

LiDAR-VGGT: Cross-Modal Coarse-to-Fine Fusion for Globally Consistent and Metric-Scale Dense Mapping

Lijie Wang¹, Lianjie Guo¹, Ziyi Xu¹, Qianhao Wang¹, Fei Gao¹, and Xieyuanli Chen²

Abstract—Reconstructing large-scale RGB point clouds is an important task in robotics, supporting perception, navigation, and scene understanding. Despite advances in LiDAR inertial visual odometry (LIVO), its performance remains highly sensitive to extrinsic calibration. Meanwhile, 3D vision foundation models, such as VGGT, suffer from limited scalability in large environments and inherently lack metric scale. To overcome these limitations, we propose LiDAR-VGGT, a novel framework that couples LiDAR inertial odometry with the state-of-the-art VGGT model through a two-stage coarse-to-fine fusion pipeline: First, a pre-fusion module robustly estimates per-session VGGT poses and RGB point clouds with a coarse metric scale. Then, a post-fusion module enhances cross-modal 3D similarity transformation, using bounding-box-based regularization to reduce scale distortions caused by inconsistent FOVs between LiDAR and camera sensors. Extensive experiments across multiple datasets demonstrate that LiDAR-VGGT achieves improved visual quality and density compared to VGGT-based methods and LIVO baselines. The implementation of our method and the proposed novel RGB point cloud evaluation toolkit are released as open source.

Index Terms—SLAM, mapping, sensor fusion.

I. INTRODUCTION

RECONSTRUCTING accurate RGB point clouds serves as the foundation for key robotics capabilities, such as perception [4], navigation [5], and scene understanding [6], in complex environments. Moreover, large-scale 3D reconstruction is crucial for embodied navigation, enabling robots to safely

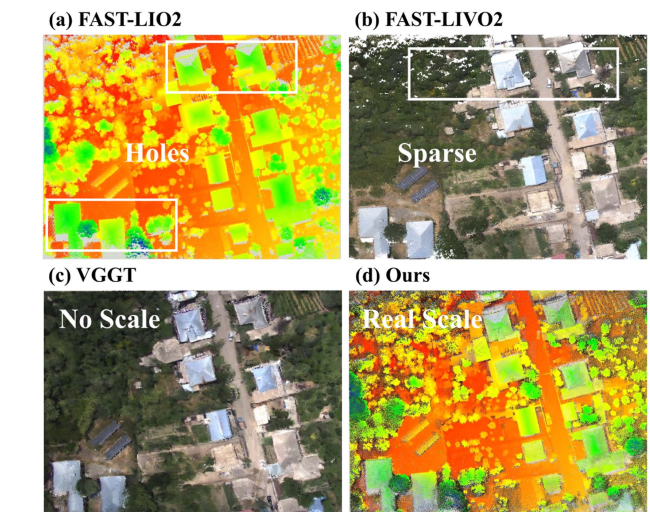


Fig. 1. (a) and (b) present point clouds from FAST-LIO2 [1] and FAST-LIVO2 [2], which are sparse and contain holes due to limited LiDAR scanning coverage. (c) shows the dense yet scale-ambiguous map produced by VGGT [3]. (d) illustrates our method, achieving dense maps with accurate metric scale.

explore unstructured environments [7]; for multi-robot collaboration, where shared maps support coordinated tasks [8], [9]; and for autonomous driving, providing precise environmental models for localization and planning [10].

Classical LiDAR inertial visual odometry (LIVO) pipelines have achieved progress by fusing multi-modal measurements for pose estimation and map reconstruction [2], [11], [12], [13], [14]. However, their performance heavily relies on precise extrinsic calibration and tight timestamp synchronization across sensors, which compromises robustness in real-world deployments. Moreover, the inherent sparsity of LiDAR scans prevents the generation of dense, high-fidelity RGB 3D reconstructions as shown in Fig. 1.

The rise of foundation models for 3D vision has opened new avenues for dense scene reconstruction. Early works such as DUST3R [15] and MAST3R [16] demonstrate the potential to directly predict dense point clouds from images, though they are limited to processing image pairs and rely on post-fusion for larger scenes. More recently, the Visual Geometry Grounded Transformer (VGGT) [3] extend this paradigm by jointly performing pose estimation, depth prediction, and dense 3D reconstruction from multiple uncalibrated RGB frames in a single inference. Although VGGT achieves dense and clean reconstructions without explicit optimization, its scalability is

Received 30 October 2025; accepted 11 February 2026. Date of publication 19 February 2026; date of current version 3 March 2026. This article was recommended for publication by Associate Editor Y. Yue and Editor J. Civera upon evaluation of the reviewers' comments. This work was supported in part by National Key R&D Program of China under Grant 2023YFB4706600, in part by Zhejiang Provincial Science and Technology Plan Project under Grant 2024C01170, and in part by the National Natural Science Foundation of China under Grant 62322314 and 62403478. (Corresponding author: Xieyuanli Chen.)

Lijie Wang, Lianjie Guo, Qianhao Wang, and Fei Gao are with the State Key Laboratory of Industrial Control Technology, Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China, and also with Differential Robot Technology Company, Ltd., Hangzhou 311100, China (e-mail: 3210101760@zju.edu.cn).

Ziyi Xu is with the State Key Laboratory of Industrial Control Technology, Institute of Cyber-Systems and Control, Zhejiang University, Hangzhou 310027, China.

Xieyuanli Chen is with the College of Intelligence Science and Technology, National University of Defense Technology, Changsha 410007, China (e-mail: xieyuanli.chen@nudt.edu.cn).

Our code is available at <https://github.com/NorwegianSmokedSalmon/LiDAR-VGGT>.

Digital Object Identifier 10.1109/LRA.2026.3666387

restricted by GPU memory, and the generated RGB point clouds lack a metric scale, limiting its practicality for large-scale robotic applications. To process long image sequences in large environments, recent works [17], [18] adopt segment processing by dividing images into smaller sessions and aligning adjacent sessions to achieve large-scale mapping. Although this reduces memory requirements, such pairwise alignment only enforces relative scales between submaps without establishing a globally consistent metric scale. Furthermore, these methods rely on point cloud correspondences inherited from a small number of shared frames for submap alignment, such that low-quality shared-frame reconstructions can lead to inaccurate or even erroneous alignment between adjacent submaps.

In contrast, LiDAR-based SLAM [1], [19] naturally provides accurate metric motion estimation and global consistency through loop detection [20], [21], [22] and pose graph optimization (PGO). By leveraging LiDAR to anchor VGGT reconstructions, dense, visually clean, and metrically accurate RGB maps are obtained.

