A Comparison of Multi-View Stereo Methods for Photogrammetric 3D Reconstruction: From Traditional to Learning-Based Approaches

Yawen Li, George Vosselman, Francesco Nex

Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, Enschede, The Netherlands
(yawen.li, george.vosselman, f.nex)@utwente.nl

Keywords: Photogrammetric 3D reconstruction, MVS, Learning-based, End-to-end, Comparative Evaluation

Abstract

Photogrammetric 3D reconstruction has long relied on traditional Structure-from-Motion (SfM) and Multi-View Stereo (MVS) methods, which provide high accuracy but face challenges in speed and scalability. Recently, learning-based MVS methods have emerged, aiming for faster and more efficient reconstruction. This work presents a comparative evaluation between a representative traditional MVS pipeline (COLMAP) and state-of-the-art learning-based approaches, including geometry-guided methods (MVSNet, PatchmatchNet, MVSAnywhere, MVSFormer++) and end-to-end frameworks (Stereo4D, FoundationStereo, DUSt3R, MASt3R, Fast3R, VGGT). Two experiments were conducted on different aerial scenarios. The first experiment used the MARS-LVIG dataset, where ground-truth 3D reconstruction was provided by LiDAR point clouds. The second experiment used a public scene from the Pix4D official website, with ground truth generated by Pix4Dmapper. We evaluated accuracy, coverage, and runtime across all methods. Experimental results show that although COLMAP can provide reliable and geometrically consistent reconstruction results, it requires more computation time. In cases where traditional methods fail in image registration, learning-based approaches exhibit stronger feature-matching capability and greater robustness. Geometry-guided methods usually require careful dataset preparation and often depend on camera pose or depth priors generated by COLMAP. End-to-end methods such as DUSt3R and VGGT achieve competitive accuracy and reasonable coverage while offering substantially faster reconstruction. However, they exhibit relatively large residuals in 3D reconstruction, particularly in challenging scenarios.

1. Introduction

With the rapid growth of vision-based systems, the ability to reconstruct 3D environments from images has become increasingly important. Accurate and efficient 3D reconstruction is essential for advanced tasks such as scene understanding, embodied intelligence, and large-scale digital twins (Jiang et al., 2024). Among various application domains, photogrammetric 3D reconstruction represents one of the most widely adopted scenarios (Jiang et al., 2021). It derives precise geometric and spatial information from overlapping images, enabling large-scale mapping and metric measurements across aerial scenarios.

3D reconstruction can be generally divided into single-view and multi-view methods (Fácil et al., 2017). Due to its simplicity and flexibility, single-view 3D reconstruction has attracted growing attention in recent years. However, such approaches heavily rely on learned priors and lack explicit geometric constraints, which limits their accuracy, generalization, and metric reliability in real-world scenarios (Kato and Harada, 2019, Sun et al., 2021). In parallel, neural rendering methods such as NeRF (Mildenhall et al., 2021) and 3D Gaussian Splatting (3DGS) (Kerbl et al., 2023) have demonstrated impressive performance in novel view synthesis. Their primary objective is to generate visually plausible images from unseen viewpoints, focusing on visual fidelity rather than geometric accuracy (Gupta et al., 2024). In contrast, Multi-View Stereo (MVS) based methods are explicitly designed to recover metric 3D structure from images. Therefore, this paper focuses on traditional and learning-based MVS frameworks, which are more suitable for high-accuracy aerial photogrammetric reconstruction.

Compared with single-view methods, multi-view reconstruction leverages geometric consistency across multiple viewpoints, which reduces shape ambiguity and improves robustness. These advantages make multi-view reconstruction approaches particularly suitable for high-precision and large-scale photogrammetric applications. Traditional multi-view reconstruction approaches, such as Structure-from-Motion (SfM) and MVS, have made significant progress in the past decades. Representative open-source systems such as COLMAP (Schönberger et al., 2016, Schonberger and Frahm, 2016), OpenMVS (Cernea, 2020), and OpenMVG (Moulon et al., 2016), as well as widely used commercial software such as Pix4Dmapper (Vallet et al., 2012) and Agisoft Metashape (Verhoeven, 2011), have become mainstream tools for photogrammetric 3D reconstruction (Liu et al., 2023). Using only images as input, these systems implement modular pipelines for feature extraction, feature matching, sparse reconstruction, and dense stereo fusion, enabling camera pose estimation and dense 3D reconstruction (Zhao et al., 2021). Nevertheless, they still face notable bottlenecks, including limited real-time performance and reliance on complex multi-stage pipelines.

With the rapid development of large-scale models, deep learning–based methods have emerged as promising alternatives for 3D reconstruction. These approaches eliminate the reliance on hand-crafted components of classical pipelines by directly estimating the 3D structure from images. Existing learning-based frameworks can be broadly categorized into geometry-guided MVS methods and end-to-end reconstruction methods, and they support both two-view and multi-view reconstruction settings. Two-view methods reconstruct 3D structures from image pairs. They are suitable for lightweight and rapid applications but offer limited geometric constraints. Multi-view methods exploit geometric consistency across multiple viewpoints to achieve more accurate and complete scene reconstruction.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-721-2026

