International Journal of Geo-Information

MDPI

Article

VGGT-Geo: Probabilistic Geometric Fusion of Visual Geometry Grounded Transformer Priors for Robust Dense Indoor SLAM

Kai Qin [1,2], Jing Li [1,3,*], Sisi Zlatanova [4], Haitao Wu [1], Hao Wu [5], Yin Gao [5], Dingjie Zhou [6], Yuchen Li [1,2], Sizhe Shen [1,2], Xiangjun Qu [1,2], Zhenxin Zhang [7], Banghui Yang [1,8] and Shicheng Xu [1]

1 Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China; qinkai21@mails.ucas.ac.cn (K.Q.)
2 School of Electronic, Electrical and Communication Engineering, University of Chinese Academy of Sciences, Beijing 101408, China
3 International Research Center of Big Data for Sustainable Development Goals, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
4 GRID, School of Built Environment, University of New South Wales, Sydney, NSW 2033, Australia; s.zlatanova@unsw.edu.au
5 National Geomatics Center of China, Beijing 100830, China
6 Yunnan Provincial Institute of Surveying and Mapping, Kunming 650011, China
7 College of Resource Environment and Tourism, Capital Normal University, Beijing 100048, China
8 National Engineering Research Center for Geoinformatics, Aerospace Information Research Institute, Chinese Academy of Sciences, Beijing 100094, China
* Correspondence: lijing202115@aircas.ac.cn

Abstract

With the rapid evolution of Digital Twins and Embodied AI, achieving fast, dense, and high-precision 3D perception in unknown environments has become paramount. However, existing Visual SLAM paradigms face a critical dilemma: geometry-based methods often fail in texture-less areas due to feature scarcity, while learning-based approaches frequently suffer from scale drift and unphysical deformations. To bridge this gap, we propose VGGT-Geo, a novel SLAM system that synergizes generative priors from Large Foundation Models with multi-modal geometric optimization. Distinguishing itself from simple cascaded architectures, we construct a Probabilistic Geometric Fusion framework, consisting of (1) Generative Warm-start, leveraging the holistic scene understanding capabilities of the VGGT, (2) Confidence-Aware Optimization to extract dense features via DINOv3 and predict their confidence map, and (3) a Multi-Modal Constraint Closure that fuses point-line features and metric depth priors to constrain rotational Degrees of Freedom in Manhattan Worlds. We conducted systematic evaluations on TUM, Replica, Tanks and Temples, and a challenging self-collected dataset featuring extreme lighting and texture-less walls. Experimental results demonstrate that VGGT-Geo exhibits superior robustness and accuracy in unseen environments. On our most challenging dataset, it achieves an Absolute Trajectory Error of 4–5 cm and a Relative Rotation Error of 0.79°, outperforming current state-of-the-art methods by approximately 50% in trajectory accuracy. This study validates that synergizing the intuition of Large Foundation Models with geometric rigor is a viable path toward next-generation robust SLAM.

Keywords: 3D reconstruction; VGGT; SLAM; indoor modeling; digital twin

Check for updates

Academic Editors: Wolfgang Kainz and Huayi Wu

Received: 24 December 2025
Revised: 6 February 2026
Accepted: 14 February 2026
Published: 16 February 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Dense Simultaneous Localization and Mapping (SLAM) constitutes a foundational capability for rapid 3D data collection in fields such as Augmented Reality (AR), robot navi-

gation, and Digital Twins. Traditional SLAM systems typically adopt a multi-stage pipeline consisting of "camera pose optimization + dense/sparse mapping": they first solve for accurate trajectories via feature matching and Bundle Adjustment (BA), followed by dense model generation utilizing multi-view geometry or Truncated Signed Distance Function (TSDF) [1] fusion. While this paradigm excels in precision, it incurs high computational complexity, requires cumbersome parameter tuning, and imposes a heavy dependency on hardware resources. For resource-constrained embedded devices or real-time interactive applications, traditional methods relying on iterative optimization often struggle to meet the dual requirements of "second-level latency and centimeter-level precision".

Early dense SLAM approaches relied heavily on depth sensors such as DGS-SLAM and MVS-SLAM [2,3]. By leveraging depth cameras to directly acquire scale-consistent depth observations, these systems bypassed the inherent challenges of traditional multi-view geometry. However, their reconstruction quality is largely contingent upon the precision of the depth sensors. Recently, dense monocular SLAM methods based on Neural Implicit Representations [4] and 3D Gaussian Splatting (3D GS) [5] have enabled high-level scene expression. Nevertheless, these methods are often confined to offline processing, failing to meet real-time reconstruction standards; furthermore, their reconstruction accuracy is fundamentally limited by the quality of the initial point cloud.

In recent years, end-to-end learning-based SLAM methods (e.g., DUSt3R, Spann3R, SLAM3R) [6–8] have emerged, directly outputting dense point clouds via feed-forward networks. While these methods bypass the explicit Bundle Adjustment (BA) [9] process and tedious geometric solving and validation [10], they continue to suffer from limited local window constraints, significant long-sequence drift, and insufficient global consistency.

Firstly, Vision Geometry Ground Transformer (VGGT) [11] methods predominantly rely on the "two-frame depth prediction to incremental fusion" paradigm. The networks learn geometric consistency only within the scope of local disparity, lacking global geometric closure constraints across frames or loops. Once depth estimation is misled by texture-less or repetitive regions, these errors are unconditionally integrated into the global model and accumulate over subsequent frames, eventually leading to geometric artifacts such as floating blocks and double-layered surfaces (as shown in Figure 1).

