

## Evaluating the synergy of hand-crafted and AI-driven feature matching in Structure-from-Motion 3D Reconstruction

Min-Lung Cheng<sup>1</sup>, Yasutaka Kuramoto<sup>1</sup>

<sup>1</sup>SkymatiX Inc., 103-0021 Tokyo, Japan - (minlung.cheng, yasutaka.kuramoto)@skymatix.co.jp

**Keywords:** Feature matching, Artificial Intelligence, Structure-from-Motion, Image alignment

### Abstract

Matching images containing repetitive patterns or uniform textures remains a persistent challenge in structure-from-motion (SfM) 3D reconstruction, as the quality of feature correspondences directly influences camera pose estimation and 3D point cloud generation. Robust feature matching is therefore essential for achieving reliable SfM results. This study investigates the impact of integrating traditional hand-crafted and AI-driven feature-matching approaches on SfM reconstruction performance. SIFT and SuperPoint are selected as representative feature extractors and combined with multiple matching algorithms, including FLANN, LightGlue, and StereoGlue, to evaluate their effectiveness. The resulting feature matches are assessed using three quantitative indicators: (1) the number of successfully matched image pairs, (2) the number of correctly aligned images, and (3) positional errors of the SfM reconstruction relative to GNSS meta-information. The experimental results demonstrate that the combination of SIFT and LightGlue consistently yields the highest number of correctly aligned images and the lowest positional errors across all test cases. StereoGlue produces partially reliable matches; however, its performance is sensitive to the characteristics of the training data. In contrast, while the SIFT–FLANN combination generates a large number of matched pairs, it frequently results in misalignments due to false correspondences in scenes with uniform textures. These findings indicate that coupling hand-crafted feature extraction with AI-assisted matching strategies provides a robust and effective solution for improving SfM reconstruction in challenging imagery datasets.

### 1. Introduction

Optical imagery acquired using unmanned aerial vehicles (UAV, or drones) has been widely recognized as a low-cost and effective means for a diverse range of environmental observation tasks and spatial applications. Representative applications include seasonal forest monitoring and change detection, disaster response and decision-making support, city planning and infrastructure maintenance, among others (Chen et al., 2025, Li et al., 2025). To meet the demands of these diverse spatial applications, extracting reliable spatial information from UAV images has become a critical task. Structure-from-Motion and Multi-View Stereo (SfM-MVS) constitute a widely-used approach for deriving spatial information from image data (Berra and Peppas, 2020). Given a set of images, SfM aims to estimate camera poses - namely exterior orientation parameters (EOPs) - by identifying and matching conjugate (common) feature points across images. The subsequent MVS stage exploits the estimated camera poses to complete dense 3D scene reconstruction. Within this two-stage workflow, SfM serves as the entry point and plays a decisive role in determining the success and quality of the final 3D scene reconstruction.

Although modern drone platforms are typically equipped with a Global Navigation Satellite System (GNSS) and Inertial Measurement Unit (IMU), the instantaneous EOPs recorded by these sensors are often insufficient to meet the accurate requirements, depending on the instrument performance (Lin and Chen, 2025). Instead, common feature points observed across multiple images provide a robust basis for estimating EOPs, as each feature point corresponds to a unique location in the real-world space. Bundle adjustment subsequently performs space resection and intersection simultaneously to approximate and refine the EOPs and compute sparse 3D points from these conjugate feature points. Consequently, the reliable identification of com-

mon feature points in images is regarded as one of the most critical stages in the SfM pipeline.

To reduce labor force requirements and time consumption, semi-automatic or fully automatic methods have been developed alongside rapid advances in computer vision. Both hand-crafted and AI-driven approaches can automate the processes of feature extraction and matching. In hand-crafted methods, the Scale-Invariant Feature Transform (SIFT) (Lowe, 1999) and its variants are capable of automatically detecting image features using predefined algorithms or detectors, such as the Difference of Gaussian (DoG). Feature descriptors are subsequently computed to characterize each feature point in image space. In contrast, AI-driven approaches rely on training data to detect feature points and generate their corresponding descriptors. Such methods are generally detector-free; however, their performance is strongly influenced by the quality and diversity of the training data.