To this end, we propose LiDAR-VGGT, a novel framework that tightly integrates LiDAR with VGGT through a coarse-to-fine fusion pipeline. In the coarse fusion stage, LiDAR inertial odometry (LIO) poses are used to refine the VGGT poses via a robust refinement mechanism that combines scale RANSAC and linearity-based validation to reject outliers in the $Sim(3)$ transformation. This enables the scale-normalized local poses of VGGT to be transformed into the real-world coordinate system while generating RGB point clouds with an approximate metric scale. Then the fine fusion module leverages the initial value to further improve the accuracy and consistency of the RGB map. Each session's VGGT point cloud is aligned with the LiDAR point cloud using a stabilized cross-modal $Sim(3)$ registration that employs bounding-box-based regularization to reduce scale distortions caused by field-of-view differences between the camera and LiDAR. This ensures stable registration and recovers the real-world metric scale. Finally, a global PGO is applied to eliminate divergence across different sessions and improve global consistency in large-scale scenes. For the evaluation of RGB point clouds, existing methods mainly rely on rendering-based metrics, which assess visual quality in 2D views and largely ignore the underlying 3D geometric structure and ground-truth maps [14], [23]. However, color quality in RGB point clouds is inherently coupled with geometric accuracy, as misalignment or structural distortion directly affects color correctness in 3D space. In contrast, we propose evaluation metrics that directly operate on ground-truth maps and comprehensively evaluate color fidelity and geometric consistency.

Our contributions can be summarized as follows.

- We propose LiDAR-VGGT, the first system that fuses LiDAR with VGGT to achieve large-scale, metric-accurate, dense, and globally consistent RGB point cloud reconstruction.
- We design a unified VGGT-LiDAR fusion pipeline that integrates a coarse pre-fusion stage and a fine post-fusion stage, enabling robust registration of VGGT reconstructions to the LiDAR map.
- We introduce an open-source toolkit¹ that evaluates point cloud color quality by jointly considering geometric consistency and ground-truth color fidelity.

¹<https://github.com/NorwegianSmokedSalmon/Color-Map-Evaluation>

II. RELATED WORKS

A. Multi-Modal SLAM and Reconstruction

Classical Multi-Modal SLAM approaches fuse information from multiple sensors such as LiDAR, camera, and IMU to estimate trajectories and reconstruct 3D scenes. The R³LIVE [12], [13] series tightly couple the LiDAR points with the image pixels to achieve localization and RGB point cloud reconstruction. FAST-LIVO and FAST-LIVO2 [2], [11] exploit the tight coupling between LiDAR points and image patches, enabling robust localization in degraded environments and producing high-precision 3D maps. In parallel, recent advances in Neural Radiance Fields (NeRF) [24] and 3D Gaussian Splatting (3DGS) [25] have inspired current works to integrate new map representations into SLAM systems. Extensions such as LiDAR-NeRF [26] and SiLVR [27] integrate LiDAR depth and SLAM trajectories to achieve scalable, geometrically accurate reconstructions with metric scales. Additionally, systems like Gaussian-LIC [28] and GS-LIVO [14] tightly fuse LiDAR, IMU, and visual data with Gaussian maps for real-time SLAM, maintaining robustness under challenging illumination, motion, and resource constraints. In summary, these pipelines achieve accurate localization and high-precision reconstruction through different sensor fusion methods and map representation. However, they typically require accurate extrinsic calibration and precise time synchronization between sensors. Moreover, LiDAR provides only sparse range measurements, which restricts the density of reconstructions.

B. Feed-forward Scene Reconstruction

Beyond explicit geometric optimization, another line of research employs feed-forward neural networks to model 3D scenes from images. DUST3R [15] first demonstrates the effectiveness of large-scale, data-driven reconstruction by predicting dense point maps from image pairs and recovering scene geometry via global alignment, but its geometric consistency degrades as the number of input images increases. To improve scalability, subsequent works explore incremental reconstruction: Spann3R [29] introduces a learned memory module for streaming sequences, MV-DUST3R+ [30] extends DUST3R to multi-frame inputs, and SLAM3R [31] incrementally aligns locally reconstructed point maps using a Local-to-World module. However, these methods lack loop-closure mechanisms to correct accumulated drift, limiting their robustness to short sequences. Fast3R [32] further scales feed-forward inference to larger image sets, but with limited accuracy.

VGGT (Visual Geometry Grounded Transformer) [3] pioneers a unified feed-forward framework for multi-frame camera pose estimation, depth prediction, and dense reconstruction, producing visually clean results without explicit optimization. However, its scalability is limited by GPU memory, restricting inference to only tens of frames. To enable large-scale mapping, VGGT-Long [17] and VGGT-SLAM [18] partition long sequences into submaps and perform $Sim(3)$ and $SL(4)$ registration. Nevertheless, these segmented pipelines enforce only relative scale consistency and lack a reliable global metric scale; their reliance on pairwise alignment of VGGT-generated submaps makes them sensitive to low-quality shared point clouds, leading to unstable or incorrect solutions under degenerate or drift-prone conditions.