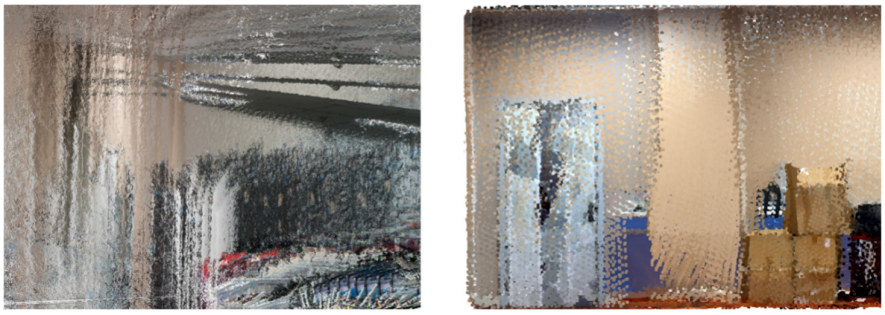


Figure 1. Qualitative Comparison of Multi-frame Point Cloud Accumulation. (**left**) VGGT (Baseline): The point cloud generated by directly accumulating multi-frame predictions from the raw VGGT model, showing inconsistent depth predictions across frames. (**right**) Ours (VGGT-Geo): The resulting point cloud eliminates the "ghosting" effects, exhibiting single-layered, sharp surfaces and correct structural consistency.

Secondly, unlike traditional SLAM which relies on Bag-of-Words (BoW) [12] or geometric retrieval for relocalization and global graph optimization, end-to-end networks often store historical states "implicitly" via external memory or recurrent networks. This makes it difficult to accurately determine global correspondences between frames, resulting in insufficient loop closure capabilities.

Finally, end-to-end methods directly regress point coordinates; without global constraints on cross-scene scale variations, "streaking artifacts" tend to occur in distant regions. Furthermore, confidence learning struggles to eliminate outliers, leaving floating noise on smooth surfaces.

In contrast our VGGT-Geo endows the system with the capability to recover fundamental visual geometry in unknown scenes and thus outperform previous approaches that simply treat network outputs as observations. The main contributions of this paper are summarized as follows:

1. A deeply coupled framework synergizing generative priors and geometric optimization: We leverage the generative capabilities of a VGGT to provide "Warm-start" initialization for the non-convex backend optimization, placing the solver within the basin of convergence. Simultaneously, we utilize dense features from DINOv3 for fine-grained association. This "Coarse Generation + Fine Optimization" dual-branch architecture effectively resolves the contradiction between the initialization difficulties of traditional methods and the poor geometric consistency of pure learning-based methods.
2. A Confidence-Aware Multi-Modal Backend: Targeting indoor texture-less and structured environments, we design a joint factor graph incorporating sparse points, semantic line segments, and metric depth priors. The core innovation lies in leveraging the confidence map predicted by large models to dynamically re-weight residuals across different modalities. This allows the system to adaptively utilize line features to constrain rotational Degrees of Freedom (e.g., in Manhattan Worlds) and employ MoGe-2 priors as adaptive Anchors, thereby strictly eliminating monocular scale drift.
3. SOTA performance and robust Zero-Shot generalization in challenging real-world scenarios: We conducted extensive evaluations on benchmarks ranging from standard datasets (TUM, Replica) to challenging self-collected scenarios (featuring long corridors and white walls). Rigorous verification against LiDAR Ground Truth demonstrates that VGGT-Geo not only outperforms existing SOTA methods in standard scenes but also maintains centimeter-level trajectory consistency in Out-of-Distribution (OOD) scenarios where baseline methods completely fail.

2. Related Work

In recent years, the field of Visual SLAM has undergone a profound paradigm shift from "rule-based hand-crafted designs" to "data-driven deep learning" paradigms. We review related advancements across three dimensions: foundation model priors, hybrid SLAM architectures, and the application of geometric constraints.

2.1. Foundation Models for 3D Vision

Traditional SLAM systems rely on hand-crafted features [13–15] (e.g., ORB, SIFT) and geometric models. While these methods excel in accuracy within controlled environments, they exhibit poor robustness in texture-less or illumination-varying scenarios. Transformers, holistic geometric regression models, represented by DUSt3R, have changed the game. DUSt3R discards traditional epipolar geometry constraints and utilizes Transformers to directly regress dense point clouds (Point maps) from image pairs, demonstrating astonishing Zero-shot generalization capabilities. Subsequently, MASt3R further enhanced feature association capabilities by introducing a matching head. VGGT introduced more comprehensive generative priors, capable of parallelly outputting camera poses, intrinsics, and dense depth. However, these foundation models are typically designed for "image pairs" or "short windows". When applied directly to long video sequences, due to the lack of Global Consistency constraints, inter-frame prediction errors accumulate unboundedly,

leading to "scale drift" and even severe geometric distortion of the scene. This indicates that relying solely on feed-forward predictions from large models is insufficient to construct a reliable, globally consistent SLAM system.

2.2. Hybrid SLAM: Bridging Learning and Optimization

To address the drift issue inherent in pure feed-forward methods, researchers have begun exploring hybrid architectures combining "Deep Priors" with "Traditional Optimization," which has become the mainstream direction for current State-of-the-Art (SOTA) approaches. Feed-forward Enhanced: SLAM3R adopts a "local prediction-global fusion" strategy, utilizing sliding windows to generate local point clouds and incrementally register them. Although it introduces smoothing terms, it remains fundamentally a pure feed-forward system, lacking an explicit Bundle Adjustment (BA) [9] backend. Experiments show that SLAM3R is prone to geometric fractures or "double-wall" phenomena in long corridors or large loop scenarios. Optimization Enhanced: MASt3R-SLAM [16] represents the current state-of-the-art. It utilizes MASt3R for frontend matching and introduces point-based graph optimization in the backend. Although incorporating BA significantly reduces drift, MASt3R-SLAM is overly dependent on the matching quality of point features. In weak-texture regions such as indoor white walls or glass curtains, point features are often sparse and unreliable, leading to optimization degradation. On the other hand, VGGT-SLAM [17] attempts global optimization of submaps on the Lie Group SL (4) manifold. However, our empirical tests reveal that this method is extremely sensitive to the quality of frontend initialization. In unknown real-world scenarios that differ significantly from the training distribution (Out-of-Distribution), the VGGT-SLAM frontend often generates numerous outliers, causing backend optimization divergence.