Automatic feature matching is to identify common feature points appearing in different images by evaluating the similarity of feature descriptors without human intervention (Cheng and Matsutaka, 2021). Image scenes containing diverse objects exhibit greater feature variability, resulting in more discriminative feature descriptors. This condition improves the reliability of feature matching and, consequently, enables more robust camera pose estimation in SfM. In contrast, images dominated by repetitive patterns or uniform textures tend to produce more homogeneous feature descriptors. Such scenarios weaken the capability of automatic feature matching algorithms to correctly establish correspondences across images, as the descriptors become too similar to be reliably distinguished. Mismatches or an insufficient number of valid correspondences can significantly deteriorate the SfM performance, particularly with respect to camera pose estimations.

This study aims to integrate and evaluate the synergy between hand-crafted and AI-driven feature extraction and matching algorithms to address challenging imagery dominated by uniform textures in SfM-based 3D reconstruction (also known as image alignment). As representative feature types, SIFT and SuperPoint (Tone et al., 2018) are selected to represent hand-crafted and AI-driven approaches, respectively. With regard to matching algorithms, FLANN, LightGlue (Lindberger et al., 2023), and StereoGlue (Barath et al., 2024) are adopted to examine their capacities in establishing reliable feature correspondences. The resulting matches are subsequently employed in SfM-based 3D reconstruction to assess their impacts on practical camera pose estimations.

## 2. Methodology

### 2.1 Incremental Structure-from-Motion

Using two-dimensional (2D) images is one of the most effective and low-cost approaches for 3D reconstruction. Among the various algorithms developed, SfM has been widely applied for many years due to its fully automatic processing capability. Figure 1 illustrates a standard SfM image-processing workflow, along with selected feature types and matching algorithms employed in this study. The initial stage involves extracting feature points and computing feature descriptors. Subsequently, matching algorithms are utilized to match these descriptors in order to identify common feature points appearing in the selected image pairs (Cheng et al., 2024). Feature extraction and matching are generally regarded as the most time-consuming stages when prior knowledge is unavailable. Moreover, the presence of both inliers (correct matches) and outliers (mismatches) in the matching results directly affect the reliability of SfM-based 3D reconstruction, particularly the accuracy of the estimated camera poses.

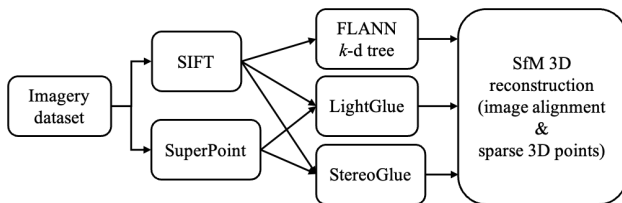


Figure 1. Structure-from-Motion workflow with various feature types and matching algorithms.

Incremental SfM 3D reconstruction is initialized by selecting an image pair, using their feature correspondences, and exploiting epipolar geometry to approximate their relative camera poses. The essential matrix, estimated using the five-point method, is used to determine the most likely solution by triangulating all matches (space intersection) and examining the resulting point positions with respect to the camera poses. The triangulated points are generally located in front of both cameras, as shown in Figure 2(a). After this initialization, the remaining images can be incrementally incorporated into the SfM reconstruction by adding one image at each step. In this study, the next best view (NBV) selection strategy is employed to determine which image to add, based on the number of triangulated points that can be projected onto candidate images. Figure 2(b) illustrates the use of space resection to estimate the camera pose of the newly added image, which can be solved using the perspective- $n$ -points (PnP) algorithm (Qiao et al., 2023).

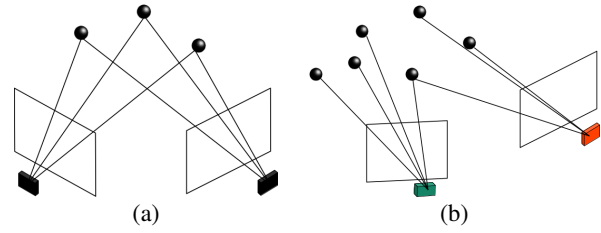


Figure 2. SfM 3D reconstruction (a) initialization by using two images (b) next best view selection based on the number of 3D points that be projected onto the new image.

