Evaluating the Performance of 3D Vision Foundation Models for DSM Reconstruction from Satellite Images

Liupeng Su[1], Yuhao Ye[1], Han Hu[1,*], Zeyuan Dai[2,3], Qianrui Guo[4], Heyi Li[4], Yulin Ding[1], Qing Zhu[1]

[1] Faculty of Geosciences and Engineering, Southwest Jiaotong University, Chengdu 611756, Sichuan, China
[2] Department of Military Oceanography and Hydrography and Cartography, Dalian Naval Academy, Dalian 116018, China
[3] Key Laboratory of Hydrographic Surveying and Mapping of PLA, Dalian Naval Academy, Dalian 116018, China
[4] Institute of Remote Sensing Satelite, China Academy of Space Technology, Beijing 100094, China
(yueyuebird@my.swjtu.edu.cn, yyh2292368993@my.swjtu.edu.cn, han.hu@swjtu.edu.cn, zeyuan_dai@zju.edu.cn, guoqianrui1005@163.com, hey1li@126.com, dingyulin@swjtu.edu.cn, zhuqing@swjtu.edu.cn)

Keywords: 3D Vision Foundation Models (3D VFMs), Satellite Imagery, Multi-view Stereo, 3D Reconstruction,Digital Surface Model (DSM).

Abstract

Three-dimensional (3D) reconstruction from satellite imagery is a critical research topic in the fields of remote sensing and geoinformation science. Although 3D Vision Foundation Models (3D VFMs) have demonstrated remarkable performance in reconstructing natural scenes, their capability to handle high-resolution satellite imagery has not been systematically evaluated. This study presents a comprehensive assessment of seven representative 3D VFMs for satellite-based 3D reconstruction and integrates four point-cloud alignment strategies. Rigorous comparisons were conducted against high-precision LiDAR-derived Digital Surface Models (DSMs) using two publicly available multi-view satellite datasets–WHU-TLC and MVS3D. The results show that Depth Anything V2 (DAV2) combined with an affine alignment strategy achieves the best overall performance among the evaluated methods. On the MVS3DM dataset, the reconstructed DSM achieves a Median Absolute Error(MedAE) of 1.693 m, a Root Mean Square Error (RMSE) of 3.649 m, and competitive reconstruction accuracy compared with several traditional photogrammetric pipelines. In contrast, on the lower-resolution WHU-TLC dataset, all 3D VFMs exhibited notable performance degradation, and the reconstructed results showed limited practical value, revealing persistent generalization challenges for current models in low-resolution scenarios. Overall, this study systematically quantifies the performance of 3D VFMs in satellite image-based 3D reconstruction, confirming their strong potential for high-resolution satellite applications and providing valuable insights for enhancing model robustness and generalization across complex urban and low-resolution environments.

1. Introduction

Automatically reconstructing large-scale, high-precision DSM from stereo or multi-view imagery remains a core challenge in photogrammetry and remote sensing. With continuous advances in satellite sensors and computer vision technologies, the generation of sub-meter resolution DSM from satellite imagery has become increasingly feasible (Hirschmüller and Hirschmüller, 2008, Rottensteiner et al., 2012, Li et al., 2023). Consequently, a variety of automated workflows have emerged, ranging from traditional stereo or multi-view geometric methods (de Franchis et al., 2014b, de Franchis et al., 2014a, Youssefi et al., 2020) to deep learning-based approaches (He et al., 2022, Li et al., 2023, Gao et al., 2023, Wei et al., 2025). However, existing methods still encounter difficulties under conditions such as large variations in viewpoint and illumination, sparse or repetitive textures, haze or shadow occlusion, cross-sensor radiometric inconsistencies, and seasonal changes. These limitations often lead to DSM artifacts, including voids, discontinuities, and blurred edges.

The recent emergence of 3D VFMs has demonstrated remarkable cross-task generalization and transfer capabilities in computer vision. Without requiring external camera pose priors, these models can directly generate high-quality depth maps or point clouds from single or multiple images (Wang et al., 2024, Keetha et al., 2026). They are typically pre-trained in a self-supervised or weakly supervised manner on large-scale real and synthetic datasets , thereby offering advantages in scale robustness, boundary fidelity, and cross-scene generalization (Yang et al., 2024). Moreover, through lightweight fine-tuning on target-domain data or by applying rigid or similarity transformations to align predictions with reference ground truth, these models can effectively mitigate scale–translation ambiguities while improving measurement accuracy. Their performance on multiple public benchmarks now approaches, and in some cases matches, that of traditional pipelines relying on real camera poses (Yang et al., 2025).

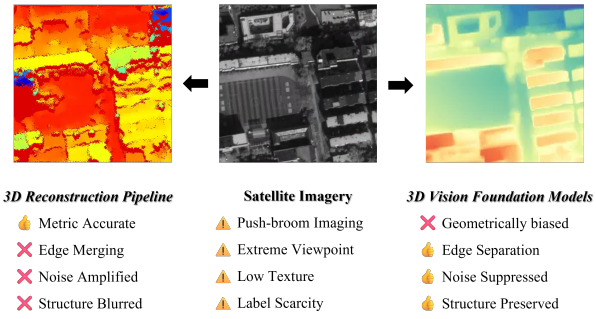


Figure 1. Illustration of the challenges inherent in satellite imagery and the performance differences between traditional reconstruction pipelines and 3D VFMs.

* Corresponding author

However, the potential of 3D VFMs in satellite image–based 3D reconstruction remains insufficiently evaluated. In practical applications, as well as in commonly used datasets, substantial variations exist in sensor types, radiometric properties, field of view, imaging geometry, and baseline length. At the same time, the scarcity of high-precision ground truth, the subjectivity involved in manually selecting alignment control points, and the inconsistent performance of different methods further complicate the tasks of metrological alignment and comparative evaluation of 3D VFMs on satellite imagery. An overview of these challenges and a qualitative comparison between traditional 3D reconstruction pipelines and 3D VFMs is presented in Fig. 1.

This study aims to systematically evaluate the applicability of 3D VFMs in satellite image–based 3D reconstruction and to assess their potential for generating high-quality DSMs. The main contributions of this work are summarized as follows: (1) A reusable and generalizable workflow is proposed that integrates raw satellite imagery, corresponding Rational Polynomial Coefficients (RPCs), and reference DSM. This framework establishes a unified procedure for the geometric alignment and quantitative evaluation of outputs generated by diverse 3D VFMs; (2) Comprehensive comparative experiments were conducted using multiple alignment strategies applied to seven representative 3D VFMs. The experiments encompass three typical urban scene categories—Single-Building, High-Rise Urban, and Dense Urban—and two representative terrain types—Plains and Mountainous regions—ensuring a broad evaluation across diverse landscape conditions; (3) A critical analysis of the evaluation results reveals both the strengths and current limitations of 3D VFMs in satellite-based 3D reconstruction. Furthermore, the study identifies potential research directions to guide future advancements in model robustness, cross-scene generalization, and high-precision geospatial reconstruction..

