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

Supplementary Material

1. More Experiments in Feature Matching

In this section, we demonstrate the effectiveness of our proposed geometry optimization module in improving the performance of the detector-free feature matcher without any geometry information from the detector-based method. The fundamental matrix is initialized by the top half of confident matches.

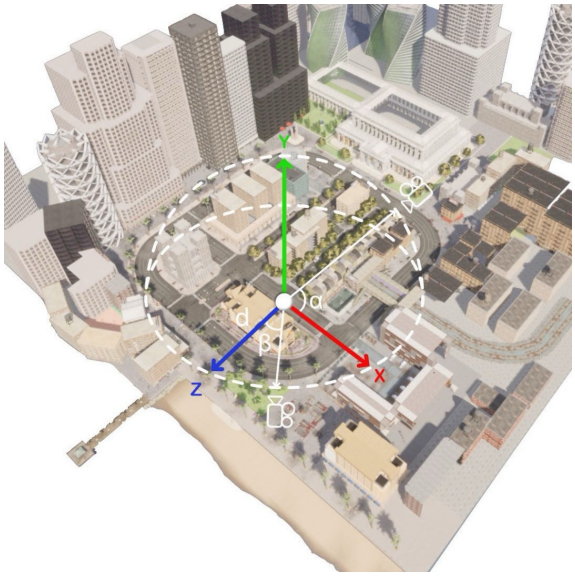

Figure 1. We propose continuous changes along three variable scale/distance d - the radius of the circle, α - the angle between the camera axis and x-axis, β - the angle between the projection of the camera axis on xz plane and z-axis.

1.1. Matching Precision

Dataset To understand the performance variations in the context of viewpoint changes, and our proposed optimization module's performance with different detector-free backbones. We generate a synthetic dataset with depth maps and camera poses using CARLA [5]. As illustrated in Figure 1, given an image pair I^A and I^B, we use the minimum value of a given variable for image I^A, 10 for distance, 0 for α and β. Then we maintain a consistent incremental unit of 1 for such variable for image I^B starting from 5 units of difference, while the other two variables remain the same for both images. We sample 25 image pairs per increment, culminating in a total of 900 image pairs per sequence. Then all images are resized with their longer dimensions adjusted to 832 for the testing.

Metrics and Comparing Methods The average matching precision of 25 image pairs is reported for each step, with correct matches defined based on a symmetric epipolar error threshold of less than $1e^{-4}$. Our methods are compared against three detector-free counterparts, namely LoFTR [17], ASpanFormer [2], and MatchFormer [19]. All models are trained on MegaDepth [10].

Results on Our Dataset. For all methods, we observed that the performance drops as the view differences between two images increase, as shown in Figure 2. Our methods consistently outperform the baseline methods, demonstrating a more gradual decline in precision as geometric disparities become more pronounced. While ASpanFormer [2] and MatchFormer [19] show robustness against scale variations and changes in the α angle, they are noticeably impacted by larger β angle differences. This is particularly apparent when the camera that captures image I^B moves horizontally away from the camera of image I^A, resulting in more significant changes in appearance features compared to the other two scenarios. In contrast, LoFTR [17], despite performing adequately under conditions with minimal geometric variation, demonstrates a marked decrease in its ability to find accurate matches as view differences grow.

The integration of our module into these methods has led to substantial improvements in precision. This is indicative of the module's ability to enhance matching accuracy, especially in scenarios where view differences are significant. The result suggests that our approach, by introducing direct geometry constraints early in the matching stage, can mitigate the impact of increasing view geometry challenges.

Method	**AUC ↑**		
	@3px	@5px	@10px
SuperGlue [14]	53.9	68.3	81.7
LoFTR [17]	65.9	75.6	84.6
TopicFM [8]	67.3	77.0	85.7
3DG-STFM [13]	64.7	73.1	81.0
ASpanFormer [2]	66.1	75.9	84.8
PDC-Net+ [18]	66.7	76.8	85.8
Ours-ASpan	<u>67.4</u>	<u>77.6</u>	<u>86.8</u>
DKM [6]	**71.3**	**80.6**	**88.5**

Table 1. Homography estimation on HPatches, measured in AUC (higher is better).

