2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Have We Mastered Scale in Deep Monocular Visual SLAM? The ScaleMaster Dataset and Benchmark

Hyoseok Ju[1], Bokeon Suh[1], and Giseop Kim[1*]

***Abstract*— Recent advances in deep monocular visual Simultaneous Localization and Mapping (SLAM) have achieved impressive accuracy and dense reconstruction capabilities, yet their robustness to scale inconsistency in large-scale indoor environments remains largely unexplored. Existing benchmarks are limited to room-scale or structurally simple settings, leaving critical issues of intra-session scale drift and inter-session scale ambiguity insufficiently addressed. To fill this gap, we introduce the *ScaleMaster Dataset*, the first benchmark explicitly designed to evaluate scale consistency under challenging scenarios such as multi-floor structures, long trajectories, repetitive views, and low-texture regions. We systematically analyze the vulnerability of state-of-the-art deep monocular visual SLAM systems to scale inconsistency, providing both quantitative and qualitative evaluations. Crucially, our analysis extends beyond traditional trajectory metrics to include a direct map-to-map quality assessment using metrics like Chamfer distance against high-fidelity 3D ground truth. Our results reveal that while recent deep monocular visual SLAM systems demonstrate strong performance on existing benchmarks, they suffer from severe scale-related failures in realistic, large-scale indoor environments. By releasing the ScaleMaster dataset and baseline results, we aim to establish a foundation for future research toward developing scale-consistent and reliable visual SLAM systems.**

I. Introduction

Visual SLAM [1, 2, 3] is experiencing a paradigm shift toward dense mapping methods that generate accurate 3D maps in real time. Beyond traditional feature-based pipelines [2], recent approaches leverage feed-forward dense pointmap estimators [4] as a front-end to directly reconstruct scenes while estimating camera poses. These methods achieve both visually detailed reconstructions and high-precision trajectory estimation. A representative example is MASt3R-SLAM [5], which employs the direct pointmap estimator [4, 6] to generate detailed 3D maps and has demonstrated state-of-the-art performance on standard indoor benchmarks.

However, the outstanding performance of these modern dense visual SLAM systems has been predominantly validated on benchmarks featuring either room-scale environments such as TUM-RGBD [7], or structurally simple spaces like those in EuRoC [8]. Consequently, their robustness in large-scale, complex indoor environments (e.g., multi-floor structures, vertical motions, long loops) remains largely unverified. Specifically, the critical issues of **scale ambiguity** and **scale inconsistency** that arise over long trajectories have not been sufficiently addressed as shown in Fig. 1. Furthermore, evaluation has traditionally focused on trajectory error (e.g., Absolute Trajectory Error (ATE)), a metric that we will demonstrate can be insufficient and even misleading. Even when poses look accurate after Sim(3) pose-graph optimization, the dense map may still be inconsistent without a joint Sim(3) bundle adjustment, as observed in systems like MASt3R-SLAM. Thus, a direct map-to-map comparison provides a more faithful way to reveal hidden inconsistencies such as warping and geometric distortion that trajectory-only metrics tend to overlook.

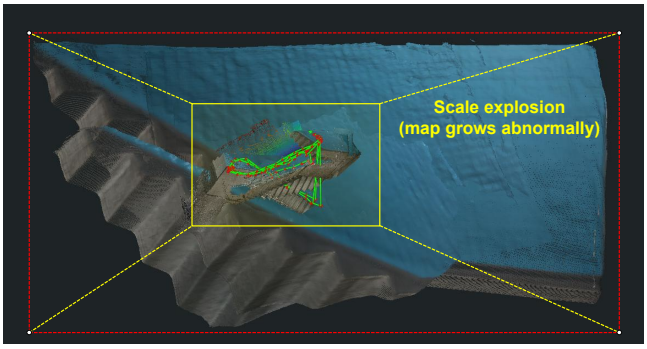

Fig. 1. Representative example of scale inconsistency. During Sim(3) pose-graph optimization, the SLAM trajectory experiences a sudden scale explosion (highlighted), resulting in an abnormally enlarged reconstruction that diverges from the previously estimated map.

This paper aims to provide an in-depth analysis of the scale inconsistency problem that arises in large-scale, complex indoor environments—involving trajectories of several hundred meters and multi-floor structures—which have not been deeply addressed by existing benchmarks (e.g., [7, 8, 9]). To this end, we introduce the **ScaleMaster** dataset[1], the first benchmark intentionally designed to target scale inconsistency by incorporating challenging scenarios where this problem becomes prominent. The ScaleMaster dataset provides a systematic empirical environment to evaluate and analyze scale inconsistency in depth, thereby overcoming the limitations of existing benchmarks.

In this paper, we make the following key contributions:

- **Systematic Analysis of Scale Inconsistencies:** We analyze how state-of-the-art monocular deep visual SLAM systems encounter **scale inconsistency** issues (as taxonomically analyzed in Fig. 2) in large-scale and complex indoor environments, highlighting vulnerabilities that are not fully revealed by existing benchmarks.
- **ScaleMaster Dataset and Baseline Evaluation:** We introduce the ScaleMaster Dataset, designed to examine

[1]H. Ju, B. Suh, and G. Kim are with the Department of Robotics and Mechatronics Engineering, DGIST, Daegu, Republic of Korea [hyoseok.ju, bokeon.suh, gsk]@dgist.ac.kr

[1]https://scalemaster-dataset.github.io/

scale consistency in challenging settings such as multi-floor structures, long trajectories, repetitive patterns, and low-texture areas. Using this dataset, we conduct baseline evaluations of representative deep-learning-based SLAM systems (DROID-SLAM [10], MASt3R-SLAM [5], VGGT-SLAM [11]), and reveal concrete cases of intra-session scale drift and inter-session scale ambiguity through both quantitative and qualitative analysis.
- **Complementary Use of Map Quality Metrics:** To address limitations of trajectory-only evaluation, we incorporate map-to-map metrics (Chamfer distance and Drop Rate), showing how they complement ATE by revealing distortions and scale collapse that remain hidden otherwise.