2. Related Work

Pipelines for Satellite Image–Based 3D Reconstruction. Numerous methods have been proposed for 3D reconstruction from satellite imagery, ranging from traditional stereo matching algorithms to deep learning–based approaches. These methods generally consist of several key stages: image preprocessing, image matching, 3D reconstruction, DSM generation, and refinement. Traditional pipelines frequently rely on the Semi-Global Matching (SGM) algorithm (Hirschmüller and Hirschmüller, 2008) to generate disparity maps from stereo pairs, which are then triangulated using RPCs to obtain 3D points (de Franchis et al., 2014b, Youssefi et al., 2020). More recent works (He et al., 2022, Li et al., 2023) employ deep learning–based feature extraction and matching networks to produce more robust and accurate disparity maps. Compared with conventional MVSNets (Yao et al., 2018, Hu et al., 2023), SatMVS (Gao et al., 2023) introduces an RPC warping module to account for the unique imaging geometry of satellite sensors and directly estimates height maps instead of depth maps, which are subsequently projected into point clouds for DSM generation. Building on this, TS-SatMVSNet (Wei et al., 2025) incorporates slope maps to guide height-map estimation and refine the resulting DSMs. Nevertheless, all these methods depend on accurate camera poses, and the quality of the generated DSMs remains highly sensitive to pose accuracy.

3D Vision Foundation Models. 3D VFMs have recently attracted significant attention in the computer vision community due to their impressive generalization capabilities across diverse 3D vision tasks. They are capable of estimating depth maps, point maps, or even camera poses from single-view or limited multi-view imagery. By employing end-to-end training, these methods eliminate the sequential dependencies inherent in the traditional Structure-from-Motion (SfM) workflow, thereby reducing the accumulation of errors and noise. *Monocular foundation models*, such as MiDaS (Ranftl et al., 2022) and its successor DAv2 (Yang et al., 2024), achieve cross-scene depth estimation through large-scale self-supervised pretraining, enabling robust generalization across diverse visual domains. *Stereo-based foundation models*, which take stereo pairs as input, jointly regress dense depth maps and relative camera poses. The pioneering DUSt3R (Wang et al., 2024) introduces a Transformer-based architecture that learns feature correspondences across multiple views and employs a confidence-weighted aggregation scheme for feature fusion. Building upon this foundation, MUSt3R (Cabon et al., 2025) further improves the accuracy and robustness of multi-scale dense correspondence matching, enabling more reliable depth estimation across complex scenes. To address computational efficiency, Fast3R (Yang et al., 2025) adopts a lightweight network design combined with knowledge distillation, achieving significantly faster reconstruction while maintaining competitive accuracy. *Multi-view foundation models* extend the input to multiple images and emphasize long-range cross-view attention and global consistency. VGGT (Wang et al., 2025) reduces the reliance on 3D geometric optimization during post-processing compared with DUSt3R, predicting a complete set of 3D attributes—including camera parameters, depth maps, point cloud maps, and 3D point trajectories—within seconds. π3 (Wang et al., 2026) overcomes the limitations of fixed-viewpoint assumptions by adopting permutation-equivariant architectures. It predicts 3D properties without requiring a reference frame, demonstrating strong robustness to input-sequence variations and high scalability. MapAnything (Keetha et al., 2026) accepts optional geometric priors (e.g., camera parameters, depth) along with image data and directly regresses depth maps, camera poses, ray directions, metric scaling factors, and camera intrinsics, thereby enabling a globally consistent and measurable framework. However, to the best of our knowledge, these methods have not yet been trained or evaluated on satellite imagery, and their potential applications and limitations in this domain remain largely unexplored.

VFM-Optimized Workflows for 3D Reconstruction. Leveraging the strong cross-scene generalization and transfer capabilities of 3D VFMs, several recent studies have attempted to integrate them into 3D reconstruction pipelines to enhance robustness and accuracy under complex conditions. (Lin et al., 2025) utilized a low-cost LiDAR sensor as a prompt to guide the depth estimation of a pre-trained VFM, achieving high-resolution and metrically accurate depth prediction. (Cheng et al., 2025, Wen et al., 2025) employed monocular VFMs to provide initial depth estimates and dense feature representations for stereo matching networks, thereby improving the accuracy and robustness of stereo depth estimation in challenging environments. Their work also established a benchmark stereo framework that demonstrates strong generalization across diverse datasets. (Wu et al., 2025) evaluated several 3D VFMs on aerial imagery, demonstrating their potential for accurate 3D reconstruction from sparse image collections, while also revealing limitations related to accuracy and scale ambiguity. However, these existing studies primarily focus on ground-based or aerial RGB imagery rather than satellite data, which differ sub-

stantially in imaging geometry and radiometric characteristics. Thus, the potential of 3D VFMs for satellite-based 3D reconstruction remains largely unexplored. Our work aims to address this gap by systematically evaluating the performance of 3D VFMs on satellite imagery and exploring their potential for generating high-quality Digital Surface Models (DSMs).

3. Evaluation Workflow

We first provide an overview of the proposed workflow, as illustrated in Fig. 2. Subsequently, each component is described in detail in the following sections, including data preparation, alignment strategies, post-processing, and accuracy evaluation.

3.1 Overview

The procedure is as follows. First, radiometric normalization and geometric correction are applied to the original imagery to ensure spectral and spatial consistency. Second, the reference DSM is projected into a pixel-aligned point cloud to generate a corresponding height map. Third, a robust transformation model based on RANSAC automatically selects control points to spatially align the depth map or point cloud generated by the 3D VFM with the reference height map. Finally, quantitative accuracy evaluation is conducted through differential analysis with the benchmark DSM, systematically assessing the performance of the 3D VFM in satellite-based 3D reconstruction.

3.2 Data Preparation

3.2.1 Image Pair Selection The accuracy of satellite-based 3D reconstruction is influenced by multiple factors, including convergence angle, incidence angle, spatial resolution, solar elevation angle, solar azimuth angle, and scene variability (Qin, 2019). In stereo satellite imagery, designing appropriate intersection angles while maintaining consistent spatial resolution enables nearly simultaneous image acquisition, thereby meeting the requirements for high-precision 3D reconstruction. Consequently, all images can be used as input. However, as the MVS3D dataset (Bosch et al., 2016) contains multi-temporal imagery, a generalized image-pairing strategy is adopted to ensure geometric and radiometric consistency: (1) images are sorted by decreasing convergence angle (15°–25° recommended); (2) filtered by increasing solar elevation and azimuth differences; and (3) ranked by acquisition time from closest to farthest. The top n images are then selected to balance perspective diversity and radiometric compatibility.

3.2.2 BA and Inverse RPC When performing 3D reconstruction from satellite imagery, precise camera poses are essential for generating high-quality DSMs. However, satellite images typically provide only approximate RPC parameters, which may contain systematic errors and inconsistencies. To improve camera pose accuracy, we perform bundle adjustment (BA) based on affine transformations in image space (Grodecki and Dial, 2003). In addition, a terrain-independent inverse RPC approach is adopted, which utilizes a virtual control grid to compute inverse projection parameters. As a result, the optimized RPCs can be converted into inverse RPCs, enabling high-precision mapping from image coordinates to geographic coordinates.

