Precision Comparison and Analysis of Multi-stereo Fusion and Multi-view Matching Based on High-Resolution Satellite Data

LIU Tengfei, *HUANG Xu**, *HUANG Zefeng*

School of Geospatial Engineering and Science, Sun Yat-sen University, Zhuhai 519082, P. R. China

(Received 15 March 2025; revised 10 June 2025; accepted 1 September 2025)

Abstract: High-resolution sub-meter satellite data play an increasingly crucial role in the 3D real-scene China construction initiative. Current research on 3D reconstruction using high-resolution satellite data primarily focuses on two approaches: Multi-stereo fusion and multi-view matching. While algorithms based on these two methodologies for multi-view image 3D reconstruction have reached relative maturity, no systematic comparison has been conducted specifically on satellite data to evaluate the relative merits of multi-stereo fusion versus multi-view matching methods. This paper conducts a comparative analysis of the practical accuracy of both approaches using high-resolution satellite datasets from diverse geographical regions. To ensure fairness in accuracy comparison, both methodologies employ non-local dense matching for cost optimization. Results demonstrate that the multi-stereo fusion method outperforms multi-view matching in all evaluation metrics, exhibiting approximately 1.2% higher average matching accuracy and 10.7% superior elevation precision in the experimental datasets. Therefore, for 3D modeling applications using satellite data, we recommend adopting the multi-stereo fusion approach for digital surface model (DSM) product generation.

Key words: multi-stereo fusion reconstruction; multi-view matching reconstruction; non-local dense matching method; occlusion detection; high-resolution satellite data

CLC number: P236 **Document code:** A **Article ID:** 1005-1120(2025)05-0577-12

0 Introduction

Dense image matching methods primarily aim to search for corresponding points pixel-by-pixel in stereo pairs, generating dense 3D point clouds through forward intersection. These methods offer significant advantages, including large observation ranges, high point cloud density, and low modeling costs, enabling widespread applications in architectural design[1], urban planning[2], disaster monitoring[3], cultural heritage preservation[4], and engineering surveying[5].

Current 3D reconstruction methods based on multi-view imagery encompass two approaches: (1) Dense matching for individual stereo pairs followed by multi-stereo fusion of digital surface model (DSM) products (referred to as the multi-stereo fusion method); and (2) direct multi-view dense matching to generate DSMs (referred to as the multi-view matching method).

The multi-stereo fusion method divides 3D reconstruction into two steps: Single-stereo dense matching and multi-view surface model fusion. For single-stereo dense matching, traditional photogrammetric stereo matching typically involves four steps: (1) Cost computation[6]; (2) cost optimization; (3) disparity calculation[6-7]; and (4) disparity refinement[8-9]. Cost computation measures the similarity between corresponding points in stereo images, commonly using metrics such as MCCNN[10],

*Corresponding author, E-mail address: huangx358@mail.sysu.edu.cn.

How to cite this article: LIU Tengfei, HUANG Xu, HUANG Zefeng. Precision comparison and analysis of multi-stereo fusion and multi-view matching based on high-resolution satellite data[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025, 42(5): 577-588.

http://dx.doi.org/10.16356/j.1005-1120.2025.05.002

census[11], mutual information[12], correlation coefficients[13], and gradient histograms[14]. Cost optimization relies on smoothness constraints between corresponding points to construct a global energy function for refining matching costs[12,15]. Disparity calculation extracts the optimal disparity for each pixel from the optimized cost using a winner-takes-all (WTA) strategy. Disparity refinement involves post-processing operations, such as noise removal and interpolation of invalid regions[16]. For multi-view DSM fusion, various algorithms exist. A simple approach involves weighted averaging of multiple DSMs along the depth direction[17]. However, weighted averaging fails to handle outliers or errors in DSMs, leading to the adoption of median fusion[18]. While median fusion and weighted averaging operate at the pixel level, they neglect local surface geometry. To address this, spatial median-based filtering methods[19] assume Gaussian distribution of elevation values within DSM grids, reducing noise and improving elevation estimation.

In contrast, the multi-view matching method directly generates DSMs from multi-view imagery by incorporating multi-view constraints into the cost computation. These constraints, such as photometric consistency and visibility, enhance matching reliability. Photometric consistency evaluates pixel similarity across images based on color or intensity, while visibility constraints ensure that only points observable across multiple views are matched[20]. However, challenges persist in occlusion detection and effective utilization of multi-view intensity constraints. Multi-view matching methods can be categorized into depth-map-based, voxel-based, and local-patch-based approaches. Depth-map-based methods improve photometric consistency by matching reference views with all visible views[21]. Voxel-based methods compute cost functions over 3D voxels and apply optimizations like graph cuts to extract DSMs[22]. Local-patch-based methods match image patches to generate semi-dense or dense point clouds[23].

Despite the maturity of both methods, no systematic comparison has been conducted for satellite data. High-resolution satellite data, with their wide coverage, long-term stability, and periodic observation capabilities, are widely used in surveying, mapping, and national security[24-27]. This study compares the accuracy of multi-stereo fusion and multi-view matching using high-resolution satellite datasets. Both methods employ non-local dense matching for cost optimization to ensure fairness. The findings provide technical guidance for satellite-based 3D reconstruction in engineering applications.

1 Multi-stereo Fusion Method

The multi-stereo fusion method first performs dense matching for each image pair and then fuses the DSM products from multiple image pairs to generate a complete 3D product. In this paper, the optimization method for dense matching adopts a non-local approach, as described in Section 1.1; the multi-view DSM fusion method uses a locally weighted median fusion approach, as introduced in Section 1.2.

