Geometry-Aware Feature Matching for Large-Scale Structure from Motion

Gonglin Chen[1,2]   Jinsen Wu[1,2]   Haiwei Chen[1,2]   Wenbin Teng[1,2]   Zhiyuan Gao[1,2]
Andrew Feng[1]   Rongjun Qin[3]   Yajie Zhao[1,2*]

{gonglinc, jinsenwu}@usc.edu, chw9308@hotmail.com, {wenbinte, gaozhiyu}@usc.edu
andrewfengusa@gmail.com, qin.324@osu.edu, zhao@ict.usc.edu

[1]Institute for Creative Technologies   [2]University of Southern California
[3]The Ohio State University

Abstract

Establishing consistent and dense correspondences across multiple images is crucial for Structure from Motion (SfM) systems. Significant view changes, such as air-to-ground with very sparse view overlap, pose an even greater challenge to the correspondence solvers. We present a novel optimization-based approach that significantly enhances existing feature matching methods by introducing geometry cues in addition to color cues. This helps fill gaps when there is less overlap in large-scale scenarios. Our method formulates geometric verification as an optimization problem, guiding feature matching within detector-free methods and using sparse correspondences from detector-based methods as anchor points. By enforcing geometric constraints via the Sampson Distance, our approach ensures that the denser correspondences from detector-free methods are geometrically consistent and more accurate. This hybrid strategy significantly improves correspondence density and accuracy, mitigates multi-view inconsistencies, and leads to notable advancements in camera pose accuracy and point cloud density. It outperforms state-of-the-art feature matching methods on benchmark datasets and enables feature matching in challenging extreme large-scale settings. Project page: https://xtcpete.github.io/geo-website/.

1. Introduction

Correspondence, as perhaps the most fundamental problem in computer vision, helps us with mapping the world. Solving correspondence is required in most applications in Augmented Reality and Robotics since it derives the camera
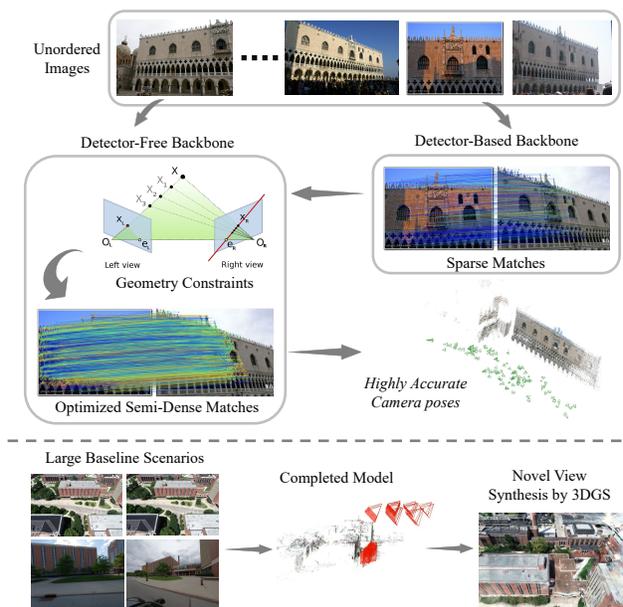

Figure 1. Our proposed method bridges detector-based feature matching with detector-free feature matching. We utilize sparse correspondences as geometric priors to iteratively optimize the matching process. By leveraging the denser matches from detector-free matchers, our method achieves highly accurate camera pose recovery and generates denser point clouds. This approach is particularly effective in challenging large baseline scenarios, such as air-to-ground imagery, and provides substantial benefits to downstream applications like novel view synthesis.

poses and 3D point locations [16].

In the 3D reconstruction from a set of uncalibrated images, the correspondence problem meets two additional demands: 1) the demand to be able to observe a corresponding point across multiple views, which is known as the *track*

---

length of a correspondence point, and 2) the demand to find a *dense* set of such points such that they are sufficient for the recovery of 3D geometry of the scene.

In this paper, we consider a very difficult correspondence problem—dense, consistent correspondences across a sequence of frames with large camera baselines. To our knowledge, no prior works have achieved satisfactory results in the presence of these challenges: both detector-based and detector-free methods struggle with significant view changes, as illustrated in Figure 2, due to their inherent limitations. Detector-based methods often match only a small number of keypoints, making accurate image registration difficult.

Figure 2. Challenging case of detector-based and detector-free feature matching methods for cross view images.

While the 3D reconstruction problem described thereof has been well studied for decades, they are notably tackled by either requiring a pre-calibration of the captured cameras [12, 27], or the requirement that consecutive frames captured are closely located to each other [8, 28, 35, 45]. In recent years, however, correspondence methods based on learned artificial neural networks (ANNs) have demonstrated significant improvements over the accuracy and robustness of the matching between a pair of in-the-wild, uncalibrated images [5, 9, 10, 21, 36, 40, 47], which has given rise to the possibility of reconstructing 3D scenes from the abundant, in-the-wild photos captured in daily setting [1, 26, 31, 38].

Recent state-of-the-art correspondence methods are separated into two major branches: the detector-based methods [9, 23, 36] that employ two independent steps to extract feature descriptor for key points first and perform matching later, and the detector-free methods [5, 10, 40, 42] that find matches in an end-to-end trained neural networks in one shot. While the detector-based methods excel in performing matching among image sets with rich texture and small view point changes, their reliance on the keypoint detection and description poses significant challenges when images of interest lack textures, or their local features are not sufficient for matching.

