Visual Consistency Enhancement for Multi-view Stereo Reconstruction in Remote Sensing

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

***Abstract*—Learnable multi-view stereo (MVS) aerial image depth estimation has obtained great success in 3D digital urban reconstruction. Currently, most depth estimation methods in the large-scale sense heavily involve adapting the general MVS framework. However, these methods often overlook the cross-view interval and limited viewpoint inherent in aerial images data. In this paper, we introduce an learning-based MVS method for aerial image depth estimation, which enhances visual consistency to address the insufficient accuracy caused by the characteristics of aerial image data, namely AggrMVS. Firstly, an Optical Flow-guided Feature Extraction Module is introduced to map the dynamic relationship between reference and source image. It explicitly captures edge information of different depth components to guide the cost volume regularization. Secondly, a Cross-view Volume Fusion Module is proposed to enhance the interaction among reference volumes, further improving the aggregation ability of the source volume. Furthermore, AggrMVS achieves refined aerial image depth estimation results with a lightweight cascade architecture. Since low-altitude oblique aerial datasets currently lack, we reconstruct a multi-category synthetic aerial scene benchmark from general MVS datasets. The benchmark dataset is available at https://github.com/ToscW/BlendedUAV. Experiments on public and proposed datasets confirm that AggrMVS outperforms other MVS depth estimation methods in terms of qualitative and quantitative aspects.**

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

I. Introduction

FINE-GRAINED 3D scene reconstruction plays an important role in agriculture [1], virtual reality [2] and mapping [3] tasks. Large-scale and highly accurate 3D reconstruction from multi-view remote sensing images is an advanced technique for obtaining terrain and ground object information on the Earth's surface. Currently, there are not only some commercial solutions that perform 3D scene reconstruction from multi-view aerial images [4]–[6], but also some open sources solutions, such as COLMAP [7] and OpenMVS [8]. For all these solutions, hand-crafted similarity metrics and semi-global matching are the core steps of dense correspondence. However, classical algorithms are usually unable to effectively handle non-Lambertian, low-textured, and specular surfaces [9], resulting in incomplete reconstruction.

In recent years, learnable methods have shown remarkable advances in addressing the MVS matching task and achieved the best results on the majority of datasets [10]–[12]. It offers a new manner for 3D reconstruction. Although it is effective for close-range object reconstruction, these deep learning approaches encounter three key limitations when applied to reconstructing outdoor large-scale scenes from multi-view aerial images. Firstly, due to the considerable distance between the camera and the ground, small movements or rotations easily result in significant cross-view intervals [13], which greatly increases the difficulty of image matching. Secondly, the aerial image typically provides limited overhead perspectives, even from varying altitudes. It requires the model to be able to extract more structural information from restricted viewpoints available. Thirdly, aerial images with expansive ground or intricate terrain may lack distinctive texture or feature points, coupled with constrained image resolution [14]. It leads to difficulty in distinguishing between buildings and terrain edges. Therefore, how to overcome these limitations, and build a new, accurate, and robust learnable depth estimation method tailored explicitly for aerial images requires further research.

In the field of data-driven deep learning methods, datasets serve as pivotal elements for model learning. Currently, only a limited number of public MVS datasets are available for remote sensing 3D reconstruction tasks, such as the WHU [15] and LuoJia-MVS [16]. While these datasets perform well in capturing expansive geographic views and distant perspectives, they have obvious limitations. Firstly, these aerial image datasets are primarily derived from synthetic high-altitude sources, so the original images or depth maps may be blurred or erroneous. Secondly, these datasets may not stem from oblique photography, which is a main technique for 3D scene reconstruction [13]. Thirdly, the scenes within these datasets are derived from segmented fixed large-scale regions [16], [17], resulting in similar content. In summary, these constraints impede the effectiveness of learning-based MVS methods in remote sensing reconstruction tasks. Therefore, there is an urgent need to construct a dataset for remote sensing 3D reconstruction with low-altitude, oblique photography, and diverse scenes.

To alleviate the poor depth map estimation results due to the characteristics of remote sensing image, we enhance the depth fusion between source and reference image based on the visual consistency. Considering that except for the existing static matching, camera position change can obtain dynamic matching results, we introduce an optical flow structure map-

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

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, Yuan Yuan, 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, y.yuan1.ieee@gmail.com, crabwq@gmail.com) (Corresponding author: Qi Wang, Qiang Li.)

ping the background change between the reference and source image as a complement to the camera extrinsic parameters. Meanwhile, the extracted optical flow results contain explicit edge information of different depth components, which can effectively help to distinguish buildings or terrain and optimize the edge relationship of the depth map. In addition, since existing methods usually ignore the connection between source volumes, we improve the matching process between source volumes and construct a cross-view interaction architecture to obtain more reconstruction details. The acquired details are then fused into cost volume for depth map estimation through a global feature consistent design. Finally, an end-to-end MVS model is provided for the aerial image depth map estimation task via a lightweight cascade network architecture. Currently, most current remote sensing MVS datasets are collected at high altitudes and sliced based on a large fixed region, whose scene content is inevitably similar. To address this issue, we reconstruct a new MVS dataset of low-altitude oblique aerial images based on the existing general MVS datasets to provide a new benchmark for learning-based model training requirements. It contains 80 models that cover a wide variety of land cover, including cities, villages, attractions, parks, etc. The main contributions of this paper are summarized below:

1) We propose a learnable MVS network (AggrMVS) to enhance visual consistency in aerial image depth estimation. AggrMVS employs multi-level modeling to deeply fuse cross-view information, which can alleviate accuracy degradation due to remote sensing scene characteristics.
2) We propose a new Optical Flow-guided Feature Extraction Module that maps the dynamic relationship between the reference and source image. It explicitly captures the edge information of different depth components in the reference image to optimize the depth map estimation.
3) We propose a Cross-view Volume Fusion Module that enhances the matching process among source volumes by combining convolution and self-attention mechanisms to obtain richer depth map details.
4) Extensive experiments on public benchmarks, such as the WHU and WHU-OMVS dataset, demonstrate that our AggrMVS method achieves excellent performance whether using three-view or five-view inputs.

II. RELATED WORK

In this section, we provide a detailed review of two key areas: general-purpose 3D reconstruction methods and aerial image depth estimation methods. In addition, some MVS open-source benchmark datasets are described.

A. MVS Methods

In recent years, deep learning has raised attention for its simplicity and effectiveness [18]–[20]. Currently, learning-based MVS approaches are categorized into two primary types: voxel-based [21] methods and depth map-based [22] methods. Here, voxel-based methods directly generate a 3d voxel representation in an integrated process, with an inevitably occupied memory. Unlike the above methods, depth map-based methods decompose the complex reconstruction process into two stages: estimation and fusion, which offer advantages in both efficiency and accuracy. Now there are many applications of aerial images based on depth maps methods, such as the generation of digital surface models.