3.2.3 Radiometric Pre-processing Compared with aerial or ground imagery, satellite imagery exhibits distinct radiometric characteristics. It is typically acquired as panchromatic (PAN) imagery, consisting of single-band radiometrically calibrated data with 10–16 bit depth, high dynamic range, and a high signal-to-noise ratio. To address the resulting radiometric inconsistencies, we adopt a preprocessing strategy similar to that proposed by (Gao et al., 2023). Specifically, pixel values are rescaled to the 0–255 range using percentage-cutoff linear stretching (PCTL). The processed image is then replicated into three channels to satisfy the input requirements of 3D VFMs.

3.2.4 Height Map Generation A height map is a two-dimensional elevation representation generated by projecting a DSM or 3D point cloud data onto the image plane using RPC parameters, where each pixel records the maximum surface elevation value at its corresponding location. The resulting height map establishes a one-to-one correspondence between pixels in the image coordinate system and those in the original imagery, enabling precise pixel-level alignment between the image and elevation domains. This alignment provides a unified spatial reference framework, forming a reliable foundation for subsequent alignment of depth maps or point maps generated by 3D VFMs, as well as for the quantitative accuracy evaluation of the alignment results.

3.3 Alignment Strategies

Based on the output type of each model, targeted alignment strategies are designed. For depth maps or point maps data containing only Z-axis components, linear or affine alignment methods are applied to ensure consistent elevation mapping. When processing 3D point map outputs with full spatial coordinate information, rigid alignment strategies are adopted to restore overall spatial consistency. All alignment procedures are implemented within the RANSAC (Random Sample Consensus) robust estimation framework, which automatically selects reliable control points and computes optimal transformation parameters, effectively reducing errors caused by outlier correspondences.

3.3.1 Linear Alignment Since 3D VFMs are typically trained using relative depth or disparity supervision rather than metric depth with real-world scale, their predictions inherently exhibit scale ambiguity. This ambiguity originates from the intrinsic underdetermination of monocular depth estimation: the absolute scene scale cannot be recovered from a single image without external geometric constraints.

In classical stereo geometry, depth is related to disparity by $d = (f \cdot B)/p$, where d denotes depth, f is the focal length, B is the physical baseline, and p represents disparity. However, this formulation assumes a known stereo configuration. In monocular depth estimation, there is no physical baseline B, and thus no explicitly defined $f \cdot B$ term. Instead, monocular models learn depth up to an unknown global scale factor, since the quantities analogous to f and B are implicitly absorbed during training when supervision is based on relative or normalized depth representations.

Consequently, the predicted depth $\hat{D}$ is related to the true metric depth D by an affine transformation:

$$D = s \cdot \hat{D} + t,$$

where s and t denote the global scale and translation parameters, respectively.

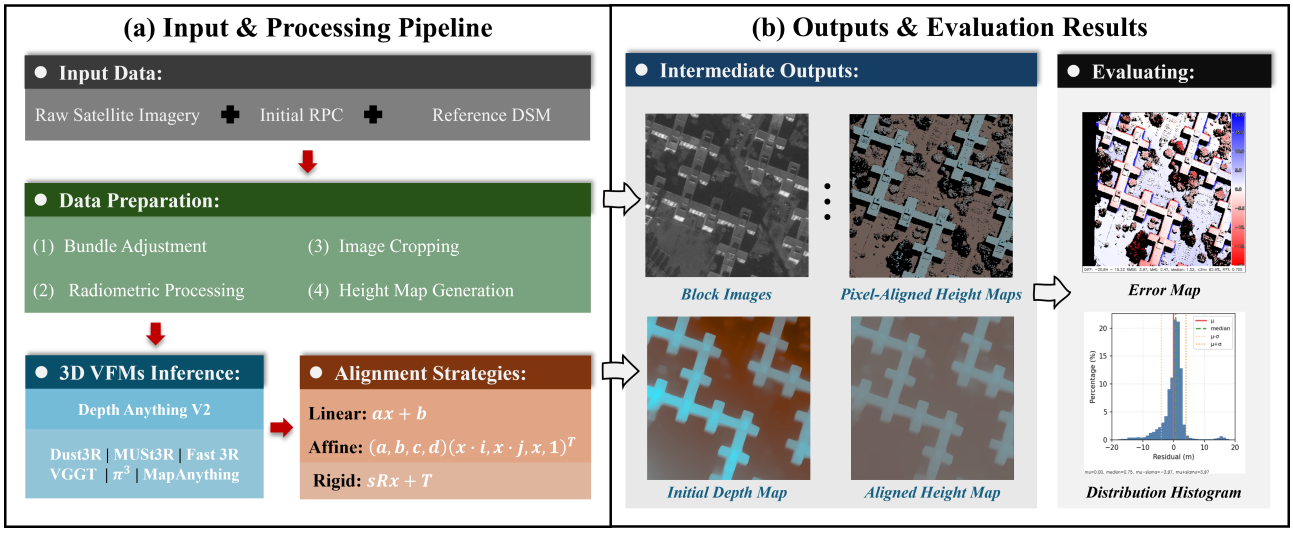


Figure 2. Overview of the proposed alignment evaluation framework. (a) The input and processing pipeline integrates raw satellite imagery, corresponding Rational Polynomial Coefficients (RPCs), and reference Digital Surface Models (DSMs) for data preparation, 3D Vision Foundation Model (3D VFM) inference, alignment, and quantitative evaluation. (b) The outputs and evaluation results—including pixel-aligned height maps, aligned depth maps, error maps, and residual histograms—are used to assess geometric consistency and accuracy across different alignment strategies.

To resolve this ambiguity and recover metric height, a linear alignment strategy is adopted following Wang et al. (2025). Let D_{pred} and D_{ref} denote the predicted and reference height maps, respectively. Using RANSAC, N pixel pairs are sampled as control points to iteratively estimate the optimal scale factor s and translation t by minimizing:

$$(s,t) = \arg\min_{s,t} \sum_{i=1}^{N} \|s \cdot D_{\text{pred}}(x_i, y_i) + t - D_{\text{ref}}(x_i, y_i)\|^2,$$

where (x_i, y_i) are the pixel coordinates of the selected control points in both the predicted depth map and the reference height map.

