Metric Pose Estimation Relative to Anchor Frame for Map-free Localization

Lili Zhao[1], Zhili Liu[2], Lei Yang[1], and Meng Guo[1]

China Mobile Research Institute, Beijing, China[1]
Anonymous Institute[2]
zllmail@foxmail.com,

Abstract. Estimating the metric relative pose between two images has become a new and popular challenge. Traditional photogrammetric methods that can only estimate poses up to scale and suffer from ambiguous pose decomposition. In this paper, we propose a novel approach, which uses zero-shot metric depth estimation and an end-to-end image matching module. It can estimate the scaled 6DoF pose given two images from a monocular camera, eliminating the need for additional 3D structural statistical analysis. Extensive experimental results have demonstrated the state-of-the-art performance of our method, compared with baselines in the map-free visual relocalization challenge. Notably, our proposed method outperform the existing methods, such as MicKey, FAR, RoMa with MicKey. Furthermore, our method exhibits strong generalization capabilities.

1 Proposed Method & Implementation Details

We employed the LoFTR [15] algorithm as a front-end matching module to obtain corresponding point pairs between the map frame and the localization frame.

To achieve accurate scale-aware spatial perception and 3D representation, we use Metric3D [17] as our depth estimation model. By leveraging Metric3D for depth estimation on anchor frames, the 3D coordinates of corresponding matching points in the anchor frame can be obtained.

Based on the geometric constraints between 3D and 2D correspondences, we can establish an analytical expression for the pose estimation problem. x_1 and x_2 denote the corresponding matching points in the anchor frame and the query frame, respectively.

$$x_2 = q_2 * (K_1^{-1} * x_1) * d + t_2, \tag{1}$$

where q_2 represents the rotation between the two frames, and t_2 represents the translation. K_1 is the homogeneous camera intrinsic matrix, while d is the depth of the matching point in the anchor frame. In our implementation, PoseLib [10] is used to solve this perspective-n-points problem.

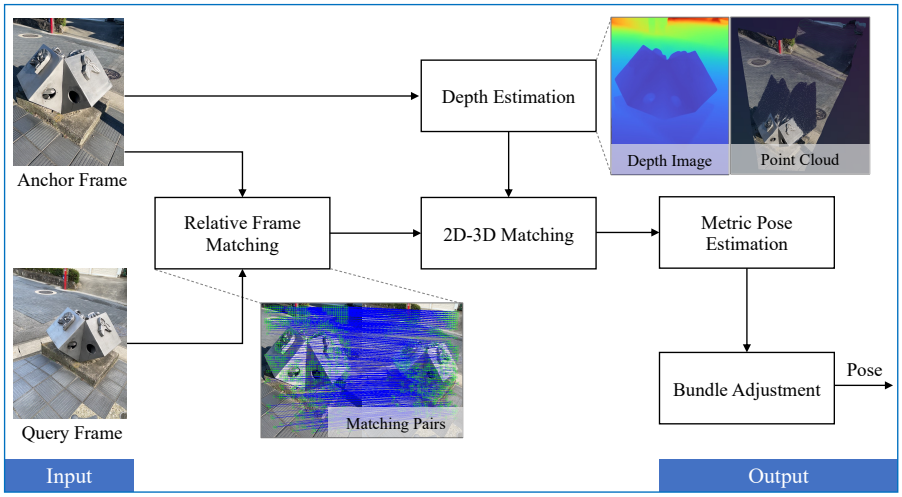

Fig. 1: The overall architecture of the proposed method.

Finally, a non-linear optimization equation for rotation q_2 and translation t_2 can be constructed based on reprojection errors. Simultaneously, the 3D coordinates of points in the anchor frame can also be optimized. The rotation quaternion q_2 and translation vector t_2 will be output as the absolute scale pose between the two frames.

$$q_2, t_2 = \text{argmin}_{d^i, q_2, t_2} \sum_i \left\| q_2 * \left(K_1^{-1} * p_1^i \right) * d^i + t_2 - p_2^i \right\|. \tag{2}$$

The optimizer is implemented based on Ceres-Solver [2] and pybind11 [9].

2 Result

To prove the effectiveness of our proposed method, we compared our method with the existing competitive works, including MicKey [3], FAR [13], and their extension by using various matching modules and 3D-2D pose calculation modules.

Specifically, there are several benchmarks in Map-free Visual Relocalization challenge 2024 [1]: 1) RoMa w/ MicKey: RoMa [7] correspondences with MicKey [3] depth maps predictions, 2) SuperGlue w/ MicKey: SuperPoint [5] + SuperGlue [14] correspondences with MicKey [3] depth maps predictions, 3) LoFTR w/ MicKey: LoFTR [15] correspondences with MicKey [3] depth maps predictions, 4) MicKey w/ DA: Mickey [3] Variant GT_Depth, 5) DPT-KITTI & ASpanFormer [4], 6) DPT-KITTI & LightGlue w/ DISK: DISK [16]+LightGlue [11] correspondences with DPT-KITTI [12] depth maps, 7) DPT-KITTI & DISK: DISK [16] correspondences with DPT-KITTI [12] depth maps, 8) DPT-KITTI & DeDoDe [6], 9) DPT-KITTI & SiLK [8].

