Uncalibrated Monocular VSLAM for Smartphone Video Benchmarking

Florian Beck Valentin Bumeder Lukas Röß

Christopher Kirschner Jan Duchscherer
j.duchscherer@hm.edu

Department of Computer Science & Mathematics
Munich University of Applied Sciences
Munich, Germany

***Abstract*—This report documents the project scaffold, evaluation protocol, and benchmark plan for uncalibrated monocular VSLAM on smartphone video. The focus is on comparing modern methods, handling unknown intrinsics, and evaluating both trajectory quality and dense reconstruction quality against public and custom datasets.**

***Index Terms*—VSLAM, monocular SLAM, dense reconstruction, trajectory evaluation, ADVIO**

I. Introduction

This project targets an off-device monocular VSLAM benchmark for smartphone video with unknown intrinsics. The main goal is to compare modern dense monocular systems, recover camera trajectories with limited calibration assumptions, and evaluate dense reconstructions under a common protocol.

The benchmark centers on two output surfaces: trajectory quality and dense 3D point cloud quality. In addition to method accuracy, the project tracks runtime and memory requirements because the final recommendation must balance precision, robustness, and practical execution cost.

II. Related Work

The initial benchmark focuses on recent monocular dense methods such as ViSTA-SLAM [1] and MASt3R-SLAM [2]. These systems provide concrete starting points for recovering camera motion and dense geometry from monocular inputs while reducing the amount of hand-designed calibration logic required from the project.

The project also relies on established evaluation and reconstruction tooling. Trajectories are expected to be compared with evo [3], while reference reconstructions can be built with COLMAP [4], [5] and inspected with Open3D [6].

III. Challenge and Scope

The challenge requires an off-device pipeline that accepts raw smartphone video, handles unknown intrinsics, and outputs both a high-precision trajectory and a dense 3D point cloud. The evaluation must compare at least two state-of-the-art methods, include ARCore as a baseline where applicable, and cover both public and custom datasets.

This repository therefore scopes the project into four major surfaces: method integration, custom data capture, trajectory evaluation, and dense reconstruction evaluation. Heavy external tools are kept outside the base Python environment and are treated as documented integrations rather than vendored code.

IV. Candidate Methods

Candidate methods are integrated behind a shared benchmark workflow so they can be compared under consistent inputs and output formats. The initial comparison focuses on ViSTA-SLAM [1] and MASt3R-SLAM [2] because both are directly referenced by the challenge brief.

Each method integration should define input expectations, output artifact locations, and any required pre-processing. The method wrappers should stay thin and should document unsupported cases explicitly instead of hiding them behind silent fallbacks.

V. Datasets

Trajectory quality is benchmarked on the ADVIO dataset [7] and on a custom smartphone capture dataset recorded for this project. The custom capture workflow is expected to store raw video and baseline ARCore logs with enough metadata to reproduce alignment and evaluation.

Dense reconstruction quality is assessed on self-recorded data because the challenge explicitly asks for a comparison against ARCore mapping results on a custom test dataset. Reference reconstructions for these captures can be generated with tools such as COLMAP [4], [5].

VI. Metrics

Trajectory evaluation should quantify both global and local behavior. Core examples are alignment error, pose drift, and sequence-level consistency. The exact metric suite should remain stable across methods and be reported in a reproducible scriptable form.

Dense reconstruction evaluation should quantify geometric fidelity, completeness, and failure modes such as missing structure or noisy surfaces. Open3D [6] and comparable point-cloud tooling can provide a practical baseline for alignment and metric computation.

VII. Experiments

The experimental plan is split into public-dataset benchmarking and custom-dataset benchmarking. For each method, the benchmark should capture run configuration, produced artifacts, runtime cost, and all trajectory or reconstruction metrics required by the final recommendation.

The initial scaffold in this repository intentionally prioritizes reproducibility over breadth. The first milestone is a functional installable package, a stable evaluation layout, and documentation that lets contributors add method-specific experiments without redesigning the repository.

VIII. Discussion

The main risks are method-specific environment complexity, custom data capture quality, and scale consistency when comparing dense outputs against ARCore or reference reconstructions. These risks should be documented early because they directly affect benchmark fairness and reproducibility.

The current repository scaffold addresses these risks by separating lightweight project code from heavy external tools, keeping weekly reporting templates in the repo, and defining a stable work package split before deeper implementation work begins.

IX. Conclusion

This report scaffold establishes the structure needed for a reproducible monocular VSLAM benchmark project. The next implementation phases are method integration, custom data capture, and the first trajectory and reconstruction comparisons on public and custom datasets.

References

[1] G. Zhang, S. Qian, X. Wang, and D. Cremers, "ViSTA-SLAM: Visual SLAM with Symmetric Two-view Association." [Online]. Available: https://arxiv.org/abs/2509.01584

[2] R. Murai, E. Dexheimer, and A. J. Davison, "MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors." [Online]. Available: https://arxiv.org/abs/2412.12392

[3] M. Grupp, "evo: Python package for the evaluation of odometry and SLAM.." 2017.

[4] J. L. Schoenberger and J.-M. Frahm, "Structure-from-Motion Revisited." [Online]. Available: https://openaccess.thecvf.com/content_cvpr_2016/html/Schonberger_Structure-From-Motion_Revisited_CVPR_2016_paper.html

[5] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm, "Pixelwise View Selection for Unstructured Multi-View Stereo," in *European Conference on Computer Vision (ECCV)*, 2016. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-46487-9_30

[6] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3D: A Modern Library for 3D Data Processing." [Online]. Available: https://arxiv.org/abs/1801.09847

[7] S. Cortés, A. Solin, E. Rahtu, and J. Kannala, "ADVIO: An authentic dataset for visual-inertial odometry." [Online]. Available: https://arxiv.org/abs/1807.09828