Refined Cascade Cost Volume for Multi-View Remote Sensing Image Reconstruction

Wei Zhang, Qiang Li, *Member, IEEE*, Qi Wang, *Senior Member, IEEE*

***Abstract*—Research on remote sensing multi-view stereo has significantly advanced the development of large-scale 3D urban reconstruction. However, existing frameworks encounter challenges with blurred edge details when processing aerial image, which impedes the accuracy of depth estimation. To address these limitations, we propose RC-MVS, the deep estimation network specifically tailored for remote sensing multi-view stereo tasks. This network aims to enhance the geometric details within the view space while effectively reducing match noise, achieving high-precision depth estimation. Specifically, we introduce a refined cascade framework that integrates geometric details with semantic information, ensuring both global structural consistency and local feature expressiveness. During the feature extraction phase, we redesign the feature space construction process and introduce a denoising feature pyramid module. This module reduces feature inconsistency and employs multiple denoising strategies to purify feature representations, thereby enhancing the accuracy of the matching process. Furthermore, to achieve progressive optimization of the depth range, we propose a progressive cross-layer fusion module. This module progressively fuses low-resolution cost volumes, reducing domain shifts between different data dimensions, thereby enhancing the understanding of fine structures within the depth map and the broader context. Experimental results show that the RC-MVS model performs exceptionally well on the LuoJia-MVS and WHU datasets, achieving superior quantitative and qualitative performance.**

***Index Terms*—Multi-view stereo, 3D reconstruction, dense image matching.**

I. Introduction

OVER the past decades, the stereo imaging capabilities of high-resolution optical remote sensing satellites have significantly improved the 3D reconstruction of Earth's surface [1], [2], becoming essential in environmental monitoring [3], [4], disaster assessment [5], and urban planning [6]. Remote sensing data based on Multi-View Stereo (MVS) technology has gained significant attention because of its ability to generate large-scale, high-resolution 3D models of urban areas. However, existing methods for large-scale 3D reconstruction primarily rely on traditional techniques, such as stereo photogrammetry [2] and the marching cubes algorithm [7] for surface reconstruction. These methods encounter challenges when dealing with occlusions and variations in illumination across different viewpoints, which lead to matching errors and necessitate extensive manual labor for post-processing. Consequently, there is an urgent need to explore MVS tasks within large-scale remote sensing scenarios to enhance the robustness and accuracy of the 3D reconstruction process.

Building upon advancements in stereo matching networks [8], learning-based MVS methods have achieved remarkable progress in depth estimation of multi-view image [9]. These methods construct a 3D cost volume by leveraging differentiable homography homography to encode camera parameters and utilize 3D convolutional neural network (CNN) for regularization. This design enables end-to-end inference of depth maps directly from reference images. Moreover, the introduction of a coarse-to-fine regularization framework allows the decomposition of the cost volume into a series of cascading stages, wherein the depth (or disparity) range is progressively refined at each stage based on predictions from the previous step [10]. This feature pyramid-based approach not only preserves fine-scale geometric and contextual details but also enables high-resolution depth map reconstruction, emerging as the dominant paradigm for close-range reconstruction with widespread applications [11], [12].

Although progress is made on close-range objects, general MVS methods encounter significant challenges in large-scale remote sensing scenarios. Recent methods often adapt the architecture by incorporating deformable convolutions, matrix decomposition, and similar enhancements to improve their suitability for aerial image [13]–[16]. However, these adaptations frequently overlook intrinsic limitations of the cascade paradigm—particularly in the construction of the cascade cost volumes. Occlusions between buildings and uneven brightness introduce inconsistencies during the coarse-to-fine feature sampling and aggregation stages, inevitably injecting matching noise into the 3D feature volume generated by homography warping. Moreover, existing cascades architecture initialize each stage with the regularized output of the previous one, neglecting potential misalignment between layers and thereby undermining depth estimation accuracy. These limitations substantially constrain the applicability of general MVS in the remote sensing and highlight the urgent need for further refinements to address occlusions and illumination variations, thereby bolstering model robustness and adaptability.

To address the aforementioned issues in remote sensing MVS tasks, we propose a novel refined cascade architecture. This architecture progressively refines the depth estimation results while performing adaptive feature aggregation of details and semantic information, ensuring both global structural

This work was supported by the National Natural Science Foundation of China under Grant 62471394, 62301385, and U21B2041.

Wei Zhang is with the School of Computer Science, and with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P. R. China. (e-mail: zhangwei707@mail.nwpu.edu.cn).

Qiang Li, and Qi Wang are with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, P.R. China. (e-mail: liqmges@gmail.com, crabwq@gmail.com) (Corresponding author: Qi Wang, Qiang Li.)

consistency and local feature expressiveness. The architecture primarily consists of two modules: a denoising feature pyramid module and a progressive cross-layer fusion module. To minimize inconsistency in feature sampling across different pyramid layers, the denoising feature pyramid module enhances the robustness of the reconstruction process and improves the accuracy of dense matching. To address the lack of sufficient transition between cascade stages, the progressive cross-layer fusion module progressively fuses coarse cost volumes with low-level geometric details and fine cost volumes with high-level semantics, generating full-resolution depth maps with rich contextual information. In summary, proposed architecture fully leverages multi-view information across different scales, while enabling sufficient feature interaction between scales, thereby improving the overall quality and accuracy of the reconstruction results. The main contributions of this paper are summarized as follows:

- We propose a novel refined cascade architecture to address the edge detail blurring caused by occlusions and uneven brightness in multi-view remote sensing tasks, enabling high-accuracy depth estimation. Comprehensive experiments conducted on the popular benchmark datasets LuoJia-MVS [16] and WHU [17] demonstrate that our RC-MVS achieves outstanding performance.
- We propose a denoising feature pyramid module that redefines the process of feature space construction to mitigate inconsistency in cascaded feature sampling. By applying filter enhancements and attention based fusion across multiple feature layers, it significantly strengthens the robustness of the subsequent dense matching pipeline.
- We propose a progressive cross-layer fusion module to achieve refined transitions between cascade stages. By adaptively aggregating coarse cost volumes with low-level details and fine cost volumes with high-level semantic information within a unified domain, enhancing the representation of both the fine structures and broader context within the depth map.

II. RELATED WORK

A. Close-range General Multi-view Stereo

Existing MVS methods can be broadly categorized into four main approaches: mesh-based, point cloud-based [18], [19], volumetric-based [20]–[23], and depth map-based methods [24]–[26]. Among these, depth map-based methods decompose the complex reconstruction process into per-view estimation and fusion of multi-view, enabling more complete surface reconstruction with enhanced robustness. For instance, Xu et al. [27] leverages multi-scale geometric consistency and adaptive sampling techniques to enhance the accuracy and robustness of depth estimation. Similarly, COLMAP [28] employs manual feature extraction while concurrently estimating pixel-level view selection, depth maps, and surface normals. These estimations are guided by photometric and geometric priors to enable dense and accurate predictions. In summary, these traditional approaches have demonstrated significant and robust results in the field of general scenarios.

Recently, learning-based methods have made significant advancements compared to traditional approaches [29]–[34]. Unlike conventional MVS techniques that rely on handcrafted features, deep learning-based methods leverage convolutional neural networks for end-to-end depth estimation. MVSNet [9] first aggregates depth features and camera parameters to construct a cost volume, which is then processed by a 3D CNN to regress the depth map. Although vanilla MVSNet can achieve pixel-level depth predictions, the dense hypothesis planes and 3D cost volumes consume a significant amount of memory. To reduce memory consumption, several subsequent works have been proposed. R-MVSNet [35] uses a GRU [36] to perform continuous regularization on the cost volume, though this results in increased runtime. Cas-MVSNet [10], CVP-MVSNet [37], and UCS-Net [38] leverage cost volume pyramids or cascaded cost volumes to estimate depth maps from coarse-to-fine. Due to its efficiency and high accuracy, the cascade architecture proposed by CasMVSNet [10] has become widely adopted.