Bundle adjustment, which combines space intersection and space resection, iteratively computes updated 3D points and estimates camera pose by minimizing the objective function shown in Equation (1). The purpose of this function is to minimize the 2D projection error between all measurements and observations. The measurements correspond to the matched feature points, whereas the observations represent the feature points projected using the estimated camera poses and 3D points. In the equation, a binary operator is used to indicate whether a 3D point is visible in a given image, where 0 denotes invisibility and 1 denotes visibility. The Levenberg-Marquardt (LM) algorithm is adopted to jointly optimize the camera poses and 3D points.

$$\min_{a_j, p_i} \sum_{i=1}^n \sum_{j=1}^m g_{i,j} (Q(a_j, P_i) - p_{i,j}) \quad (1)$$

where

$a_j = j$ -th camera

$P_i = i$ -th 3D point

$p_{i,j} = i$ -th 3D point on  $j$ -th camera

$g_{i,j} =$  binary operator, 0 or 1

$Q =$  prediction of projecting a 3D point onto an image

### 2.2 Feature Types

Although SIFT has been available for more than two decades, it remains the gold standard for automatic feature recognition (Lowe, 1999). Prior knowledge of defining feature points is not required, as SIFT uses the Difference of Gaussian (DoG) to identify the most stable feature points within an image at the sub-pixel level. SIFT employs image gradient magnitudes and orientation histograms to construct 128-dimensional descriptors that characterize each feature point. One of the most significant properties of SIFT is its invariance to scale and rotation, which ensures robustness and adaptability across diverse imaging scenarios and datasets.

SuperPoint is a learning-based, self-supervised feature extraction approach that does not require human-generated annotations (Tone et al., 2018). The base interest point detector utilized for training is MagicPoint. An encoder is employed to generate a multi-scale image pyramid prior to feature extraction. Unlike SIFT, which detects feature points and then computes their descriptors, SuperPoint adopts a dual-head architecture, consisting of interest point decoder and descriptor decoder, which operate in parallel. During training, homographic adaptation is introduced to improve the invariance of interest points to viewpoint changes. Consequently, SuperPoint is not fully invariant to scale and rotation, and the detected feature points are generally localized at the pixel-level. In addition, SuperPoint produces 256-dimensional descriptors, and its performance is constrained by the diversity and quality of the pre-trained data.

### 2.3 Matching Algorithms

The Fast Library for Approximate Nearest Neighbors (FLANN) is one of the most widely used frameworks for feature matching when a large number of feature points are present in images (Muja and Lowe, 2009). Instead of performing exhaustive linear matching, this study employs a randomized  $k$ -d tree to accelerate the feature matching process. This implementation evaluates the Euclidean distances between feature points detected in two images where a shorter Euclidean distance indicates greater similarity between two feature points. When using the randomized  $k$ -d tree algorithm, FLANN typically finds a match that is very close to the true nearest neighbor; however, the result is not guaranteed to be the exact nearest neighbor. To identify the most likely true nearest neighbor, this study sets the parameter  $k=2$  to the randomized  $k$ -d tree search to obtain two candidate matches with the smallest Euclidean distances ( $i$  and  $j$ ). Lowe’s ratio test (Lowe, 1999) is then applied to their Euclidean distances, with the ratio threshold set to 0.6 in this study. As described in Equation (2), two outcomes are possible: (1) one of two candidates represents the true nearest neighbor, or (2) neither candidate corresponds to the true nearest neighbor.

$$match = \begin{cases} feature_i, d_i < ratio \times d_j \\ None, otherwise \end{cases} \quad (2)$$

where  $feature_i$  = likely true nearest neighbor  
 $d$  = Euclidean distance  
 $ratio$  = Lowe’s ratio

LightGlue is derived from SuperGlue with several improvements in both efficiency and accuracy. It employs a stack of transformer layers that combine self-attention (within an image) and cross-attention (between images) mechanisms to enrich feature descriptors with contextual information (Lindberger et al., 2023). To identify common features between two images, a confidence classifier integrates descriptor similarity and matchability to dynamically and adaptively infer feature correspondences. For pairs of feature correspondences exhibiting significant overlap or limited appearance changes, the inference process terminates early; otherwise, it proceeds through additional layers. However, matching all feature points between two images can be computationally expensive. To address this limitation, LightGlue prunes unmatchable features to reduce computation costs. As a result, LightGlue achieves efficient feature matching while maintaining high matching accuracy.

