Semantic-guided Multi-view Stereo Reconstruction for Aerial Image

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

***Abstract*—The application of learning-based Multi-view Stereo (MVS) depth estimation methods has achieved significant results in large-scale 3D reconstruction benchmarks. However, adjacent terrains in aerial image interfere with depth estimation along building edges during matching process, leading to inaccurate results. To address these challenges, we propose a new end-to-end MVS network, named FuS-MVSNet, which fuses monocular depth probability as a semantic guidance into the multi-view geometry-based MVS framework. By combining the strengths of geometric consistency and local semantics, FuS-MVSNet achieves notable enhancements in both accuracy and robustness. Specifically, we first construct a monocular branch based on the pre-trained Depth Anything model to perform monocular metric depth estimation. The non-shared parameters ensure that the depth estimation process is independent of multi-view branch, focusing exclusively on semantic depth inference. Subsequently, to incorporate monocular features into the multi-view network, we introduce a volume adaptive fusion module, which adaptively integrates monocular feature volumes into the standard cost volume via an attention mechanism and guides the cost volume regularization. Finally, confidence-based dynamic selection between the two prediction branches ensures the selection of the more robust branch result under challenging conditions. Qualitative and quantitative results indicate that we achieve competitive performance on multiple benchmarks, including the WHU and LuoJia-MVS datasets.**

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

I. Introduction

As a fundamental technology in contemporary geographic information systems (GIS) and remote sensing, large-scale 3D reconstruction based on aerial imagery is of critical importance for the accurate capture and analysis of Earth's surface features. By interpreting aerial remote sensing images, this technique generates high-resolution 3D models covering extensive geographical areas, thereby providing reliable data for key applications such as topographic mapping [1], urban planning [2], [3], disaster assessment [4], [5], and environmental monitoring [6]. The majority of existing reconstruction methods rely on commercial software such as ContextCapture [7], SURE [8], and Pix4D [9]. However, due to the limitation of traditional dense matching techniques [10], including semi-global matching [11] and PatchMatch [12], these methods may encounter matching errors in complex scenes, necessitating manual post-processing. Consequently, further research into multi-view reconstruction for large-scale remote sensing scenarios is imperative.

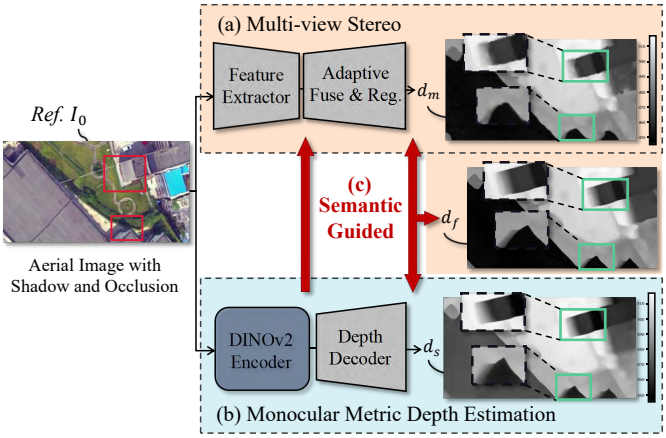

Fig. 1. Overview of our depth estimation framework for the combined local semantic and geometric consistency: (a) The general MVS framework includes an volume adaptive fusion module designed to align and fuse features from different branches and viewpoints. (b) The monocular branch implement based the Depth Anything [13], with fine-tuning of the decoder for metric depth estimation. (c) The final estimation results are obtained by utilizing semantic guidance to fuse the features and output images from the two branches.

In recent years, learning-based MVS methods have introduced a new paradigm for 3D reconstruction through the regularization of cost volumes constructed from convolution features [14], [15]. For example, MVSNet [14] encodes camera parameters through differentiable homography warping and employs 3D CNN to eliminate noise, thereby enabling the regression of depth estimation. Cas-MVSNet [15] proposes a multi-stage depth estimation architecture that incrementally constructs a coarse-to-fine depth hypothesis range, significantly reducing memory and computation costs. While considerable progress has been made in the field of object-level reconstruction, the application of neural networks to scene-level reconstruction from aerial images presents new challenges. REDNet [16] is the first network for large-scale MVS matching and surface reconstruction, which employs a recurrent encoder-decoder architecture to encode and decode cost volumes obtained from multi-view images. HDC-MVSNet [17] introduces a hierarchical deformable cascade network for high-resolution multi-scale feature extraction,

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

Zhigang Yang, 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: zgyang@mail.nwpu.edu.cn, liqmges@gmail.com, crabwq@gmail.com) (Corresponding author: Qi Wang, Qiang Li.)

creating full-resolution depth maps enriched with contextual information. SDL-MVS [18] introduces a novel paradigm that learns feature interactions across view spaces and models depth ranges for better depth estimation. While current aerial MVS methods have enhanced precision through framework optimizations, they remain fundamentally limited in addressing depth estimation inaccuracies caused by spectral-structural homogeneity among urban elements (e.g., buildings, pavements, vegetation) within aerial image.

In contrast to multi-view methods, monocular depth estimation relies on semantic understanding of the scene and perspective cues to estimate depth without the need for additional viewpoints [19]. It is more robust in weakly textured areas, dynamic objects, and independent of camera pose. Even in regions with complex elements, monocular methods can capture subtle depth variations and local features with greater precision, providing per-pixel depth estimation. However, due to scale ambiguity, monocular methods generally lack the metric accuracy of multi-view approaches. Recent research aim to combine monocular and multi-view depth estimation to improve performance in weakly textured areas [20]–[22]. For example, MaGNet [20] improves multi-view matching efficiency by sampling depth candidates near the monocular depth predictions and enforces consistency with the monocular depth to prevent multi-view matching failures, such as on textureless or reflective surfaces. AFNet [21] constructs a dual-branch network for monocular and multi-view predictions, where the monocular features in the decoder part are fused. Finally, the predicted confidence is used to fuse the monocular and multi-view depth to achieve more robust and accurate depth estimation. However, these approaches overlook a critical issue: monocular and multi-view branches sharing feature parameters may produce similar predictions, failing to fully leverage the complementary strengths of both methods.

To address these limitations, we propose a multi-view depth estimation method with semantic guidance. This method is designed to combine the strengths of both monocular and multi-view depth estimation, with the aim of achieving accurate and robust depth predictions. As illustrated in Fig. 1, we propose a dual-branch network architecture, wherein each branch is dedicated to a specific depth cue: one branch extracts local semantic depth information, while the other leverages geometric information from multiple viewpoints. The monocular branch is constructed upon the Depth Anything framework, with an independent feature extraction module that ensures discrete focal points between the monocular and multi-view branches. Furthermore, in order to bridge the feature domain gap between the aforementioned branches and integrate semantic features into the multi-view branch, we introduce an attention-based volume adaptive fusion module, which aggregates cross-domain features during the decoding process. To further exploit the potential of semantic guidance, we implement a dynamic pixel-level fusion of the outputs from both branches based confidence. The FuS-MVSNet employs this innovative architecture, effectively combining the local semantic insights of the monocular branch with the geometric consistency constraints of the multi-view branch. This significantly enhances depth estimation accuracy and robustness, particularly in handling complex edge regions by mitigating mismatches or erroneous matches. The main contributions are summarized as follows:

- We propose FuS-MVSNet, which leverages local semantic to guide multi-view stereo matching in challenging areas for more robust and accurate depth estimation. It achieves excellent performance on datasets such as WHU [16] and Luojia-MVS [17].
- We propose a volume adaptive fusion module that seamlessly integrates semantic features into a standard cost volume to bootstrap depth inference. In addition, we fuse monocular branching results based on confidence at the pixel level. This two-stage semantic guidance guarantees prediction accuracy in complex edge regions.
- We construct an independent monocular branch based on Depth Anything [13], which guarantees depth inference by focusing only on semantics, even though the non-shared parameters enhance the difficulty of metric depth estimation. Stable quantitative predictions in camera noise experiments validate the effectiveness of this branch.