B. Large-scale Aerial Multi-view Stereo

Although progress has been made in the reconstruction of close-range objects, differences between aerial and close-range image(such as scale range and scene types) pose new challenges for the application of general models [39], [40]. To address these challenges and enable accurate and efficient MVS depth estimation in aerial contexts, RED-Net [17] pioneered the development of a large-scale scene MVS reconstruction model. This model demonstrates superior performance, surpassing all traditional MVS methods on benchmark datasets such as WHU. Ada-MVS [41] introduced a novel depth estimation architecture that combines an adaptive multi-view cost aggregation mechanism with an effective regularization process, specifically designed for reconstructing large-scale scenes from multi-view images. HDC-MVSNet [16] presents a hierarchical deformable cascaded MVS network for depth estimation from aerial images, which simultaneously executes high-resolution multi-scale feature extraction and constructs hierarchical cost volume modules. CSC-MVS [42] incorporates a non-negative matrix factorization branch and a depth spectral decomposition branch, designed to produce local and global semantic guidance, respectively. Additionally, an uncertainty-aware multi-task optimization approach is proposed to adaptively integrate matching and semantic metrics. SDL-MVS [15] re-examines deformable learning approaches in the multi-view stereo task, introducing a new paradigm based on viewpoint space and depth deformable learning.

Existing remote sensing methods have made some progress by improving conventional methods through techniques such as deformable convolutions and matrix decomposition. However, these methods often overlook inherent limitations of the existing paradigm, such as the propagation of noise within the feature extraction and depth transitions structures in current cascade architectures. Therefore, our goal is to optimize the existing cascade architecture to enhance the robustness and applicability of general MVS models in remote sensing tasks.

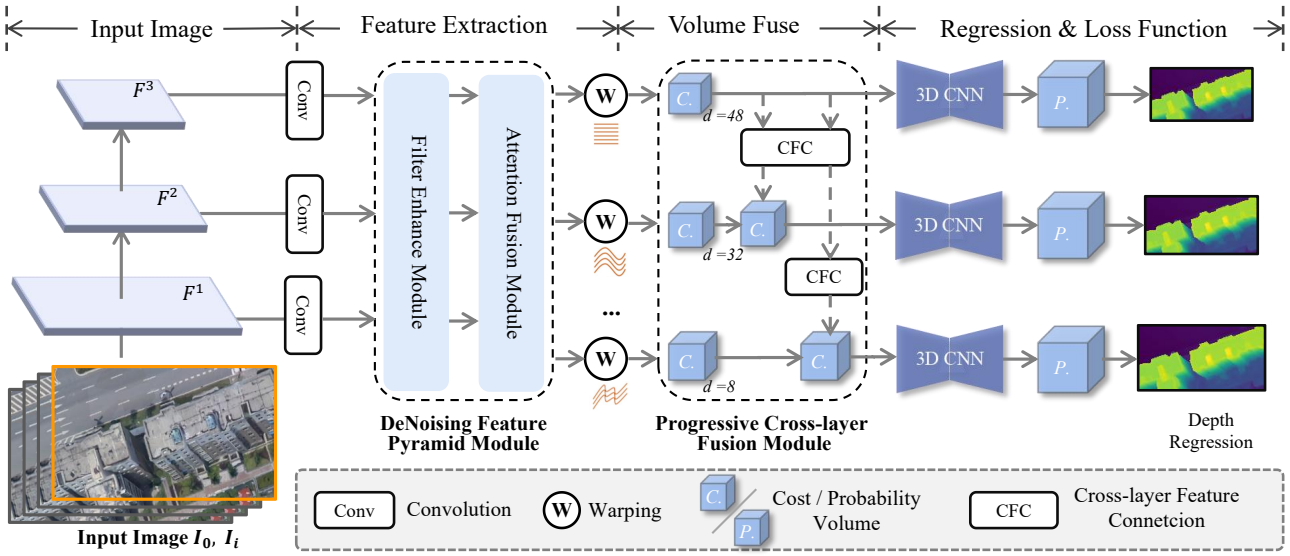

Fig. 1. Overview of the RC-MVS architecture. The proposed Denoising Feature Pyramid Module consists of two key submodules: the Filter Enhance Module and the Attention Fusion Module. The proposed Progressive Cross-Layer Fusion Module consists of multiple cross-layer feature connection operations.

III. Proposed Method

This section presents a comprehensive description of the RC-MVS and illustrates its pipeline in Fig.1. Similar to MVS-Net [9], [43], the key operation involves constructing a plane-sweep volume on the reference image and computing the dense matching costs between the source image and the reference images. Given the reference image $I_0 \in \mathcal{R}^{3\times H\times W}$ and corresponding $n-1$ source images $\{I_i \in \mathcal{R}^{3\times H\times W}\}_{i=1}^{N-1}$, along with the corresponding camera extrinsics $\{[R_{0,i}|t_{0,i}]\}_{i=1}^{N-1}$ and intrinsics $\{K_i\}_{i=1}^{N-1}$, the proposed RC-MVS is designed to predict a pixel-level depth map $d \in \mathcal{R}^{H\times W}$ for the reference image. Following popular practices [10], the overall architecture of RC-MVS is implemented using a cascaded cost volume. However, we have developed a novel refined cascade architecture, and the specific modifications will be detailed below. This section includes three parts: the overall architecture of RC-MVS, the denoising feature pyramid module, and the progressive cross-layer fusion module.

A. RC-MVS Framework

The RC-MVS architecture consists of several key steps: multi-scale feature extraction, cascaded cost volume generation, volume regularization, and depth prediction.

Feature Pyramid Extraction. Followed existing deep learning-based MVS methods [10], we use a weight-sharing feature extractor to extract depth features from the source and reference images for subsequent dense matching. To facilitate the cascaded matching process, the feature extractor adopts a multi-scale Feature Pyramid Network (FPN) [44], generating feature maps in varying spatial resolutions, corresponding to 1/16, 1/4, 1 of the input image resolution. The feature map with the highest spatial resolution is referred to as the first-stage feature map, while the coarsest-resolution feature map is utilized at the final stage of processing.

However, due to the occlusions and illumination variations in remote sensing imagery, feature sampling across multi-view images becomes inconsistent during the progress of FPN multi-scale aggregation, introducing matching noise into the feature volume constructed by homography warping. To address these issues, we propose the Denoising Feature Pyramid Module (DFPM), which uses filtering operations for spatial alignment and feature restoration. DFPM guides the original features toward structurally clear and semantically stable regions while mitigating channel-shift induced information loss and noise. As shown in Fig. 1, the DFPM process begins with multi-scale feature extraction from all input images. The extracted features are then enhanced and resampled through a dual-path filtering process. Finally, the geometric and semantic information are adaptively fused, while residuals are used to reintroduce low-level details, yielding the final output channels.

$$\mathbf{F}_{\text{out}}^{(s)} = \text{DFPM}\left(\left\{I_i^{(s)}\right\}_{i=0}^{N}\right), s \in \{1, 2, \ldots, S\} \quad (1)$$

where the $F_{\text{out}}^{(s)}$ denotes the refined feature map at scale s, and S is set to 3. A more detailed description of this module is provided in Sec. III-B.