On the other hand, the detector-free methods [5, 10, 40, 42] naturally find much denser correspondences between a pair of images with a dense search, but they typically lack control of the locations of keypoints across views and

are optimized for sub-pixel correspondences that depend on specific image pairs. As shown in Figure 3, this pair dependency makes them unsuitable for maintaining long-track correspondence across multiple images. Therefore, it leads to poor performance when these methods are directly applied to Structure-from-Motion (SfM) reconstruction. Though quantization on the found matches can mitigate the problem of short track length, it comes at the price of accuracy and requires additional refinement that makes the entire system more computationally expensive.
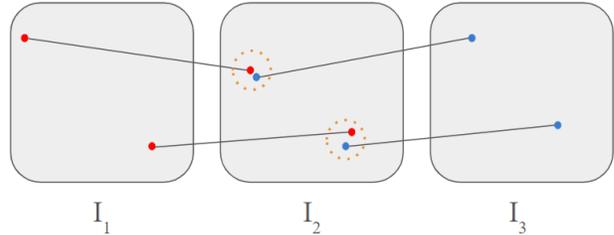

Figure 3. Matches from detector-free methods suffered from the problem of short track length (the track breaks at I_2). The location of the matches in I_2 depends on the images that it pairs with.

We aim to bridge the gap between the detector-based and detector-free methods in order to obtain **dense** correspondences that can be tracked **across multiple views** with **large baselines** - addressing a very difficult correspondence problem in SfM reconstruction from in-the-wild image collections. Specifically, we propose a novel optimization-based method that refines the pair-wise dense correspondences by iteratively filtering and reassigning correspondence under a geometric constraint formulated by the Sampson Distance [13]. This method allows the early rejection of outliers during the matching steps, improving the reliability of correspondences in large-baseline scenarios. Interestingly, we have observed that incorporating both priors from the existing detector-based method and enforcing the geometric constraint helps our model achieve better camera prediction and longer track length among the found correspondences, as demonstrated in Section 4.4. The designed optimization module thus consistently improves upon any detector-based and detector-free backbones, by providing denser correspondences compared to the detector-based methods, while refining the detector-free results to be more accurate and consistent across multiple views. 3D point cloud reconstructed with our feature matcher thus excels in both density and accuracy compared to previous state-of-the-art approaches.

To validate the effectiveness of our proposed method, we have conducted comprehensive evaluations using several publicly available datasets, including Image Matching Competition Benchmark [15] and ScanNet [7]. We compared the SfM reconstruction results using matches from

our feature matcher against several state-of-the-art baseline methods. The results demonstrated better accuracy in estimated camera poses and longer track length compared to those from the baselines. Additionally, we have extended our evaluation to a challenging large-baseline setting that requires mapping from the UAV to street view perspectives (Figure 6).

2. Related work

2.1. Detector-based Feature Matching

A conventional detector-based matching framework typically detects keypoints in images, describes them using high-dimensional vectors, and further applies a matching algorithm to find the correspondences between these image sets [25]. Historically, numerous handcrafted methods like SIFT [23], SURF [3], and ORB [34] have been used for the keypoint detection. Lately, learning-based methods [9, 36, 44], particularly those leveraging convolution neural networks (CNNs), have significantly improved the performance over the handcrafted features and shown robustness under large viewpoint and illumination changes. For instance, SuperPoint [9] achieved a good performance by creating a vast dataset of pseudo-ground-truth interest points and using a self-supervised interest point detector. SuperGlue [36] learns feature matching using point features from a graph neural network (GNN) and attention layers that integrate global context information.

Patch2Pix [48] has made further improvement with patch-level matches and geometry-guided refinement that has enhanced the correspondence accuracy. Patch2Pix [48] utilizes the Sampson distance [13] to refine initial patch-level matches to pixel-level accuracy, ensuring geometric consistency and precision. However, Patch2Pix [48] is limited to refining existing sparse matches and add new matches or extend beyond the initial patch-level matches. Despite their success, detector-based methods are limited by their dependency on feature detectors, which struggle to consistently identify distinct interest points in areas that lack clear and distinguishable features.

2.2. Detector-free Feature Matching

Detector-free approaches [5, 6, 14, 37, 40] bypass the feature detection stage and directly generate matches, following a concept that originates from SIFT Flow [22]. The introduction of learning-based methods that employ contrastive loss to train pixel-wise feature descriptors [6, 37] marked a significant development in this field. These methods - similar to detector-based approaches - employ the nearest neighbor search for post-processing to match the dense descriptors.

NCNet [32] learns dense correspondences in an end-to-end manner, creating 4D cost volumes that account for all possible inter-image matches. 4D convolutions are applied to refine these volumes, promoting consistency across the correspondence field. Building on this work, SparseNC-Net [33] improved the efficiency with sparse convolutions. DRCNet [19] improved dense matching accuracy with a coarse-to-fine architecture. While NCNet addresses all potential matches, its 4D convolution's receptive field is limited to small local neighborhoods.

LoFTR [40], in contrast, aims for a broader match consensus, leveraging the expansive receptive field of transformers. ASpanFormer [5], MatchFormer [42], and Efficient LoFTR [43] follows the transformer-based design to utilize multilevel cross-attention to mutually update image features, thereby incorporating dual-view contexts for achieving both global and local consensus. While these methods represent significant improvements to the state-of-the-art, they still struggle to uphold high-level contexts such as epipolar geometries. Therefore, we propose a geometry-aware approach that enforces epipolar consistency in dense feature matching.