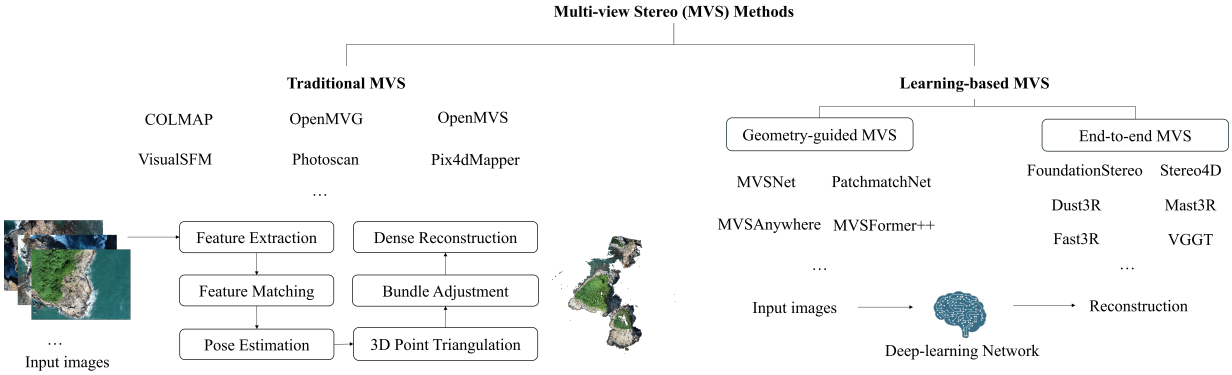


Figure 1. Multi-View Stereo Methods.

Geometry-guided methods, such as MVSNet (Yao et al., 2018), CasMVSNet (Gu et al., 2020), PatchMatchNet (Wang et al., 2021), MVSAnywhere (Izquierdo et al., 2025), and MVSFormer++ (Cao et al., 2024) integrate traditional multi-view geometry with learned cost volumes to predict depth maps efficiently. These methods focus on densification, relying on images and externally estimated camera poses as input.

In contrast, end-to-end models, including FoundationStereo (Wen et al., 2025), Stereo4D (Jin et al., 2025), DUSt3R (Wang et al., 2024), MASt3R (Leroy et al., 2024), VGGT (Wang et al., 2025), and Fast3R (Yang et al., 2025), leverage large-scale vision models to infer 3D structure directly from image collections without explicit geometric modeling. These approaches demonstrate strong capabilities in terms of speed, robustness, and generalization, showing great potential for scalable and real-time photogrammetric reconstruction.

Despite the rapid progress of learning-based MVS methods, few studies (Hermann et al., 2024) have systematically compared them with traditional frameworks in the field of photogrammetry. This work bridges that gap by evaluating learning-based approaches against a representative traditional approach. For experimental evaluation, we use two aerial datasets: the MARS-LVIG dataset with LiDAR-based ground truth (Li et al., 2024), and a public Pix4D dataset with reference point clouds generated by Pix4Dmapper (Pix4D, 2025). Quantitative and qualitative assessments are performed in CloudCompare (Girardeau-Montaut et al., 2016), reporting metrics such as processing time, accuracy, and coverage. Our goal is to provide insights into their strengths and weaknesses and offer useful references for downstream 3D reconstruction tasks.

The remainder of this paper is organized as follows. Section II provides an overview of the MVS methods, including traditional and learning-based MVS methods. Section III discusses the datasets, experimental results, and analysis. Finally, the conclusions and future work are summarized in the last section.

2. Related Works

In this section, we review the traditional and learning-based reconstruction frameworks. Fig. 1 shows the overall 3D reconstruction pipeline. The process starts from a set of aerial images captured from a UAV platform. These aerial images are subsequently processed through two distinct methodological branches: (1) traditional MVS, which is based on classical pipelines, and (2) learning-based MVS, which leverages recent advances in deep learning.

2.1 Traditional MVS Methods

Traditional MVS frameworks take only images as input and typically consist of several well-defined stages, including feature extraction and matching, camera pose estimation, 3D point triangulation, bundle adjustment, and dense reconstruction.

COLMAP (Schönberger et al., 2016, Schonberger and Frahm, 2016) is one of the most representative open source systems. It used incremental SfM combined with global BA optimization and generated high-quality dense point clouds. OpenMVG primarily focused on extracting dense feature matches from input images and used these matches to estimate camera pose and reconstruct sparse point clouds, providing a complete SfM framework (Moulon et al., 2016). OpenMVS focused on dense point cloud processing and surface reconstruction, generating high-quality 3D models through multi-view depth fusion and texture mapping (Cernea, 2020). It can also directly use OpenMVG output as input, achieving more refined reconstruction results.

In the field of aerial photogrammetric mapping, several mature software platforms have been widely adopted for practical applications. Agisoft Metashape (Verhoeven, 2011) and Pix4Dmapper (Vallet et al., 2012) are two widely used commercial software packages that provide highly automated and robust photogrammetric workflows. They integrated aerial triangulation, dense image matching, surface modeling, and texture mapping within a unified framework.

Traditional MVS pipelines rely on accurate image orientation, which necessitates precise registration and parameter tuning. Any failure during this stage can propagate through subsequent reconstruction steps, leading to degraded results. Moreover, they often suffer from limited real-time capabilities, which can be computationally expensive and take hours on large-scale datasets. They are also sensitive to textureless or repetitive regions, illumination changes, and image quality, which often lead to visible artifacts and distorted reconstructions. Such limitations are further amplified in aerial photogrammetry, where high-resolution imagery, wide-area coverage, and stringent accuracy requirements impose additional challenges on both computational efficiency and reconstruction quality.