Cascade Cost Volume Construction. Using the feature outputs from DFPM, a 3D feature volume is constructed to assess the similarity between corresponding image patches and achieve the dense matching results. The core operation involves distorting the depth features from the source view and compute the corresponding position in the reference image using a plane sweep algorithm.

$$H_i(p') = K_i \cdot [R_i \cdot ({K_0}^{-1}\text{p}d) + t_i], \quad (2)$$

where $H_i(p')$ refers to the homography between the feature map of the source image i and the reference image at depth d, p represents a pixel . The sets K_0, K_i are the intrinsics of the

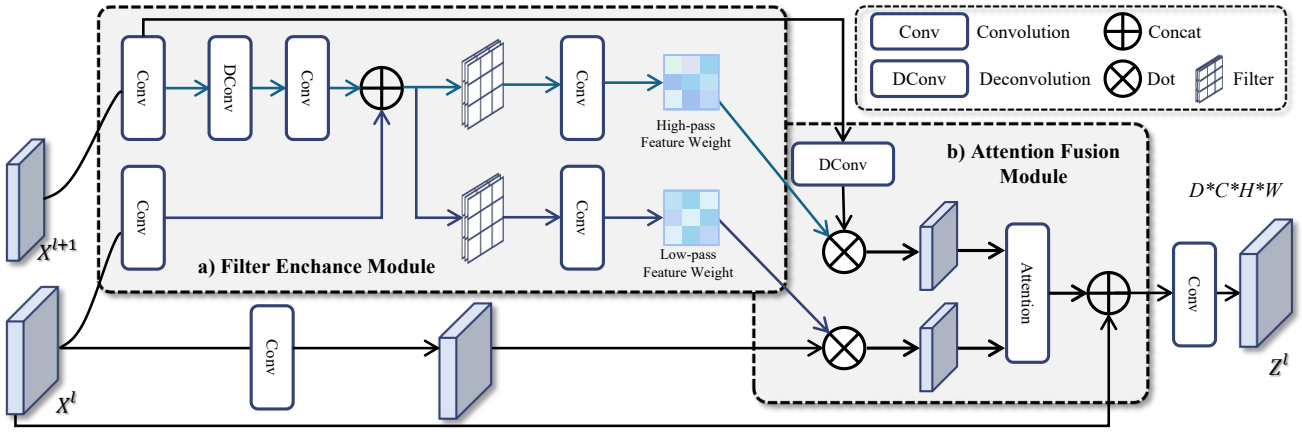

Fig. 2. Overview of proposed Denoising Feature Pyramid Module. It first performs dual-path filter reconstruction and enhancement in the Filter Enhance Module, followed by dual-path fusion in the Attention Fusion Module, ultimately producing the final fused result.

reference and i-th source cameras. And the R_i, t_i represent the rotation matrices, and translation vectors of the source image i and the reference camera, respectively.

The feature volumes obtained at each layer are then aggregated using an adaptive cost metric to form a cost volume. Specifically, the reference volume is first compared with the source volume to compute the difference, and a set of weights is derived using two layers of 3D convolutions. The computed weights are then applied to the differences to obtain the final adaptive volume. After aggregating the volumes from all viewpoints, the final cost volume $C^{(s)}$ is obtained.

$$\begin{aligned} \overline{V}^{(s)} &= \left(V_{\text{ref}}^{(s)} - V_{\text{warped}}^{(s)}\right)^2, \\ C^{(s)} &= \left(1 + \text{Conv}\left(\overline{V}^{(s)}\right)\right) \cdot \overline{V}^{(s)}, \end{aligned} \tag{3}$$

where V_{warped} and V_{ref} are the warped source and reference feature volumes, $\overline{V}$ denotes the weights, and s represents the scale of feature pyramid network.

In the original cascaded cost volume formulation, the first stage spans the entire depth (or disparity) range of the input scene, resulting in a large initial hypothesis interval and the lowest spatial resolution. In subsequent stages (for instance, the second or third), the depth of the hypothesis plane at the kth stage is determined by combining the estimate from the $k+1$th stage with the depth residual computed at the current stage. Accordingly, both the depth range and the hypothesis interval between adjacent planes are refined. As the hypothesis range narrows significantly at each stage while still covering the entire output range, this approach effectively reduces the total number of hypothesis planes. Consequently, this cascaded structure has been widely adopted in MVS related tasks.

However, each stage of feature pyramid extraction adopts different feature dimensions across layers, causing domain offsets that simple stage transition strategies fail to resolve effectively. Furthermore, the hypothesis planes generated in earlier stages are already regularized, leading to inconsistencies across feature spaces at different depths. As a result, existing methods that directly fuse these heterogeneous features tend to introduce considerable noise. Furthermore, the extended depth process causes the loss of important information in the cost volume. To address these problems, we propose a progressive cross-layer fusion module that progressively integrates multiple low-resolution dense cost volumes. This approach reduces domain shift between different data dimensions to facilitate accurate initial disparity estimation. Specifically, we first construct low-resolution cost volumes at each scale, and then use residual and alignment operations to fuse cost volumes across different layers. This entire cost volume fusion process can be seamlessly incorporated into the existing cascaded architecture. A more detailed description of the structure can be found in Sec. III-C.

Regularization and Depth Estimation. For the cost volumes obtained at each stage, regularization is performed using a 3D version of U-Net, which transforms $C_k \in \mathbb{R}^{H\times W\times D}$ into the probability of the hypothesis depth plane $P_k \in \mathbb{R}^{H\times W\times D}$. Due to inherent ill-posed regions such as occlusions and uneven brightness, pixel-level cost calculations are often ambiguous. The 3D U-Net, with its ability to aggregate neighboring information through a large receptive field, effectively consolidates contextual information, reduces matching errors, and preserves spatial smoothness. Furthermore, by employing depth regression, the final depth for each pixel is computed based on its depth probability distribution. Finally, the overall loss is computed as a weighted sum of the losses from all stages, where each individual loss is defined by the L1 norm of the difference between the reference and the estimated depth.

B. Denoising Feature Pyramid Module

As a fundamental component of cascaded architectures, feature pyramids have been widely adopted [10]. However, in remote sensing scenarios the fusion methods used within these pyramids introduce two detrimental issues for dense prediction: matching inconsistencies and boundary shifts. Standard fusion operations are inadequate to correct these mismatched features and can even exacerbate the problem by up-sampling a single inconsistent feature into multiple erroneous pixels. This interpolation tends to over-smooth the output, resulting in boundary displacement. Furthermore, the detailed boundary information in low-level features is not fully utilized.

Inspired by the pyramid adjacency in FPN [44], full-scale connections in HDC-MVSNet [16], and frequency-aware feature fusion in FreqFusion [45], we propose a denoising feature pyramid module and illustrates its pipeline in Fig. 2. This module consists of two components: the filter Enhance module and the attention fusion module. The former uses dual-path filters to reconstruct semantic and geometric feature separately, while the latter further fuses shallow encoded features with the final reconstructed features to obtain the final feature output.

$$\begin{aligned} X^L, X^H &= \text{FilterEnhance}(X^l, X^{l+1}), \\ Z^l &= \text{AttentionFusion}(X^l, X^{l+1}, X^H, X^L), \end{aligned} \tag{4}$$

where X_i and X_{i+1} represent the low-level features and high-level fused features generated by the backbone. X_h and X_l denote the intermediate features, Z signifies the final output.

Filter Enhance Module. The input X_i and X_{i+1} represent the low-level features and high-level fused features generated by the backbone, respectively. First, a simple upsampling operation is applied to X_{i+1} using deconvolution layer to maintain scale consistency. Then, convolution operation is used to compress and fuse X_i and X_{i+1} initially, yielding X_f. Next, X_f is used as input to predict the spatial variations of the low-pass filter, with a softmax constraint applied to the kernel, resulting in the final smoothed low-pass filtered features. Finally, content-aware feature reorganization is performed, where the low-level features are restructured based on the low-pass filtering to obtain the final low-pass reconstructed features. Similarly, the initially fused X_f is used as input to predict the spatial variations of the high-pass filter, which consists of a 3×3 convolution layer followed by a softmax layer and a filter inversion operation, yielding the final high-pass reconstructed features.

Attention Fusion Module. The results of the dual-path filter enhancement obtained in the Filter Enhance Module are subsequently combined with their respective shallow encoded features, followed by attention enhancement, to yield the final fused result. Specifically, the dual-branch features are initially fused and then processed through a feature construction module to generate weights. These weights are used to perform an inner product with the high-level features and are further integrated with the low-level features via a long skip connection. Finally, a 1×1 convolution is applied for channel compression to ensure consistency with the features X_i.