2.3. Structure from Motion

Structure from Motion (SfM) methods have been extensively studied [4, 26] in the past. Some of these efforts focus on the scaling to large-scene reconstruction [1, 38, 39]. SfM methods mainly use detector-based matching at the beginning of their pipelines, which limits the overall SfM performance in challenging cases. Some end-to-end SfM methods propose to directly model camera poses [30, 41, 46]. These methods avoid the bottleneck of feature detection but have limited scalability to real-world scenarios. Although Recent differentiable rendering methods [17, 24] can optimize poses and the scene representation simultaneously, they rely on the traditional SfM pipelines like COLMAP[38, 39] to provide the initial camera poses. Lindenberger et al. [20] improves the accuracy of SfM and visual localization by refining keypoints, camera poses, and 3D points post to the traditional SfM pipelines. Unlike previous SfM methods, our proposed feature matcher ensures robustness and local geometry consistency at the first step of an SfM system.

3. Methods

The core of our methods is an optimization module that operates on existing sparse matches M_s and two feature maps f_c^A and f_c^B from a pair of images. The module optimizes for the relative camera poses between the two images and refines the input coarse correspondences into refined coarse correspondence $M_{c_{final}}$ that is geometric consistent under a Sampson distance [13] constraint. In Section 3.1, we provide an overview of how the sparse matches and feature maps are obtained, as well as the definition of Sampson distance. In Section 3.2, we discuss the details of our optimization module.
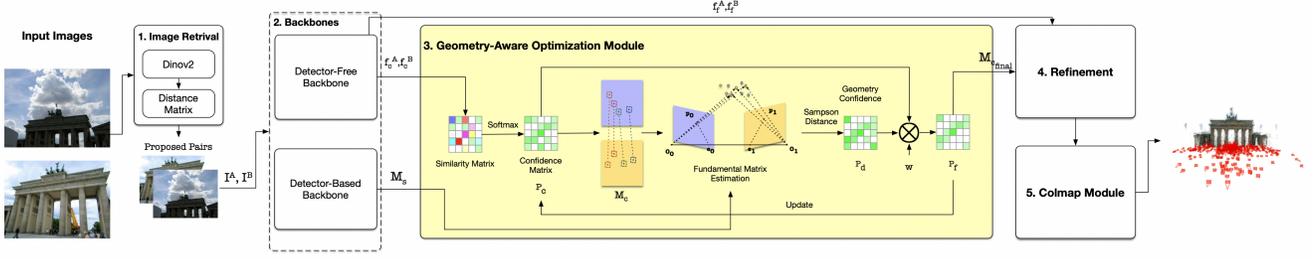
Figure 4. An overview of our pipeline for SfM reconstruction. *1.* the pipeline runs image retrieval based on global embeddings generated by dinov2 [29]. *2.* A backbone module takes image pairs as input. The image pairs will be processed by a Detector-Free Backbone and a Detector-Based backbone. *3.* A geometry-aware optimization module is applied to iteratively optimize the fundamental matrix and matches with anchor points from detector-based methods. *4.* The final matched coarse points are refined using a correlation-based refinement block. *5.* Final refined matches are then fed into COLMAP [38, 39] for SfM.

3.1. Preliminaries

3.1.1 Detector Based Method

Given two images I^A and I^B, our method accepts a detector-based backbone F_{db} that first identifies the sparse interest points with a keypoint detector. Then, local features are extracted around the identified interest points, encoding the local visual appearance in high-dimensional vectors. These features are then matched between I^A and I^B using either learned neural networks [21, 36] or nearest-neighbor search to generate a set of sparse correspondences M_s. In our implementation, we use SuperPoint [9] + SuperGlue [36] to obtain sparse correspondence between a given image pair as it has been shown that the two methods combined can provide anchor points that introduce robust and good priors to guide dense matching [18]. We also evaluate the performance of different detector-based backbones in Section 4.4.

3.1.2 Detector Free Method

Given two images I^A and I^B, our method accepts a detector-free backbone F_{df} that processes the entire image to output semi-dense correspondences without relying on explicit keypoint detection. Instead, dense feature extraction is performed using deep neural networks. The output features f_c^A and f_c^B are then matched using similarity metrics, which are generated to quantify the confidence and select the final correspondences. In our implementation, we use AspanFormer [5] as our detector-free backbone.

3.1.3 Sampson Distance

Given a fundamental matrix F_m and the correspondence P_i^A, P_j^B, the Sampson Distance $d(i,j)$ is the distance of P_i^A, P_j^B to their true epipolar line, which is given by:

$$d(i,j) = \frac{\left(\mathbf{P}_j^{B\top}\mathbf{F_m}\mathbf{P}_i^A\right)^2}{(\mathbf{F_m}\mathbf{P}_i^A)_1^2 + (\mathbf{F_m}\mathbf{P}_i^A)_2^2 + (\mathbf{F_m}^\top\mathbf{P}_j^B)_1^2 + (\mathbf{F_m}^\top\mathbf{P}_j^B)_2^2},$$

$$(1)$$

where point $\mathbf{P}_i^A = (x_i, y_i, 1)^\top$ and point $\mathbf{P}_i^B = (x_j, y_j, 1)^\top$ are two points in I^A and I^B respectively and $(\mathbf{F_m}\mathbf{P}_i^A)_k^2$ is the k-th entry of the vector $(\mathbf{F_m}\mathbf{P}_i^A)$. Therefore, the Sampson distance [13] gives the first-order approximation of the reprojection error. The error is minimized when two points are a perfect match.