Motivated by the concept of binocular stereo matching [23], Yao et al. [22] propose the MVSNet. It leverages differentiable deformations to encode camera parameters and constructs a 3D cost volume. To address the high memory consumption of MVSNet, R-MVSNet [24] employs GRU [25] mechanism to regularize the cost volume sequentially, alleviating the cost of increased runtime. Another significant advance is the CasMVSNet [26], which decomposes the depth estimate process into a cascade architecture, consisting of multiple progressive stages. This cascade architecture preserves fine contextual information while enabling high-resolution reconstruction, which is a popular framework in the field. Recent work further enhances network performance by integrating Transformer [27] architecture. For example, the TransMVSNet [28] employs a Transformer-based feature-matching strategy to capture contextual correlation across extended ranges. MVSTER [29] learns semantic and spatial associations based on an epipolar Transformer. MVSFormer [30] introduces a visual Transformer to enhance the abilities of feature extraction. These advancements highlight the ongoing efforts to improve the efficiency and accuracy of MVS methods through innovative network architectures.

Although some advances in the field of reconstructing at close-range objects, the differences between close-range and aerial images present new challenges for the application of the above models. To the best of our knowledge, RED-Net [15] is the first model built for large-scale multi-view stereo reconstruction, outperforms all traditional MVS methods on the WHU [31] and München [32] datasets, and its efficiency as well. Unlike MVSNet and R-MVSNet [25], RED-Net employs a Recurrent Encoder-Decoder framework to perform a sequence of convolutions on aerial images. This is followed by sequentially regularized operations to produce the final depth map. Furthermore, MS-REDNet [33] employs a high-resolution encoder-decoder module based on a cascaded architecture to fully exploit the detailed features extracted. Ada-MVS [13] presents a novel depth estimation architecture that incorporates adaptive mechanisms for multi-view cost aggregation along with an efficient regularization process, tailored specifically for the reconstruction of large-scale scenes from multi-view images. HDC-MVSNet [16] is also based on a cascade network design and introduces a deformable mechanism to cope with aerial image scale variations, which improves the ability to estimate the depth of aerial images. Different from the above studies, we takes into account the dependence of large-scale scene between reference and source images and introduce a cross-view consistency enhancement method to achieve remote sensing oblique image scene reconstruction.

B. MVS Dataset

Dataset serve as pivotal elements for model performance [34], especially for the learnable methods. Currently,

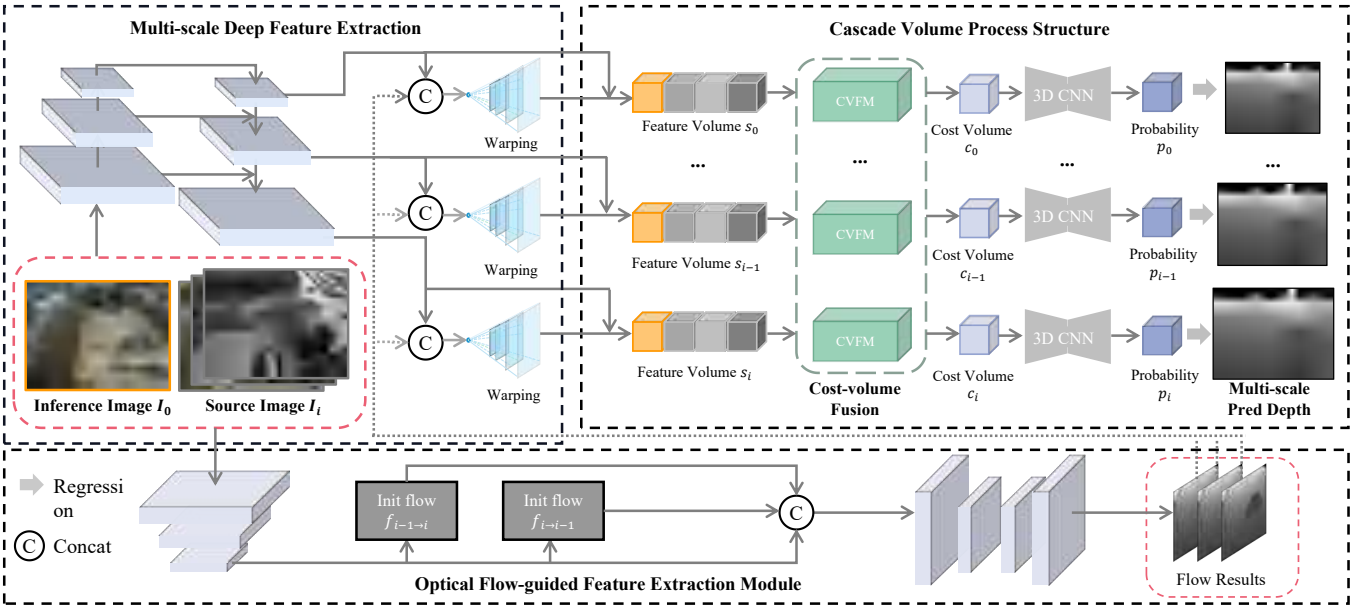

Fig. 1. Overall architecture of AggrMVS. AggrMVS firstly extracts features via multi-scale deep feature extraction, then the multi-view features are aggregated by the cascade volume process structure. Subsequently, the aggregated cost volume is regularized by 3D CNN, producing multi-scale depth map. The "CVFM" means Cross-view Volume Fusion Module.

there are some public MVS benchmark datasets: Middlebury [35], Tanks and Temples [36], DTU [37], ETH3D [38], BlendedMVS [39], LuoJia-MVS [16], WHU [15] and WHU-OMVS [13]. The first five datasets contain models that are primarily composed of indoor or close-range images. These datasets are not suitable as a benchmark for large-scale scene depth estimation due to their significant differences in viewing angle and camera parameters. The recent WHU dataset [15] is the first large-scale multi-view aerial dataset and comes from synthetic technology. It has a variety of terrain features, including high-rise buildings, sparse factories, mountains covered by forest, bare land, and rivers. Due to the lack of oblique aerial image datasets, Liu et al. [13] establish a large-scale synthetic multi-view oblique aerial image dataset, namely WHU-OMVS. However, these remote sensing MVS datasets are collected at high altitudes and sliced based on a large fixed region, whose scene content is inevitably similar. To solve it, we reconstruct a new low altitude oblique aerial image MVS dataset as a complement to the WHU [15] and WHU-OMVS [13] datasets, to further promote the application of deep learning algorithms in 3D scene reconstruction.

III. Proposed Method

This section provides a detailed description of the proposed method and outlines the overall architecture, as illustrated in Fig. 1. Our approach, AggrMVS, firstly extracts 2D multi-scale features via a Multi-scale Deep Feature Extraction. Leveraging the corresponding source image camera parameters, AggrMVS warps the source image features into the reference camera frustum, and further constructs the source volumes after combining the optical flow features. Subsequently, the Cross-view Volume Fusion Module, which is based on convolution and self-attention, matches and merges these source volumes to generate a refined cost volume. Finally, the cost volume is regularized using the 3D CNN module to predict accurate depth estimation results. In this process, AggrMVS pipeline adopts a cascade architecture that propagates the depth map from coarse to fine, improving the overall efficiency and accuracy of the framework.