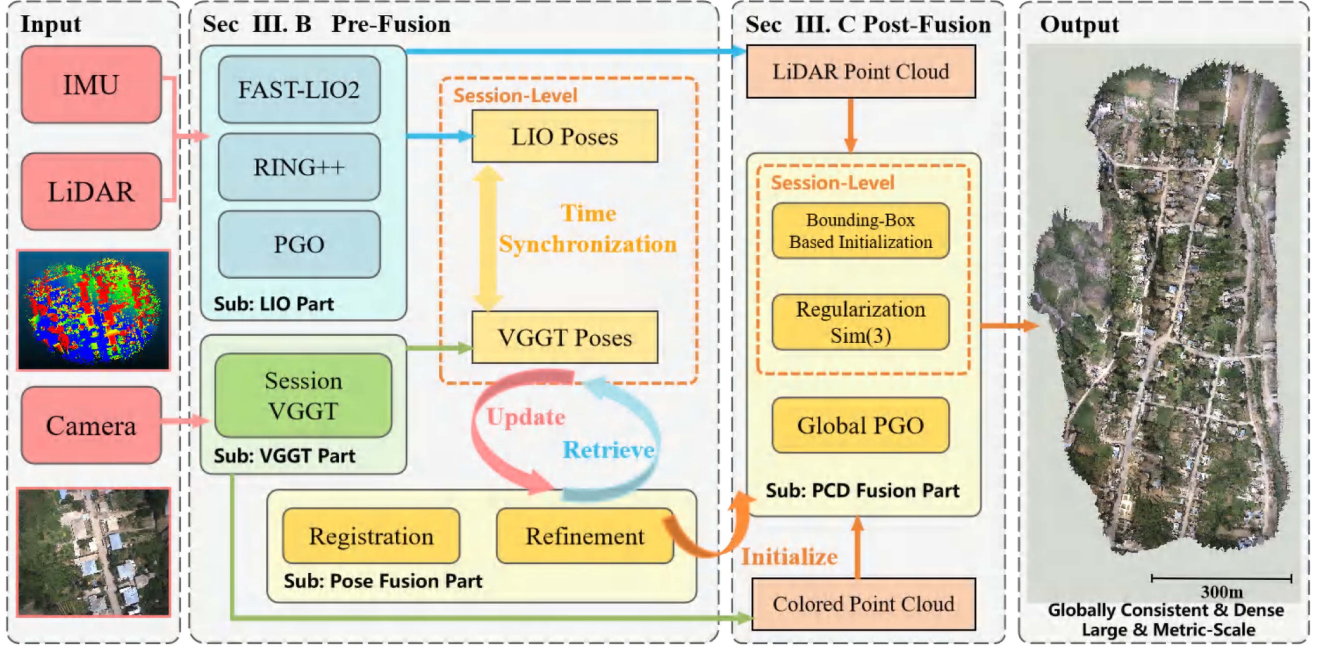


Fig. 2. Overview of the proposed LiDAR-VGGT system. It takes multi-modal data from IMU, LiDAR, and camera as input and produces globally consistent, dense reconstructions with metric scale. Our pipeline follows a coarse-to-fine strategy, consisting of a pre-fusion module (Section III-B) and a post-fusion module (Section III-C). The former fuses LIO and VGGT poses to establish an approximate real-world scale for VGGT-generated scene attributes, while the latter refines the VGGT RGB point cloud through stabilized cross-modal $Sim(3)$ registration and global pose graph optimization.

III. METHODOLOGY

A. Problem Formulation and System Overview

VGGT takes N RGB images, $S_N = \{I_1, I_2, \dots, I_N\}$ as input, where each image is denoted as $I_i \in \mathbb{R}^{3 \times H \times W}$. It outputs scene properties and reconstruction formulated as:

$$f(S_N) = (\mathbf{g}_i, D_i, P_i, Tr_i)_{i=1}^N. \quad (1)$$

For image I_i , $\mathbf{g}_i \in \mathbb{R}^9$ represents the camera intrinsics and extrinsics, $D_i \in \mathbb{R}^{H \times W}$ is the predicted depth map, $P_i \in \mathbb{R}^{3 \times H \times W}$ is the corresponding point map, and $Tr_i \in \mathbb{R}^{C \times H \times W}$ is a grid of features for point tracking. In large-scale scenarios, the image sequence S_N can be divided into K overlapping sessions $\mathbb{S}_K = \{S_1, S_2, \dots, S_K\}$, where each session $S_k \in \mathbb{S}_K$ contains L frames. Each session is independently processed by VGGT, producing a set of RGB point cloud sessions $\mathbb{P}_K = \{\mathcal{P}_1, \mathcal{P}_2, \dots, \mathcal{P}_K\}$, where $\mathcal{P}_k \in \mathbb{P}_K$ represents the map of sequence S_k aligned to the coordinate frame of its first image.

Our goal is to register each session $\mathcal{P}_k$ with the corresponding LiDAR point cloud $\mathcal{L}_k$ from LIO and then merge all sessions into a consistent global map. We estimate a similarity transformation $\mathcal{T}_k^{V \rightarrow L} \in Sim(3)$ between $\mathcal{P}_k$ and $\mathcal{L}_k$, which includes a $SE(3)$ transformation and a scale $s \in \mathbb{R}^+$. This $Sim(3)$ transformation aligns each VGGT session with its LiDAR counterpart. Next, the relative transformations $T_{k \rightarrow k+1} \in SE(3)$ between adjacent VGGT sessions are refined to achieve seamless fusion and a globally consistent reconstruction. Our proposed LiDAR-VGGT addresses these challenges, as illustrated in Fig. 2. It fuses LiDAR scans, IMU measurements, and images to produce a dense, globally consistent RGB map with accurate metric scale. Our system consists of two modules: the Coarse Pre-Fusion Module (Section III-B), and the Fine Post-Fusion Module (Section III-C).

B. Coarse Pre-Fusion Module

For the VGGT component, each image sequence is down-sampled to preserve sufficient inter-frame disparity for reliable VGGT inference while reducing computational cost. The selected images are then processed independently to estimate camera poses and reconstruct a dense RGB map. For the LIO component, we use FAST-LIO2 [1] to build a LiDAR point cloud map. To ensure long-term global consistency, RING++ [20], a fast LiDAR scan descriptor, is employed for loop closure detection. The detected loop constraints are incorporated into a pose graph, which is subsequently optimized through PGO to obtain a globally consistent LiDAR map.

Pose Registration: We estimate a $Sim(3)$ transformation for each session $S_k \in \mathbb{S}_K$ using the LiDAR poses to align the VGGT poses $\{T_i^{VGGT}\}_{i=1}^L$ to the world coordinate frame. For brevity, the session index k is omitted below. Each LiDAR pose is converted into the camera pose in the world system $\{T_i^{Cam}\}_{i=1}^L$ with the same timestamp using extrinsic parameters. Each VGGT pose is then paired with the nearest camera pose in time:

$$C_k = \left\{ (T_i^{VGGT}, T_{j^*(i)}^{Cam}) \mid i = 1, \dots, L \right\},$$

$$j^*(i) = \arg \min_j |t_i^{VGGT} - t_j^{Cam}|, \quad (2)$$

where t_i^{VGGT} and t_j^{Cam} are the timestamp of the two poses. Given the poses in pairs, we denote that $\{\mathbf{x}_i\}_{i=0}^L$ and $\{\mathbf{y}_i\}_{i=0}^L$ are the translations of VGGT poses and camera poses. We estimate a standard $Sim(3)$ transformation between the two pose sets using the closed-form Umeyama method [33], which jointly recovers scale, rotation, and translation.