3.2. Geometry Aware Optimization Module

Our proposed optimization module aims to enforce geometric consistency among correspondences and guide the feature matching with geometric information. We utilize M_s from F_{db} to initialize the fundamental matrix. With dense features f_c^A and f_c^B from F_{df}, our module first computes a confidence map P_c using the similarity metrics. Then the Sampson Distance [13] is used as a geometry constraint of an image pair to iteratively select and reassign the correspondences by updating P_c. Such constraints can reduce mismatches and address missed matches due to insufficient appearance features and repeating texture patterns.

Initial Step Given the features f_c^A and f_c^B, the initial confidence matrix is obtained with a dual-softmax operator. For the i-th feature in I^A and the j-th feature in I^B, the similarity score is calculated as

$$S(i,j) = \langle \mathbf{f}_c^A(i), \mathbf{f}_c^B(j) \rangle,$$

$$(2)$$

where $\langle a, b \rangle$ denotes the dot product. Then the confidence level P_c is calculated as

$$P_c(i,j) = \text{softmax}(S(i,\cdot))_j \cdot \text{softmax}(S(\cdot,j))_i,$$

$$(3)$$

Based on the initial confidence matrix,

$$M_c = \{(\tilde{i}, \tilde{j}) | \forall (\tilde{i}, \tilde{j}) \in MNN(P_c), P_c(\tilde{i}, \tilde{j}) \geq \theta_{iter}\}.$$

$$(4)$$

we follow LoFTR [40] to apply mutual nearest neighbor (MNN) criteria using equation 4 to obtain the initial matches. The row index i represents the index of the patch from I^A and the column index j represents the index from I^B. $(\tilde{i}, \tilde{j})$ is a match if and only if the confidence

level is the maximum of both ith row and jth column. We choose points where their confidences are larger than θ_{iter} as matches to enter iterative steps.

The initial fundamental matrix F_m is calculated using the normalized 8-points method if there exist enough good anchor points in M_s, which is at least 10 points with confidence larger than 0.5. Otherwise, the initial fundamental matrix will be calculated using the top half matches from equation 4.

Iterative Steps Our iterative steps begin with initialized coarse matches M_s and fundamental matrix F_m and update them in each iteration. The Sampson distance error is calculated for all initial matches according to Equation 1. Geometric confidence is then estimated from the Sampson distances as

$$P_d(i,j) = \text{sigmoid}(\text{relu}(\tau - d(i,j)), \qquad (5)$$

where τ is a distance threshold, and smaller values of τ will lead to more strict constraints. The geometric confidence for correspondences that have distances larger than the distance threshold will be 0.5. In other words, the confidence of matches that violate the geometry constraint will be penalized.

Finally, the confidence map is updated by the geometric confidence map by a given weight w. Therefore, the final confidence is

$$P_c(i,j)_i = P_c(i,j)_{i-1} \cdot P_d(i,j)_{i-1} \cdot w. \qquad (6)$$

At every iteration, the updated confidence map is normalized using min-max normalization to avoid extreme values. When the iteration ends, the module will select the final coarse-level matches $M_{c_{final}}$ based on the final confidence map from the last iteration with mutual nearest neighbor criteria according to equation 4, but with a different threshold θ_c to avoid selecting a large amount of correspondences with very small confidences. The refined matches $M_{c_{final}}$ are further fed into a correlation-based refinement block, which is the same with LoFTR [40], to obtain the final matching results.

3.3. Structure from Motion Reconstruction

The final matching results are fed into COLMAP [38] for estimating structure from motion. By integrating our robust and dense matches into COLMAP [38], our method is shown to enhance the accuracy and completeness of the camera poses and reconstructed point cloud, resulting in superior 3D models, even in challenging scenarios with significant viewpoint changes, as discussed in Section 4.3.

3.4. Implementation Details

The Sampson Distance threshold τ is set to 10 which allows a few pixels of error. The iteration of the optimization module is set to 10 and the weight w is 1.2. The confidence threshold θ_{iter} is set to a small value of 0.01. The idea is to conduct a comprehensive search across the space. The effect of different hyperparameters will be discussed in Section 4.4.

4. Experiments

In this section, we demonstrate the performance of our method in the SfM reconstruction task in terms of the accuracy of camera pose estimation and its application in challenging real-world scenarios. We also conduct a comprehensive ablation study. All experiments were run on an RTX 4090 GPU with 24GB VRAM and a 13th Gen Intel® Core™ i9-13900KF CPU with 32 cores and 126GB RAM.

4.1. Dataset

We utilize the Image Matching Challenge (IMC) 2021 [15] and ScanNet [7] benchmark datasets to evaluate our method's performance in SfM reconstruction for outdoor and indoor settings respectively. As there are no publicly available air-to-ground datasets, we collected two additional air-to-ground datasets to demonstrate our method's ability to find reliable correspondences in this challenging scenario.

The IMC Phototourism benchmark [15] encompasses multiple large-scale outdoor scenes. The test set includes 9 sites, and each set has 100 images. The ground truth or pseudo ground truth camera poses are obtained from COLMAP [38] on the entire collection of images and then sampled into test sets. ScanNet [7] contains 1613 monocular indoor sequences with ground truth poses and depth maps. We use the data split of the test set following SuperGlue [36] and LoFTR [40], which includes 15 sequences. The primary challenge lies in accurately recovering the camera poses from in-the-wild images with significant viewpoint and illumination changes. For all test scenes, the images are considered unordered.