II. Related Works

A. Deep Learning-based Visual SLAM

Following traditional modular visual SLAM systems [12, 13], research integrating deep learning into the SLAM pipeline has become mainstream [10, 14, 15]. A prominent example of this trend is DROID-SLAM [10], which set a new standard by achieving high accuracy through a fully differentiable architecture. More recently, a subsequent wave of feed-forward dense SLAM systems has emerged, which leverage pre-trained two-view reconstruction foundation models. MASt3R-SLAM [5] and VGGT-SLAM [11], which attempt to address the ambiguity of the scale, are notable examples of this approach. The aforementioned works aim to maintain the robustness of modern learning-based SLAM systems in real-world environments. However, their robustness to scale drift, particularly in large-scale and complex indoor environments, remains largely unexplored.

B. Benchmark Datasets for Visual SLAM

The progress and evaluation of visual SLAM algorithms have been critically dependent on the role of standard benchmark datasets that provide precise ground truth. The rigorous performance assessment in SLAM research is primarily based on three key datasets. TUM-RGBD dataset [7] established the standard for evaluating ATE, while the EuRoC MAV dataset [8] has served as a crucial benchmark for the robustness of visual-inertial systems. Lastly, 7-Scenes [9] is specialized for evaluating re-localization performance. These are collectively regarded as the de facto standards for performance validation in the visual SLAM community. Despite their utility, standard benchmarks are mostly room-scale and do not support systematic evaluation of long-term **intra-session** scale drift or **inter-session** scale consistency. ARKitScenes [16]—though a comparatively recent dataset—also falls short, as it does not encompass complex indoor environments with long trajectories. This lack of a scale-focused benchmark has limited systematic analysis of scale failure modes; this is a gap we address in this work.

C. Evaluation Metrics for Visual SLAM

A commonly reported trajectory metric is the ATE [17], computed after Sim(3) alignment of the estimated poses to ground truth; it summarizes global trajectory agreement as a Root Mean Square Error (RMSE) over pose errors. While invaluable, ATE only assesses the accuracy of the localization component (L in SLAM). It provides no direct information about the quality of the mapping component (M), which is the primary output of dense SLAM systems [18]. To directly assess 3D structure quality, we adopt the Chamfer distance [5] (and Drop Rate will be defined in Sec. V-A) between the reconstructed map and a high-fidelity reference point cloud. These map-oriented metrics provide geometry-focused signals that complement trajectory-only evaluation and have been used in prior dense visual SLAM work. Although many SLAM datasets exist (e.g., [19, 20]), the benchmarks that deep, dense monocular visual SLAM systems have been primarily evaluated on are those highlighted in Sec. II-B (e.g., [7, 8, 9]). These room-scale or structurally simple indoor datasets do not capture challenging large-volume environments (e.g., multi-floor lobbies with tens of meters of ceiling height), where long-range intra- and inter-session scale inconsistency becomes apparent. Because dense visual SLAM yields dense maps as well as poses, map-oriented evaluation is crucial: ATE can look correct while the map suffers scale distortions. With Chamfer- and Drop Rate–based evaluation on ScaleMaster, we provide a protocol that exposes such failures under conditions existing benchmark datasets cannot reveal.

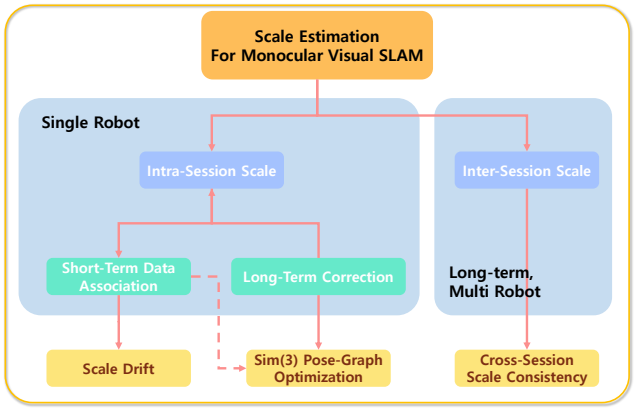

Fig. 2. A diagram illustrating the components of the scale estimation problem.

III. Problem Definition

The scale consistency problem in monocular visual SLAM can be categorized into two primary challenges as in Fig. 2: Intra-session inconsistency, which occurs within a single run, and inter-session ambiguity, which arises between multiple runs (e.g., long-term map management or multi-robot collaborative mapping).

A. Intra-Session Scale Inconsistency

Intra-session scale inconsistency refers to the scale errors that accumulate during a single SLAM session. This problem arises primarily from the interplay of the following two factors:

TABLE I. Comparison of public benchmark datasets with ours.