1.1 Non-local dense matching method

This paper employs the concept of dynamic programming to achieve efficient dense matching. To address the limitation of weak one-dimensional constraints in traditional dynamic programming methods, a novel non-local dynamic programming path is adopted. This allows the dense matching result of each pixel to be influenced by surrounding global pixels, thereby achieving both high efficiency and high accuracy in dense matching. Traditional semi-global matching (SGM) methods only consider cost aggregation in eight directions, meaning that each pixel's matching result is constrained solely by scanning lines in those eight directions. As illustrated in Fig.1, pixel Q can only propagate cost information to pixels on the scanning line (e.g. P), but cannot influence pixels outside the scanning line. In weakly textured regions where the signal-to-noise ratio is low, relying solely on scanning line constraints is insufficient for robust dense matching. Compared to traditional SGM algorithms, this study enhances the matching robustness in weakly textured areas through a two-stage iterative approach. In the first iteration, the matching cost of pixel Q is propagated

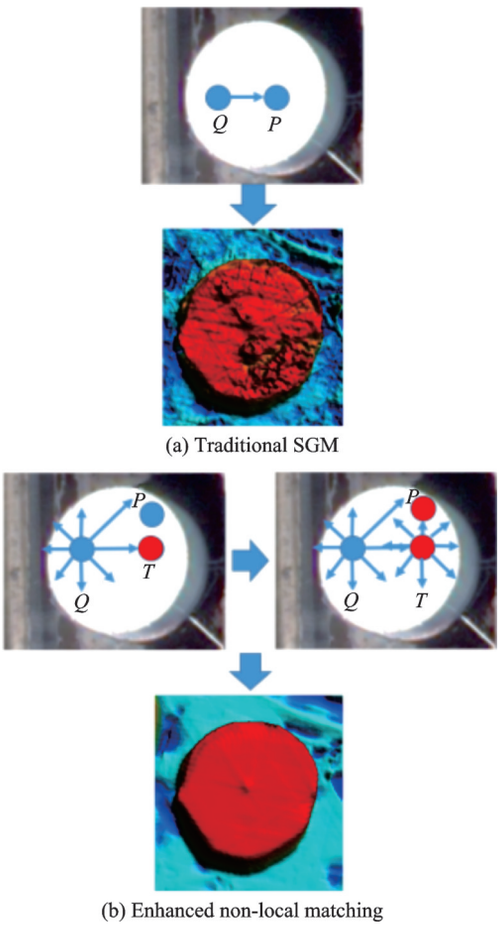


Fig.1 Cost aggregation path

along the eight scanning directions to pixel T. In the second iteration, the cost is further propagated from pixel T to pixel P along its own eight scanning directions. Therefore, by performing two iterations, each pixel in the weakly textured region is constrained by global pixels, which significantly enhances the robustness of the matching process[11].

The SGM algorithm formulates dense matching as a labeling problem by establishing a global energy function, which is optimized through 1D dynamic programming in eight directions, shown as

$$L_r(p,d)=C(p,d)+\min\begin{Bmatrix} L_r(p-1,d) \\ L_r(p-1,d-1)+P_1 \\ L_r(p-1,d+1)+P_1 \\ \min\limits_k L_r(p-1,k)+P_2 \end{Bmatrix}-\min_i L_r(p-1,i) \tag{1}$$

where $L_r(p,d)$ denotes the accumulated cost of the pixel p at disparity d along the current path; r the direction of the path; $C(p,d)$ the matching cost of pixel p at disparity d; and $p-1$ the previous pixel of p along the current path direction.

According to the SGM theory[12], Eq.(2) is typically used to perform cost aggregation in eight directions across the entire image. The aggregated results from all directions are then summed to obtain an approximate optimal solution of the global energy function, which serves as the final result of dense matching, shown as

$$S^1(p,d)=\sum_r L_r(p,d) \tag{2}$$

where $S^1(p,d)$ represents the overall cost aggregation result obtained by summing the cost aggregation outcomes from all directions during the first iteration.

After the first iteration, for each pixel, valid paths exist only along the scanning line directions, while pixels outside the scanning lines remain unconnected. Therefore, the aggregated result from the first iteration is used as the new cost input for the second iteration, leading to

$$L_r^2(p,d)=S^1(p,d)+\min\begin{Bmatrix} L_r^2(p-1,d) \\ L_r^2(p-1,d-1)+P_1 \\ L_r^2(p-1,d+1)+P_1 \\ \min\limits_k L_r^2(p-1,k)+P_2 \end{Bmatrix} \tag{3}$$

where $L_r^2(p,d)$ denotes the accumulated cost of the pixel p at disparity d along direction r during the second iteration.

Finally, the cost aggregation results from all eight directions are summed, shown as

$$S^2(p,d)=S(p,d)+\sum_{r=1}^{8}(L_r^2(p,d)-S(p,d)) \tag{4}$$

where $S^2(p,d)$ represents the total cost aggregation result from all eight directions during the second iteration. Finally, WTA strategy is applied to extract the final disparity map from the optimized cost volume.

After two iterations, each pixel in the image is connected via paths to all other pixels across the entire image, thereby significantly enhancing the overall performance of dense matching.

1.2 Local window weighted median fusion method

Traditional fusion of homologous DSMs typically employs a point-based median filtering strategy, that is, the number of 3D points appearing at the same location is counted, and the median of their elevations is taken as the fusion result, as illustrated in Fig.2(a). Although the traditional method is simple and effective, it does not consider information from surrounding points, resulting in residual noise on the model surface after fusion. Moreover, in challenging areas, the scarcity of valid 3D points leads to inaccurate fusion results.

