MVSR3D: An End-to-End Framework for Semantic 3-D Reconstruction Using Multiview Satellite Imagery

Xuejun Huang, Xinyi Liu, Yi Wan, *Member, IEEE*, Zhi Zheng, Bin Zhang, Yameng Wang, Haoyu Guo, and Yongjun Zhang, *Member, IEEE*

***Abstract*— Semantic 3-D reconstruction from multiview images is essential for applications such as 3-D city modeling and robot navigation. However, existing methods treat semantic segmentation (SS) and height estimation (HE) as separate tasks, leading to suboptimal reconstruction results. To bridge this gap, we introduce MVSR3D, the first end-to-end framework for semantic 3-D reconstruction using multiview satellite images. MVSR3D employs a dual-stream architecture, consisting of the segmentation branch (MVSAM) based on segment anything model (SAM) and the HE branch based on multiview stereo (MVS). To enhance multiview feature fusion, we propose the epipolar cross attention (ECA) module in the MVSAM branch, which integrates image embeddings primarily along epipolar line to exploit complementary multiview information. Unlike conventional multitask learning approaches, we design dedicated interaction modules—the SAM feature-guided (SAM-FG) module and the elevation-guided sparse prompts generator (EGSPG)—to facilitate multitask interaction and feature fusion. Extensive evaluations on the DFC19 and SpaceNet4 datasets demonstrate that MVSR3D significantly outperforms the state-of-the-art multiview multitask learning (MV-MTL) method, improving the mIoU3 metric at a 2.5-m threshold by 37.09%–45.11%.**

Received 2 March 2025; accepted 18 April 2025. Date of publication 23 April 2025; date of current version 1 May 2025. This work was supported in part by China Railway Group Laboratory Basic Research Project under Grant L2023G014; in part by the National Natural Science Foundation of China under Grant 42030102, Grant 42471470, and Grant 42201474; in part by the Major special projects of Guizhou under Grant [2022]001; in part by the Key Laboratory of Spatial-Temporal Big Data Analysis and Application of Natural Resources in Megacities, Ministry of Natural Resources (MNR), under Grant KFKT-2024-04; in part by the Postdoctoral Research Project for China Railway Siyuan Survey and Design Group Company Ltd. under Grant KY2023127S; and in part by China Postdoctoral Science Foundation under Grant 2024M753810. *(Corresponding authors: Xinyi Liu; Yongjun Zhang.)*

Xuejun Huang and Yameng Wang 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).

Xinyi Liu, Yi Wan, and Yongjun Zhang 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 the Greater Bay Area (GBA), Ministry of Natural Resources, Yuexiu District, Guangzhou, Guangdong 510075, China (e-mail: liuxy0319@whu.edu.cn; yi.wan@whu.edu.cn; zhangyj@whu.edu.cn).

Zhi Zheng is with the Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China (e-mail: zhizheng@cuhk.edu.hk).

Bin Zhang is with China Railway Siyuan Survey and Design Group Company 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).

Digital Object Identifier 10.1109/TGRS.2025.3563498

***Index Terms*— 3-D reconstruction, multitask learning, multiview fusion, semantic 3-D reconstruction, semantic segmentation (SS).**

I. Introduction

SEMANTIC 3-D reconstruction is an essential task in remote sensing image processing, aiming to generate 3-D models enriched with semantic labels. This technique is widely applied in various fields, including 3-D city modeling [1], robot navigation [2], and autonomous driving [3]. However, most existing methods primarily rely on monocular remote sensing images or epipolar rectified image pairs, often resulting in suboptimal reconstruction results. Given that multiview observation is a fundamental capability of satellites [4], fully leveraging multiview image information is crucial for improving reconstruction accuracy.

Most existing multiview semantic 3-D reconstruction methods treat height estimation (HE)—defined as the absolute elevation of objects relative to a reference surface—and semantic segmentation (SS) as two independent tasks [1], [5], [6]. This separation overlooks the potential benefits of multitask learning, where SS can enhance HE accuracy, and vice versa [7], [8]. In addition, none of these methods exploit the potential benefits of multiview information for SS.

Although some methods employ multitask learning to jointly handle SS and HE using monocular images [9], [10], [11] or epipolar rectified image pairs [12], [13], [14], [15], they typically rely on a shared encoder followed by separate decoders for each task. While this strategy improves performance to some extent, they still have significant limitations: 1) they do not explicitly establish interactions between different task branches within the multitask network and 2) they fail to fully exploit multiview information to enhance SS.

To this end, we propose MVSR3D, a novel end-to-end framework for semantic 3-D reconstruction using multiview satellite images. As illustrated in Fig. 1, MVSR3D consists of two branches: a HE branch based on multiview stereo (MVS) and a SS branch, MVSAM. A key aspect of our approach is the bidirectional interaction between these tasks, enabling more effective information exchange. Additionally, we leverage the geometric constraints of multiview images to enhance semantic information aggregation.

1558-0644 © 2025 IEEE. All rights reserved, including rights for text and data mining, and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.
Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

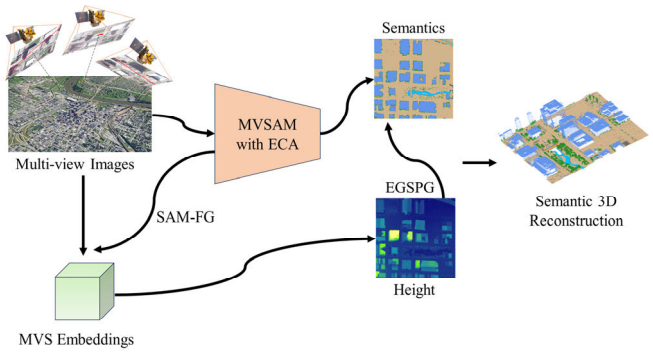

Fig. 1. Main idea of this article. MVSAM fuses multiview image embeddings primarily along the epipolar line and guides the HE based on MVS, and the HE result in turn facilitates SS.

Compared to previous studies in this field, our method offers the following contributions.

1) We propose MVSR3D, the first end-to-end framework for semantic 3-D reconstruction using multiview satellite images, leveraging multiview fusion and multitask learning.
2) We apply epipolar constraints for multiview semantic information fusion and propose a novel epipolar cross attention (ECA) module to aggregate segment anything model (SAM)-encoded features from multiview satellite images.
3) We enhance multitask learning through bidirectional task interaction: the elevation-guided sparse prompts generator (EGSPG) utilizes HEs as sparse prompts for the MVSAM branch, while the SAM feature-guided (SAM-FG) module integrates rich SAM features into the HE branch.
4) Extensive quantitative and qualitative experiments on the DFC19 and SpaceNet4 datasets demonstrate consistent improvements over standalone HE and segmentation models, as well as SOTA semantic 3-D reconstruction methods.

The remainder of this article is structured as follows: Section II reviews related literature; Section III details the proposed framework and its three key modules; Section IV presents the dataset and our experimental results; and Section V concludes the study.

II. Related Work

In this section, we briefly review related previous works, including segmentation foundation models, 3-D reconstruction, and semantic 3-D reconstruction.

A. Segmentation Foundation Model

Meta proposed SAM [16], which sparked the rapid growth of segmentation foundation models. SAM is a versatile model capable of both interactive and fully automatic segmentation. Pretrained with a large amount of data, it possesses extreme generality. In recent years, numerous foundation models have been introduced in the segmentation domain, often exceeding 100 billion parameters and demonstrating exceptional zero-shot generalization performance [17], [18], [19].

Due to significant differences between remote sensing images and natural images in resolution, spectral characteristics, and imaging mechanisms, directly applying generic segmentation foundation models to remote sensing images yields suboptimal results. To bridge this gap, researchers have proposed various improvements, primarily focusing on fine-tuning or adapting SAM [20], [21]. For example, SAMRS fine-tuned SAM on classic remote sensing datasets to address the issue of SAM's segmentation results lacking semantic categories [22]. Additionally, other studies have introduced task-specific decoders to enable SAM-based models to produce semantic masks for domain-specific applications [23], [24].

Beyond SAM-based fine-tuning approaches, SkySense [25] has introduced a new paradigm as a segmentation foundation model specifically tailored for remote sensing applications. Unlike SAM, SkySense is pretrained on a multimodal remote sensing dataset comprising 21.5 million time-series images and incorporates billions of parameters, making it more suitable for remote sensing tasks.

However, existing segmentation foundation models are primarily tailored for single-view remote sensing images, neglecting the valuable multiview information that could enhance segmentation accuracy. Our work aims to address this limitation by integrating multiview information into the foundation model, thereby fully unlocking its potential in image segmentation.

B. 3-D Reconstruction

Most deep learning-based 3-D reconstruction algorithms built upon MVSNet [26], [27], with various extensions such as Mono-MVS [28], R-MVSNet [29], and AACVP-MVSNet [30], demonstrating enhanced performance in general 3-D reconstruction tasks. However, when applied to aerial imagery, these methods face significant challenges due to differences in image characteristics, such as resolution and imaging mechanisms, necessitating specialized adaptations. To address this, Liu and Ji [31] introduced RED-Net, a recursive architecture for cost map regularization, which was later extended by Yu et al. [32] into an automatic 3-D building reconstruction method specifically designed for multiview aerial imagery.

