Comparative Analysis of Mainstream Image Matching Methods for Georeferencing Tianwen-1 HiRIC Imagery without Ground Control Points

Zhuolu Hou[1]; Aomei Zhang[1]; Yunfei Hu[1]; Xulei Shi[1]; Pei Mi[1]; Pengjie Tao[1, *]; Tao Ke[1]

[1]School of Remote Sensing and Information Engineering, Wuhan University, 430079, Wuhan, China - pjtao@whu.edu.cn

Keywords: Image Matching, Tianwen-1, Mars, Bundle Adjustment, Deep Learning.

Abstract

High-precision mapping of planetary surfaces, such as Mars, relies on matched control points derived from existing georeferenced data, as ground control points (GCPs) cannot be obtained through field measurement. However, image matchers like SIFT limit the robustness of this approach, particularly on texture-scarce and self-similar Martian terrain. While deep learning-based matchers offer a new paradigm, their performance gain for bundle adjustment remains inadequately quantified. This paper systematically evaluates four matchers (hand-crafted SIFT and deep learning-based DOG+HardNet+LightGlue, DISK+LightGlue, and LoFTR), assessing their impact on georeferencing tasks using Tianwen-1 high-resolution imagery. Deep learning methods, such as LoFTR, generate more correspondence points with a more uniform spatial distribution, halving the outlier rate and improving bundle adjustment accuracy by 10%. Our study provides a benchmark for planetary mapping and shows that powerful, learning-based image matchers are pivotal for next-generation automated mapping systems.

1. Introduction

Bundle Adjustment (BA) is the cornerstone of modern photogrammetry (Schönberger and Frahm, 2016), yet its efficacy is critically dependent on the quality of observations, namely control points and tie points. The quantity, precision, and spatial distribution of these observations significantly impact the robustness and accuracy of the entire automated pipeline.

In the past, classical algorithms like the Scale-Invariant Feature Transform (SIFT) succeeded in conventional scenarios, leading to the widespread perception that image matching had become a largely solved problem (Lowe, 2004). By detecting scale- and rotation-invariant keypoints and encoding their neighborhood gradients into a high-dimensional descriptor, SIFT provides a robust matching solution. Consequently, it has been widely adopted over the past two decades and deeply integrated into mainstream photogrammetric software, such as COLMAP (Schönberger and Frahm, 2016), and even official planetary data processing pipelines, such as USGS ISIS3 (Edmundson et al., 2012). However, the efficacy of such methods is contingent upon salient local gradient information. This dependency becomes their primary scale invariance. extended into challenging environments (Zhong et al., 2023), such as deserts, polar ice regions, and extraterrestrial surfaces like Mars. In these texture-deficient or highly self-similar landscapes, conventional matchers often fail to generate sufficient density and reliability for geometric correspondences, resulting in frequent failures during downstream optimization processes. Consequently, image matching has re-emerged not merely as a resolved challenge but as a critical bottleneck hindering the progress of automated photogrammetric systems.

Concurrently, a revolution in matching technology driven by deep learning has emerged as a promising solution to this bottleneck (Ma et al., 2021). This technological evolution has trended from optimizing individual components, such as using Convolutional Neural Networks (CNNs) to learn more discriminative feature descriptors, such as HardNet (Mishchuk et al., 2017), to developing end-to-end sparse matching methods that jointly optimize detection and description, such as SuperPoint (Detone et al., 2018) and DISK (Tyszkiewicz et al., 2020). More advanced paradigms have followed, such as the Graph Neural Network (GNN)-driven SuperGlue (Sarlin et al., 2020) and LightGlue (Lindenberger et al., 2023a), which frames feature matching as a graph matching problem, and the "detector-free" approach of LoFTR (Sun et al., 2021), which utilizes a Transformer architecture to capture long-range dependencies between dense pixels. These novel methods are achieving breakthrough performance in scenarios where traditional methods struggle.

Despite their notable success in computer vision, a significant technology gap persists: a systematic and quantitative assessment of their end-to-end performance improvement within conventional photogrammetric workflows, particularly in high-reliability applications, is still lacking. Currently, planetary remote sensing image matching still relies heavily on SIFT and its variants (Tao, 2022), while the application of deep learning has been predominantly concentrated on semantic tasks, such as automated crater detection (Lee, 2019), rather than on the core geometric processing for BA. To our knowledge, a study that systematically applies state-of-the-art deep learning matchers (e.g., LoFTR) to planetary imagery and provides an end-to-end evaluation of their true impact on the BA pipeline remains conspicuously absent from the literature.

Therefore, the primary objective of this study is to bridge this identified gap. We adopt a comprehensive evaluation framework to conduct a systematic investigation on one of today's most demanding scenarios: the georeferencing of high-resolution imagery from the Tianwen-1 Mars mission. This task involves orienting the imagery utilizing control derived from existing, multi-sensor georeferenced data (Smith et al., 2001), in the complete absence of ground control points (GCPs). The Martian surface poses a significant challenge, encapsulating a confluence of formidable issues, including non-linear radiometric distortions from varying illumination, vast low-texture regions, and morphologically similar impact craters (Woicke et al., 2018). This study examines the influence of bundle adjustment for three representative deep learning methods: 1) DoG-HardNet+LightGlue (representing the "hybrid" paradigm combining a classical detector with learned

* Corresponding Author

descriptors); 2) DISK (Tyszkiewicz et al., 2020)+LightGlue (representing the data-driven, end-to-end sparse matching paradigm); and 3) LoFTR (representing the dense, "detector-free" paradigm that diverges from the classical pipeline).