Unlike the mentioned methods, VGGT-Geo does not attempt to replace geometry with networks but rather dedicates itself to quantifying geometric confidence via networks. We not only utilize VGGT for initialization but also introduce a Probabilistic Multi-Modal BA, thereby finding a new equilibrium between generalization and precision.

2.3. Geometric Priors and Uncertainty Modeling

In indoor man-made environments (Manhattan World [18]), point features alone are often insufficient to constrain all Degrees of Freedom (DoF) [19]. Line Constraints: Traditional methods like PL-SLAM [20] have long demonstrated the effectiveness of Line Features in structured scenes. However, traditional line detection (e.g., LSD [21]) is not only prone to fragmentation but also struggles to describe the "confidence" of line segments. Depth Constraints: With the maturation of monocular depth estimation (e.g., MoGe-2 [22]), introducing depth maps as regularization terms has become standard practice to resolve monocular scale ambiguity. However, existing combination methods mostly employ Hard Constraints, assuming extracted lines or depth values are perfectly trustworthy. This is error prone in practice because incorrect line matching or depth artifacts can directly destroy optimization results.

3. Method

3.1. System Overview and Probabilistic Formulation

Our proposed VGGT-Geo system is designed to tackle the inherent challenges of geometric ambiguity and cumulative drift in monocular visual SLAM. Distinguishing itself from traditional pure geometric approaches or purely end-to-end learning methods, we introduce a novel paradigm of Probabilistic Geometric Fusion (Figure 2).

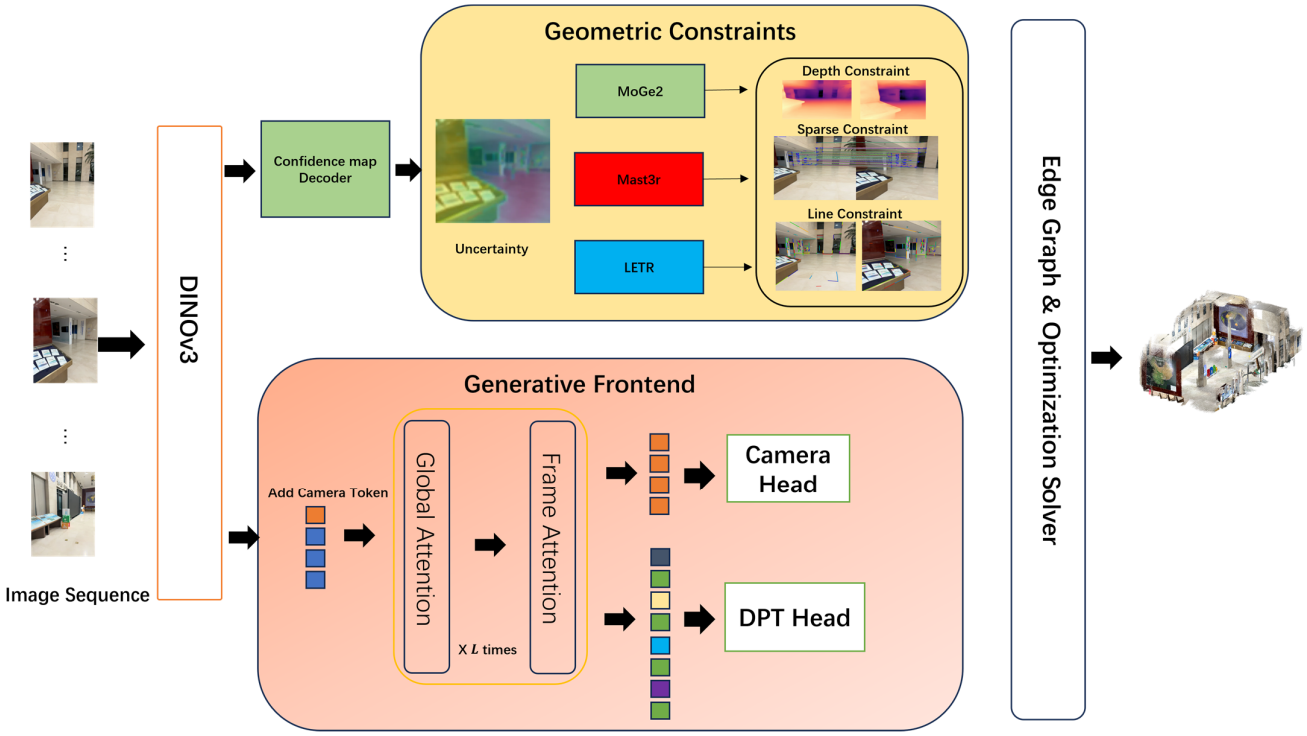


Figure 2. System Overview of VGGT-Geo. Given a RGB video stream, our pipeline leverages a Generative Frontend to provide warm-start initialization, followed by a Backend that performs geometric optimization using depth, points, and lines, significantly enhancing the robustness and generalization of dense scene reconstruction.

We formulate the SLAM problem as a joint posterior probability estimation of the system state $\mathcal{X}$ over a Factor Graph. Let the input image sequence be $\mathcal{I} = \{I_1, \ldots, I_N\}$. The system state variables $\mathcal{X}$ to be estimated are defined as:

$$\mathcal{X} = \{T, \mathcal{K}, \mathcal{D}\} \tag{1}$$

where $\mathcal{T} = \{T_i \in SE(3)\}_{i=1}^{N}$ denotes the set of camera poses, $\mathcal{K}$ represents the camera intrinsics (assumed to be globally shared), and $\mathcal{D} = \{D_i\}_{i=1}^{N}$ denotes the dense inverse depth map for each frame.

