Evaluating Classical and Deep Keypoint Detectors for SfM Reconstruction in Arctic UAV Imagery

Nicholas Sansoterra[1]*, Maria G. Lenzano[2], William J. Shuart[3], John E. Anderson[1], Alper Yilmaz[1], Charles Toth[1]*

[1] SPIN Laboratory, Department of Civil, Environmental and Geodetic Engineering, The Ohio State University, Columbus, OH, USA – (sansoterra.2, yilmaz.15, toth.2)@osu.edu
[2] Resp. Lab. Geomatica Andino (LAGEAN), Mendoza, Argentina – mlenzano@mendoza-conicet.gob.ar
[3] USACE ERDC GRL Corbin Field Station, USA – (William.J.Stewart, John.Anderson)@usace.army.mil

KEYWORDS: Structure-from-Motion, keypoint detection, SIFT, SuperPoint, DISK, Arctic UAV imagery

ABSTRACT:

Structure-from-Motion (SfM) pipelines rely heavily on the detection and matching of repeatable keypoints across images, yet the performance of modern learned feature extractors in challenging environments remains insufficiently understood. This paper evaluates classical and deep keypoint detectors for SfM reconstruction using winter Arctic UAV imagery, a domain characterized by low texture, repetitive patterns, and limited man-made structure. We compare three feature pipelines within a shared PyCOLMAP-based framework: SIFT with nearest-neighbor matching (SIFT+NN), SuperPoint, and DISK, along with a hybrid approach combining SuperPoint and DISK correspondences.

Quantitative evaluation is conducted using standard SfM metrics, including number of observations, track length, observations per image, and reprojection error, complemented by qualitative analysis of keypoint distributions and reconstruction interpretability. Results show that SIFT+NN consistently achieves the most complete and stable reconstructions, producing the highest number of matched observations and lowest reprojection error across aggregate experiments. However, on more challenging subsets lacking clear structural features, learned methods demonstrate improved robustness, successfully reconstructing multiple views where SIFT fails. SuperPoint provides broader spatial coverage, while DISK produces denser clusters in high-confidence regions, highlighting complementary behaviors between learned approaches.

Overall, the findings indicate that classical methods remain strong baselines for Arctic UAV photogrammetry under standard SfM pipelines, while learned detectors offer advantages in difficult conditions. The observed performance gap is attributed to domain mismatch and backend optimization for handcrafted features. These results suggest that domain-specific training and improved spatial feature distribution are promising directions for advancing learned keypoint methods in Arctic reconstruction tasks.

1. INTRODUCTION

The Arctic regions have become an increasingly important region of interest in recent years. Monitoring the terrain of these regions brings many unique challenges. One important tool is the use of drones to construct 3D maps of the terrain using Structure-from-Motion (SfM) pipelines. SfM relies heavily on the detection and matching of repeatable keypoints across images. Classical handcrafted features such as SIFT remain widely used for this task. Additionally, modern deep learning-based pipelines have shown impressive results on urban and indoor benchmarks. However, their behavior on challenging domains such as snow-covered Arctic environments is not well documented. This work compares classical and deep SfM methods of 3D reconstruction using winter Arctic UAV imagery and analyzes their strengths, weaknesses, and failures.

Our data consists of multi-view aerial imagery collected over several Alaskan sites in winter conditions, including roads, snow-covered terrain, trees, fences, rooftops, and vehicles, provided by USARMY CEERD-GRL. We evaluate three keypoint pipelines within the same PyCOLMAP-based SfM framework: (i) SIFT with nearest-neighbor matching (SIFT+NN) as a baseline; (ii) SuperPoint, a fully convolutional detector-descriptor trained in two stages on synthetic shapes and real natural images; and (iii) DISK, a U-Net-like architecture trained end-to-end with policy-gradient reinforcement learning, where rewards reflect geometric consistency of matches. We additionally test a simple combined method that merges SuperPoint and DISK matches before triangulation. All methods use identical geometric verification and bundle adjustment settings, allowing a fair comparison of the keypoint front-end.