This paper offers three key contributions:

(1) We provide quantitative evidence of the performance leap afforded by modern matching techniques for automated Martian mapping. We demonstrate that an advanced image matcher significantly enhances matching robustness, nearly halving the outlier rate while also improving accuracy, as reflected by a 10% reduction in the reprojection error (RMSE).

(2) We establish a clear performance benchmark for algorithm selection in such challenging environments, delineating the distinct capability tiers from SIFT to LoFTR.

(3) We substantiate the thesis that a powerful, learning-based image matcher is pivotal for engineering robust and automated planetary mapping systems.

2. Method

This section elaborates on the core algorithms and mathematical models employed in this study. We begin by introducing the overall technical framework, followed by a detailed description of the various image matching techniques under evaluation. Finally, we present the rigorous sensor model and bundle adjustment formulation for line-scan push-broom imagery, which serves as the unified back-end.

2.1 Overall Technical Framework

The overall framework of this study is designed to establish a fair and comprehensive quantitative evaluation system, aimed at isolating and quantifying the impact of different front-end matching algorithms in the downstream task of image positioning. Figure 1 shows that the framework has three main parts. illustrated in Figure 1.

(1) Matched Control Points (MCPs) Generation: Initially, the Tianwen-1 HiRIC images (Li et al., 2021) are pre-processed and partitioned into tiles. Subsequently, the initial rational polynomial coefficients (RPCs) of the HiRIC are used to locate corresponding tiles from the Mars reference orthoimage, which is derived from CTX orthoimages (Malin et al., 2007). A coarse-to-fine matching strategy, employing the various matching algorithms, is then used to establish correspondences between the HIRISE and CTX orthoimages. Finally, by leveraging the georeferenced information of the CTX orthoimages and the MOLA (Smith et al., 2001) digital elevation model (DEM), these correspondences are converted into MCPs, each containing both image coordinates and object-space coordinates.

(2) Bundle Adjustment: The MCPs generated in the previous stage are fed as observations into a unified bundle adjustment module. This module, based on a rigorous sensor model, refines the exterior orientation parameters of the camera.

(3) Performance Evaluation: The performance metrics are collected at various stages of the pipeline to conduct a comprehensive, quantitative assessment of the different matching algorithms.

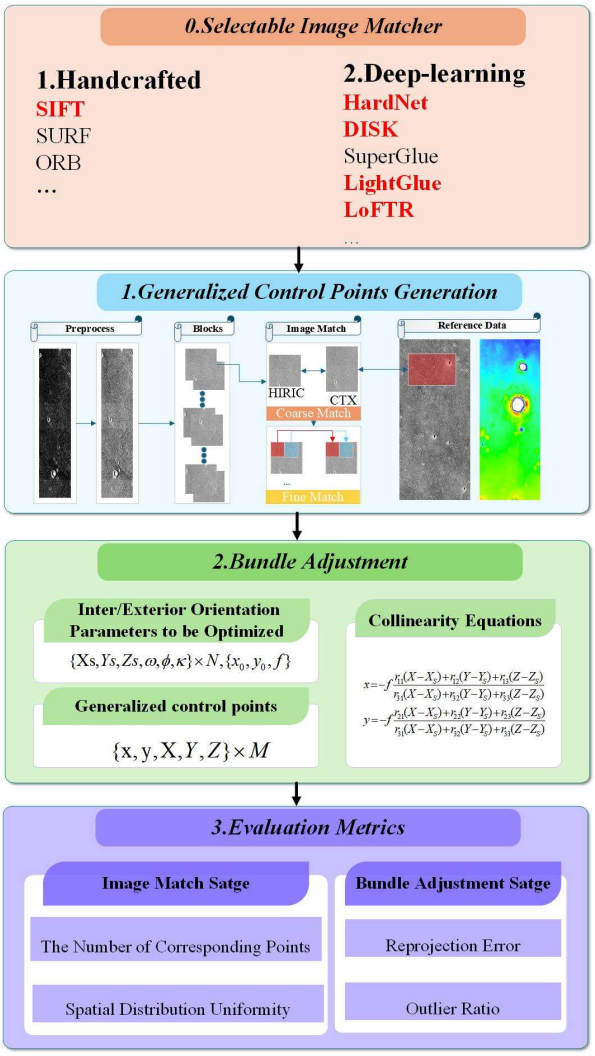

Figure 1. The flowchart of the proposed evaluation framework for image matchers. The framework consists of four main stages: (0) Selection of an image matcher from the selectable module (categorized as handcrafted and deep-learning based); (1) Generation of Matched control points (MCPs) using the selected matcher; (2) Bundle adjustment based on the collinearity equations; and (3) Calculation of evaluation metrics from both the image match and bundle adjustment stages.

2.2 Image Matching Techniques

The fundamental task of image matching is to establish stable and reliable correspondences between two images of the same scene captured from different viewpoints. Based on their technical paradigms, matching techniques can be broadly categorized into handcrafted and deep learning-based methods.

2.2.1 The Handcrafted Image Matching: Representing the classical approach, this study employs the SIFT (Lowe, 2004) algorithm. This method adheres to a well-defined three-stage pipeline:

(1) Feature Detection: This stage aims to identify keypoints in an image that are repeatable and invariant to changes in scale, rotation, and illumination. SIFT achieves this by constructing a Difference-of-Gaussians (DoG) pyramid to simulate scale-space

(2) Feature Description: For each detected keypoint, SIFT generates a high-dimensional numerical vector (a descriptor) by

https://doi.org/10.5194/isprs-annals-XI-2-2026-9-2026

computing a histogram of gradient orientations within the keypoint's neighborhood. This descriptor is used for subsequent similarity measurement.