StereoGlue builds upon LightGlue with several enhancements designed to improve the robustness of matching results as presented in Figure 3. It introduces a joint matching and model estimation strategy that iteratively evaluates and refines the model parameters. For an image pair, the model represents the geometric relationship between the two images, such as homography or relative camera pose. Starting from an initial model, a guided matching algorithm is employed to search for the next best match and incorporate it to the estimated model. Since guided matching operates in a pointwise manner, the AI-driven approach DepthAnything (Yang et al., 2024) is used as an intermediate module to assist and enhance its performance. Owing to the integration of additional AI modules aimed at improving feature-matching robustness, StereoGlue incurs a significantly higher computational cost than FLANN and LightGlue.

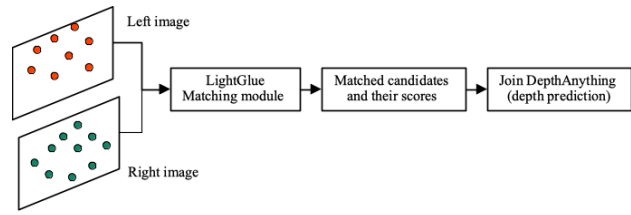


Figure 3. StereoGlue algorithm that integrates AI-powered modules for robust feature matching.

With two feature types and three matching algorithms, this study investigates five combinations, as presented in Table 1, to evaluate their performances in matching challenging imagery datasets for SfM image alignment. For image pairs selection, this study follows the graph-tracking strategy (Cheng et al., 2024) to obtain the most spatially correlated image pairs. Due to computational resource constraints, the maximum number of SIFT features detected per image is limited to 1,0000, while 2,048 features are extracted for SuperPoint. After feature descriptor matching, both inliers and outliers may be present. Since outliers are likely to degrade the camera pose estimation, an additional stage for their detection and removal is required. In this study, MAGSAC ++ (Barath et al., 2020) is applied to address this task and to preserve reliable inliers for SfM image alignment.

| Feature type | Matching algorithm | AI-driven |
|--------------|--------------------|-----------|
| SIFT         | FLANN              | N         |
| SIFT         | LightGlue          | Y         |
| SIFT         | StereoGlue         | Y         |
| SuperPoint   | LightGlue          | Y         |
| SuperPoint   | StereoGlue         | Y         |

Table 1. Feature extraction and matching method combinations

### 2.4 Evaluation between SfM camera position and GNSS meta-information

SfM estimates camera positions in a relative coordinate system; therefore, a comparison with GNSS meta-information provides a means of evaluating the reliability of the SfM reconstruction. Since camera positions are presented in relative and absolute coordinate systems, a 3D conformal transformation, expressed in Equation (3), is applied to estimate the seven transformation parameters. The camera positions in the relative coordinate system are then transformed into the absolute coordinate system for evaluation. To identify potentially misaligned images, accuracy standards derived from GNSS surveying are adopted. According to (Isik et al., 2020), positional differences of 2–5 m indicate an acceptable level of accuracy. In this study, a threshold of 2.5 m is used to detect misalignments.

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} = s \cdot R \begin{bmatrix} x \\ y \\ z \end{bmatrix} + t \quad (3)$$

where  $X, Y, Z$  = GNSS-surveyed camera position  
 $x, y, z$  = SfM-estimated camera position  
 $s$  = scale factor  
 $R$  = rotation matrix  
 $t$  = translation vector

### 3. Experimental Results and Analyses

This paper exploits two UAV imagery datasets capturing ground scenes over large areas of repetitive textures to examine the performances of the combinations of the aforementioned approaches. This first dataset consists of 211 images covering a field equipped with solar panels, while the second dataset comprises 424 images depicting a wide area of vegetation and low brushes. This study employs three evaluation indicators: (1) the number of robustly matched image pairs with a threshold of 20 matches gained after outlier removal, (2) the number of correctly aligned images, and (3) the positional errors (POS error) of the SfM reconstruction relative to GNSS meta-information, i.e., root mean square error (RMSE). These indicators are used to analyze the impacts of integrating different feature types and matching algorithms for SfM image alignment.