II. RELATED WORK

A. Monocular Depth Estimation

Monocular depth estimation (MDE) is an important research domain in the fields of computer vision and machine learning. It aims to infer the 3D depth information of a scene from a single 2D image. This technology has extensive applications in autonomous driving [23], [24], robot navigation [25], virtual reality [26], and augmented reality. Due to the absence of direct depth cues provided by stereo vision, monocular methods rely on semantic information, texture content, contextual relationships, and shape cues within the image to infer depth. In recent years, with the development of deep learning technology [27]–[29], MDE has made significant progress, with convolutional neural network (CNN)-based methods gradually becoming mainstream in research [30]. Additionally, the introduction of the Transformer architecture has brought new opportunities to MDE. Transformers, with their powerful sequence modeling capabilities, can better capture global information in images, thereby improving depth estimation accuracy [31], [32]. Recent studies have shown that hybrid models combining Transformers and CNNs exhibit exceptional performance in MDE tasks [33], [34]. Despite their continuous performance improvement across various benchmarks, the most advanced accuracy of these methods still falls short of multi-view geometry-based approaches. In our study, an independently fine-tuned monocular depth estimation branch is introduced to calibrate MVS matching results. This dual-level integration of localized semantic cues at both feature and image levels improves the precision and robustness of MVS predictions in boundary challenged regions.

B. Multi-view Depth Estimation

Multi-view depth estimation holds a significant position in the field of computer vision, inferring the 3D structure

of a scene from multiple images taken from different viewpoints. Traditional methods for multi-view depth estimation rely on stereo matching and disparity calculation techniques. These methods typically involve finding corresponding points between images, calculating disparity maps, and inferring depth information. However, these traditional approaches face considerable challenges when dealing with complex scenes, occlusions, or texture-less areas, limiting the accuracy and robustness of depth estimation. Recently, integrating multi-view geometric information with deep learning models has become a prominent research direction. MVSNet maps features from multiple images into the same depth space and uses 3D convolution for joint optimization, significantly improving the accuracy and robustness of depth estimation [14]. Subsequent improvements such as R-MVSNet [35] and Cas-MVSNet [15] have further enhanced model performance and efficiency.

However, despite some progress in close-range target reconstruction, the differences between close-range images and aerial images present new challenges for applying these models. REDNet is the first model to outperform all traditional MVS methods in large-scale multi-view stereo reconstruction and has excelled on datasets like WHU [16] dataset. Ada-MVS [36] introduce a new depth estimation architecture that combines an adaptive multi-view cost aggregation mechanism with an efficient regularization process, specifically designed for reconstructing large-scale scenes from multi-view images. HDC-MVSNet [17], based on a cascaded network design, introduces a deformable mechanism to address scale variations in aerial imagery, enhancing depth estimation capabilities for aerial images. EG-MVSNet [37] addresses depth adhesion of building and terrain edges under insufficient lighting and low image resolution by combining building edge information and jointly estimating depth and edge maps. CSC-MVS [38] propose a network for MVS in remote sensing that integrates clustering-based semantic consistency for depth estimation. Inspired by these advancements, our research focuses primarily on improving the accuracy of depth estimation in the edge regions of buildings within multi-view stereo. To achieve this, we innovatively combine monocular semantics with multi-view stereo, achieving collaborative optimization for high-precision aerial image depth estimation.

III. Proposed Method

This section provides a comprehensive description of the proposed FuS-MVSNet and illustrates its pipeline in Fig. 2. The FuS-MVSNet method aims to predict pixel-level depth map $d \in \mathcal{R}^{H\times W}$ of the reference image $I_0 \in \mathcal{R}^{3\times H\times W}$ using the given reference image, corresponding $n-1$ source images $\{I_i \in \mathcal{R}^{3\times H\times W}\}_{i=1}^{N-1}$, and the associated camera extrinsics $\{[R_{0,i}|t_{0,i}]\}_{i=1}^{N-1}$ and intrinsics $\{K_i\}_{i=1}^{N-1}$. Three parts are included, monocular branch, multi-view branch and loss function.

A. Monocular Branch

The monocular branch takes the reference image I_0 and its corresponding camera intrinsics as input and outputs the depth map d_s and the corresponding confidence probability c_s. As mentioned in the Sec. I, the depth estimation results of the monocular branch based on shared parameters are influenced by the multi-view branch and cannot fully rely on semantic feature for depth reasoning. Therefore, we chose the Depth Anything [13] to construct the monocular branch. With its strong generalization capability, we fine-tuned the monocular branch using metric depth labels, ensuring that the predicted depth results are in the same scale range as the multi-view branch results (as shown in Fig. 3). This branch consists mainly of two parts: the encoder module and the decoder module.

The Encoder Module uses the pretrained of DINOv2 [39] for feature extraction. To balance performance and efficiency, we select the version of ViT-β as the backbone network. The embedding dimension is set of 768 with 12 heads and the output channels to 768. The Decoder Module uses the Dense Prediction Transformer (DPT) [40] for depth regression. The DPT can recover image-like representations from the output tokens of any layer in the Encoder Module, gradually fusing all feature representations into the final dense prediction. Specifically, it first apply a spatial concatenation operation to produce a feature map of size $H_p \times W_p$ with D channels. Next, we project the input representation to a fixed $\hat{D}$ feature using a 1×1 convolution and rescale all representations to $H/2\times W/2$ using either 3×3 convolutions or transposed convolutions. Finally, the RefineNet Feature Fusion [41] is used to combine feature blocks extracted from consecutive stages, progressively upsampling the representation by a factor of 2 at each fusion stage. The final representation has a size of half the resolution of the input image. An additional output head for the depth estimation task is appended to produce the final prediction.

B. Multi-view Branch

Following the framework of existing deep learning-based MVS methods [15], the multi-view branch takes the reference image I_0, source images $\{I_i\}_{i=1}^{N-1}$, and the corresponding camera intrinsics K and extrinsics R as input, and estimates the depth map d_m along with the corresponding confidence probability c_m. Unlike the monocular branch, which directly decodes and regresses the depth, the multi-view branch performs warping operations within the reference camera frustum to construct a 3D volume. The volumes from all views are then merged before regressing the depth. Furthermore, the multi-view branch adopts a cascaded architecture, employing a coarse-to-fine strategy. This approach significantly reduces memory consumption, improves computational efficiency, and incrementally refines the depth estimation results.