3.3.2 Affine Alignment Unlike the linear alignment assumption, the affine alignment model directly transforms the depth maps generated by 3D VFMs into geographically meaningful height maps or DSMs, establishing a geometric mapping between image-plane and geographic coordinates. For a pinhole camera model, the mapping between image coordinates (u, v) and camera coordinates (X_c, Y_c, Z_c) can be expressed as:

$$Z_c \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} \Rightarrow \begin{cases} X_c = \frac{u}{f_x} \\ Y_c = \frac{v}{f_y} \\ Z_c = D \end{cases} \quad (1)$$

Furthermore, assuming that a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{t}$ exist between the camera and world coordinate systems, the relationship between the camera coordinates (X_c, Y_c, Z_c) and the world coordinates (X_w, Y_w, Z_w) can be expressed as:

$$[X_w, Y_w, Z_w, 1]^T = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ 0 & 1 \end{bmatrix} [X_c, Y_c, Z_c, 1]^T \quad (2)$$

By combining Eq. (1) and Eq. (2), the mapping relationship between the depth value D in the camera coordinate system and the elevation value Z_w in the world coordinate system can be expressed as:

$$\begin{aligned} Z_w &= R_{31}X_c + R_{32}Y_c + R_{33}Z_c + T_3 \\ &= \frac{R_{31}}{f_x} D \cdot u + \frac{R_{32}}{f_y} D \cdot v + R_{33} D + T_3 \end{aligned} \quad (3)$$

Previous studies have demonstrated that, for satellite imagery, the RPC model can be approximated by a pinhole camera model within a local area of approximately $1km^2$, achieving high geometric accuracy (Zhang et al., 2019). Therefore, to align the depth map generated by a 3D VFM with the georeferenced elevation map, the aforementioned affine alignment strategy can be adopted. For simplicity, define $a = \frac{R_{31}}{f_x}$, $b = \frac{R_{32}}{f_y}$, $c = R_{33}$, and $d = T_3$, which are treated as constant parameters. The coefficients (a, b, c, d) are referred to as affine transformation parameters, describing the linear mapping relationship between the depth map and the elevation map. This process can be formulated as the following minimization problem:

$$\begin{aligned} (a,b,c,d) = \arg\min_{a,b,c,d} \sum_{i=1}^{N} \big\| & D_{\text{pred}}(x_i, y_i) \cdot (a\, x_i \\ & + b\, y_i + c) + d - D_{\text{ref}}(x_i, y_i) \big\|_2 \end{aligned} \quad (4)$$

3.3.3 Rigid Alignment For the $(H, W, 3)$ point map output by the 3D VFM, each point's three-dimensional coordinates are expressed in the camera coordinate system, whereas the reference DSM data derived from satellite imagery is defined in the geographic coordinate system. According to Eq. (2), the relationship between these two coordinate systems can be represented by a combination of rotation and translation transformations. This optimization problem can be expressed as:

$$\begin{aligned} (\mathbf{R},\ \mathbf{T},\ \mathbf{S}) = \arg\min_{\mathbf{R},\,\mathbf{T},\,\mathbf{S}} \sum_{i=1}^{N} \big\| & \mathbf{S}\,\mathbf{R}\,\mathbf{P}_{\text{pred}}(x_i, y_i) \\ & + \mathbf{T} - \mathbf{P}_{\text{ref}}(x_i, y_i) \big\|_2^2 \end{aligned} \quad (5)$$

Among these, $\mathbf{R}$ denotes the rotation matrix, $\mathbf{T}$ represents the translation vector, and $\mathbf{S}$ signifies the scaling matrix.In the subsequent implementation, the single-scale version is referred to as the Rigid Strategy, whereas the version with independent scaling factors along the x, y, and z axes is denoted as the Rigid* Strategy.

3.4 Accuracy Evaluation

After completing the alignment, a point-to-point error statistical method is employed to quantitatively evaluate the accuracy of the results. Based on the registration outcomes, elevation differences at corresponding pixel locations are calculated to generate an error distribution map. Subsequent statistical analysis of this map produces quantitative metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Median Absolute Error(MedAE), Percentage of Absolute Error within a threshold (PAE) , and the Coefficient of Determination (R^2).

4. Experimental and Results

4.1 Experimental Setup

4.1.1 Datasets We utilized two publicly available stereo satellite image datasets for evaluation: MVS3D (Bosch et al., 2016) and WHU-TLC (Gao et al., 2023). The MVS3D dataset consists of 50 panchromatic images captured by the WorldView-3 satellite, featuring a spatial resolution of 0.31 m. It covers complex urban environments and is well suited for high-precision 3D reconstruction tasks. The WHU-TLC dataset is constructed from ZY-3 satellite three-line-array imagery with a spatial resolution of 2.1 m. It contains multi-view observations from both mountainous and urban regions, effectively representing diverse topographic and elevation characteristics. Both datasets provide high-precision airborne LiDAR data as ground truth, offering a reliable benchmark for evaluating model accuracy.

4.1.2 Implementation Details This study systematically evaluated seven representative 3D VFMs, including **Depth Anything v2**, **DUSt3R**, **MUSt3R**, **Fast3R**, **VGGT**, **π^3**, and **MapAnything**. Due to varying input size requirements among these models, all satellite images were cropped into 512×512 pixel patches. For models requiring input dimensions divisible by 14, the images were first scaled to 518×518 pixels and then resized back to 512×512 after inference. To ensure diverse evaluation conditions, three representative urban scenes were selected from the MVS3D dataset, while two representative terrain types were chosen from the WHU-TLC dataset, as illustrated in Fig. 3. For the MVS3D dataset, five images per scene were selected according to the ranking strategy described in Section 3.2.1. In contrast, since each WHU-TLC scene provides only three available viewing angles, all images were used for testing. All models were executed on a single NVIDIA RTX 4090 GPU using their default hyperparameters and official pre-trained weights. For comparison, a baseline DSM was reconstructed using Agisoft Metashape (Agisoft LLC, 2022), with the matching accuracy set to *Very High* to ensure that the DSM resolution matched the native image resolution.

4.2 Patch-Level Alignment Evaluation

This section analyzes the performance variations of different 3D VFMs under various alignment strategies across diverse scenarios. Given that each model has distinct requirements

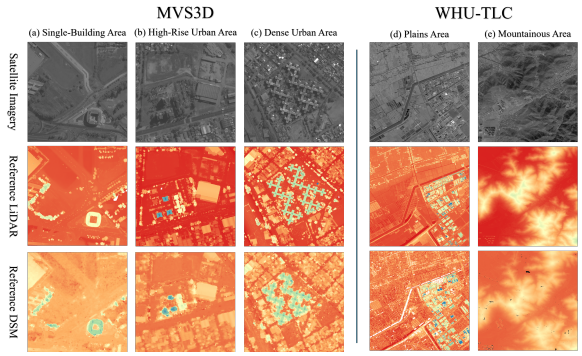


Figure 3. Representative experimental regions selected from the MVS3D and WHU-TLC datasets.

for the number of input images, the experiments set the input image count to maximize each model’s multi-view reconstruction capability—one image for Depth Anything v2, two for MUSt3R and DUSt3R, and multiple for Fast3R, MapAnything, and π^3—while fully utilizing the available image information. During the accuracy evaluation phase, only outputs corresponding to low off-nadir images were selected for registration and error analysis to ensure comparability and fairness among different models.

