Improving Map-Free Localization with Depth Supervision

Hitesh Jain[1] and Sagar Verma[1]

Granular AI, Cambridge, MA, USA
{hitesh,sagar}@granular.ai

Abstract. We propose an enhancement to the MicKey architecture, a state-of-the-art method for map-free localization, by integrating the Depth Anything network [10] to improve depth estimation. Our approach uses Depth Anything-generated ground truth depth maps during MicKey's training, resulting in more accurate depth predictions, especially in challenging regions like depth discontinuities. Experimental results show that our method maintains competitive performance across key metrics while requiring minimal supervision, thereby improving the versatility and accessibility of the MicKey framework for diverse datasets. Code: https://github.com/micropilot/dualmapfree

1 Proposed Method & Implementation Details

We propose modifying the MicKey architecture, which currently demonstrates exceptional performance on the map-free localization benchmark. MicKey utilizes a feature extractor that partitions the image into patches and computes four different types of features for each patch: 1) a 2D offset (U), 2) keypoint confidence (C), 3) depth value (Z), and 4) descriptor vector (D).

To improve depth learning during MicKey's training, we explored integrating the Depth Anything network [10], which is known for its impressive zero-shot capabilities and reduced generalization error, into MicKey. This network has been trained on a large-scale dataset of approximately 62 million samples and demonstrates lower generalization error. We use this network to generate ground truth depth images from the map-free dataset.

By incorporating Depth Anything-generated ground truth depth maps into MicKey's training, we anticipate several key advantages:

- The supervision provided by these high-quality depth maps can guide MicKey in learning more accurate depth estimations, particularly in regions where traditional feature detectors might struggle, such as areas with depth discontinuities or corners. This additional supervision helps mitigate one of the primary issues with conventional depth estimators, which often fail in such challenging areas due to their independent operation from feature detectors.
- Secondly, the use of Depth Anything reduces the reliance on extensive ground truth depth data, which can be difficult to obtain in certain domains. This aligns with MicKey's design philosophy of requiring minimal supervision,

making it even more versatile and accessible for training across various datasets without the need for comprehensive depth annotations.

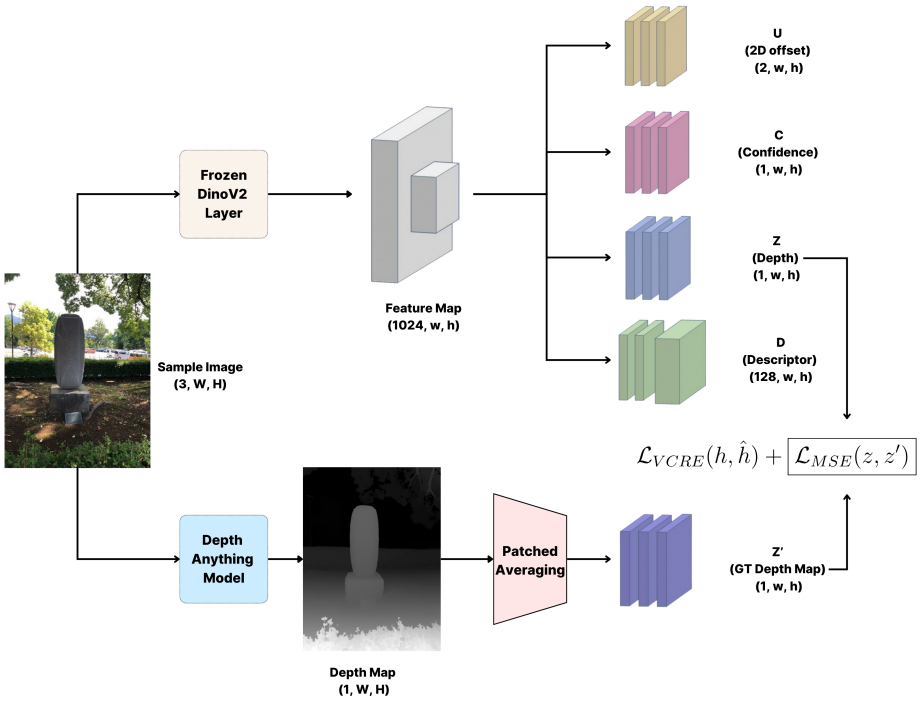

Fig. 1: Depth Anything network integrated with MicKey training.