#### 3.1 Solar Panel Field Reconstruction

With the growing demand for green energy, large-scale solar panel installations have become widely adopted. Managing these facilities using UAV imagery combined with 3D reconstruction techniques offers an effective digital solution. However, UAV flight planning and data acquisition conditions are crucial, as they strongly influence image quality. For instance, sunlight reflections from solar panels alter illumination across different viewing angles, while large numbers of images depicting nearly identical scenes, such as areas dominated by solar panels, lead to high similar feature descriptors, increasing the difficulty of reliable image matching.

In this experiment, 1195 image pairs are generated for matching. The numbers of robustly matched image pairs are reported in Table 2. The results show that SIFT-FLANN produces the highest number of robustly matched pairs. Among the AI-powered methods, LightGlue yields more matched pairs than StereoGlue, regardless of the feature types employed. Further analysis indicates that LightGlue produces varying numbers of matched pairs depending on the number of detected feature points. In contrast, the number of matched pairs obtained by StereoGlue remains nearly identical when different maximum numbers are specified for SIFT and SuperPoint feature detection. Figure 4 illustrates a critical image pair dominated by repeated patterns in this study, for which only 42 robust matches are identified using SIFT-StereoGlue pipeline.

| Method                | Matched pairs | Ratio (%) |
|-----------------------|---------------|-----------|
| SIFT-FLANN            | 1160          | 97.9      |
| SIFT-LightGlue        | 940           | 78.6      |
| SIFT-StereoGlue       | 609           | 50.9      |
| SuperPoint-LightGlue  | 700           | 58.6      |
| SuperPoint-StereoGlue | 608           | 50.9      |

Table 2. Solar panel dataset: number of robustly matched image pairs using the five designed approaches

Figure 5 presents the SfM image alignment results by using the proposed combinations of feature types and matching algorithms. Although all approaches are capable of performing SfM image alignment, it is evident that SIFT combined with the three matching algorithms produces more complete results than SuperPoint-based methods. Table 3 reports the quantitative results of the SfM reconstruction. Among all combinations, the SIFT-StereoGlue approach achieves the highest reconstruction completeness, followed by SIFT-FLANN and SIFT-LightGlue. However, as shown in Table 2, the number of robustly matched image pairs produced by SIFT-StereoGlue is lower than that



Figure 4. An image pair dominated by repeated patterns for matching. By using SIFT-StereoGlue method, only 42 robust matches were found.

of the other two methods. These results indicate that while the number of matched pairs is an important factor influencing SfM image alignment, a larger number does not necessarily guarantee higher reconstruction completeness.

SuperPoint-based methods exhibit the smallest positional error among all pipelines. One likely reason is that improper observations, i.e., mismatches, can be more easily detected and removed during bundle adjustment. In addition, error propagation is reduced when a relatively small number of images is aligned. In contrast, SIFT-based methods align a larger number of images, which increases the difficulty of identifying improper observations during bundle adjustment. As a result, larger positional errors are observed when using this feature type. In this case study, the SIFT-FLANN method yields the largest positional error in SfM reconstruction compared with the other methods.

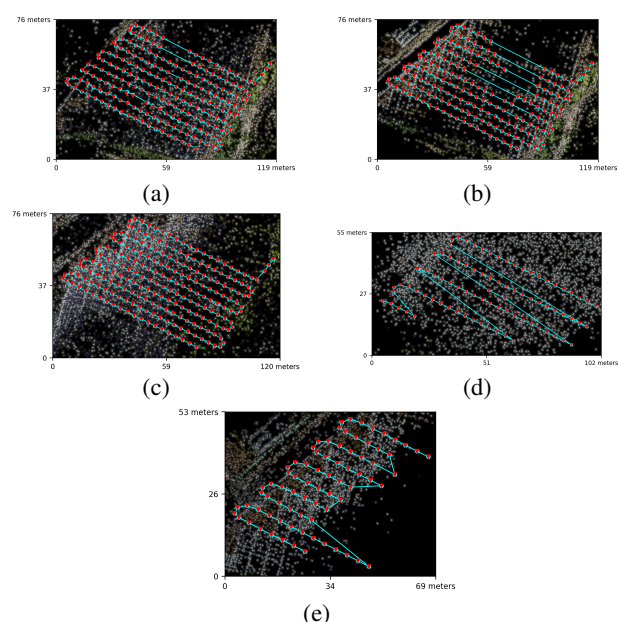


Figure 5. SfM image alignment trajectory of solar panel imagery dataset using (a) SIFT-FLANN (b) SIFT-LightGlue (c) SIFT-StereoGlue (d) SuperPoint-LightGlue (e) SuperPoint-StereoGlue algorithms. Red triangles: GNSS meta-information; Blue dots: SfM estimations