To address these issues, this paper adopts a DSM fusion method based on local grayscale consistency[28], as shown in Fig.2(b). The core idea of this method is as follows: First, a local window is established centered at the current fusion position; second, in each layer of DSM and digital orthophoto map (DOM), 3D points with grayscale values similar to the central pixel are identified; finally, the elevations of all grayscale-similar pixels across all layers are combined, and their median is taken as the final fusion result. This method fully incorporates the 3D information of surrounding points, thus achieving higher fusion accuracy.

The core of the DSM fusion method based on local grayscale consistency lies in how to extract high-precision fusion results by utilizing the similarity of surrounding pixels to the central pixel. This paper adopts a weighted median filtering strategy to achieve this goal. First, within a window, each pixel is assigned a weight ranging from 0 to 1 based on its grayscale similarity to the central pixel—The more similar the grayscale, the higher the weight. Second, the weight of each pixel in the window is multiplied by a fixed factor (defaulted to 2 in the system), and the number of times each pixel's elevation appears in the median filtering process is calculated. The higher the weight, the more frequently the corresponding elevation appears, and thus the greater the likelihood it will be selected by the median filter. Third, the weighted results from all DSMs are aggregated to form an elevation sequence, such as $\{d_1^1, d_1^1, d_2^1, \cdots, d_i^j, \cdots\}$, where d_i^j represents the elevation information of the ith pixel in the jth DSM window. When the weight of a pixel is relatively high, the corresponding elevation is more likely to appear in the sequence (e.g. d_1^1), whereas when the weight is low, the corresponding elevation appears less frequently or even not at all (e.g. d_2^1), thereby ensuring the continuity and fusion accuracy of building edges. Finally, to achieve fast weighted median filtering, this paper employs a strategy based on histogram representations.

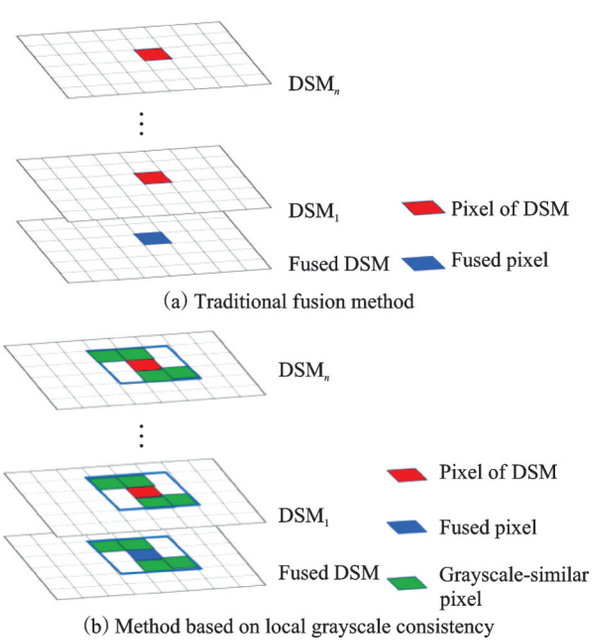


Fig.2 Schematic of DSM fusion

In the fast computation method based on histogram representations, the DSM space is first discretized into histogram bins according to spatial resolution, and the frequency of each elevation value in the elevation sequence is quickly counted. Then, starting from the lowest elevation, the elevation value at which the cumulative frequency reaches half of the total is selected as the median, and the corresponding bin height represents the final result of the median filtering.

2 Multi-view Matching Method

The multi-view matching method directly utilizes multiple images and operates in object space. It defines the elevation range through the vertical line locus (VLL) and employs a non-local dense match-

ing optimization approach to directly generate DSM products. The advantage of this object-space-based multi-view direct optimization method lies in its high computational efficiency and low resource requirements. However, challenges remain in determining pixel occlusion and fully utilizing grayscale constraints from multiple views.

To leverage the respective strengths of the multi-stereo fusion method and the multi-view matching method while overcoming their limitations, this paper proposes a hierarchical image matching method that integrates both approaches. The technical framework is illustrated in Fig.3. First, image pyramids are constructed for the multi-view images, as shown in Fig.3(a). At the upper level of the pyramid, a robust initial disparity map is generated using the multi-stereo fusion method, as shown in Fig.3(b). Finally, at the bottom level of the pyramid, high-precision multi-view matching is performed by determining the visible images and grayscale constraints for each ground pixel based on the initial disparity map, as shown in Fig.3(c).

In this proposed method, applying the multi-stereo fusion approach at the upper pyramid level significantly reduces computational resources and improves efficiency. For example, in a 2×2 pyramid, the image matching time at the upper level is only one-eighth of that at the lower level. Furthermore, to address the visibility issue of ground pixels in multi-view images, the initial surface model provided by the multi-stereo fusion method is used as the basis for occlusion judgment. This effectively resolves the visibility and grayscale constraint issues in the multi-view matching process.