Firstly, multi-scale features are extracted from the input images using a feature extraction network, capturing geometric details at different scales. Specifically, the reference image and source images are processed through a feature extraction network with shared weights as the backbone, extracting image features from all images and obtaining feature representations at different scales. The feature extraction network is composed of an FPN network:

$$F_0^l, \{F_i^l\}_{i=1}^{N-1} = f(I_0, \{I_i\}_{i=1}^{N-1}), \tag{1}$$

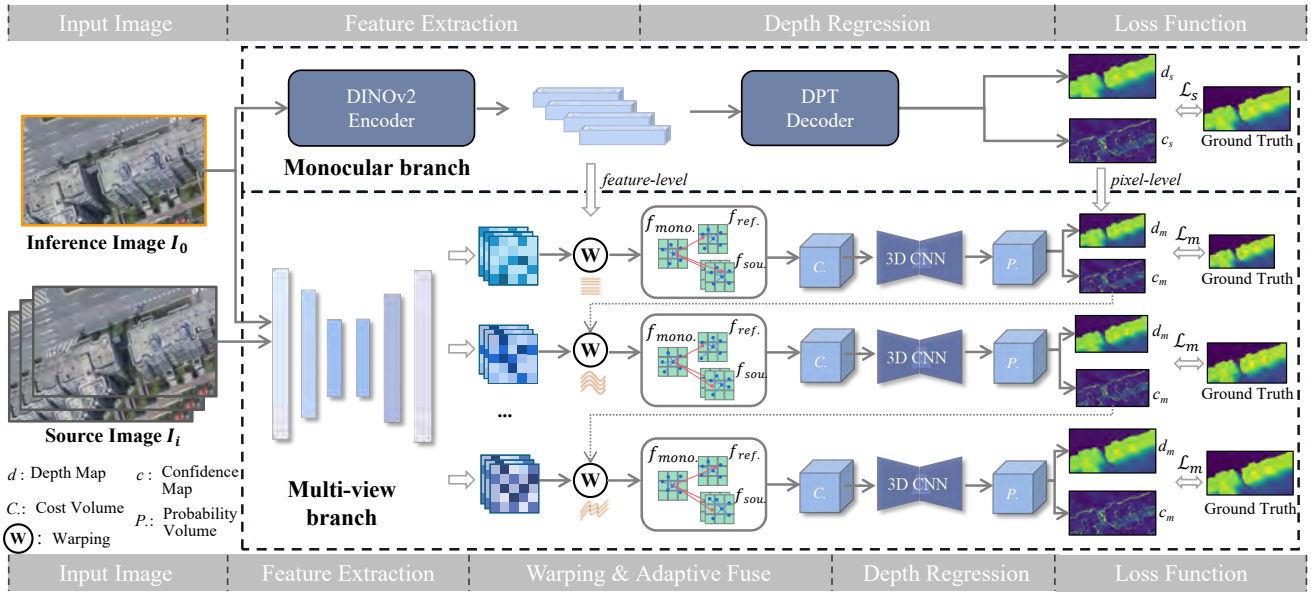

Fig. 2. Overview over our framework for the combined semantic and geometric consistency. By combining monocular branch metric depth estimate result, guide the MVS depth estimation branch. Where the upper part is the monocular branch and the lower part is the multi-view branch. It can be seen that the multi-view branch has an additional warping and fusion stage.

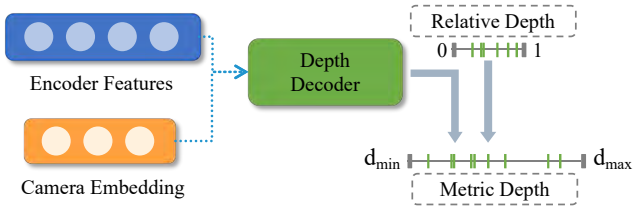

Fig. 3. Illustration of monocular metric depth estimation. By introducing camera parameters, it directly outputs metric depth results within the same depth range as the multi-view branch.

among them, F_0^l and $\{F_i^l\}$ represent the feature representations of the reference image and source images at the l-th scale, respectively. The dimensions of the extracted multi-scale features are C= 96, 192, 384, and 768.

Next, depth hypothesis sampling is performed at each scale by constructing a 3D volume v_i. We uniformly sample the depth range of the reference view to obtain $\mathcal{D}$ hypothetical depth planes that cover possible terrain variations. Following the classical plane sweep methods [42], the features of each source image are warped into the reference camera frustum, incorporating multiple depth hypotheses. The homography between the feature map of the i-th view and the reference feature map at $\mathcal{D}$ is represented as:

$$F_{proj}(\mathcal{D}) = warp(\{F_i^l\}, K_i, R_{0,i}, t_{0,i}, \mathcal{D}), \tag{2}$$

where $F_{proj}(\mathcal{D})$ represents the result of projecting the source image features F_i^l into the frustum of reference view for the given set of hypothetical depth planes $\mathcal{D}$. The $warp(\cdot)$ denotes the warping operation, which transforms the source features into the coordinate frame of the reference image.

All feature maps are downsampled and formed into N feature volumes v_i and fused into the cost volume c_i. During this process, features from the monocular branch are integrated to provide localized semantic guidance for dense matching operations. This operation can be considered as treating the monocular branch features as source volumes from the same viewpoint. However, since these feature volumes come from different feature spaces, there are differences that prevent direct fusion. Based on this observation, we introduce an volume adaptive fusion module to simultaneously fuse feature volumes from different views or feature spaces(as shown in Fig. 4). To better capture the spatial relationships between features, we first introduce deformable convolutions. These convolutions can flexibly adjust the position of the convolution kernels by learning offsets, thereby improving the flexibility and accuracy of feature transformations. Additionally, the results of the deformable convolutions serve as inputs to a view-weighting module that learns pairwise weight maps for each adjacent image, and then aggregates the feature volumes v_i into a cost volume c. Through a data-driven strategy, the model assigns higher weights to views that are more conducive to matching, and suppresses inconsistent information between views caused by occlusions or noise, thus achieving robust and adaptive cost aggregation.

To mitigate noise in the fused cost volume, a 3D CNN is employed for regularization of the cost volume. Specifically, the constructed cost volume is inputted into a 3D convolutional network, where multi-layered 3D convolutions and skip connections are utilized to generate a probability volume p for depth inference. A 3D convolutional network structure with a depth of 8 is employed, with all convolutional kernels set to $3 \times 3 \times 3$. Following integration with intermediate features from the backbone network at the 1-th, 3-th, and 4-th layers, a final result with single channel is outputted. Each 3D convolution includes convolutional, regularization, and activation layers to progressively extract and fuse information

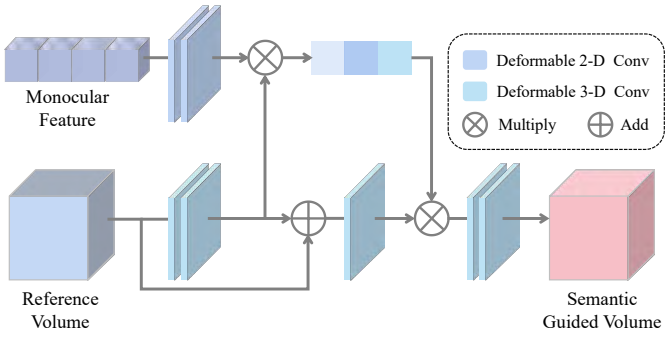

Fig. 4. The pipeline of volume adaptive fusion module. Deformable convolutions are used to reduce feature misalignment between monocular features and the reference feature, and attention mechanisms is applied to generate a weight matrix, achieving the correction of incorrectly estimation regions.