This study further examines the overall reprojection error of the SfM reconstruction to assess the impacts of SIFT and SuperPoint. Corresponding to the method configurations listed in Table 3, the reprojection errors are 1.44, 1.04, 1.42, 1.96, and 1.81 pixels, respectively. None of the results meets the com-



| Method                | Images aligned | POS error (m) |
|-----------------------|----------------|---------------|
| SIFT-FLANN            | 211/221        | 0.282         |
| SIFT-LightGlue        | 185/221        | 0.122         |
| SIFT-StereoGlue       | 221/221        | 0.096         |
| SuperPoint-LightGlue  | 84/221         | 0.018         |
| SuperPoint-StereoGlue | 77/221         | 0.004         |

Table 3. Solar panel dataset: number of images aligned and positional error

monly accepted SfM reconstruction criterion that the reprojection error should be less than one pixel. Nevertheless, SIFT-based methods consistently exhibit smaller reprojection errors than SuperPoint-based methods. Although SuperPoint is designed as a pixel-level feature detector, achieving the conventional SfM reprojection error threshold remains challenging in this case.

### 3.2 Vegetation and Low Brushes Area Reconstruction

Smart agriculture using UAV images and 3D reconstruction techniques has emerged as an effective means for seasonal monitoring and management. However, reconstructing natural landscapes remains challenging due to repetitive patterns, uniform textures, and varying illumination conditions caused by different viewing angles. As a result, feature descriptors of identical ground objects may vary across images. To evaluate these challenges, this experiment uses an imagery dataset dominated by vegetations and low brushes to assess the designed five combinations of feature detection and matching methods for SfM reconstruction.

In this study case, 2519 image pairs are generated for matching. Table 4 reports the number of robustly matched image pairs. The overall matching success rate is low, indicating that the image matching for this dataset is more challenging. Unlike the previous study case, SuperPoint combined with LightGlue and StereoGlue yields similar or higher numbers of robustly matched pairs than SIFT-based methods. In particular, the SuperPoint-StereoGlue combination outperforms all other approaches in this experiment. For SIFT-based methods, a trend similar to that observed in the previous study case is evident: FLANN produces the highest number of robustly matched pairs, followed by LightGlue and StereoGlue.

| Method                | Matched pairs | Ratio (%) |
|-----------------------|---------------|-----------|
| SIFT-FLANN            | 542           | 21.5      |
| SIFT-LightGlue        | 477           | 18.9      |
| SIFT-StereoGlue       | 275           | 10.9      |
| SuperPoint-LightGlue  | 524           | 20.8      |
| SuperPoint-StereoGlue | 782           | 31.0      |

Table 4. Vegetation and low bushes dataset: number of robustly matched image pairs using the five designed approaches

Figure 6 illustrates three critical example image pairs matched using SIFT-StereoGlue and SuperPoint-StereoGlue, yielding 42 and 43 robust matches, respectively. A key factor is that the descriptors of highly similar features are often too alike to be reliably distinguished. Consequently, the likelihood of mismatches or insufficient matches increases in this imagery dataset. Although outlier detection is applied, mismatches still cannot be eliminated in these image pairs.

