A Comparison of SuperPoint + SuperGlue and SIFT-Based Pipelines for Structure-from-Motion

Shoya Morizaki[1], Masayuki Matsuoka[1*]

[1]*Mie University, 1577 Kurimamachiya, Tsu, Mie 514-0102, Japan. matsuoka@info.mie-u.ac.jp*

Abstract: We evaluated two SfM configurations using COLMAP: a default baseline pipeline with SIFT features and a fully learned pipeline that couples SuperPoint with SuperGlue (SP+SG) for context-aware matching. Using 100 images, we evaluated total keypoints, candidate matches, inlier rate after geometric verification, reconstructed 3D points, track length, reprojection error, and reconstructed 3D shape. Despite using substantially fewer keypoints (~0.59M vs. ~1.3M), SP+SG yielded more total matches and inliers, enhancing image-pair connectivity under limited overlap. In contrast, SIFT achieved a higher inlier rate (proportion of verified matches), longer mean track length, and a lower reprojection error (~0.82 pixel vs. ~1.27 pixel), indicating tighter geometric consistency. These results reveal a practical trade-off: SP+SG improves connectivity and reconstruction chances in sparse-view settings, whereas SIFT provides stronger multi-view support per point.

Keywords: Structure-from-Motion, SuperPoint, SuperGlue, Sparse-view Reconstruction

1. Introduction

Structure-from-Motion (SfM) reconstructs camera poses and sparse 3D structure by detecting local features, matching them across views, estimating two-view geometry, and refining all variables with bundle adjustment; this is the standard pipeline popularized in COLMAP and related work (Schönberger 2016, COLMAP 2025). In sparse-view capture—where overlap is limited and baselines are short—classical pipelines can fragment due to insufficient verified correspondences.

Recent learned components address this bottleneck at the pairwise matching stage. SuperPoint jointly detects and describes interest points via a self-supervised network, yielding repeatable keypoints and descriptors robust to moderate viewpoint and appearance changes (DeTone et al. 2018). SuperGlue then performs context-aware matching with attention and graph neural networks, formulating correspondence assignment with a differentiable optimal transport layer; this substantially improves match quality on challenging pairs (Sarlin et al. 2020). The HLoc toolbox operationalizes these modules within a modular pipeline for localization/SfM, while PyCOLMAP provides Python bindings to COLMAP's reconstruction routines, enabling end-to-end processing from feature matching to mapping/BA within Python (HLoc 2025, PyCOLMAP 2025).

We quantified the end-to-end impact of a learned pipeline—SuperPoint+SuperGlue via PyCOLMAP—against a SIFT baseline (Lowe 2004) on an outdoor dataset from the Image Matching Challenge 2021 (2021). Our results made explicit a practical trade-off between connectivity and geometric fidelity.

2. Methodology

2.1 Pipelines

- Baseline (SIFT). We followed the standard COLMAP workflow: SIFT detection/description, nearest-neighbor matching with geometric verification, incremental mapping, and bundle adjustmen.
- Learned (SP+SG via PyCOLMAP). Using HLoc, we extracted features with SuperPoint and match with SuperGlue. From the matching stage onward, we called PyCOLMAP to execute COLMAP's two-view estimation, triangulation, incremental mapping, and bundle adjustment as Python bindings of COLMAP's algorithms. In other words, the reconstruction steps are algorithmically equivalent to the COLMAP CLI pipeline; any differences stem from configuration rather than different solvers.

2.2 Dataset and settings

We used 100 images of a mid-scale outdoor object from Image Matching Challange 2021. All images were used at their native resolutions (no upscaling). We imposed no explicit cap on detected keypoints in either pipeline.

2.3 Evaluation metrics

We measured: Features per image denotes the number of detected local features. Matches are putative correspondences; inliers are matches accepted by geometric verification. The inlier rate is the fraction of matches that become inliers (inliers/matches). 3D points result from triangulation. Observations are the total count of 2D measurements attached to those 3D points—i.e., how many images actually see each reconstructed point after triangulation and resectioning. In other words, every time a triangulated point is associated with a keypoint in an image, that counts as one observation. The mean track length is therefore observations per 3D point and indicates multi-view support (longer tracks = the same point is seen in more views). The mean reprojection error is computed after bundle adjustment in pixels. The reconstructed 3D shapes were laid out side by side for visual comparison.

3. Results/Findings

As summarized in Table 1, the learned pipeline detects fewer keypoints than the SIFT baseline, yet produces more total matches and more total inliers. In contrast, SIFT achieved a higher inlier rate, both in mean and median. This indicated that SP+SG broadens pair connectivity under limited overlap, while SIFT yields purer correspondences once a pair is connected.

Table 2, shows that the number of 3D points was comparable between methods. However, observations were higher with SIFT, resulting in a longer mean track length. Consistently, mean

reprojection error was lower for SIFT, reflecting tighter multi-view geometry after bundle adjustment.

