DUSt3R-SLAM

Overview

Simultaneous Localization and Mapping (SLAM) is the task of estimating the trajectory of a camera and mapping the environment, given a stream of camera images. SLAM is a key component in many computer vision systems, with applications in autonomous driving/navigation, augmented reality and 3D mapping.

Early SLAM methods, such as in ORB-SLAM [2], are based on tracking hand-crafted image features and optimizing for geometric consistency. However, purely hand-crafted approaches lead to complicated pipelines that require expertise to design and operate/debug. As an alternative, there has in recent years been much interest in designing learning-based SLAM methods, such as in DROID-SLAM [4]. Yet, learning-based methods for SLAM only work well in scenarios that are similar to the training data, which has so far limited their applicability.

In this project, we will evaluate whether recent developments in large-scale vision model training can enable learning-based SLAM, without the need for additional data collection or fine-tuning. In particular, we will focus on DUSt3R [5]. At its core, DUSt3R uses *scene coordinate regression* [3], meaning that they directly predict the 3D scene coordinates of each image pixel. Other geometric properties, such as relative poses, intrinsics, depth, optical flow, etc. can then be derived from the scene coordinates using optimization. Importantly, they make use of pre-training with two-view image completion [7, 6] to obtain a highly adaptable base network. This network is then fine-tuned with depth and pose supervision, achieving impressive performance across a wide range of geometric vision tasks.

Goals

The goal of this project is to design a SLAM pipeline based on DUSt3R. The SLAM problem presents unique challenges which were not tackled in the original paper. For instance, consecutive

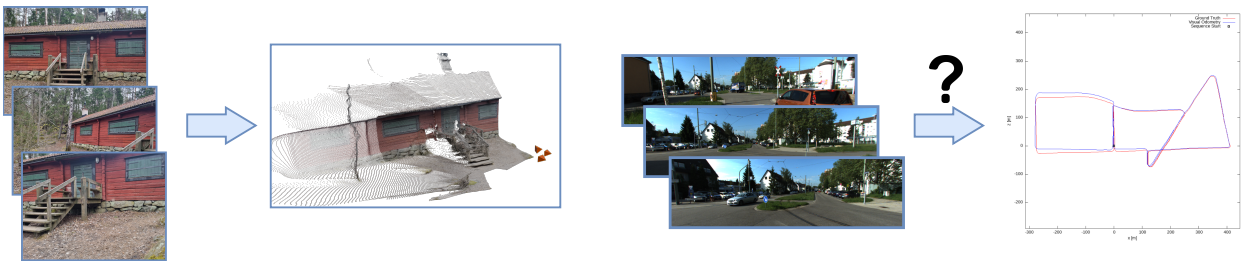

Figure 1: DUSt3R performs well out of the box on SfM scenarios with sparse camera views (left) we will investigate whether the model can be adapted to perform visual odometry (right).

frames have a high level of overlap making it infeasible to run inference on every pair of views, and moving objects may complicate the reconstruction task.

Tasks

1 Getting familiar with DUSt3R and related literature (~ 50h).

2 Setting up dataset and baseline (for instance KITTI [1] and DROID-SLAM [4]) (~ 50h).

2 Investigate how to adapt DUSt3R for SLAM, with the goal of balancing execution time and tracking performance (~ 100h).

3 Conducting experiments and evaluating trajectory metrics (ATE, RPE, etc.) (~ 50h).

4 Writing a project report and prepare presentations (~ 50h).

Contact

Linus Härenstam-Nielsen, Weirong Chen
Email: `linus.nielsen@tum.de`, `weirong.chen@tum.de`

References

[1] Andreas Geiger, Philip Lenz, and Raquel Urtasun. "Are we ready for Autonomous Driving? The KITTI Vision Benchmark Suite". In: *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2012.

[2] Raúl Mur-Artal and Juan D. Tardós. "ORB-SLAM2: an Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras". In: *IEEE Transactions on Robotics* 33.5 (2017), pp. 1255–1262. DOI: 10.1109/TRO.2017.2705103.

[3] Jamie Shotton et al. "Scene coordinate regression forests for camera relocalization in RGB-D images". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2013, pp. 2930–2937.

[4] Zachary Teed and Jia Deng. "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras". In: *Advances in neural information processing systems* 34 (2021), pp. 16558–16569.

[5] Wang, Shuzhe and Leroy, Vincent and Cabon, Yohann and Chidlovskii, Boris and Revaud Jerome. *DUSt3R: Geometric 3D Vision Made Easy*. 2023.

[6] Philippe Weinzaepfel et al. "CroCo v2: Improved Cross-view Completion Pre-training for Stereo Matching and Optical Flow". In: *ICCV*. 2023.

[7] Weinzaepfel, Philippe and Leroy, Vincent and Lucas, Thomas and Brégier, Romain and Cabon, Yohann and Arora, Vaibhav and Antsfeld, Leonid and Chidlovskii, Boris and Csurka, Gabriela and Revaud Jérôme. "CroCo: Self-Supervised Pre-training for 3D Vision Tasks by Cross-View Completion". In: *NeurIPS*. 2022.