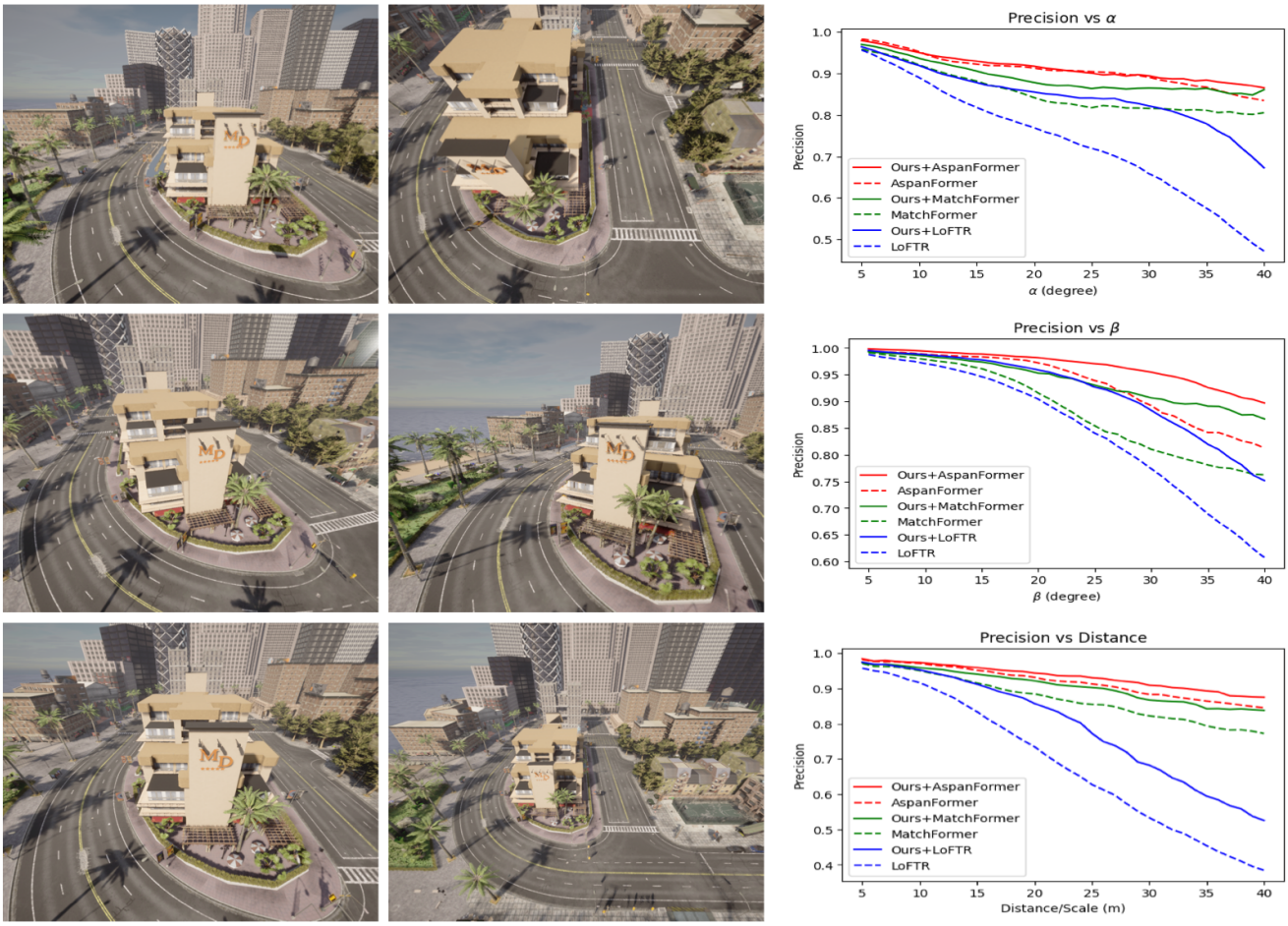

Figure 2. **Results on our Dataset.** The above illustrates our dataset and the precision evaluation of view changes. In the first row, α difference ranges from 5° to 40°; in the second row, β difference ranges from 5° to 40°; in the third row, distance/scale difference ranges from 5m to 40m. All examples show the largest differences 40° for α, 40° for β, and 40m for distance. As view differences increase, there is a notable decline in the performance of all methods tested. Notably, our proposed method exhibits superior performance, and the margins of this advantage further escalate with augmented view differences.

1.2. Homography Estimation

Datasets Following [14, 17, 19], we evaluate our feature matching method in widely adopted HPatches dataset [1] for homography estimation. Hpatches contain a total of 108 sequences with significant illumination changes and large viewpoint changes. We follow the evaluation protocol of LoFTR [17], resizing the shorter size of the image to 480. AUCs at 3 different thresholds are reported.

Results In Table 1, we can see that our proposed module can improve the performance of our baseline ASpanFormer [2] on HPatches in homography estimation under all error thresholds and only worse than the dense feature matching method DKM [6] which is optimized for two view pose estimation.

Method	Pose Estimation AUC ↑		
	@3°	@5°	@10°
LoFTR [17]	52.0	68.4	80.5
Ours-LoFTR	52.9	68.8	80.5
MatchFormer [19]	51.2	68.5	81.2
Ours-MatchFormer	52.8	69.3	81.3
ASpanFormer [2]	53.0	69.8	81.8
Ours-Aspan	**55.1**	**70.9**	**82.4**

Table 2. **Evaluation on MegaDepth [10]** The best performance is highlighted by bold text. The results show that our proposed method can significantly improve the feature matching performance of Detector-free Methods [2, 17, 19]