Dataset	Sequences	Avg. Length	Image Resolution	Multi-floor & Elevation	Pure Rotation	Complex Indoor	Scale Study Feasibility
EuRoC	11	81.2 m	752 x 480	△	X	X	X
TUM-RGBD	∼39	12.2 m	640 x 480	X	X	X	X
7-Scenes	7	64.3 m	640 x 480	X	X	X	X
ARKitScenes	>5,000	<100 m	1920 x 1440	X	O	X	X
Ours (ScaleMaster dataset)	25	152.2 m	1920 x 1440	O	O	O	O

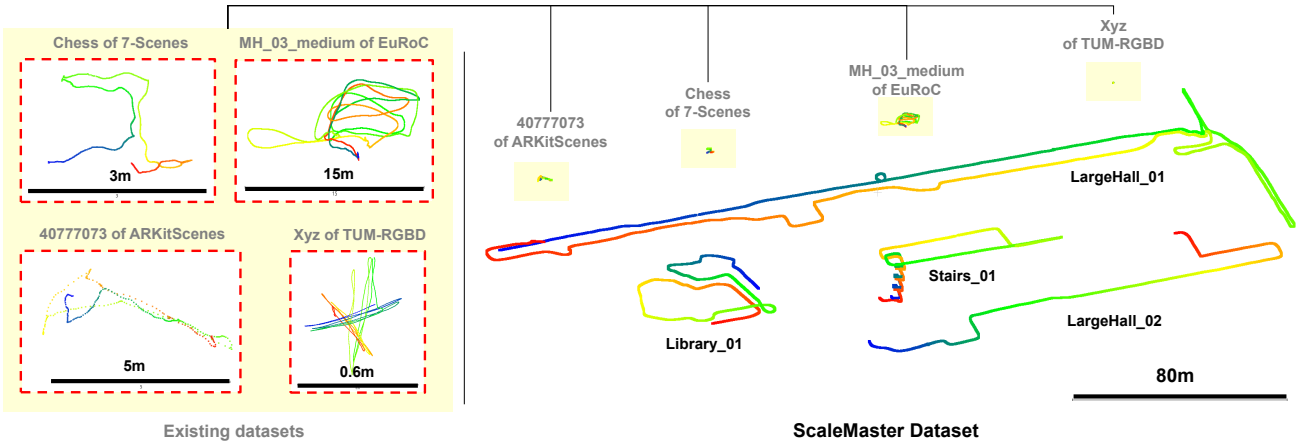

Fig. 3. A visual comparison of trajectory scales between our proposed ScaleMaster dataset and existing standard benchmarks. This distinct contrast visually demonstrates why existing room-scale benchmarks are insufficient for evaluating long-term scale consistency failures.

- **Short-Term Data Association & Scale Drift:** Short-term data association establishes odometry edges by relating consecutive image frames. Due to the inherent unobservability of metric scale in monocular visual SLAM, relative scale drift inevitably arises between them [21].
- **Long-Term Correction & Limitations of Sim(3) pose-graph optimization (PGO):** Owing to the Sim(3) gauge in monocular visual SLAM, scale cannot be fixed by SE(3) optimization [22]; while Sim(3) pose-graph optimization is standard, its small, Gaussian-error assumption makes it brittle to large, discontinuous loop-closure errors, leaving residual scale inconsistency.

B. Inter-Session Scale Ambiguity

Inter-session scale ambiguity arises when attempting to merge multiple maps generated at different times or when trying to re-localize within a map from a previous session [23]. If each session already contains its own unique and uncorrected scale drift, determining the relative scale between these maps becomes an extremely challenging problem. This ultimately leads to the failure of **Cross-Session Scale Consistency**, making it impossible to integrate multiple maps into a single, globally consistent representation.

In conclusion, this research critically addresses the challenges of intra-session scale drift and inter-session scale ambiguity, which are rooted in the fundamental scale ambiguity of monocular visual SLAM systems.

IV. The ScaleMaster Dataset

To address the inadequacy of existing benchmarks for evaluating robustness for scale consistency in monocular visual SLAM, we introduce and publicly release a new challenging dataset, **ScaleMaster Dataset**.

A. Hardware

For data acquisition, we built a custom handheld rig (see Fig. 4) equipped with an iPhone 14 Pro, a Livox HAP LiDAR sensor, and an Orbbec Gemini 335L camera. This setup ensures precise temporal synchronization between visual frames and LiDAR scans, leading to improved data quality and reliability. The iPhone provides ARKit odometry, while the LiDAR sensor captures dense 3D geometry. This hardware configuration enables the collection of long, large-scale trajectories with consistent temporal alignment.

B. Baseline Pose Generation

The baseline camera trajectories were obtained using Apple's ARKit framework, which is known to provide centimeter-level accuracy in large indoor spaces. Since our primary goal is to highlight the fundamental scale failures, ARKit trajectories that may potentially exhibit long-term drift are sufficient, where residual centimeter-scale errors are tolerable. To build reference maps, we projected high-resolution LiDAR point clouds onto the ARKit trajectories. As shown in Fig. 4, this process yields dense maps that are qualitatively well aligned and preserve fine structural

TABLE II. Summary of collected sequences and their characteristics.