While these advancements have improved MVS-based approaches for aerial imagery, applying them to satellite images remains challenging due to the complex rational polynomial coefficient (RPC) model, which governs satellite imaging geometry [33]. To address this issue, researchers have explored various strategies, including approximating the RPC model as a perspective camera model [34]. Additionally, Gao et al. [33], [35] proposed a rigorous RPC warping module to extend feature warping from pinhole models to RPC models, leading to the development of SatMVSF—a deep learning-based MVS framework tailored for satellite images.

More recently, multitask learning has been leveraged to enhance HE in 3-D reconstruction [36], [37], while attention-based stereo-matching networks have achieved SOTA performance in satellite image reconstruction [38]. However,

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

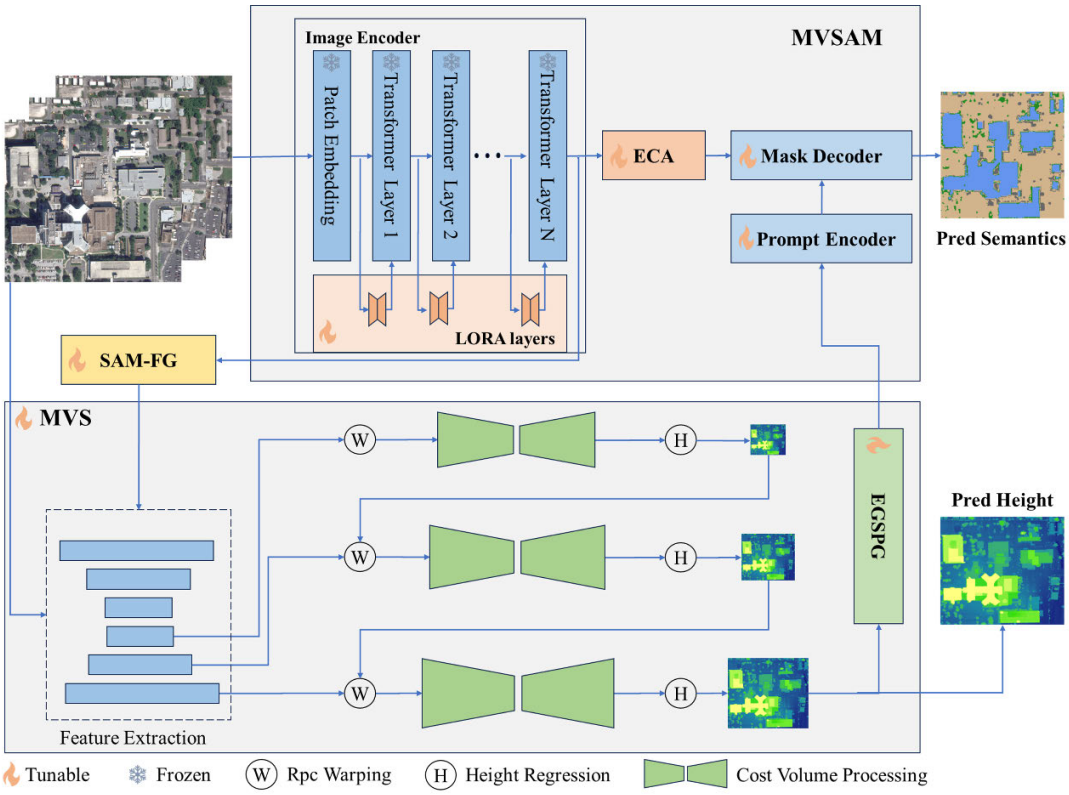

Fig. 2. Overall framework of MVSR3D. It consists of the MVSAM branch and the MVS branch. Among them, the ECA module aggregates multiview image embeddings, SAM-FG enhances feature extraction for MVS, and EGSPG provides sparse prompts to prompt the encoder through estimated height.

the feature encoding backbone of existing MVS models for satellite imagery may still be a limiting factor, leading to suboptimal performance in certain scenarios. To address this, we integrate the semantic features of SAM into the MVS encoder, aiming to enhance the accuracy of 3-D reconstruction.

C. Semantic 3-D Reconstruction

The semantic 3-D reconstruction task was first introduced in [39] to explore the complementary nature of SS and HE in satellite imagery. Since single-view semantic 3-D reconstruction struggles to fully exploit geometric constraints, research focus has gradually shifted toward pairwise and multiview semantic 3-D reconstruction.

In the pairwise setting, the top two winners of the semantic 3-D reconstruction challenge [40], [41] demonstrated that SS and HE can mutually benefit. Building upon these insights, researchers have recently developed an end-to-end pairwise semantic 3-D reconstruction network to improve information fusion [13], [14]. Additionally, S2Net and S3Net have enhanced performance by integrating a multitask learning framework [12], [15].

While recent studies have explored multiview semantic 3-D reconstruction, they do not adopt a multitask learning framework, leaving multiview semantic 3-D reconstruction an underdeveloped area [42]. For instance, competition-winning approaches achieved semantic 3-D reconstruction through postprocessing fusion [5], [6], while Leotta et al. [1] employed a sequential processing pipeline. Although these methods enable semantic 3-D reconstruction, SS and HE remain independent tasks, failing to fully exploit the benefits of multitask learning, which leads to suboptimal performance.

Although multitask learning has proven effective in various domains [43], [44], [45], existing methods still fall short in fully exploiting multiview information, resulting in suboptimal performance. Additionally, they primarily rely on a shared encoder, which limits the interaction between different task branches. In contrast, our approach not only integrates multiview information using geometric constraints but also facilitates deep interactions between different tasks, thereby improving semantic 3-D reconstruction performance.

III. Proposed Method

In this section, we introduce our proposed method, MVSR3D, which adopts a dual-stream structure comprising the MVSAM and MVS branches. The framework consists of four components: multiview semantic encoding, semantic-enhanced 3-D reconstruction, height-guided SS, and semantic 3-D reconstruction. To ensure effective interaction between the two tasks, we integrate the encoded embeddings from the MVSAM branch into the MVS branch's encoder via the SAM-FG module, while utilizing HE results as sparse prompts for the MVSAM branch through EGSPG. To further enhance multiview semantic representation, we introduce the innovative ECA module, which primarily integrates multiview information along the epipolar line. The proposed method enables semantic 3-D reconstruction in an end-to-end manner, processing multiview images as input, where one serves as

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

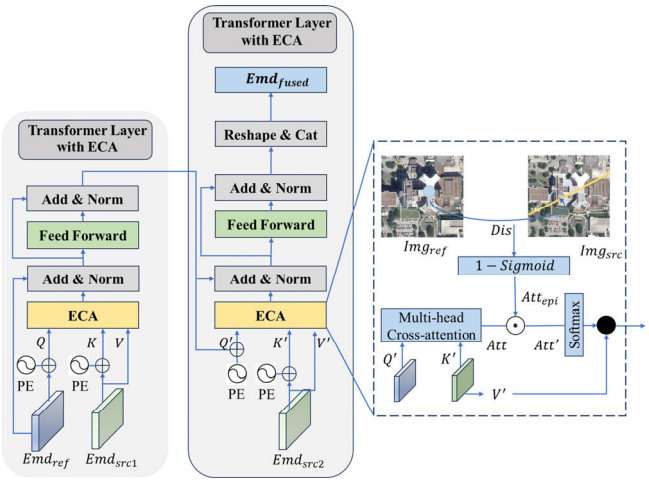

Fig. 3. Pipeline of the ECA module, and PE represents position encodings.

the reference image while the others act as source images. The overall framework is illustrated in Fig. 2.

A. Multi Multiview Semantic Encoding

We employ a low-rank adaptation (LoRA) [46] strategy to fine-tune the image encoder while applying full fine-tuning to the prompt encoder and decoder of MVSAM. At the same time, we designed the ECA module to fully integrate the semantic encoding of multiview images. The core idea of this module is to leverage feature fusion and geometric constraints via the RPC, ensuring accurate feature extraction.

The ECA module is embedded after the image encoder of MVSAM. For clarity, we illustrate our algorithm using a standard three-view stereo image setup, where one image serves as the reference and the remaining two act as source images. Our proposed ECA module is shown in Fig. 3.

We first apply the cross-view attention [47], [48] to compute the affinity matrix $\text{Att} \in R^{HW\times HW}$

$$\text{Att} = \text{softmax}\left(\frac{QK^T}{\sqrt{C}}\right) \tag{1}$$

where $Q \in R^{H\times W\times C}$ is the query matrix, obtained by adding positional encoding (PE) to the encoded features of the reference image (Emd_{ref}), while $K \in R^{H\times W\times C}$ represents the key matrix, containing encoded features of the source image (Emd_{src1}) with PE. Here, H, W, C represents the height, width, and number of channels of the feature map, respectively.