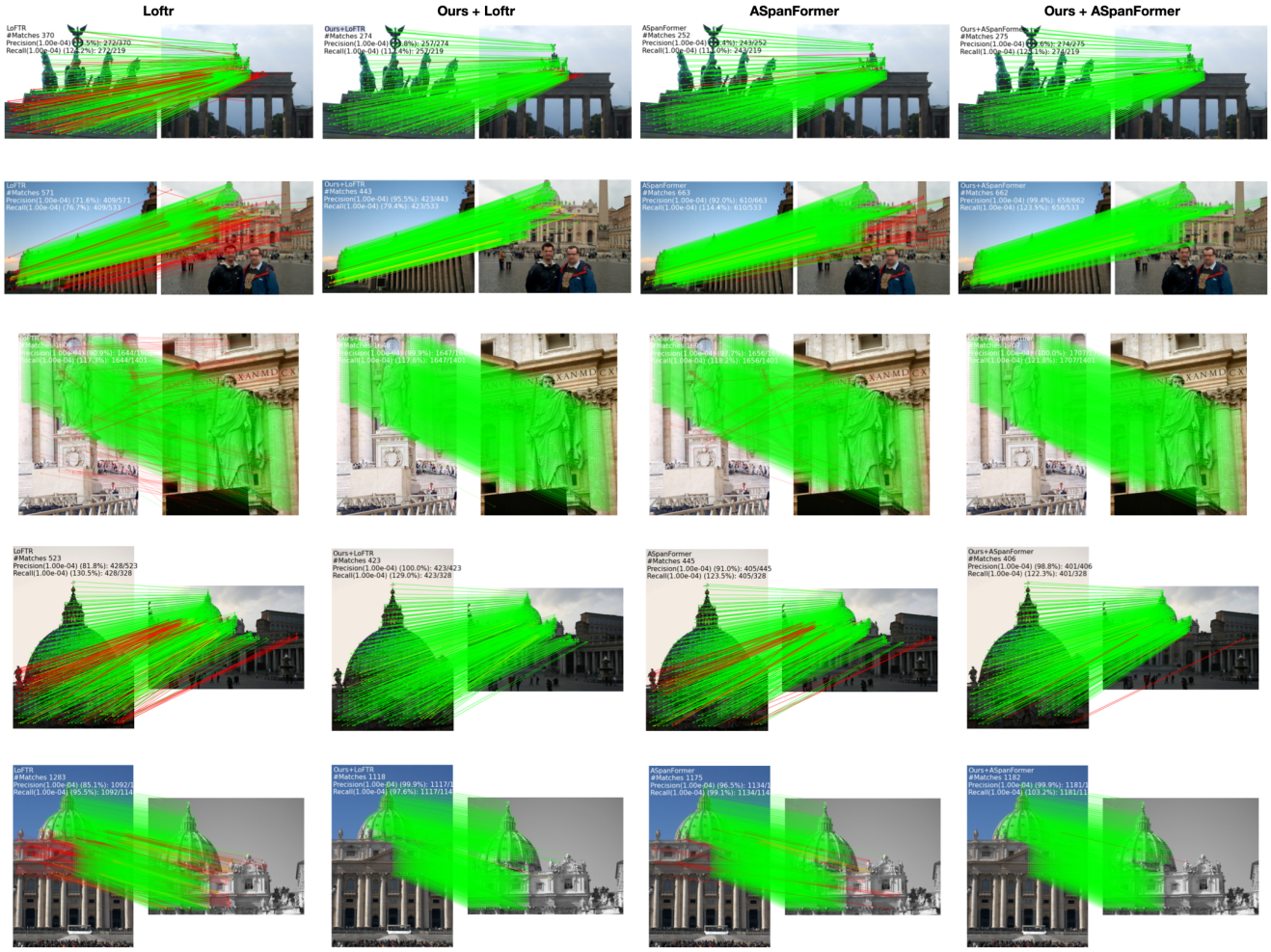

Figure 3. **Qualitative Results on MegaDepth**. Qualitative Comparison on MegaDepth [10]. The first column is LoFTR [17], the second column is Ours+LoFTR, the third is ASpanFormer and the Forth is Ours+AspanFormer [2]. The green line indicates correct matches in which the symmetric epipolar error is less than $1e^{-4}$, and the red line indicates wrong matches. By introducing the geometry constraints, one can see that the accuracy of matches is improved noticeably.

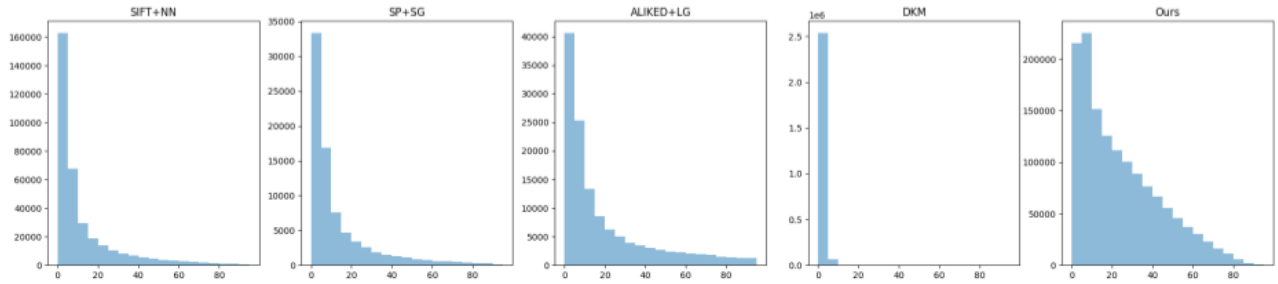

Figure 4. The distribution of track length using different feature matchers on IMC 2021 phototourism dataset [9]

1.3. Relative Pose Estimation

Dataset Following [14, 17], we use MegaDepth [10] for outdoor pose estimation. MegaDepth [10] is a large outdoor dataset containing over 1 million internet images from 196 different outdoor scenes. The camera pose is reconstructed by COLMAP [15, 16], and the depth maps are calculated from the multi-view stereo. We follow SuperGlue [14] to obtain the ground truth matches from the depth map and

camera pose. We use the same test split as [2, 17, 19]. Images are resized such that their longer dimension is equal to 840/832 for training and 832 for testing on all methods. (ASpanFormer [2] use 832 due to their need for an image resolution divisible by 16).