Taken together, SP+SG improved coverage/connectivity (more matches and inliers from a smaller keypoint pool), which is advantageous under sparse overlap; SIFT provided higher inlier purity, translating into longer tracks and lower reprojection error, i.e., greater geometric fidelity. This connectivity–fidelity trade-off aligns with the design of SuperPoint/SuperGlue for robust pairwise matching and with the classical SIFT pipeline's strengths in precise geometry once correspondences are verified.

From Figure 1, the reconstructed point cloud using SP+SG exhibits low local density but is continuously distributed across the entire target area, faithfully representing the outline without gaps. In contrast, SIFT concentrates correspondences on high-contrast facial regions (around the nose and mouth), forming dense point clusters, while showing missing points in peripheral areas.

Table 1, Pair-Level Matching Statistics

Method	Keypoints(total)	Matches(total)	Inliers(total)	Mean inlier rate	Median inlier rate
SIFT	1,336,994	601,629	578,073	0.773	0.911
SP+SG	588,061	745,123	661,972	0.755	0.838

Table 2, SfM Reconstruction Quality

Method	3D Points(total)	Observations(total)	Mean track length	Mean reprojection error
SIFT	57,741	338,840	5.868	0.82 pixel
SP+SG	58,924	243,638	4.135	1.27 pixel

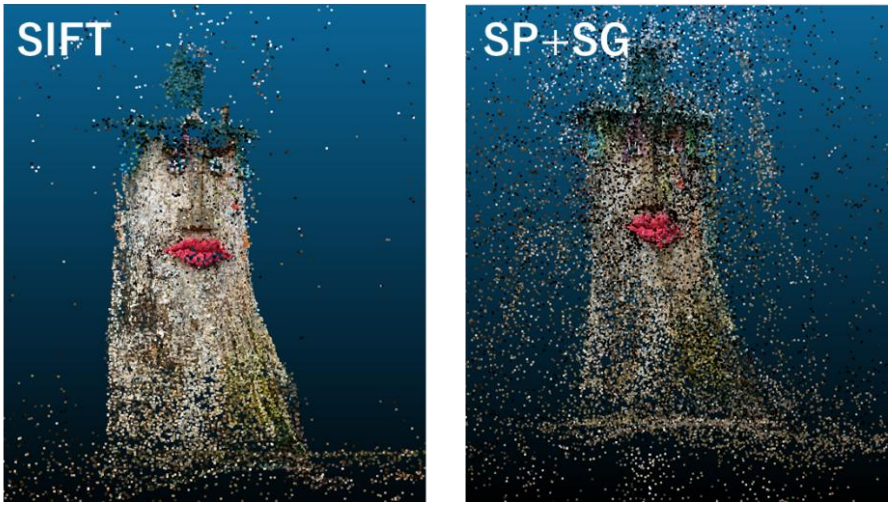

Figure 1: Reconstructed 3D models

This comparison quantitatively demonstrated a clear trade-off under sparse-view conditions: SuperPoint+SuperGlue (SP+SG) excels at ensuring connectivity by increasing the total number of matches and inliers, whereas SIFT holds advantages in geometric consistency and multi-view support. In particular, SP+SG strengthens image-pair connections even from a smaller keypoint pool, expanding the set of verified pairs and reducing the risk of early reconstruction failure. By

contrast, SIFT facilitates tighter reconstructions, supported by higher inlier rates, lower mean reprojection errors, and longer mean track lengths, which together indicate stronger per-point multi-view evidence.

4. Conclusion

We evaluated two SfM configurations (SIFT and SP+SG) using COLMAP. Practically, when overlap is scarce and robustness is the priority, SP+SG is preferable; when accuracy and geometric fidelity dominate, SIFT is the better choice. Looking ahead, we plan to probe the failure boundary by further reducing the number of inputs images, and validate generality on self-captured datasets across scene types and capture protocols.

Acknowledgment

This work was supported by JSPS KAKENHI Grant Numbers JP21K05669, JP22H05004, and JP20K20487.

References

COLMAP, 2025. COLMAP: Official website and documentation. Available at: colmap.github.io. Accessed: Sept. 22, 2025.

DeTone, D., T. Malisiewicz and A. Rabinovich, 2018. SuperPoint: Self-Supervised Interest Point Detection and Description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Salt Lake City, USA, June 18–22, 2018.

Hierarchical-Localization (HLoc), 2025. Official GitHub repository. Available at: github.com/cvg/Hierarchical-Localization. Accessed: Sept. 22, 2025.

Image Matching Challenge 2021, 2021. Dataset website. Available at: cs.ubc.ca/research/image-matching-challenge/2021/. Accessed: Sept. 22, 2025.

Lowe, D. G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), pp. 91–110.

PyCOLMAP, 2025. Official documentation (Python bindings for COLMAP). Available at: colmap.github.io/pycolmap/. Accessed: Sept. 22, 2025.

Sarlin, P.-E., D. DeTone, T. Malisiewicz and A. Rabinovich, 2020. SuperGlue: Learning Feature Matching with Graph Neural Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Virtual (Seattle, USA), June 14–20, 2020.

Schönberger, J. L., and J.-M. Frahm, 2016. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, USA, June 26–July 1, 2016.