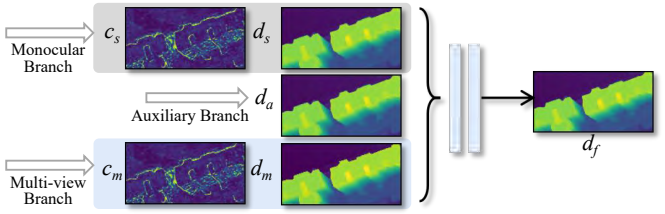

Fig. 5. The pipeline of confidence-based multi-branch fusion. Three depth maps and the corresponding confidence map are merged and passed through the convolution layer, and the final prediction is obtained.

from the cost volume. Through this regularization, noise and uncertainty within the cost volume are effectively suppressed, enhancing the robustness of the disparity estimation. Finally, by pixel-wise computation of weighted depth hypotheses, the comprehensive depth estimation result d_m is determined. Simultaneously, confidence probability c_m is generated to indicate the confidence level of depth estimation for each pixel.

To obtain the final accurate and robust prediction, we also fuse the two branches at the pixel level, as shown in Fig. 5. We fuse three depth estimation maps, including the depth maps d_s and d_m generated by the dual branches and their corresponding confidence maps c_s and c_m. The third component is an auxiliary branch inspired by MVSTER [43], which directly aggregates the feature maps extracted by the FPN and forwards them to the regression head to produce the depth estimation result d_a. This branch not only enhances the semantic information for depth differentiation during the matching stage but also serves as a weight to guide fusion during the view aggregation stage. Finally, all branch results are merged and passed through a convolutional layer to obtain the final fused depth $d_f \in \mathcal{R}^{H\times W}$.

C. Loss Function

To effectively optimize the results of depth estimation, we introduce a composite loss function that integrates multi-view geometric consistency and monocular semantic depth priors. The depth loss is formulated using a standard L1 loss function, ensuring simplicity and effectiveness. Maintaining geometric consistency in the depth map is essential for achieving accurate 3D reconstruction. We first introduce a geometric consistency loss to ensure that the generated depth map remains consistent with the real scene in both local and global geometric structures. The definition of this loss function is represented as:

$$\mathcal{L}_m = \sum_{l=1}^{L} c_m |d_m - d_{gt}|, \tag{3}$$

where d_m and d_{gt} are the predicted depth and ground truth depth, respectively, and c_m is the confidence at each pixel. L is the total scale. Through a weighted loss function, we can reduce the contribution of low-confidence regions to the loss function, thereby enhancing the model focus on high-confidence regions. Subsequently, leveraging information provided by monocular estimation branch can improve the robustness and accuracy of depth estimation in complex edge regions:

$$\mathcal{L}_s = c_s |d_s - d_{gt}|. \tag{4}$$

To comprehensively consider optimization objectives that incorporate geometric consistency and local feature consistency, we define the final composite loss function and is represented as:

$$\mathcal{L}_{total} = \lambda_m * \mathcal{L}_m + \lambda_s * \mathcal{L}_s + \lambda_a * \mathcal{L}_a + \lambda_f * \mathcal{L}_f, \tag{5}$$

where λ are weighting coefficients for the loss functions, used to balance the impact of multiple loss on model training. $\mathcal{L}_a$ and $\mathcal{L}_f$ represent the L1 loss of d_a and d_f with respect to the ground truth, respectively. By optimizing the composite loss function, the performance and generalization capability of the depth estimation model in complex scenarios can be improved. Furthermore, a multi-stage training strategy is employed throughout the training process, where the monocular and multi-view branches are initially fine-tuned or trained separately, followed by joint training in the final stage.

IV. Experiments

In this section, we conduct detailed experiments on the proposed model to demonstrate its feasibility and effectiveness. Firstly, the implementation details in the experiment are introduced. Then, comparison and ablation experiments are conducted and analyzed. Finally, we provide some visual results for direct comparison.

A. Datasets

We conduct experimental validation using two large-scale remote sensing MVS datasets, LuoJia-MVS [17] and WHU [16]. The WHU dataset is a high-resolution multi-view image dataset designed for research in MVS and stereo matching. This dataset encompasses collections of images from various scenes, each scene composed of high-resolution photos captured from different viewpoints. Additionally, the dataset includes accurate camera parameters and ground truth depth maps for some scenes, serving as standardized benchmarks for evaluating algorithm performance. The Luojia-MVS dataset is a high-resolution collection of outdoor scenes, providing a robust platform for MVS research and benchmark. It features extensive urban landscapes captured from various viewpoints, complete with precise camera calibrations and ground truth data for accurate 3D reconstruction validation.

TABLE I
THE QUANTITATIVE RESULT OF DEPTH ESTIMATION ON THE WHU DATASET. SOME RESULTS ARE OBTAINED FROM HDC-MVSNET [17] AND REDNET [16]. BOLD REPRESENTS THE BEST WHILE UNDERLINED REPRESENTS THE SECOND-BEST.

Method	$N=3$			$N=5$		
	MAE↓	<3-interval ↑	<0.6-thres↑	MAE↓	<3-interval↑	<0.6-thres↑
COLMAP [44]	0.154	0.949	0.956	0.154	0.949	0.956
SURE [8]	0.224	0.936	0.920	0.224	0.920	0.936
MVSNet [14]	0.190	0.943	0.950	0.160	0.955	0.958
R-MVSNet [35]	0.183	0.935	0.953	0.173	0.938	0.954
PatchmatchNet [45]	0.173	0.948	0.965	0.160	0.950	0.969
Fast-MVSNet [46]	0.184	0.941	0.955	0.157	0.956	0.961
Cas-MVSNet [15]	0.111	0.976	0.977	0.095	0.978	0.978
REDNet [16]	0.112	0.979	0.981	0.104	0.979	0.981
HDC-MVSNet [17]	**0.101**	0.978	0.979	0.087	0.980	0.981
Ada-MVS [36]	0.109	0.953	0.976	0.102	0.964	0.980
AggrMVS [3]	0.154	0.980	0.981	0.103	0.980	0.980
RingMo-Aerial [47]	0.121	0.975	0.978	0.091	0.979	0.980
FuS-MVSNet(ours)	0.102	**0.980**	**0.981**	**0.086**	**0.983**	**0.986**

TABLE II
THE QUANTITATIVE RESULT OF DEPTH ESTIMATION ON THE LUOJIA-MVS DATASET. SOME RESULTS ARE OBTAINED FROM HDC-MVSNET [17].