Based on Bayes' rule, given the observation set $\mathcal{Z}$ (comprising image features and pre-trained priors), the optimal state estimation $\mathcal{X}^*$ corresponds to the Maximum A Posteriori (MAP) [23] estimation:

$$\mathcal{X}^* = \underset{\mathcal{X}}{argmax}\ P(\mathcal{X}|\mathcal{Z}) \propto \underbrace{P(\mathcal{Z}|\mathcal{X})}_{Geometric\ Likelihood} \cdot \underbrace{P(\mathcal{X})}_{Generative\ Prior} \tag{2}$$

We transform this probability maximization problem into the minimization of a negative log-likelihood energy function $E_{total}(\mathcal{X})$:

$$E_{total}(\mathcal{X}) = \sum_{(i,j)\in\mathcal{E}} E_{geo}(i,j) + \lambda \sum_{i\in\mathcal{V}} E_{prior}(D_i) \tag{3}$$

Here, $E_{geo}(i,j)$ represents the geometric observation term constructed based on DINOv3 features, incorporating sparse point E_{sparse} and dense line E_{line} constraints to enforce multi-view geometric consistency; E_{prior} is the depth prior regularization term based on MoGe-2, serving as a Metric Scale Anchor, and regarding initialization (implicit in the starting values), we leverage the predictions generated by VGGT as the initial guess for

non-linear optimization, ensuring the solver resides within the Basin of Convergence of the global optimum. In this manner, VGGT-Geo achieves a deep coupling between generative foundation model priors and traditional multi-view geometric constraints.

Finally, regarding initialization, we leverage the holistic predictions generated by VGGT as the "Warm-start" initial guess for the non-linear optimization. This ensures that the solver resides within the Basin of Convergence of the global optimum, achieving a deep coupling between generative foundation model priors and traditional multi-view geometric constraints.

3.2. Generative Frontend: Initialization & Confidence-Aware Perception

To facilitate the backend optimization, we design a dual-branch generative frontend dedicated to "coarse trajectory initialization" and "fine-grained feature perception," respectively.

a. Holistic Initialization via VGGT:

The success of non-linear optimization (e.g., Bundle Adjustment) is heavily contingent upon the quality of initialization. Traditional SLAM methods often struggle to provide reliable initial estimates in scenarios featuring pure rotation, large baselines, or texture-less regions. We leverage the pre-trained VGGT as the system's startup engine. For a given input sliding window of frames, VGGT parallelly regresses the initial camera extrinsics T^{vggt}, intrinsics K^{vggt}, and the corresponding dense point clouds P^{vggt}. Although the direct outputs of VGGT lack metric precision over long sequences, by learning perspective laws from massive datasets, it provides initial solutions $\mathcal{X}_{init}$ with correct global structures. We utilize the depth maps projected from P^{vggt} as initial guesses for the state variables $\mathcal{D}$, thereby avoiding local minima traps associated with "cold-start" optimization.

b. Perceptual Confidence map via DINOv3 Features:

Constructing high-precision geometric constraints necessitates robust feature association. We adopt DINOv3 as our dense feature extractor. Beyond high-dimensional semantic descriptors, we leverage the attention mechanism inherent in the Vision Transformer architecture to derive a pixel-wise Confidence Map ($\Omega \in [0,1]^{H\times W}$). We formulate Ω as a measure of Feature Saliency and Matching Reliability. High confidence values typically correspond to structurally distinct regions (e.g., edges, corners) where the attention heads focus intensely, while low values indicate ambiguous areas (e.g., texture-less walls, specular highlights). This metric serves as a robust, data-driven prior to guiding the backend optimization, allowing the system to dynamically down-weight noisy observations in the energy function.

3.3. Construction of Geometric Constraints

Let $\mathcal{V}$ be the set of keyframes, and graph edges $\mathcal{E}$ represent frame pairs with high contribution scores. The camera extrinsic parameter of the i-th frame is $T_i \in SE(3)$, and the intrinsic parameter is $k = (f_x, f_y, c_x, c_y)$. The depth field of the i-th frame is denoted as D_i. In the core part of VGGT-Geo, we solve the Bundle Adjustment (BA) problem for the following variables:

$$E_{total}(\{T_i\},\{D_i\},k) = \sum_{(i,j)\in\mathcal{E}} \left[E_{line}(T_i,T_j,D_i,k) + E_{sparse}(T_i,T_j,D_i,k)\right] + \alpha \sum_{i\in\mathcal{V}} E_{depth}(D_i) \quad (4)$$

where E_{line} utilizes line matching constraints between frames i and j, E_{sparse} utilizes sparse key point matching constraints, and E_{depth} is dense depth regularization. These terms ensure the global consistency and robustness of pose estimation through weighted summation. The hyperparameter α (corresponding to the λ in Equation (3)) controls the strength of the adaptive Scale Anchor. We employ the Levenberg-Marquardt (LM) solver to minimize this energy function. The details of these three constraints are described below.

3.3.1. Confidence-Aware Line Constraints

Indoor man-made environments (Manhattan World) typically abound with linear edges that are structurally distinct yet texturally sparse (e.g., wall corners, door frames, and ceiling boundaries). In such regions, point features are often difficult to extract or degenerate into a single direction, resulting in reduced observability of the system state, particularly in constraining the camera's rotational Degrees of Freedom (DoF). To bridge this observability gap, we construct geometric constraints based on line features.