(3) Feature Matching: Initial candidate matches are found by measuring the distance (e.g., Euclidean distance) between descriptors. To enhance matching accuracy, the Nearest Neighbor Distance Ratio (NNDR) or RANSAC (Fischler and Bolles, 1987) is typically used for initial filtering.

2.2.2 Learned Sparse Feature Matching: These methods follow the classical "Detect-Describe-Match" pipeline but replace one or more stages with deep learning-based networks to learn more robust and discriminative features. In this study, we fix the matcher to the highly efficient LightGlue and compare two different feature detection and description modules.

Hybrid Detection and Learned Description (DoG-HardNet): Keypoints are detected using the handcrafted difference-of-gaussians (DoG) operator, identical to the detection stage of SIFT. HardNet is a deep convolutional network trained with Metric Learning to output a 128-D descriptor. Through an optimized Triplet Loss function, HardNet (Mishchuk et al., 2017) maximizes the distance between non-matching features while minimizing it for matching ones, resulting in a more discriminative representation than classical descriptors like SIFT.

End-to-End Learned Detection and Description: DISK (Tyszkiewicz et al., 2020) represents a fully data-driven strategy. DISK utilizes a unified neural network, trained end-to-end with reinforcement learning, to jointly optimize keypoint detection and description. The design objective is to produce keypoints that are not only highly repeatable but whose descriptors also provide maximal information for the matching task. This integrated learning approach enables the detector and descriptor to work in better synergy, especially in challenging scenes.

Learned Image Matcher: For both methods above, we employ LightGlue (Lindenberger et al., 2023b), as the feature matcher. LightGlue is a lightweight yet powerful matching network based on Graph Neural Networks (GNNs) and attention mechanisms. It takes the keypoint sets and their descriptors from two images as input. Through self-attention and cross-attention, it considers both the local feature (from descriptors) and the global position to predict the optimal matches. Its efficient design allows it to achieve very high matching speeds while maintaining high accuracy.

2.2.3 Learned Detector-Free Dense Matching: LoFTR (Sun et al., 2021) completely abandons the keypoint detection stage. Firstly, it extracts coarse-level feature maps from two images using a CNN, then directly feeds these maps into a Transformer module. Leveraging the powerful self-attention and cross-attention mechanisms of the Transformer, LoFTR models dense correspondences on the coarse feature maps and obtains more precise matches on the fine feature maps. LoFTR has achieved breakthrough performance in low-texture regions compared to classical methods.

2.3 Rigorous Bundle Adjustment Model for Line-Scan Push-broom Cameras

Bundle adjustment is the most widely used method in photogrammetry for the optimization of camera parameters. It is essentially a large-scale, non-linear least-squares optimization problem, with the objective of minimizing the sum of reprojection errors of all matched control points across all images.

For a central-projection frame camera, the geometric imaging process is described by the collinearity equations:

$$x = -f\frac{r_{11}(X-X_s)+r_{12}(Y-Y_s)+r_{13}(Z-Z_s)}{r_{31}(X-X_s)+r_{32}(Y-Y_s)+r_{33}(Z-Z_s)}+x_0 \quad (1)$$

$$y = -f\frac{r_{21}(X-X_s)+r_{22}(Y-Y_s)+r_{23}(Z-Z_s)}{r_{31}(X-X_s)+r_{32}(Y-Y_s)+r_{33}(Z-Z_s)}+y_0 \quad (2)$$

Where (x_0, y_0, f) are the interior orientation parameters. (X, Y, Z) are the object point coordinates, and the exterior orientation parameters consist of the sensor's position (X_s, Y_s, Z_s) and attitude (described by the nine elements r_{ij} of the rotation matrix R).

For a line-scan push-broom sensor, imaging occurs line by line. Each line l has a unique imaging time t_l, and thus its exterior orientation parameters are functions of time, denoted as $(X_s(t), Y_s(t), Z_s(t))$ and $R(t)$.

These time-dependent position and attitude parameters are derived by interpolating a set of N discrete, high-precision observations. These observations are sampled at the specific imaging times t_i corresponding to N selected image lines, and are collectively defined as the Orientation Image Model (OIM). Each individual observation image lines within this OIM provides the parameter vector $P(t_l) = [X(t_l)_s, Y(t_l)_s, Z(t_l)_s, \varphi(t_l), \omega(t_l), \kappa(t_l)]$ at time t_l. The parameter $P(t_l)$ for any scan line l at its imaging time t_l can be computed by selecting k adjacent OIM nodes and applying the Lagrange polynomial interpolation formula:

$$P(t_i) = \sum_{j=k}^{k+m-1} P_j . L_j(t_i) \quad (3)$$

where $L_j(t_i)$ is the Lagrange basis polynomial, defined as:

$$L_j(t_i) = \prod_{q=k,\ q\neq j}^{k+m-1} \frac{t_i - t_q}{t_j - t_q} \quad (4)$$

Using these equations, we can precisely solve for the exterior orientation parameters corresponding to any image line, thereby constructing a complete and time-dependent rigorous sensor model. In the bundle adjustment, the set of unknown parameters to be optimized comprises both the interior and exterior orientation parameters.

Specifically, we estimate a single set of interior orientation parameters (IOP) and N sets of exterior orientation parameters for the entire image strip. IOP includes the principal point offsets (x_p, y_p) and the focal length (f). The N sets of EOPs represent the sensor's position (X_s, Y_s, Z_s) and attitude (ω, ϕ, κ) sampled at N predetermined, constant scan line numbers $(l_1, l_2, \ldots, l_n)$ distributed throughout the imaging.