Since we cropped the image relatively small, we can approximate the oblique parallel projection of the satellite images as a central projection [33]. Specifically, we first obtain the geographic coordinates of the center point using RPC_{src} and then sample 1000 geographic 3-D coordinate points (Lat, Lon, Hei) around it. Next, we transform these points into the corresponding pixel coordinates X_{src}, X_{ref} through the RPC_{src} and RPC_{ref}, respectively, by

$$X_{\text{src}} = \text{RPC}_{\text{src}}^{-1}\begin{pmatrix}\text{Lat}\\ \text{Lon}\\ \text{Hei}\end{pmatrix} \tag{2}$$

$$X_{\text{ref}} = \text{RPC}_{\text{ref}}^{-1}\begin{pmatrix}\text{Lat}\\ \text{Lon}\\ \text{Hei}\end{pmatrix}. \tag{3}$$

Subsequently, we estimate the fundamental matrix F using a point correspondence-based method that minimizes algebraic error. This allows us to establish the epipolar geometry between the images, facilitating accurate multiview fusion. We also calculate the cross-attention affinity matrix Att_{epi} as follows:

$$E_{\text{ref}} = FX_{\text{src}} \tag{4}$$

$$\text{Att}_{\text{epi}} = 1 - \text{Sigmoid}(\text{Dis}(X_{\text{ref}}, E_{\text{ref}})) \tag{5}$$

where $\text{Dis}(\cdot,\cdot)$ denotes the distance from a point to a line, E_{ref} represents the epipolar line corresponding to the X_{src}, and $\text{Sigmoid}(\cdot)$ denotes sigmoid function. Note that a larger value indicates a higher correlation between the corresponding pixels of the source and reference encoding feature maps. Then, the affinity matrix Att′ and the output of the $\text{ECA}(Q, K, V) \in R^{H\times W\times C}$ is computed as

$$\text{Att}' = \text{Att} \odot \text{Att}_{\text{epi}} \tag{6}$$

$$\text{Fusion} = \text{ECA}(Q, K, V) = \text{Att}'V \tag{7}$$

where $V \in R^{H\times W\times C}$ is the value matrices, which represents the Emd_{src1} without PE.

Finally, the fused semantic encoding features fusion are processed through residuals, normalization, and a multilayer perceptron (MLP) before being input as the query matrix Q' into the next ECA

$$Emd_{\text{img}} = \text{ECA}(Q', K', V') \tag{8}$$

where K' represents the key matrix, which corresponds to the encoded features of the second source image with PE, while V' represents the value matrix without PE.

B. Semantic-Enhanced 3-D Reconstruction

Given the limitations of traditional MVS models for 3-D reconstruction in remote sensing, we leverage the Sat-MVS model [33] as the foundation of our approach. Our pipeline, shown in the bottom part of Fig. 2, follows a coarse-to-fine HE strategy. First, each image undergoes feature extraction and cost volume construction with RPC warping, followed by regularization. Finally, a height map (DSM) is inferred using a soft argmin operation along the height direction.

During the training phase, the pyramid network generates height maps at three different resolutions, and the loss is defined as

$$L_i = \frac{1}{|M|}\begin{cases}\sum_{x\in M} 0.5\left(h_{i,x} - h^*_{i,x}\right)^2, & \text{if } \left|h_{i,x} - h^*_{i,x}\right| < 1\\ \sum_{x\in M}\left|h_{i,x} - h^*_{i,x}\right|, & \text{otherwise}\end{cases} \tag{9}$$

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

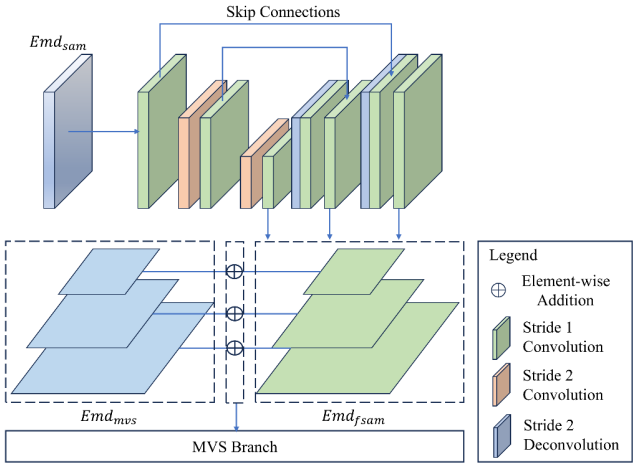

Fig. 4. Pipeline of the SAM-FG module.

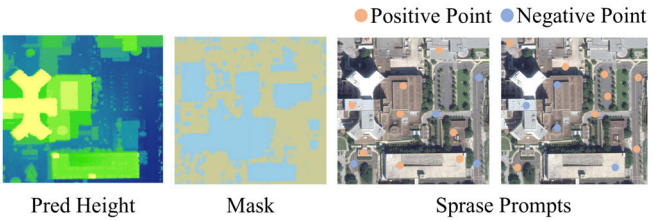

Fig. 5. Schematic illustrates the generated sparse prompts, with the images presented from left to right: the predicted height map, the mask m_{high} obtained from the H_{ave}, the sparse prompts for the m_{high} region, and the sparse prompts for the m_{low} region.

where M denotes the valid grid cells in the true height map, and $h_{i,x}$ and $h^*_{i,x}$ denote the predicted height value and the corresponding ground truth value at position x in stage i, respectively. The MVS loss is formulated as a weighted combination of the multistage loss terms L_i

$$L_{\text{MVS}} = \sum_{i=1}^{N} w_i \times L_i \tag{10}$$

where $w_i = \{0.5, 1, 2\}$ represents the weight assigned to each stage.

It should be noted that the accuracy of feature extraction directly affects the accuracy of 3-D reconstruction. To address this, we design the SAM-FG module to embed the rich encoded features of the MVSAM encoder into the MVS encoder. Fig. 4 illustrates the feature enhancement process. Specifically, feature enhancement begins with the multiscale extraction of Emd_{sam}, where Emd_{sam} represents the semantic encoding features from the MVSAM. This process can be expressed as follows:

$$Emd_{U-\text{sam}} = U\text{net}(Emd_{\text{sam}}) \tag{11}$$

where $Emd_{U-\text{sam}}$ is then resized to match the dimensions of the feature pyramid network features from the MVS encoder, Emd_{mvs} [35]

$$Emd_{f\text{sam}} = \text{Resize}(Emd_{U-\text{sam}}, \text{size}(Emd_{\text{mvs}})). \tag{12}$$

Subsequently, the resized SAM features are fused with the Emd_{mvs} through an elementwise addition operation

$$Emd_{\text{enhanced}} = Emd_{f\text{sam}} + Emd_{\text{mvs}}. \tag{13}$$

C. Height-Guided SS

Given the observation that the height of the building is generally higher than that of the ground and water, we use the height estimated by the MVS branch as the basis for creating sparse prompts for MVSAM.

Specifically, we first calculate the average height value H_{ave} of the reference image based on the MVS prediction H_{ref}. Then, we obtain the mask m_{high} for regions with heights greater than H_{ave} and the mask m_{low} for regions with heights less than H_{ave}, respectively, by

$$\begin{cases} m_{\text{high}} = H_{\text{ref}} \geq H_{\text{ave}} \\ m_{\text{low}} = H_{\text{ref}} < H_{\text{ave}}. \end{cases} \tag{14}$$

We then randomly sample points (termed sparse prompts) from the foreground and background regions of the target object within the masked areas, denoted as $P = \{P^{\text{for}}_{\text{high}}, P^{\text{back}}_{\text{high}}, P^{\text{for}}_{\text{low}}, P^{\text{back}}_{\text{low}}\}$. Fig. 5 visually demonstrates the randomized sparse prompts generation method.

Finally, we feed these points into the prompt encoder, $\text{Encoder}_{\text{prompt}}$, to obtain the prompt-encoded features Emd_{sparse}. Meanwhile, the ECA module extracts the image-encoded features Emd_{img}. Both feature sets are then fed into the decoder, with full fine-tuning applied to both the prompt encoder and the decoder

$$Emd_{\text{sparse}} = \text{Encoder}_{\text{prompt}}(P) \tag{15}$$

$$\text{Cls} = \text{Decoder}(Emd_{\text{img}}, Emd_{\text{sparse}}). \tag{16}$$

In addition, due to the pronounced seasonal variation in foliage within the dataset and the indistinguishable height of elevated roadways, sparse prompts are not added for these two categories. Therefore, in the DFC19 dataset, we add $\{P^{\text{for}}_{\text{high}}, P^{\text{back}}_{\text{high}}\}$ to the building category, and add $\{P^{\text{for}}_{\text{low}}, P^{\text{back}}_{\text{low}}\}$ to the ground and water categories. In the SpaceNet4 dataset, the building category is also added with $\{P^{\text{for}}_{\text{high}}, P^{\text{back}}_{\text{high}}\}$, and the ground category by adding $\{P^{\text{for}}_{\text{low}}, P^{\text{back}}_{\text{low}}\}$. Here, we add 1200 points to the foreground and 600 points to the background, with the number of points determined based on the image size.

D. Semantic 3-D Reconstruction

To achieve semantic 3-D reconstruction, we use multitask learning to train the network. Specifically, we combine the loss function of the SS branch L_{MVSAM} and the loss function of the HE branch L_{MVS} to train the network together

$$L = \alpha L_{\text{MVSAM}} + \beta L_{\text{MVS}} \tag{17}$$

where α and β denote the weights of the MVSAM and MVS branch loss functions, respectively. We set $\alpha = 20$ and $\beta = 1$ to appropriately weight the segmentation and MVS losses. Fixed values provided stable results for our multitask learning framework.

Finally, during 3-D point cloud generation, the postprocessing procedure follows the approach described in [35].