Linearity Validation and Rotation Correction: Although the Umeyama method provides a closed-form $Sim(3)$ solution, its

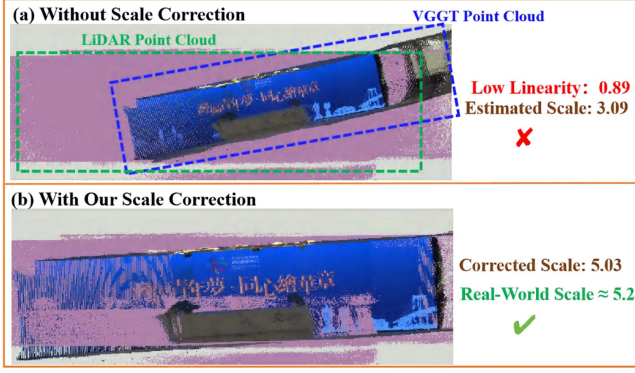


Fig. 3. (a) demonstrates an example with low linearity of LIO poses and thus yielding incorrect scale after pose registration. (b) shows the result with our scale correction which recovers an accurate scale close to the real value.

rotation estimation becomes unstable under weak geometric constraints, particularly when the trajectory is close to linear. To detect degenerate motion, we perform Principal Component Analysis (PCA) on the camera pose sequence and define a linearity score using the eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3$:

$$\ell = 1 - \frac{\lambda_2 + \lambda_3}{\lambda_1}. \quad (3)$$

A larger ℓ indicates near-linear motion. For sequences with high linearity, we refine the initial rotation by applying a standard rotation averaging procedure between the VGGT rotations and the camera rotations. Specifically, relative rotations are aggregated across frames and projected onto $SO(3)$ via singular value decomposition, following common practice in rotation averaging. The resulting averaged rotation is then used to correct the initial estimate.

Scale RANSAC and Refinement: Inaccurate VGGT poses can distort scale estimation, as shown in Fig. 3(a). We thus employ a RANSAC-based approach to identify reliable scales and refine session-level scaling. Empirically, pose sequences exhibiting high linearity are more likely to yield accurate scale estimates, since smooth and near-linear motions facilitate stable pose estimation in VGGT. Conversely, aggressive motions with low linearity introduce larger pose estimation errors, which in turn degrade the reliability of the recovered scale. Therefore, we prioritize sequences with high linearity as inlier priors in our scale RANSAC. Let $\{s_k\}_{k=1}^K$ denote initial scale estimates and $\{\ell_k\}_{k=1}^K$ their linearity scores. We define a scale threshold based on the standard deviation:

$$\text{threshold} = k_\sigma \cdot \sigma_s, \quad \sigma_s = \text{std}(\{s_k\}), \quad (4)$$

where k_σ is set to 2. High-linearity frames are prioritized by converting ℓ_k to sampling probabilities via softmax:

$$p_k = \frac{\exp(\alpha \ell_k)}{\sum_{k=1}^K \exp(\alpha \ell_k)}, \quad (5)$$

where α is calculated by the coefficient of variation cov :

$$\alpha = \frac{1}{cov} = \frac{\frac{1}{K} \sum_{k=1}^K s_k}{\sigma_s}. \quad (6)$$

For a candidate scale s_c , inliers are defined as:

$$\mathcal{I}(s_c) = \{k \mid |s_k - s_c| < \text{threshold}\}. \quad (7)$$

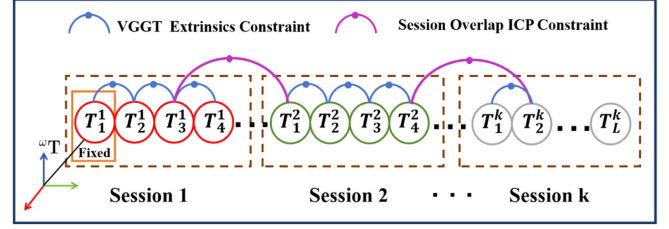


Fig. 4. Structure of our global PGO. The first frame of the first session is fixed as the world reference frame.

After iterations, scale outliers are corrected by combining the nearest inlier scale with the $Sim(3)$ estimated scale from their overlapping poses, thus ensuring consistent and correct scale propagation, as shown in Fig. 3(b).

C. Fine Post-Fusion Module

This module further refines registration through stabilized cross-modal $Sim(3)$ and global PGO.

Stabilized Cross-Modal $Sim(3)$ Registration: We initialize the registration between the VGGT session $\mathcal{P}_k$ and the corresponding LiDAR session $\mathcal{L}_k$ by transforming $\mathcal{P}_k$ using the frontend estimate (Section III-B) to obtain $\mathcal{P}'_k$. However, due to the field-of-view (FOV) discrepancy between camera and LiDAR, unconstrained $Sim(3)$ may overfit and cause scale drift. To stabilize scale, we introduce a regularized objective:

$$\min_{s_2 > 0, \mathbf{R}_2 \in SO(3), \mathbf{t}_2 \in \mathbb{R}^3} \sum_{i=1}^M \|\mathbf{q}_i - (s_2 \mathbf{R}_2 \mathbf{p}_i + \mathbf{t}_2)\|^2 + \lambda (s_2 - s_1^*)^2, \quad (8)$$

where s_1^* denotes the scale estimated in the pre-fusion stage, and the regularization weight λ is defined as:

$$\lambda = \beta \cdot n \cdot D^2, \quad \beta \in (0, 1], \quad (9)$$

and n denotes the number of points in the source map, D represents the diagonal length of the bounding box, and β is a tunable coefficient controlling the regularized strength.

Unlike normal $Sim(3)$, the scale regularization couples s_2 with the alignment residuals, preventing a joint closed-form solution. Thus the optimization is solved through an alternating strategy, where point correspondences are first established via nearest-neighbor search and iteratively updated after each step, since direct cross-modal correspondences are not available. Specifically, we fix $\mathbf{R}_2$ and $\mathbf{t}_2$ to update the scale s_2 , followed by refining $\mathbf{R}_2$ and $\mathbf{t}_2$ based on the updated correspondences. The closed-form solution for s_2 is:

$$s_2 = \frac{\sum_{i=1}^M (\mathbf{q}_i - \mathbf{t}_2)^\top \mathbf{R}_2 \mathbf{p}_i + \lambda s_1^*}{\sum_{i=1}^M \|\mathbf{R}_2 \mathbf{p}_i\|^2 + \lambda}. \quad (10)$$

The regularization term constrains scale drift from the initial estimate s_1^* , stabilizing cross-modal alignment.

Global Pose Graph Optimization: While session-level registration ensures local accuracy, global consistency is achieved through PGO as shown in Fig. 4. The optimization includes intra-session constraints from VGGT poses and inter-session constraints obtained via ICP on overlapping regions. Implemented in GTSAM [34], the optimized poses are used to re-project all RGB point clouds into a common global frame, resulting in a globally consistent reconstruction.

D. RGB Point Cloud Map Evaluation

Color Distance and Color Fidelity: The CD measures the average bidirectional color difference between the reconstructed and reference point clouds, and CF expresses it in decibel form for interpretability.

$$\text{CD} = \frac{1}{2N_s} \sum_{i=1}^{N_s} \|c_i^s - c_{\pi(i)}^r\|_2 + \frac{1}{2N_r} \sum_{j=1}^{N_r} \|c_j^r - c_{\pi'(j)}^s\|_2,$$

$$\text{CF} = -20 \log_{10}(\text{CD}), \quad (11)$$

where c_i^s and c_j^r denote the RGB color vectors of the source and reference point clouds, N_s and N_r are their point numbers, and $\pi(i)$ and $\pi'(j)$ represent the nearest-neighbor correspondences between the two point sets. Smaller CD and higher CF indicate better color fidelity.