C. Progressive Cross-layer Fusion Module

Existing MVS methods have demonstrated the importance of utilizing multi-scale cost volumes. A common cascade architecture directly assumes that the depth of the hypothetical plane at the k-th stage is equal to the previous estimate from the $(k+1)$-th stage, plus the depth residual at the current stage. However, in a cascade structure, each decoder layer of the feature pyramid extraction has different feature dimensions. Due to the differences in feature layers, a simple depth residual fusion approach may lead to inconsistencies in the feature space. Furthermore, as the depth process extends (from the first stage to the third stage), crucial cost volume information may be lost.

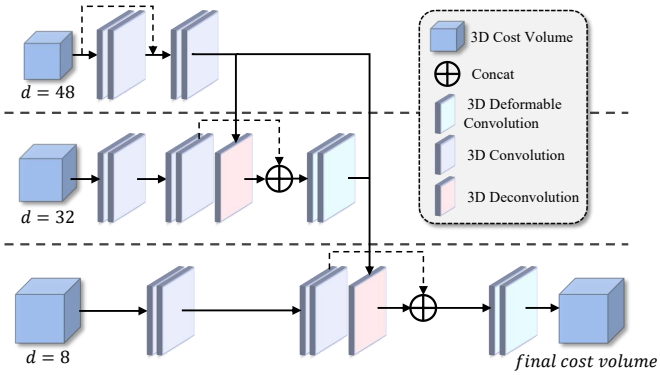

Fig. 3. Illustration of proposed Progressive Cross-layer Fusion Module. It facilitates feature transfer and alignment across multiple stages using 3D convolutions and 3D deformable convolution, enabling progressive fusion of low-resolution and high-resolution cost volumes.

To mitigate this problem, we propose a progressive cross-layer fusion module that integrates multiple low-resolution dense cost volumes. This strategy aims to reduce domain shifts across varying data dimensions, thereby facilitating more accurate initial disparity estimation. The detailed structure of this module is depicted in Fig. 3. Rather than simply using low-resolution cost volumes as depth residuals, we argue that cost volumes at different scales provide multi-scale receptive fields, guiding the network to attend to image regions of varying sizes. By fusing these cost volumes, the network can simultaneously capture both global and structural information, resulting in a more accurate initial disparity map than would be achieved with sparse, high-resolution cost volumes alone. Specifically, we first construct low-resolution cost volumes at each scale, and then design a fusion mechanism to incorporate them into the existing coarse-to-fine cascade process.

Building upon the implementations of HDC-MVSNet [16], MSMD-Net [46], and CFNet [47], we improve the cascade architecture to fuse low-resolution cost volumes. Specifically, we first reduce the depth hypothesis range from 48 to 32 in the first stage using 3D max pooling with a kernel and stride of 2, and then apply 3D transposed convolution to upsample the low-resolution cost volume from 1/4 to 1/2 of the input image resolution, matching the second stage. Next, the upsampled cost volume is concatenated with the cost volume of the next stage along the feature dimension. We apply similar operations to the second-stage cost volume, using pooling and transposed convolution to align it with the depth hypothesis range and resolution of the third-stage cost volume. Additionally, the first-stage cost volume undergoes multiple layers of pooling and transposed convolution to align with the third stage, enabling long-range residual connections. After all the concatenations, an additional 3D convolution layer is applied to reduce the feature channels to a fixed size, ensuring consistency with the original cost volume dimensions.

$$C_{\text{fused}}^{(s)} = \text{Conv3D}\left(\text{Concat}\left(\text{UpSample}(C^{(s-1)}), C^{(s)}\right)\right), \tag{5}$$

where UpSample denotes the upsampling operation, Concat refers to the concatenation, Conv3D represents the alignment

operation, and s signifies different scales. In the third stage, fusion occurs not only with the previous stage but also with the first stage.

IV. Experiments

A. Datasets

Three aerial multi-view benchmarks is used in our experiments: WHU-MVS [17] covers 6.7 × 2.2 km at 0.1 m ground resolution, producing 1,776 high-resolution (5376 × 5376 px) images via a five-camera oblique UAV and Smart3D, all RGB and corresponding depth maps cropped into 768 × 384 px tiles (80 five-view pairs, 400 sub-images). LuoJia-MVS [16] comprises 7,972 five-view units from Baiyun, Guiyang, and Guizhou in China, covering cultivated areas, forests, urban/rural settlements, industrial zones, and idle land; each view provides a 768 × 384 px 10 cm-resolution RGB image, a pixel-level depth map, and calibrated camera parameters. WHU-OMVS [13] extends WHU-MVS into a city-scale oblique stereo over the same 6.7 × 2.2 km urban area at 550 m altitude at 10 cm resolution, providing 768 × 384 px PNG renders, float EXR depth maps, and per-image TXT calibration files. All datasets include precisely annotated camera intrinsics and extrinsics.

B. Evaluation Metrics

In the MVS task, we evaluate the quality of the depth estimation using various types of metrics: Mean Absolute Error (MAE), percentage of pixels with depth error <0.6 meters, and percentage of pixels with depth error within a <3-interval threshold.

- The mean absolute error (MAE) quantifies the discrepancy between the estimated depth map and the ground truth. It is computed as the average L1 distance between the estimated and ground truth depth values, restricted to 100 depth intervals to mitigate the influence of extreme outliers.
- The <0.6 m metric assesses the accuracy of the estimated depth map by computing the proportion of pixels whose L1 error falls below 0.6m.
- The <3-interval metric quantifies the accuracy of the estimated depth map by computing the proportion of pixels whose L1 error remains within three depth intervals.

C. Implementation Details

We implement our method based on PyTorch 1.7.1 and conduct training and testing on a server equipped with four NVIDIA RTX 3090 GPU. We use the AdamW optimizer and employ a one-cycle learning rate scheduling strategy, setting the maximum learning rate to lr_{max} = 0.001. For all experiments, we set the batch size to 1. During both training and testing, the network uses five-view images as input.

For all cascade-based networks, we follow the configuration of existing multi-stage architectures. From the first to the third stage, the number of depth hypotheses $D_1 - D_3$ are 48, 32, and 8, the depth intervals $I_1 - I_3$ are 4, 2, and 1, and the multi-scale feature map resolutions are set to 1/16, 1/4, 1 of the input image size. During training, the weights of the depth estimation loss functions at the three stages are assigned as 0.5, 1.0, 2.0. In contrast, for a fair comparison, the non-cascade network uses 192 depth hypotheses, a depth interval of 1, and the original reference image resolution for its feature maps. All other parameters are consistent with those reported in the respective original publications.

TABLE I
THE QUANTITATIVE RESULT OF DEPTH ESTIMATION ON THE WHU. SOME RESULTS ARE OBTAINED FROM HDC-MVSNET [16]. BOLD REPRESENTS THE BEST WHILE UNDERLINED REPRESENTS THE SECOND-BEST.

Method	MAE↓	<3-interval↑	<0.6m↑
COLMAP [48]	0.154	0.949	0.956
SURE [7]	0.224	0.920	0.936
MVSNet [9]	0.160	0.955	0.958
R-MVSNet [35]	0.173	0.938	0.954
RED-Net [17]	0.104	0.979	0.981
PatchmatchNet [49]	0.160	0.950	0.969
Fast-MVSNet [50]	0.157	0.956	0.961
Cas-MVSNet [10]	0.095	0.978	0.978
Ada-MVS [41]	0.102	0.964	0.980
HDC-MVSNet [16]	<u>0.087</u>	0.980	0.981
AggrMVS [13]	0.102	<u>0.980</u>	<u>0.986</u>
RC-MVS (Ours)	**0.084**	**0.984**	**0.987**

D. Benchmark Performance

Comparison Results on WHU dataset. In the experiment, we compare the performance of our RC-MVS framework with other methods on the WHU dataset, including COLMAP [48], SURE [7], MVSNet [9], PatchmatchNet [49], Fast-MVSNet [50], R-MVSNet [35], RED-Net [17], Cas-MVSNet [10], Ada-MVS [41], AggrMVS [13] and HDC-MVSNet [16]. The comparison of depth estimation performance between our RC-MVS framework and other methods is presented in Table I. As can be observed, our proposed refined coarse-to-fine framework achieves the best depth estimation performance, surpassing all other methods with the lowest MAE error and the highest accuracy.