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

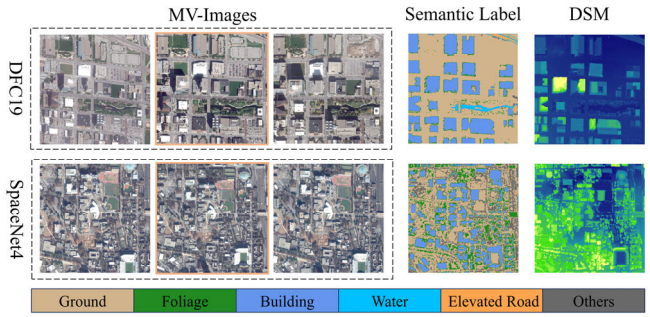

Fig. 6. Samples from the DFC19 and SpaceNet4 datasets and the reference images are surrounded by orange boxes.

Specifically, the reference image point P_1 is projected onto the source image P_2 using RPC and estimated height, and is then back-projected back onto the reference image to generate P_3. The two views are considered geometrically consistent if $||P_3 - P_1||_2 < \psi$. To ensure geometric consistency, the number of source views is represented by the parameter z. In addition, for each geometrically consistent point, we use the majority voting method to assign class labels. Specifically, the most frequently occurring label among segmentation results is assigned as the final class label for the corresponding point.

IV. Experiments

We first introduce two datasets for semantic 3-D reconstruction. Next, Section IV-B details the experimental setting and evaluation metrics. In Section IV-C, we provide a comparison with SOTA methods, followed by an ablation study in Section IV-D. Finally, Section IV-E presents an analysis of the strengths and weaknesses of our approach.

A. Data Preparation

We used two optical satellite image datasets to evaluate our proposed method: the DFC19 dataset and the SpaceNet4 dataset. Partial samples of the two datasets are shown in Fig. 6. In addition, geometric processing is often used to correct inaccuracies between RPC models [49], [50]. We used the bundle adjustment method [39] to refine the RPC parameters.

1) *The DFC19:* dataset contains 26 images collected in Jacksonville, Florida (JAX) from 2014 to 2016 and 43 images collected in Omaha, Nebraska (OMA) from 2014 to 2015. These images are WorldView-3's visible images with approximately 0.3 m ground sampling distance (GSD). In addition, the dataset provides DSM and semantic labels for each image, including six semantic categories: ground, foliage, building, water, elevated roadway, and others [51], [52], [53], [54].
The dataset covers diverse regions, including urban areas, forests, water bodies, and road networks. To mitigate the impact of seasonal variations on label accuracy, we selected images with different off-nadir angles from the JAX and OMA city datasets. The corresponding index values are $(4, 15, 19)$ for JAX and $(1, 33, 38)$ for OMA. We then combined these images with their height maps and semantic labels to create a standardized three-view semantic 3-D reconstruction dataset. Each image in the dataset was cropped to 512×512 pixels. The training dataset consists of 2976 sets of images, and the test dataset has 368 sets of images.
2) *The SpaceNet4:* dataset consists of 27 WorldView-2 satellite images with approximately 0.5 m GSD from different viewpoints. The images were acquired within a 5-min timeframe and cover an extensive 665-km^2 area in downtown Atlanta. The dataset also contains the corresponding DSM and semantic classification labels for four semantic categories: ground, foliage, building, and others [55]. It exhibits significant diversity, not only in its multiview characteristics but also in its extensive coverage of urban structures, parks, forests, roadways, and water bodies. We selected original images with different off-nadir angles, with corresponding index values $(7, 25, 32)$. Then, we combined them with the corresponding height maps and semantic labels to generate a set of standardized three-view semantic 3-D reconstruction datasets. Each image in the dataset was cropped to 512×512 pixels. The training dataset comprises 4528 images, while the validation and test datasets each contain 528 images.

B. Experimental Setting and Evaluation Metrics

1) Experimental Setting: To ensure fair comparisons among methods, all experiments were conducted on a server equipped with eight NVIDIA[1] GeForce RTX 4090 GPUs (24GB VRAM each), running on an Ubuntu 22.04 operating system. All codes were implemented using the PyTorch framework. The model was trained using the RMSprop optimizer ($\alpha = 0.9$) with a learning rate of 0.001 and a total batch size of 2.

For the MVS branch, referring to the settings in the original paper [33], [35], the height interval is determined based on the image height range provided in the RPC parameters, the number of depth hypotheses planes for the three stages is set to $\{64, 32, 8\}$, and the down-sampling rate of the image is set to $\{1/4, 1/2, 1\}$. Since the depth resolution of each stage is different, their loss weights are set to $\{0.5, 1.0, 2.0\}$, respectively.

2) Evaluation Metrics: In this article, the following metrics are used to assess the quality of SS, HE, and semantic 3-D reconstruction.

a) Metrics for semantic segmentation: To evaluate the accuracy of SS, we used mean intersection over union (mIoU) and mean$F1$ − score (mean$F1$).

1) *mIoU:* the measurement of overlap between predicted outcomes and GT

$$\text{IoU} = \frac{\text{TP}^k}{\text{TP}^k + \text{FP}^k + \text{FN}^k} \tag{18}$$

$$\text{mIoU} = \frac{1}{K}\sum_{k=1}^{K} \frac{\text{TP}^k}{\text{TP}^k + \text{FP}^k + \text{FN}^k} \tag{19}$$

where TP^k, FP^k, TN^k, and FN^k denote the true positive number, the false positive number, the true negative

[1]Registered trademark.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

number, and the false negative number for kth class, respectively, in the confusion matrix. K denotes the number of categories.

2) *mean F1:* the harmonic mean of precision and recall

$$\text{mean } F1 = 2 \times \frac{1}{K} \sum_{k=1}^{K} \frac{\text{Precision}^k \times \text{Recall}^k}{\text{Precision}^k + \text{Recall}^k} \tag{20}$$

where Precision^k and Recall^k are defined as follows:

$$\text{Precision}^k = \frac{\text{TP}^k}{\text{TP}^k + \text{FP}^k}, \quad \text{Recall}^k = \frac{\text{TP}^k}{\text{TP}^k + \text{FN}^k}. \tag{21}$$

b) Metrics for HE: The mean absolute error (MAE), root mean-square error (RMSE), and percentage of accurate grids in total (PAG) [35] are used to evaluate the accuracy of HE.

1) *MAE:* the average over all pixels of the $L1$ distance between the true value of the height and the predicted height

$$\text{MAE} = \frac{\sum_{(i,j)\in G\cap G^*} |h_{ij} - h_{ij}^*|}{\sum_{(i,j)\in G\cap G^*} I((i,j) \in G \cap G^*)} \tag{22}$$

where $I(A)$ denotes the Iverson bracket, $I(A) = 1$ if A is true and $I(A) = 0$, otherwise. G and G^* denote the valid grid cells in the height map predicted and true height values, and h_{ij} and h_{ij}^* denote the predicted and true values of the heights in the pixels of the ith row and jth column.

2) *RMSE:* the standard deviation of the $L1$ distance between the true height value and the estimated height

$$\text{RMSE} = \sqrt{\frac{\sum_{(i,j)\in G\cap G^*} \left(h_{ij} - h_{ij}^*\right)^2}{\sum_{(i,j)\in G\cap G^*} I((i,j) \in G \cap G^*)}}. \tag{23}$$

3) *PAG:* the percentage of grid cells where the L1 distance error is less than the threshold α

$$\text{PAG}_\alpha = \frac{\sum_{(i,j)\in G\cap G^*} I\left(\left|h_{ij} - h_{ij}^*\right| < \alpha\right)}{\sum_{(i,j)\in G\cap G^*} I((i,j) \in G^*)}. \tag{24}$$

c) Metrics for semantic 3-D reconstruction:

1) *mIoU3$_\beta$:* as a combined metric of mIOU and MAE, denotes the measurement of overlap between semantic prediction and SS GT with MAE less than threshold β [39]

$$\text{mIoU3}_\beta = \frac{1}{K} \sum_{k=1}^{K} \frac{\text{TP3}_\beta^k}{\text{TP3}_\beta^k + \text{FP}^k + \text{FN}^k} \tag{25}$$

$$\text{TP3}_\beta^k = \text{TP}^k(\text{MAE} < \beta). \tag{26}$$

C. Comparison With SOTA Methods

In this study, MVSR3D was compared with several existing SOTA methods, including single-view SS (SV-SS), multiview HE (MV-HE), single-view multitask learning (SV-MTL), and multiview multitask learning (MV-MTL). Brief descriptions of these methods are as follows.

1) *SV-SS:* Includes classical CNN-based methods, such as PSPNet [56] and DeepLab V3+ [57], as well as SOTA transformer-based methods, such as SAM-Frozen [16] and SAM-LoRA [17].
2) *MV-HE:* Includes SOTA MVS methods, such as A-SATMVSNet [38] and SatMVS [35].
3) *SV-MTL:* Includes SOTA methods like Carvalho et al. [9], Mti-net [10] and MQTransformer [11].
4) *MV-MTL:* Includes baselines like Bosch et al. [39] and Qin et al. [6], which decompose multiview images into pairs and apply pairwise semantic stereo processing.