We report standard SfM metrics, including the number of registered images, total matched observations, mean track length, and mean reprojection error, alongside qualitative assessments of keypoint spatial distribution and point cloud interpretability. On a 151-image sequence, SIFT+NN reconstructs all images and produces the largest number of matched 3D points, with mean reprojection error on the order of 1.3 pixels. SuperPoint and DISK achieve comparable reprojection errors but register only a subset of the images and yield fewer total points; DISK in particular tends to cluster keypoints in high-confidence regions, leaving other areas sparse. Conversely, on a more difficult 56-image subset lacking clear man-made structures, SIFT+NN collapses to only two registered images and very few matches, whereas SuperPoint and DISK still reconstruct several views and hundreds of points. Across all methods, dense tree canopy and homogeneous snow remain consistent failure cases: keypoints are detected but cannot be matched reliably enough to support stable 3D reconstruction.

Our results indicate that SIFT+NN remains a strong and robust baseline for Arctic UAV imagery under a standard COLMAP-style pipeline, while deep keypoint methods provide complementary behavior on the most challenging sequences. We attribute the performance gap mainly to domain mismatch in the training data of SuperPoint and DISK and to the SfM backend

being tuned for SIFT-like features. The study suggests that domain-specific pretraining on aerial winter imagery and modified objectives encouraging more spatially diverse keypoints, especially for DISK, are promising directions to improve deep keypoint detectors for Arctic photogrammetric applications.

2. RELATED WORK

2.1 Classical and Learned Local Features

SIFT (Lowe, 2004) is one of the most established local feature methods in computer vision and photogrammetry. It identifies extrema in a scale-space representation and computes descriptors from local gradient histograms. Its long-standing use in SfM pipelines is due not only to its invariance properties, but also to the fact that many reconstruction systems and matching heuristics were developed around SIFT-like descriptors and point distributions.

Learned local features attempt to replace handcrafted detectors and descriptors with models trained from image data. SuperPoint (DeTone et al., 2018) jointly predicts keypoint locations and descriptors in a fully convolutional framework. It is trained in two stages, beginning with synthetic shapes and then adapting to real imagery through homographic adaptation. In practice, SuperPoint often produces relatively uniform spatial coverage, which can be advantageous when broad image support is needed for downstream geometry estimation.

DISK (Tyszkiewicz et al., 2020) formulates local-feature learning as a reinforcement learning problem. The network predicts keypoints and descriptors, while training rewards are based on the geometric validity of resulting matches. This objective makes DISK particularly relevant to correspondence-driven tasks. However, the training distribution and optimization target can also bias the detector toward concentrated high-confidence regions rather than globally uniform coverage.

2.2 Learned Matching and Challenging Reconstruction Domains

Beyond feature extraction, recent work has also improved feature matching itself. SuperGlue (Sarlin et al., 2020), for example, incorporates graph neural networks and attention-based context aggregation to produce more discriminative correspondence assignments. Such methods demonstrate that learned matching can improve performance even when the underlying detector is fixed. Nevertheless, the present study isolates the feature front-end as much as possible by keeping the geometric backend consistent across experiments.

Within photogrammetry and remote sensing, winter and polar imagery remain difficult operating conditions for feature-based reconstruction. Low-texture snow, repetitive vegetation, and limited man-made structure can all reduce matching reliability. These challenges make Arctic UAV imagery a useful testbed for examining whether learned local features truly generalize beyond the benchmarks on which they are commonly reported.

3. DATASET AND METHODOLOGY

3.1 Dataset

The dataset used in this study was provided by the U.S. Army Corps of Engineers Engineer Research and Development Center and contains multi-view aerial imagery captured over several Alaskan sites during winter conditions. The scenes include roads, snow-covered terrain, fences, trees, rooftops, and vehicles. Image collections were acquired using multiple UAV platforms and flight configurations, producing both relatively structured scenes with roads or buildings and more difficult subsets dominated by snow and vegetation. Representative examples of the winter Arctic UAV imagery are shown in Figure 1.