Method	$N=3$			$N=5$		
	MAE↓	<3-interval ↑	<0.6-thres↑	MAE↓	<3-interval↑	<0.6-thres↑
PatchmatchNet [45]	0.255	0.872	0.927	0.283	0.841	0.904
Fast-MVSNet [46]	0.194	0.920	0.957	0.357	0.749	0.846
MVSNet [14]	0.172	0.924	0.961	0.270	0.818	0.912
R-MVSNet [35]	0.177	0.935	0.960	0.259	0.867	0.923
Cas-MVSNet [15]	0.103	0.971	0.984	0.141	0.954	0.979
HDC-MVSNet [17]	**0.089**	0.978	0.987	**0.121**	0.966	**0.983**
REDNet [16]	0.109	0.969	0.984	0.156	0.905	0.949
RingMo-Aerial [47]	0.095	0.978	0.87	0.122	0.965	0.981
FuS-MVSNet(ours)	0.102	**0.980**	**0.989**	0.133	**0.967**	0.980

B. Evaluation Metrics

We evaluate the model results using various metrics in MVS tasks, including:

- Mean Absolute Error (MAE): This is calculated as the average distance between predicted depth values and ground truth, and only calculates the distance within 100 depth intervals to exclude extreme outliers.
- <3-interval: This measures the percentage of pixels where the error is less than 3 depth intervals, it represent the accuracy of the estimated depth map.
- <0.6-thres: This measures the percentage of pixels where the error is less than 0.6 meter threshold, it also represent the accuracy of the estimated depth map.

C. Implementation Details

We implement our method using PyTorch and conduct experiments on the NVIDIA RTX 3090 GPU. We employ the AdamW optimizer and scheduled the learning rate with a one-cycle policy, setting lr_{max} = 0.001. During training, we used a sequence of three views or five views as input, denoted as $N = 3$ or $N = 5$. The comparison model uses the same parameters as in the original literature. We compare FuS-MVSNet with a range of other MVS models and software, including traditional methods such as COLMAP [44] and the commercial software SURE, as well as deep learning-based approaches like MVSNet [14], Cas-MVSNet [15], and R-MVSNet [35], among others. Then, we also compare with some recent remote sensing models on several aerial image datasets, including REDNet [16], Ada-MVS [36], HDC-MVSNet [17], AggrMVS [3], and RingMo-Aerial [47].

D. Benchmark Performance

Results on WHU. As shown in Table I, we test our model on the WHU dataset to demonstrate the generalization and flexibility of our FuS-MVSNet. We can observe that our method achieves excellent performance compared to other methods. For instance, our approach improves upon MVSNet by 0.064 (40% performance improvement) on the MAE metric. For the pixel-wise depth evaluation, our method achieves 0.983 and 0.986 on the <3-interval and <0.6-thres metrics, respectively (3% performance improvement). Due to the presence of numerous boundary regions with coexisting buildings and vegetation within the WHU dataset, accurate estimation through multi-view approaches becomes challenging. By deeply fusing monocular branch semantic cues with multi-view geometric cues, errors in matching complex boundary regions can be reduced, thereby improving the accuracy and robustness of depth estimation. Although introducing an additional monocular branch incurs some parameter overhead and increases inference time, our method employs a lightweight network architecture to maintain computational efficiency comparable to existing approaches, as shown in Table III.

We also qualitatively compare our approach with MVSNet, R-MVSNet, and RED-Net, and visualize the depth map results in Fig. 6. While the existing depth results are generally

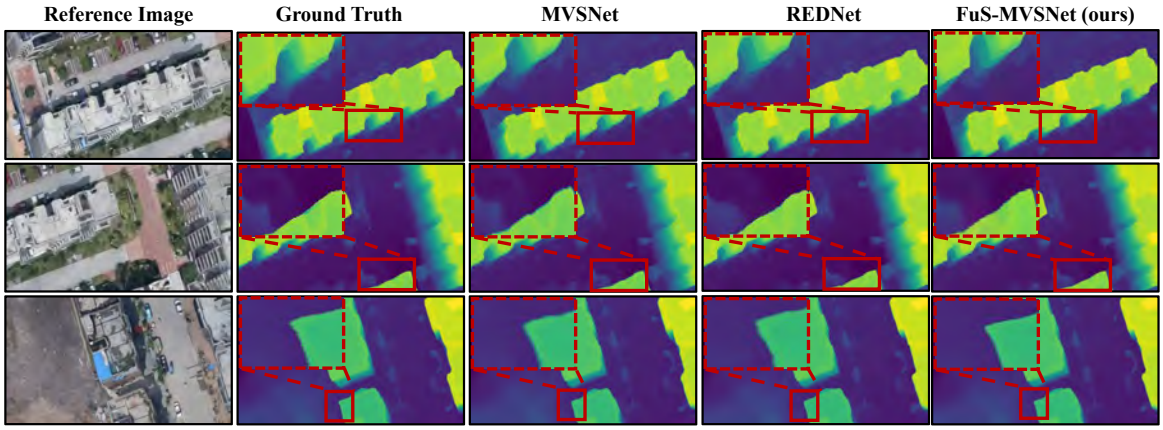

Fig. 6. The qualitative result of depth estimation on the WHU dataset. The red solid box indicates the area of focus for comparison, and its zoomed-in area is shown in the red dashed box.

TABLE III
EFFICIENCY EXPERIMENTS ON THE WHU DATASET. ALTHOUGH THE INCLUSION OF THE MONOCULAR BRANCH LEADS TO HIGHER MEMORY CONSUMPTION FOR OUR FUS-MVSNET UNDER THE SAME CONFIGURATION, IT ACHIEVES THE BEST PERFORMANCE WHILE MAINTAINING INFERENCE TIME ON PAR WITH EXISTING METHODS.

Method	MAE↓	GPU memory↓	Inference time↓
RED-Net [16]	0.103	2481 MB	22 min
AggrMVS [3]	0.091	2873 MB	4.2 min
Fus-MVSNet(ours)	0.083	5627 MB	5.4 min

accurate in predicting depths range, they exhibit blurriness in some edge regions due to the inability to obtain precise match result. Our approach introduces local semantic to ascertain the plane assignment for each pixel, and achieves more accurate results in edge regions. To further demonstrate the robustness of the proposed method, we conduct the comparison between the baseline and our approach in more challenging regions, as shown in Fig. 8. These regions, characterized by the presence of tall buildings, trees, and low-rise structures, not only exhibit mutual occlusions but also cast shadows. The results indicate that these areas are challenging for obtaining accurate depth results through direct matching, particularly in the locations highlighted by the red bounding boxes. Compared to the baseline, our method leverages semantic reasoning to produce pixel-level component layering results. This capability enhances the robustness of predictions in these challenging regions, enabling more accurate depth estimation results.

Results on Luojia-MVS. We conduct further experiments on the LuoJia-MVS dataset to thoroughly verify the performance of FuS-MVSNet in more complex environments. As shown in Table II, we can observe that our model achieves competitive performance. Specifically, compared to the MVSNet-like method, our proposed method significantly improves performance. The MAE decreases from 0.172 for MVSNet to 0.102, corresponding to a 40% performance improvement. Additionally, compared to other aerial image depth estimation methods, such as REDNet and HDC-MVSNet, our method improves the <3-interval from 0.969 and 0.978 to 0.980, and <0.6-thres from 0.984 and 0.987 to 0.989, respectively. In the case of five-view input, similarly competitive experimental results are achieved. The Luojia-MVS dataset primarily consists of mountainous and forested regions, where issues of detail blurring also occur. The introduced monocular branch is capable of performing pixel-wise depth discrimination, thereby enhancing the accuracy of boundary estimation.

