**	**66.79**	**96.15**	**80.09**

shown to be reasonable. Comparing the experimental results of the last and third lines of the DeepLab-based model on the SpaceNet4-MVSeg dataset, its performance improved by 1.19% on IoU, 0.28% on OA, and 0.79% on the F1 score compared to method PRA.

2) Effect of Early Fusion Strategy: In this section, we conduct a series of experiments to explore the impact of the early fusion strategy on the multiview segmentation task, with the experimental results presented in Table VI. Specifically, we perform an exhaustive study of five variants of the multiview segmentation network, including the following.

1) L-PSDA (late fusion utilizing the PSDA module).
2) L-MVtrain (late fusion utilizing the MV-train).
3) E-L-PSDA (both early and late fusion utilizing the PSDA module).
4) E-PSDA-L-MVtrain (early fusion with the PSDA module and late fusion with the MV-train).
5) E-PSDA (early fusion utilizing the PSDA module).

To ensure the fairness of the experiment, the position of the PSDA module used for late fusion was kept consistent with the MV-train. In addition, we provide the results of single-view segmentation (SV) for comparison. This comprehensive setup enables a clear assessment of each variant's contribution to overall segmentation accuracy and provides valuable insights into the advantages of early fusion in the multiview context.

From the experimental results, we can draw four main conclusions.

1) The segmentation effect of early fusion using the PSDA module is better than that of late fusion using the PSDA module (compare the last and second rows of Table VI in each BaseNet). This may be because the offset learning of the PSDA module requires strong feature information as input, and it is difficult to extract strong feature information from the classification scores, thus leading to the failure of the PSDA module.
2) The segmentation effect of applying the PSDA module to both early fusion and late fusion is not as good as that of

TABLE VII
ABLATION STUDY OF THE NUMBER OF VIEWS ON THE SPACENET4-MVSEG AND DFC19-MVSEG DATASETS

Method	View number	SpaceNet4-MVSeg			DFC19-MVSeg		
		IoU	OA	F1 Score	IoU	OA	F1 Score
PSDA-DeepLab	1	70.87	95.84	82.95	66.66	96.01	79.99
	2	73.19	96.28	84.52	66.47	95.95	79.86
	3	73.58	96.36	84.78	**68.31**	**96.25**	**81.17**
	4	**74.70**	96.57	**85.52**	---	---	---
	5	74.55	**96.60**	85.42	---	---	---
PSDA-FastFCN	1	71.80	96.29	83.59	65.56	95.98	79.20
	2	72.77	96.39	84.24	63.92	95.76	77.99
	3	**73.86**	**96.59**	**84.96**	**66.79**	**96.15**	**80.09**
	4	72.13	96.39	83.81	---	---	---
	5	73.51	96.51	84.73	---	---	---

using it only for early fusion (compare the fourth and last rows of Table VI in each BaseNet). From the conclusion of 1), it is clear that PSDA applied to late fusion has a negative effect, so this result is not unexpected.

3) The segmentation effect of late fusion using MV-train is even less effective than the SV effect (compare the third and first rows of Table VI in each BaseNet). This is because the MV-train method is designed for multiview true orthoimages and utilizes regular convolution for late fusion. It is unable to accommodate feature offsets in multiview orthophotos, causing the algorithm to fail.
4) The joint PSDA module for early fusion and MV-train for late fusion strategy is not as effective as PSDA alone for early fusion (compare the penultimate and last rows of Table VI in each BaseNet).

Since 3) shows that the MV-train method does not apply to multiview orthoimages, late fusion with MV-train brings negative effects. The experiments also show that integrating late and early fusion within a single network for end-to-end training can hinder task performance.

3) Effect of the Number of Views: Our method aims to make full use of the multiview information's complementarity and constraints to obtain accurate semantic segmentation results. We used different quantities of images as the input to explore the impact of the number of multiview images on the model. Our experimental results are shown in Table VII.

Table VII confirms that the accuracy of our semantic segmentation generally increased as we increased the number of views for multiview images in most cases. For example, in the PSDA-DeepLab model on the DFC19-MVSeg dataset, the model with three multiview image inputs outperformed the single-view input model by 1.65% on IoU, 0.24% on OA, and 1.18% on the F1 Score. Similarly, the model with three multiview image inputs outperformed the model with two multiview image inputs by 1.84% on IoU, 0.30% on OA, and 1.31% on the F1 Score. However, the PSDA-FastFCN model in the SpaceNet4-MVSeg dataset produced better segmentation results with three multiview image inputs than with five. This incongruity could be attributed to the fact that when the shooting angle difference was too large, the deformable convolution failed to accurately identify the offset eigenvalue position, resulting in incorrect feature fusion.