A. Cascade Architecture

The cascade architecture has consistently demonstrated its effectiveness in stereo depth estimation [33], monocular reconstruction [40], and binocular reconstruction [26], significantly enhancing both efficiency and performance. To balance accuracy and speed, we take advantage of a cascade design with a sequential architecture that progressively refines the depth mapping from coarse to fine. According to established methodologies rooted in cascade-based approaches, our network is meticulously configured across three distinct stages. Here, the number of depth hypotheses is systematically defined as 48, 32, and 8 for each stage, with building a progressive refinement in precision. Meanwhile, the corresponding depth interval ratios are intricately set at 4, 2, and 1. To maintain spatial consistency throughout the refinement process, the resolutions of feature maps are proportionally adjusted to 1/16, 1/4, and 1 of the input image resolution at each stage. Furthermore, a weighting scheme is implemented across these stages to ensure that the influence of each stage contribution within the pipeline is maximized.

B. Multi-scale Deep Feature Extraction

Compared to the hand-crafted features, deep features that are enriched with contextual information through convolutional operations recently gained significant attention for their robustness in the feature extraction stage [22], [26], [29]. In

this section, we first use convolutional networks to directly extract features for all viewpoint images. Meanwhile, we obtain optical flow features for the reference image and all source images to achieve the initial perception of depth space. In this respect, the proposed Multi-scale Deep Feature Extraction contains two parts: a Multi-scale FPN-like Module is used to extract different sizes of static features, and an Optical Flow-guided Feature Extraction Module for extracting global dynamic flow results.

The Multi-scale FPN-like Module consists of a series of 2D convolutions and an FPN network, augmented with skip connections at each stage, which facilitate the seamless integration of different scale feature representations. At each stage of the FPN, rich hierarchical contextual information is meticulously incorporated. Firstly, a convolution operation with a 2 stride is used to compress the feature maps to $1/2$ and $1/4$ of their original size, respectively. Then, the compressed feature maps are upsampled to match the dimensions of the multi-scale shallow features using nearest-neighbor interpolation, after which they are merged with the original feature map. Finally, this composite of features is subjected to a fusion process through a convolutional operation with a 3×3 kernel size. The upsampling and merged operations are repeated m times to extract the pyramid features, where the module depth setting determines m. Given a reference image $I_{i=0}$ and its neighboring source images $I_{i=1,\ldots,N-1}$, all N-view images are processed through a shared-weight Multi-scale FPN-like Module to obtain the corresponding feature maps.

The Optical Flow-guided Feature Extraction Module maps the dynamic relationship between reference and source image. It explicitly captures the edge information of different depth components and enriches the semantic information of depth features. Specifically, optical flow denotes the two-dimensional vector field that describes the movement of brightness patterns between two continuous video frames. In a standard MVS system, a series of images is captured from multiple vantage points as the cameras are positioned around a static object. However, the relative motion between the cameras and the object can be conceptualized as the object moving towards a singular, stationary virtual camera (see Fig. 2). This conceptualization facilitates the computation of a dense optical flow between any two arbitrary views. Due to the large temporal and spatial intervals between the closest views of both BlendedMVS [39] and WHU [15], we employ feature maps extracted by a deep neural network to compute optical flow. In the specific implementation, we follow [41] to build this module. Firstly, the source image I_0 and each reference image I_i form an image pair as input. All image pairs are extracted with 8-fold downsampled dense features F_0, F_i. Then, we compare the feature similarities for each location in F_0 to all locations in F_i by computing their global matching. Furthermore, a Softmax matching layer is used to obtain the matching distribution. Finally, the weighted pixel grid is adapted to subtract the initial pixel grid to obtain the final optical flow result FR. Based on the definition of the source images, we form multiple image pairs of reference image with all source images to compute the optical flow. The output of the optical flow computation yields a pair of directional flow

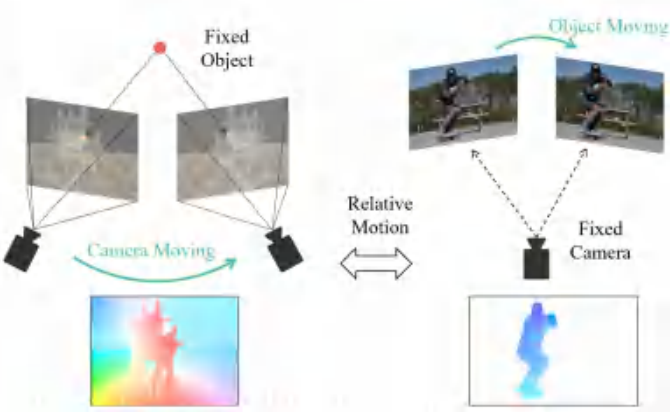

Fig. 2. Illustration of correlation between camera moving and object moving. The optical flow representation can be used to view the relative motion of a moving camera as a special case of a moving object [42].

vectors that delineate the motion between the corresponding points in two distinct views. Specifically, the forward flow result $FR_{0,i}$ represents the displacement to project points from the reference view onto the corresponding points in the source view. Conversely, the backward flow result $FR_{i,0}$ represents the optical flow trajectory that traces the reverse path, mapping points from the source view back to the reference view. All the computed optical flow outcomes are concatenated with the results of multi-scale feature extraction, resulting in a feature extraction result that integrates both dynamic and static characteristics.

C. Differentiable Homography Warping

To construct geometric consistency between multiple views, the subsequent phase involves the synthesis of a 3D volume. This is achieved by leveraging the extracted feature maps and the input parameters specific to the camera setup. Following previous learning-based methods [22], [29], the construction of the 3D volume is predicated upon the reference camera frustum. Specifically, by utilizing the camera intrinsic parameters K_i and transformation parameters $\{[\mathrm{R}_{0,i} \mid \mathrm{t}_{0,i}]\}$, source features map can be warped into the reference camera frustum, i.e.,

$$\mathrm{p}_{k,j} = \mathrm{K}_i \cdot \left(\mathrm{R}_{0,i} \cdot (\mathrm{K}_0{}^{-1} \cdot \mathrm{p}_r \cdot d_j\right) + \mathrm{t}_{0,i}\Big), \qquad (1)$$

where d_j denotes j-th hypothesized depth of pixel p_r in the reference feature, and $\mathrm{p}_{k,j}$ is the corresponding pixel in the i-th source features. After the warping operation, N-1 source volumes v_i are constructed. This operation enables the transformation of image features from one view to another while maintaining end-to-end differentiable in the deep learning network.