We report the quantitative evaluation results for the DFC19 dataset in Table I. Compared to the SOTA methods, the MVSR3D achieves better performance in both SS and HE. Specifically, compared to SV-SS, the MVSR3D is 0.70% higher on mean $F1$ T and 1.06% higher on mIoU metric. Compared with MV-HE, the MVSR3D has a 0.076 m lower error on RMSE metric, and 0.96% higher on $\text{PAG}_{2.5}$ metric. Compared to the SV-MTL model, the MVSR3D is 3.00% higher on mean $F1$ T, 4.40% higher on mIoU metric, 10.39% higher on $\text{mIoU3}_{1.0}$ metric, and 8.69% higher on $\text{mIoU3}_{2.5}$ metric. Compared with the SOTA method of the MV-MTL model, the MVSR3D is 41.04% higher on mean $F1$ T, 39.34% higher on mIoU metric, 12.673 m lower on RMSE metric, 52.16% higher on $\text{PAG}_{2.5}$ metric, 44.56% higher on $\text{mIoU3}_{1.0}$ metric, and 45.11% higher on $\text{mIoU3}_{2.5}$ metric. It is worth noting that we did not do the comparison experiment of single-view HE and did not record the HE metrics from the SV-MTL model, as single-view approaches can only estimate height above ground level (AGL) rather than the true height. In addition, for comparison purposes, we calculated the mIoU3 metrics by summing the ground level height (DEM) with the AGL. The visualization results are shown in Fig. 7.

We further conducted experiments on the SpaceNet4 dataset for comparison. As shown in Table II, similar to the DFC19 dataset, the MVSR3D outperforms the current SOTA methods in all metrics. In particular, compared with the SOTA MV-MTL, the MVSR3D is 25.35% higher on mean $F1$ T, 30.80% higher on mIoU metric, 11.760 m lower on RMSE metric, 43.29% higher on $\text{PAG}_{2.5}$ metric, 40.35% higher on $\text{mIoU3}_{1.0}$ metric, and 37.09% higher on $\text{mIoU3}_{2.5}$ metric. Fig. 8 also shows the qualitative results on the SpaceNet4 dataset, which proves the effectiveness of the MVSR3D.

Overall, as shown in Tables I and II, our method outperforms all SV-SS and SV-MTL approaches. This can likely be attributed to not only the fine-tuning of our MVSAM branch based on SAM but also the effective integration of multiview information from MVS data. Furthermore, our approach surpasses all MV-HE and MV-MTL methods, which may be due to the deep interaction between the two task branches in an end-to-end manner, rather than merely sharing an encoder.

In addition, we selected the top ten sets of three-view image pairs with the highest scores based on the commonly used principle [35]. The obtained HE results were postprocessed with thresholds ψ and z set to 1 each. After fusing the ten sets of prediction results, the semantic 3-D reconstruction results were obtained, as shown in Fig. 9. From this figure, it is evident that the reconstructed point cloud is assigned semantic categories, resulting in a more detailed, categorized

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

TABLE I
QUANTITATIVE RESULTS OF COMPARATIVE EXPERIMENTS ON THE DFC19 DATASET. THE BEST PERFORMANCE IS MARKED IN BOLD

	Methods	Task: Semantic Seg.		Task: Height Est.		Task: Semantic 3D Reconstruction.	
		$mean\ F1$	$mIoU$	$RMSE$	$PAG_{2.5}$	$mIoU3_{1.0}$	$mIoU3_{2.5}$
SV-SS	PSPNet [56]	43.17	33.19	--	--	--	--
	DeepLab V3+ [57]	42.95	32.63	--	--	--	--
	SAM-Frozen [16]	69.18	54.80	--	--	--	--
	SAM-LoRA [17]	86.28	76.41	--	--	--	--
MV-HE	A-SATMVSNet [38]	--	--	11.370	46.65	--	--
	SatMVS [35]	--	--	3.884	78.77	--	--
SV-MTL	Carvalho *et al.* [9]	51.32	42.44	--	--	26.56	34.07
	Mti-net [10]	81.32	69.42	--	--	48.45	58.39
	MQTransformer [11]	83.98	73.07	--	--	54.31	64.44
MV-MTL	Bosch *et al.* [39]	45.94	38.13	16.481	27.57	20.14	28.02
	Qin *et al.* [6]	40.71	31.47	16.553	27.95	14.27	21.63
	MVSR3D (Ours)	**86.98**	**77.47**	**3.808**	**79.73**	**64.70**	**73.13**

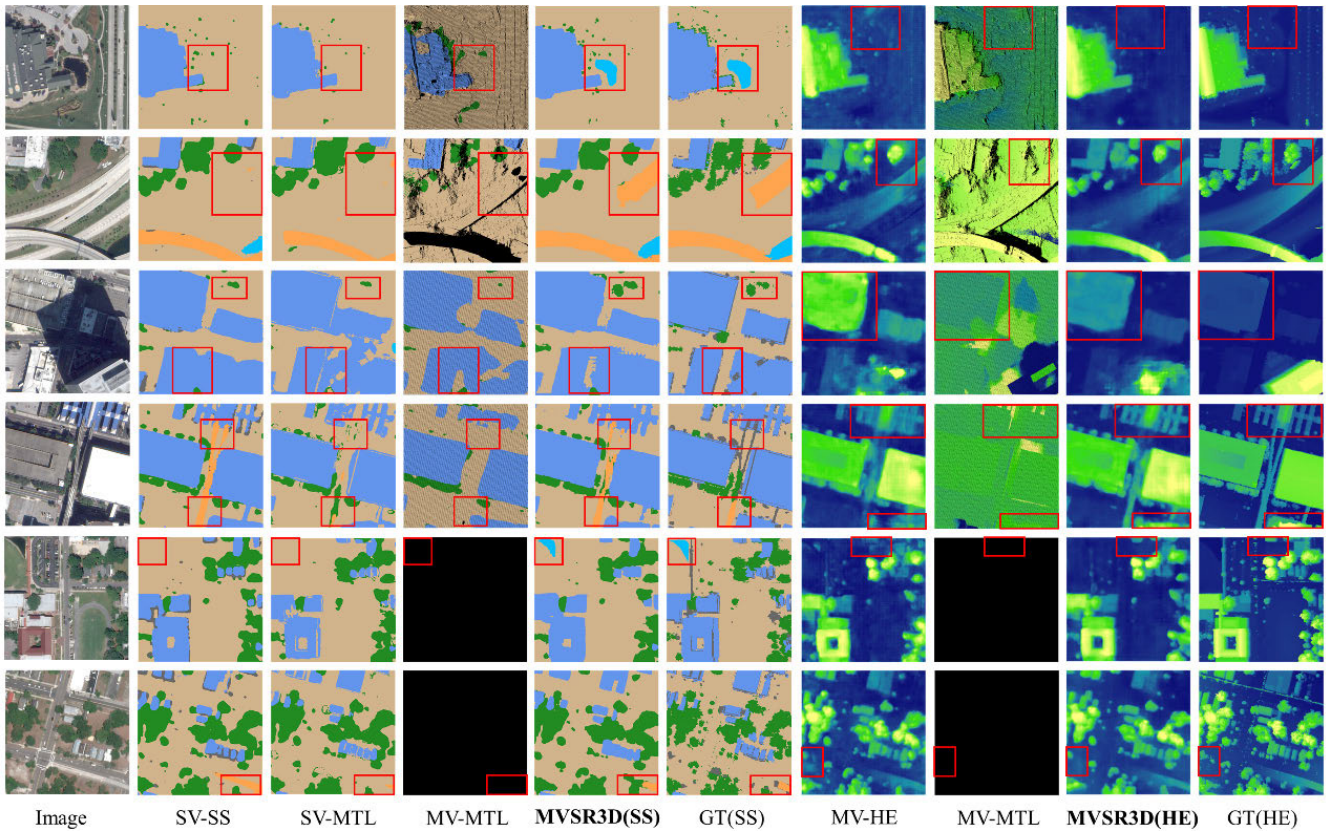

Fig. 7. Qualitative results of HE and SS on the DFC19 dataset, where black pixels indicate no-data values.

point cloud. These results highlight the potential of MVSR3D for practical applications.

D. Ablation Study

In this section, we conducted extensive ablation experiments on two datasets to validate the effectiveness of our proposed modules and core ideas. Specifically, we set up ablation studies on the semantic 3-D reconstruction task, HE task, SS task, and multiview fusion strategy, respectively.

1) Effectiveness of Components on Semantic 3-D Reconstruction: To assess the impact of each component, we report the quantitative results for the various network variants in Table III, along with SS performance. We can draw several conclusions from Table III.