Distinguishing ourselves from traditional methods that require complex 3D line parameterizations (such as Plücker coordinates), we adopt a more efficient Implicit 3D Line Constraint strategy. Let l_i denote a 2D line segment (Figure 3) which we used LETR decoder [24] detected in the reference frame I_i. While the corresponding 3D line is physically determined by the inverse depths of its endpoints, during optimization, we implicitly index it utilizing the dense inverse depth map D_i from DINOv3. Let l_j denote the corresponding matched line segment in the current frame I_j; we parameterize it using the normalized plane equation:

$$a_j u + b_j v + c_j = 0, s.t. a_j^2 + b_j^2 = 1 \tag{5}$$

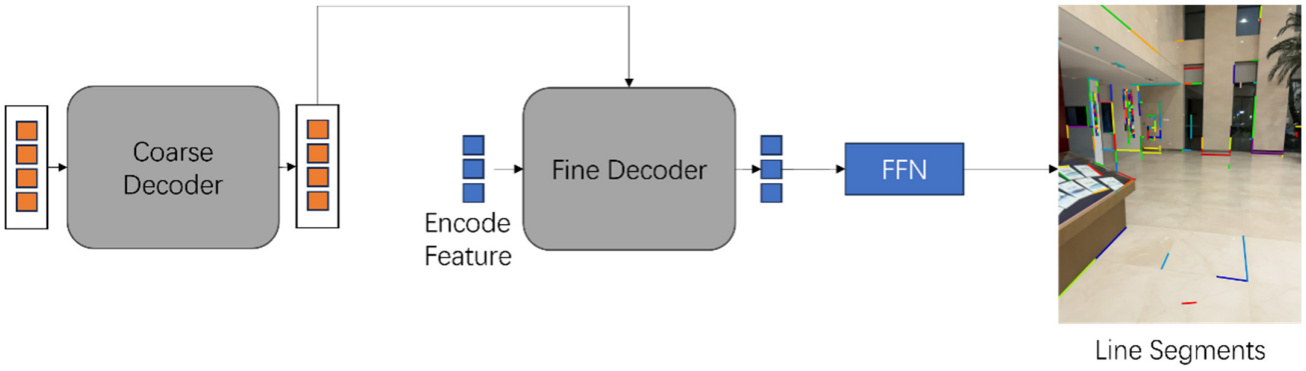


Figure 3. Illustration of the LETR-based Line Extraction Decoder. This component is responsible for detecting semantic lines to constrain rotational degrees of freedom in the Confidence-Aware Line Constraints module.

To construct residuals, we uniformly sample a set of points $u_{i,s} \in \mathcal{S}(l_i)$ along the reference line segment l_i with a step size Δs. We transform these sampled points to the pixel coordinate system of the current frame I_j via the relative pose $T_{ij} = T_j T_i^{-1}$ and inverse depth $D_i(u_{i,s})$, obtaining the projected points $u'_{i,s}$.

The line reprojection residual $r_{line}(u_{i,s})$ is defined as the algebraic distance from the projected point to the target line:

$$r_{line}(u_{i,s}) = a_j u'_{i,s} + b_j v'_{i,s} + c_j \tag{6}$$

Traditional geometric methods operate under the assumption that all extracted line segments are equally reliable—an assumption that often fails in real-world scenarios containing shadow boundaries or dynamic occlusions. To address this, we leverage the feature decoding head of DINOv3 to estimate a Reliability Weight Ω for each pixel along the line segment. We incorporate this weight into the energy function, transforming the pure geometric error into a probabilistically weighted loss function:

$$E_{line}(i,j) = \sum_{m \in \mathcal{L}_{ij}} \sum_{s \in \mathcal{S}(l_i^m)} \rho\left(\Omega(u_{i,s}) \cdot \| r_{line}(u_{i,s}) \|^2\right) \tag{7}$$

where $\rho(\cdot)$ represents the Huber kernel function. This mechanism enables the optimizer to adaptively focus on geometric edges that are semantically clear and structurally stable, while automatically suppressing edges that are texturally blurred or ambiguous, thereby significantly enhancing the robustness of pose estimation in low-texture regions.

3.3.2. Joint Optimization with Sparse Keypoints

While line features enhance rotational constraints, they suffer from an inherent limitation regarding translational constraints along the line direction (i.e., the aperture problem). To secure isotropic positional constraints, we retain sparse keypoints match obtained by MASt3R as the foundation of geometric optimization (Figure 4).

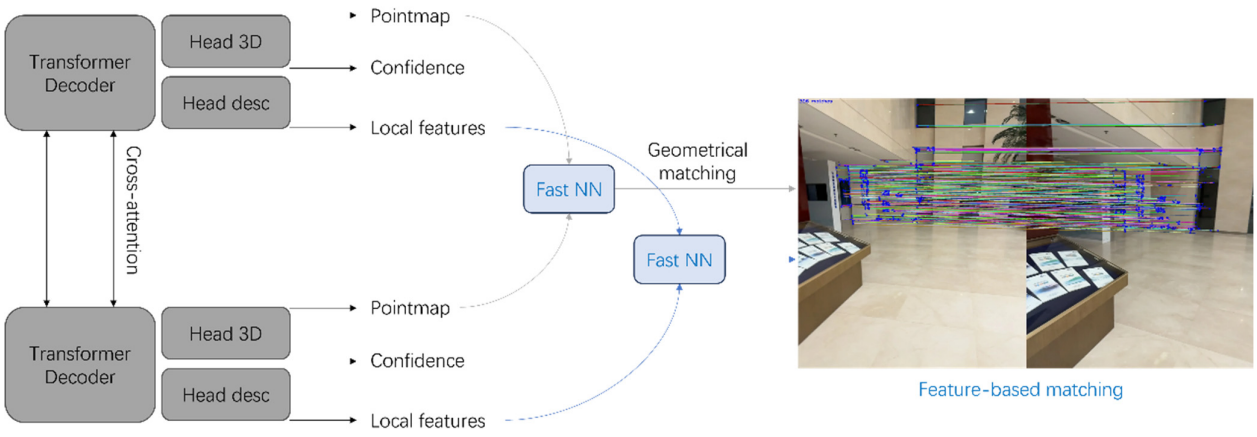


Figure 4. We utilize the MASt3R matching head to establish robust point correspondences between frames, providing essential isotropic positional constraints that complement the directional constraints of line features, thereby resolving the aperture problem.