3. Experiment

We present a systematic experimental design to quantitatively evaluate the performance of four distinct image matching algorithms on the georeferencing task for Mars Tianwen-1 imagery, including the classical handcrafted method SIFT and

three deep learning-based methods, DoG + HardNet + LightGlue, DISK + LightGlue, and LoFTR. The three deep learning-based methods are further subdivided into three paradigms: hybrid matching (DoG + HardNet + LightGlue) combining geometric detectors with learned descriptors, end-to-end sparse matching (DISK + LightGlue), and detector-free dense matching (LoFTR).

3.1 Datasets

The Martian surface poses a formidable challenge for image matching algorithms due to its sparse texture, harsh illumination conditions (including large, sharp shadows), and the repetitive nature of geomorphological features (e.g., impact craters).

Data: The dataset comprises two High-Resolution Imaging Camera (HiRIC) images from China's Tianwen-1 mission, specifically covering the Zhurong rover landing site. These images feature a ground sampling distance (GSD) of approximately 0.5 meters. The raw HiRIC data are publicly accessible via the China Lunar and Deep Space Exploration Data Release Service (https://moon.bao.ac.cn).

To achieve georeferencing without GCPs, we introduced two widely recognized global Martian datasets as reference baselines:

(1) The 5 m/px global image mosaic generated by the Context Camera (CTX) aboard NASA's Mars Reconnaissance Orbiter (MRO).

(2) The ~463 m/px global DEM from the Mars Orbiter Laser Altimeter (MOLA) on the Mars Global Surveyor (MGS).

3.2 Evaluation Framework

We constructed an evaluation framework to quantify the performance of four representative image matching paradigms and their subsequent impact on bundle adjustment. The algorithms compared are the classical SIFT benchmark, the "hybrid" DoG-HardNet+LightGlue (classical detector +learned descriptor/matcher), the end-to-end sparse DISK+LightGlue and the end-to-end "detector-free" LoFTR.

The framework's first stage involves a rigorously standardized matched control points extraction. As a critical preprocessing step, all HiRIC images (~0.5 m) are downsampled to the 5 m/px resolution of the CTX reference DOM, which eliminates challenging scale ambiguities. A coarse-to-fine matching strategy is then applied to generate corresponding points, which are subsequently converted into MCPs using the reference planimetric and vertical data. For implementation, all deep learning methods used their official pre-trained models and default parameters, while SIFT's feature detection was dynamically optimized (5k-50k keypoints) to maintain competitiveness.

In the second stage, bundle adjustment, the MCPs from each method are input into an identical, fixed-parameter BA engine (Ceres Solver) using the exact same initial poses. The final stage is the metrics calculation. This entire pipeline, from matching to adjustment, is designed to isolate the matcher as the sole variable, ensuring all observed performance differences are attributable only to the algorithm itself.

3.3 Evaluation Metrics

We defined quantitative metrics for both the matching and adjustment stages.

The Metrics of Image Matching: 1). The Number of Corresponding Points (NCP): NCP is the count of valid corresponding points between two views after RANSAC robust estimation. 2). Spatial Distribution Uniformity (SDU): We employ a Multi-Scale Grid Coverage metric, defined as the mean coverage $\mathrm{Cov(s)}$ from three distinct grid resolutions, $\mathrm{s} = \{32\mathrm{x}32, 64\mathrm{x}64, 128\mathrm{x}128\}$. Let P be the set of valid corresponding points2. For a given scale $s \in S$, the image is divided into a grid $\mathrm{G_s}$ with $N_s = s \times s$ total cells. The coverage $\mathrm{Cov(s)}$ for a single scale s is calculated as:

$$Cov(s) = \frac{1}{N_s}\sum_{i=1}^{N_s} I(C_i) \tag{5}$$

where C_i is the i-th cell in the grid $\mathrm{G_s}$, and $I(C_i)$ is an indicator function defined as:

$$I(C_i) = \begin{cases} 1, & if\ C_i\ contains\ at\ least\ one\ point\ p \in P \\ 0, & otherwise \end{cases} \tag{6}$$

The final SDU score is the mean of the coverages from all three scales:

$$SDU = \frac{1}{3}\sum_{s \in S} Cov(s) \tag{7}$$

The Metric of Bundle Adjustment: 1). Reprojection Error: To evaluate the bundle adjustment, we first compute the reprojection error for each of the n inlier observations. Let $(\mathrm{x_{obs,k}}, \mathrm{y_{obs,k}})$ be the observed image coordinates of the k-th point, and $(\mathrm{x_{porj,k}}, \mathrm{y_{proj,k}})$ be its coordinates re-projected using the final optimized model parameters. We compute four statistical metrics: The Root Mean Square Error (RMSE), Mean, Max, and Min; RMSE is then formulated as:

$$RMSE = \sqrt{\frac{1}{n}\sum_{k=1}^{n}\left(\left(x_{obs,k} - x_{obs,k}\right)^2 + \left(\mathrm{y_{porj,k}} - \mathrm{y_{porj,k}}\right)^2\right)} \tag{8}$$

2). Outlier Ratio: The outlier ratio (OR) is the percentage of matched control points rejected (outliers) during the bundle adjustment process. Let N_{total} be the total number of corresponding points (NCP) fed into the BA module, and $N_{inliers}$ be the number of inliers. A matched control point is classified as an inlier if its final reprojection error is less than three times the overall Root Mean Square Error (RMSE) of the bundle adjustment. The ratio is defined by the follow equation:

$$OR = \frac{N_{total} - N_{inliers}}{N_{total}} \times 100\% \tag{9}$$

3.4 Image Matching Performance

3.4.1 Quantitative Comparison: The results in Table 1 clearly reveal the performance hierarchy of the four matching algorithms in the challenging Martian scenario. LoFTR exhibits superior overall performance, generating 3-to-4 times more valid corresponding points than other deep learning methods and an order of magnitude more than SIFT. Coupled with its excellent spatial uniformity (over 93%), this result establishes it as the premier choice for robust, automated reconstruction.