4) Transportability of PSDA Module: Our PSDA module integrated smoothly with existing single-view semantic segmentation methods, creating a multiview enhanced end-to-end framework. We conducted experiments by integrating our PSDA module with popular frameworks such as DeepLab and FastFCN on both datasets. Table VIII shows that the PSDA module improved the semantic segmentation of buildings, demonstrating its portability and effectiveness.

VI. DISCUSSION

We conduct an experimental analysis of the generalization of the PSDA module using the SpaceNet4-MVSeg and DFC19-MVSeg datasets to investigate the potential limitations and applicability of PSDA across varying environmental conditions. We selected three challenging scenarios for segmentation:

1) poorly lit imaging conditions,
2) buildings with textures similar to their surroundings, and
3) imaging environments obscured by clouds.

The segmentation results for these scenarios are visualized in Fig. 13.

As shown in Fig. 13, the first and second rows correspond to the case of insufficient lighting. The reference image is heavily affected by black pixels, which poses a significant challenge for the PSDA module. The visualization results indicate that the PSDA module experiences segmentation errors and incomplete segmentation in this difficult environment. Nonetheless, it still achieves better performance compared to SV, thanks to its ability to aggregate clearer features from other viewpoint images. The third row of Fig. 13 illustrates the challenge posed by buildings having similar texture features to their surroundings.

Here, the PSDA-based multiview segmentation network exhibits some segmentation errors, particularly in differentiating the building roof from the ground. However, the presence of vehicles in the other two viewpoint images helps improve the results compared to SV by aggregating multiview information. Finally, the final row of Fig. 13 depicts a scenario where the reference image is occluded by clouds. Although there are some

TABLE VIII
QUANTITATIVE COMPARISON OF BASELINE METHODS AND THEIR MULTIVIEW ENHANCED VERSIONS ON THE SPACENET4-MVSEG AND DFC19-MVSEG DATASETS

Dataset	Method	IoU	OA	F1 Score
SpaceNet4-MVSeg	DeepLab	70.87	95.84	82.95
	MV-DeepLab	**74.55**	**96.60**	**85.42**
	FastFCN	71.80	96.29	83.59
	MV-FastFCN	**73.51**	**96.51**	**84.73**
DFC19-MVSeg	DeepLab	66.66	96.01	79.99
	MV-DeepLab	**68.31**	**96.25**	**81.17**
	FastFCN	65.56	95.98	79.20
	MV-FastFCN	**66.79**	**96.15**	**80.09**

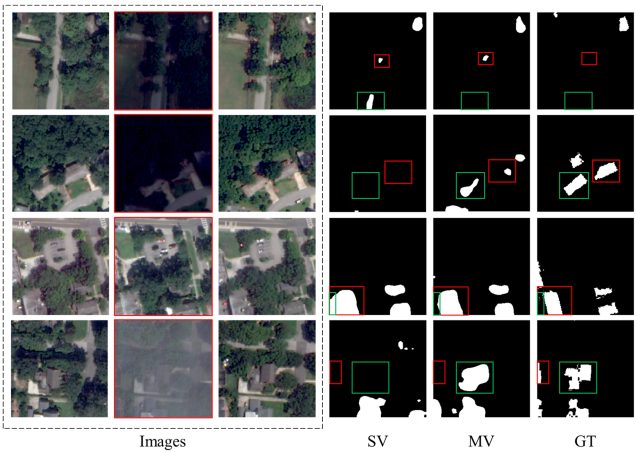

Fig. 13. Performance of PSDA under different environmental conditions, where areas with poor segmentation are marked with red boxes. The areas with improved results over SV are marked with green boxes.

misdetections and omissions, the PSDA module still outperforms SV in this case.

All in all, while some challenging environments present difficulties in building segmentation, the PSDA-based multiview segmentation network outperforms the SV network.

VII. CONCLUSION

The study we presented in this article fully analyzed the complementarity and constraints of multiview information for enhancing the feature representation of a reference image. Based on our analysis, we introduce the novel EMVSeg framework, which simplifies multiview segmentation by avoiding the complex and time-consuming production of true orthoimages. Its central component is the PSDA module, which we demonstrated is capable of using aggregate feature information from multiview images. Due to its use of an early-fusion strategy, we showed that our approach achieved superior performance compared to current SOTA methods, particularly in semantically ambiguous building areas. It is important to note that the PSDA module is highly portable and can be easily integrated with existing SV networks. In addition, we constructed the first MVSeg dataset.

The primary limitation of our PSDA module is that multiview images require semantic annotation for network learning, and most current multiview image datasets lack semantic labels. In future work, we plan to address this limitation by introducing unsupervised or self-supervised strategies. In addition, we plan to integrate techniques such as 3-D Gaussian Splatting to achieve joint estimation of semantics and building heights.

REFERENCES

[1] R. Alshehhi, P. R. Marpu, W. L. Woon, and M. D. Mura, "Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks," *ISPRS J. Photogrammetry Remote Sens.*, vol. 130, pp. 139–149, 2017.