Specifically, in terms of the MAE error metric, the proposed RC-MVS yields a depth estimation error of 0.084, which is 3.8% lower than that of the HDC-MVSNet method. In terms of accuracy evaluation, the proposed method achieves 0.987 and 0.984 for the accuracy metrics in the <0.6m and <3-interval, respectively, showing an stable improvement compared to the classic method HDC-MVSNet. The WHU dataset contains a large number of buildings, leading to common issues such as occlusion and Uneven brightness. Due to insufficient views in occluded areas, the available consistent information between views is not sufficient to support accurate depth estimation. The refined cascade framework proposed addresses this challenge by reducing inconsistencies in feature fusion at different levels through denoising-based feature extraction, enhancing the robustness of the reconstruction process and effectively reducing matching errors. Additionally, the design of progressive cross-layer fusion enables the gradual integration of coarse cost volumes with fine cost volumes that have higher-level semantics, which significantly improves the overall quality and accuracy of depth estimation.

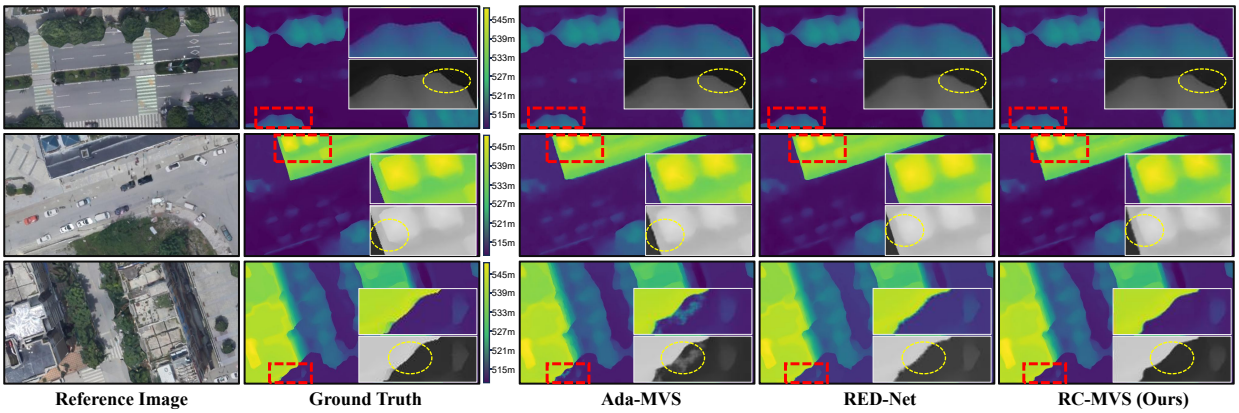

Fig. 4. The qualitative result of depth estimation on the WHU dataset. The red dashed box outlines the region selected for detailed comparison. The white-bordered box provides a magnified view of this region (in both color and grayscale), where the yellow dashed box further emphasizes the key differences.

TABLE II
THE COMPARISON OF COMPUTATIONAL COMPLEXITY. "GPU MEM." INDICATES GPU MEMORY USAGE, "RUN-TIMES" REFERS TO INFERENCE TIME, AND "I/O. SIZE" REPRESENTS THE INPUT/OUTPUT DIMENSIONS.

Model	Input data		Running cost		MAE↓
	View	I/O. size	GPU Mem.↓	Run-time↓	
Ada-MVS [41]	5	384 × 768	2927 MB	9.05 min	0.102
AggrMVS [13]	5	384 × 768	3014 MB	7.14 min	0.102
RC-MVS (Ours)	5	384 × 768	**2742 MB**	**6.06 min**	**0.084**

TABLE III
THE QUANTITATIVE RESULT OF DEPTH ESTIMATION ON THE LUOJIA-MVS. SOME RESULTS ARE OBTAINED FROM HDC-MVSNET [16].

Method	MAE↓	<3-interval↑	<0.6m↑
PatchmatchNet [49]	0.283	0.841	0.904
Fast-MVSNet [50]	0.357	0.749	0.846
MVSNet [9]	0.270	0.818	0.912
R-MVSNet [35]	0.259	0.867	0.923
RED-Net [17]	0.156	0.905	0.949
Cas-MVSNet [10]	0.141	0.954	0.979
HDC-MVSNet [16]	<u>0.121</u>	<u>0.966</u>	<u>0.983</u>
RC-MVS (Ours)	**0.120**	**0.970**	**0.985**

To provide a more intuitive demonstration of the performance advantages of our RC-MVS, we further conduct a qualitative analysis of the predicted depth maps on the WHU. Fig. 4 presents the visual results of RED-Net, Ada-MVS, and the proposed method, alongside the ground truth. For the qualitative analysis, various typical land cover types are selected, including low-rise building areas, high-rise building areas, highways, and unused land areas. Due to the characteristics of the WHU dataset, which includes densely built-up areas and sparse natural geographic instances, occlusion issues are more prominent. Additionally, the high building density makes the acquired multi-view image data highly susceptible to Uneven brightness. However, the proposed depth estimation network demonstrates robust performance in both densely built-up areas and regions where buildings are mixed with irregular textures. Overall, the proposed method achieves impressive depth estimation performance, demonstrating accurate predictions even in dense high-rise building areas.

In addition, we conduct a comparative analysis of the inference efficiency of RC-MVS in the WHU dataset under the same input and output setting, with results summarized in Table II. Although the proposed method introduces additional computation through the newly designed modules, it essentially optimizes the existing architecture. Consequently, the overall runtime remains within an acceptable range, while achieving a better trade-off between performance and accuracy compared to some existing methods.

Comparison Results on Luojia-MVS dataset. Table III presents the depth estimation errors and accuracy performance of the proposed RC-MVS architecture on the LuoJia-MVS dataset. From the experimental results, it can be concluded that our RC-MVS method achieves superior depth estimation performance, demonstrating advantages in both MAE error and accuracy. Specifically, in terms of MAE metric, the proposed RC-MVS architecture achieves an estimation error of 0.120, which is 14% lower than that of Cas-MVSNet. In the accuracy metric, the proposed method achieved 0.970 and 0.985 performance for the <0.6m and <3-interval, respectively, compared to RED-Net, the classic work in remote sensing multi-view stereo, showing an improvement of at least 4%. This also demonstrates the effectiveness of the refined cascade network in improving depth estimation accuracy for multi-view stereo. Since the LuoJia-MVS dataset often includes rural and natural scene types that lack regular textures, the progressive cascade approach in this work better adapts to depth estimation in multiple complex scenarios. The denoising feature extraction not only mitigates the effects of inconsistent matching but also enables adaptive fusion of deep semantic information and shallow geometric features, demonstrating significant advantages in model detail reconstruction.

To provide a more intuitive demonstration of the performance advantages of our RC-MVS method, we further conduct a qualitative analysis of the predicted depth maps obtained on the LuoJia-MVS dataset. Fig. 5 shows the visual results of RED-Net, Ada-MVS, and the proposed method, alongside the ground truth. To highlight the effectiveness of the proposed RC-MVS in various scenes, we focused on instances such as fields, grasslands, and courtyards for depth estimation.