On the other hand, the DoG + HardNet + LightGlue hybrid method possesses the highest spatial uniformity (up to 97–98%), making it the ideal choice for applications prioritizing completeness of spatial coverage.

In contrast, the performance of DISK + LightGlue is highly unstable, with significant matching failures in several scenes leading to poor spatial distribution (as low as 63.28%), rendering it ill-suited for high-precision mapping tasks. Finally, the classical SIFT benchmark produces the fewest points, and its effectiveness in sparse-texture environments like Mars has been fully superseded by modern learning-based approaches.

Data	Metric	Image Matching Method			
		SIFT	DH+LG	DISK+ LG	LoFTR
24_1	NCP	65681	209613	239851	746730
	SDU	76.23	97.14	89.44	93.64
24_2	NCP	72221	205574	217815	738254
	SDU	84.42	97.79	86.47	93.56
24_3	NCP	94621	237246	189154	766827
	SDU	88.63	98.11	63.28	93.10
26_1	NCP	73723	193756	189081	698145
	SDU	87.34	97.02	80.38	93.02
26_2	NCP	74376	178452	195612	662274
	SDU	92.72	97.83	88.44	93.97
26_3	NCP	99325	224119	169220	836371
	SDU	95.11	98.34	65.03	93.46

Table 1. Quantitative comparison for the four image matching algorithms across six scenes. DH and LG stand for DoG+HardNet and LightGlue, respectively. The test scenes (24_CCD1, 2, 3 and 26_CCD1, 2, 3) represent the three CCD images acquired by the Tianwen-1 mission on 2021-03-24 and 2021-03-26, respectively. For each metric (row), the best-performing result is highlighted in red, and the second-best result is shown in blue.

3.4.2 Qualitative Comparison: Figure 2 shows the coarse-to-fine matching results of four methods on a HiRIC sub-image block. SIFT fails completely in the coarse matching stage (a), which prevents it from obtaining any valid matches in the fine matching stage (b). In contrast, the other three methods successfully establish robust coarse correspondences. Among all methods, LoFTR (g, h) demonstrates clear superiority. Specifically, the correspondences of LoFTR in both the coarse (g) and fine (h) stages are the most superior in terms of both quantity and spatial uniformity.
Figure 3 and 4 show the spatial distribution of MCPs extracted by four different methods across three CCD image strips (20210326). Among them, LoFTR (d) exhibits a clear overall advantage, significantly outperforming the other methods in both the number and spatial distribution of the MCPs. In contrast, SIFT (a) and DISK+LG (c) exhibit poorer performance, with MCPs missing in some local areas. Specifically, SIFT (a) extracts the fewest points, which highlights its strong reliance on gradient information. This results in difficulties in areas of the Martian surface that have weak or repetitive textures. DISK+LG (c) shows the poorest distribution, a failure clearly attributable to a domain gap between its natural-image training and the distinct features of Martian remote sensing imagery. The matching control points of DH+LG (b) are also evenly distributed, but the number lags significantly behind that of LoFTR.

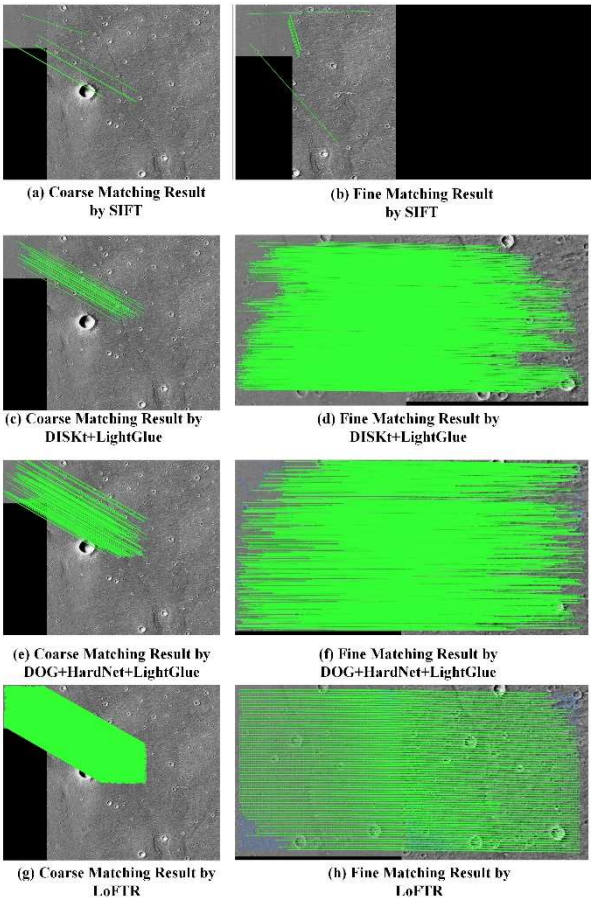

Figure 2. Coarse-to-fine matching results comparison of four methods on a HiRIC sub-image block and a CTX DOM sub image block. Each row presents one method: (a, b) SIFT, (c, d) DISK+LightGlue, (e, f) DOG+HardNet+LightGlue, and (g, h) LoFTR. The left column (a, c, e, g) displays the results from the coarse matching stage, while the right column (b, d, f, h) displays the results from the fine matching stage.