4.2.1 Performance in MVS3D Tab. 1 presents the alignment results obtained using both LiDAR and DSM data. When LiDAR data were used for alignment, the Affine alignment strategy achieved the best performance in most scenarios, significantly outperforming the Linear, Rigid, and Rigid* methods. In terms of model performance, DAV2 achieved the lowest $RMSE$ and highest R^2 values across multiple scenes, demonstrating overall superiority, followed by VGGT and π^3. When DSM data were used for alignment, the results were largely consistent with those based on LiDAR data. The Affine alignment strategy again exhibited optimal performance in most scenarios, particularly in scenes (a) and (b), where its accuracy was comparable to that achieved with LiDAR-based alignment. However, in scene (c), the complex urban environment introduced DSM data issues such as clumping and noise, leading to a noticeable decline in overall alignment accuracy. Fig. 4 illustrates qualitative results before and after alignment under different alignment strategies for scene (c). Consistent with the quantitative findings, DAV2 combined with the Affine alignment strategy effectively mitigates systematic bias and achieves more stable geometric consistency. Furthermore, the pixel-wise residual histograms in Figs. 5(a), 5(b) and 5(c) reveal a more compact error distribution for this combination, further validating its superior alignment accuracy. Although VGGT combined with the Rigid* strategy achieves high accuracy in Fig. 5(d), the presence of numerous negative residuals (15 m to 5 m) reduces the overall error concentration, resulting in lower alignment stability compared with the more consistent performance of the Affine strategy.

4.2.2 Performance in WHU-TLC Tab. 2 presents the alignment results obtained using LiDAR data. The results show that existing alignment strategies perform poorly across almost all scenarios in the WHU-TLC dataset. Although Scene (d) yields a relatively low RMSE, its R^2 value approaches zero or even becomes negative. In Scene (e), while R^2 shows slight improvement, the RMSE remains excessively high. These results indicate that current 3D VFMs still encounter substantial challenges when processing low-resolution satellite imagery. Under

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-2-2026-811-2026 | © Author(s) 2026. CC BY 4.0 License.

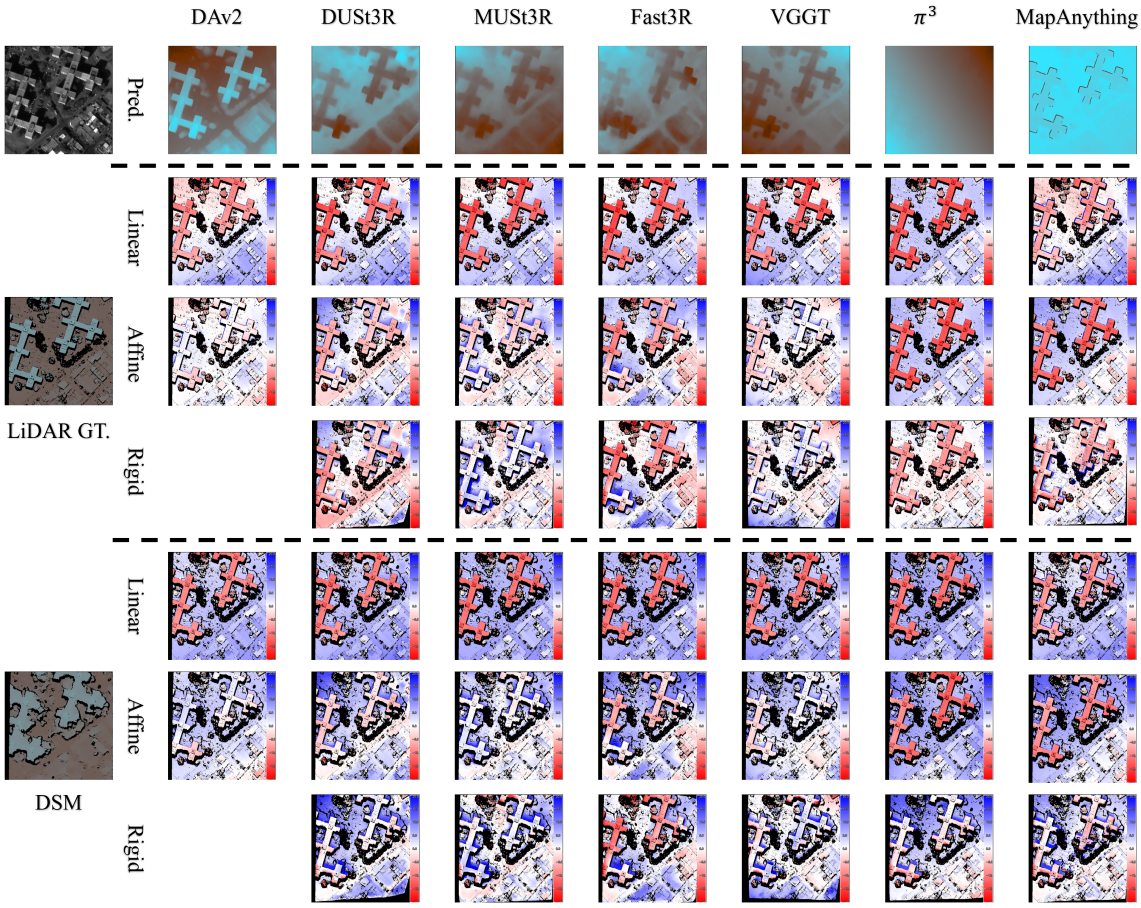
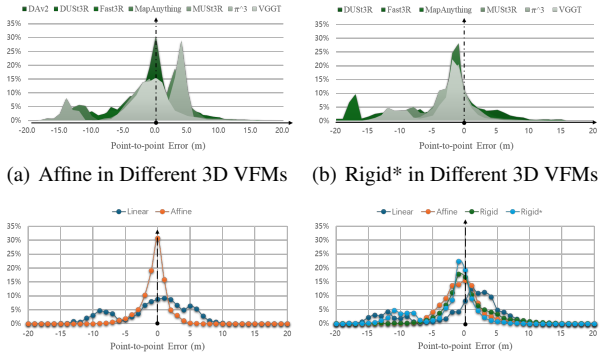


Figure 4. Comparison of error distributions for different monocular and multi-view depth models under various alignment strategies in scenario (c). The first row shows the raw depth predictions of each model, pseudo-colored for visualization. The following rows report the results after applying Linear, Affine, and Rigid alignment to LiDAR ground truth (GT) and DSM, respectively, along with the corresponding error maps. Errors are computed as (prediction - reference) and clipped to [-20, 20]: blue indicates positive values, red indicates negative values, and white denotes zero error (perfect agreement).



(a) Affine in Different 3D VFMs (b) Rigid* in Different 3D VFMs

(c) Different Strategies in DAv2 (d) Different Strategies in VGGT

Figure 5. Elevation error distribution of selected alignment results on the MVS3D dataset. Values closer to zero indicate higher alignment accuracy, while a more Gaussian-like distribution suggests better model fitting performance.

the Affine alignment strategy, the qualitative results for each scenario are presented in Fig. 6, further illustrating the impact of low-resolution imagery on reconstruction accuracy.