D. Cross-view Volume Fusion Module

After homography of the obtained source features, N-1 wrapped source volumes v_i are generated. All volumes v_i are merged to expand reference volume, forming a cost volume c, which allows for the final depth estimation. In the processing of fuse of volumes, existing cost metrics directly add up all

the source volumes with the reference volume. It ignores conflicting information caused by occlusion or noise from different viewpoints. These may impact the quality of the resulting depth map estimations. In this case, we propose a Cross-view Volume Fusion Module, which covers the hybrid of convolution and self-attention operation to fuse the warped feature volumes. Following the previous work [29], a monocular depth estimation branch is used to enrich the implicit content of the reference volume. Then, a hybrid structure of self-attention and convolution is employed to implement the association between wrapped source volumes and reference volume, and to generate the attention guidance for aggregating the feature volume from different views. In the proposed hybrid structure, we first employ convolution operations to pre-align the input volumes. Subsequently, source volumes are considered the key value of self-attention. In the proposed hybrid structure, source volumes are considered the key value of self-attention. The key value is utilized to construct the deep association of query value with self-attention operation, i.e.,

$$w_i = Softmax(Conv({v_i}^{\mathrm{T}} p_r)), \tag{2}$$

where v_i denotes the wrapped source volume, and w_i is the attention weights obtained from query and keys. $Conv(\cdot)$ denotes the convolution layer. Additionally, following [29], [43], we also employ a group-wise design of value computation to realize the visual similarity between volumes efficiently, i.e.,

$$\mathrm{s}_i^g = \frac{(v_i^g, p_r^g)}{G}, \tag{3}$$

where $g = 0, ..., G$-1. v_i^g denotes the g-th group feature in v_i and p_r^g denotes the g-th group feature in p_r. s_i^g is then merged back along the channel dimensions to yield s_i. $(,)$ denotes the inner product. Then, these values are aggregated with w for the final cost volume. The multi-view aggregated cost volume c with the shape of $D \times G \times H \times W$ is obtained by weighting all grouped relevant cost quantities and then being fed into the convolution layer, where G is the number of groups. It is defined as:

$$c = Conv(\mathbf{w}_i \mathbf{s}_i). \tag{4}$$

In summary, the proposed structure effectively combines the contents of all volumes by utilizing a hybrid approach with self-attention and convolution operation. Within this process, elements such as query values and key values are generated using both the reference and source volumes. The integration of deep channels by using convolution structures. This design not only enables efficient merging between the reference and source volumes but also promotes interaction among different source volumes. Such a design can enhance the utilization of multiple perspectives while alleviating issues like occluded regions. Such a design can enhance the utilization of multiple perspectives while alleviating issues like reconstruction defects.

E. Depth Estimation

After obtaining the cost volume c, it is first processed through a regularization network to extract probability volume p. This regularization network mainly consists of a lightweight 3D CNN. The probability volume can then be considered as the weights for the depth hypotheses, with each pixel value representing the weighted sum of these hypotheses. Once the depth map d is obtained, it is constrained using the L1 loss between the estimated depth and the ground truth depth. In the cascade network architecture, each depth estimation stage generates depth maps of different resolutions. For the results at different downsampling levels, we apply different weighted losses for computation.

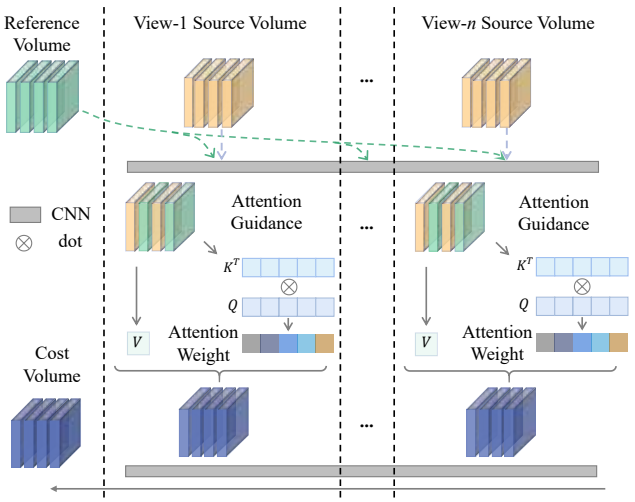

Fig. 3. Cross-view Volume Fusion Module. The reference volume achieves fusion with all other source volumes by means of multi-layer convolutional and self-attention operations.

Fig. 4. BlendedUAV vs. WHU. Left is WHU, where it is difficult to acquire images from other angles of the building. Right is BlendedUAV, where multiple viewpoints of the building are accessible. Note that there is a difference in the detail display.

IV. Dataset

This section introduces the BlendedUAV dataset, tailored for large-scale aerial image depth estimation. It consists of a comprehensive detail of data sources, dataset construction, and an in-depth exploration of its special characteristics.

A. Data Source

The components of BlendedUAV dataset are mainly from open source general MVS datasets such as BlendedMVS [44] and BlendedMVG [45], which contains many types of scenes such as cities, villages, attractions, parks, etc. We reconstruct the BlendedUAV by organizing the low-altitude oblique remote sensing scenarios and following certain requirements. As mentioned in [44], to avoid the high costs associated with active scanning technology, images and depth maps on BlendedUAV are rendered by textured 3D models to various viewpoints. During the training process, the texture mesh of scenes is reconstructed from input images and combined with images to incorporate ambient light data. This combination ensures that the model benefits from the detailed visual information present in the rendered images and the realistic ambient lighting conditions of the input data, enhancing the generalization of the model to real-world settings.

B. Dataset Construction

The BlendedUAV dataset contains 80 well-selected reconstructed 3D models. Among them, 47 for training purposes, 20 for validation, and 16 for testing. In contrast to popular remote sensing MVS datasets such as the WHU dataset, which only slices the scene over a large region, the texture models in the BlendedUAV dataset offer a range of different scenes. These include cities, buildings, sculptures, and the countryside. To facilitate direct use by researchers, BlendedUAV is constructed following the paradigm of the WHU dataset. Additionally, top-down views and redundant perspectives were filtered out while ensuring enough input images to achieve more robust and reliable evaluation results. Each scene contains raw images, camera parameters, and depth maps. The image resolution is fixed at 768×576 pixels.

C. Dataset Characteristic

While some existing remote sensing MVS datasets such as WHU, WHU-OMVS [13], and Luojia-MVS [16] offer various reconstruction scenarios, they mainly comprise image captured from high altitudes. Such data often fail to capture specific buildings from multiple perspectives, potentially degrading the MVS reconstruction task into the monocular depth estimation task. Moreover, the simple building structure reduces the difficulty of estimating depth. In contrast, BlendedUAV is constructed with a focus on low-altitude (less than 100 meters) oblique photography, which empowers researchers to tackle large-scale ground scenes while preserving intricate textural details of ground objects. Fig. 4 illustrates the perceptible disparities between the BlendedUAV dataset and the WHU dataset. Beyond these distinctions, the BlendedUAV dataset contains a multitude of additional advantages. It contains as follows:

1) *Scene Abundance*: existing datasets such as WHU are basically collected from a fixed area, and thus contain similar scenes. In contrast, BlendedUAV has independent scenes from different regions.
2) *Oblique Photography*: most existing datasets are not derived from oblique photography, which is the mainstream technology for fine-grained scene reconstruction. The BlendedUAV is obtained from oblique photography.
3) *Rich Details*: compared with the high-altitude dataset, the low-altitude BlendedUAV dataset can capture more ground details.
4) *Different Trajectories*: the scenes on BlendedUAV contain a diverse of camera trajectories. These unstructured camera tracks offer better simulated varied image-capturing styles and enhance the ability of the network to perform robustly to real-world reconstruction.