3.5 Bundle Adjustment Performance

We evaluated the performance of four image matching algorithms within a bundle adjustment (BA) framework. Table 2 compares the reprojection error and outlier rate of the MCPs provided by each algorithm, as determined by the bundle adjustment.

The results in Table 2 demonstrate that the quality of MCPs provided by different matching algorithms decisively impacts the final accuracy and robustness of the bundle adjustment. It

must first be noted that the overall Root Mean Square Error (RMSE) for all methods lies in the 6~8 pixel range. The result is primarily attributed to the significant ~10 times resolution disparity between the reference CTX DOM (~5 m) and the target Tianwen-1 imagery (~0.5 m), which amplifies the pixel-level reprojection error from the lower-resolution reference. Among the four methods, the end-to-end deep learning frameworks, DISK+LG and LoFTR, exhibit the best overall

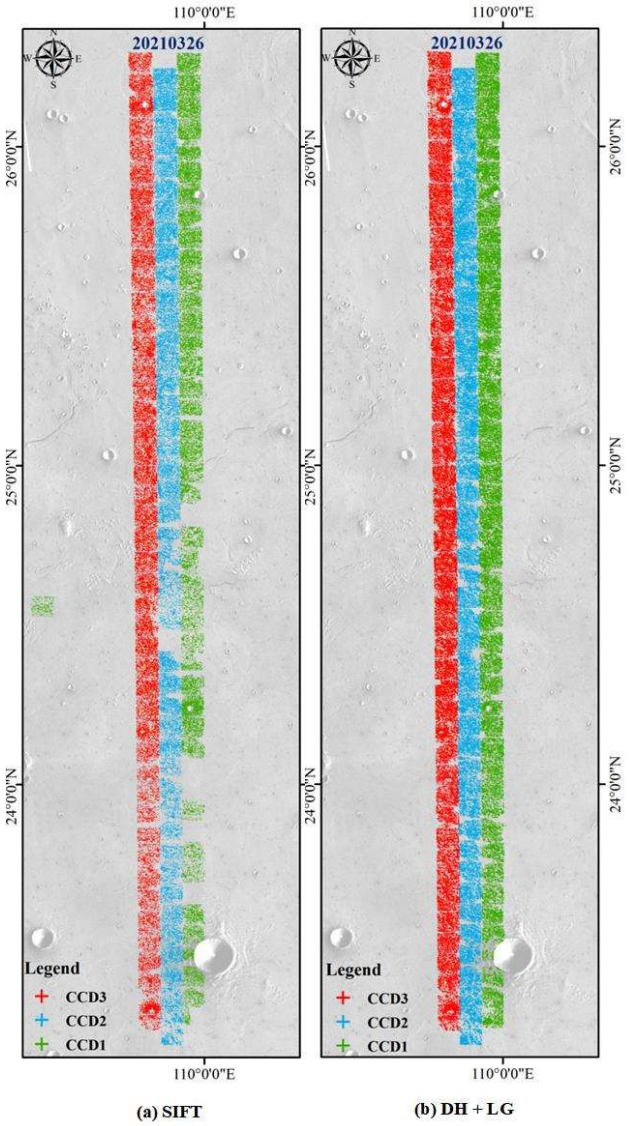

Figure 3. The distribution of Matched Control Points (MCPs) for the SIFT and DoG + HardNet + LightGlue (DH+LG) on the three-line push-broom imagery acquired on March 26, 2021. Red, blue, and green represent the MCPs of CCD3, CCD2, and CCD1, respectively.

performance, while the handcrafted SIFT algorithm is the worst. On the other hand, SIFT gives convergent constraints but has the highest RMSE (7.6–7.7 pixels) and the highest outlier rate (7.3%). This reflects the limited discriminative power of descriptors and localization accuracy of keypoints in complex scenes. In contrast, the DH+LG (hybrid) method reveals a critical insight: by incorporating the advanced HardNet and LigheGlue matcher, it significantly improves matching robustness, reducing the outlier rate to 4.1%–4.7%. However, its accuracy (RMSE) shows no improvement. This phenomenon demonstrates that the final accuracy bottleneck lies in the keypoint detection stage, the limited localization precision of the DoG detector. This accuracy bottleneck is successfully overcome by DISK+LG and LoFTR, both of which achieve the lowest RMSE (6.7–6.8 pixels) alongside an excellent low outlier rate. There are two key factors at play: DISK optimizes keypoint detection and description through learning, enhancing localization accuracy at the source; LoFTR achieves high-precision matching with its detector-free Transformer architecture.