To ensure computational efficiency and preserve the consistency of dense reconstruction, we do not explicitly maintain independent 3D landmarks in the backend. Instead, we directly optimize the inverse depth maps of keyframes. This implies that the sparse point constraints (E_{sparse}), line constraints (E_{line}), and depth priors (E_{depth}) share the same set of geometric state variables, thereby guaranteeing the strict coupling of multi-modal constraints within the physical space.

Let c_{ij} denote the set of matched point pairs between frame I_i and I_j. For a matched pair (u_i, u_j), the reprojection residual is defined as:

$$r_{sparse}(u_i) = \pi(T_j, T_i, D_i(u_i), u_i) - u_j \tag{8}$$

where $\pi(\cdot)$ represents the pinhole camera projection function. To mitigate the impact of unreliable feature matches common in indoor environments, we incorporate the frontend confidence scores directly into the bundle adjustment. We formulate the sparse point constraint as a Confidence-Weighted Least Squares problem. The energy term is defined as:

$$E_{sparse}(i,j) = \sum_{(u_i,u_j)\in C_{ij}} \rho(\Omega(u_i)' \cdot \| r_{sparse}(u_i) \|^2) \tag{9}$$

where $\rho(\cdot)$ represents the Huber robust kernel. In our implementation, we map the normalized confidence score to the optimization weight via $\Omega(u_i)' = \alpha\Omega(u_i) + \varepsilon$, effectively down-weighting residuals in regions with low perceptual confidence. This strategy enables the optimizer to "softly" reject outliers in texture-less regions without requiring hard thresholds or complex probabilistic inference.

3.3.3. Adaptive Depth Regularization with Metric Priors

Recovering the true Metric Scale is a prerequisite for applications in indoor robot navigation and Digital Twins. However, monocular vision suffers from inherent Gauge Freedom, rendering pure geometric optimization highly susceptible to scale drift. To break this ill-posedness, we introduce the absolute depth map $D_i^{prior}(u)$ provided by the pre-trained monocular depth estimation network (MoGe-2) as the system's Scale Anchor.

To seamlessly integrate this data-driven prior into the geometric optimization framework, we design a confidence-weighted depth regularization term. For a pixel u in keyframe i, the energy function is defined as:

$$E_{depth}(i) = \sum_{u \in \Omega_i} \rho(\Omega(u_i) \cdot \| D_i(u) - D_i^{prior}(u) \|^2) \tag{10}$$

where $D_i(u)$ denotes the state variable to be optimized (inverse depth). In standard monocular SLAM, scale errors accumulate indefinitely over time. By explicitly penalizing the deviation between the estimated state $D_i(u)$ and the metric prior D_i^{prior} across the sliding window, we effectively suppress this accumulation. This formulation endows the system with an adaptive behavior to achieve the best of both worlds.

In texture-less regions (Low Ω): Where geometric observations are vanishingly weak (e.g., white walls), the optimizer is primarily constrained by the prior. The system relies on the Soft Anchor to lock the absolute scale and prevent surface collapse, effectively mitigating the drift typical of monocular methods. In texture-rich regions (High Ω): The dominance of high-fidelity geometric residuals increases. The formulation allows the optimized depth $D_i(u)$ to deviate elastically from the prior $D_i^{prior}(u)$ to satisfy strict multi-view geometric consistency. This ensures that potentially noisy network predictions are corrected by reliable visual observations, preserving local metric accuracy.

3.4. Backend: Graph Construction and Optimization

To balance the trade-off between reconstruction accuracy and real-time performance, we do not optimize every incoming frame. Instead, we dynamically construct a Factor Graph based on keyframe selection and maintain a bounded computational load using a sliding window strategy.

We employ a strategy based on relative motion thresholds to select informative keyframes from the input video stream. Let T_{last} be the pose of the last inserted keyframe and T_{cur} be the estimated pose of the current frame (initialized by the VGGT frontend). A current frame is selected as a new keyframe K_{new} if the relative motion satisfies either of the following criteria:

$$\| \Delta t \| > \tau_t \quad or \quad \| Log(\Delta R) \| > \tau_R \tag{11}$$

where Δt and ΔR represent the relative translation and rotation between T_{cur} and T_{last}. τ_t and τ_R are pre-defined thresholds. This strategy ensures that the system captures sufficient parallax for triangulation while avoiding redundant computations in stationary scenarios. Additionally, frames with a tracking inlier ratio below a distinct threshold are explicitly forced as keyframes to prevent tracking failure in difficult regions.

Once a new keyframe K_i is added to the graph, we establish edges $\mathcal{E}$ to enforce geometric constraints. Unlike simple temporal chaining, we construct a covisibility graph based on feature correspondence to strengthen global consistency. For the new keyframe K_i, we retrieve a set of candidate reference keyframes K_j from the optimization window. We utilize the matching head of MASt3R/DINOv3 to calculate the number of valid feature matches $N_{match}^{(i,j)}$ between K_i and each candidate K_j. An edge (i, j) is added to the factor graph only if the number of matches exceeds a significance threshold:

$$(i,j) \in \mathcal{E} \ \ if \ \ N_{match}^{(i,j)} > \tau_{match} \tag{12}$$

This criterion defines the "contribution score" mentioned in Equation (4), ensuring that optimization is driven only by reliable multi-view connections.