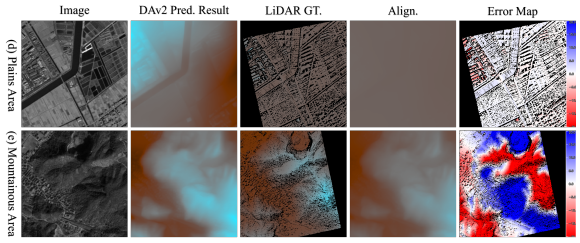


Figure 6. Illustration of affine alignment results based on LiDAR data in the WHU-TLC dataset

4.3 Large-Scale DSM Reconstruction

After obtaining the local alignment results from various models, this study further developed and validated an integrated reconstruction workflow based on geometric consistency correction and point cloud fusion, thereby demonstrating the potential of 3D VFMs for large-scale DSM reconstruction.

For each block, geometric consistency correction was applied to the alignment results using the corresponding RPC parameters derived from the imagery. By analyzing the elevation differences of homologous points across multiple viewpoints,

Scenario	Methods	DaV2		DUSt3R		MUSt3R		Fast3R		VGGT		π^3		MapAnything	
		RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑
Align with LiDAR															
(a)	*Affine*	3.94	0.69	4.46	0.61	4.67	0.57	5.86	0.33	4.04	0.68	7.04	0.03	6.70	0.12
	Linear	5.06	0.50	5.40	0.43	6.89	0.07	7.16	-0.01	5.95	0.31	7.16	-0.01	7.08	0.02
	Rigid	-	-	6.07	0.29	5.76	0.35	6.51	0.19	5.15	0.49	5.70	0.38	7.55	-0.08
	*Rigid**	-	-	6.29	0.23	7.30	-0.05	7.63	-0.12	5.41	0.43	11.01	-0.12	7.19	0.00
(b)	*Affine*	1.38	0.92	2.61	0.73	2.37	0.78	4.15	0.31	2.07	0.83	4.60	0.16	4.07	0.34
	Linear	1.47	0.91	4.71	0.12	4.65	0.14	5.00	0.01	3.71	0.45	4.83	0.07	4.96	0.02
	Rigid	-	-	5.50	-0.16	5.46	-0.18	10.99	-3.77	2.85	0.68	3.23	0.59	4.75	0.11
	*Rigid**	-	-	4.65	0.16	4.14	0.32	5.11	-0.04	2.84	0.68	3.25	0.58	4.36	0.25
(c)	*Affine*	2.46	0.88	4.39	0.60	3.71	0.72	5.11	0.46	3.05	0.81	6.91	0.01	6.36	0.16
	Linear	5.33	0.41	6.39	0.16	6.70	0.07	6.67	0.08	6.36	0.16	6.94	0.00	15.07	-3.71
	Rigid	-	-	4.96	0.50	4.59	0.57	5.85	0.29	3.91	0.69	4.06	0.66	7.93	-0.28
	*Rigid**	-	-	5.87	0.31	4.25	0.63	5.95	0.27	4.83	0.53	2.93	0.82	7.50	-0.16
Align with DSM															
(a)	*Affine*	3.91	0.66	4.36	0.58	4.66	0.52	5.79	0.26	3.87	0.67	6.91	-0.05	6.43	0.09
	Linear	4.44	0.57	4.75	0.50	6.45	0.08	6.82	-0.03	5.50	0.33	6.81	-0.02	6.59	0.04
	Rigid	-	-	5.57	0.33	5.88	0.24	6.29	0.17	5.39	0.38	5.59	0.33	7.51	-0.21
	*Rigid**	-	-	5.57	0.32	6.44	0.09	7.02	-0.06	5.15	0.43	5.44	0.35	5.75	0.28
(b)	*Affine*	1.78	0.86	2.68	0.67	2.54	0.71	4.10	0.24	2.40	0.74	4.53	0.07	4.02	0.27
	Linear	1.73	0.87	4.55	0.06	4.44	0.11	4.79	-0.04	3.78	0.35	4.64	0.02	4.46	0.10
	Rigid	-	-	5.39	-0.22	5.13	-0.19	10.39	-3.84	3.27	0.52	4.02	0.29	5.42	-0.30
	*Rigid**	-	-	4.23	0.20	3.61	0.41	4.55	0.07	3.30	0.51	3.35	0.50	4.10	0.25
(c)	*Affine*	3.56	0.74	4.94	0.49	4.24	0.62	5.59	0.35	3.81	0.70	7.24	-0.09	6.63	0.08
	Linear	6.28	0.18	6.87	0.02	7.16	-0.07	7.12	-0.06	6.82	0.03	7.30	-0.11	9.38	-0.84
	Rigid	-	-	6.29	0.21	5.31	0.42	5.94	0.27	5.24	0.44	5.84	0.31	8.36	-0.43
	*Rigid**	-	-	6.56	0.14	5.01	0.48	5.59	0.36	5.81	0.31	5.58	0.37	6.37	0.16

Table 1. Accuracy comparison of different alignment methods applied to various models across multiple scenes in the MVS3D dataset. The best and second-best results in each column are highlighted with orange and blue backgrounds, respectively. Within each scene, the best result is underlined in red, and the second-best is underlined in blue.

Scenario	Methods	DaV2		DUSt3R		MUSt3R		Fast3R		VGGT		π^3		MapAnything	
		RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑	RMSE ↓	R^2 ↑
(d)	*Affine*	1.49	0.12	1.53	0.07	1.53	0.07	1.52	0.08	1.49	0.12	1.49	0.12	1.53	0.07
	Linear	1.59	-0.01	1.58	0.01	1.61	-0.02	1.57	0.02	1.56	0.04	1.56	0.03	1.59	-0.01
	Rigid	-	-	32.95	-444.14	28.42	-314.19	19.76	-153.85	12.15	-58.17	13.80	-74.82	13.08	-65.56
	*Rigid**	-	-	1.32	-0.01	1.63	-0.04	1.62	-0.05	1.57	-0.03	1.31	-0.01	1.63	-0.04
(e)	*Affine*	22.35	0.68	11.04	0.92	10.2	0.93	17.73	0.80	22.63	0.68	34.19	0.26	27.59	0.52
	Linear	23.44	0.65	19.74	0.75	30.43	0.41	39.46	0.02	39.68	0.01	34.44	0.25	39.96	-0.01
	Rigid	-	-	85.42	-3.47	80.78	-3.20	22.76	0.68	40.84	-0.02	22.70	0.67	39.44	0.02
	*Rigid**	-	-	18.71	0.78	31.05	0.37	20.83	0.73	38.04	0.12	35.81	0.26	36.83	0.15

Table 2. Comparison of accuracy across different models and alignment methods in various scenarios of the WHU-TLC dataset. Within each scenario, the best and second-best results are underlined in red and blue, respectively.