Metrics and Comparing Methods We follow SuperGlue [14] to report the AUC of recovered pose under threshold (5°, 10°, and 20°). The camera poses are recovered from solving RANSAC with predicted matches. We compared our method to several state-of-the-art methods SuperPoint+SuperGlue [14], LoFTR [17], MatchFormer [19], AspanFormer [2].

Results on MegaDepth As depicted in Table 2, our proposed method outperforms baseline methods. This improvement can be attributed to the early engagement of epipolar constraints, which effectively eliminates geometry inconsistencies among matches. The incorporation of geometry verification during prediction results in more accurate and robust matches, as visually demonstrated in Figure 3.

2. Track Length Distribution

Track length measures the number of consecutive frames in which a feature point in the scene can be reliably tracked, reflecting the quality of a reconstructed model. Longer track length implies more accurate reconstruction and more robust feature points. Table 3 shows that our method achieves longer track length on average. The distribution of the track length for all methods is shown in Figure 4, where the x-axis is the track length and the y-axis is the frequency. Our method demonstrates a more consistent track length distribution compared to other methods.

	SIFT [12]	SP [4]+SG [14]	ALIKED [20]+LG [11]	DKM [6]	RoMa [7]	Ours
Track Length	12.74	13.06	21.3	3.04	3.12	23.74

Table 3. Average track length on IMC 2021 [9]. Results are averaged across different scenes.

3. Failure Cases

Distortion When images are from different camera models, such as fisheye and perspective cameras, our pipeline would generate two separate models. One model primarily consists of perspective images with a few fisheye images, while the other predominantly contains fisheye images with a few perspective images.

Incorrect Initial Matches Our pipeline relies on a backbone model for fundamental matrix estimation. If the initial matches provided by the backbone model are misleading, the pipeline may produce dense correspondences that are not necessarily correct, leading to inaccurate 3D reconstructions. A common failure case occurs with repetitive patterns, particularly symmetrical buildings.

References

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

[2] 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. 1, 2, 3, 4

[3] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes, 2017. 7, 8

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

[5] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 1

[6] 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. 1, 2, 4

[7] 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. 4

[8] Khang Truong Giang, Soohwan Song, and Sungho Jo. Topicfm: Robust and interpretable topic-assisted feature matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2447–2455, 2023. 1

[9] 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. 3, 4, 6

[10] Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 1, 2, 3

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

[12] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 4

[13] Runyu Mao, Chen Bai, Yatong An, Fengqing Zhu, and Cheng Lu. 3dg-stfm: 3d geometric guided student-teacher feature matching, 2022. 1

[14] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 1, 2, 3, 4

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

[16] 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

[17] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1, 2, 3, 4

[18] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network, 2021. 1

[19] 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. 1, 2, 4

[20] 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. 4

Figure 5. **Qualitative results**. Our method is qualitatively compared with other feature matching methods on IMC 2021 Phototourism Dataset[9]. Green cameras have less than 3° absolute pose error, while red cameras have an error larger than 3°.

Figure 6. **Qualitative results**. Our method is qualitatively compared with other feature matching methods on ScanNet Dataset[3]. Green cameras have less than 3°absolute pose error, while red cameras have an error larger than 3°.