[2] F. Fang et al., "Incorporating superpixel context for extracting building from high-resolution remote sensing imagery," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 17, pp. 1176–1190, 2024.

[3] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2018.

[4] B. Bischke, P. Helber, J. Folz, D. Borth, and A. Dengel, "Multi-task learning for segmentation of building footprints with deep neural networks," in *Proc. IEEE Int. Conf. Image Process.*, 2019, pp. 1480–1484.

[5] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 183, pp. 240–252, 2022.

[6] H. Zhu, W. Ma, L. Li, L. Jiao, S. Yang, and B. Hou, "A dual–branch attention fusion deep network for multiresolution remote–sensing image classification," *Inf. Fusion*, vol. 58, pp. 116–131, 2020.

[7] O. Manas, A. Lacoste, X. Giró-i-Nieto, D. Vazquez, and P. Rodriguez, "Seasonal contrast: Unsupervised pre-training from uncurated remote sensing data," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 9414–9423.

[8] Y. Fu, M. Zheng, P. Chen, and X. Liu, "Mono-MVS: Textureless-aware multi-view stereo assisted by monocular prediction," *Photogrammetric Rec.*, vol. 39, no. 185, pp. 183–204, 2024.

[9] H. Hu et al., "Adaptive region aggregation for multi-view stereo matching using deformable convolutional networks," *Photogrammetric Rec.*, vol. 38, no. 183, pp. 430–449, 2023.

[10] R. Qin, X. Huang, W. Liu, and C. Xiao, "Semantic 3D reconstruction using multi-view high-resolution satellite images based on U-Net and image-guided depth fusion," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5057–5060.

[11] M. Shvets, D. Zhao, M. Niethammer, R. Sengupta, and A. C. Berg, "Joint depth prediction and semantic segmentation with multi-view sam," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2024, pp. 1328–1338.

[12] Q. Hu et al., "Utilizing unsupervised learning, multi-view imaging, and CNN-based attention facilitates cost-effective wetland mapping," *Remote Sens. Environ.*, vol. 267, 2021, Art. no. 112757.

[13] B. Comandur and A. C. Kak, "Semantic labeling of large-area geographic regions using multiview and multidate satellite images and noisy OSM training labels," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4573–4594, 2021.

[14] Q. Chen, W. Gan, P. Tao, P. Zhang, R. Huang, and L. Wang, "End-to-end multiview fusion for building mapping from aerial images," *Inf. Fusion*, vol. 111, 2024, Art. no. 102498.

[15] Y. Zhang et al., "Modality-aware mutual learning for multi-modal medical image segmentation," in *Proc. 24th Int. Conf. Med. Image Comput. Comput. Assist. Intervention*, 2021, pp. 589–599.