Figure 1. Typical UAV imagery from Arctic environment

The visual characteristics of the data make it challenging for local feature pipelines. Snow-covered regions often contain large areas with weak gradients and limited texture. Forested areas may produce visually distinct points in a single image but still lead to ambiguous matches because many local patches appear similar across views. These properties make the dataset well suited for evaluating not only numerical reconstruction quality, but also the relationship between keypoint placement and reconstruction interpretability.

3.2 Reconstruction Pipeline

All methods were evaluated within the same reconstruction framework. Feature extraction and matching were performed through the HLOC-based pipeline, while geometric verification, triangulation, and bundle adjustment were handled by PyCOLMAP, the Python interface to COLMAP (Schönberger and Frahm, 2016). This shared backend is important for a fair comparison: differences in the final models should primarily reflect the properties of the front-end features rather than different geometric solvers or bundle-adjustment settings.

Three primary pipelines were evaluated. The baseline was SIFT with nearest-neighbor matching (SIFT+NN). The learned alternatives were SuperPoint and DISK. In addition, a combined method was tested by merging the matched correspondences produced by SuperPoint and DISK before triangulation. The combined experiment was motivated by qualitative observations that SuperPoint tends to distribute keypoints broadly across the image, whereas DISK often concentrates them into dense, locally confident clusters.

3.3 Evaluated Methods

SIFT+NN serves as the classical reference point. In practice, it remains tightly coupled to mature SfM software and often benefits from geometric settings and heuristics that have been refined around SIFT-like correspondences. SuperPoint represents a modern learned detector-descriptor with relatively uniform image coverage, while DISK represents a geometry-aware learned alternative that directly optimizes feature usefulness through reinforcement learning.

The goal of the comparison is not to evaluate every possible learned matching module or backend variation, but to study how these feature extraction strategies behave under identical

reconstruction settings in a difficult, domain-shifted environment.

3.4 Evaluation Metrics

Four quantitative metrics were used throughout the experiments: (1) number of observations, defined as the total number of matched 2D keypoints used for 3D triangulation; (2) mean track length, defined as the average number of images in which a triangulated 3D point is observed; (3) mean observations per image, defined as the average number of matched 2D points contributed by each registered image; and (4) mean reprojection error in pixels, which measures the geometric consistency between the reconstructed 3D points and their image projections.

Quantitative results were complemented by qualitative inspection of keypoint spatial distribution and sparse point-cloud structure. This second component is important because two methods can produce similar reprojection errors while leading to reconstructions that differ substantially in completeness, point placement, and ease of interpretation.

Table 1. Average statistics across all experiments

	Number of Observations [points]	Mean Track Length	Mean Observations per Image [points]	Mean Reprojection Error [pixels]
SuperPoint	1778.75	2.32	301.95	1.32
DISK	4209.75	2.57	748.06	1.26
SIFT + NN	7,929.38	2.77	1,078.46	0.80

4. EXPERIMENTS AND RESULTS

4.1 Aggregate Quantitative Performance

Across the aggregate statistics summarized in Table 1, SIFT+NN achieved the strongest overall performance. The method produced the highest number of observations, the highest mean observations per image, the longest average track length, and the lowest mean reprojection error. The reported averages across all experiments were 7,929.38 observations, mean track length 2.77, mean observations per image 1,078.46, and mean reprojection error 0.80 pixels for SIFT+NN. The corresponding values for DISK were 4,209.75 observations, 2.57 mean track length, 748.06 mean observations per image, and 1.26 pixels reprojection error. SuperPoint produced 1,778.75 observations, 2.32 mean track length, 301.95 mean observations per image, and 1.32 pixels reprojection error.