In conclusion, BlendedUAV, as a complement to the WHU and WHU-OMVS datasets, can further promote the application of the learnable MVS method in aerial image depth estimation and large-scale scene reconstruction.

V. Experiments

In this section, extensive experiments are conducted to demonstrate the effectiveness of the proposed AggrMVS model and the BlendedUAV dataset. 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. Implementation Details

Datasets: Three datasets are used in our article, including the DTU [22], WHU [15], WHU-OMVS [13] and BlendedUAV datasets. Among these datasets, DTU is an indoor dataset and photographs small objects at close range, which is the most commonly used dataset in the MVS field currently. WHU and WHU-OMVS are aerial image 3D reconstruction dataset that features large-scale scenes that incorporate the complexity of natural and urban landscapes, advancing the use of MVS techniques for outdoor 3D reconstruction tasks.

Comparison Methods: To evaluate the superiority of our proposed AggrMVS model, we demonstrate its performance on the DTU dataset. Subsequently, we conduct a comparison between AggrMVS and a selection of other MVS models and software, including COLMAP and the commercial software SURE based on the traditional method, deep neural networks such as MVSNet [22], Cas-MVSNet [26], and R-MVSNet [24], etc. Then, we also compare with some recent remote sensing models on several aerial image datasets, including RED-Net [15], Ada-MVS [13], and HDC-MVSNet [16].

Experiments Setting: The experiments are executed on the NVIDIA RTX 3090 using the PyTorch framework. To ensure the robustness of the findings, the methods are tested under both Three-View and Five-View reconstruction conditions. The networks undergo training for a total of 16 epochs, commencing with an initial learning rate set to 0.001. In this particular study, the focus is solely on predicting the depth map for the reference image. The comparison model uses the same parameters as in the original literature.

TABLE I
COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE BLENDEDUAV DATASET, BOLD REPRESENTS THE BEST WHILE UNDERLINED REPRESENTS THE SECOND-BEST.

Method	Three-View				Five-View			
	MAE↓	<1-Thres(%)↑	<3-Interval(%) ↑	<0.6-Thres(%)↑	MAE↓	<1-Thres(%)↑	<3-Interval(%)↑	<0.6-Thres(%)↑
RED-Net [15]	0.5120	0.9752	0.9659	0.9720	0.4432	0.9801	0.9633	0.9793
Ada-MVS [13]	0.5093	0.9787	0.9676	0.9739	0.5103	0.9792	0.9705	0.9647
MVSTER [29]	0.6312	0.9614	0.9525	0.9556	0.5526	0.9758	0.9701	0.9756
GeoMVSNet [46]	0.6577	0.9756	0.9645	0.9736	0.4723	0.9632	0.9632	0.9547
UniMVSNet [47]	0.4412	0.9788	0.9665	0.9745	0.4256	0.9712	0.9711	0.9814
AggrMVS (Ours)	**0.4293**	**0.9810**	**0.9709**	**0.9800**	**0.3824**	**0.9849**	**0.9732**	**0.9843**

TABLE II
COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE WHU [15] DATASET. SOME RESULTS ARE OBTAINED FROM HDC-MVSNET [16]. BOLD REPRESENTS THE BEST WHILE UNDERLINED REPRESENTS THE SECOND-BEST.

Method	Three-View				Five-View			
	MAE↓	<1-Thres(%)↑	<3-Interval(%) ↑	<0.6-Thres(%)↑	MAE↓	<1-Thres(%)↑	<3-Interval(%)↑	<0.6-Thres(%)↑
COLMAP [7]	0.1548	0.9587	0.9495	0.9567	0.1548	0.9587	0.9495	0.9567
SURE [4]	0.2245	0.9489	0.9369	0.9567	0.2245	0.9489	0.9209	0.9369
MVSNet [22]	0.1974	0.9432	0.9322	0.9474	0.1543	0.9654	0.9536	0.9582
R-MVSNet [24]	0.1882	0.9567	0.9400	0.9490	0.1505	0.9673	0.9564	0.9599
PatchmatchNet [48]	0.1730	0.9558	0.9480	0.9650	0.1600	0.9701	0.9500	0.9690
Fast-MVSNet [49]	0.1840	0.9651	0.9410	0.9650	0.1570	0.9671	0.9560	0.9610
Cas-MVSNet [26]	0.1110	0.9645	0.9760	0.9770	0.0950	0.9841	0.9780	0.9780
HDC-MVSNet [16]	0.1010	–	0.9780	0.9790	**0.0870**	–	0.9800	0.9810
RED-Net [15]	0.1120	0.9827	**0.9790**	0.9810	0.1041	0.9849	0.9793	0.9808
Ada-MVS [13]	**0.1093**	**0.9849**	0.9539	0.9766	0.1025	0.9859	0.9647	0.9802
AggrMVS (Ours)	0.1549	0.9829	0.9709	**0.9810**	0.1032	**0.9869**	**0.9803**	**0.9863**

Evaluation Metrics: Five types of metrics are used in this paper as follows:

1) The Mean Absolute Error (MAE) is calculated as the average L1 distance between the ground truth and the depths map predicted by the model, and only calculates the distance within 100 depth intervals to exclude extreme outliers.
2) The <3-Interval (%) metric is used to assess the percentage of pixels for which the L1 error is less than three times the depth interval values.
3) The <1-Thres(%) and <0.6-Thres(%) metrics are used to assess the percentage of pixels for which the L1 error is less than 1m and 0.6m threshold.
4) GPU Memory represents the GPU computational resources required for the training phase, while Inference Time represents the time required to perform inference on all test sets.
5) The Mean Acc, Mean Comp, and Overall metrics come from the [13]. These metrics indicate the accuracy, completeness, and their average score of the point cloud results, respectively.

B. Performance Comparison

In this part, we first evaluate the AggrMVS with other models in terms of quantitative aspects. Secondly, we present qualitative results on the BlendedUAV and WHU datasets. Finally, we also compare the performance of Peak GPU Memory and Inference Time metrics.

Tables I-III list the reconstruction results of the MVS networks on three remote sensing MVS datasets. Tables I-III provide a detailed comparison of reconstruction accuracy for various depth map-based Multi-View Stereo (MVS) networks across three distinct remote sensing MVS datasets. We achieve excellent results in most metrics on the BlendedUAV dataset. Specifically, compared to the second best method Ada-MVS [15], our study yields diverse improvement in the most concerned metrics in Three-View, i.e. <1-Thres(%), <0.6-Thres(%) and <3-Interval(%). We also obtain the best performance in Five-View. In our view, BlendedUAV is a low-altitude remote sensing dataset. It is more concerned about the interaction between reference and source images. Interestingly, AggrMVS implements a Cross-view Volume Fusion Module based on the hybrid attention mechanism. The method not only enhances the connection between source and reference volumes but also increases the interaction among source volumes. With respect to the WHU dataset, we achieve comparable experimental results. Although lags behind Ada-MVS on the MAE metric, we obtain the best results on all <0.6-Thres (%) metrics and <1-Thres (%) in Five-View. The reason for this phenomenon is that the WHU dataset relies less on the reference image, which reduces the advantage of the Cross-view Volume Fusion Module.