2.2 Learning-based MVS Methods

Recently, several learning-based approaches have been developed for generating depth maps and reconstructing 3D models. These methods can be broadly categorized into geometry-

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-721-2026 | © Author(s) 2026. CC BY 4.0 License.

Method	Number of Views	Characteristic	Category
COLMAP	N	Classical SfM+MVS pipeline with explicit geometry	Traditional
PatchmatchNet	N	Learned PatchMatch for cost volume–based depth regression	Geomery-Guided
MVSFormer++	N	Transformer-based cost aggregation with explicit geometry	Geomery-Guided
MVSAnywhere	N	Cost Volume Patchifier with view-independent and scale-independent geometry	Geomery-Guided
FoundationStereo	2	Zero-shot generalization	End-to-end
Stereo4D	2	Monocular priors + 4D cost volume with temporal consistency	End-to-end
DUSt3R	2	ViT-based dense correspondence with pointmap regression	End-to-end
MASt3R	2	Extension of DUSt3R with enhanced feature extraction and fast reciprocal matching	End-to-end
Fast3R	N	Efficient and scalable end-to-end matching via positional interpolation	End-to-end
VGGT	N	Large Transformer with global and frame attention	End-to-end

Table 1. Overview and categorization of representative MVS methods considered in this study

guided MVS methods and end-to-end MVS methods. As summarized in Table 1, the number of views indicates whether the method supports two-view (2) or multi-view (N) reconstruction, which directly impacts its scalability to largescale mapping scenarios. The characteristic briefly describes the core mechanism of each method, highlighting whether it relies on explicit geometric modeling, geometry-guided learning, or end-to-end reasoning. The category column classifies methods into three main classes: traditional, geometry-guided, and end-to-end.

2.2.1 Geometry-guided MVS methods

By integrating multi-view geometry into the learning framework, these methods take images together with externally estimated poses as inputs, and construct cost volumes from camera parameters to enable accurate dense depth estimation.

MVSNet (Yao et al., 2018) constructed a cost volume based on the reference image and employed CNNs to learn 3D regularization. To improve memory consumption, R-MVSNet (Yao et al., 2019) replaced MVSNet's depth prediction with a recurrent network to make improvements. Subsequently, many approaches began to explore how to reduce computing resources and improve operating efficiency. Cas-MVSNet (Gu et al., 2020) put forward a cascade cost volume and VA-MVSNet (Yi et al., 2020) introduced a self-adjusting view aggregation to determine the contribution of each input image on a voxel-by-voxel basis. PatchmatchNet (Wang et al., 2021) adapted patchmatch-based propagation to replace heavy 3D convolutions, enabling iterative depth estimation in a coarse-to-fine manner.

More recent transformer-based or hybrid methods, such as MVSFormer, MVSFormer++, and MVSAnywhere, leveraged Vision Tranformers (ViTs) to capture long-range dependencies and enhance cross-view reasoning. MVSFormer (Cao et al., 2023) enhanced MVS depth estimation by fusing pretrained ViT features and employing multi-scale training with hybrid classification–regression for improved robustness. MVSFormer++ (Cao et al., 2024) further integrated DINOv2 (Oquab et al., 2023) features with a proposed side view attention, achieving higher accuracy and generalization in challenging multi-view scenarios. MVSAnywhere (Izquierdo et al., 2025) was a transformer-based MVS framework that introduced a cost volume patchifier, enabling tokenization of cost volumes while fusing single-view ViT features. It employed a view-independent and scale-independent cost volume construction mechanism, achieving state-of-the-art depth accuracy and 3D consistency across arbitrary numbers of input views.

The aforementioned methods are mostly evaluated on indoor or simulated datasets. These scenes typically have decent lighting, limited scale, and clear geometric structure, which simplifies the reconstruction problem. The generalization ability of these methods to large-scale, high-resolution outdoor scenes remains underexplored.

2.2.2 End-to-end MVS methods

End-to-end MVS methods integrate correspondence estimation, pose recovery, and 3D structure generation into a unified learning pipeline using only images. By leveraging large-model architectures, these approaches are more flexible and directly produce dense and coherent 3D reconstructions.

Two-view methods such as Stereo4D and FoundationStereo focus on lightweight and efficient stereo inference, relying on rectified image pairs as input. Stereo4D (Jin et al., 2025) focused on real-time stereo depth estimation for dynamic 3D scenes by jointly modeling spatial and temporal cues, enabling fast and robust 3D reconstruction under motion. FoundationStereo (Wen et al., 2025) leveraged large foundation models to achieve accurate and robust stereo matching with minimal task-specific training, providing strong generalization across diverse scenes.

DUSt3R (Wang et al., 2024) was a milestone in end-to-end 3D reconstruction, enabling direct and efficient point cloud generation from image pairs. It adopted a ViT-based architecture that was pre-trained for cross-view correspondence reasoning. Two input images were first fed into a shared encoder, and their features were then processed by a transformer-based decoder using cross-attention to establish dense correspondences. At the end of the decoder, two independent prediction heads output point maps for both views, with the second point map aligned to the coordinate frame of the first view. MASt3R (Leroy et al., 2024) built upon DUSt3R by introducing a new feature extraction head that produced dense local features and employed a matching loss to enhance correspondence accuracy while maintaining robustness under challenging viewpoint changes. DUSt3R and MASt3R were limited to pairwise reconstruction, requiring additional fusion and geometric post-processing, and became inefficient when extended to multi-view settings.