Sequence Name	Frames	Path Length (m)	Duration (s)	Environment	Tags
Basement_01	2036	29.11	67	Basement area followed by a staircase ascent.	Indoor, Short trajectory, 3D Map
HotelRoom_01	4217	29.07	141	Interior traversal of a hotel room.	Indoor, Repetitive view, Short trajectory
Lab_01	2167	25.85	72	In-place rotations inside a lab room.	Indoor, Repetitive view, Short trajectory
LargeHall_01	22830	884.12	761	Full loop covering the entire LargeHall.	Indoor, Very long trajectory
LargeHall_02	6576	241.89	219	Evening traversal of a large open hall, low-light.	Indoor, Long trajectory, 3D Map
LargeHall_03	2764	109.89	92	Short loop inside the LargeHall at night.	Indoor, Low-texture risk, Medium trajectory
LargeHall_04	4331	179.69	144	Loop around the E1 section of LargeHall.	Indoor, Medium trajectory
LargeHall_05	1912	54.21	63	Traversal under low-light conditions.	Indoor, Short trajectory, 3D Map
Library_01	6515	254.98	217	Multi-floor descent from 5F to 3F with loops per floor.	Indoor, Long trajectory, Repetitive view, 3D Map
Library_02	5001	163.58	166	Single-floor loop on 4F with repetitive bookshelves.	Indoor, Medium trajectory, 3D Map
Library_03	3136	105.81	105	Loop on the 3rd floor of the library.	Indoor, Medium trajectory
Library_04	5450	146.22	182	Walking paths between library bookshelves.	Indoor, Repetitive view, Medium trajectory
Library_05	2540	78.24	85	Large open central atrium (depth sensor limitation).	Indoor, Low-texture risk, Short trajectory
Library_06	2026	13.27	67	360-degree in-place rotation at the library center.	Indoor, Short trajectory, 3D Map
Library_07	1580	5.23	52	Static panoramic survey from the 1st floor.	Indoor, Short trajectory, 3D Map
Library_08	2241	3.51	75	Short traversal around the open central viewpoint.	Indoor, Low-texture risk, Short trajectory
Library_09	2303	20.65	77	In-place rotation in front of a 3F glass room.	Indoor, Repetitive view, Short trajectory
Lobby_01	2893	104.83	96	Traversal inside a lobby.	Indoor, Medium trajectory
Lounge_01	6823	199.09	228	Loop trajectory inside a lounge area.	Indoor, Medium trajectory
Office_01	6009	154.50	200	Traversal inside a repetitive office view.	Indoor, Repetitive view, Medium trajectory
Parking_01	8218	323.01	274	Full loop inside the underground parking lot (B2).	Underground, Long trajectory
Parking_02	2270	88.06	76	Loop in an indoor parking area.	Indoor, Short trajectory
Stairs_01	9394	298.42	313	Ascending from 2F to 6F, then looping on 6F.	Indoor, Vertical motion, Repetitive view, Long trajectory
Stairs_02	4143	122.23	139	Repeated ascending and descending of stairs.	Indoor, Vertical motion, Repetitive view, Medium trajectory
Station_01	1715	170.84	229	Escalator traversal inside a train station.	Indoor/Outdoor, Vertical motion, Repetitive view, Medium trajectory

Trajectory length tags : 0–100 m = Short; 100–200 m = Medium; $>$ 200 m = Long.

details, making them a reliable reference for analyzing scale consistency and geometric distortion.

C. Dataset Summary and Comparison

The ScaleMaster dataset comprises 25 sequences that specifically address scale-related challenges rarely captured in previous benchmarks. Of these, 18 sequences, acquired solely with an iPhone 14 Pro, provide trajectory-only ground truth, while the remaining 7 sequences, collected using the rig shown in Fig. 4, also include dense LiDAR-based 3D maps. The dataset covers multi-floor motion, long loops with repetitive structures, large open spaces, and low-texture conditions.

- **Table I** compares ScaleMaster against standard benchmarks, highlighting its unique support for scale evaluation in complex indoor environments.
- **Table II** presents representative sequences with their scale, duration, and environmental tags, illustrating the dataset's diversity—from nearly 900 m building-scale loops to vertical stair traversals.

Each sequence is named by its location followed by an index number (e.g., Library_01, Parking_02). The environment description and trajectory characteristics for each sequence are summarized as tags in Table II, covering attributes such as trajectory type, vertical motion, repetitive views, and low-texture conditions. These characteristics establish ScaleMaster as a challenging benchmark for testing SLAM robustness to intra- and inter-session scale inconsistency.

V. Benchmark Evaluation

In this section, we present our comprehensive evaluation of state-of-the-art monocular deep visual SLAM systems, designed to test their limits on ScaleMaster benchmark. Our complete experimental workflow—from data acquisition with our custom rig, through metric SE(3) pose estimation and baseline metric map generation, to the final map-to-map error calculation—is illustrated in Fig. 4.

Our analysis proceeds in three stages: we begin by applying conventional trajectory metrics to uncover Sim(3) pose estimation failure cases, then demonstrate the inherent