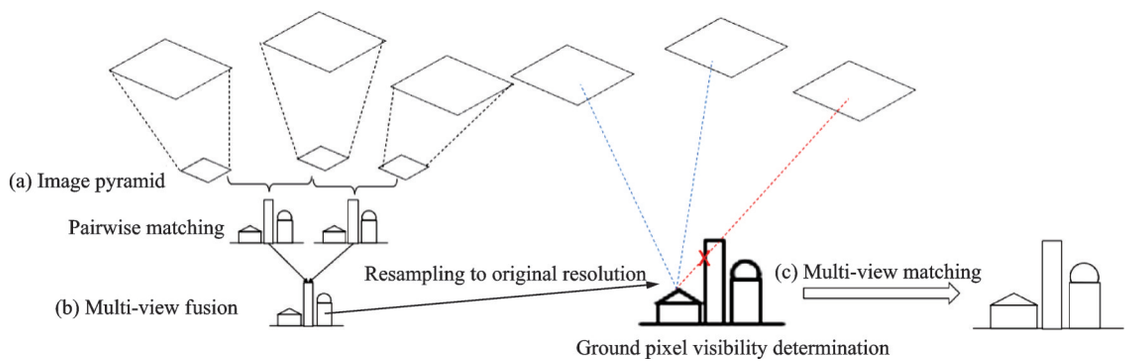


Fig.3 Flowchart of image matching technology

2.1 Visible image set for ground pixels

The major challenge in multi-view matching lies in determining the visible image set for each ground pixel. Including occluded or invisible images in the matching process can severely degrade the accuracy of image matching. To address this issue, this paper uses the multi-stereo fusion result from the upper level of the pyramid as the basis for visibility determination and applies the Z-buffer algorithm to assess whether each pixel is visible in a given image. The specific steps are as follows:

(1) For a given satellite image, all 3D points within the study area are projected onto the image using the rational function model (RFM).

(2) Projected points falling outside the image extent are discarded, and only those within the image bounds are retained.

(3) For each pixel on the satellite image, the elevation of the projected 3D point is recorded. If multiple elevation values are projected onto the same pixel, only the highest elevation point is retained, while the others are considered occluded.

(4) All images are processed sequentially, and the visible image set for each ground pixel is finally obtained.

2.2 Image matching based on multi-view constraints

Assuming the visible image set for ground pixel p is $V(p)=\{I_i(p)\}$, the image matching cost term can be expressed as the average matching cost of pixel p across all visible images, shown as

$$C(p,d)=\frac{\sum_{I_i,I_j\in V(p)}\text{cost}(p,d,I_i,I_j)}{\left(|V(p)|\cdot\frac{|V(p)|-1}{2}\right)} \tag{5}$$

where $C(p,d)$ denotes the total matching cost for pixel p at elevation d; $|V(p)|$ the number of visible images in the visible image set; and $\text{cost}(p,d,I_i,I_j)$ the matching cost for pixel p at elevation d between the image pair I_i and I_j. Eq.(5) shows that the matching cost for pixel p is expressed as the average cost over multiple stereo image pairs in the visible image set. The Census transform is used in this paper as the cost similarity measure.

In designing the cost smoothing term for image matching, this paper assumes, as in prior work, that neighboring pixels with similar appearance should have consistent elevation values. Therefore, the smoothing term is formulated as a function of the elevation disparity between neighboring pixels, shown as

$$T(p,q)=\begin{cases}0 & |d_p-d_q|=0\\ P_1 & |d_p-d_q|\leqslant 1\\ P_1+P_2\cdot w(p,q) & |d_p-d_q|>1\end{cases} \tag{6}$$

where $T(p,q)$ denotes the smoothness term between pixels p and q. When the smoothness term is strong, the probability that pixels p and q share similar elevations increases. d_p and d_q represent the estimated elevations of pixels p and q, respectively; P_1 is the smaller penalty coefficient, P_2 the larger penalty coefficient, and $w(p,q)$ the weight defined by the grayscale difference between pixels p and q. A larger grayscale difference results in a smaller weight; conversely, a smaller grayscale difference yields a larger weight. The value of $w(p,q)$ lies within the range $[0,1]$, and it is used to improve reconstruction accuracy at elevation discontinuities. To achieve high-precision weighted computation, the grayscale values of pixels p and q must be obtained. In this paper, the average grayscale across all visible images is used as the grayscale value of the ground pixel, shown as

$$S^1(p,d)=\sum_r L_r(p,d) \tag{7}$$

$$\overline{G(p)}=\frac{\sum_{I_i\in V(p)}g(p,I_i)}{|V(p)|} \tag{8}$$

where $\overline{G(p)}$ denotes the average grayscale of pixel p across the visible image set; $g(p,I_i)$ the grayscale of pixel p in the visible image I_i. Therefore, the definition of the weight $w(p,q)$ is

$$w(p,q)=\exp\left(-\frac{\left(\overline{G(p)}-\overline{G(q)}\right)^2}{\sigma^2}\right) \tag{9}$$

where σ^2 represents the smoothing factor that adjusts the weight of grayscale similarity.

Based on the definitions of the data term and the smoothness term, this paper ultimately combines the two to construct a global energy function, and a non-local energy optimization strategy is adopted to directly obtain the DSM of the surveyed area, shown as

$$E(D)=\sum_{p\in M}C(p,d)+\sum_{p,q\in M}T(p,q) \tag{10}$$

where E denotes the global energy function for multi-view matching; D the set of elevation values for all ground pixels; and M the spatial extent of the survey area.

For the energy function (Eq.(10)), this paper continues to employ a non-local dense matching strategy for optimization. The specific optimization method for the energy function can be found in Section 1.1.

3 Experimental Areas

To comprehensively evaluate the geometric accuracy of the multi-stereo fusion method and the multi-view matching method, three representative experimental regions with distinct terrain characteristics were selected, as shown in Fig.4. The specific details are as follows.

(1) Stereo satellite data of Shandong University of Science and Technology, China

This dataset uses stereo imagery from the GF-7 satellite, with a spatial resolution ground sampling distance (GSD), which refers to the distance between the centers of two consecutive pixels on the

ground, of approximately 0.7 m, captured on April 17, 2021. The surface of this region is primarily covered by buildings and mountainous terrain, with some vegetation, resulting in moderate terrain complexity. The stereo image pair has an intersection angle of 25°, high solar elevation angle, minimal shadow regions, and an image overlap exceeding 80%, indicating high data quality. The corresponding ground truth data consist of laser point cloud data from the same region (GSD of 13 cm), resampled to generate DSM ground truth products matching the image resolution.