Given these challenging feature-matching results, Figure 7 visualizes the resulting SfM image alignment trajectories. It can be observed that an insufficient number of matches significantly affects the number of images that can be aligned, leading to

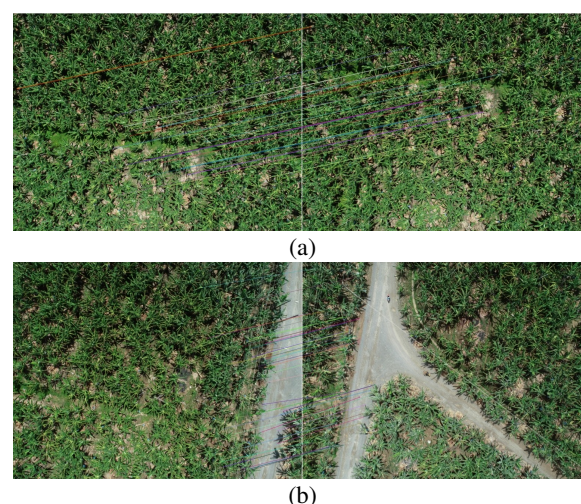


Figure 6. Feature matching results using (a) SIFT-StereoGlue (b) SuperPoint-StereoGlue for image scenes of vegetation and low bushes

unstable alignment quality for this dataset. Table 5 presents the quantitative results of the SfM reconstruction using the proposed approaches. Although SuperPoint-StereoGlue achieves the highest number of matched image pairs, these matches do not translate into improved SfM image alignment performance, as this method aligns the fewest images among all approaches.

In addition, the flight trajectory computed from matches produced by SIFT-StereoGlue results in several misaligned images, leading to larger positional error than those of the other methods. As shown in Table 5, the flight paths estimated using matches derived from SIFT-FLANN and SuperPoint-LightGlue are comparable in terms of both the number of aligned images and the positional error. Overall, SIFT-LighGlue appears to provide the best SfM image alignment performance, as it aligns the largest number of images while maintaining a moderate positional error compared with the other approaches.

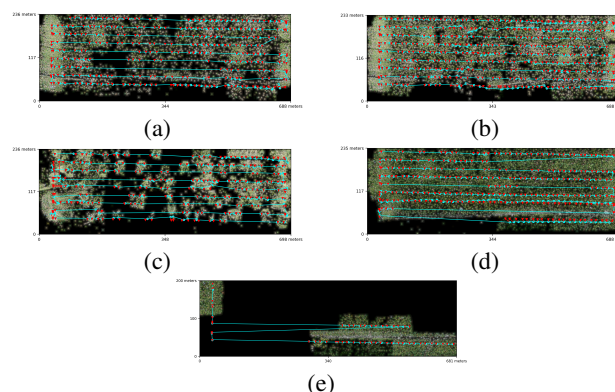


Figure 7. SfM image alignment trajectory of vegetation and low brushes imagery dataset using (a) SIFT-FLANN (b) SIFT-LightGlue (c) SIFT-StereoGlue (d) SuperPoint-LightGlue (e) SuperPoint-StereoGlue algorithms. Red triangles: GNSS meta-information; Blue dots: SfM estimations

Similarly, this paper examines the overall reprojection error of the SfM reconstruction to analyze the impacts of SIFT and SuperPoint. Corresponding to the method configurations listed in Table 5, the reprojection errors are 0.42, 0.43, 2.62, 1.53,

| Method                | Images aligned | POS error (m) |
|-----------------------|----------------|---------------|
| SIFT-FLANN            | 287/424        | 2.610         |
| SIFT-LightGlue        | 380/424        | 3.400         |
| SIFT-StereoGlue       | 231/424        | 34.150        |
| SuperPoint-LightGlue  | 278/424        | 3.090         |
| SuperPoint-StereoGlue | 53/424         | 1.570         |

Table 5. Vegetation and low brushes dataset: number of images aligned and positional error

and 1.62 pixels, respectively. Compared to the previous study case, the reprojection errors are smaller for this imagery dataset. These results indicate that SuperPoint-based methods generally fail to meet sub-pixel accuracy requirement for SfM image alignment. In contrast, two approaches - SIFT-FLANN and SIFT-LightGlue - demonstrate that SIFT-based methods are capable of satisfying such a requirement.

#### 4. Conclusions

Extracting geospatial information from optical imagery for 3D reconstruction is fundamental to a wide range of spatial applications. Owing to the flexibility of UAV-based image acquisition, UAV imagery has become an effective data source for geospatial computation. SfM, combined with photogrammetry techniques for UAV image processing, provides a fully automatic workflow to meet these demands. Feature extraction and matching play a critical role in automating the pipeline and determining the reliability of the results. With advances in computer vision and AI, this study aims to evaluate their performances in feature extraction and matching, and consequently in SfM image alignment. In addition to employing hand-crafted and AI-powered methods independently, this research also integrates their respective advantages to examine their effectiveness in SfM image alignment on two critical imagery datasets.