Local Color Recall: The LCR measures the proportion of reference points whose local color difference in the reconstructed map is below a threshold:

$$\text{LCR}_{r \rightarrow s}(\tau, r_g) = \frac{1}{N_r} \sum_{j=1}^{N_r} \mathbb{I} \left(\min_{i \in \mathcal{N}_s(x_j^r; r_g)} \|c_i^s - c_j^r\|_2 \leq 3\tau \right), \quad (12)$$

where $\mathcal{N}_s(x_j^r; r_g)$ is the neighborhood of x_j^r within radius r_g , τ is the color threshold, and $\mathbb{I}(\cdot)$ is the indicator function that returns 1 if the condition holds and 0 otherwise.

Color Consistency Score: We assume that points within a small voxel of a map have similar RGB values due to spatial continuity. The CCS quantifies this local color homogeneity by averaging the RGB covariance trace of all voxels:

$$\text{CCS} = \frac{1}{N_v} \sum_{k=1}^{N_v} C_i^{(k)}, \quad C_i^{(k)} = \text{tr}(\Sigma_c^{(k)}),$$

$$\Sigma_c^{(k)} = \frac{1}{n_k - 1} \sum_{j=1}^{n_k} (c_j^{(k)} - \bar{c}^{(k)})(c_j^{(k)} - \bar{c}^{(k)})^\top, \quad (13)$$

where N_v is the number of occupied voxels, n_k is the number of points in the k -th voxel, $c_j^{(k)}$ and $\bar{c}^{(k)}$ are the RGB vector and its mean in voxel k , and $\Sigma_c^{(k)}$ is the RGB covariance matrix. Smaller CCS indicates more consistent colors.

IV. EXPERIMENTS

A. Experimental Setup

To evaluate the performance of LiDAR-VGGT, we conduct experiments on several datasets, including the public MARS-LVIG [35], FAST-LIVO2 [2] and MUN_FRL [36] dataset, and a self-collect dataset acquired using the hardware platform shown in Fig. 5, their parameters are shown in Table I. Specifically, MARS-LVIG is a large-scale aerial dataset, with each sequence covering on the order of 10^5 m^2 . We first assess the overall mapping quality, benchmarking our approach against VGGT-Long, VGGT-SLAM and SLAM3R. We then perform a registration experiment for ablation to evaluate the individual contributions of our pre-fusion and post-fusion modules, with particular attention to the impact of the proposed enhanced *Sim(3)* with regularization. Finally, a multi-modal comparison and robustness

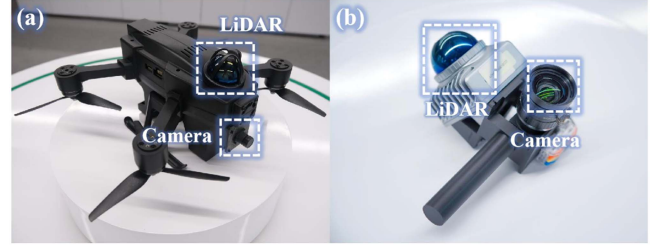


Fig. 5. Our two data collection platforms. The UAV device (a) is used for outdoor data collection, while the handheld device (b) is used for indoor.

TABLE I
OVERVIEW OF DATASET FOR EXPERIMENTS

Dataset	Sequence	Trajectory Length (m)	LiDAR
MARS-LVIG	HKAirport01	2058	Avia
	HKIsland01	1877	Avia
	AMvalley01	4377	Avia
	AMtown01	4915	Avia
	HKAirport03	2073	Avia
FAST-LIVO2	Brightwall	24	Avia
	CBDBuilding	34	Avia
MUN_FRL	LightHouse	168	Velodyne
	Flight	319	Velodyne
Self-Collect	Company	75	MID360
	Street	69	MID360
	TechnologyPark	45	MID360

analysis is conducted, providing both quantitative and qualitative comparisons with LIO and LIVO methods.

B. Mapping Quality Experiment

For brevity, we denote benchmark methods VGGT-Long as V-Long and the two modes of VGGT-SLAM as V-Sim(3) and V-SL(4). We evaluate both geometric fidelity and color quality against ground-truth RGB maps. For MARS-LVIG, the ground truth is a high-precision point cloud generated by the DJI L1 and DJI Terra system, while for MUN_FRL it is obtained by interpolating ground-truth poses with known extrinsics and projecting images onto the point cloud. For benchmarking, we use CloudCompare to match the scale to the ground-truth map and align the first frame. Detailed definitions of the geometric metrics are provided in [37].

The geometric fidelity results are summarized in Table II. Our method achieves the best performance across all four metrics, including Chamfer distance (CD), ICP overlap ratio, average Wasserstein distance (AWD), and Fitness. Notably, a lower AWD indicates better global geometric consistency by jointly capturing large-scale structure and local geometry. Except for the most challenging AMvalley01 sequence with a coverage area exceeding 450000 m^2 and many regions with similar textures, our method achieves an overlap ratio above 50% on the remaining datasets, demonstrating effective preservation of metric scale and geometric structure. On large-scale sequences spanning hundreds to thousands of meters, LiDAR-only SLAM accumulates noticeable drift, whereas LiDAR-VGGT maintains an AWD close to 1 m, indicating strong global consistency and close alignment with ground truth. Our method also achieves

TABLE II
MAPPING GEOMETRIC FIDELITY COMPARISON

	Data	V-Long	V-Sim(3)	SLAM3R	V-SL(4)	Ours
CD# (m)↓	AMtown01	52.01	51.34	184.4	×	2.23
	AMvalley01	55.14	51.12	135.4	×	10.2
	HKisland01	26.2	17.14	35.5	×	6.33
	LightHouse	25.49	34.9	15.5	23.25	2.34
	Flight	20.09	16.92	87.9	14.0	3.7
AC (m)↓	AMtown01	0.044	0.044	0.04	×	0.04
	AMvalley01	0.04	0.043	0.036	×	0.035
	HKisland01	0.033	0.035	0.033	×	0.039
	LightHouse	0.03	0.03	0.036	0.028	0.04
	Flight	0.044	0.042	0.04	0.041	0.04
ICP Overlap ↑	AMtown01	8%	1.2%	9.3%	×	53%
	AMvalley01	2.3%	2.2%	4.9%	×	36.8%
	HKisland01	13.5%	28.2%	21.6%	×	58.6%
	LightHouse	8.8%	16.5%	32.5%	12.9%	91%
	Flight	33%	23.5%	8.3%	22.4%	78.9%
AWD (m)↓	AMtown01	1.67	1.4	1.37	×	0.8
	AMvalley01	1.89	1.75	1.79	×	1.59
	HKisland01	1.64	1.34	1.42	×	1.12
	LightHouse	1.88	1.67	1.92	1.80	1.41
	Flight	1.59	1.66	1.69	1.78	1.50
Fitness ↑	AMtown01	0.007	0.001	0.004	×	0.05
	AMvalley01	0.0012	0.002	0.003	×	0.023
	HKisland01	0.006	0.015	0.01	×	0.04
	LightHouse	0.0043	0.009	0.03	0.0064	0.13
	Flight	0.036	0.022	0.006	0.022	0.08

CD# denotes Chamfer distance, × denotes system fails.