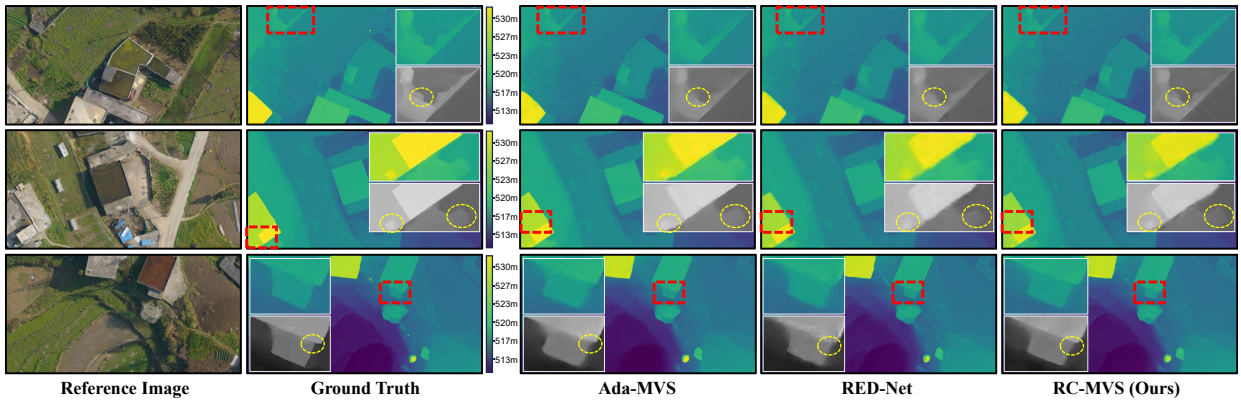

Fig. 5. The qualitative result of depth estimation on the Luojia-MVS dataset. The white box shows the magnified detail area.

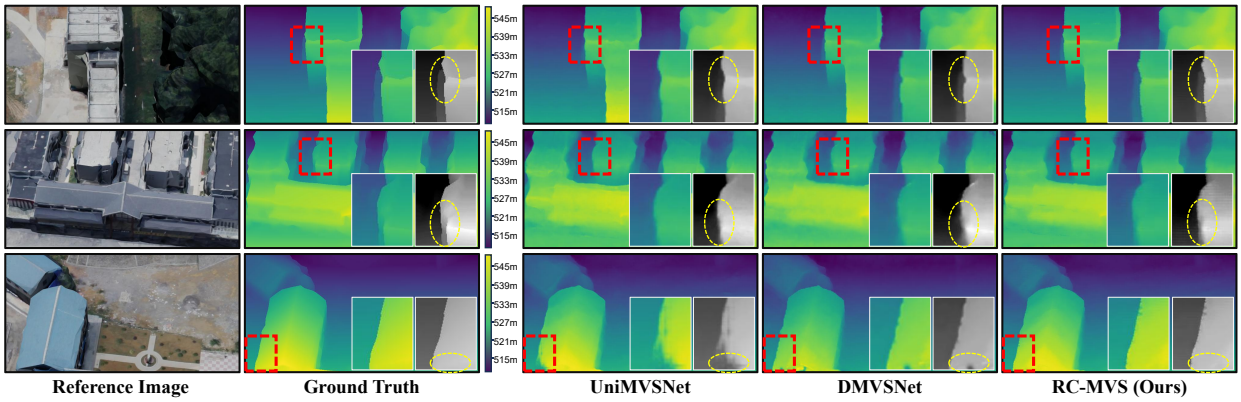

Fig. 6. The qualitative result of depth estimation on the WHU-OMVS dataset. The white box shows the magnified detail area.

TABLE IV
THE QUANTITATIVE RESULT OF DEPTH ESTIMATION ON THE WHU-OMVS DATASET.

Method	MAE↓	<3-interval↑	<0.6m↑
Cas-MVSNet [10]	0.267	0.902	0.911
UniMVSNet [11]	0.240	0.880	0.881
DMVSNet [51]	0.240	0.911	0.910
RC-MVS (Ours)	**0.210**	**0.939**	**0.934**

Compared to the depth estimation results of RED-Net and Cas-MVSNet, the proposed RC-MVS method yields more accurate depth predictions in regions prone to noise. The other two methods suffer from blurry predictions and lack precision in detail reconstruction. This results further demonstrates the advantages of the proposed cascade architecture in noise suppression and detail reserved.

Comparison Results on WHU-OMVS dataset. To further evaluate the generalization ability of the proposed model, we conducted cross-dataset testing by applying the model trained on the WHU dataset directly to the WHU-OMVS dataset. As shown in Table IV, the quantitative results demonstrate that our method maintains strong performance and good adaptability even without additional training on the new dataset. Furthermore, we provide a visual comparison in Fig. 6, which further highlights the accuracy of our model's predictions.

E. Ablation Study

To validate the effectiveness of the proposed denoising feature pyramid module and progressive cross-layer fusion module, we conduct a series of ablation experiments on two datasets to assess the impact of each component on MVS task. In the following, the term 'baseline' refers to the basic model, Cas-MVSNet, which employs an FPN as the feature extraction module and uses relatively independent coarse-to-fine branches in a cascading architecture. The **last row** (in bold) of each table represents the proposed model, RC-MVS, which utilizes a denoise feature pyramid as the feature extraction module and a cross-layer fusion cascading structure as the backbone network.

Effectiveness of the denoising feature pyramid module. By aligning and reconstructing features during extraction, the proposed method effectively reduces the impact of unavoidable feature noise, including occlusion-induced noise and pixel noise caused by uneven brightness. This results in significant improvements in multi-view stereo depth estimation performance. Table V demonstrates the effectiveness of the proposed denoising feature pyramid mechanism. Compared to the baseline model, the method proposed shows notable improvements in both error and accuracy after utilizing denoising feature extraction. Specifically, in the WHU dataset, the MAE decreases from 0.096 to 0.084, and the percentages of pixels within the <3-interval and <0.6m improve from 0.978 and 0.979

TABLE V
EFFECTIVENESS OF THE DENOISING FEATURE PYRAMID MODULE. THE FPN AND U-NET MODULES ARE DERIVED FROM CAS-MVSNET, WHILE FULL-FPN IS TAKEN FROM HDC-MVSNET [16].

Module	MAE↓	<3-interval↑	<0.6m↑
FPN [10]	0.096	0.978	0.979
Unet [10]	0.098	0.972	0.971
Full-FPN [16]	0.091	0.979	0.980
Filter Enhance Module	0.087	0.982	0.984
Attention Fusion Module	0.086	0.980	0.983
DFPM (Ours)	**0.084**	**0.984**	**0.987**

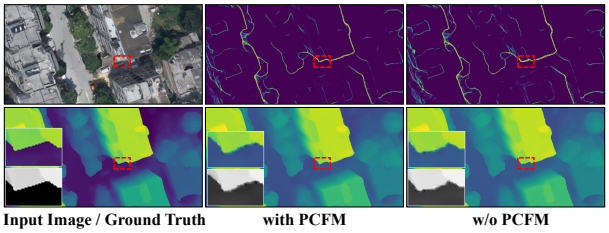

Fig. 7. Visual Comparison of the DFPM Ablation. The upper part shows the confidence map, indicating that the edge regions correspond to areas of lower confidence. By contrasting the predicted depth and confidence maps before and after integrating the proposed module, we observe that DFPM delivers significantly sharper and more accurate edge delineation in occluded regions.

TABLE VI
EFFECTIVENESS OF THE PROGRESSIVE CROSS-LAYER FUSION MODULE. 'ADJACENT-LAYER' INDICATES FUSION BETWEEN NEIGHBORING LAYERS ONLY, WHILE 'CROSS-LAYER' DENOTES CONNECTIONS SPANNING NON-ADJACENT LAYERS.