Similarly, AggrMVS still attains the advantage in the <1-Thres(%), <0.6-Thres(%) and 3-Interval(%) metrics on the WHU-OMVS [13] dataset, respectively. Additionally, Ada-MVS [15] also demonstrate the usability of their models, achieving better results on the MAE metrics. However, it need a longer inference time compared to AggrMVS. Moreover, we also demonstrate that AggrMVS can surpass other general methods on the close-range DTU [22] dataset, even if AggrMVS is mainly developed for large-scale aerial images. Table IV presents that the AggrMVS outperforms that of

TABLE III
COMPARISON OF THE DIFFERENT DEPTH MAP-BASED MVS METHODS ON THE WHU-OMVS [13] DATASET. WE RETRAINED ALL MODELS ON WHU-OMVS AND COMPARED THEM ON DEPTH MAP RESULTS TO ENSURE FAIRNESS.

Method	Accuracy				Efficiency	
	MAE↓	<1-Thres(%)↑	<3-Interval(%) ↑	<0.6-Thres(%)↑	GPU Memory↓	Inference Time↓
Cas-MVSNet [26]	0.1544	0.9781	0.9662	0.9621	**9921 MB**	20 min
MS-REDNet [33]	0.1562	0.9784	0.9651	0.9611	11203 MB	24 min
AdaMVS [13]	**0.1500**	0.9793	0.9685	0.9632	10721 MB	22 min
AggrMVS (Ours)	0.1534	**0.9796**	**0.9705**	**0.9700**	11987 MB	**20min**

TABLE IV
QUANTITATIVE RESULTS ON DTU EVALUATION SET. SOME RESULTS ARE OBTAINED FROM RED-NET [31].

Method	Mean Acc.↓	Mean Comp.↓	Overall(mm)↓
R-MVSNet [24]	0.3850	0.4590	0.4220
RED-Net [15]	0.4560	0.3260	0.3910
CasMVSNet [26]	**0.3250**	0.3850	0.3550
MVSNet [22]	0.3960	0.5270	0.4620
AggrMVS(Ours)	0.3520	**0.2780**	**0.3141**

R-MVSNet [24], RED-Net [31] by significant improvement when subjected to identical post-processing techniques, which include photometric and geometric filtering.

Fig. 5 and Fig. 6 present a visual comparison of AggrMVS on the BlendedUAV and WHU datasets, respectively. In Fig. 5, the depth map predicted by other methods is highly susceptible to breakage or missing regions. Actually, existing methods ignore the large viewing angle interval of aerial images, which makes it difficult to obtain complete structural features from the source image. In contrast, AggrMVS is more accurate in the prediction of edge positions due to the introduction of optical flow information by explicitly guiding spatial perception. Therefore, the proposed model can deal with this situation. In addition, one can observe on the WHU dataset that all models can achieve similar plausible results. This reason is mainly that the WHU dataset was collected from a high altitude where a large amount of object detail is lost, which leads to an unreliable visual comparison of the reconstruction results.

C. Ablation Study

To confirm the effectiveness of the Optical Flow-guided Feature Extraction Module and Cross-view Volume Fusion Module, a set of experiments are performed on the BlendedUAV dataset in Table V. Here, "baseline" refers to the Cas-MVSNet [26]. The last row in the table represents the proposed AggrMVS. It uses the Optical Flow-guided Feature Extraction Module to map the dynamic link between reference and source image, and the Cross-view Volume Fusion Module is introduced to enhance the interaction among reference volumes.

As shown in Table V, the model performance on the BlendedUAV dataset is improved by the combination of the Optical Flow-guided Feature Extraction Module and Cross-view Volume Fusion Module. Specifically, the Optical Flow-guided Feature Extraction Module and Cross-view Volume Fusion Module gain 8% and 11% improvement in MAE metric. Furthermore, The combination of the two proposed modules produces a 14% improvement in the MAE metric. This reason mainly is the simplicity of the existing UNet or FPN architecture and the lack of emphasis on edges, which affects the effectiveness of the model in reconstructing weakly textured regions. The proposed module establishes an optical flow between reference and source image, explicitly capturing the edge semantics information of the components at different depths, which alleviates the modeling failure for this aspect. Moreover, existing aggregation operations do not take into account the interactions between source volumes, resulting in insufficient estimation of regions such as occlusions. The proposed module enhances the interaction between reference volumes through the attention mechanism to construct more robust cost volumes. In summary, the visual consistency enhancement design achieves superior results with more source volumes by exploring the correlation between reference and source images.

D. Study of the BlendedUAV Dataset

To complement the lack of existing MVS aerial datasets at low altitudes, we reconstruct BlendedUAV based on the general MVS dataset. Quantitative experiments in Table I and qualitative experiments in Fig. 5 demonstrate the usability of this dataset. Some conclusions can be summarised as follows:

1) The quantitative results indicate the satisfactory aerial image depth estimation performance on WHU and BlendedUAV datasets. Besides, BlendedUAV has finer viewpoint variations, and more reference images can get richer reconstruction details. Therefore, BlendedUAV shows greater improvement in Five-View in the experiment.
2) All models achieve similar qualitative results on the WHU dataset. It is difficult to visually assess them (see Fig. 6). Different from this phenomenon, the fine reconstruction ability of the models faces the challenges on the BlendedUAV dataset and can obtain a visual model assessment.
3) Fig. 8 presents the results of fine-tuning the model on WHU and WHU-OMVS datasets after pre-training on BlendedUAV. It indicates that the models with pre-training not only converge faster but also achieve improved final scores. This suggests that the BlendedUAV dataset may have a similar spatial distribution to the existing WHU dataset, despite being collected from different altitudes.