TABLE III
RGB POINT CLOUD COLOR QUALITY EVALUATION

	Data	V-Long	V-Sim(3)	SLAM3R	Ours
CD* ↓	AMtown01	130.49	130.96	122.22	82.93
	AMvalley01	87.04	85.76	85.50	84.48
	HKisland01	158.21	142.29	152.83	112.90
	LightHouse	67.92	70.80	72.23	58.56
	Flight	59.84	57.98	59.23	57.55
CF (dB)↑	AMtown01	11.87	11.84	12.41	15.81
	AMvalley01	15.39	15.52	15.54	15.57
	HKisland01	10.21	11.09	10.50	13.12
	LightHouse	11.49	11.13	10.96	12.78
	Flight	12.59	12.87	12.68	12.93
CCS ↓	AMtown01	1.53	2.07	1.76	1.19
	AMvalley01	0.83	0.88	0.73	0.67
	HKisland01	1.99	1.76	1.81	1.11
	LightHouse	1.54	2.18	1.72	1.62
	Flight	1.31	1.75	1.27	1.12
LCR ↑	AMtown01	2.6%	0.4%	0.86%	24.1%
	AMvalley01	1.32%	1.27%	0.5%	17.8%
	HKisland01	3.2%	5.96%	2.8%	23.0%
	LightHouse	1.7%	4.6%	6.1%	36.7%
	Flight	7.9%	3.1%	0.8%	18.6%

CD* denotes color distance.

an average CD of only a few meters; slightly higher CD values in some cases are mainly caused by field-of-view differences, where bidirectional evaluation amplifies errors in non-overlapping regions. Finally, LiDAR-VGGT achieves the best accuracy distance (AC) on three datasets, reflecting higher local positional accuracy.

The color quality evaluation results are shown in Table III. Our method achieves near-optimal performance on all metrics,



Fig. 6. Our large-scale mapping result in the MARS-LVIG HKisland01 dataset. The enlarged details show that the reconstructed RGB point cloud is dense, consistent, and maintains an accurate metric scale.

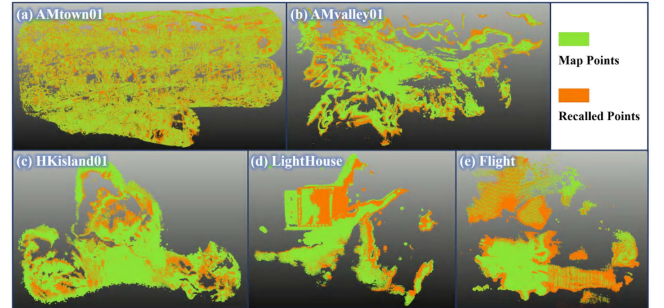


Fig. 7. The visualization of LCR on five sequences shows that our mapping results maintain good local color consistency compared to the ground-truth map. The holes in the middle are mainly caused by the lower geometric quality of the VGGT point clouds, which deviate from the truth map.

demonstrating excellent color fidelity. Fig. 7 shows the LCR at $\tau = 0.1$ and $r_g = 0.5m$ in all five datasets, illustrating that our map maintains overall color consistency with the ground-truth, while the empty areas mainly result from the poor geometric quality of the VGGT point cloud with little overlap. Since the benchmark methods fail on the challenging dataset, only our results are shown in Fig. 6. As seen, our method produces high-quality, dense maps over a large-scale island scene.

C. Registration Experiment for Ablation

We use the LiDAR point cloud as the reference to evaluate registration performance and compare three variants: the normal *Sim(3)*, our proposed stabilized cross-modal *Sim(3)*, and the pre-fusion module with only pose registration. We compute the root mean square error (RMSE), CD, AWD, and spatial consistency score (SCS) for each session and report their averages for

TABLE IV
REGISTRATION COMPARISON FOR ABLATION

Data	Pre-Fusion				Normal Sim(3)				Regularization Sim(3)			
	RMSE (m)↓	AWD (m)↓	CD (m)↓	SCS↓	RMSE↓	AWD↓	CD↓	SCS↓	RMSE↓	AWD↓	CD↓	SCS↓
HKairport03	72.87	5.89	2.45	×	1.03	1.81	0.051	0.54	1.02	1.34	0.048	0.6
HKisland01	37.31	6.76	22.1	×	1.01	0.74	0.05	0.6	0.968	0.73	0.046	0.56
AMtown01	17.08	14.26	10.02	×	0.901	0.786	0.057	0.53	0.893	0.675	0.056	0.52
HKairport01	38.23	7.77	23.0	×	1.03	1.29	0.042	0.5	0.97	1.25	0.032	0.5
AMvalley01	224.9	10.39	2.53	×	2.95	2.49	0.031	0.66	0.942	1.33	0.018	0.64
BrightWall	2.06	0.88	0.101	0.61	0.205	0.66	0.09	0.60	0.196	0.63	0.08	0.57
CBDBuilding	2.47	0.94	0.12	×	0.78	0.74	0.06	0.60	0.49	0.64	0.06	0.60
Company	1.04	0.45	0.064	×	0.29	0.41	0.07	0.54	0.28	0.40	0.06	0.49
Street	1.24	0.56	0.079	×	0.28	0.49	0.064	0.52	0.24	0.46	0.058	0.5

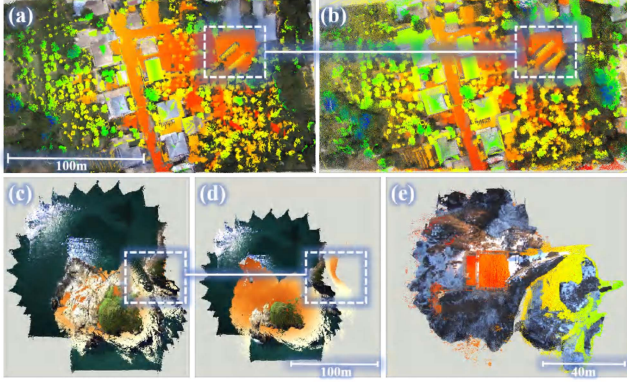


Fig. 8. These figures show registration comparisons on a single session of the MARS-LVIG dataset between normal *Sim(3)* and our regularized *Sim(3)*. The former suffers from excessive scale distortion, while the latter achieves accurate alignment with minimal scale error. (e) presents the registration result of our method on the MUN_FRL dataset.