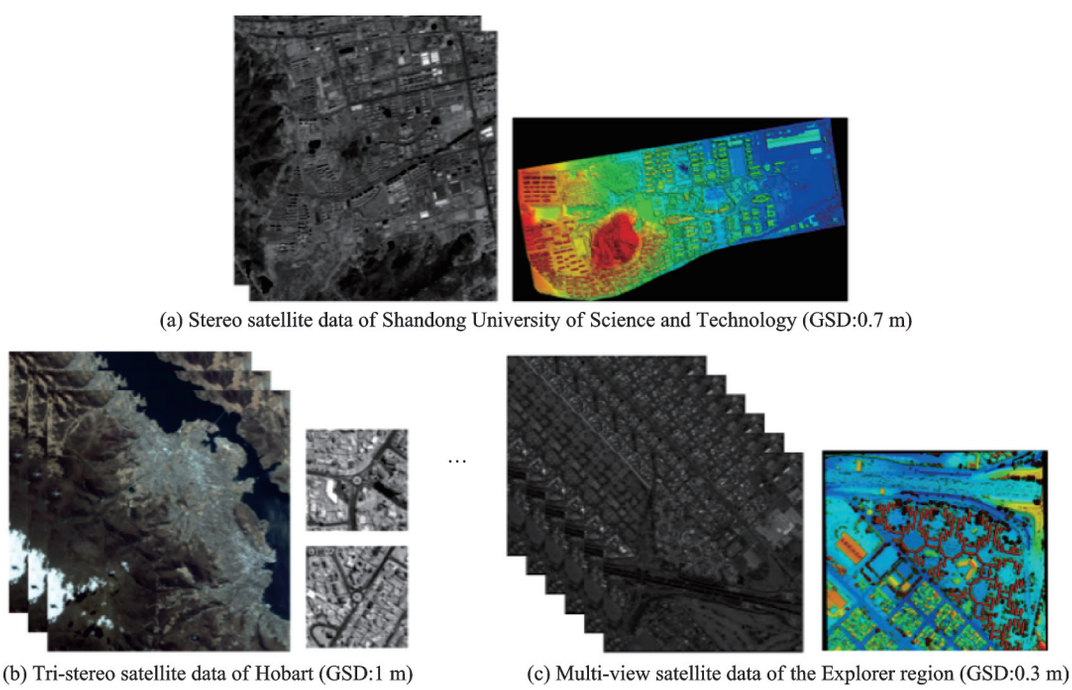
(2) Tri-stereo satellite data of Hobart, Australia

This dataset uses tri-view imagery from the IKONOS satellite, with a spatial resolution GSD of approximately 1.0 m, captured in 2003. The surface of this region includes mountainous areas, forests, exposed surfaces, and a small number of buildings, contributing to high terrain complexity. The intersection angles of the three images range from 15° to 30°, with an image overlap exceeding 70%, high solar elevation angle, minimal shadow regions, and high data quality, making it suitable for the experimental requirements of this study. The ground truth data consist of 114 high-precision control points and their corresponding annotations in the same region.

(3) Multi-view high-resolution satellite data of the Explorer region, Argentina

This dataset uses multi-view off-nadir imagery from the WorldView-3 satellite, comprising a total of 19 images captured over a time span of approximately two years (2015—2016), with a spatial resolution GSD of approximately 0.3 m. The surface of this region is primarily flat, with no mountainous terrain, many buildings, and some vegetation, resulting in low terrain complexity. The image overlap ranges from 60% to 80%, and the abundant number of images makes it suitable for multi-view matching experiments. The corresponding ground truth data consist of laser point cloud data from the same region, resampled to generate DSM ground truth products matching the image resolution.

By selecting satellite data with varying terrain complexities as experimental samples and incorporating diverse image parameters, this study enables a more comprehensive evaluation of the geometric accuracy of the multi-stereo fusion method and the multi-view matching method under different terrain conditions.



(a) Stereo satellite data of Shandong University of Science and Technology (GSD:0.7 m)

(b) Tri-stereo satellite data of Hobart (GSD:1 m)

(c) Multi-view satellite data of the Explorer region (GSD:0.3 m)

Fig.4 Schematic diagrams of the experimental area

4 Results and Analysis

4.1 Accuracy evaluation methods

The accuracy evaluation metrics used in this study include matching accuracy and elevation accuracy.

For matching accuracy, the ground truth data are used as references to independently evaluate the DSM accuracy produced by the multi-stereo fusion method and the multi-view matching method. A predefined threshold is set: If the absolute error of a pixel is less than this threshold, the pixel is considered correctly matched; otherwise, it is considered a mismatched pixel. The ratio of correctly matched pixels to the total number of pixels in each DSM is calculated as the matching accuracy, defined as

$$\mathrm{Acc}_{\mathrm{per}}=\frac{\mathrm{COUNT}(p|\mathrm{dis}(p)<2)}{\mathrm{COUNT}(p|\mathrm{dis}(p)\in N)} \tag{11}$$

where p denotes a pixel in the DSM; $\mathrm{dis}(p)$ the elevation difference between the estimated elevation of pixel p and the ground truth; COUNT the number of pixels satisfying a specific condition; and $\mathrm{Acc}_{\mathrm{per}}$ the proportion of correctly matched pixels relative to the total number of valid DSM pixels.