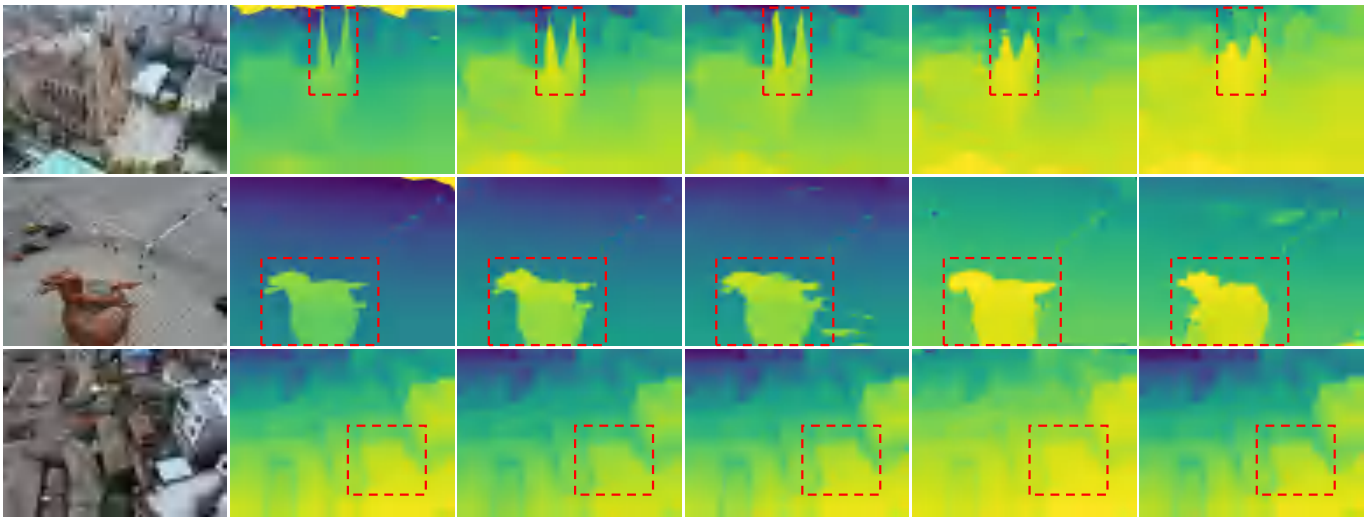

Fig. 5. Comparison of the depth estimation results of the proposed AggrMVS and other methods on the BlendedUAV dataset. The red dashed box indicates the key area of interest.

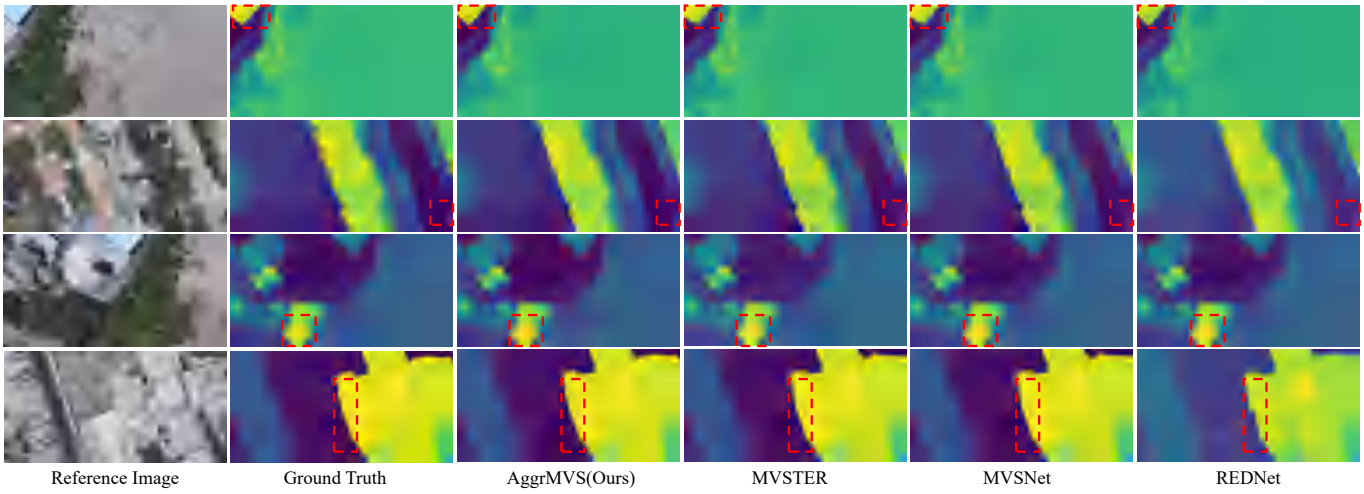

Fig. 6. Comparison of the depth estimation results of the proposed AggrMVS and other methods on the WHU [16] dataset. The red dashed box indicates the key area of interest.

TABLE V
ABLATION RESULTS ON BLENDEDUAV EVALUATION SET. "OFFE" AND "CVFM" REFER TO THE OPTICAL FLOW-GUIDED FEATURE EXTRACTION MODULE AND CROSS-VIEW VOLUME FUSION MODULE RESPECTIVELY. "BASELINE" IS THE ORIGINAL CAS-MVSNET.

Method	**Representation**			**Aggregation**		**Metrics**			
	UNet	FPN	OFFE	Variance	CVFM	MAE↓	<1-Thres(%)↑	<0.6-Thres(%)↑	<3-Interval(%)↑
baseline	✓			✓		0.1225	0.9682	0.9650	0.9663
baseline		✓		✓		0.1204	0.9697	0.9584	0.9677
baseline+OFFE		✓			✓	0.1103	0.9799	0.9772	0.9782
baseline+CVFM		✓	✓	✓		0.1069	0.9829	0.9709	0.9801
baseline+OFFE+CVFM		✓	✓		✓	**0.1032**	**0.9869**	**0.9803**	**0.9863**

Fig. 7. Visual output obtained by the Optical Flow-guided Feature Extraction Module.

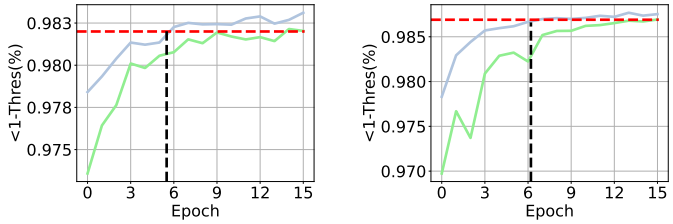

Fig. 8. Comparation results of models with BlendedUAV pre-training on the WHU (left) and WHU-OMVS (right). The black line indicates with pre-training and the green line indicates without pre-training.

E. Study of Point Cloud

To validate the effectiveness of the predicted depth maps by AggrMVS, we perform reconstruction on four representative land cover types. Specifically, AggrMVS first derives the estimated depth of the aerial image on the BlendedUAV dataset. Then, the depth maps are fused using the gipuma method [50]. According to these steps, we can obtain the point cloud reconstruction results. As can be seen from Fig. 9, the proposed AggrMVS performs well on the point cloud generation results. It can show the content of the scene clearly both in general and in detail.

VI. Conclusion

In this paper, we introduce an end-to-end MVS network for aerial image depth estimation, namely AggrMVS. Firstly, the proposed Optical Flow-Guided Feature Extraction Module explicitly captures edge information of different depth components and guides cost volume regularization. Secondly, a Cross-View Volume Fusion Module is proposed to efficiently achieve inter-frame feature fusion for both reference and source image, further improving the aggregation ability of the source volume. The experiment results demonstrate that the AggrMVS outperforms other popular MVS methods for estimating depth in aerial images across several datasets. Additionally, we reconstruct a benchmark dataset from the existing general MVS datasets. It enriches the existing aerial image MVS datasets by introducing diversity in capture altitude, photography mode, and landscape type. Our future plans include developing more robust reconstruction models that can ensure stable reconstruction even in scenarios with varying focal lengths, limited views, or unstructured environments.

Fig. 9. Point clouds results on the BlendedUAV dataset.