To maintain constant-time complexity for long-term operation, we employ a Sliding Window optimization strategy.

We maintain an optimization window of fixed size N. The total energy function (Equation (4)) is minimized over the active state variables within this window. When a new keyframe is inserted and the window size exceeds N, we apply a marginalization strategy based on the Schur Complement. Specifically, we marginalize out the oldest keyframe and its corresponding measurements, converting their information into a linearized prior term (comprising the Hessian matrix and gradient vector) which is added to the cost function. This mechanism effectively preserves historical geometric information to mitigate drift while keeping computational resources bounded.

The resulting non-linear least squares problem is solved using the Levenberg-Marquardt (LM) algorithm. We choose LM over the standard Gauss-Newton method for its superior stability and robustness. The damping term in LM effectively handles the potential ill-conditioning of the Hessian matrix and the inherent gauge freedom in monocular systems, ensuring reliable convergence even under challenging initialization conditions.

4. Experiment

4.1. Experiments

4.1.1. Datasets

To comprehensively evaluate the generalization capability of VGGT-Geo across diverse domains, we employ a three-tiered dataset strategy: Standard Academic Benchmarks: Comprising TUM RGB-D (real-world indoor) [25] and Replica (high-fidelity synthetic) [26] datasets, serving to quantitatively assess the system's limit accuracy within controlled environments. Large-Scale Generalization Stress-Test: Utilizing Tanks and Temples (Barn, Palace, Temple) [27] to challenge the system's trajectory consistency in large-scale scenes. Challenging Self-Collected Dataset (Ours-Real): This constitutes the core testing ground of this work. We captured office building scenarios using a handheld RGB camera, featuring distinct Geometric Degeneracy characteristics such as long corridors, glass curtain walls, and texture-less white walls.

To ensure the rigor of the evaluation on our self-collected dataset, we eschewed traditional COLMAP-based [28] pseudo-ground truth. Instead, we employed a Leica P50 3D terrestrial laser scanner for high-fidelity ground truth generation. The dataset was collected across an eight-story office building, covering multiple floors and diverse indoor topologies. The overall scene spans a significantly large area, with the maximum continuous trajectory distance reaching approximately 200 m.

We utilized a combination of traverse survey methods and resection methods across a total of 40 scanning stations. Multiple control points were established outdoors to guide the indoor scans, strictly controlling indoor accuracy. The input images were captured at 4K resolution (4096×3072). Crucially, the dataset features scenarios designed to challenge the system:

Long Corridors: Characterized by low texture and repetitive visual patterns, testing robustness against geometric degeneracy.

Glass Walls & White Walls: Introducing challenges from specular reflection and texture-less surfaces.

Ultimately, by rigorously adhering to professional surveying techniques, we achieved an exceptional average point cloud error of ± 2 mm, confirming the high quality of our

geometric benchmark. We employ a multi-sensor system combining LiDAR and an RGB camera to reconstruct high-precision geometry. To ensure comprehensive evaluation, the data collection spans both outdoor and indoor environments across multiple floors (8 stories) of an office building, explicitly incorporating complex vertical topologies such as staircases. These reconstructions serve as the geometric ground truth benchmark for evaluating our self-collected dataset.

4.1.2. Baselines & Metrics

We benchmark VGGT-Geo against three representative classes of state-of-the-art methods: MASt3R-SLAM: The current SOTA, representing the paradigm of strong optimization based on point features. SLAM3R: Representing the "pure end-to-end feed-forward" paradigm without explicit BA. VGGT-SLAM: Representing the "generative paradigm based on manifold optimization.

We employ standard Absolute Trajectory Error (ATE) [29] to evaluate global consistency, Relative Rotation Error (RRE) [29] to assess local geometric constraint capabilities and calculate the Root Mean Square Error (RMSE) [29] between reconstructed depth and the scanner ground truth to verify scale accuracy.

4.1.3. Implementation Details

Our system is implemented in PyTorch 1.2.1 and executed on a desktop PC equipped with an NVIDIA RTX 5090 GPU (NVIDIA Corporation, USA). To ensure real-time performance and evaluate zero-shot generalization, both the DINOv3 (ViT-L/16) [30] feature extractor and the LETR line detector are loaded with official pre-trained weights and kept strictly frozen without online fine-tuning. Input images are resized to 640×480 pixels for inference. For the backend, we employ a sliding window optimization strategy with a size of $N = 10$, using the Levenberg-Marquardt algorithm to solve the nonlinear least squares problem; when the window is full, the oldest keyframe is marginalized using the Schur complement to preserve historical priors. New keyframes are inserted based on relative motion thresholds of translation $\tau_t > 0.15$ m and rotation $\tau_R > 5^\circ$. If the number of valid tracked features falls below $\tau_{match} = 50$. The depth weight α is set 2.0 to strictly enforce the global scale anchor provided by MoGe-2, thereby preventing monocular scale drift, while still retaining sufficient flexibility for multi-view geometric constraints to refine local structural details. We utilized Huber robust loss with a threshold of $\delta = 1.5$ pixels for all residuals. Finally, line segments are sampled with a step size of $\Delta s = 10$ pixels to balance computational efficiency with constraint density.

4.2. Evaluation

4.2.1. Results on Self-Capture Datasets

We present a comprehensive comparison of different methods (including MASt3R-SLAM, SLAM3R, VGGT-SLAM, and our proposed method). Specifically, we evaluate the performance from two dimensions: we qualitatively assess the structural completeness of the reconstructed point clouds (Figure 5) and quantitatively evaluate the geometric consistency via trajectory and depth error metrics (Table 1). As illustrated in the qualitative results below, our method demonstrates superior performance when dealing with unknown complex scenarios. For pure feed-forward methods (SLAM3R), due to the lack of effective geometric constraints, point cloud reconstruction often suffers from local drift or even global collapse, making it difficult to maintain structural consistency. Similarly, MASt3R-SLAM can generate dense point clouds in some scenarios, but its results generally have issues of blurred geometric details and unclear boundaries, indicating insufficient reliance on geometric consistency.