This paper summarizes the following findings based on the experimental results. First, a large number of robustly matched image pairs does not necessarily guarantee satisfactory SfM image alignment. This observation also suggests that matching most spatially correlated image pairs play a crucial role in achieving correct image alignment. Second, although hand-crafted and AI-driven methods each have their own advantages and limitations, selecting an appropriate pipeline for feature extraction and matching remains critical when dealing with challenging imagery datasets. Based on the two study cases, integrating a hand-crafted feature type (SIFT) and AI-powered matching algorithm (LightGlue or StereoGlue) appears to be a potential approach for overcoming these challenges. The feature-matching results produced by these hybrid pipelines lead to more successful SfM image alignment for both two imagery datasets.

For future work, it is necessary to further investigate the impacts of AI-based approaches, including the reliability of the pre-trained data or model. For particularly challenging image scenes, analyzing the spatial distribution of matched features within each image may help enforce geometry constraints during space resection and intersection, thereby yielding more reliable camera pose estimations.

#### References

Barath, D., Mishkin, D., Cavalli, L., Sarlin, P.-E., Hruby, P., Pollefeys, M., 2024. StereoGlue: Robust estimation with

single-point solvers. *Proceedings of the European Conference on Computer Vision*, 15115, 421–441.

Barath, D., Nuskova, J., Ivashechkin, M., Matas, J., 2020. Mag-sac++, a fast, reliable and accurate robust estimator. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1304–1312.

Berra, E., Peppia, M., 2020. Advances and challenges of uav sfm mvs photogrammetry and remote sensing: Short review. *2020 IEEE Latin American GRSS & ISPRS Remote Sensing Conference (LAGIRS)*, 533–538.

Chen, Y., Liu, X., Zhu, B., Zhu, D., Zuo, X., Li, Q., 2025. UAV Image-Based 3D Reconstruction Technology in Landslide Disasters: A Review. *Remote Sensing*, 17(17), 3117.

Cheng, M.-L., Fujita, Y., Kuramoto, Y., H., M., Matsuoka, M., 2024. Associating UAS images through a graph-based guiding strategy for boosting structure from motion. *Photogrammetry Record*, 39(185), 157–182.

Cheng, M.-L., Matsuoka, M., 2021. An Efficient and Precise Remote Sensing Optical Image Matching Technique Using Binary-Based Feature Points. *Sensors*, 21(6035).

Isik, O., Hong, J., Petrunin, I., Tsourdos, A., 2020. Integrity Analysis for GPS-Based Navigation of UAVs in Urban Environment. *Robotics*, 9(66).

Li, H., Lin, B., Zhu, H., 2025. UAV Photogrammetry with Low Image Overlap for Panoramic Imaging and River Morphology Monitoring. *Journal of Geovisualization and Spatial Analysis*, 9(34).

Lin, H.-Y., Chen, L., 2025. GNSS-denied UAV localization with satellite and aerial image matching. *Results in Engineering*, 28(108132).

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. Lightglue: Local feature matching at light speed. *2023 IEEE/CVF International Conference on Computer Vision*, 17581–17592.

Lowe, D., 1999. Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 2, 1150–1157.

Muja, M., Lowe, D., 2009. Fast approximate nearest neighbors with automatic algorithm configuration. *Proceedings of the Fourth International Conference on Computer Vision Theory and Applications*, 1, 331–340.

Qiao, R., Xu, G., Wang, P., Cheng, Y., Dong, W., 2023. An Accurate, Efficient, and Stable Perspective-n-Point Algorithm in 3D Space. *Applied Sciences*, 13(2), 1111.

Tone, D., Malisiewicz, T., Rabinovich, A., 2018. Superpoint: Self-supervised interest point detection and description. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 337–33712.

Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H., 2024. Depth anything v2. *NIPS '24: Proceedings of the 38th International Conference on Neural Information Processing Systems*, 21875–21911.