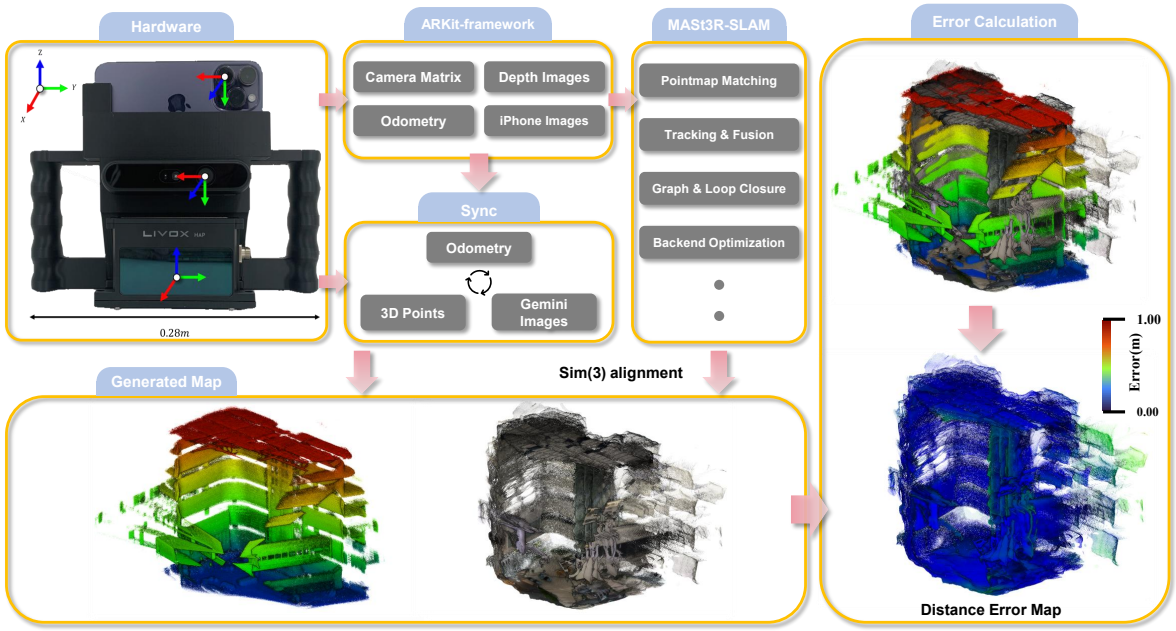

Fig. 4. Our overall experimental pipeline, illustrating the process from data acquisition with our custom rig (left), through ground truth map generation and SLAM processing (center), to the final map-to-map error calculation (right).

limitations of relying on these metrics alone, and finally introduce our direct map-to-map quality evaluation [24], which provides evidence of geometric inconsistencies further supported by qualitative visualizations.

A. Experimental Setup

- **Baselines:** We evaluate three representative deep learning-based SLAM systems: DROID-SLAM [10], MASt3R-SLAM [5], and VGGT-SLAM [11]. All experiments were conducted using the official implementations provided by the authors, on a desktop with an AMD Ryzen 9 9900X CPU and an NVIDIA RTX 5090 GPU.
- **(Pose-oriented) Trajectory Evaluation:** All trajectories are evaluated using the ATE in RMSE (m) using the `evo` evaluator [25]. To ensure comparability, a Sim(3) alignment [26] was applied for scale correction because our primary focus is on intra- and inter-scale consistency challenges, rather than on metric scale evaluation (i.e., a single global scale parameter estimation).
- **(Map-oriented) 3D Reconstruction Quality Evaluation:** We assess geometric fidelity by aligning the SLAM-generated map to a LiDAR point cloud, which we treat as a near–ground-truth metric map. We apply the scale, rotation, and translation (s, R, t) parameters derived from the aforementioned `evo` trajectory alignment based on Umeyama algorithm. This method tests the hypothesis that a correct trajectory should yield a correct map under the same transformation. After alignment, we compute the Chamfer distance and the Drop Rate(%); we define the Drop Rate as the percentage of points in the SLAM-generated map that have no corresponding point in the ground truth metric map within a predefined distance threshold. This metric is designed to quantify severe outliers and regions of complete map failure.

B. Trajectory Evaluation

We begin our quantitative analysis with the standard ATE to first establish a performance baseline and then reveal the critical failure modes that our benchmark uniquely exposes. As shown in Table III, leading algorithms like MASt3R-SLAM reconfirm their state-of-the-art performance on the ARKitScenes dataset [16], which establishes a baseline of their high performance under controlled, small-sized room conditions. This result can overstate general robustness. However, this perception of robustness does not hold when evaluated on the targeted challenges of our ScaleMaster dataset, as detailed in Table IV. For instance, in long-range sequences like `LargeHall_01`, the trajectory error surges to the 80-90 meter range. This is not a random tracking loss but a direct consequence of accumulated scale drift over a long trajectory. A similar phenomenon can be observed in a different sequence, as illustrated in Fig. 8.

These contrasting ATE results serve as a clear numerical demonstration that an algorithm's stability can be severely compromised by the scale-related challenges present in real-world environments, a weakness that remains hidden when evaluated on simpler, less complex benchmarks.

C. 3D Reconstruction Quality: Exposing Geometric Failures

While the dramatic ATE failures are informative, ATE provides only a partial characterization of system performance. It does not describe the geometric fidelity and completeness

TABLE III. Absolute Trajectory Error (ATE (m)) on ARKitScenes sequences [16].

Sequence	DROID SLAM	VGGT SLAM	MASt3R SLAM	MASt3R SLAM*
40777073	0.65	0.24	0.18	**0.13**
40958754	0.20	0.03	0.09	**0.02**
40958756	0.21	0.07	0.08	**0.03**
41007589	0.32	0.05	0.09	**0.03**
41045408	**0.01**	0.06	0.05	**0.01**
41048083	0.77	0.82	**0.08**	**0.08**
41048120	0.06	0.39	0.06	**0.05**
Average	0.32	0.24	0.09	**0.05**

* : calibrated mode

TABLE IV. Absolute Trajectory Error (unit: meter) on our dataset.