Module	MAE↓	<3-interval↑	<0.6m↑
Baseline	0.096	0.978	0.979
Adjacent-layer	0.090	0.980	0.983
Cross-layer	0.088	0.981	0.982
(PCFM) Ours	**0.084**	**0.984**	**0.987**

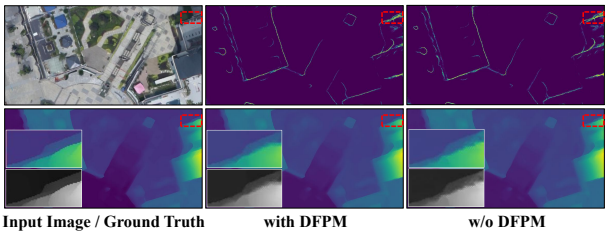

Fig. 8. Visual Comparison of the PCFM Ablation. By contrasting the predicted depth and confidence maps before and after integrating the PCFM module, we observe that the proposed module delivers markedly sharper and more precise edge delineation in regions affected by illumination changes.

to 0.984 and 0.987, respectively. These performance gains are attributed to the superiority of the denoising feature pyramid, which reduces feature inconsistency by redesigning the feature space construction process. Additionally, the use of diversified denoising strategies purifies feature representations, effectively mitigating the negative impact of noise on reconstruction results. This significantly enhances the accuracy of the matching process and greatly improves the overall performance. This further underscores the importance of feature denoising during the encoder stage for depth optimization from coarse to fine.

Additionally, we conducted isolated evaluations of each submodule within the proposed modules. The ablation results confirm that every individual component contributes to performance gains, and their combined integration achieves the highest overall accuracy. Moreover, as shown in Fig. 7, we present visual comparisons of the outputs before and after module incorporation. These visualizations clearly demonstrate that—especially in regions affected by illumination variations—the enhanced framework produces significantly improved predictions.

Effectiveness of the progressive cross-layer fusion module. Multi-stage MVS depth estimation achieves progressive optimization from coarse to fine depth through inter-stage cascading, where the depth range can be initialized based on the results of the previous stage. The initial depth range, being broader, is suitable for depth estimation of low-resolution, low-frequency features. As the stages progress, the reduction in depth range facilitates the fine learning of high-frequency features at higher resolutions. However, in the cascade structure, each decoder layer of the feature pyramid extraction has different feature dimensions. Simple depth residual fusion methods may lead to inconsistencies in the feature space, which can introduce noise. By introducing a progressive fusion structure, the depth range compression can be adaptively adjusted. Therefore, building on the baseline model, we use progressively deformable depth range assumptions in the depth hypothesis, and refine the compression of the depth range assumption based on the predicted depth map from the previous stage. This significantly improves the depth estimation in the subsequent stage. With better priors, the current method achieves a performance improvement of 0.084 in MAE error and 0.984 accuracy at the <3-interval, as well as 0.987 accuracy at <0.6m.

We also conduct an independent evaluation of each submodule. As shown in Table VI, both the addition of adjacent-layer and cross-layer connections lead to performance improvements, with the combination of modules achieving the best overall scores. Furthermore, as illustrated in Fig. 8, we provide a visual comparison before and after integrating the modules. The visual results demonstrate that after adding the modules, predictions in occluded areas are significantly improved. These performance gains are attributed to the advantages of the cascade structure, where the introduction of learnable transition strategies effectively reduces domain offset and enhances the representation of fine structures and broader contextual information within the depth maps.

F. Visualization Analysis of Point Cloud Fusion

To further demonstrate the effectiveness of the model in predicting depth maps, we reconstructed point cloud data for large-scale scenes based on the predicted pixel depths. Specifically, RC-MVS first generates estimated depth for aerial images from the dataset. Then, by using the camera's intrinsic and extrinsic parameter matrices, the original image coordinate system is mapped to the world coordinate system. We subsequently fuse the depth map results from the test sets of the LuoJia-MVS and WHU datasets and perform a visual analysis of the resulting point clouds.

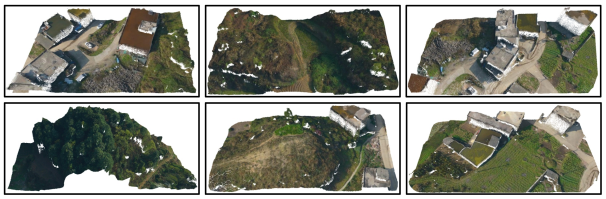

Fig. 9. Local visualization results of point cloud fusion of the proposed RC-MVS on the Luojia-MVS dataset. We selected a variety of regions, including residential areas, farmland, forests, and hills, for visual analysis.

Fig. 10. Local visualization results of point cloud fusion of the proposed RC-MVS on the WHU dataset.

Point cloud results from the Luojia-MVS dataset. Results on Luojia-MVS Dataset. As shown in Fig. 9, we visualize the point cloud reconstruction results for several typical regions in the Luojia-MVS dataset test set, including hills, buildings, ravines, farmland, and forests. Among these, building structures exhibit regular texture patterns, while areas like forests contain large amounts of irregular textures, making reconstruction more challenging. However, the visual results demonstrate that our approach achieves satisfactory point cloud reconstruction, confirming the effectiveness of the proposed architecture in depth estimation. Additionally, it is worth noting that, unlike the WHU dataset, which includes facades capture data, the Luojia-MVS dataset lacks sufficient information on facades, leading to gaps in the reconstruction of many building facades.

Point cloud results from the WHU dataset. As shown in Fig. 10, we select regions from various scenes for visual analysis of point cloud, including trees, roads, buildings, open spaces, and residential areas with a mix of low and high-rise buildings. Through qualitative analysis, we observe that in large-scale reconstructions of areas with regular architectural structures, our proposed MVS reconstruction method achieves superior point cloud fusion and reconstruction results for urban-scale regions. Additionally, we performed local zoom-ins on different scene regions, as shown in Fig. 10. The WHU dataset construction process used multiple cameras positioned at a 40° tilt angle, ensuring that most scenes, including building facades, are captured effectively. As a result, the side facades of these zoomed-in areas are clearly reconstructed. Furthermore, for large areas with irregular textures, such as trees, our proposed multi-view stereo reconstruction method also delivers satisfactory results, strongly demonstrating the superiority of our depth estimation approach.

V. Conclusion

In this paper, we introduce a depth estimation method specifically tailored for MVS tasks in remote sensing, aiming to achieve high-precision depth map estimation for aerial imagery through a refined cascade architecture. Our improved cascade framework integrates geometric details with high-level semantic, ensuring global structural consistency while enriching local features. A denoising pyramid feature module is proposed to minimize feature inconsistencies and purify representations through multiple denoising strategies, significantly enhancing the robustness of subsequent dense matching. In addition, a progressive cross-layer fusion module is introduced, which adaptively aggregates coarse cost volumes rich in low-level details and fine cost volumes carrying high-level semantics within a unified domain, progressively optimizing the fine structures and broader contextual representations during the depth propagation process. Experimental results on the LuoJia-MVS and WHU demonstrate that our MVSNet method outperforms recent aerial MVS depth estimation approaches. Future work will focus on further minimizing detail loss and enhancing the robustness of model in aerial depth estimation.

References

[1] L. Zhao, Y. Men, Y. Zhu, H. Wang, and C. Men, "A cascade domain clustering algorithm for multiview dsm fusion from urban satellite images," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–21, 2024.

[2] Y. Mao, K. Chen, L. Zhao, W. Chen, D. Tang, W. Liu, Z. Wang, W. Diao, X. Sun, and K. Fu, "Elevation estimation-driven building 3-d reconstruction from single-view remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–18, 2023.

[3] Q. Yuan, H. Shen, T. Li, Z. Li, S. Li, Y. Jiang, H. Xu, W. Tan, Q. Yang, J. Wang *et al.*, "Deep learning in environmental remote sensing: Achievements and challenges," *Remote Sens. Environ.*, vol. 241, p. 111716, 2020.

[4] Y. Liu, Z. Xiong, Y. Yuan, and Q. Wang, "Transcending Pixels: Boosting Saliency Detection via Scene Understanding from Aerial Imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[5] A. Sarkar, T. Chowdhury, R. R. Murphy, A. Gangopadhyay, and M. Rahnemoonfar, "Sam-vqa: Supervised attention-based visual question answering model for post-disaster damage assessment on remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–16, 2023.