Table 1: Quantitative comparisons between the proposed method and the existing competitive works, including MicKey [3], FAR [13], RoMa [7] w/ MicKey [3], SuperGlue [14] w/ MicKey [3], LoFTR [15] w/ MicKey [3], MicKey [3] w/ DA, DPT-KITTI [12] & ASpanFormer [4], DPT-KITTI [12] & LightGlue [11] w/ DISK [16], DPT-KITTI [12] & DISK [16], DPT-KITTI [12] & DeDoDe [6], and DPT-KITTI [12] & SiLK [8].

Method	VCRE<45°		Median Reproj.	Err <25cm, 5°		Median Error	
	AUC	Precision	Error (px)	AUC	Precision	Trans.(m)	Rot.(°)
MicKey [3]	0.558	30.1%	126.9	0.283	12.0%	1.59	26.0
FAR [13]	0.481	25.3%	137.1	0.392	17.7%	1.48	17.3
RoMa w/ MicKey	0.604	37.8%	**111.9**	0.546	31.4%	**1.18**	**15.6**
SuperGlue w/ MicKey	0.556	29.8%	139.9	0.490	23.5%	1.70	26.1
LoFTR w/ MicKey	0.550	27.2%	155.0	0.467	20.3%	1.92	33.6
MicKey w/ DA	0.548	28.0%	142.0	0.273	10.8%	1.84	30.8
DPT & ASpanFormer	0.414	20.8%	161.8	0.361	16.3%	1.90	29.2
DPT & LightGlue w/ DISK	0.355	19.5%	138.8	0.314	15.9%	1.44	18.5
DPT & DISK	0.346	15.1%	208.2	0.264	10.2%	2.59	52.0
DPT & DeDoDe	0.325	16.9%	167.4	0.265	12.5%	2.02	30.3
DPT & SiLK	0.192	9.8%	176.4	0.157	7.3%	2.21	33.8
Ours	**0.681**	**39.9%**	125.6	**0.593**	**31.9%**	1.75	31.2

References

1. Map-free Visual Relocalization: Metric Pose Relative to a Single Image, ECCV 2024 Workshop & Challenge: https://research.nianticlabs.com/mapfree-reloc-benchmark/leaderboard?t=single&f=2024/, 2024 2
2. Agarwal, S., Mierle, K., Team, T.C.S.: Ceres Solver (10 2023), https://github.com/ceres-solver/ceres-solver 2
3. Barroso-Laguna, A., Munukutla, S., Prisacariu, V., Brachmann, E.: Matching 2D Images in 3D: Metric Relative Pose from Metric Correspondences. In: CVPR (2024) 2, 3
4. Chen, H., Luo, Z., Zhou, L., Tian, Y., Zhen, M., Fang, T., Mckinnon, D., Tsin, Y., Quan, L.: Aspanformer: Detector-free image matching with adaptive span transformer. In: ECCV (2022) 2, 3
5. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-Supervised Interest Point Detection and Description. In: CVPRW (2018) 2
6. Edstedt, J., Bökman, G., Wadenbäck, M., Felsberg, M.: DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In: 3DV. IEEE (2024) 2, 3
7. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust Dense Feature Matching. In: CVPR (2024) 2, 3
8. Gleize, P., Wang, W., Feiszli, M.: Silk: Simple learned keypoints. In: ICCV (2023) 2, 3
9. Jakob, W., Rhinelander, J., Moldovan, D.: pybind11 — seamless operability between c++11 and python (2016), https://github.com/pybind/pybind11 2

10. Larsson, V., contributors: PoseLib - Minimal Solvers for Camera Pose Estimation (2020), https://github.com/vlarsson/PoseLib 1
11. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023) 2, 3
12. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021) 2, 3
13. Rockwell, C., Kulkarni, N., Jin, L., Park, J.J., Johnson, J., Fouhey, D.F.: FAR: Flexible accurate and robust 6dof relative camera pose estimation. In: CVPR (2024) 2, 3
14. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning feature matching with graph neural networks. In: CVPR (2020) 2, 3
15. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-free local feature matching with transformers. In: CVPR (2021) 1, 2, 3
16. Tyszkiewicz, M., Fua, P., Trulls, E.: DISK: Learning local features with policy gradient. NeurIPS (2020) 2, 3
17. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In: ICCV (2023) 1