To overcome these scalability limitations, Fast3R (Yang et al., 2025) leveraged positional interpolation to train on a small number of views and generalized to large-scale multi-view inference. It integrated FlashAttention with parallel training and combined local point maps with global alignment to improve efficiency and consistency. VGGT (Wang et al., 2025) employed a large-scale transformer architecture with a global attention and frame attention mechanism. In addition, a camera-

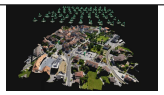
Dataset	Scene	Total images	Groundtruth	Image Resolution	Display
MARS-LVIG	Island	149	LiDAR	5472*3648	
Pix4D Example	Urban area	100	Pix4Dmapper	6000*4000	

Table 2. Datasets used for our tests.

head module was introduced to estimate camera parameters and 3D structure simultaneously.

3. Experiments

3.1 Method Selection

To ensure a fair and representative comparison, this study selects a set of representative MVS methods, including COLMAP, PatchMatchNet, MVSFormer++, MVSAnywhere, FoundationStereo, Stereo4D, DUSt3R, MASt3R, VGGT, and Fast3R. In this study, we use the publicly available pretrained models without any fine-tuning.

3.2 Dataset

To evaluate the performance of different reconstruction paradigms, we use two datasets shown in Table 2. The MARS-LVIG datasets are UAV-based aerial photogrammetry datasets collected in complex environments (Li et al., 2024). These datasets are well-suited for testing reconstruction robustness and scalability due to their high-resolution imagery, large-scale coverage, and complex structures. A challenging island scene is used in this study, with image resolutions of 5472 × 3648. The ground-truth reference is derived from high-precision LiDAR point clouds.

In addition, we include a publicly available Pix4Dmapper official dataset, consisting of 100 aerial frames (Pix4D, 2025). This dataset provides well-structured image acquisition with sufficient overlap, enabling stable and reliable reconstruction. A typical urban scene is used for evaluation, with image resolutions of 6000 × 4000. The ground-truth point clouds are generated using Pix4Dmapper.

For geometry-guided methods, the inputs are derived from COLMAP-generated results, including camera poses, sparse reconstruction, undistorted images, and view selection files. To align the input conditions among different methods, all input images for DUSt3R/MASt3R/Fast3R/VGGT in our experiments were undistorted beforehand. For FoundationStereo and Stereo4D, the input image pairs were rectified beforehand to meet their stereo inference requirements.

DUSt3R, MASt3R, and Fast3R are limited to a maximum input size of 512 pixels, and VGGT requires 518 pixels. To ensure a fair and systematic comparison, all images were rescaled to 512 pixels to the maximum dimension. These rescaled images were then used as input for the rest methods (Wu et al., 2025).

We employed a progressive subset selection strategy, sampling images in increasing numbers (2, 10, 20, 50, and 100 frames). It allows us to evaluate how each method scales with the number of input views and analyze their reconstruction accuracy, completeness, and runtime under different input conditions. For the subsets with 2, 10, and 20 images, the reference point clouds were cropped to match the corresponding coverage area.

3.3 Evaluation Indicators

We conduct a comprehensive and systematic evaluation of representative methods across all key aspects. Specifically, we assess their performance in terms of processing time, reconstruction accuracy, and coverage, ensuring a fair and consistent comparison under the same experimental conditions (Wu et al., 2025).

Processing time (seconds [s]) reflects the overall computational efficiency of each method. It provides a practical measure of algorithm scalability and suitability for real-time and large-scale applications. The reported runtime for each image setting is averaged over three runs.

Accuracy quantifies the geometric closeness between the reconstructed point cloud and the reference ground truth. In this study, it is assessed using two commonly used indicators: Root Mean Square (RMS), which measures the overall deviation between corresponding points, and Mean Distance (MD), which reflects the average alignment accuracy across the entire point cloud. The island aerial images (5472 × 3648 pixels) have an average ground sampling distance (GSD) of 2.5 cm/pixel. After downsampling to 512 × 341 pixels, the effective GSD becomes approximately 0.267 m/pixel. The urban aerial images (6000 × 4000 pixels) have an average ground GSD of 2.41 cm/pixel. After downsampling to 512 × 341 pixels, the effective GSD becomes approximately 0.282 m/pixel. All residual errors (RMS and MD) are therefore reported in normalized units.

Coverage measures the proportion (%) of the reference surface that is successfully reconstructed. It is defined as the percentage of ground truth points whose distances to the reconstructed surface are within a predefined tolerance. In our experiments, a threshold of 1 meter is applied.

All metrics were computed using CloudCompare (Girardeau-Montaut et al., 2016), ensuring a unified and objective evaluation framework across different methods.

3.4 Results

The experimental evaluations were conducted using a single NVIDIA A40 GPU. These evaluation metrics are thoroughly analyzed in our experiments, providing quantitative insights among different approaches. In all tables, the numerically best performance for each metric is shown in bold, and the second-best performance is shown underlined.