Figure 1 illustrates the integration of Depth Anything ground truth during MicKey's training by computing the mean square loss between the ground truth (z') and predicted depth features (z) and adding this loss to the VCRE loss. To ensure compatibility between the ground truth depth map and the predicted depth map's shape, we partition the ground truth depth map into patches and compute the average depth for each patch. This new term ensures that the network's depth predictions align with the high-quality depth estimates provided by the Depth Anything model. By doing so, we can guide the network to learn more accurate and robust depth representations, leveraging the strengths of the ground truth depth maps without altering the core architecture of MicKey.

2 Result

In our evaluation, we compared the modified MicKey architecture with several leading methods in the map-free localization benchmark, as detailed in Table 1.

Method	VCRE<45°		Median Reproj. Err	<25cm, 5°		Median Error	
	AUC	Precision	Error (px)	AUC	Precision	Trans. (m)	Rot. (°)
MicKey [1]	0.558	30.1%	126.9	0.283	12.0%	1.59	26.0
FAR [7]	0.481	25.3%	137.1	0.392	17.7%	1.48	17.3
RoMa [4] w/ MicKey	0.604	37.8%	111.9	0.546	31.4%	1.18	15.6
SuperGlue [8] w/ MicKey	0.556	29.8%	139.9	0.490	23.5%	1.70	26.1
LoFTR [9] w/ MicKey	0.550	27.2%	155.0	0.467	20.3%	1.92	33.6
DPT & ASpanFormer [2]	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 [6]	0.346	15.1%	208.2	0.264	10.2%	2.59	52.0
DPT & DeDoDe [3]	0.325	16.9%	167.4	0.265	12.5%	2.02	30.3
DPT & SiLK [5]	0.192	9.8%	176.4	0.157	7.3%	2.21	33.8
MicKey (GT-Depth)	0.548	28.0%	141.96	0.273	10.765%	1.842	30.78

Table 1: Comparison of different methods from map-free localization single frame leaderboard.

Specifically, we compared our approach against the original MicKey [1], FAR [7], RoMa [4] combined with MicKey, SuperGlue [8] integrated with MicKey, and LoFTR [9] paired with MicKey. Additionally, we assessed the performance of combinations of DPT with ASpanFormer [2], LightGlue integrated with DISK, DISK [6] alone, DeDoDe [3], and SiLK [5]. The results demonstrate the competitive performance of our approach, particularly when using Depth Anything-generated ground truth depth maps, highlighting its effectiveness in improving depth estimation within the MicKey framework.

References

1. Arnold, E., Wynn, J., Vicente, S., Garcia-Hernando, G., Monszpart, Á., Prisacariu, V.A., Turmukhambetov, D., Brachmann, E.: Map-free Visual Relocalization: Metric Pose Relative to a Single Image. In: ECCV (2022)
2. 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)
3. Edstedt, J., Bökman, G., Wadenbäck, M., Felsberg, M.: DeDoDe: Detect, Don't Describe — Describe, Don't Detect for Local Feature Matching. In: 3DV (2024)
4. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust Dense Feature Matching (2024)
5. Gleize, P., Wang, W., Feiszli, M.: SiLK: Simple Learned Keypoints (2023)
6. Michał J. Tyszkiewicz, Pascal Fua, E.T.: DISK correspondences with DPT-KITTI depth maps (2024)
7. Rockwell, C., Kulkarni, N., Jin, L., Park, J.J., Johnson, J., Fouhey, D.F.: FAR: Flexible, Accurate and Robust 6DoF Relative Camera Pose Estimation (2024)
8. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching with Graph Neural Networks. In: CVPR (2020)
9. Sun, J., Shen, Z., Wang, Y., Bao, H., Zhou, X.: LoFTR: Detector-Free Local Feature Matching with Transformers (2021)
10. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. In: CVPR (2024)