Previously, different datasets typically required the selection of thresholds based on their spatial resolution, as higher spatial resolution and greater image coverage often necessitate smaller threshold values. However, to simplify the computational methodology and ensure stronger consistency and comparability of results across different datasets, this study opted to use a unified threshold of $T=2$ meters for accuracy evaluation. This choice minimizes evaluation biases caused by threshold differences and provides a more intuitive reflection of the performance of the matching algorithms.

For evaluating elevation accuracy, the ground truth DSM of the Explorer region from 2017 is used as a reference. The elevation accuracy of the DSMs produced by the multi-stereo fusion method and the multi-view matching method is separately calculated, defined as

$$\mathrm{Acc}_{\mathrm{H}}=\mathrm{AVG}\left(|H_p-\overline{H_p}|\right) \tag{12}$$

where $\mathrm{Acc}_{\mathrm{H}}$ denotes the DSM elevation accuracy metric; AVG the averaging function; H_p the estimated elevation of a given pixel; and $\overline{H_p}$ the ground truth elevation of that pixel.

4.2 Comparison and analysis of modeling accuracy

Using ground truth DSMs and control point data as references, the elevation accuracy of DSMs generated by two types of algorithms, multi-stereo fusion and multi-view matching, is evaluated.

In the case of the Shandong University of Science and Technology dataset, only two satellite images are available. Therefore, the multi-stereo fusion method in this experiment essentially corresponds to a dense matching result derived from a stereo pair. Under this two-image condition, the difference between the multi-stereo fusion and multi-view matching methods lies in their processing domains: The former performs dense matching in image space, while the latter operates in object space.

As shown in Table 1, overall, the geometric accuracy of the multi-stereo fusion method surpasses that of the multi-view matching method across all experimental datasets. Specifically:

(1) Shandong University of Science and Technology Dataset: The accuracy of the two algorithms is very close for this dataset, primarily because the dataset contains only two images, which limits the use of multi-view constraints.

(2) Hobart Dataset: In the experiments involving the Hobart dataset, both the multi-stereo fusion method and the multi-view matching method achieved a matching accuracy ($\mathrm{Acc}_{\mathrm{per}}$) of 100%. This is attributed to the ground truth data in the Hobart region, which consist of 114 high-precision control points and their corresponding annotations. These control points are typically located at prominent and easily matched features, such as building corners or road intersections, resulting in high matching accuracy. Both methods achieved perfect matching within the threshold of $T=2$ m. However, matching accuracy alone does not fully reflect the differences in geometric precision. A comparison of elevation errors reveals that the elevation error of

Table 1 Accuracy statistics

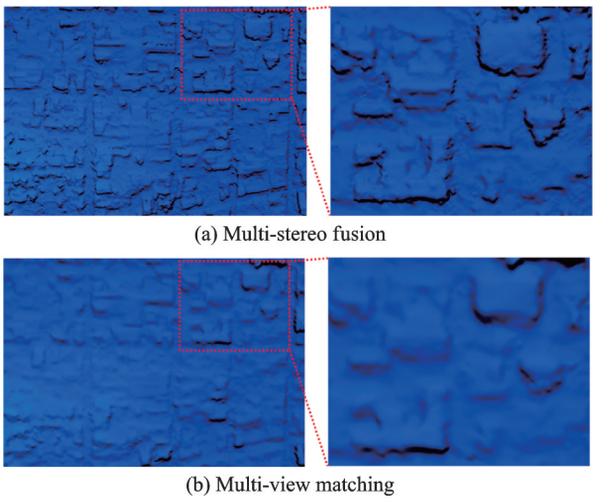
Dataset	Acc_{per}/%		Acc_H/m	
	Multi-stereo fusion	Multi-view matching	Multi-stereo fusion	Multi-view matching
Shandong University of Science and Technology	84.06	83.78	1.71	1.72
Hobart	100	100	0.66	0.79
Explorer	86.43	83.00	1.18	1.39

the multi-stereo fusion method is 0.66 m, while that of the multi-view matching method is 0.79 m. This discrepancy indicates that the multi-stereo fusion method exhibits superior precision and robustness in elevation estimation.

(3) Explorer Dataset: The geometric accuracy of the multi-stereo fusion method is significantly higher than that of the multi-view matching method. This is because the multi-view matching method concentrates redundant observational constraints in the cost computation stage, while the subsequent cost optimization and disparity computation stages lack sufficient utilization of these constraints, leading to lower accuracy. In contrast, the multi-stereo fusion method effectively leverages redundant observational constraints across multiple stages, including cost computation, cost optimization, and multi-stereo fusion, resulting in higher DSM product accuracy.

Through a systematic analysis of datasets with varying terrain complexities, an important conclusion can be drawn: The DSM accuracy generated by the multi-stereo fusion method consistently outperforms that of the multi-view matching method under different terrain complexity conditions. This conclusion is universal and demonstrates the broader applicability of the multi-stereo fusion method in handling 3D reconstruction tasks for satellite imagery. It provides strong evidence to support the selection of methods in practical applications.

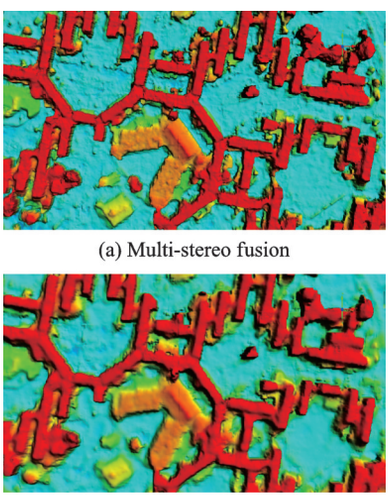
To further compare the modeling performance of the two methods, this study selects local regions from the Hobart and Explorer datasets to visualize the DSM results produced by both approaches, as shown in Figs.5 and 6. From the comparison, it can be observed that both the multi-stereo fusion method and the multi-view matching method achieve good reconstruction performance: The model surfaces are smooth and continuous, with minimal noise, and the building structures are clearly and accurately reconstructed.