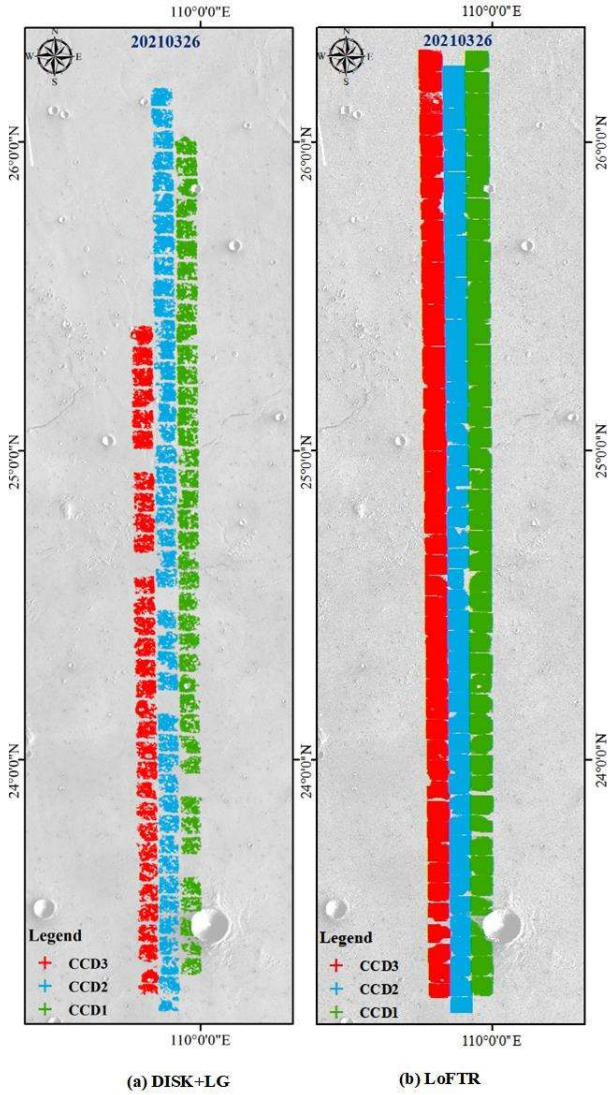

Figure 4. The distribution of Matched Control Points (MCPs) for DISK + LightGlue (DISK + LG) and LoFTR. on the three-line push-broom imagery acquired on March 26, 2021. Red, blue, and green represent the MCPs of CCD3, CCD2, and CCD1, respectively.

Date	Metric	Method			
		SIFT	DH -LG	DISK -LG	LoFTR
324	RMSE (px)	7.6	7.8	6.8	6.8
	Mean (px)	6.7	6.7	6.08	6.07
	Max_X (px)	16.9	16.8	14.8	14.3
	Max_Y (px)	15.2	15.3	14.3	13.4

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-2-2026-9-2026 | © Author(s) 2026. CC BY 4.0 License.

	Min_X (px)	-17.0	-16.8	-14.8	-14.2
	Min_Y (px)	-15.2	-15.3	-14.2	-13.9
	OR (%)	7.3	4.1	4.2	4.4
326	RMSE (px)	7.7	7.4	6.9	6.7
	Mean (px)	6.6	6.6	6.2	5.9
	Max_X (px)	15.9	15.6	14.6	14.0
	Max_Y (px)	15.9	15.9	14.9	14.3
	Min_X (px)	-15.9	-15.6	-14.6	-14.0
	Min_Y (px)	-15.9	-15.9	-14.9	-14.3
	OR (%)	7.3	4.7	4.1	4.4

Table 2. Comparison of Bundle Adjustment (BA) performance using the Matched Control Points (MCPs) derived from the four matching algorithms. DH and LG stand for DoG+HardNet and LightGlue, respectively. The "324" and "326" scenes represent the two Tianwen-1 mission blocks acquired on 2021-03-24 and 2021-03-26, respectively. For each image strip, the three CCD images were adjusted together as a single unit. For all metrics, the best-performing result is highlighted in red, and the second-best result is highlighted in blue.

In summary, the experimental results demonstrate that deep learning-based image matching methods generate more accurate and robust MCPs than handcrafted methods, making them the optimal choice for enhancing bundle adjustment performance.

4. Discussion

Our experimental results consistently demonstrate that the performance of the handcrafted SIFT degrades significantly in the Martian environment with sparse, repetitive textures and non-linear radiometric distortions. In contrast, deep learning-based methods are more adept at handling these image distortions, successfully extracting a large number of high-quality and evenly distributed correspondences. As shown in Table 2, when these high-quality observations are used as MCPs, they directly and significantly improve the accuracy and robustness of the bundle adjustment (BA).

However, these results reveal a critical accuracy limitation, 6–8 pixel RMSE in BA, dictated by the resolution of reference imagery. A 1-pixel reprojection error in the CTX data (5 m) will theoretically project to a 10-pixel error in the HiRIC image (0.5 m). More importantly, the vertical reference (Z) for the MCPs is derived from the ultra-low-resolution MOLA DEM (~463 m/px). The substantial elevation uncertainty from this source inevitably propagates into the final projection accuracy. Therefore, the observed 6–8 pixel RMSE actually signifies a high-precision, sub-pixel match (approx. 0.6~0.8 pixels relative to the reference source), which is already approaching the theoretical accuracy limit imposed by this multi-source, multi-resolution reference dataset.

The DH+LG (DoG + HardNet + LightGlue) method presented a contradiction: it achieved the highest Spatial Distribution Uniformity (SDU) (Table 1) while simultaneously yielding the worst BA RMSE (Table 2) among the deep learning-based methods, performing only marginally better than SIFT. The high SDU score indicates that the method possesses remarkable robustness in establishing coarse matches across image blocks, successfully covering the vast majority of the area. However, its poor localization accuracy (high RMSE) likely stems from the limitations of its handcrafted keypoint detector. We infer that the DoG detector is highly sensitive to the inherent sensor noise within the CTX reference imagery, leading to this insufficient keypoint localization.

In contrast, the performance of DISK+LG exposes the risks of deep learning-based methods. Its unstable SDU scores (dropping as low as 63% in some scenes) collectively demonstrate a severe deficit in robustness of matching sub-image blocks. This instability may likely be attributed to the classic domain gap problem: DISK's training strategy (likely on terrestrial datasets) may have caused it to overfit to Earth-based features, failing to generalize to Mars's alien geomorphology. This failure underscores a key point: not all deep learning methods are universally applicable without domain-specific fine-tuning.