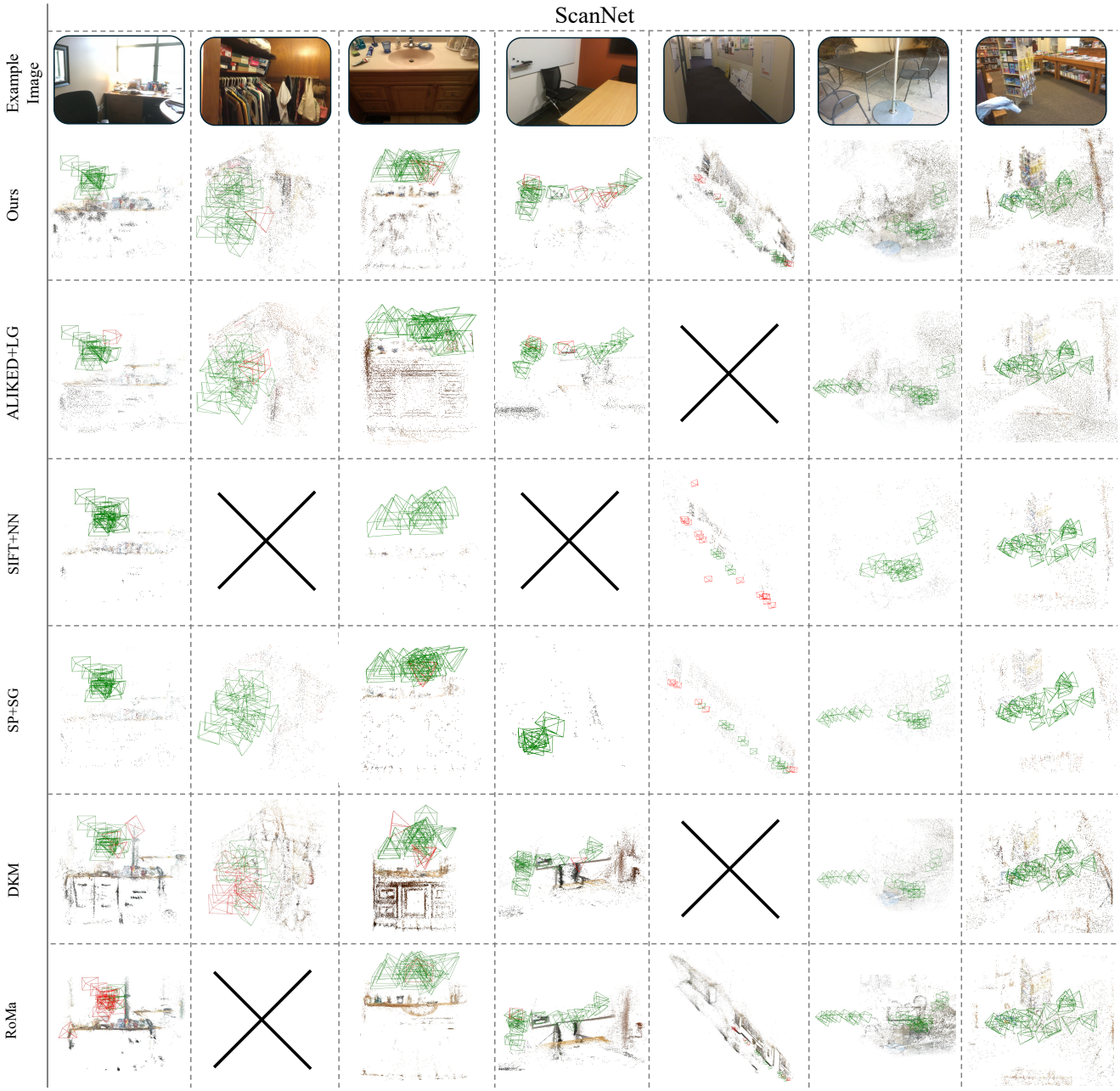

Figure 7. **Qualitative results**. Our method is qualitatively compared with other feature matching methods on ScanNet Dataset[3]. Green cameras have less than 3°absolute pose error, while red cameras have an error larger than 3°.