all sessions. Especially, The SCS measures local geometric consistency, where smaller values indicate more uniform local errors and better overall map quality. Table IV presents a comparison of three variants, two of which use the same number of iterations. Our method consistently outperforms the normal *Sim(3)* across nearly all metrics, whereas the pre-fusion approach performs the worst due to the lack of direct point cloud fusion. Fig. 8 provides a direct explanation for the superior performance of our method. In cross-modal matching with differing point cloud FOVs, unconstrained scale updates in normal *Sim(3)* cause unbounded scale distortion, as seen in (b) and (d). In contrast, the proper incorporation of the regularization term in our method effectively preserves an accurate scale as demonstrated in Fig. 8(a), (c), and (e).

D. Multi-Modal Comparison and Robustness Analysis

We compare our method with representative LIO and LIVO baselines, including FAST-LIO2 and two famous LIVO systems, R³LIVE and FAST-LIVO2, denoted as LIO2, R3, and LIVO2, respectively. As shown in Table V, benefiting from highly accurate localization, LIVO2 achieves the best performance on all color metrics and most geometric metrics. Our method consistently attains second-best results across nearly all color quality metrics and outperforms R3, demonstrating strong potential for RGB reconstruction. For geometric metrics such as AWD and CD#, our method is inferior to LIO2 and LIVO2, mainly due to the limited geometric fidelity of VGGT, but still outperforms R3

TABLE V
COMPREHENSIVE EVALUATION WITH LiDAR METHODS

Data	LIO2	R3	LIVO2	Ours
CD* ↓	AMtown01	×	131.4	64.28
	AMvalley01	×	117.0	59.76
	HKisland01	×	146.1	88.4
CF (dB)↑	AMtown01	×	11.78	17.99
	AMvalley01	×	12.79	18.62
	HKisland01	×	10.86	15.22
CCS ↓	AMtown01	×	1.30	0.77
	AMvalley01	×	0.68	0.38
	HKisland01	×	2.14	1.24
LCR ↑	AMtown01	×	0.84%	94.6%
	AMvalley01	×	0.57%	86.9%
	HKisland01	×	11.6%	98%
AWD (m) ↓	AMtown01	0.74	4.3	0.51
	AMvalley01	0.80	1.92	0.46
	HKisland01	0.38	1.20	0.15
AC (m) ↓	AMtown01	0.045	0.079	0.042
	AMvalley01	0.037	0.041	0.074
	HKisland01	0.078	0.041	0.07
CD# (m) ↓	AMtown01	1.92	212.8	0.72
	AMvalley01	9.28	248.9	1.19
	HKisland01	0.69	197.5	0.29
Points ↑	AMtown01	1.1×10^8	6.4×10^7	3.3×10^8
	AMvalley01	9.1×10^7	6.3×10^7	2.5×10^8
	HKisland01	4.1×10^7	2.8×10^7	1.1×10^8

CD* denotes color distance. CD# denotes Chamfer distance.

across all sequences. Notably, our method produces much denser reconstructions. Across all sequences, we reconstruct billions of points (e.g., 1.5×10^9 — 6.0×10^9), compared to point cloud produced by LIVO2 and R3, yielding up to 4.5 to 125 times higher point counts. This high density provides substantially more correspondence candidates during alignment, enabling our method to achieve the best AC performance among all compared methods.

Furthermore, we evaluate our method and LIVO2 on the self-collect and MUN_FRL datasets. The former features roughly calibrated extrinsics but reliable hardware time synchronization, whereas the latter offers accurately calibrated extrinsics but lacks synchronization. Fig. 9 compares the results of the two methods. It can be observed that in the absence of either time synchronization or precise calibration, LIVO2 suffers significant performance degradation, producing blurred maps. In contrast, our method remains largely unaffected, producing sharper, cleaner, and denser point clouds.

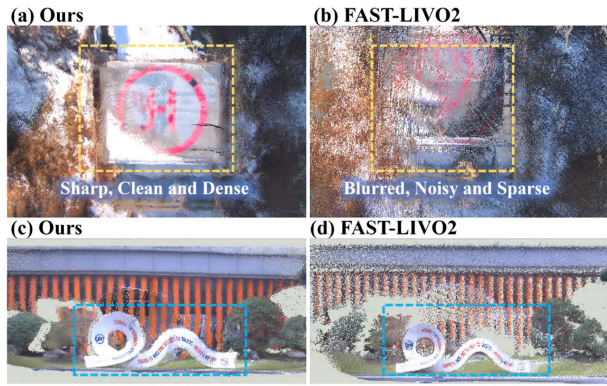


Fig. 9. (a) and (b) compare our method with FAST-LIVO2 on MUN_FRL, while (c) and (d) show mapping results on our dataset TechnologyPark. Benefiting from the correspondence between points and pixels, our reconstructions are significantly denser, sharper, and cleaner.

V. CONCLUSION

This letter presents LiDAR-VGGT, a novel framework that integrates LiDAR with VGGT to achieve large-scale, dense, metric-accurate, and globally consistent reconstruction. We design a coarse-to-fine fusion pipeline that enables LiDAR-inertial odometry to robustly initialize and refine VGGT poses and maps through cross-modal registration. To our best knowledge, this is the first work to combine LiDAR and VGGT, demonstrating that LiDAR fusion can successfully recover the real-world metric attributes of VGGT reconstructions. Future work will explore cross-modal integration by directly embedding LiDAR cues into the VGGT backbone for end-to-end reconstruction.