1) The ECA module and the EGSPG module do bring gains to the SS task and semantic 3-D reconstruction task. However, using them alone brings negative gains in some cases, which may be caused by the large seasonal differences in the datasets. In addition, experiments

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

TABLE II
QUANTITATIVE RESULTS OF COMPARATIVE EXPERIMENTS ON THE SPACENET4 DATASET. THE BEST PERFORMANCE IS MARKED IN BOLD

	Methods	Task: Semantic Seg.		Task: Height Est.		Task: Semantic 3D Reconstruction.	
		$mean\ F1$	$mIoU$	$RMSE$	$PAG_{2.5}$	$mIoU3_{1.0}$	$mIoU3_{2.5}$
SV-SS	PSPNet [56]	82.05	70.11	--	--	--	--
	DeepLab V3+ [57]	82.18	70.27	--	--	--	--
	SAM-Frozen [16]	76.33	62.49	--	--	--	--
	SAM-LoRA [17]	86.66	76.68	--	--	--	--
MV-HE	A-SATMVSNet [38]	--	--	5.685	57.66	--	--
	SatMVS [35]	--	--	4.396	65.74	--	--
SV-MTL	Carvalho *et al.* [9]	84.32	73.35	--	--	53.48	65.62
	Mti-net [10]	86.78	76.97	--	--	60.34	71.00
	MQTransformer [11]	85.95	75.78	--	--	57.18	68.95
MV-MTL	Bosch *et al.* [39]	62.04	47.07	16.081	23.52	23.23	35.25
	Qin *et al.* [6]	62.54	47.81	15.957	23.80	24.26	36.51
	MVSR3D (Ours)	**87.89**	**78.61**	**4.197**	**67.09**	**64.61**	**73.60**

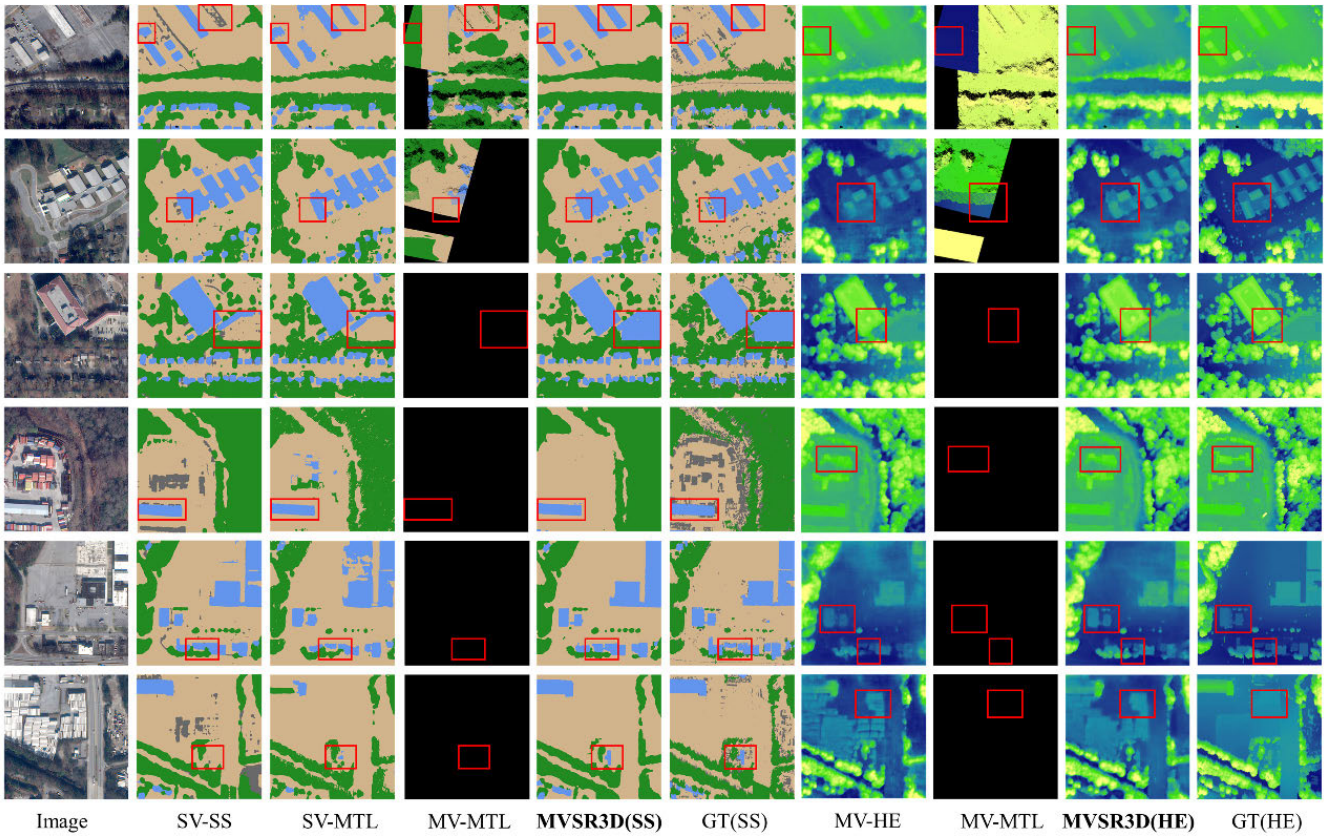

Fig. 8. Qualitative results of HE and SS on the SpaceNet4 dataset, where black pixels indicate no-data values.

demonstrate that using both modules together resulted in greater gains, indicating that the two models are mutually reinforcing (comparing the fourth row with the first row and the fifth row with the last row in each dataset). These improvements stem from two complementary aspects—ECA enhances multiview information integration, while EGSPG facilitates learning with sparse prompts—as confirmed by the experimental results.

2) The combination of the SAM-FG module with either the ECA or EGSPG module gives better results in most cases (comparing the third row with the seventh row and the second row with the sixth row in each dataset). This may be attributed to the improved accuracy of HE, which enables the EGSPG module to provide more precise prompts and enhances the learning of multiview semantic features.
3) Our full model achieves the best performance on both tasks (comparing the first row with the last row in each dataset), which demonstrates the effectiveness of the MVSR3D and its three modules. It is important to note

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

TABLE III
ABLATION EXPERIMENTS FOR EACH MODULE IN THE MODEL. THE BEST PERFORMANCE IS MARKED IN BOLD

Components			DFC19 Dataset. (Higher Better)			SpaceNet4 Dataset. (Higher Better)		
ECA	EGSPG	SAM-FG	$mean\ F1$	$mIoU$	$mIoU3_{2.5}$	$mean\ F1$	$mIoU$	$mIoU3_{2.5}$
×	×	×	86.78	77.15	72.10	87.76	78.40	73.34
√	×	×	86.52	76.83	71.99	87.77	78.42	73.36
×	√	×	86.19	76.27	71.34	87.74	78.37	73.27
√	√	×	86.36	76.51	71.64	87.89	78.60	73.45
×	×	√	86.86	77.27	72.86	87.64	78.22	73.29
√	×	√	86.69	77.03	72.64	87.84	78.53	73.44
×	√	√	86.64	76.99	72.09	87.79	78.44	73.43
√	√	√	**86.98**	**77.47**	**73.13**	**87.89**	**78.61**	**73.60**

TABLE IV
ABLATION EXPERIMENTS FOR SAM-FG. THE BEST PERFORMANCE IS MARKED IN BOLD

Methods	DFC19 Dataset. (Lower Better)		DFC19 Dataset. (Higher Better)		SpaceNet4 Dataset. (Lower Better)		SpaceNet4 Dataset. (Higher Better)	
	MAE	$RMSE$	$PAG_{2.5}$	$PAG_{7.5}$	MAE	$RMSE$	$PAG_{2.5}$	$PAG_{7.5}$
MTL	2.134	3.753	79.29	94.86	2.666	4.267	66.98	91.74
MTL+SAM-FG	**1.989**	**3.400**	**80.20**	**95.73**	**2.629**	**4.226**	**67.60**	**92.02**

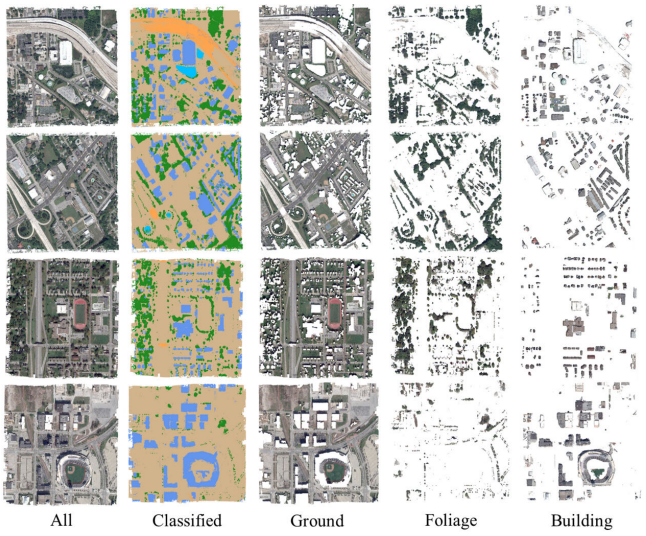

Fig. 9. Visualization of semantic 3-D reconstruction. The point clouds for each study area are divided into three semantic categories (ground, foliage, and building) for visualization.

that the first row in each dataset represents the baseline model in our ablation study, which follows a standard multitask pipeline but lacks our proposed two-branch deep interaction mechanism and multiview information fusion module.

2) Effectiveness of SAM-FG on HE: In this section, we set up a series of experiments to validate the effectiveness of the SAM-FG module for the HE task, with the results presented in Table IV. We report quantitative results for two network variants, including the multitasking network without any modules (MTL), and the MTL network with SAM-FG modules (MTL + SAM-FG).

From Table IV, it can be seen that better HE results are achieved by adding the SAM-FG module, which suggests that enriched SAM semantic features can indeed facilitate the learning of the HE branch.