(a) Multi-stereo fusion

(b) Multi-view matching

Fig.5 Local comparison in the Hobart Area



(a) Multi-stereo fusion

(b) Multi-view matching

Fig.6 Local comparison in the Explorer region

However, the multi-stereo fusion method demonstrates superior capability in modeling fine details. Due to the absence of redundant observational constraints during the cost optimization and disparity es-

timation stages, the multi-view matching method exhibits weaker detail modeling performance compared to the multi-stereo fusion method. Overall, the difference in detail reconstruction ability between the two methods contributes directly to the variation in their final geometric modeling accuracy.

Although the multi-stereo fusion method demonstrates superior DSM accuracy compared to the multi-view matching method, it also has certain limitations. For instance, the DSM accuracy can be significantly affected by the quantity and quality of matching image pairs. To validate this, the Explorer region dataset was used as an example, where the matching image pairs were ranked by quality, and the top n pairs were selected for matching and DSM fusion. The results are presented in Table 2.

Table 2 Influence of different number of matching image pairs on DSM fusion accuracy in multi-stereo fusion methods (Explorer)

Number of matching image pairs	Acc_{per}/%	Acc_H/m
5	86.37	1.28
10	86.43	1.26
15	85.65	1.37
20	85.08	1.43
25	84.28	1.46
30	82.83	1.51

The experiments show that for this dataset, both matching accuracy and elevation accuracy improve steadily as the number of matching image pairs increases up to 10 pairs. However, once the number exceeds 10, the accuracy begins to decline. When the number of matching image pairs reaches 30, the accuracy is even lower than that of the multi-view matching method. This indicates that an excessive number of matching image pairs may introduce redundant data and noise, leading to error propagation and negatively impacting fusion accuracy.

Therefore, in practical applications, it is recommended to select the number of matching image pairs based on the specific characteristics of the data to achieve a balance between accuracy and error control.

5 Conclusions

Mainstream approaches to dense image matching include multi-stereo fusion methods and multi-view matching methods. However, to date, no studies have systematically compared the accuracy of these two types of algorithms using high-resolution satellite imagery. In this paper, comparative experiments and analyses are conducted using datasets from three regions: the Shandong University of Science and Technology campus in China, the Hobart area in Australia, and the Explorer region in Argentina. The satellite data sources include Gaofen-7, IKONOS, and WorldView-3, covering both urban plains and mountainous areas.

Using average matching accuracy and elevation accuracy as evaluation metrics, this study considers both stereo reconstruction and multi-view reconstruction scenarios. The experimental results indicate that, for stereo datasets, the performance difference between the two methods is minimal. However, in multi-view datasets, when the number of matching image pairs is kept within a reasonable range, the geometric accuracy of the multi-stereo fusion method is significantly superior to that of the multi-view matching method. Across the experimental results of the three datasets, the average matching accuracy of the multi-stereo fusion method is approximately 1.2% higher than that of the multi-view matching method, while the elevation accuracy is improved by approximately 10.7%. Therefore, for 3D modeling applications using satellite imagery, the multi-stereo fusion method is recommended for the production of DSM products.

References

[1] QUAN Changwen, LI Zhenghong, PANG Baining, et al. Application of UAV image matching point clouds in monitoring of illegal land and illegal construction[J]. Bulletin of Surveying and Mapping, 2023(4): 111-114. (in Chinese)

[2] ZHONG Liang, ZHAO Shengzhi, CHEN Yanzhi. Application of UAV tilt photogrammetry in digital city construction[J]. Geomatics & Spatial Information Technology, 2023, 46(9): 202-204. (in Chinese)

[3] FENG Xiao. Application of UAV oblique photogrammetry technology in geological hazard monitoring: A

case study of the landslide in Diexi Town, Mao County, Sichuan Province[J]. Huabei Natural Resources, 2022(4): 98-101. (in Chinese)

[4] LIU Yang, LIAO Dongjun, WANG Chaogang, et al. 3D modeling of ancient buildings based on UAV close-range photography[J]. Bulletin of Surveying and Mapping, 2020(11): 112-115. (in Chinese)

[5] CHEN Xin. Application of oblique photogrammetry technology in power engineering[J]. Construction And Budget, 2023(10): 79-82. (in Chinese)

[6] JEONG W, PARK S Y. UGC-Net: Uncertainty-guided cost volume optimization with contextual features for satellite stereo matching[J]. Remote Sensing, 2025, 17(10): 1772.

[7] HE S, LI S, JIANG S, et al. HMSM-Net: Hierarchical multi-scale matching network for disparity estimation of high-resolution satellite stereo images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 188: 314-330.

[8] SCHARSTEIN D, SZELISKI R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms[J]. International Journal of Computer Vision, 2002, 47(1): 7-42.

[9] PENG J W, ZHOU Z Y, GUO J S, et al. Stereo matching with disparity space map compensation[J]. Arabian Journal for Science and Engineering, 2025. DOI 10.1007/s13369-025-10218-6.

[10] ZBONTAR J, LECUN Y. Stereo matching by training a convolutional neural network to compare image patches[J]. Journal of Machine Learning Research, 2016, 17: 1-32.