Furthermore, we present the visualization of the depth maps, as illustrated in Fig. 7. While existing methods demonstrate the capability to generate visually plausible depth maps, persistent challenges remain in addressing edge prediction ambiguity caused by illumination variations or occlusions. These challenges can lead to less precise depth estimation in regions where clear boundaries are essential for accurate interpretation. To overcome these limitations, our proposed method strengthens edge distinction by incorporating a monocular depth estimation branch with semantic features. With the introduction of this additional auxiliary information, our approach effectively mitigates the issue of edge blurring, resulting in more accurate and reliable depth estimation outcomes.

E. Ablation Study

In this section, we substantiate the effectiveness of each component in the FuS-MVSNet model through ablation studies on the WHU dataset. The comparison results are shown in Table IV. It is observed that incorporating the monocular branch based on MVS2D [48] has a negligible impact on performance. This suggests that not all monocular branches contribute to improved performance due to the inherent limitations of monocular depth estimation. In contrast, our proposed monocular branch significantly enhances predictive accuracy. Despite its limitations in depth range compared to matching methods, the fine-tuned monocular metric depth estimation offers clear advantages in precise semantic delineation.

F. Visualization Analysis

To evaluate the effectiveness of our model in predicting accurate depth maps, we conducted a reconstruction of point cloud data for large-scale scenes based on the predicted metric

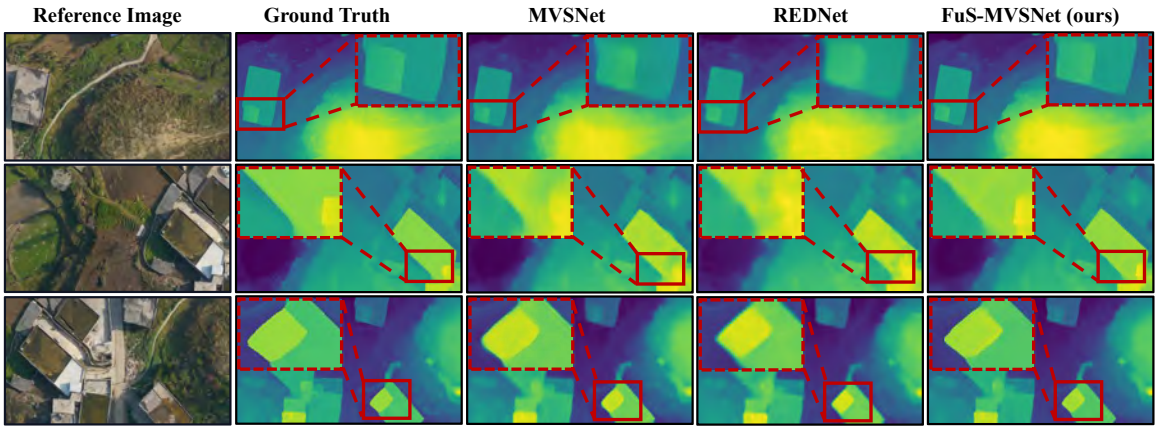

Fig. 7. The qualitative result of depth estimation on the Luojia-MVS dataset. The red solid box indicates the area of focus for comparison, and its zoomed-in area is shown in the red dashed box.

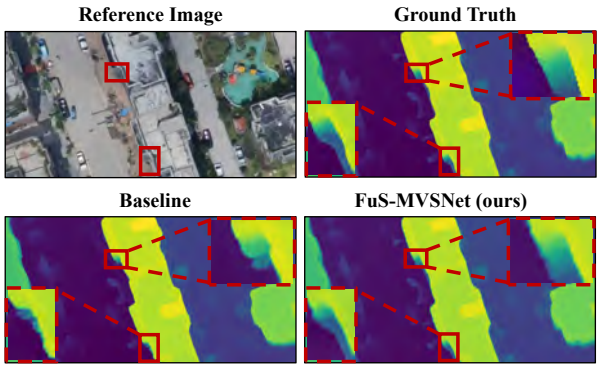

Fig. 8. The qualitative comparison between the baseline and our approach in challenging regions.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS. "BASE" DENOTES BASELINE, "MONO." DENOTES OUR MONOCULAR DEPTH ESTIMATION BRANCH. "MVS2D" DENOTES MONOCULAR BRANCH CONSTRUCTED WITH REFERENCE TO MVS2D [48]. VAFM DENOTES THE PROPOSED VOLUME ADAPTIVE FUSION MODULE. "+" REPRESENT ADD OPERATION.

Base	MVS2D	Mono.	+	VAFM	MAE
✓	-	-	-	-	0.108
-	✓	-	✓	-	0.106
-	-	✓	✓	-	0.095
-	-	✓	-	✓	**0.083**

pixel depths. The process begins with FuS-MVSNet estimating the depth of aerial images drawn from the dataset. Subsequently, we utilize the camera intrinsic and extrinsic parameter matrices to map the image coordinate system to the world coordinate system, ensuring accurate spatial representation. By integrating the depth map outputs from the test sets of the WHU and Luojia-MVS dataset, we generate comprehensive point clouds that encapsulate the 3D structure of the scenes. A detailed visual analysis of these reconstructed point clouds is then performed to assess the quality and precision of the depth estimations provided by our model.

Fig. 9 and Fig. 10 provide a visualization of the point cloud reconstruction results for areas 2 and 3 from the test set, which encompass numerous building complexes as well as extensive natural geographic features. The qualitative analysis of these visualizations reveals that in densely populated regions with regularly textured building structures, our proposed semantic-guided MVS reconstruction method excels in producing highly accurate point cloud fusion and reconstruction. The method effectively captures the fine details and complex geometries of these areas, resulting in superior reconstruction quality. Furthermore, even in regions characterized by large expanses of irregularly textured trees, the proposed MVS reconstruction method demonstrates robust performance, maintaining a high level of detail and accuracy. These findings underscore the effectiveness and superiority of our depth estimation approach, particularly in its ability to handle a diverse range of textures and structures in large-scale, complex scenes.

Fig. 11 presents the reconstructed point clouds for several representative parts of the Luojia-MVS dataset, including hills, houses, ravines, fields, and forests. Among these, buildings typically feature regular texture information, while regions like forests contain a substantial amount of irregular textures, making reconstruction more challenging. Despite these difficulties, the visualized results demonstrate that our method achieves high-quality point cloud reconstruction, underscoring the effectiveness of the proposed architecture in depth estimation. To further evaluate the practical applicability of our model in real-world scenarios, we conducted tests on several uncalibrated aerial datasets, including both publicly available datasets and data collected using our own UAV device. As shown in Fig. 12, our model achieves satisfactory reconstruction results on these real-world scenes, demonstrating strong generalization capability.

G. Discussion

Potential of Monocular Methods. Currently, mainstream 3D reconstruction methods predominantly rely on MVS frameworks, largely due to their high precision in generating accurate depth information. However, our research reveals that with appropriate metric fine-tuning, monocular depth estimation can also produce highly promising results, as illustrated in Fig. 13.