3.4.1 Processing time

The processing time was measured separately for each fixed subset of input images (2, 10, 20, 50, and 100 views) selected from the same image block. As shown in Table 3 and Table 4, the runtime of all methods grows rapidly with the increase in the number of input images. Traditional methods such

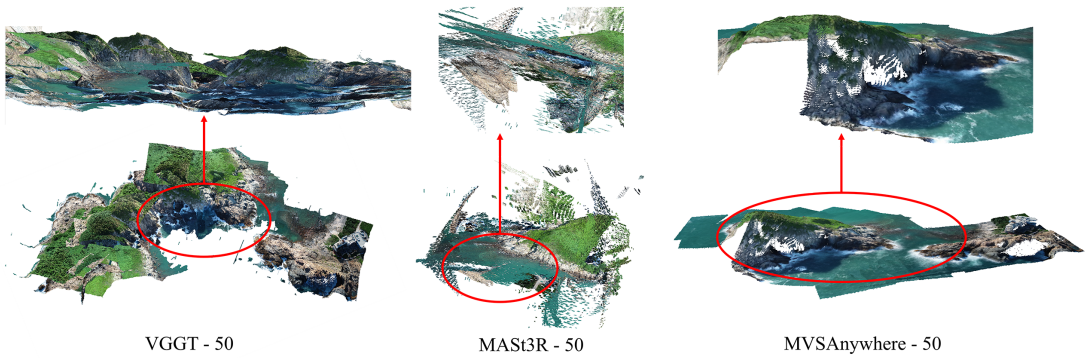


Figure 2. Failed results on Island. The notation “Method–N” (e.g., MASt3R-50, VGGT-50) indicates the reconstruction result obtained from different image samples (N). The red circle marks a local area with discontinuities and layering effects in the island reconstructed point cloud. The zoomed-in view above highlights these artifacts.

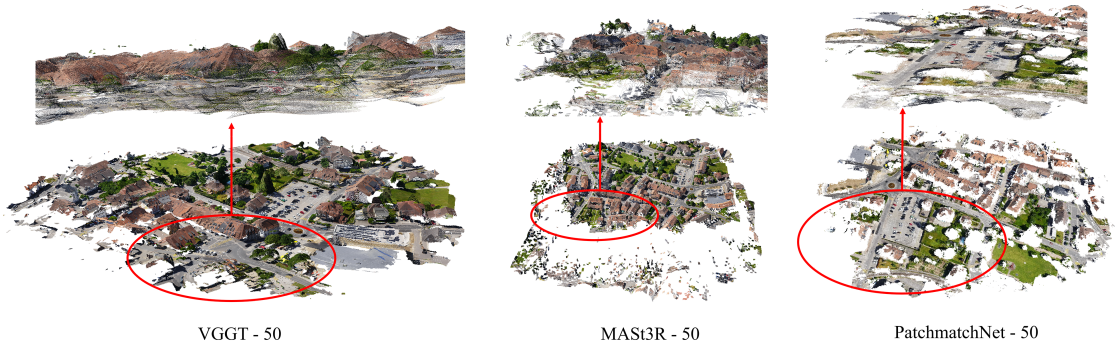


Figure 3. Failed results on Urban. The notation “Method–N” (e.g., MASt3R-50, VGGT-50) indicates the reconstruction result obtained from different image samples (N). The red circle marks a local area with discontinuities and layering effects in the urban reconstructed point cloud. The zoomed-in view above highlights these artifacts.

Methods	**2**	**10**	**20**	**50**	**100**
DUSt3R	3.15	35.91	120.16		
MASt3R	8.60	39.49	68.63	1075.49	
VGGT	1.52	6.10	16.59	28.36	60.79
Fast3R	**0.41**	**0.82**	**1.70**	**5.57**	**16.47**
MVSAnywhere	N/A	N/A	N/A	42.07	100.28
MVSFormer++	15.34	64.99	79.29	80.42	304.28
PatchmatchNet	5.67	15.51	23.76	34.45	238.10
FoundationStereo	2.82				
Stereo4D	0.63				
COLMAP	21.18	86.92	183.24	343.85	1326.38

Table 3. Runtime (s) of different methods on Island across all subsets with varying numbers of input images (2, 10, 20, 50, and 100). “N/A” indicates no result is available.

Methods	**2**	**10**	**20**	**50**	**100**
DUSt3R	3.39	42.93	142.54		
MASt3R	13.16	39.17	88.61	1086.80	
VGGT	1.63	5.42	10.56	29.67	63.06
Fast3R	0.71	**0.83**	**1.72**	**5.47**	**15.59**
MVSAnywhere	N/A	N/A	N/A	57.47	101.51
MVSFormer++	14.57	34.47	55.53	121.09	268.59
PatchmatchNet	4.32	9.91	21.84	121.09	268.59
FoundationStereo	2.81				
Stereo4D	**0.65**				
COLMAP	21.65	104.89	239.41	634.99	1358.68

Table 4. Runtime (s) of different methods on Urban area across all subsets with varying numbers of input images (2, 10, 20, 50, and 100). “N/A” indicates no result is available.

as COLMAP exhibit extremely high computational cost in the dense reconstruction stage, reaching 1326.38s on 100 images for the Island dataset. Moreover, COLMAP frequently faces issues such as incomplete registered frames and inaccurate feature correspondences, which lead to limited scene coverage and potential reconstruction artifacts.

Geometry-guided MVS methods are built on top of COLMAP’s sparse reconstruction pipeline, relying on its camera registration and sparse point cloud as input. When COLMAP fails to generate a complete sparse model, the downstream geometry-guided methods cannot proceed with dense reconstruction. As a result, MVSAnywhere, MVSFormer++, and PatchmatchNet fail to produce any results in cases where COLMAP cannot establish the sparse reconstruction. In Table 3, when the number of input views is small, the large viewpoint gaps between neighboring images cause significant appearance changes and reduce the overlap area, making correspondence estimation unstable.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-721-2026 | © Author(s) 2026. CC BY 4.0 License.