[16] X. Huang et al., "A multispectral and multiangle 3-D convolutional neural network for the classification of ZY-3 satellite images over urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 12, pp. 10266–10285, Dec. 2021.
[17] N. Lu et al., "An assessment of multi-view spectral information from UAV-based color-infrared images for improved estimation of nitrogen nutrition status in winter wheat," *Precis. Agriculture*, vol. 23, no. 5, pp. 1653–1674, 2022.
[18] W. Ma, S. Xu, W. Ma, and H. Zha, "Multiview feature aggregation for facade parsing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8003505.
[19] D. Yu, S. Ji, J. Liu, and S. Wei, "Automatic 3D building reconstruction from multi-view aerial images with deep learning," *ISPRS J. Photogrammetry Remote Sens.*, vol. 171, pp. 155–170, 2021.
[20] P. d'Angelo et al., "3D semantic segmentation from multi-view optical satellite images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 5053–5056.
[21] M. J. Leotta et al., "Urban semantic 3D reconstruction from multiview satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1451–1460.
[22] J. Dai et al., "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 764–773.
[23] J. Deng, L. Wang, S. Pu, and C. Zhuo, "Spatio-temporal deformable convolution for compressed video quality enhancement," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 07, pp. 10696–10703.
[24] S. Dong et al., "DeU-Net 2.0: Enhanced deformable U-net for 3D cardiac cine MRI segmentation," *Med. Image Anal.*, vol. 78, 2022, Art. no. 102389.
[25] Y. Tian, Y. Zhang, Y. Fu, and C. Xu, "Tdan: Temporally-deformable alignment network for video super-resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3360–3369.
[26] J. Zhu, L. Fang, and P. Ghamisi, "Deformable convolutional neural networks for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 8, pp. 1254–1258, Aug. 2018.
[27] Y. Liu, W. Wang, Q. Li, M. Min, and Z. Yao, "DCNet: A deformable convolutional cloud detection network for remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 8013305.
[28] L. Deng, M. Yang, H. Li, T. Li, B. Hu, and C. Wang, "Restricted deformable convolution-based road scene semantic segmentation using surround view cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4350–4362, Oct. 2020.
[29] D. Zheng, X. Zheng, L. T. Yang, Y. Gao, C. Zhu, and Y. Ruan, "Mffn: Multi-view feature fusion network for camouflaged object detection," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 6232–6242.
[30] H. Wu, J. Zhang, K. Huang, K. Liang, and Y. Yu, "Fastfcn: Rethinking dilated convolution backbone for semantic segmentation," 2019. [Online]. Available: http://arxiv.org/abs/1903.11816
[31] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.
[32] X. Wang, K. C. Chan, K. Yu, C. Dong, and C. C. Loy, "Edvr: Video restoration with enhanced deformable convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2019, pp. 1954–1963.
[33] H. Hao et al., "Improving building segmentation for Off-Nadir satellite imagery," 2020. [Online]. Available: http://arxiv.org/abs/2109.03961
[34] N. Weir et al., "Spacenet MVOI: A multi-view overhead imagery dataset," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 992–1001.
[35] M. Bosch, K. Foster, G. Christie, S. Wang, G. D. Hager, and M. Brown, "Semantic stereo for incidental satellite images," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2019, pp. 1524–1532.
[36] G. Christie, K. Foster, S. Hagstrom, G. D. Hager, and M. Z. Brown, "Single view geocentric pose in the wild," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1162–1171.
[37] G. Christie, R. R. R. M. Abujder, K. Foster, S. Hagstrom, G. D. Hager, and M. Z. Brown, "Learning geocentric object pose in oblique monocular images," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14512–14520.
[38] B. L. Saux, N. Yokoya, R. Hänsch, and M. Brown, "2019 IEEE GRSS data fusion contest: Large-scale semantic 3D reconstruction," *IEEE Geosci. Remote Sens. Mag.*, vol. 7, no. 4, pp. 33–36, Dec. 2019.
[39] P. Anilkumar and P. Venugopal, "Research contribution and comprehensive review towards the semantic segmentation of aerial images using deep learning techniques," *Secur. Commun. Netw.*, vol. 2022, no. 1, 2022, Art. no. 6010912.

Xuejun Huang received the B.S. degree in automation from Harbin Engineering University, Harbin, China, in 2022. He is currently working toward the Ph.D. degree in remote sensing science and technology with Wuhan University, Wuhan, China.

His research interests include deep learning, computer vision, semantic segmentation, and 3-D reconstruction.

Yi Wan (Member, IEEE) was born in 1991. He received the B.S. degree in remote sensing science and technology from Wuhan University, in 2013 and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, in 2018.

He is currently an Associate Research Fellow with Wuhan University. His research interests include digital photogrammetry, computer vision, 3-D reconstruction, and change detection in remote sensing imagery.

Yongjun Zhang (Member, IEEE) received the B.S. degree in geodesy, the M.S. degree in geodesy and surveying engineering, and the Ph.D. degree in geodesy and photogrammetry from Wuhan University, Wuhan, China, in 1997, 2000, and 2002, respectively.

He is currently a Professor and the Dean with the School of Remote Sensing and Information Engineering, Wuhan University. He has authored or coauthored more than 180 research articles and three books. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource data sets, object information extraction and modeling with artificial intelligence, integration of LiDAR point clouds and images, and 3-D city model reconstruction.

Dr. Zhang is the Coeditor-in-Chief of *The Photogrammetric Record*.

Xinyi Liu received the B.S. degree in remote sensing science and technology from Wuhan University, in 2014 and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, in 2020.

She is currently a Postdoctoral Researcher with Wuhan University. Her research interests include 3-D reconstruction, LiDAR and Image Integration, and texture mapping.

Bin Zhang received the B.S. degree in remote sensing science and technology from Liaoning Technical University, Fuxin, China, in 2017, and the M.S. and Ph.D. degrees in photogrammetry and remote sensing with Wuhan University, Wuhan, China, in 2019 and 2023, respectively.

His research interests include high spatial resolution remote sensing image processing, computer vision, and pattern recognition.

Yameng Wang received the B.S. degree in remote sensing science and technology, in 2018, from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

Her research interests include multimodal remote sensing data processing and deep learning.

Haoyu Guo received the B.S. degree in surveying and mapping engineering from Information Engineering University, Zhengzhou, China, in 2015, and the M.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2021.

His research interests include satellite imagery photogrammetry and remote sensing.

Yingying Pei received the B.S. degree in remote sensing science and technology, in 2022, from Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

Her research interests include deep learning, monocular height estimation, and 3-D reconstruction.

Zhonghua Hu received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2021, where he is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

His research interests include deep learning, computer vision, 3-D scences generation, and 3-D reconstruction.