[11] MEI X, SUN X, ZHOU M, et al. On building an accurate stereo matching system on graphics hardware [C]//Proceedings of the 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). Barcelona, Spain: IEEE, 2011. DOI: 10. 1109/iccvw. 2011. 6130280.

[12] HIRSCHMULLER H. Stereo processing by semiglobal matching and mutual information[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(2): 328-341.

[13] LIN C, LI Y, XU G, et al. Optimizing ZNCC calculation in binocular stereo matching[J]. Signal Processing: Image Communication, 2017, 52: 64-73.

[14] HUANG X, ZHANG Y, YUE Z. Image-guided non-local dense matching with three-steps optimization[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2016, Ⅲ-3: 67-74.

[15] TANIAI T, MATSUSHITA Y, SATO Y, et al. Continuous 3D label stereo matching using local expansion moves[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2725-2739.

[16] HUANG X, QIN R. Post-filtering with surface orientation constraints for stereo dense image matching[J]. The Photogrammetric Record, 2020, 35(171): 375-401.

[17] REINARTZ P, MÜLLER R, HOJA D, et al. Comparison and fusion of DEM derived from SPOT-5 HRS and SRTM data and estimation of forest heights[J]. Proceedings of the Earsel Symposium. Porto, Portugal: [s.n.], 2005.

[18] KUSCHK G, D'ANGELO P, GAUDRIE D, et al. Spatially regularized fusion of multiresolution digital surface models[J]. IEEE Transactions on Geoscience and Remote Sensing, 2017, 55(3): 1477-1488.

[19] QIN R, LING X, FARELLA E M, et al. Uncertainty-guided depth fusion from multi-view satellite images to improve the accuracy in large-scale DSM generation [J]. Remote Sensing, 2022, 14(6): 1309.

[20] ZHANG Y, ZHANG Y, MO D, et al. Direct digital surface model generation by semi-global vertical line locus matching[J]. Remote Sensing, 2017, 9(3): 214.

[21] ZHU Z, STAMATOPOULOS C, FRASER C S. Accurate and occlusion-robust multi-view stereo[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2015, 109: 47-61.

[22] VOGIATZIS G, TORR P H S, CIPOLLA R. Multi-view stereo via volumetric graph-cuts[C]//Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR' 05). San Diego, CA, USA: IEEE, 2005.

[23] FURUKAWA Y, PONCE J. Accurate, dense, and robust multiview stereopsis[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32 (8): 1362-1376.

[24] GAO Yu, HU Zhaoling, FAN Ru. Effect of high-resolution image fusion algorithm on the classification precision of land utilization in coastal wetland[J]. Bulletin of Surveying and Mapping, 2022 (1): 116-120. (in Chinese)

[25] LIU Dongzhi, REN Yi, ZHU Wanxiong. Terrain-level geographic scene data production based on WorldView-3 satellite strip image[J]. Bulletin of Surveying and Mapping, 2023(9): 135-138. (in Chinese)

[26] LIU Min, HUANG Jinhui, PENG Zhenhua, et al. Extraction of shallow water bodies in Shenzhen using 0.5 m resolution satellite data[J]. Bulletin of Survey-

ing and Mapping, 2023(6): 146-149. (in Chinese)

[27] LIU Dongzhi, XU Qingling. GF-7 and ZY-3 satellites jointly serve airport clearance monitoring[J]. Bulletin of Surveying and Mapping, 2023(6): 138-141. (in Chinese)

[28] QIN Rongjun. Automated 3D recovery from very high resolution multi-view satellite images[C]//Proceedings of the ASPRS 2017 Annual Conference. [S.l.]: [s.n.], 2017: 12-16.

Authors

The first author Mr. **LIU Tengfei** received the B.S. degree in optoelectronic information science and engineering from Ocean University of China, Qingdao, China, in 2023. He is now a postgraduate student at Sun Yat-sen University, Zhuhai, China. His research interest is satellite photogrammetry.

The corresponding author Dr. **HUANG Xu** received the Ph.D. degree at Wuhan University in 2016. Now he is an associate professor at School of Geospatial Engineering and Science, Sun Yat-sen University. He has long been engaged in research on Gaofen satellite image data processing.

Author contributions Mr. **LIU Tengfei** designed the study and wrote the manuscript. Mr. **HUANG Zefeng** participated in the design and assisted with the experiments. Dr. **HUANG Xu** provided data and guided the research. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: SUN Jing)

基于高分辨率卫星数据的多立体融合方法和多视匹配方法精度对比和分析

刘腾飞，黄　旭，黄泽锋

（中山大学遥感科学与技术学院，珠海 519082，中国）

摘要：高分辨率亚米级卫星数据在实景三维中国建设当中发挥着越来越重要的作用，当前基于高分辨率卫星数据的三维重建研究主要包括多立体融合方法及多视匹配方法。目前基于这两种方法进行多视影像三维重建的算法已较为成熟，但针对卫星数据来分析比较多立体融合和多视匹配两类方法优劣的研究仍不多见。本文针对不同地区的高分辨率卫星数据集，对比分析了两类方法的实际精度。为了保证精度对比的公平性，两类方法均采用非局部密集匹配方法进行代价优化。结果表明，多立体融合方法的精度在各方面均优于多视匹配方法，在所用数据集中其平均匹配准确率提高约1.2%，而高程精度提高约10.7%。因此，在卫星数据三维建模应用中，推荐采用多立体融合方法进行数字表面模型（Digital surface model，DSM）产品生产。

关键词：多立体融合重建；多视匹配重建；非局部密集匹配方法；遮挡检测；高分辨率卫星数据