The first collected air-to-ground dataset consists of 16 UAV images and 21 ground images, while the second dataset contains 16 UAV images and 24 ground images. The primary challenge for these datasets is to accurately register UAV images with ground view images. For all scenes, the images are considered unordered and are matched exhaustively. When aerial images are matched with ground images, to avoid failure in finding correspondences due to resizing, the aerial images are cropped into quarters plus a center region to generate five pairs. The matches are subsequently merged to obtain the final matches. This strategy is used for all methods.

4.2. Metrics and Comparing Methods

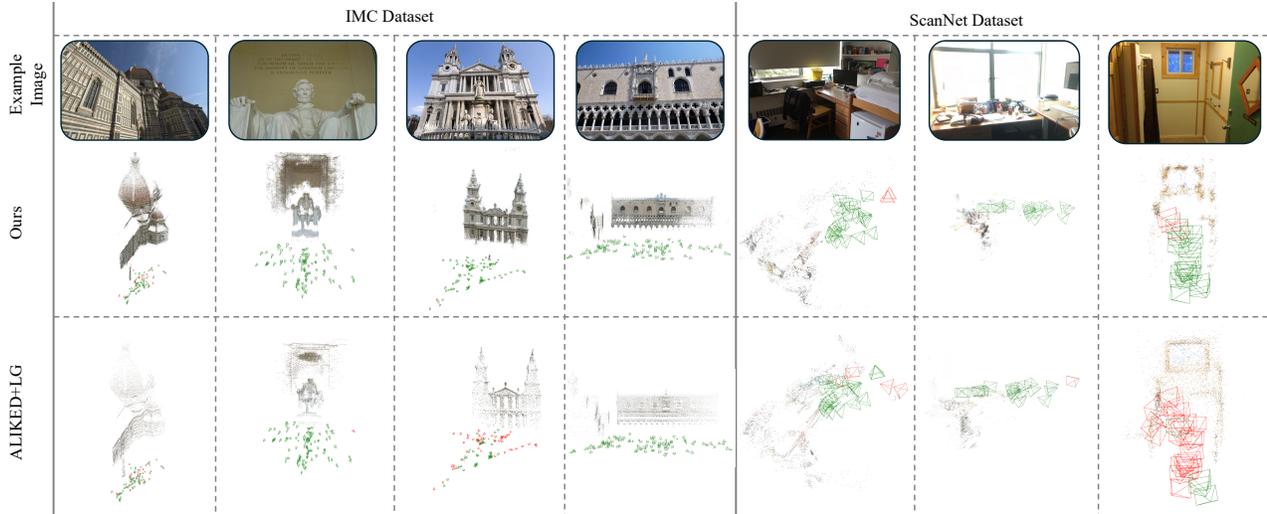We report the metric area-under-curve (AUC) to evaluate camera pose accuracy for IMC 2021 [15] and ScanNet [7].

Figure 5. **Qualitative Results**. Our method is qualitatively compared with ALIKED [47] + LG [21] on multiple scenes. Green cameras have less than 3° absolute pose error, while red cameras have an error larger than 3°. More results can be found in supplementary material.

Method	Pose Estimation AUC ↑		
	@3°	@5°	@10°
SIFT [23]+NN	51.5	65.3	78.9
SP [9]+SG [36]	59.6	71.0	83.3
ALIKED [47]+LG [21]	<u>66.0</u>	<u>76.8</u>	<u>87.6</u>
DKM [10]	42.3	58.5	76.3
RoMa [11]	36.3	51.6	69.5
Ours	**69.1**	**80.5**	**90.1**

Table 1. Estimated pose errors on IMC 2021 phototourism [15] in outdoor scences. Results are averaged across all scenes.

Method	Pose Estimation AUC ↑		
	@3°	@5°	@10°
SIFT [23]+NN	18.6	24.1	30.4
SP [9]+SG [36]	<u>45.9</u>	<u>61.1</u>	73.8
ALIKED [47]+LG [21]	44.0	59.5	<u>74.3</u>
DKM [10]	23.5	39.0	60.7
RoMa [11]	28.5	47.3	67.6
Ours	**46.0**	**63.4**	**77.6**

Table 2. Estimated pose errors on ScanNet [7] in indoor scences. Results are averaged across all scenes.

Specifically, for every reconstructed model, we aligned the predicted camera poses with the ground truth camera poses. We then computed the absolute angular translation and rotation errors. The AUC selects the maximum between these two errors and outputs the area under the accuracy threshold curve. As the collected air-to-ground datasets lack ground-truth camera poses, we follow [38] to report the number of registered images, the number of 3D points, the average track length, and the average projection error (px) in this setting to evaluate the methods by the completeness of the reconstructed model.

We compared our method against four different methods: SIFT [23]+NN, which is the default feature matching method in COLMAP [38, 39]; Super-Point [9]+Superglue [36], a robust method widely used in SfM; ALIKED [47]+LightGlue [21], a state-of-the-art detector-based feature matching method; DKM [10] and RoMa [11], two state-of-the-art dense feature matching methods. We omit RoMa [11] for air-to-ground datasets since DKM [10] appears to be superior to RoMa [11] in

terms of 3D reconstruction for outdoor scenes.