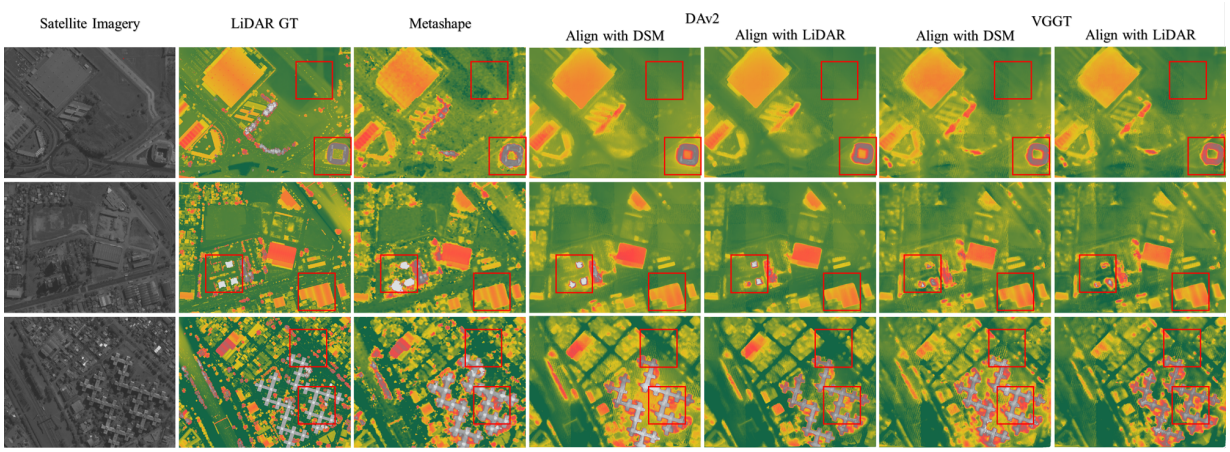


Figure 7. Visualization of large-scale DSM reconstruction results for three scenes in the MVS3D dataset.

geometrically inconsistent or anomalous points were automatically identified and removed, thereby improving the overall accuracy and reliability of the reconstructed point cloud. After point cloud generation, a voxel-based discretization approach was employed to produce a DSM. Finally, the quantitative assessment of DSM quality was conducted using the evaluation metrics described in Section 3.4.

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-2-2026-811-2026 | © Author(s) 2026. CC BY 4.0 License.

	-	DaV2	DaV2-5	CATALYST	Metashape	S2P	SDRDIS	JHU/APL	Adapted COLMAP	Sat-MVSF (WHU-TLC)	Sat-MVSF (WHU-MVS)	Sat-MVSF (self-ref.)
Mean of all sites	$PAE_{1m}(\%)\uparrow$	42.66	23.92	58.92	56.73	59.49	56.67	55.19	50.38	55.90	51.09	60.48
	MedAE(m) ↓	1.693	3.406	0.767	0.495	0.400	0.503	0.883	0.371	0.587	0.368	0.397
	RMSE(m) ↓	3.649	4.355	4.323	3.464	4.778	4.166	4.896	8.397	3.867	2.957	3.242

Table 3. Quantitative comparison on the MVS3DM benchmark (mean over all sites). For each metric, the best and second-best results are highlighted with orange and blue backgrounds, respectively.

	-	Metashape	DaV2		VGGT	
			DSM-based	LiDAR-based	DSM-based	LiDAR-based
(a)	RMSE ↓	3.00	2.90	2.88	2.74	2.79
	MAE ↓	1.52	1.57	1.32	1.57	1.39
	$PAE_{1m}(\%)\uparrow$	63.78	58.76	69.83	52.30	65.55
	$PAE_{3m}(\%)\uparrow$	88.72	89.91	90.84	89.58	89.72
(b)	RMSE ↓	3.71	2.91	3.08	3.23	3.29
	MAE ↓	1.70	1.77	1.68	1.90	1.84
	$PAE_{1m}(\%)\uparrow$	64.81	45.13	54.24	42.05	47.07
	$PAE_{3m}(\%)\uparrow$	83.25	85.46	86.85	84.69	85.49
(c)	RMSE ↓	4.76	4.06	3.83	4.13	3.95
	MAE ↓	2.70	2.79	2.43	2.91	2.69
	$PAE_{1m}(\%)\uparrow$	51.05	26.44	36.50	24.70	28.74
	$PAE_{3m}(\%)\uparrow$	73.54	68.84	76.81	66.89	72.04
Mean	RMSE ↓	3.75	3.25	3.24	3.32	3.31
	MAE ↓	1.91	1.98	1.75	2.05	1.90
	$PAE_{1m}(\%)\uparrow$	59.53	41.24	51.71	37.87	44.59
	$PAE_{3m}(\%)\uparrow$	81.59	80.87	84.62	79.76	82.06

Table 4. Comparison of large-scale DSM reconstruction accuracy across three scenarios. For each metric within each scenario, the best and second-best results are highlighted with orange and blue backgrounds, respectively.

4.3.1 Typical Urban Scenes As shown in Tab. 4, quantitative evaluations were performed on large-scale DSMs generated by various 3D VFMs. The 3D VFMs consistently achieved lower RMSE values than Metashape under both DSM-based and LiDAR-based alignment strategies, while MAE and PAE_{3m} remained comparable. However, the PAE_{1m} values revealed noticeable deficiencies. Specifically, in three representative scenes, the DSM-based alignment results of DAV2 and VGGT achieved 13.3% and 11.46% reductions in RMSE, respectively, compared with Metashape. Fig. 7 presents the qualitative comparisons. The DSMs reconstructed by 3D VFMs exhibit sharper building-edge structures and fewer noise artifacts in shadowed regions, confirming their advantages in large-scale DSM reconstruction. Nevertheless, in texture-sparse regions such as grasslands and roads, the reconstructed DSMs show limited elevation variation, likely due to the underrepresentation of such surface types in the models' training data.

4.3.2 MVS3DM Benchmark Evaluation To evaluate the reconstruction capability of the proposed alignment strategy, experiments were conducted on the MVS3DM dataset. As summarized in Table 3, in addition to assessing the overall reconstruction performance across all sites, we further investigated the influence of input image resolution on depth alignment accuracy. Two configurations were considered: **Dav2**, which directly aligns the predicted depth maps, and **Dav2-5**, where the input images are first downsampled by a factor of five prior to alignment.

Impact of resolution. The results in Table 3 show a noticeable degradation in accuracy when the input resolution is reduced. Specifically, PAE_{1m} decreases from 42.66% to 23.92%, corresponding to a reduction of 18.74 percentage points. Meanwhile, the MedAE increases from 1.693 m to 3.406 m, and the RMSE rises from 3.649 m to 4.355 m. These results indicate that high-resolution imagery provides important structural information that supports reliable depth alignment. When the spatial resolution is reduced, the loss of fine-scale geometric features—such as building boundaries and terrain discontinuities—limits the structural consistency of the predicted depth field and leads to larger reconstruction errors.

Gap to existing methods. From a dataset-level perspective, Dav2 demonstrates moderate reconstruction accuracy across the entire MVS3DM benchmark (Table 3). In terms of RMSE, the method achieves competitive performance (3.649 m), outperforming several traditional stereo processing pipelines. However, metrics reflecting local geometric precision, including PAE_{1m} and Median, remain lower than those achieved by advanced multi-view stereo approaches such as the Sat-MVSF series (best PAE_{1m} = 60.48%, RMSE = 2.957 m). This result highlights the trade-off between the global structural consistency provided by monocular VFM-based depth prediction and the fine-scale geometric accuracy achieved by multi-view stereo reconstruction.