References

[1] H. Freeman, E. Schneider, C. H. Kim, M. Lee, and G. Kantor, "3D reconstruction-based seed counting of sorghum panicles for agricultural inspection," in *IEEE Int. Conf. Integr. Technol.*, 2023, pp. 9594–9600.
[2] M. G. Bevilacqua, M. Russo, A. Giordano, and R. Spallone, "3D reconstruction, digital twinning, and virtual reality: Architectural heritage applications," in *Proc. IEEE Conf. Virtual Real. 3D User Interfaces Abstr. Workshops, VRW*, 2022, pp. 92–96.
[3] R. Fan, U. Ozgunalp, Y. Wang, M. Liu, and I. Pitas, "Rethinking road surface 3d reconstruction and pothole detection: From perspective transformation to disparity map segmentation," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 5799–5808, 2021.
[4] 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.
[5] Bentley, "Contextcapture," [EB/OL], https://www.bentley.com/software/contextcapture/. 2018.
[6] Agisoft, "Agisoft metashape," [EB/OL], https://www.agisoft.com/. 2022.
[7] J. L. Schönberger, E. Zheng, J.-M. Frahm, and M. Pollefeys, "Pixelwise view selection for unstructured multi-view stereo," in *Euro. Conf. on Comput. Vis.*, 2016, pp. 501–518.
[8] OpenMVS, "Openmvs: open multi-view stereo reconstruction library," [EB/OL], https://github.com/cdcseacave/openMVS. 2020.
[9] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, "Neural 3D scene reconstruction with the manhattan-world assumption," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2022, pp. 5511–5520.
[10] H. Laga, L. V. Jospin, F. Boussaid, and M. Bennamoun, "A survey on deep learning techniques for stereo-based depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, p. 1738–1764, 2022.
[11] J. Liu, S. Ji, C. Zhang, and Z. Qin, "Evaluation of deep learning based stereo matching methods: From ground to aerial images," *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, vol. XLII–2, p. 593–597, 2018.
[12] Y. Zhang, J. Zhu, and L. Lin, "Multi-view stereo representation revist: Region-aware mvsnet," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2023, pp. 17 376–17 385.
[13] 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.
[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] 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.
[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 Sensing*, 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] Q. Li, M. Gong, Y. Yuan, and Q. Wang, "Symmetrical feature propagation network for hyperspectral image super-resolution," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–12, 2022.
[19] 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.
[20] Q. Li, Y. Yuan, and Q. Wang, "Multi-scale factor joint learning for hyperspectral image super-resolution," *IEEE Transactions on Geoscience and Remote Sensing*, 2023.
[21] 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 Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 2307–2315.
[22] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, *MVSNet: Depth Inference for Unstructured Multi-view Stereo*, 2018, p. 785–801.
[23] M. Mahato, S. Gedam, J. Joglekar, and K. M. Buddhiraju, "Dense stereo matching based on multiobjective fitness function—a genetic algorithm optimization approach for stereo correspondence," *IEEE Trans. Geosci. Remote Sensing*, vol. 57, no. 6, pp. 3341–3353, 2019.
[24] 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.
[25] R. Dey and F. M. Salem, "Gate-variants of gated recurrent unit (gru) neural networks," in *Inter. Midwest Symposium Circuits Sys.*, 2017, pp. 1597–1600.
[26] 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.
[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *Adv. Neural. Inf. Process Syst.*, 2017.
[28] Y. Ding, W. Yuan, Q. Zhu, H. Zhang, X. Liu, Y. Wang, and X. Liu, "Transmvsnet: Global context-aware multi-view stereo network with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2022, pp. 8585–8594.
[29] 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. Eur. Conf. Comput. Vis.*, 2022, pp. 573–591.
[30] C. Cao, X. Ren, and Y. Fu, "Mvsformer: Learning robust image representations via transformers and temperature-based depth for multi-view stereo," *arXiv preprint arXiv:2208.02541*, 2022.
[31] X. Peng, R. S. Feris, X. Wang, and D. N. Metaxas, "Red-net: A recurrent encoder–decoder network for video-based face alignment," *Int. J. Comput. Vis.*, vol. 126, pp. 1103–1119, 2018.
[32] N. Haala, "The landscape of dense image matching algorithms." paper presented at photogrammetric week 2013, wichmann verlag, berlin/offenbach," 2013.
[33] D. Yu, S. Ji, J. Liu, and S. Wei, "Automatic 3D building reconstruction from multi-view aerial images with deep learning," *ISPRS J. Photogramm. Remote Sens*, vol. 171, pp. 155–170, 2021.
[34] W. Jing, Y. Yuan, and Q. Wang, "Dual-field-of-view context aggregation and boundary perception for airport runway extraction," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, pp. 1–12, 2023.
[35] D. Scharstein, H. Hirschmüller, Y. Kitajima, G. Krathwohl, N. Nešić, X. Wang, and P. Westling, *High-Resolution Stereo Datasets with Subpixel-Accurate Ground Truth*, 2014, p. 31–42.
[36] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples," *ACM Transactions on Graphics*, p. 1–13, 2017.
[37] R. Jensen, A. Dahl, G. Vogiatzis, E. Tola, and H. Aanaes, "Large scale multi-view stereopsis evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2014.
[38] T. Schops, J. L. Schonberger, S. Galliani, T. Sattler, K. Schindler, M. Pollefeys, and A. Geiger, "A multi-view stereo benchmark with high-resolution images and multi-camera videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017.
[39] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2020.
[40] W. Jiang, W. Fan, J. Chen, H. Shi, and X. Luo, "Self-supervised cascade training for monocular endoscopic dense depth recovery," in *Chinese Conference on Pattern Recognition and Computer Vision*, 2023, pp. 480–491.
[41] H. Xu, J. Zhang, J. Cai, H. Rezatofighi, F. Yu, D. Tao, and A. Geiger, "Unifying flow, stereo and depth estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023.
[42] H. Xu, Z. Zhou, Y. Wang, W. Kang, B. Sun, H. Li, and Y. Qiao, "Digging into uncertainty in self-supervised multi-view stereo," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6078–6087.
[43] 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. [Online]. Available: http://dx.doi.org/10.1109/cvpr46437.2021.01397
[44] Y. Yao, Z. Luo, S. Li, J. Zhang, Y. Ren, L. Zhou, T. Fang, and L. Quan, "Blendedmvs: A large-scale dataset for generalized multi-view stereo networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2020, pp. 1790–1799.
[45] Z. Luo, L. Zhou, X. Bai, H. Chen, J. Zhang, Y. Yao, S. Li, T. Fang, and L. Quan, "Aslfeat: Learning local features of accurate shape and localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2020.
[46] Z. Zhang, R. Peng, Y. Hu, and R. Wang, "Geomvsnet: Learning multi-view stereo with geometry perception," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2023, pp. 21 508–21 518.
[47] 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.
[48] 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.
[49] 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.
[50] S. Galliani, K. Lasinger, and K. Schindler, "Massively parallel multiview stereopsis by surface normal diffusion," 2015, pp. 873–881.

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.

Yuan Yuan (M'05-SM'09) is currently a Full Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. She has authored or co-authored over 150 papers, including about 100 in reputable journals, such as the IEEE TRANSACTIONS AND PATTERN RECOGNITION, as well as conference papers in CVPR, BMVC, ICIP, and ICASSP. Her current research interests include visual information processing and image/video content analysis.

Qi Wang (M'15-SM'15) 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. https://crabwq.github.io/