Images	Methods	RMS (GSD)	MD (GSD)	Coverage (%)
2	DUSt3R	**2.88**	**2.21**	**84.89**
	MASt3R	4.53	3.26	71.85
	Fast3R	8.76	6.85	38.26
	VGGT	3.71	2.58	77.77
10	DUSt3R	9.33	7.19	38.85
	Fast3R	12.73	7.72	46.02
	VGGT	**5.73**	**3.82**	**80.65**
20	DUSt3R	9.70	7.38	39.88
	VGGT	**4.46**	**2.92**	**74.32**

Table 5. Quantitative results of different methods on the Island scene under various numbers of input images. GSD = 0.267m/pixel.

Images	Methods	RMS (GSD)	MD (GSD)	Coverage (%)
2	DUSt3R	5.50	4.11	55.26
	MASt3R	**3.87**	**2.62**	**78.53**
	Fast3R	6.74	5.14	45.91
	VGGT	5.60	3.83	65.06
10	DUSt3R	**3.69**	**2.55**	**75.69**
	Fast3R	5.35	3.83	62.09
	VGGT	5.46	3.69	66.71
20	DUSt3R	**4.33**	**3.09**	**68.26**
	Fast3R	5.21	3.69	63.58
	VGGT	5.64	3.72	66.46
	COLMAP	8.55	6.06	52.75

Table 6. Quantitative results of different methods on the Urban scene under various numbers of input images. GSD = 0.282 m/pixel

The island scene contains visually similar textures such as vegetation and rooftops, these repetitive patterns further increase the ambiguity of feature matching under large viewpoint differences. In contrast, with more input views, the accumulated overlap improves the spatial consistency and allows the model to recover a coherent 3D structure.

In contrast, end-to-end MVS methods only require input images and directly infer scene geometry. Among them, Stereo4D and FoundationStereo are fundamentally stereo-based methods, relying on pairwise image matching. While they achieve fast inference on two-view inputs, they are unable to handle multi-view geometry. The limitation highlights their restricted applicability in real-world aerial photogrammetry scenarios. DUSt3R achieves stable results across different subset sizes, but its computational cost increases rapidly as the number of images grows, which limits its scalability. MASt3R exhibits the highest runtime, reaching over 3 hours for 100 images on each dataset, making it unsuitable for large-scale aerial mapping tasks. In contrast, VGGT and Fast3R demonstrate superior efficiency and scalability, with significantly lower runtime even at large image counts, as summarized in Table 3 and Table 4.

3.4.2 Accuracy

For large-scale inputs (50 and 100 images), several methods failed to produce valid point clouds or suffered from severe layering artifacts. Therefore, we only report quantitative results for the 2, 10, and 20 image subsets. Fig. 2 and Fig. 3 illustrate these failed results on the island and urban scenes using different methods. Certain areas are zoomed in to facilitate clearer observation. The reconstructed point clouds exhibit noticeable layering and deformation artifacts, especially over vegetation, rock surfaces, and water regions. These issues mainly arise from unreliable depth estimation and weak geometric constraints in textureless and highly repetitive areas, resulting in distorted 3D structures.

The quantitative accuracy evaluation is reported in Table 5 and Table 6, where RMS and MD are used to measure reconstruction errors. The reconstructed point clouds are first coarsely aligned to the reference model, followed by a fine registration using the ICP algorithm in CloudCompare, with the C2C distance threshold set to 5 m.

FoundationStereo and Stereo4D fail to produce usable results on our datasets due to strong distortions and discontinuities in the reconstructed point clouds. Consequently, we exclude them from the quantitative analysis. Among the end-to-end approaches, DUSt3R achieves the best performance in sparse-view conditions. With only 2 input images, it reaches 2.88 GSD RMS and 2.21 MD on the Island scene, and 5.50 GSD RMS and 4.11 GSD MD on the Urban scene. However, as the number of input views increases, DUSt3R's accuracy drops significantly, particularly in the Island scene, where RMS rises from 2.88 GSD to 9.70 GSD. It indicates reduced geometric consistency and incomplete reconstruction in overlapping regions. VGGT demonstrates superior robustness to increasing view counts, maintaining relatively stable accuracy. MASt3R performs competitively in 2-view cases but struggles to scale. Fast3R generally produces the largest RMS and MD due to insufficient global consistency.

These trends are also visually evident in the maps shown in Fig. 4 and Fig. 5, where reconstruction errors increase significantly in the outer regions of the reconstructed area as the number of input views grows. The errors are particularly pronounced along the scene boundaries and edges, where image overlap is lower and geometric constraints are weaker.

Overall, these results indicate that DUSt3R excels in sparse-view settings, delivering the highest accuracy when image overlap is limited. VGGT shows greater robustness as the number of input views increases. MASt3R and Fast3R exhibit larger RMS and MD values, consistent with the layering and local misalignment artifacts observed in their reconstructed point clouds.

3.4.3 Coverage

The coverage values reported in Table 5 and Table 6 provide a quantitative measure of the proportion of ground truth points effectively reconstructed by different methods.

DUSt3R achieves the highest coverage in sparse-view settings, with coverage rates of 84.89% in the Island scene and 55.26% in the Urban scene using only 2 input images. However, its coverage collapses when more inputs are introduced, particularly on the Island scene.