Fig. 9. Point cloud reconstruction results of the proposed method on WHU dataset. The upper section represents the sampling area of the entire dataset, while the lower section depicts the 3D reconstruction results for the test set region. Region 0 indicates the entire data collection area [16]. Region 1, 4, 5, and 6 denote the sampling locations within the training set, while regions 2 and 3 indicate the sampling locations within the test set.

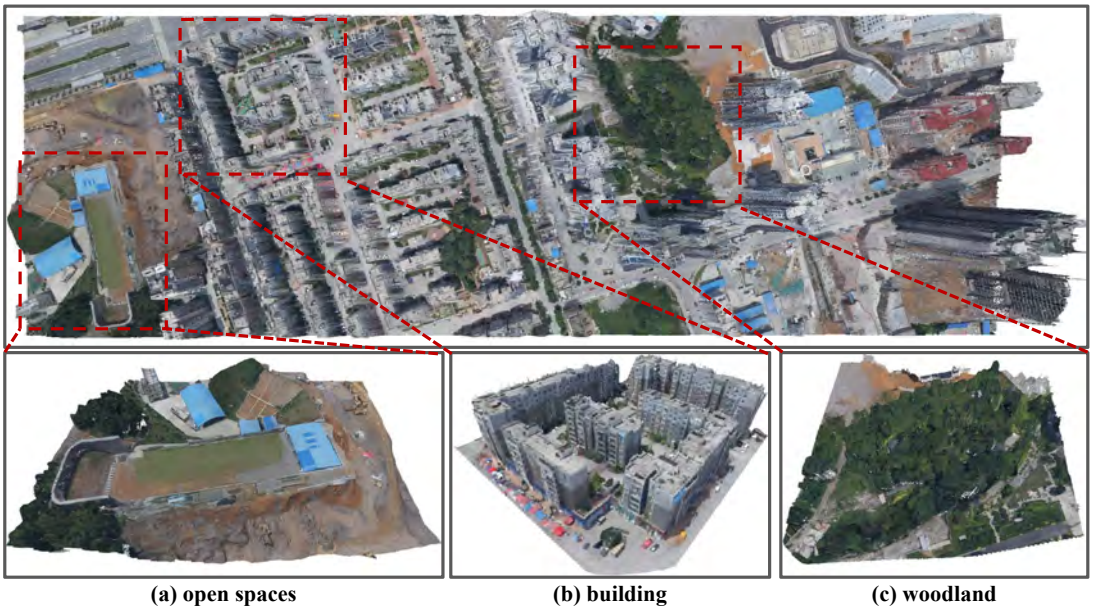

Fig. 10. Zoomed-in visualisations of local areas in test area 3 selected for (a) open spaces, (b) building, and (c) woodland, represent the majority of the reconstructed scenes. At different scales, the fused point clouds exhibit high-quality visualization results.

This finding opens up new possibilities for achieving high-precision dense 3D reconstruction using monocular depth estimation techniques in future applications. A notable challenge in aerial imaging is the difficulty of obtaining precise camera parameters, which complicates the stereo matching process in multi-view aerial images. Monocular depth estimation offers a compelling solution to this problem, allowing us to bypass the complexities of multi-view data acquisition. By doing so, we can significantly enhance the flexibility and efficiency of aerial image processing, enabling the 3D reconstruction of complex scenes with greater ease and precision.

Robustness under Noise Poses. The camera parameters in existing aerial datasets are obtained through simulated and corrected accordingly. However, in practical applications, noise is inevitable, and the robustness of MVS depth estimation networks under noisy conditions is crucial. Therefore, we test the robustness of the proposed FuS-MVSNet by adding different levels of synthetic noise. We convert the relative

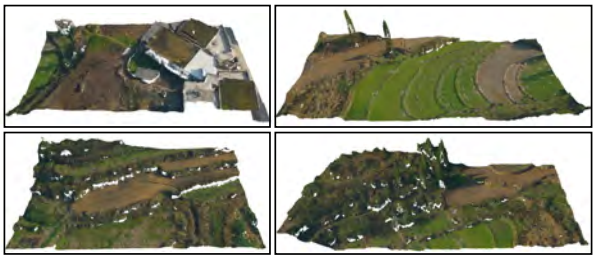

Fig. 11. Reconstruction results of the proposed method on Luojia-MVS.

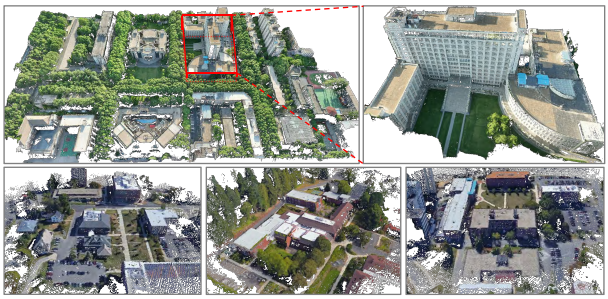

Fig. 12. Additional visualization results across different scenarios. First row: Reconstruction results from our collected aerial data—left: the full-scene reconstruction; right: a zoomed-in view of a selected region. Second row: Selected point-cloud reconstructions from the University-1652 [49] dataset.

pose to Euler angles and translation vectors, then applied perturbation coefficients to simulate various noise levels: σ = 0, σ = 0.001 and σ = 0.005. The results presented in Table V demonstrate that FuS-MVSNet exhibits superior performance compared to other MVS methods in terms of robustness against input pose noise. Across all tested noise levels, FuS-MVSNet consistently achieves higher accuracy than others methods, highlighting its resilience to pose inaccuracies.

V. Conclusion

In this paper, we introduce a novel MVS framework, FuS-MVSNet, for large-scale aerial image depth estimation that incorporates semantic priors. By integrating local semantic features with geometric consistency constraints, FuS-MVSNet achieves enhanced robustness and accuracy in depth estimation for edge regions. Furthermore, we explore the potential of fine-tuning monocular metric depth estimation on aerial images, demonstrating reliable prediction performance even with noisy camera pose conditions. Experimental results, both quantitative and qualitative, on datasets such as WHU and Luojia-MVS, demonstrate the superior performance of FuS-MVSNet, exceeding other MVS methods on standard benchmarks. Future work will focus on further narrowing the gap between monocular and multi-view depth estimation, reducing reliance on camera parameters, and improving generalization.

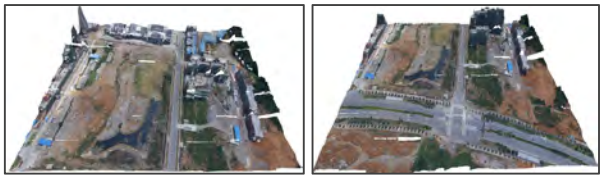

Fig. 13. Visualization of reconstructed scenes based on monocular branch. Although the lack of geometric constraints provided by additional views results in inaccurate metric estimation and noticeable misalignment at the boundaries of the point cloud stitching, the overall reconstruction quality remains reliable.

TABLE V
Performance comparison under different levels of pose noise. $\sigma = 0$ indicates no noise, with larger values representing higher noise levels.