In order to reduce the computational time, the number of matches per pair is limited to 4K for all methods. Bold text indicates the best results, and underlined text indicates the second best.

4.3. Comparison to State of the Art Methods

SfM results using our feature matching method on both the IMC 2021 [15] and the ScanNet [7] datasets have shown significant improvements in performance. As shown in Table 1 and Table 2, our feature matcher contributes to the highest average pose estimation AUC in both datasets. Specifically, Table 1 indicates an increase of +3.1 in AUC@3° for estimated camera poses compared to ALIKED [47] + LG [21], as can be visually shown in Figure 5. Similar gains are observed for ScanNet [7], as shown in Table 2. These findings underscore the efficacy of our method in achieving superior SfM results.

As shown in Table 3 and Figure 6, compared to other methods, our method is able to register all images and is the
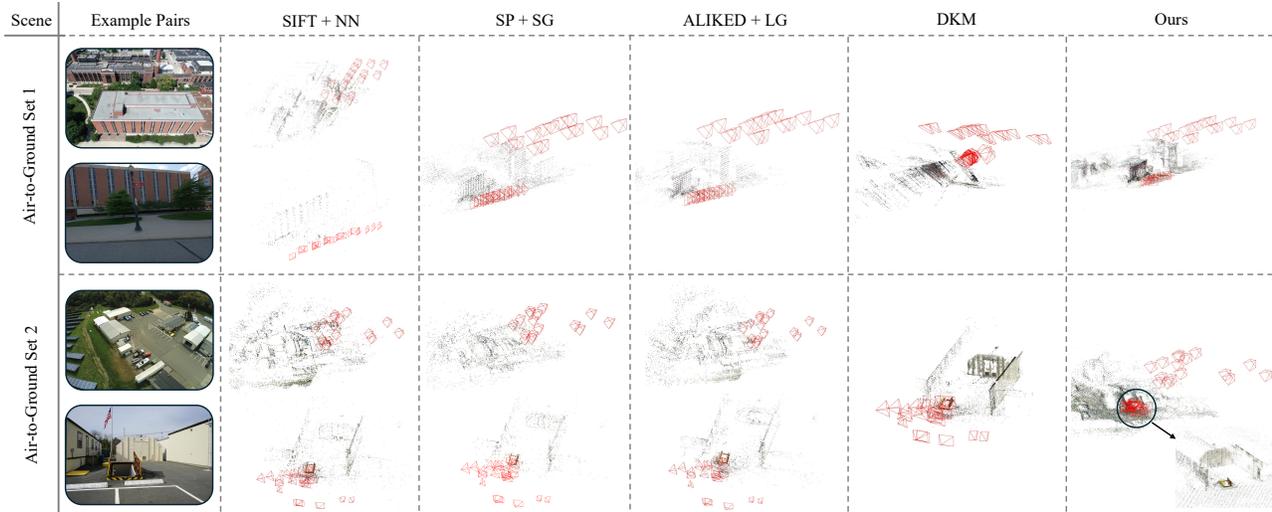
Figure 6. **Qualitative Results**. Our method is qualitatively compared with other feature matching methods on collected air-to-ground datasets. Red cameras are recovered poses.

Metrics	Air to Ground Set 1 (37 Images)					Air to Ground Set 2 (40 Images)				
	SIFT+NN	*SP+SG*	*ALIKED+LG*	*DKM*	*Ours*	*SIFT+NN*	*SP+SG*	*ALIKED+LG*	*DKM*	*Ours*
# Registered ↑	16/18	36	36	36	**37**	16/24	16/24	16/24	25	**40**
# 3D Points ↑	11.2K/1.27K	11.5K	12.5K	22.1K	**53.2K**	13.0K/11.3K	5.52K/3.60K	7.95K/8.75	48.0K	**74.4K**
Avg. Track Length ↑	3.4/2.7	3.08	**3.7**	2.2	3.1	3.5/3.6	3.4/3.2	3.9/3.5	2.8	**3.6**
Avg. Reproj. Error ↓	1.05/0.99	1.32	1.25	**0.81**	1.22	1.15/1.00	1.39/1.36	1.35/1.33	**0.84**	1.28

Table 3. SfM results with different feature matchers on Air-to-Ground datasets. For methods that generate 2 separate models, we report them as "air model/ground model".

only method that can align the ground images with UAV images for set 2. We have noticed that reconstructed camera poses using DKM [10] in set 1 are apparently wrong with ground images pointing down and speculate that it is due to the lack of multi-image matches for feature tracking. In addition, DKM [10] also fails in reconstructing a model for UAV images in set 2. These results further validate the practical utility and effectiveness of our feature matcher in real-world scenarios.

4.4. Ablation Study

We have conducted an ablation study to investigate the impact of different parameters and design choices on both feature matching and 3D reconstruction.

Ablation study on Homography Estimation. Feature matching is evaluated on the widely adopted HPatches dataset [2]. HPatches [2] contains a total of 108 sequences with significant illumination changes and large viewpoint changes. We follow the evaluation protocol of LoFTR [40] to resize the shorter size of the image to 480 and report AUCs at 3 different thresholds. Table 4 shows results on different parameters and design choices on HPathes [2]. The "Baseline" denotes only using ASpanFormer [5]. The

results show the effects of different parameters in feature matching. Additionally, the full method outperforms other design variations, confirming the robustness and effectiveness of our approach in enhancing feature matching accuracy. More experiments related to feature matching can be found in supplementary material.