REFERENCES

- [1] W. Xu, Y. Cai, D. He, J. Lin, and F. Zhang, "FAST-LIO2: Fast direct LiDAR-inertial odometry," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2053–2073, Aug. 2022.
- [2] C. Zheng et al., "FAST-LIVO2: Fast, direct LiDAR-inertial-visual odometry," *IEEE Trans. Robot.*, vol. 41, pp. 326–346, 2025.
- [3] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "VGGT: Visual geometry grounded transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 5294–5306.
- [4] C. Cadena et al., "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.
- [5] Y. Cui et al., "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 722–739, Feb. 2022.
- [6] F. Verdoja, D. Thomas, and A. Sugimoto, "FAST 3D point cloud segmentation using supervoxels with geometry and color for 3D scene understanding," in *Proc. IEEE Int. Conf. Multimedia Expo.*, 2017, pp. 1285–1290.
- [7] X. Lei, M. Wang, W. Zhou, and H. Li, "GaussNav: Gaussian splatting for visual navigation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 4108–4121, May 2025.
- [8] Y. Tian, Y. Chang, F. H. Arias, C. Nieto-Granda, J. P. How, and L. Carlone, "Kimera-Multi: Robust, distributed, dense metric-semantic SLAM for multi-robot systems," *IEEE Trans. Robot.*, vol. 38, no. 4, pp. 2022–2038, Aug. 2022.
- [9] L. Wang et al., "Lemon-mapping: Loop-enhanced large-scale multi-session point cloud merging and optimization for globally consistent mapping," 2025, *arXiv:2505.10018*.
- [10] S. Chen, B. Liu, C. Feng, C. Vallespi-Gonzalez, and C. Wellington, "3D point cloud processing and learning for autonomous driving: Impacting map creation, localization, and perception," *IEEE Signal Process. Mag.*, vol. 38, no. 1, pp. 68–86, Jan. 2021.
- [11] C. Zheng, Q. Zhu, W. Xu, X. Liu, Q. Guo, and F. Zhang, "FAST-LIVO: Fast and tightly-coupled sparse-direct LiDAR-inertial-visual odometry," in *Proc. IEEE/RSJ Intl. Conf. Intell. Robots Syst.*, 2022, pp. 4003–4009.
- [12] J. Lin and F. Zhang, "R3LIVE: A robust, real-time, RGB-colored, LiDAR-inertial-visual tightly-coupled state estimation and mapping package," in *Proc. IEEE Intl. Conf. Robot. Automat.*, 2022, pp. 10672–10678.
- [13] J. Lin and F. Zhang, "R3LIVE++: A robust, real-time, radiance reconstruction package with a tightly-coupled LiDAR-inertial-visual state estimator," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 11168–11185, Dec. 2024.
- [14] S. Hong et al., "GS-LIVO: Real-time LiDAR, inertial, and visual multi-sensor fused odometry with Gaussian mapping," *IEEE Trans. Robot.*, vol. 41, pp. 4253–4268, 2025.
- [15] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "DUST3R: Geometric 3D vision made easy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 20697–20709.
- [16] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3D with MAST3R," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 71–91.
- [17] K. Deng, Z. Ti, J. Xu, J. Yang, and J. Xie, "VGGT-Long: Chunk it, loop it, align it—pushing VGGT's limits on kilometer-scale long RGB sequences," 2025, *arXiv:2507.16443*.
- [18] D. Maggio, H. Lim, and L. Carlone, "VGGT-SLAM: Dense RGB SLAM optimized on the SL (4) manifold," *Proc. Thirty-ninth Annu. Conf. Neural Inf. Process. Syst.*, 2025.
- [19] W. Xu and F. Zhang, "FAST-LIO: A fast, robust LiDAR-inertial odometry package by tightly-coupled iterated Kalman filter," *IEEE Robot. Automat. Lett.*, vol. 6, no. 2, pp. 3317–3324, Apr. 2021.
- [20] X. Xu et al., "RING: Roto-translation invariant gram for global localization on a sparse scan map," *IEEE Trans. Robot.*, vol. 39, no. 6, pp. 4616–4635, Dec. 2023.
- [21] X. Chen et al., "Overlapnet: Loop closing for LiDAR-based SLAM," in *Proc. Robot.: Sci. Syst.*, vol. 16, no. 009, pp. 1–10, 2020.
- [22] C. Shi, X. Chen, J. Xiao, B. Dai, and H. Lu, "Fast and accurate deep loop closing and relocalization for reliable LiDAR SLAM," *IEEE Trans. Robot.*, vol. 40, pp. 2620–2640, 2024.
- [23] C. Jiang et al., "RISED: Accurate and efficient RGB-colored mapping using image selection and point cloud densification," in *Proc. IEEE Intl. Conf. Robot. Automat.*, 2025, pp. 3277–3283.
- [24] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing scenes as neural radiance fields for view synthesis," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [25] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 1–14, 2023.
- [26] T. Tao et al., "LiDAR-NeRF: Novel LiDAR view synthesis via neural radiance fields," in *Proc. Int. Conf. Multimedia*, 2024, pp. 390–398.
- [27] Y. Tao, Y. Bhalgat, L. F. T. Fu, M. Mattamala, N. Chebrolu, and M. Fallon, "SILVR: Scalable LiDAR-visual reconstruction with neural radiance fields for robotic inspection," in *Proc. IEEE Intl. Conf. Robot. Automat.*, 2024, pp. 17983–17989.
- [28] X. Lang et al., "Gaussian-LIC: Real-time photo-realistic SLAM with Gaussian Splatting and LiDAR-inertial-camera fusion," in *Proc. IEEE Intl. Conf. Robot. Automat.*, 2024, pp. 8500–8507.
- [29] H. Wang and L. Agapito, "3D reconstruction with spatial memory," in *Proc. 2025 Int. Conf. 3D Vis.*, 2024, pp. 78–89.
- [30] Z. Tang et al., "MV-2053DUST3R: Single-stage scene reconstruction from sparse views in 2 seconds," in *Proc. Comput. Vis. Pattern Recognit. Conf.*, 2024, pp. 5283–5293.
- [31] Y. Liu et al., "SLAM3R: Real-time dense scene reconstruction from monocular RGB videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 16651–16662.
- [32] J. Yang et al., "FAST3R: Towards 3D reconstruction of 1000 images in one forward pass," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2025, pp. 21924–21935.
- [33] S. Umeyama, "Least-squares estimation of transformation parameters between two point patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 13, no. 4, pp. 376–380, Apr. 1991.
- [34] F. Dellaert, "Factor graphs and GTSAM: A hands-on introduction," Georgia Inst. Technol., Atlanta, GA, USA, Tech. Rep., vol. 2, no. 4, 2012.
- [35] H. Li et al., "MARS-LVIG dataset: A multi-sensor aerial robots SLAM dataset for LiDAR-visual-inertial-GNSS fusion," *Int. J. Robot. Res.*, vol. 43, no. 8, pp. 1114–1127, 2024.
- [36] R. G. Thalagala, O. De Silva, A. Jayasiri, A. Gubbels, G. K. Mann, and R. G. Gosine, "MUN-FRL: A visual-inertial-LiDAR dataset for aerial autonomous navigation and mapping," *Int. J. Robot. Res.*, vol. 43, no. 12, pp. 1853–1866, 2024.
- [37] X. Hu et al., "MapEval: Towards unified, robust and efficient SLAM map evaluation framework," *IEEE Robot. Automat. Lett.*, vol. 10, no. 5, pp. 4228–4235, May 2025.