5. Conclusions

This study systematically evaluated the performance of several 3D Vision Foundation Models (3D VFMs) in satellite-image-based 3D reconstruction tasks. The results reveal that: (1) For very-high-resolution satellite imagery (e.g., 0.3 m), 3D VFMs demonstrated reconstruction performance surpassing that of existing commercial photogrammetric solutions, achieving excellent geometric accuracy and detail fidelity with only minimal scale adjustment. (2) However, this advantage does not extend to medium- and low-resolution imagery (approximately 2–3 m), where the models struggle to accurately reconstruct terrain features due to their limited cross-resolution generalization capability. (3) In addition, existing alignment strategies were observed to introduce systematic biases, manifested as a distinct see-saw effect, which can be attributed to the linear assumptions and error propagation inherent in the registration process.

Future research will advance in two primary directions: (1) In practical applications, since initial DSMs are typically generated from stereo imagery, future work may explore integrating 3D vision foundation models with more reliable and readily available ground-elevation data—such as laser altimetry or sparse point clouds—to improve scale consistency and geometric accuracy. (2) Given that current 3D VFMs are fundamentally monocular in structure, developing a scale-calibration mechanism that avoids the introduction of systematic bias could enable 3D reconstruction from monocular satellite imagery, paving the way for low-cost and fully automated spatial modeling.

Acknowledgment

This research received funding from the National Natural Science Foundation of China (Project No. U25A20772) and the Natural Science Foundation of Sichuan Province (Project No. 2026NSFSCZY0054).

References

Agisoft LLC, 2022. Agisoft metashape – intelligent photogrammetryenhanced with lidar data processing, version 1.8.4. https://www.agisoft.com/.

Bosch, M., Kurtz, Z., Hagstrom, S., Brown, M., 2016. A multiple view stereo benchmark for satellite imagery. 2016 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), 1–9.

Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V., 2025. Must3r: Multi-view network for stereo 3d reconstruction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 1050–1060.

Cheng, J., Liu, L., Xu, G., Wang, X., Zhang, Z., Deng, Y., Zang, J., Chen, Y., Cai, Z., Yang, X., 2025. Monster: Marry monodepth to stereo unleashes power. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6273–6282.

de Franchis, C., de Franchis, C., de Franchis, C., de Franchis, C., Meinhardt-Llopis, E., Meinhardt-Llopis, E., Michel, J., Michel, J., Morel, J., Morel, J.-M., Morel, J.-M., Facciolo, G., Facciolo, G., 2014a. Automatic sensor orientation refinement of Pléiades stereo images. 2014 IEEE Geoscience and Remote Sensing Symposium.

de Franchis, C., de Franchis, C., de Franchis, C., de Franchis, C., Meinhardt-Llopis, E., Meinhardt-Llopis, E., Michel, J., Michel, J., Morel, J., Morel, J.-M., Morel, J.-M., Facciolo, G., Facciolo, G., 2014b. On stereo-rectification of pushbroom images. International Conference on Information Photonics.

Gao, J., Liu, J., Ji, S., 2023. A General Deep Learning Based Framework for 3D Reconstruction from Multi-View Stereo Satellite Images. ISPRS Journal of Photogrammetry and Remote Sensing, 195, 446–461.

Grodecki, J., Dial, G., 2003. Block Adjustment of High-Resolution Satellite Images Described by Rational Polynomials. Photogrammetric Engineering and Remote Sensing, 69(1), 59–68.

He, S., Li, S., Jiang, S., Jiang, W., 2022. HMSM-net: Hierarchical Multi-Scale Matching Network for Disparity Estimation of High-Resolution Satellite Stereo Images. ISPRS Journal of Photogrammetry and Remote Sensing, 188, 314–330.

Hirschmüller, H., Hirschmüller, H., 2008. Stereo Processing by Semiglobal Matching and Mutual Information. IEEE Transactions on Pattern Analysis and Machine Intelligence.

Hu, H., Su, L., Mao, S., Chen, M., Pan, G., Xu, B., Zhu, Q., 2023. Adaptive Region Aggregation for Multi-View Stereo Matching Using Deformable Convolutional Networks. The Photogrammetric Record, 38, 430–449.

Keetha, N., Müller, N., Schönberger, J., Porzi, L., Zhang, Y., Fischer, T., Knapitsch, A., Zauss, D., Weber, E., Antunes, N. et al., 2026. MapAnything: Universal feed-forward metric 3d reconstruction. International Conference on 3D Vision (3DV), IEEE.

Li, S., He, S., Jiang, S., Jiang, W., Zhang, L., 2023. WHU-stereo: A Challenging Benchmark for Stereo Matching of High-Resolution Satellite Images. IEEE Transactions on Geoscience and Remote Sensing, 61, 1–14.

Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B., 2025. Prompting depth anything for 4k resolution accurate metric depth estimation. Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR), 17070–17080.

Qin, R., 2019. A Critical Analysis of Satellite Stereo Pairs for Digital Surface Model Generation and a Matching Quality Prediction Model. ISPRS Journal of Photogrammetry and Remote Sensing, 154, 139–150.

Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V., 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3).

Rottensteiner, F., Rottensteiner, F., Sohn, G., Sohn, G., Jung, J., Jung, J., Jung, J., Gerke, M., Gerke, M., Baillard, C., Baillard, C., Benitez, S., Benitez, S., Breitkopf, U., Breitkopf, U., 2012. The ISPRS benchmark on urban object classification and 3D building reconstruction. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences.

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 (CVPR), 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 (CVPR), 20697–20709.

Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T., 2026. π^3: Permutation-equivariant visual geometry learning. The Fourteenth International Conference on Learning Representations.

Wei, Z., Zhang, S., Xu, W., Zhang, L., Wang, Y., Zhang, J., Liu, J., 2025. TS-SatMVSNet: Slope Aware Height Estimation for Large-Scale Earth Terrain Multiview Stereo. IEEE Transactions on Geoscience and Remote Sensing, 63, 1-15.

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 (CVPR), 5249–5260.

Wu, X., Landgraf, S., Ulrich, M., Qin, R., 2025. An evaluation of DUSt3R/MASt3R/VGGT 3D reconstruction on photogrammetric aerial blocks. Geo-spatial Information Science, 0(0), 1–19.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 21924–21935.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (eds), Advances in Neural Information Processing Systems, 37, Curran Associates, Inc., 21875–21911.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. MVSNet: Depth inference for unstructured multi-view stereo. Computer Vision – ECCV 2018, Springer International Publishing, 785–801.

Youssefi, D., Michel, J., Sarrazin, E., Buffe, F., Cournet, M., Delvit, J.-M., L’Helguen, C., Melet, O., Emilien, A., Bosman, J., 2020. Cars: A photogrammetry pipeline using dask graphs to construct a global 3d model. IGARSS 2020 - 2020 IEEE International Geoscience and Remote Sensing Symposium, 453–456.

Zhang, K., Snavely, N., Sun, J., 2019. Leveraging vision reconstruction pipelines for satellite imagery. 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), IEEE, Seoul, Korea (South), 2139–2148.

https://doi.org/10.5194/isprs-annals-XI-2-2026-811-2026