These results indicate that, on average, the classical baseline provided both more complete geometric support and lower reconstruction error than the learned alternatives. The gap is especially clear in the observation counts, where SIFT+NN substantially exceeded the learned methods. Because observation count and track length directly affect the stability of downstream geometry estimation, this aggregate advantage is not merely cosmetic; it materially affects the completeness and reliability of the reconstructed model.

4.2 Representative Large-Scene Reconstruction

A representative large sequence containing 151 input images illustrates the best overall behavior of the three approaches. The corresponding quantitative comparison is summarized in Table 2, and the SIFT+NN sparse reconstruction for this sequence is shown in Figure 2. On this sequence, SIFT+NN registered all 151 images and reconstructed 95,817 matched points with mean reprojection error 1.27 pixels. DISK registered 97 images and reconstructed 63,472 matched points with mean reprojection error 1.34 pixels. SuperPoint also registered 97 images, but reconstructed only 51,320 matched points and had a mean reprojection error of 1.44 pixels.

This example shows that the strongest method is not simply the one with the densest local clusters or the most visually appealing isolated regions, but the one that provides sufficiently stable correspondences across a large portion of the image set. In scenes containing roads, buildings, and other man-made structures, SIFT+NN consistently produced the clearest and most complete reconstruction.

Table 2. Representative comparison on the 151-image sequence

	SIFT+NN	DISK	SuperPoint
Points Matched	95817	63472	51320
Points Detected	494716	239956	159981
Mean Reprojection Err [pixels]	1.27	1.34	1.44
Input Images	151	151	151
Matched Images	151	97	97

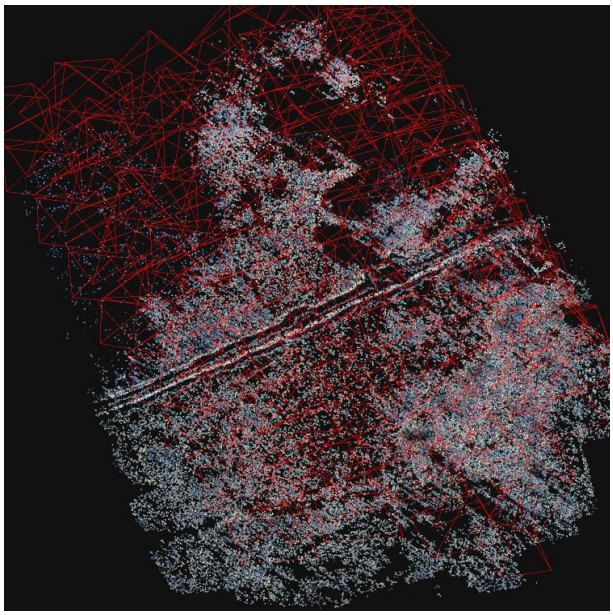

Figure 2. SIFT + Nearest Neighbor SFM Reconstruction of the 151-image sequence

4.3 Representative Difficult Subset

A second representative case demonstrates that the ranking is not uniform across all subsets. We have included an example of this difficult terrain in Figure 3. The corresponding quantitative comparison is summarized in Table 3. On a more difficult 56-image sequence lacking clear roads and buildings, SIFT+NN effectively collapsed: only 2 images were registered, with 98 matched points and mean reprojection error 0.94 pixels. In contrast, DISK registered 6 images with 811 matched points and mean reprojection error 1.31 pixels, while SuperPoint registered 8 images with 544 matched points and mean reprojection error 1.33 pixels.

Although the reprojection error of SIFT+NN remained numerically low on this difficult subset, the reconstruction itself was practically unusable because too few images entered the model. This illustrates why reprojection error must be interpreted together with image registration count and observation statistics. A method that reconstructs almost nothing can still appear geometrically clean on the few points that survive.