Sequence	DROID SLAM	VGGT SLAM	MASt3R SLAM	MASt3R SLAM*
Basement_01	**0.08**	1.44	0.38	0.42
HotelRoom_01	0.05	–	0.10	**0.06**
Lab_01	0.36	–	0.36	**0.09**
LargeHall_01	89.35	–	**80.54**	91.62
LargeHall_02	**3.78**	21.69	6.12	5.89
LargeHall_03	13.21	–	1.99	**1.96**
LargeHall_04	4.01	1.12	**0.57**	0.92
LargeHall_05	0.56	0.51	0.45	**0.33**
Library_01	**1.68**	–	5.29	3.61
Library_02	1.45	–	0.54	**0.63**
Library_03	0.09	–	0.09	**0.06**
Library_04	4.86	–	3.54	**3.22**
Library_05	4.35	13.26	**3.08**	4.00
Library_06	0.05	–	0.05	**0.04**
Library_07	0.13	0.22	0.13	**0.12**
Library_08	0.09	–	0.09	**0.06**
Library_09	0.04	–	0.07	**0.05**
Lobby_01	0.76	3.18	0.54	**0.27**
Lounge_01	4.51	–	0.47	**0.16**
Office_01	5.61	–	8.03	**0.65**
Parking_01	**10.21**	–	32.37	26.13
Parking_02	**0.20**	–	0.39	0.21
Stairs_01	20.20	–	4.60	**2.30**
Stairs_02	5.59	1.05	1.00	**0.14**
Station_01	11.66	–	13.21	**4.37**

* : calibrated mode
– : Runs that terminated with an invalid pose update (negative determinant during SL(4) normalization) were marked as failures; ATE is therefore undefined for those sequences.

of a resulting dense 3D map, which is the primary output of a dense visual SLAM system. As our subsequent analysis will prove, a SLAM system can produce a trajectory with low ATE, yet the corresponding map can be severely warped or incorrectly scaled due to internal scale inconsistencies. Therefore, a more direct evaluation of the 3D map is essential [27].

To overcome the limitations of ATE, we use Chamfer distance and Drop Rate mentioned in Sec. V-A. These direct

TABLE V. 3D Reconstruction Quality of MASt3R-SLAM

Sequence Name	Threshold	Chamfer Distance(m)	Drop Rate(%)	Analysis
Library_01	1	0.36	42.5	**Severe Failure**
	10	**6.43**	0.9	Significant map distortion.
Library_06	1	0.08	1.1	**Success**
	10	**0.10**	0.0	High-fidelity reconstruction.
Library_07	1	0.52	**89.1**	**Catastrophic Failure**
	10	**9.99**	0.0	Near-total map collapse.

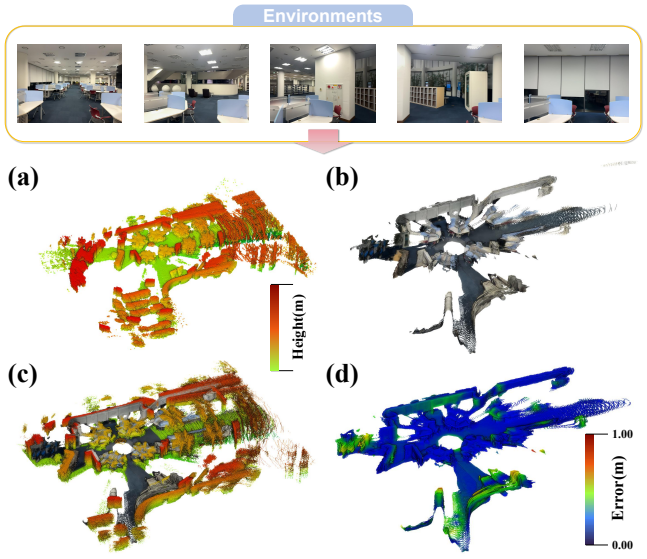

Fig. 5. Qualitative comparison of 3D reconstruction of the `Library_06` sequence. (a) Ground truth point cloud from LiDAR, color-coded by height. (b) The map reconstructed by MASt3R-SLAM. (c) The alignment of the MASt3R-SLAM map onto the ground truth. (d) Point-to-point distance error visualization, where warmer colors (red) indicate larger geometric inconsistencies between the two maps

map-to-map quality evaluations are presented in Table V. This analysis reveals the metric reconstruction failures.

- **Success Case**: On the `Library_06` sequence (see Fig. 5), the baseline algorithm (e.g., MASt3R-SLAM) performs well in both pose estimation (0.04 m in Table IV) and map reconstruction up to a single global scale, achieving a low Chamfer distance of 0.10 m. This example is the case where the scale is well maintained, thus both the trajectory and the map are correct.
- **Catastrophic Failure Case**: In contrast, Fig. 6 shows that the `Library_07` sequence suffers a geometric failure even though the pose error is low (0.12 m in Table IV). With the correspondence distance threshold set to 1 m, 89.1% of the generated map points were discarded as outliers, and when the threshold was increased to 10 m, the Chamfer distance reached a massive 9.99 m. This scale consistency-aware reconstruction failure is fundamentally invisible to ATE but is captured perfectly by the map quality metrics.