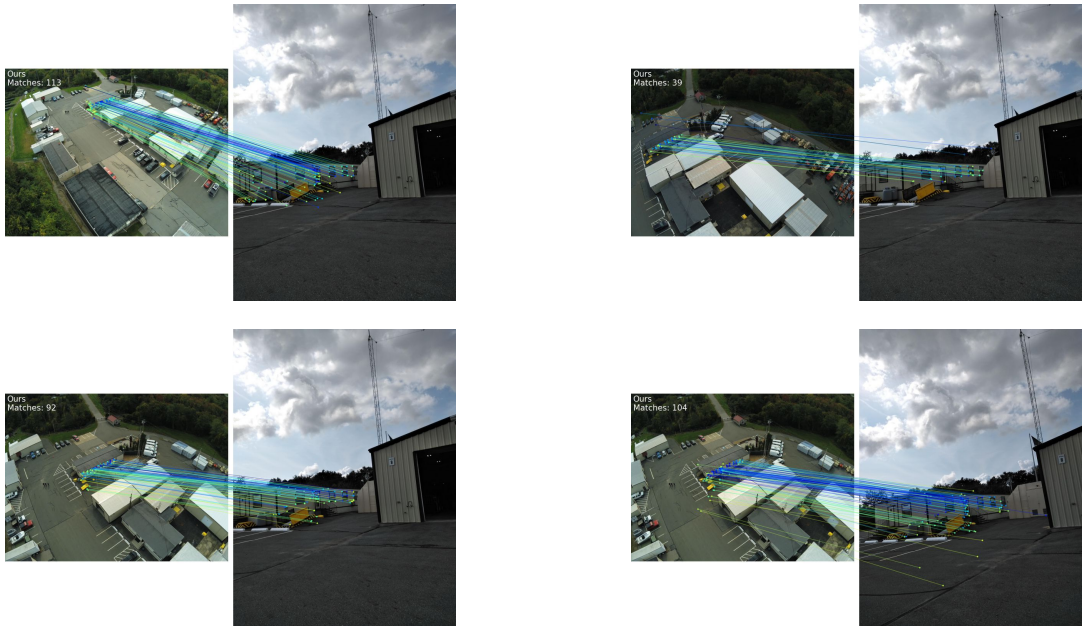

Figure 8. **Qualitative Matching Results on Collected Air-to-Ground Set 2**. Our method is able to find reliable matches for cross-view images.

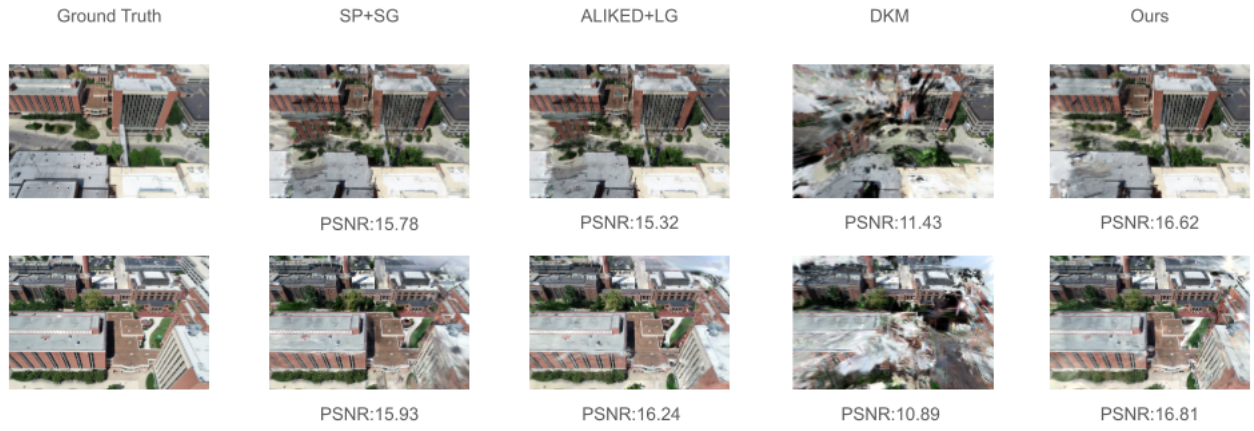

Figure 9. **Qualitative Results of 3D gaussain splatting on Collected Air-to-Ground Set 2**. All 3D Gaussians are trained for 30,000 iterations. SIFT+NN fails in registering ground images with UAV images