Finally, we acknowledge the limitations of our experimental design. In this study, our dataset lacks significant rotational variations between the multi-source Mars sub-image blocks. This experimental setup inadvertently suppressed SIFT's core strength (i.e., rotation invariance) and masked the potential weakness of learning-based methods in handling such transformations. However, scale and rotational variations are typically compensated for using the known attitude parameters of satellite imagery.

5. Conclusion

Our systematic evaluation on the challenging georeferencing task of Tianwen-1 HiRIC imagery suggests that image matchers impact the performance of bundle adjustment (BA) in two primary aspects: accuracy and robustness. To ensure a fair comparison, all matching methods were benchmarked within our unified coarse-to-fine image matching framework.

Specifically, the hybrid method combining a handcrafted detector (e.g., DoG) with a deep learning-based descriptor exhibited strong matching robustness, but suffered from poor reprojection error due to the inherent localization inaccuracies of the DoG detector. In contrast, end-to-end frameworks, exemplified by LoFTR, addressed both challenges more effectively. They not only generated more corresponding points with a evenly spatial distribution but also significantly reduced the final RMSE in BA.

Therefore, developing deep learning-based image matching capable of generating large quantities of accurate and robust corresponding points is critical for georeferencing task of Tianwen-1 Imagery. In the future, we will explore the application of these deep learning-based techniques to large-scale planetary terrain reconstruction.

References

DeTone, D., Malisiewicz, T., and Rabinovich, A., 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 337-33712. 10.1109/CVPRW.2018.00060.

Edmundson, K., Cook, D., Thomas, O., Archinal, B., and Kirk, R., 2012. Jigsaw: The ISIS3 bundle adjustment for extraterrestrial photogrammetry. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 1, 203-208.

Fischler, M. A. and Bolles, R. C.: Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, in: Readings in Computer Vision, edited by: Fischler, M. A., and Firschein, O., Morgan Kaufmann, San Francisco (CA), 726-740, https://doi.org/10.1016/B978-0-08-051581-6.50070-2, 1987.

Lee, C., 2019. Automated crater detection on Mars using deep learning. *Planetary and Space Science*, 170, 16-28.

Li, C., Zhang, R., Yu, D., Dong, G., Liu, J., Geng, Y., Sun, Z., Yan, W., Ren, X., and Su, Y., 2021. China’s Mars exploration mission and science investigation. *Space Science Reviews*, 217, 57.

Lindenberger, P., Sarlin, P.-E., and Pollefeys, M., 2023a. LightGlue: Local Feature Matching at Light Speed. *arXiv preprint arXiv:2306.13643*.

Lindenberger, P., Sarlin, P. E., and Pollefeys, M., 2023b. LightGlue: Local Feature Matching at Light Speed. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 17581-17592. 10.1109/ICCV51070.2023.01616.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60, 91-110, 10.1023/b:Visi.0000029664.99615.94.

Ma, J., Jiang, X., Fan, A., Jiang, J., and Yan, J., 2021. Image matching from handcrafted to deep features: A survey. *International Journal of Computer Vision*, 129, 23-79.

Malin, M. C., Bell III, J. F., Cantor, B. A., Caplinger, M. A., Calvin, W. M., Clancy, R. T., Edgett, K. S., Edwards, L., Haberle, R. M., and James, P. B., 2007. Context camera investigation on board the Mars Reconnaissance Orbiter. *Journal of Geophysical Research: Planets*, 112.

Mishchuk, A., Mishkin, D., Radenovic, F., and Matas, J., 2017. Working hard to know your neighbor's margins: Local descriptor learning loss. *Advances in neural information processing systems*, 30.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., and Rabinovich, A., 2020. Superglue: Learning feature matching with graph neural networks. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4938-4947.

Schönberger, J. L. and Frahm, J. M., 2016. Structure-from-Motion Revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104-4113. 10.1109/CVPR.2016.445.

Smith, D. E., Zuber, M. T., Frey, H. V., Garvin, J. B., Head, J. W., Muhleman, D. O., Pettengill, G. H., Phillips, R. J., Solomon, S. C., and Zwally, H. J., 2001. Mars Orbiter Laser Altimeter: Experiment summary after the first year of global mapping of Mars. *Journal of Geophysical Research: Planets*, 106, 23689-23722.

Sun, J., Shen, Z., Wang, Y., Bao, H., and Zhou, X., 2021. LoFTR: Detector-free local feature matching with transformers. *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8922-8931.

Tao, Y.: On Martian surface exploration: development of automated 3D reconstruction and super-resolution restoration techniques for Mars orbital images, UCL (University College London), 2022.

Tyszkiewicz, M., Fua, P., and Trulls, E., 2020. Disk: Learning local features with policy gradient. *Advances in neural information processing systems*, 33, 14254-14265.

Woicke, S., Moreno Gonzalez, A., El-Hajj, I., Mes, J., Henkel, M., and Klavers, R., 2018. Comparison of crater-detection algorithms for terrain-relative navigation. *2018 aiaa guidance, navigation, and control conference*, 1601.

Zhong, J., Yan, J., Li, M., and Barriot, J.-P., 2023. A deep learning-based local feature extraction method for improved image matching and surface reconstruction from Yutu-2 PCAM images on the Moon. *ISPRS Journal of Photogrammetry and Remote Sensing*, 206, 16-29.