These results show that trajectory error alone, even after a single scalar scale adjustment, is insufficient and often misleading. Direct map quality evaluation is therefore necessary

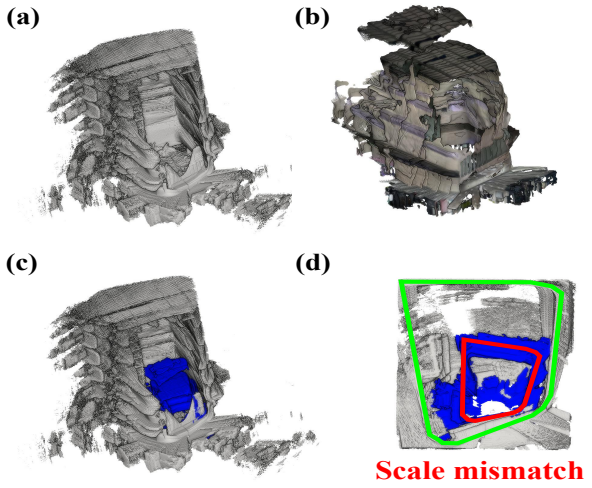

Fig. 6. Qualitative comparison on `Library_07` demonstrating that low trajectory error does not guarantee correct map scale. (a) The LiDAR ground-truth and (b) the MASt3R-SLAM reconstruction. (c) The overlay and (d) top-down cross-section are generated after applying the single global Sim(3) transformation calculated to align the camera trajectories.

to assess dense visual SLAM reliably.

D. Qualitative Analysis of Scale Inconsistency

Building on the taxonomy of scale inconsistency outlined in Fig. 2, we now illustrate three representative failure cases, each corresponding to a distinct category of the problem. The following three figures (Fig. 7, Fig. 8, and Fig. 9) serve as visual counterparts to the taxonomy in Fig. 2, grounding the abstract categories in concrete empirical evidence.

Intra-session Scale Inconsistency: Fig. 7 illustrates short-term scale failure in a vertical motion sequence, `Station_01`. Repetitive stair patterns disrupt data association between consecutive frames, causing immediate scale drift and severe map distortions like multi-layer overlaps. Furthermore, Fig. 8 shows a long-term optimization failure on the `Parking_01` sequence. Although a loop closure is detected, the massive accumulated scale drift traps the optimizer in a local minimum, resulting in a geometrically inconsistent map.

Inter-session Scale Ambiguity: Fig. 9 demonstrates another critical issue. When a single video is processed as three independent sessions, each resulting map fragment is generated at a different, inconsistent scale. While internally coherent, they cannot be merged into a single, globally scale-consistent map, highlighting a major challenge for long-term mapping or collaborative SLAM.

VI. Conclusion

In this work, we examined the fundamental challenges of three types of scale inconsistencies in deep monocular visual SLAM. With ScaleMaster, which exposes intra-session scale drift and inter-session scale ambiguity in large indoor environments, we coupled traditional trajectory evaluation with a direct map-to-map assessment (e.g., Chamfer distance and Drop Rate). Across multiple baselines (including DROID-SLAM, MASt3R-SLAM, and VGGT-SLAM), our quantita-

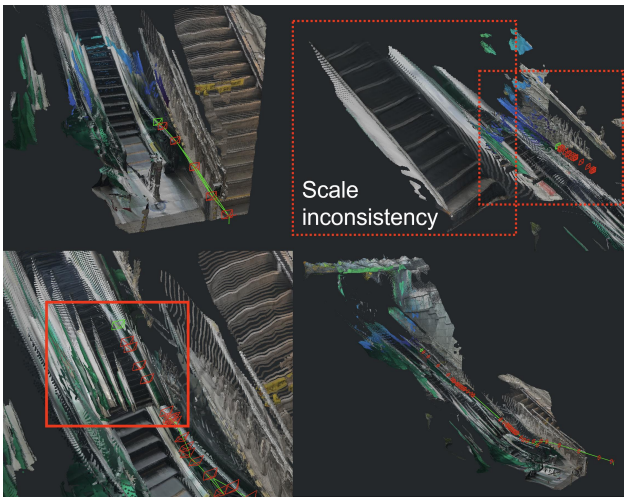

Fig. 7. Short-term scale inconsistency of MASt3R-SLAM on the `Station_01` sequence. The repetitive structure and vertical motion cause a severely distorted and overlapping 3D map. This demonstrates the system's vulnerability even in short, challenging trajectories.

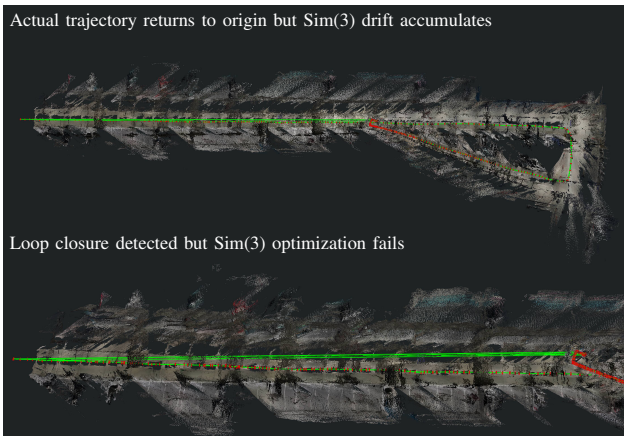

Fig. 8. A long-term optimization failure case for MASt3R-SLAM on the `Parking_01` sequence.