Table 3. Representative comparison on the difficult 56-image subset

	SIFT+NN	DISK	SuperPoint
Points Matched	98	811	544
Points Detected	196	1946	1260
Mean Reprojection Err [pixels]	0.94	1.31	1.33
Input Images	56	56	56
Matched Images	2	6	8

Figure 3. Difficult Subset Terrain Example

4.4 Qualitative Analysis

Qualitative inspection, illustrated in Figure 4, reveals consistent differences in how the methods allocate keypoints. SuperPoint generally distributes features more evenly across the image, resembling the broad spatial coverage of SIFT. DISK, by contrast, often forms dense clusters in regions it considers highly reliable, while leaving other regions comparatively sparse. This behavior can make certain point clouds easier to interpret locally because coherent object regions emerge as compact point clusters. However, the same clustering can reduce global coverage and limit the number of stable correspondences needed for full-scene reconstruction.

The combined SuperPoint+DISK experiment supports the idea that the two learned methods provide complementary behavior. In representative examples, the merged correspondences produced reconstructions that were visually easier to interpret than those of either learned method alone, because they combined broader coverage with denser local clusters. Even so, the combined approach still fell short of SIFT+NN in the clearest large-scene reconstructions, especially when measured by total registered images and overall structural completeness.

Across all methods, dense tree canopy and homogeneous snow remained persistent failure cases. In many such scenes, detectors could still place keypoints, but the resulting descriptors were not distinctive enough to support reliable matching across views. The failure therefore was not only in detection, but in the inability to establish consistent multi-view correspondences that survive geometric verification.

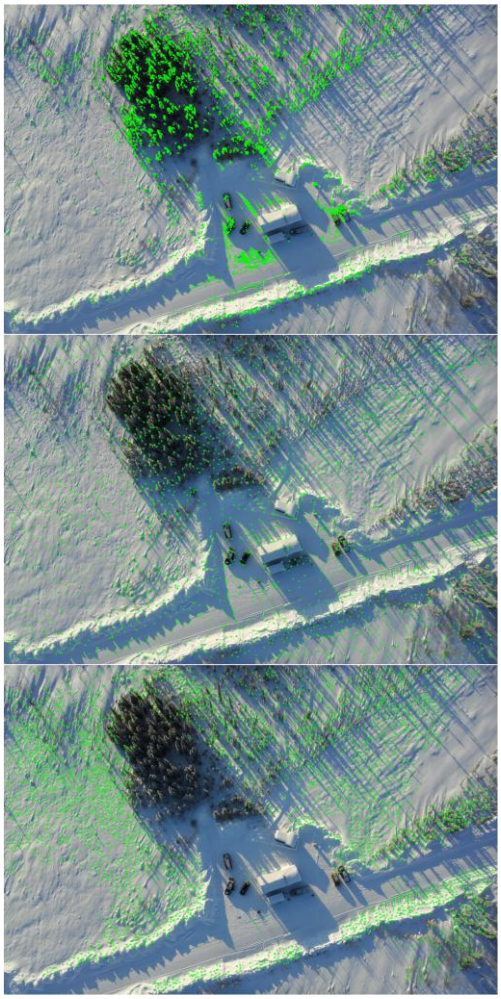

Figure 4. Example keypoint distributions for SIFT (top), SuperPoint (middle), and DISK (bottom)

4.5 Combined Method

The combined SuperPoint+DISK reconstruction was tested to determine whether the broad coverage of SuperPoint and the dense local clustering of DISK could be exploited simultaneously. An example combined reconstruction is shown in Figure 5. In representative examples, the combined method produced visually improved sparse structure and, in the presentation materials, yielded mean reprojection errors of approximately 1.26, 1.34, 1.31, and 1.31 pixels across four showcased scenes. These values generally fell between the

individual learned methods, indicating that the merged correspondences did not destabilize geometry while still improving qualitative scene interpretability.

Even so, the combined method did not surpass SIFT+NN in the most complete large-scene reconstructions. The hybrid result is therefore best interpreted as evidence of complementarity between the two learned methods rather than as a replacement for the classical baseline.