Ablation study on 3D reconstruction. In order to comprehensively assess the efficacy of our proposed components within the SfM system, we also measure the impact of the aforementioned design changes on 3D reconstruction. To this end, we randomly sample 20 images from each scene of IMC 2021 phototourism datasets [15] and compare the SfM reconstruction results based on pose estimation. We follow IMC [15] to report relative pose errors for all possible pairs since we are using subsets for this experiment. All results are averaged across all scenes and all methods are evaluated on the same subset of data.

"SG" denotes anchor points from SuperGlue [36]; if the optimization ("Geo") module is not enabled, anchor points are naively concatenating to final matches. If there are no anchor points, the optimization ("Geo") module is initialized with the top half most confident matches from only the detector-free backbone. As shown in Tables 5 and 6, the

Types	Values	AUC		
		@3px	@5px	@10px
(1) Updating weight	$w = 1.2$	**70.3**	**79.5**	**88.0**
	$w = 1.5$	70.2	79.4	**88.0**
	$w = 1.8$	**70.3**	**79.5**	**88.0**
(2) Number of iterations	No optimization.	68.9	78.0	87.4
	5 iter	68.2	78.1	87.1
	10 iter	**70.3**	**79.5**	**88.0**
	20 iter	68.2	78.1	87.1
(3) Distance threshold	$\tau = 1$	69.0	78.5	87.0
	$\tau = 10$	70.3	79.5	**88.0**
	$\tau = 50$	**71.2**	**79.8**	87.6
	$\tau = 200$	70.7	79.73	87.7
(4) Designs	Full method	**70.3**	**79.5**	**88.0**
	No anchor points	67.4	77.6	86.8
	No optimization	68.9	78.0	87.4
	Baseline	66.1	75.9	84.8

Table 4. Ablation study of our proposed method in Homography Estimation on Hpatches [2].

SG	Geo	Pose Estimation AUC↑		
		@3°	@5°	@10°
		50.0	62.4	75.6
✓		56.0	66.3	76.0
	✓	57.0	69.8	82.1
✓	✓	**58.6**	**70.8**	**82.9**

Table 5. Ablation study of our proposed method in 3D reconstruction of sampled images from IMC 2021 phototourism dataset. Results are averaged across all scenes.

SG	Geo	# Registered	# Points ↑	Reproj. Err(px) ↓	Mean Track Length ↑
		15	12.7K	**0.334**	3.3
✓		**20**	**72.9K**	0.581	5.3
	✓	19	55.9K	0.468	5.0
✓	✓	**20**	64.2K	0.508	**5.6**

Table 6. Ablation study of our proposed method in 3D reconstruction of sampled images from IMC 2021 phototourism dataset. Results are averaged across all scenes.

Detector Based Method	Pose Estimation AUC↑		
	@3°	@5°	@10°
SP [9]+SG [36]	**69.1**	**80.5**	**90.1**
ALIKED [47]+LG [21]	65.3	78.3	89.1

Table 7. Ablation study of using different detector-based backbone in SfM on IMC dataset [15].

results again demonstrate the effectiveness of our proposed method as it brings improvement over the baseline methods in the accuracy of SfM reconstruction.

Ablation study on detector-based backbones. Following the setting of Section 4.3, Table 7 compares the performance of using two detector-based methods on the IMC dataset [15]. This demonstrates that SP[9] + SG[36] is more effective in providing priors in guiding dense matching.

5. Limitation and Future Work

	Image Retrieval	Feature Matching	Reconstruction
Avg. Runtime (s)	3	1769	1478

Table 8. Average runtime analysis on IMC 2021 [15]. Each site contains around 100 images and image retrieval proposes an average of 1750 pairs.

The main limitation of our method is efficiency, due to the computational cost of the transformer-based architecture and the significant number of correspondences produced, as shown in Table 8. Utilizing a newly developed backbone model with a focus on efficiency, such as Efficient LoFTR [43], can improve it with slightly degraded performance.

Due to our method's reliance on the backbone model to provide initial matches for fundamental matrix estimation and its heuristic design, our method could fail with poor initial matches from backbone models. Nevertheless, our evaluations against baselines demonstrate a marked improvement in accuracy, as well as the capability to find reliable matches in challenging scenarios. Furthermore, we believe the accuracy of SfM reconstruction can be further improved by adopting multi-view refinement techniques, such as those proposed by pixSfM [20].

6. Conclusion

We propose the geometry-aware optimization module, an innovative optimization approach that bridges detector-based matcher and detector-free matcher to improve SfM reconstruction and accuracy of estimated camera poses. This module leverages epipolar constraints, iteratively optimizes for the best relative pose, and applies Sampson distance weighting to refine match precision. The introduction of geometry optimization enhances the system's ability to handle challenging scenarios, including drastic changes in perspective, scale, and illumination. Our method is shown to be also effective in 3D reconstruction for large baselines, a scenario where conventional methods often fail and perform unsatisfied.