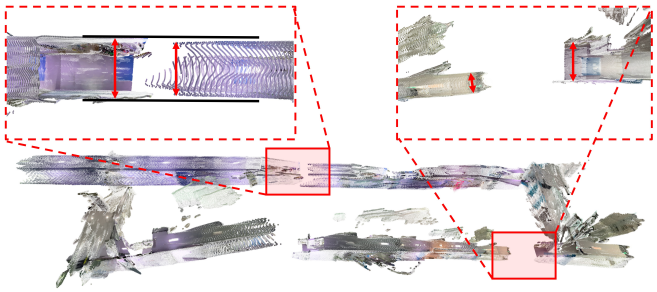

Fig. 9. A failure case of inter-session scale ambiguity in MASt3R-SLAM. The image shows the result of processing a single sequence as three independent sessions. The resulting map fragments cannot be aligned into a consistent map as each was generated with a different, inconsistent scale.

tive and qualitative results reveal severe scale-related failures under ScaleMaster's more realistic conditions, despite good performance on existing benchmarks. We hope this dataset and the accompanying baseline evaluations provide a solid platform for measuring and improving scale-consistent, reliable SLAM.

ACKNOWLEDGMENT

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.RS-2025-25420118), Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2025-02219277, AI Star Fellowship Support (DGIST)), and the InnoCORE program of the Ministry of Science and ICT (26-InnoCORE-01).

REFERENCES

[1] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2017.

[2] Carlos Campos, Richard Elvira, Juan J. Gomez, José M. M. Montiel, and Juan D. Tardós. ORB-SLAM3: An accurate open-source library for visual, visual-inertial and multi-map SLAM. *IEEE Transactions on Robotics*, 37(6):1874–1890, 2021.

[3] Luca Carlone, Ayoung Kim, Timothy Barfoot, Daniel Cremers, and Frank Dellaert, editors. *SLAM Handbook: From Localization and Mapping to Spatial Intelligence*. Cambridge University Press, 2025.

[4] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[5] Riku Murai, Eric Dexheimer, and Andrew J Davison. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16695–16705, 2025.

[6] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024.

[7] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of RGB-D SLAM systems. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 573–580, 2012.

[8] Michael Burri, Janosch Nikolic, Pascal Gohl, Thomas Schneider, Joern Rehder, Sammy Omari, Markus W Achtelik, and Roland Siegwart. The EuRoC micro aerial vehicle datasets. *The International Journal of Robotics Research*, 35(10):1157–1163, sep 2016.

[9] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time RGB-D camera relocalization. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2013.

[10] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. *Advances in Neural Information Processing Systems*, 2021.

[11] Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. *arXiv preprint arXiv:2505.12549*, 2025.

[12] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 31(5):1147–1163, 2015.

[13] Georg Klein and David Murray. Parallel tracking and mapping for small ar workspaces. In *2007 6th IEEE and ACM international symposium on mixed and augmented reality*, pages 225–234. IEEE, 2007.

[14] Margarita N Favorskaya. Deep learning for visual slam: the state-of-the-art and future trends. *Electronics*, 12(9):2006, 2023.

[15] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12786–12796, 2022.

[16] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitScenes: A diverse real-world dataset for 3D indoor scene understanding using mobile RGB-D data. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2021.

[17] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018.

[18] Heng Li, Xiaodong Gu, Weihao Yuan, Luwei Yang, Zilong Dong, and Ping Tan. Dense rgb slam with neural implicit maps. *arXiv preprint arXiv:2301.08930*, 2023.

[19] Hexiang Wei, Jianhao Jiao, Xiangcheng Hu, Jingwen Yu, Xupeng Xie, Jin Wu, Yilong Zhu, Yuxuan Liu, Lujia Wang, and Ming Liu. Fusionportablev2: A unified multi-sensor dataset for generalized slam across diverse platforms and scalable environments. *The International Journal of Robotics Research*, 44(7):1093–1116, 2025.

[20] Zhiqiang Chen, Yuhua Qi, Dapeng Feng, Xuebin Zhuang, Hongbo Chen, Xiangcheng Hu, Jin Wu, Kelin Peng, and Peng Lu. Heterogeneous lidar dataset for benchmarking robust localization in diverse degenerate scenarios. *The International Journal of Robotics Research*, page 02783649251344967, 2024.

[21] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[22] Hauke Strasdat, J Montiel, and Andrew J Davison. Scale drift-aware large scale monocular SLAM. *Robotics: Science and Systems (RSS)*, 2010.

[23] Been Kim, Michael Kaess, Luke Fletcher, John Leonard, Abraham Bachrach, Nicholas Roy, and Seth Teller. Multiple relative pose graphs for robust cooperative mapping. In *2010 IEEE International Conference on Robotics and Automation*, pages 3185–3192. IEEE, 2010.

[24] Tong Wu, Liang Pan, Junzhe Zhang, Tai Wang, Ziwei Liu, and Dahua Lin. Density-aware chamfer distance as a comprehensive metric for point cloud completion. *arXiv preprint arXiv:2111.12702*, 2021.

[25] Michael Grupp. evo: Python package for the evaluation of odometry and SLAM. https://github.com/MichaelGrupp/evo, 2017.

[26] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 13(4):376–380, 2002.

[27] Xiangcheng Hu, Jin Wu, Mingkai Jia, Hongyu Yan, Yi Jiang, Binqian Jiang, Wei Zhang, Wei He, and Ping Tan. MapEval: Towards unified, robust and efficient SLAM map evaluation framework. *IEEE Robotics and Automation Letters (RA-L)*, 2025.