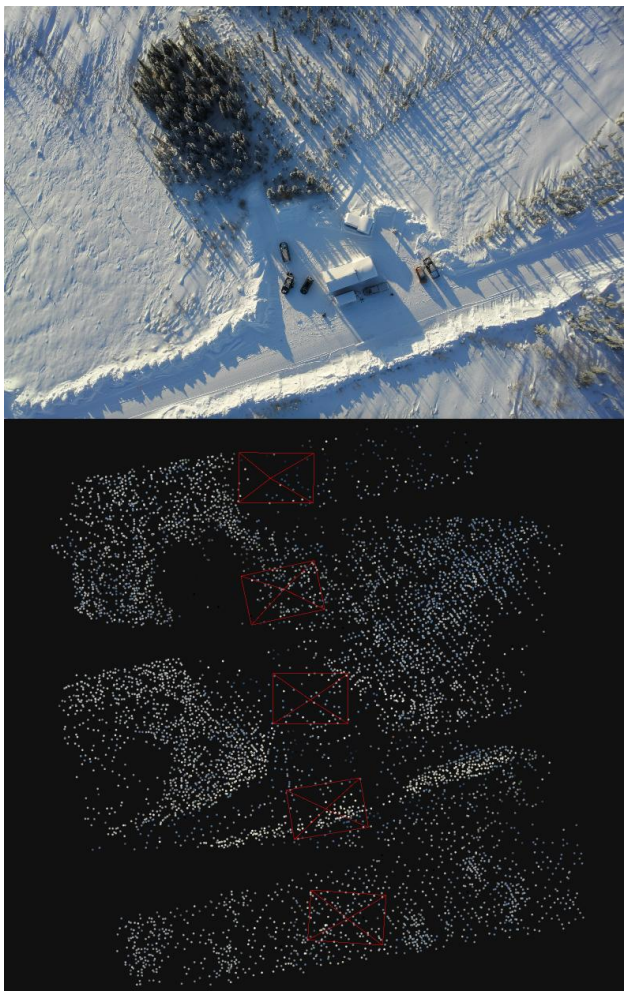

Figure 5. Combined SuperPoint+DISK reconstruction example

4.6 Sequence-Level Trends

The per-sequence materials also reveal a consistent trend between the learned approaches themselves. DISK produced more observations than SuperPoint in seven of the eight representative small-sequence outputs included in the supporting tables, often by a substantial margin. However, this observation advantage did not always translate into broader image registration or more globally interpretable point clouds, because many of DISK's additional observations were concentrated in a limited set of image regions.

SuperPoint, despite producing fewer observations overall, often maintained more even image coverage. This pattern reinforces the qualitative interpretation seen in the overlays: learned performance cannot be summarized by a single scalar metric, because observation density, image coverage, and structural interpretability each emphasize different aspects of the reconstruction problem.

5. DISCUSSION

5.1 Interpretation of the Comparative Results

Our results indicate that SIFT+NN remains a strong and robust baseline for Arctic UAV imagery under a standard COLMAP-style pipeline. Across the aggregate statistics, it registered more images, produced more matched observations, longer tracks, and lower reprojection errors than the learned methods in most structured scenes. This advantage was especially visible in larger sequences containing roads, rooftops, and other man-made structure, where stable and spatially diverse correspondences supported more complete 3D reconstruction.

At the same time, the results also show that the learned methods provide complementary behavior on the most challenging subsets. On the difficult 56-image sequence lacking clear man-made structure, SIFT+NN collapsed to only two registered images, whereas SuperPoint and DISK still reconstructed several views and hundreds of points. SuperPoint generally produced more even spatial coverage, while DISK often concentrated detections in high-confidence regions. These different behaviors suggest that the learned methods are not simply weaker versions of SIFT, but respond differently to low-texture and repetitive content.