In contrast, VGGT maintains the most stable and balanced coverage across different input sizes and scenes. Its reconstruction completeness decreases moderately with increasing views. When the number of input images increases, other methods exhibit less favorable scaling behavior in terms of reconstruction accuracy and stability. Fast3R maintains low coverage in general. MASt3R performs well only in sparse-view settings but degrades with more inputs.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-721-2026 | © Author(s) 2026. CC BY 4.0 License.

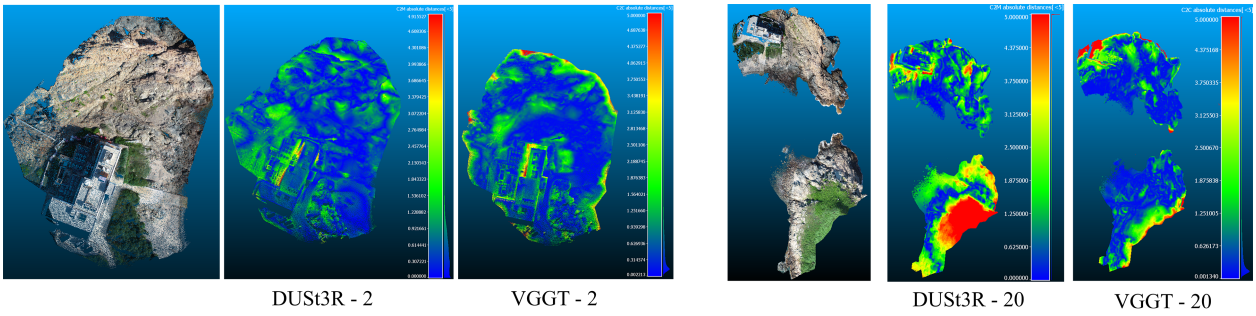


Figure 4. Comparison of reconstruction accuracy on the Island scene using DUSt3R and VGGT with 2 and 20 input images. The colored maps show point-wise distance errors with respect to the LiDAR ground truth. Blue regions indicate small deviations (higher accuracy), whereas red regions denote larger errors. The color scale represents the error magnitude in meters.

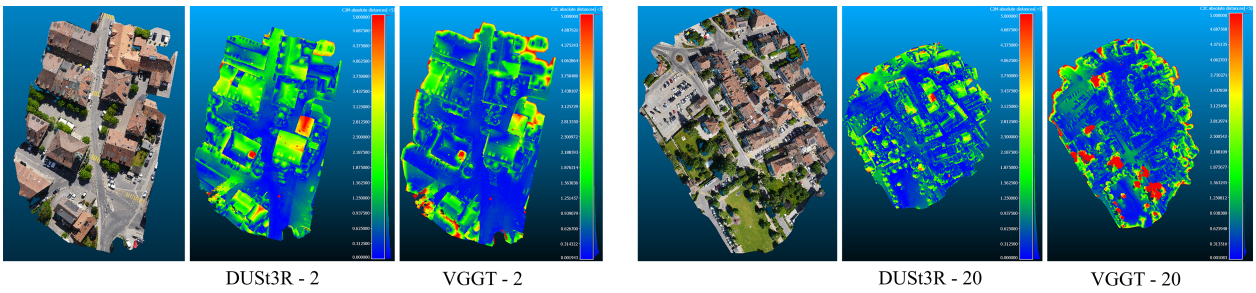


Figure 5. Comparison of reconstruction accuracy on the Urban scene using DUSt3R and VGGT with 2 and 20 input images. The colored maps show point-wise distance errors with respect to the Pix4dmapper-generated ground truth. Blue regions indicate small deviations (higher accuracy), whereas red regions denote larger errors. The color scale represents the error magnitude in meters.

4. Conclusions and Future Work

This paper presents a comparative study of traditional, geometry-guided, and end-to-end learning-based MVS methods for aerial photogrammetry, evaluated on island and urban scenes. The results show clear differences among the three paradigms. Traditional methods, represented by COLMAP, provide relatively sparse reconstruction results under controlled conditions. These also suffer from high computational cost and poor scalability in large-scale outdoor environments. Geometry-guided methods inherit this dependency on upstream sparse reconstruction, leading to frequent failures when image registration is incomplete. In contrast, end-to-end approaches demonstrate higher robustness and better generalization, with VGGT achieving the best balance between reconstruction quality and runtime efficiency. In contrast, end-to-end approaches provide faster reconstruction and better coverage. However, they still produce large reconstruction errors, limiting their practical usability. Common issues, such as layering artifacts and local structural inconsistencies, still limit their practical deployment in high-precision aerial mapping.

In the future, we will focus on integrating pose-aware priors such as SLAM trajectories to enhance large-scale reconstruction accuracy, developing scalable inference strategies to process large image collections efficiently.

References

Cao, C., Ren, X., Fu, Y., 2023. MVSFormer: Multi-View Stereo by Learning Robust Image Features and Temperature-based Depth. *Transactions of Machine Learning Research*.

Cao, C., Ren, X., Fu, Y., 2024. MVSFormer++: Revealing the devil in transformer’s details for multi-view stereo. *arXiv preprint arXiv:2401.11673*.

Cernea, D., 2020. Openmvs: Multi-view stereo reconstruction library.

Fácil, J. M., Concha, A., Montesano, L., Civera, J., 2017. Single-view and multi-view depth fusion. *IEEE Robotics and Automation Letters*, 2(4), 1994–2001.