[6] J. Chen, Y. Xu, S. Lu, R. Liang, and L. Nan, "3-d instance segmentation of mvs buildings," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2022.

[7] M. Rothermel, K. Wenzel, D. Fritsch, and N. Haala, "Sure: Photogrammetric surface reconstruction from imagery," in *Proc. LC3D Workshop, Berlin*, vol. 8, no. 2, 2012.

[8] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, "End-to-end learning of geometry and context for deep stereo regression," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 66–75.

[9] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Proc. Euro. Conf. on Comput. Vis.*, 2018, pp. 767–783.

[10] X. Gu, Z. Fan, S. Zhu, Z. Dai, F. Tan, and P. Tan, "Cascade cost volume for high-resolution multi-view stereo and stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2495–2504.

[11] R. Peng, R. Wang, Z. Wang, Y. Lai, and R. Wang, "Rethinking depth estimation for multi-view stereo: A unified representation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 8645–8654.

[12] X. Wang, Z. Zhu, G. Huang, F. Qin, Y. Ye, Y. He, X. Chi, and X. Wang, "Mvster: Epipolar transformer for efficient multi-view stereo," in *Proc. Euro. Conf. on Comput. Vis.* Springer, 2022, pp. 573–591.

[13] W. Zhang, Q. Li, Y. Yuan, and Q. Wang, "Visual consistency enhancement for multiview stereo reconstruction in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.

[14] S. Zhang, Z. Wei, W. Xu, L. Zhang, Y. Wang, J. Zhang, and J. Liu, "Edge aware depth inference for large-scale aerial building multi-view stereo," *ISPRS J. Photogramm. Remote Sens.*, vol. 207, pp. 27–42, 2024.

[15] Y.-Q. Mao, H. Bi, L. Xu, K. Chen, Z. Wang, X. Sun, and K. Fu, "Sdl-mvs: View space and depth deformable learning paradigm for multi-view stereo reconstruction in remote sensing," *arXiv preprint arXiv:2405.17140*, 2024.
[16] J. Li, X. Huang, Y. Feng, Z. Ji, S. Zhang, and D. Wen, "A hierarchical deformable deep neural network and an aerial image benchmark dataset for surface multiview stereo reconstruction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
[17] 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 Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6050–6059.
[18] Y. Furukawa and J. Ponce, "Accurate, dense, and robust multiview stereopsis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 8, pp. 1362–1376, 2009.
[19] M. Lhuillier and L. Quan, "A quasi-dense approach to surface reconstruction from uncalibrated images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 3, pp. 418–433, 2005.
[20] S. M. Seitz and C. R. Dyer, "Photorealistic scene reconstruction by voxel coloring," *Int. J. Comput. Vis.*, vol. 35, pp. 151–173, 1999.
[21] K. N. Kutulakos and S. M. Seitz, "A theory of shape by space carving," *Int. J. Comput. Vis.*, vol. 38, pp. 199–218, 2000.
[22] A. Kar, C. Häne, and J. Malik, "Learning a multi-view stereo machine," *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.
[23] M. Ji, J. Gall, H. Zheng, Y. Liu, and L. Fang, "Surfacenet: An end-to-end 3d neural network for multiview stereopsis," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2307–2315.
[24] Y. Yao, S. Li, S. Zhu, H. Deng, T. Fang, and L. Quan, "Relative camera refinement for accurate dense reconstruction," in *Int. Conf. 3D Vis.* IEEE, 2017, pp. 185–194.
[25] V.-C. Miclea and S. Nedevschi, "Dynamic Semantically Guided Monocular Depth Estimation for UAV Environment Perception," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.
[26] C. Wang, H. Xu, G. Jiang, M. Yu, T. Luo, and Y. Chen, "Underwater monocular depth estimation based on physical-guided transformer," *IEEE Trans. Geosci. Remote Sens.*, 2024.
[27] Q. Xu and W. Tao, "Multi-scale geometric consistency guided multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun 2019.
[28] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun 2016.
[29] W. Jing, K. Chi, Q. Li, and Q. Wang, "3-d neighborhood cross-differencing: A new paradigm serves remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–11, 2024.
[30] Q. Li, M. Zhang, Z. Yang, Y. Yuan, and Q. Wang, "Edge-guided perceptual network for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–10, 2024.
[31] C. Yang, K. Zhuang, M. Chen, H. Ma, X. Han, T. Han, C. Guo, H. Han, B. Zhao, and Q. Wang, "Traffic sign interpretation via natural language description," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 18 939–18 953, 2024.
[32] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Rgb-induced feature modulation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–11, 2023.
[33] Q. Li, Y. Yuan, and Q. Wang, "Multi-scale factor joint learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, 2023.
[34] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multi-branch mutual-guiding learning for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, 2025.
[35] 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 Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 5525–5534.
[36] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *IEEE Int. Midwest Symp. Circuits Syst. (MWSCAS)*. IEEE, 2017, pp. 1597–1600.
[37] J. Yang, W. Mao, J. M. Alvarez, and M. Liu, "Cost volume pyramid based depth inference for multi-view stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 4877–4886.
[38] S. Cheng, Z. Xu, S. Zhu, Z. Li, L. E. Li, R. Ramamoorthi, and H. Su, "Deep stereo using adaptive thin volume representation with uncertainty awareness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 2524–2534.
[39] G. Zhang, C. Xue, and R. Zhang, "Supernerf: High-precision 3-d reconstruction for large-scale scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–13, 2024.
[40] C. Zhang, Y. Yan, C. Zhao, N. Su, and W. Zhou, "Fvmd-isre: 3-d reconstruction from few-view multidate satellite images based on the implicit surface representation of neural radiance fields," *IEEE Trans. Geosci. Remote Sens.*, vol. 62, pp. 1–14, 2024.
[41] J. Liu, J. Gao, S. Ji, C. Zeng, S. Zhang, and J. Gong, "Deep learning based multi-view stereo matching and 3d scene reconstruction from oblique aerial images," *ISPRS J. Photogramm. Remote Sens.*, vol. 204, pp. 42–60, 2023.
[42] 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.
[43] W. Zhang, Z. Yang, Q. Li, and Q. Wang, "Semantic-guided multi-view stereo reconstruction for aerial image," *IEEE Transactions on Geoscience and Remote Sensing*, pp. 1–1, 2025.
[44] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.
[45] L. Chen, Y. Fu, L. Gu, C. Yan, T. Harada, and G. Huang, "Frequency-aware feature fusion for dense image prediction," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2024.
[46] Z. Shen, Y. Dai, and Z. Rao, "Msmd-net: Deep stereo matching with multi-scale and multi-dimension cost volume," *arXiv preprint arXiv:2006.12797*, vol. 4, no. 6, 2020.
[47] Z. Shen, Y. Dai, and Z. Rao, "Cfnet: Cascade and fused cost volume for robust stereo matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 906–13 915.
[48] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Proc. Euro. Conf. on Comput. Vis.* Springer, 2016, pp. 501–518.
[49] F. Wang, S. Galliani, C. Vogel, P. Speciale, and M. Pollefeys, "Patchmatchnet: Learned multi-view patchmatch stereo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 194–14 203.
[50] Z. Yu and S. Gao, "Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1949–1958.
[51] X. Ye, W. Zhao, T. Liu, Z. Huang, Z. Cao, and X. Li, "Constraining depth map geometry for multi-view stereo: A dual-depth approach with saddle-shaped depth cells," in *in Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 17 661–17 670.

Wei Zhang is pursuing a Ph.D. in computer science and technology at the School of Computer Science and the School of Artificial Intelligence, Optics, and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, remote sensing, and 3D reconstruction.

Qiang Li (Member, IEEE) is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University. His research interests include remote sensing image processing, particularly for image quality enhancement, object/change detection.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively. He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, machine learning, pattern recognition and remote sensing. For more information, visit the link (https://crabwq.github.io/).