5.2 Domain Mismatch and Method Limitations

We attribute the performance gap mainly to two factors already suggested by the experimental setup. First, SuperPoint and DISK were trained on generic synthetic, natural, or multi-view datasets rather than aerial winter imagery. This domain mismatch is likely important in scenes dominated by snow, tree canopy, and repetitive natural textures. Second, the reconstruction backend is tuned around SIFT-like features and descriptors, which may favor the broader and more stable distributions produced by the classical baseline.

A further factor contributing to the observed performance gap is the interaction between feature spatial distribution and SfM backend assumptions. Classical pipelines such as COLMAP implicitly assume a relatively uniform spatial distribution of keypoints to ensure stable geometric verification and bundle adjustment. In contrast, learned detectors such as DISK tend to concentrate keypoints in high-confidence regions, which can lead to locally dense but globally sparse correspondences. While these clusters may be individually reliable, they reduce the spatial coverage necessary for robust pose estimation across wide baselines. This mismatch suggests that improving learned feature performance in SfM is not solely a matter of detector quality, but also of aligning feature distribution with the geometric requirements of the reconstruction pipeline. Future work may therefore benefit from incorporating explicit spatial regularization or coverage-aware objectives during training.

This interpretation should be kept within the limits of the present study. The experiments isolate the keypoint front-end within one shared HLOC/PyCOLMAP pipeline, but they do not exhaustively vary learned matching strategies, detector thresholds, or backend settings tailored to learned features. The study therefore shows comparative behavior within a standard SfM workflow, rather than an absolute ceiling on what SuperPoint or DISK could achieve after domain adaptation or pipeline redesign.

5.3 Practical Implications

For practical Arctic UAV mapping, the results support a conservative recommendation: SIFT+NN should remain an essential baseline and may still be the preferred front-end when the objective is the most complete and stable reconstruction under standard COLMAP-style settings. However, the learned methods remain relevant because they provide useful

complementary behavior on subsets where classical features can fail.

6. CONCLUSIONS

This paper compared SIFT+NN, SuperPoint, and DISK for SfM reconstruction on challenging Arctic UAV winter imagery within a shared HLOC and PyCOLMAP pipeline. The results indicate that SIFT+NN remains a strong and robust baseline under a standard COLMAP-style workflow. On a 151-image sequence, it reconstructed all images and produced the most matched observations, while on average it achieved the lowest reprojection error and the greatest reconstruction completeness.

SuperPoint and DISK nevertheless showed complementary behavior, particularly on the most difficult subsets lacking clear man-made structure. SuperPoint produced more even spatial coverage, whereas DISK often formed dense clusters in high-confidence regions. Across all methods, homogeneous snow and dense tree canopy remained persistent failure cases, showing that keypoint detection alone is insufficient when reliable cross-view matching cannot be maintained.

Overall, the study suggests that the current performance gap is driven less by a fundamental limitation of learned features than by domain mismatch and by a backend tuned for SIFT-like correspondences. Domain-specific pretraining on aerial winter imagery and modified objectives encouraging greater spatial diversity remain promising directions for improving learned keypoint detectors for Arctic photogrammetric applications.

REFERENCES

DeTone, D., Malisiewicz, T., Rabinovich, A., 2018. SuperPoint: Self-supervised interest point detection and description. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 224-236, https://doi.org/10.48550/arXiv.1712.07629.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis., 60(2), 91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94.

Sarlin, P.-E., DeTone, D., Malisiewicz, T., Rabinovich, A., 2020. SuperGlue: Learning feature matching with graph neural networks. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 4938-4947.

Schönberger, J.L., Frahm, J.-M., 2016. Structure-from-Motion revisited. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 4104-4113. https://doi.org/10.1109/CVPR.2016.445.

Tyszkiewicz, M.J., Fua, P., Trulls, E., 2020. DISK: Learning local features with policy gradient. Adv. Neural Inf. Process. Syst., 33, 14254-14265.