Girardeau-Montaut, D. et al., 2016. CloudCompare. *France: EDF R&D Telecom ParisTech*, 11(5), 2016.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2495–2504.

Gupta, M., Borrmann, A., Czerniawski, T., 2024. Comparison of 3D Reconstruction between Neural Radiance Fields and Structure-from-Motion-Based Photogrammetry from 360° Videos. *Computing in Civil Engineering 2023*, 429–436.

Hermann, M., Weinmann, M., Nex, F., Stathopoulou, E., Remondino, F., Jutzi, B., Ruf, B., 2024. Depth estimation and 3D reconstruction from UAV-borne imagery: Evaluation on the UseGeo dataset. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 13, 100065.

Izquierdo, S., Sayed, M., Firman, M., Garcia-Hernando, G., Turmukhambetov, D., Civera, J., Mac Aodha, O., Brostow, G., Watson, J., 2025. MVSAnywhere: Zero-Shot Multi-View Stereo. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 11493–11504.

Jiang, S., Jiang, W., Wang, L., 2021. Unmanned Aerial Vehicle-Based Photogrammetric 3D Mapping: A survey of techniques, applications, and challenges. *IEEE Geoscience and Remote Sensing Magazine*, 10(2), 135–171.

Jiang, S., You, K., Li, Y., Weng, D., Chen, W., 2024. 3D reconstruction of spherical images: a review of techniques, applications, and prospects. *Geo-spatial Information Science*, 27(6), 1959–1988.

Jin, L., Tucker, R., Li, Z., Fouhey, D., Snavely, N., Holynski, A., 2025. Stereo4D: Learning How Things Move in 3D from Internet Stereo Videos. *Proceedings of the Computer Vision and Pattern Recognition Conference*.

Kato, H., Harada, T., 2019. Learning view priors for single-view 3D reconstruction. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9778–9787.

Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G., 2023. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4), 139–1.

Leroy, V., Cabon, Y., Revaud, J., 2024. Grounding image matching in 3d with MASt3r. *European Conference on Computer Vision*, Springer, 71–91.

Li, H., Zou, Y., Chen, N., Lin, J., Liu, X., Xu, W., Zheng, C., Li, R., He, D., Kong, F. et al., 2024. MARS-LVIG dataset: A multi-sensor aerial robots SLAM dataset for LiDAR-visual-inertial-GNSS fusion. *The International Journal of Robotics Research*, 43(8), 1114–1127.

Liu, J., Gao, J., Ji, S., Zeng, C., Zhang, S., Gong, J., 2023. Deep learning based multi-view stereo matching and 3D scene reconstruction from oblique aerial images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 204, 42–60.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1), 99–106.

Moulon, P., Monasse, P., Perrot, R., Marlet, R., 2016. OpenMVG: Open multiple view geometry. *International Workshop on Reproducible Research in Pattern Recognition*, Springer, 60–74.

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Howes, R., Huang, P.-Y., Xu, H., Sharma, V., Li, S.-W., Galuba, W., Rabbat, M., Assran, M., Ballas, N., Synnaeve, G., Misra, I., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2023. DINOv2: Learning Robust Visual Features without Supervision.

Pix4D, 2025. Example datasets – Pix4Dmatic. https://support.pix4d.com/hc/en-us/articles/360048957691.

Schonberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European conference on computer vision*, Springer, 501–518.

Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H., 2021. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 15598–15607.

Vallet, J., Panissod, F., Strecha, C., Tracol, M., 2012. Photogrammetric performance of an ultra light weight swinglet" UAV". *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 38, 253–258.

Verhoeven, G., 2011. Taking computer vision aloft–archaeological three-dimensional reconstructions from aerial photographs with photoscan. *Archaeological prospection*, 18(1), 67–73.

Wang, F., Galliani, S., Vogel, C., Speciale, P., Pollefeys, M., 2021. PatchmatchNet: Learned multi-view patchmatch stereo. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14194–14203.

Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D., 2025. VGGT: Visual geometry grounded transformer. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5294–5306.

Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J., 2024. DUSt3r: Geometric 3d vision made easy. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wen, B., Trepte, M., Aribido, J., Kautz, J., Gallo, O., Birchfield, S., 2025. Foundationstereo: Zero-shot stereo matching. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 5249–5260.

Wu, X., Landgraf, S., Ulrich, M., Qin, R., 2025. An Evaluation of DUSt3R/MASt3R/VGGT 3D Reconstruction on Photogrammetric Aerial Blocks. *arXiv preprint arXiv:2507.14798*.

Yang, J., Sax, A., Liang, K. J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M., 2025. Fast3R: Towards 3d reconstruction of 1000+ images in one forward pass. *Proceedings of the Computer Vision and Pattern Recognition Conference*, 21924–21935.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth inference for unstructured multi-view stereo. *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent mvsnet for high-resolution multi-view stereo depth inference. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5525–5534.

Yi, H., Wei, Z., Ding, M., Zhang, R., Chen, Y., Wang, G., Tai, Y.-W., 2020. Pyramid multi-view stereo net with self-adaptive view aggregation. *European conference on computer vision*, Springer, 766–782.

Zhao, Y., Chen, L., Zhang, X., Xu, S., Bu, S., Jiang, H., Han, P., Li, K., Wan, G., 2021. RTSfM: real-time structure from motion for mosaicing and DSM mapping of sequential aerial images with low overlap. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–15.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-721-2026 | © Author(s) 2026. CC BY 4.0 License.