Method	$\text{MAE}_{\sigma=0}$	$\text{MAE}_{\sigma=0.001}$	$\text{MAE}_{\sigma=0.005}$
MVSNet	0.160	0.388	0.520
FuS-MVSNet	**0.082**	**0.320**	**0.408**

References

[1] B. Li, H. Xie, X. Tong, H. Tang, S. Liu, Y. Jin, C. Wang, and Z. Ye, "High-accuracy laser altimetry global elevation control point dataset for satellite topographic mapping," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

[2] X. Zheng, M. Gao, Z.-L. Li, K.-S. Chen, X. Zhang, and G. Shang, "Impact of 3-d structures and their radiation on thermal infrared measurements in urban areas," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8412–8426, 2020.

[3] 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.

[4] A. Calantropio, F. Chiabrando, M. Codastefano, and E. Bourke, "Deep learning for automatic building damage assessment: application in post-disaster scenarios using uav data," *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. 1, pp. 113–120, 2021.

[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] L. Zhang and L. Zhang, "Artificial intelligence for remote sensing data analysis: A review of challenges and opportunities," *IEEE Geoscience and Remote Sensing Magazine*, vol. 10, no. 2, pp. 270–294, 2022.

[7] Bentley, "Contextcapture," [EB/OL], https://www.bentley.com/software/contextcapture/. 2018.

[8] 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.

[9] "Pix4d," [EB/OL], https://www.pix4d.com/. 2018.

[10] L. Zhang, L. Song, B. Du, and Y. Zhang, "Nonlocal low-rank tensor completion for visual data," *IEEE transactions on cybernetics*, vol. 51, no. 2, pp. 673–685, 2019.

[11] H. Hirschmuller, "Stereo processing by semiglobal matching and mutual information," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 328–341, 2007.

[12] M. Bleyer, C. Rhemann, and C. Rother, "Patchmatch stereo-stereo matching with slanted support windows." in *Bmvc*, vol. 11, 2011, pp. 1–11.

[13] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 10 371–10 381.

[14] 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. (ECCV)*, 2018, pp. 767–783.

[15] 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 Int. Conf. Comput. Vis. (ICCV)*, 2020, pp. 2495–2504.

[16] 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. (CVPR)*, 2020, pp. 6050–6059.

[17] 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.
[18] 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," *IEEE Trans. Geosci. Remote Sens.*, pp. 1–1, 2024.
[19] V. Arampatzakis, G. Pavlidis, N. Mitianoudis, and N. Papamarkos, "Monocular depth estimation: A thorough review," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.
[20] G. Bae, I. Budvytis, and R. Cipolla, "Multi-view depth estimation by fusing single-view depth probability with multi-view geometry," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 2842–2851.
[21] J. Cheng, W. Yin, K. Wang, X. Chen, S. Wang, and X. Yang, "Adaptive fusion of single-view and multi-view depth for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2024, pp. 10 138–10 147.
[22] Y. Fu, M. Zheng, P. Chen, and X. Liu, "Mono-mvs: textureless-aware multi-view stereo assisted by monocular prediction," *The Photogrammetric Record*, vol. 39, no. 185, pp. 183–204, 2024.
[23] Y. Ming, X. Meng, C. Fan, and H. Yu, "Deep learning for monocular depth estimation: A review," *Neurocomputing*, vol. 438, pp. 14–33, 2021.
[24] S. Grigorescu, B. Trasnea, T. Cocias, and G. Macesanu, "A survey of deep learning techniques for autonomous driving," *Journal of field robotics*, vol. 37, no. 3, pp. 362–386, 2020.
[25] X. Dong, M. A. Garratt, S. G. Anavatti, and H. A. Abbass, "Towards real-time monocular depth estimation for robotics: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 16 940–16 961, 2022.
[26] M. Rey-Area, M. Yuan, and C. Richardt, "360monodepth: High-resolution 360deg monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 3762–3772.
[27] Q. Li, Y. Yuan, X. Jia, and Q. Wang, "Dual-stage approach toward hyperspectral image super-resolution," *IEEE Trans. Image Process.*, vol. 31, pp. 7252–7263, 2022.
[28] Q. Li, W. Zhang, W. Lu, and Q. Wang, "Multibranch mutual-guiding learning for infrared small target detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 63, pp. 1–10, 2025.
[29] Q. Li, Y. Yuan, and Q. Wang, "Multi-scale factor joint learning for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, 2023.
[30] V.-C. Miclea and S. Nedevschi, "Monocular depth estimation with improved long-range accuracy for uav environment perception," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–15, 2021.
[31] Z. Cheng, Y. Zhang, and C. Tang, "Swin-depth: Using transformers and multi-scale fusion for monocular-based depth estimation," *IEEE Sensors Journal*, vol. 21, no. 23, pp. 26 912–26 920, 2021.
[32] 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.
[33] N. Zhang, F. Nex, G. Vosselman, and N. Kerle, "Lite-mono: A lightweight cnn and transformer architecture for self-supervised monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023, pp. 18 537–18 546.
[34] C. Zhao, Y. Zhang, M. Poggi, F. Tosi, X. Guo, Z. Zhu, G. Huang, Y. Tang, and S. Mattoccia, "Monovit: Self-supervised monocular depth estimation with a vision transformer," in *Proc. Int. Conf. 3D Vis. (3DV)*. IEEE, 2022, pp. 668–678.
[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. (CVPR)*, 2019, pp. 5525–5534.
[36] 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.
[37] 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.
[38] X. Huang, S. Zhang, J. Li, and L. Wang, "A multi-task network for multi-view stereo reconstruction: When semantic consistency based clustering meets depth estimation optimization," *IEEE Trans. Geosci. Remote Sens.*, 2024.
[39] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, "Dinov2: Learning robust visual features without supervision," *arXiv preprint arXiv:2304.07193*, 2023.
[40] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 12 179–12 188.
[41] G. Lin, A. Milan, C. Shen, and I. Reid, "Refinenet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 1925–1934.
[42] R. T. Collins, "A space-sweep approach to true multi-image matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*. Ieee, 1996, pp. 358–363.
[43] 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. (ECCV)*. Springer, 2022, pp. 573–591.
[44] 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. (ECCV)*, 2016, pp. 501–518.
[45] 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. (CVPR)*, 2021, pp. 14 194–14 203.
[46] 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. (CVPR)*, 2020, pp. 1949–1958.
[47] W. Diao, H. Yu, K. Kang, T. Ling, D. Liu, Y. Feng, H. Bi, L. Ren, X. Li, Y. Mao *et al.*, "Ringmo-aerial: An aerial remote sensing foundation model with a affine transformation contrastive learning," *arXiv preprint arXiv:2409.13366*, 2024.
[48] Z. Yang, Z. Ren, Q. Shan, and Q. Huang, "Mvs2d: Efficient multi-view stereo via attention-driven 2d convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2022, pp. 8574–8584.
[49] Z. Zheng, Y. Wei, and Y. Yang, "University-1652: A multi-view multi-source benchmark for drone-based geo-localization," in *ACM intern. confer. on Multim.(ACMMM)*, 2020, pp. 1395–1403.

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.

Zhigang Yang is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include remote sensing and computer vision.

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/).