7. Acknowledgement

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building rome in a day. In *2009 IEEE 12th International Conference on Computer Vision*, pages 72–79, 2009. 2, 3

[2] Vassileios Balntas, Karel Lenc, Andrea Vedaldi, and Krystian Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *CVPR*, 2017. 7, 8

[3] Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. Speeded-up robust features (surf). *CVIU*, 110(3):346–359, 2008. 3

[4] Paul Beardsley, Phil Torr, and Andrew Zisserman. 3d model acquisition from extended image sequences. In *Computer Vision — ECCV '96*, pages 683–695, Berlin, Heidelberg, 1996. Springer Berlin Heidelberg. 3

[5] Hongkai Chen, Zixin Luo, Lei Zhou, Yurun Tian, Mingmin Zhen, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Aspanformer: Detector-free image matching with adaptive span transformer. *ECCV*, 2022. 2, 3, 4, 7

[6] Christopher B Choy, JunYoung Gwak, Silvio Savarese, and Manmohan Chandraker. Universal correspondence network. *NeurIPS*, 2016. 3

[7] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. 2, 5, 6

[8] Andrew J. Davison, Ian D. Reid, Nicholas D. Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067, 2007. 2

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. *CVPR Workshops*, pages 224–236, 2018. 2, 3, 4, 6, 8

[10] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. DKM: Dense kernelized feature matching for geometry estimation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 6, 7

[11] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 6

[12] Andreas Geiger, Julius Ziegler, and Christoph Stiller. Stereoscan: Dense 3d reconstruction in real-time. In *2011 IEEE Intelligent Vehicles Symposium (IV)*, pages 963–968, 2011. 2

[13] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2 edition, 2003. 2, 3, 4

[14] Dihe Huang, Ying Chen, Yong Liu, Jianlin Liu, Shang Xu, Wenlong Wu, Yikang Ding, Fan Tang, and Chengjie Wang. Adaptive assignment for geometry aware local feature matching. *CVPR*, 2023. 3

[15] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision*, 2020. 2, 5, 6, 7, 8

[16] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering, 2023. 3

[18] Z. Kuang, J. Li, M. He, T. Wang, and Y. Zhao. Densegap: Graph-structured dense correspondence learning with anchor points. In *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022. 4

[19] Xinghui Li, Kai Han, Shuda Li, and Victor Prisacariu. Dual-resolution correspondence networks. *NeurIPS*, 2020. 3

[20] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 3, 8

[21] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 2, 4, 6, 8

[22] Ce Liu, Jenny Yuen, and Antonio Torralba. SIFT Flow: Dense correspondence across scenes and its applications. *T-PAMI*, 2010. 3

[23] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 2, 3, 6

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis, 2020. 3

[25] M.Karpushin, G.Valenzise, and F.Dufaux. Keypoint detection in rgbd images based on an anisotropic scale space. *IEEE Transactions on Multimedia*, 18(9):1762–1771, 2016. 3

[26] Roger Mohr, Long Quan, and Françoise Veillon. Relative 3d reconstruction using multiple uncalibrated images. *The International Journal of Robotics Research*, 14(6):619–632, 1995. 2, 3

[27] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real time localization and 3d reconstruction. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 363–370, 2006. 2

[28] Richard A. Newcombe, Steven J. Lovegrove, and Andrew J. Davison. Dtam: Dense tracking and mapping in real-time. In *2011 International Conference on Computer Vision*, pages 2320–2327, 2011. 2

[29] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 4

[30] Chethan M. Parameshwara, Gokul Hari, Cornelia Fermüller, Nitin J. Sanket, and Yiannis Aloimonos. Diffposenet: Direct differentiable camera pose estimation, 2022. 3

[31] Michal Polic, Wolfgang Förstner, and Tomas Pajdla. Fast and accurate camera covariance computation for large 3d reconstruction, 2018. 2

[32] Ignacio Rocco, Mircea Cimpoi, Relja Arandjelovic é, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Neighbourhood consensus networks. *NeurIPS*, 2018. 3

[33] Ignacio Rocco, Relja Arandjelovic é, and Josef Sivic. Efficient neighbourhood consensus networks via submanifold sparse convolutions. *ECCV*, 2020. 3

[34] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. *ICCV*, 2011. 3

[35] Renato F. Salas-Moreno, Richard A. Newcombe, Hauke Strasdat, Paul H.J. Kelly, and Andrew J. Davison. Slam++: Simultaneous localisation and mapping at the level of objects. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1352–1359, 2013. 2

[36] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 3, 4, 5, 6, 7, 8

[37] Tanner Schmidt, Richard Newcombe, and Dieter Fox. Self-supervised visual descriptor learning for dense correspondence. *RAL*, 2016. 3

[38] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2, 3, 4, 5, 6

[39] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 3, 4, 6

[40] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 2, 3, 4, 5, 7

[41] Sudheendra Vijayanarasimhan, Susanna Ricco, Cordelia Schmid, Rahul Sukthankar, and Katerina Fragkiadaki. Sfm-net: Learning of structure and motion from video, 2017. 3

[42] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *Asian Conference on Computer Vision*, 2022. 2, 3

[43] Yifan Wang, Xingyi He, Sida Peng, Dongli Tan, and Xiaowei Zhou. Efficient LoFTR: Semi-dense local feature matching with sparse-like speed. In *CVPR*, 2024. 3, 8

[44] Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, and Pascal Fua. Lift: Learned invariant feature transform. *ECCV*, 2016. 3

[45] Chao Yu, Zuxin Liu, Xin-Jun Liu, Fugui Xie, Yi Yang, Qi Wei, and Qiao Fei. Ds-slam: A semantic visual slam towards dynamic environments. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1168–1174, 2018. 2

[46] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild, 2022. 3

[47] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter C. Y. Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation, 2023. 2, 6, 8

[48] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *CVPR*, 2021. 3