3) Effectiveness of EGSPG on SS: In this section, we conducted a quantitative study to evaluate the EGSPG module, as shown in Table V. For the DFC19 dataset, we added sparse prompts to the ground, building, and water categories, respectively. For the SpaceNet4 dataset, we added sparse prompts to the ground and building categories, respectively. We used the MTL network with the SAM-FG module as the baseline. The experimental results show that adding EGSPG either independently or in combination with ECA improves the SS performance across most categories.

4) Effectiveness of ECA on Multi-Multiview Fusion: Here, we validated the benefits of multiview image feature fusion for both the SS and semantic 3-D reconstruction tasks, as consistently emphasized throughout this article. We conducted a set of comparative experiments, including MVSR3D without the ECA module [MVSR3D (w/o ECA)] and the full MVSR3D with the ECA module [MVSR3D (w/ECA)]. The experimental results are shown in Table VI.

From Table VI, we can see that using the ECA module for multiview feature fusion not only enhances the semantic 3-D reconstruction task but also significantly improves SS quality. We further visualized the impact of multiview information fusion in Fig. 10, which is mainly reflected in the following three aspects.

1) It effectively detects buildings with indistinct features, especially those that are challenging to identify from a single-view image, even for human observers (e.g., a rooftop parking lot). As illustrated in the first row of Fig. 10, multiview observations correctly classify

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

TABLE V
ABLATION EXPERIMENTS FOR EGSPG. THE BEST PERFORMANCE IS MARKED IN BOLD

Methods	DFC19 Dataset.			SpaceNet4 Dataset.	
	$IoU\uparrow$			$IoU\uparrow$	
	Ground	Building	Water	Ground	Building
MTL+SAM-FG	85.47	80.36	87.88	78.70	70.23
MTL+SAM-FG+EGSPG	85.49	80.34	**88.43**	78.77	70.84
MTL+SAM-FG+ECA+EGSPG	**85.60**	**80.87**	88.05	**78.98**	**71.01**

TABLE VI
ABLATION EXPERIMENTS FOR MULTIVIEW FUSION. THE BEST PERFORMANCE IS MARKED IN BOLD

Methods	DFC19 Dataset.			SpaceNet4 Dataset.		
	(Higher Better)			(Higher Better)		
	$mean\ F1$	$mIoU$	$mIoU3_{2.5}$	$mean\ F1$	$mIoU$	$mIoU3_{2.5}$
MVSR3D (w/o ECA)	86.64	76.99	72.09	87.79	78.44	73.43
MVSR3D (w/ ECA)	**86.98**	**77.47**	**73.13**	**87.89**	**78.61**	**73.60**

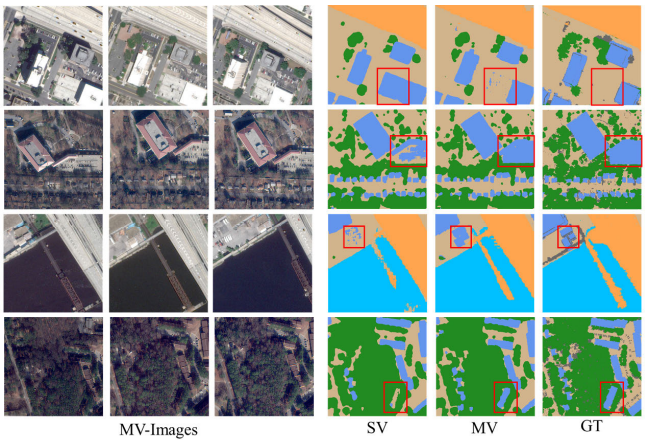

Fig. 10. Segmentation results before and after multiview fusion, highlighting the improvement in SS achieved through multiview information fusion. MV and SV denote MVSR3D with and without the ECA module, respectively.

a rooftop parking lot instead of mislabeling it as a building, while the second row demonstrates an accurate identification of a building instead of a parking lot.

2) It significantly enhances the recognition of textureless regions, as shown in the third row of Fig. 10.
3) It has the potential to effectively recognize hidden or occluded objects, as shown in the fourth row of Fig. 10. These surprising results above are mainly attributed to the fact that the multiview observation provides additional information from different angles, and we have successfully achieved the fusion of this information.

Additionally, we compared the experimental results of the current SOTA SS method, SAM-LORA, with those of our proposed MVSAM. It should be noted that MVSAM builds on SAM-LORA by incorporating the ECA module, which integrates multiview semantic information into this foundational model. All models here were trained in a single-task manner, without adding height branch loss. The experimental results are shown in Table VII.

The experimental results indicate that our MVSAM, which incorporates multiview information, outperforms SAM-LORA across all metrics. This demonstrates the effectiveness of our proposed model and highlights our success in embedding multiview information into the foundational model, thereby unlocking greater potential from MVS data.

E. Strengths and Weaknesses

In this study, we propose a novel framework MVSR3D that combines MVS with the foundation model in an end-to-end manner, effectively leveraging multiview information to enhance SS and HE. A key innovation of our approach is integrating multiview features into the segmentation foundation model, unlocking the potential of multiview satellite imagery within it. Consequently, our method achieves superior semantic 3-D reconstruction and surpasses traditional SOTA methods across all evaluation metrics. Furthermore, unlike conventional multitask learning methods, we establish effective interactions between the MVSAM and HE branches, forming a well-structured multitask learning framework.

However, progress in this area remains constrained by the limitations of existing datasets. The two datasets utilized in this study are currently the only publicly available datasets for this task, but they require extensive preprocessing, including pair selection, due to their significant temporal differences. Future advances in dataset quality are expected to further drive improvements in this field.

In addition, the illumination of images taken on different dates varies greatly, which poses a great challenge to this task. Since the segmentation branch of the MVSR3D framework is built on the foundation model, it is inherently robust to changes in input color and illumination, as shown in the first row of Fig. 11. However, this capability is not absolute. As shown in the second row of Fig. 11, we observe that it still leads to suboptimal segmentation in some extreme cases. In the future, we can further mitigate this problem through

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

TABLE VII

ABLATION EXPERIMENTS FOR MULTIVIEW FUSION (FOR THE SINGLE-TASK SS). THIS TABLE COMPARES THE PERFORMANCE OF MVSAM, WHICH INCORPORATES EMBEDDED MULTIVIEW INFORMATION, WITH THE SOTA METHOD. THE BEST PERFORMANCE IS HIGHLIGHTED IN BOLD

Methods	DFC19 Dataset. (Higher Better)		SpaceNet4 Dataset. (Higher Better)	
	mean F1	*mIoU*	*mean F1*	*mIoU*
SAM-LORA	86.28	76.41	86.66	76.68
MVSAM (Ours)	**86.79**	**77.17**	**87.82**	**78.50**

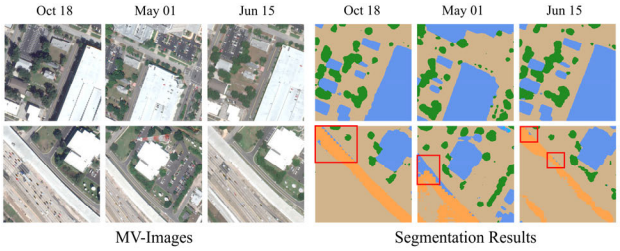

Fig. 11. Segmentation results for images affected by drastic illumination.

more advanced preprocessing techniques or by incorporating specialized network components.

V. CONCLUSION

In this work, we propose a new technical approach based on multiview fusion and multitask learning, introducing the MVSR3D framework—the first end-to-end framework for semantic 3-D reconstruction using multiview remote sensing images. This marks a pioneering contribution in the field of satellite remote sensing. Specifically, the network fully fuses the multiview image features primarily along the epipolar line by the ECA module. In addition, the SAM-FG module and the EGSPG module facilitate effective interaction between the MVSAM branch and the HE branch, forming a well-structured multitask learning framework. The experimental results show that the MVSR3D substantially leads the SOTA method of MV-MTL. In future work, we will further explore the feasibility of semantic 3-D reconstruction in an unsupervised or self-supervised manner to get rid of the dependence on labels. In addition, the MVSR3D framework holds great promise for advancing 3-D modeling processes, potentially enabling the generation of finer-grained 3-D building models, thereby increasing its wider impact and practical applications.

REFERENCES

[1] M. J. Leotta et al., "Urban semantic 3D reconstruction from multiview satellite imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 1451–1460.

[2] V. Vineet et al., "Incremental dense semantic stereo fusion for large-scale semantic scene reconstruction," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2015, pp. 75–82.

[3] D. Fernandes et al., "Point-cloud based 3D object detection and classification methods for self-driving applications: A survey and taxonomy," *Inf. Fusion*, vol. 68, pp. 161–191, Apr. 2021.

[4] 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, Nov. 2024, Art. no. 102498.

[5] P. d'Angelo et al., "3D semantic segmentation from multi-view optical satellite images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2019, pp. 5053–5056.

[6] 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.*, Jul. 2019, pp. 5057–5060.

[7] S. Liu, E. Johns, and A. J. Davison, "End-to-end multi-task learning with attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1871–1880.

[8] Y. Feng et al., "Height aware understanding of remote sensing images based on cross-task interaction," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 233–249, Jan. 2023.

[9] M. Carvalho, B. Le Saux, P. Trouvé-Peloux, F. Champagnat, and A. Almansa, "Multitask learning of height and semantics from aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1391–1395, Aug. 2020.

[10] S. Vandenhende, S. Georgoulis, and L. Van Gool, "MTI-Net: Multi-scale task interaction networks for multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 527–543.

[11] Y. Xu, X. Li, H. Yuan, Y. Yang, and L. Zhang, "Multi-task learning with multi-query transformer for dense prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 2, pp. 1228–1240, Feb. 2024.

[12] P. Liao et al., "S2Net: A multitask learning network for semantic stereo of satellite image pairs," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024, Art. no. 5601313.

[13] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "SDBF-net: Semantic and disparity bidirectional fusion network for 3D semantic detection on incidental satellite images," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Nov. 2019, pp. 438–444.

[14] Z. Rao, M. He, Z. Zhu, Y. Dai, and R. He, "Bidirectional guided attention network for 3-D semantic detection of remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 6138–6153, Jul. 2021.

[15] Q. Yang, G. Chen, X. Tan, T. Wang, J. Wang, and X. Zhang, "S3Net: Innovating stereo matching and semantic segmentation with a single-branch semantic stereo network in satellite epipolar imagery," 2024, *arXiv:2401.01643*.

[16] A. M. Kirillov et al., "Segment anything," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 4015–4026.

[17] K. Zhang and D. Liu, "Customized segment anything model for medical image segmentation," 2023, *arXiv:2304.13785*.

[18] X. Zou et al., "Segment everything everywhere all at once," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 19769–19782.

[19] J. Oin et al., "FreeSeg: Unified, universal and open-vocabulary image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19446–19455.

[20] L. P. Osco et al., "The segment anything model (SAM) for remote sensing applications: From zero to one shot," *Int. J. Appl. Earth Observ. Geoinformation*, vol. 124, Nov. 2023, Art. no. 103540.

[21] K. Chen et al., "RSPrompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–17, 2024, Art. no. 4701117.

[22] D. Wang, J. Zhang, B. Du, D. Tao, and L. Zhang, "SAMRS: Scaling-up remote sensing segmentation dataset with segment anything model," in *Proc. Adv. Neural Inf. Process. Syst.*, Jan. 2023, pp. 8815–8827.

[23] Z. Yan et al., "RingMo-SAM: A foundation model for segment anything in multimodal remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023, Art. no. 5625716.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

[24] S. Julka and M. Granitzer, "Knowledge distillation with segment anything (SAM) model for planetary geological mapping," in *Proc. Int. Conf. Mach. Learn., Optim., Data Sci.* Cham, Switzerland: Springer, Jan. 2023, pp. 68–77.

[25] X. Guo et al., "SkySense: A multi-modal remote sensing foundation model towards universal interpretation for Earth observation imagery," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 27662–27673.

[26] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "MVSNet: Depth inference for unstructured multi-view stereo," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 785–801.

[27] E. K. Stathopoulou and F. Remondino, "A survey on conventional and learning-based methods for multi-view stereo," *Photogramm. Rec.*, vol. 38, no. 183, pp. 374–407, Sep. 2023.

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

[29] Y. Yao, Z. Luo, S. Li, T. Shen, T. Fang, and L. Quan, "Recurrent MVSNet for high-resolution multi-view stereo depth inference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5520–5529.

[30] A. Yu et al., "Attention aware cost volume pyramid based multi-view stereo network for 3D reconstruction," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 448–460, May 2021.

[31] J. Liu and S. Ji, "A novel recurrent encoder–decoder structure for large-scale multi-view stereo reconstruction from an open aerial dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 6049–6058.

[32] D. Yu, S. Ji, J. Liu, and S. Wei, "Automatic 3D building reconstruction from multi-view aerial images with deep learning," *ISPRS J. Photogramm. Remote Sens.*, vol. 171, pp. 155–170, Jan. 2021.

[33] J. Gao, J. Liu, and S. Ji, "Rational polynomial camera model warping for deep learning based satellite multi-view stereo matching," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 6128–6137.

[34] K. Zhang, N. Snavely, and J. Sun, "Leveraging vision reconstruction pipelines for satellite imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2139–2148.

[35] J. Gao, J. Liu, and S. Ji, "A general deep learning based framework for 3D reconstruction from multi-view stereo satellite images," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 446–461, Jan. 2023.

[36] 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. (WACV)*, Jan. 2024, pp. 1317–1327.

[37] X. Huang, S. Zhang, J. Li, and L. Wang, "A multitask network for multiview stereo reconstruction: When semantic consistency-based clustering meets depth estimation optimization," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–16, 2024, Art. no. 5612816.

[38] L. Lin, Y. Zhang, Z. Wang, L. Zhang, X. Liu, and Q. Wang, "A-SATMVSNet: An attention-aware multi-view stereo matching network based on satellite imagery," *Frontiers Earth Sci.*, vol. 11, Apr. 2023, Art. no. 1108403.

[39] 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. (WACV)*, Jan. 2019, pp. 1524–1532.

[40] H. Chen et al., "Multi-level fusion of the multi-receptive fields contextual networks and disparity network for pairwise semantic stereo," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 4967–4970.

[41] R. Qin, X. Huang, W. Liu, and C. Xiao, "Pairwise stereo image disparity and semantics estimation with the combination of U-Net and pyramid stereo matching network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2019, pp. 4971–4974.

[42] S. Kunwar et al., "Large-scale semantic 3-D reconstruction: Outcome of the 2019 IEEE GRSS data fusion contest—Part a," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 922–935, 2021.

[43] Q. Wu, Y. Wan, Z. Zheng, Y. Zhang, G. Wang, and Z. Zhao, "CAMP: A cross-view geo-localization method using contrastive attributes mining and position-aware partitioning," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024, Art. no. 5637614.

[44] P. Xia, Y. Wan, Z. Zheng, Y. Zhang, and J. Deng, "Enhancing cross-view geo-localization with domain alignment and scene consistency," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 12, pp. 13271–13281, Dec. 2024.

[45] L. He, J. Lu, G. Wang, S. Song, and J. Zhou, "SOSD-net: Joint semantic object segmentation and depth estimation from monocular images," *Neurocomputing*, vol. 440, pp. 251–263, Jun. 2021.

[46] E. J. Hu et al., "LoRA: Low-rank adaptation of large language models," 2021, *arXiv:2106.09685*.

[47] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 13750–13759.

[48] H.-Y. Tseng, Q. Li, C. Kim, S. Alsisan, J.-B. Huang, and J. Kopf, "Consistent view synthesis with pose-guided diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 16773–16783.

[49] Q. Fu et al., "GPU-accelerated PCG method for the block adjustment of large-scale high-resolution optical satellite imagery without GCPs," *Photogramm. Eng. Remote Sens.*, vol. 89, no. 4, pp. 211–220, Apr. 2023.

[50] Y. Zhang, N. Yang, and Q. Luo, "A matching optimization algorithm about low-altitude remote sensing images based on geometrical constraint and convolutional neural network," *Photogramm. Eng. Remote Sens.*, vol. 88, no. 8, pp. 527–533, Aug. 2022.

[51] 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. (CVPR)*, Jun. 2020, pp. 14500–14508.

[52] 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. Workshops (CVPRW)*, Jun. 2021, pp. 1162–1171.

[53] B. Le 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, Jan. 2019.

[54] W. Li, L. Meng, J. Wang, C. He, G.-S. Xia, and D. Lin, "3D building reconstruction from monocular remote sensing images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 12528–12537.

[55] H. Hao et al., "Improving building segmentation for off-nadir satellite imagery," 2021, *arXiv:2109.03961*.

[56] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[57] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Jan. 2018, pp. 833–851.

Xuejun Huang received the B.S. degree from Harbin Engineering University, Harbin, China, in 2022, where he is currently pursuing the Ph.D. degree with the School of Remote Sensing and Information Engineering.

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

Xinyi Liu received the B.S. and Ph.D. degrees from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2014 and 2020, respectively.

She is currently a Post-Doctoral Researcher with Wuhan University. Her research interests include 3-D reconstruction, LiDAR and image integration, and texture mapping.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.

Yi Wan (Member, IEEE) was born in 1991. He received the B.S. and Ph.D. degrees from Wuhan University, Wuhan, China, in 2013 and 2018, respectively.

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.

Zhi Zheng received the B.S. degree in remote sensing and the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017 and 2023, respectively.

He is currently a Post-Doctoral Fellow at the Department of Geography and Resource Management, The Chinese University of Hong Kong, Hong Kong, China. He has published more than ten research articles. His research interests include satellite remote sensing, stereo matching, change detection, and geohazard monitoring using deep learning technology.

Dr. Zheng was awarded the Research Fellowship Scheme by the Chinese University of Hong Kong, in January 2024. In recent years, he has frequently served as a referee for several international journals.

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 from 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 from Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the Ph.D. degree in photogrammetry and remote sensing.

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

Haoyu Guo was born in 1992.

His research interests are in satellite imagery photogrammetry and remote sensing.

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 of the School of Remote Sensing and Information Engineering, Wuhan University. He has published more than 180 research articles and three books. His research interests include aerospace and low-attitude photogrammetry, image matching, combined block adjustment with multisource datasets, 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 Co-Editor-in-Chief of *The Photogrammetric Record*.

Authorized licensed use limited to: Wuhan University. Downloaded on May 26,2025 at 07:24:52 UTC from IEEE Xplore. Restrictions apply.