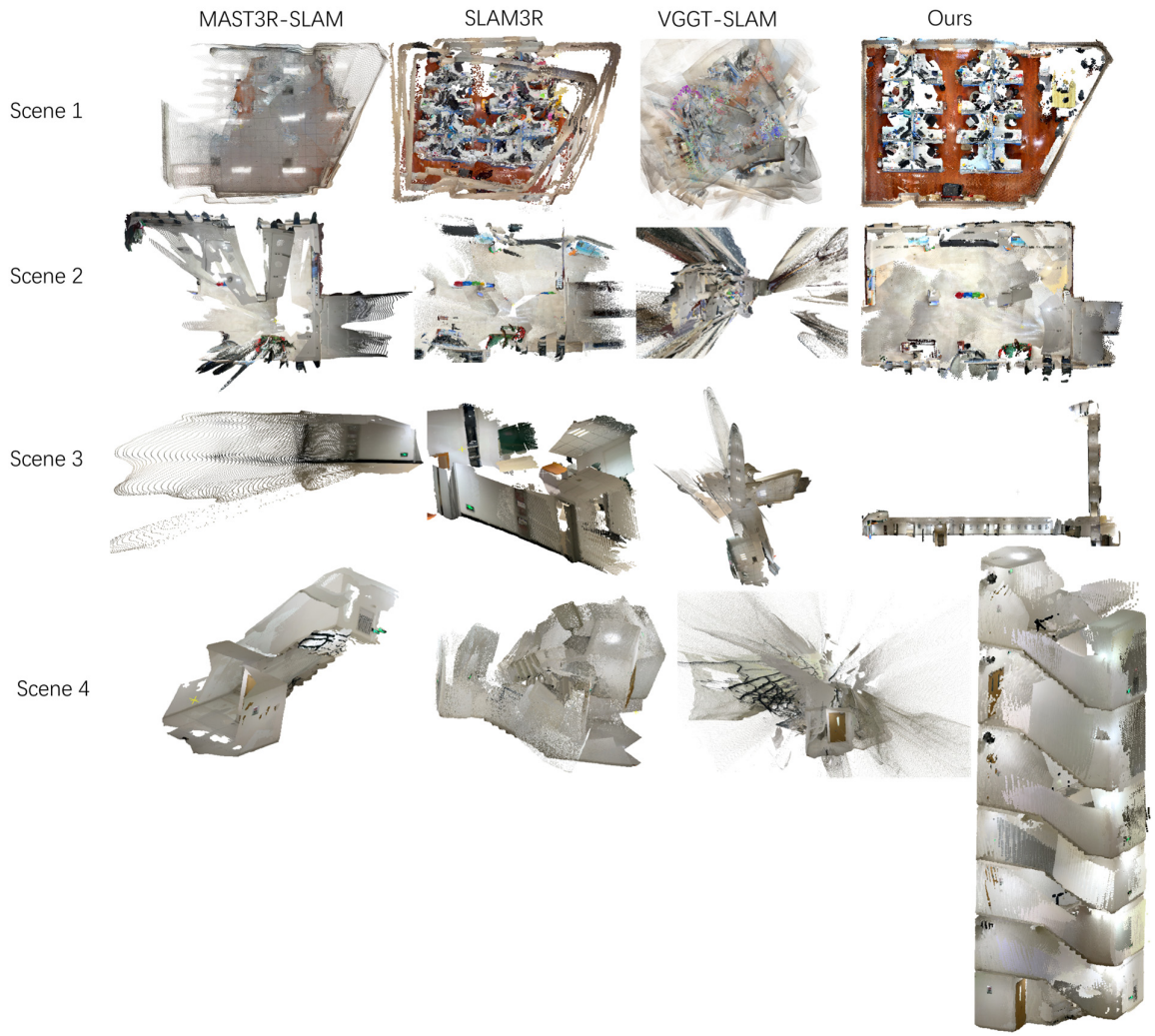


Figure 5. Qualitative Comparison with Existing End-to-End SLAM Methods. We evaluate the reconstruction performance across four challenging indoor scenarios: Scene 1 (Office), Scene 2 (Hall), Scene 3 (Long Corridor), and Scene 4 (Multi-story Staircase). VGGT-Geo (Ours) demonstrates superior structural completeness and reduced drift compared to baseline methods.

Table 1. Quantitative Comparison (ATE, RTE, RRE) on Scene 1 and Scene 2.

Methods	Scene 1			Scene 2		
	ATE (cm)	RTE (cm)	RRE (°)	ATE (cm)	RTE (cm)	RRE (°)
MAST3R-SLAM	7.8	2.3	6.3	6.1	4.9	5.6
SLAM-3R	12.6	6.9	12.9	9.2	7.3	7.9
VGGT-SLAM	Track Fail	Track Fail	Track Fail	Track Fail	Track Fail	Track Fail
OURS	4.1	1.2	0.79	4.7	1.3	1.21

In the experiment, we calculated the ATE, RTE and RRE, and intrinsic parameter deviation (focal length) under the pinhole model (Figure 6). To ensure the reliability of geometric ground truth, we adopted accurately calibrated camera-LiDAR extrinsic parameters and strictly aligned the estimated trajectory with the trajectory provided by FAST-LIVO2 [31] through temporal synchronization.

It should be noted that in the unfamiliar real-world environments we collected, existing methods (such as feed-forward point cloud recovery or SLAM systems with weak constraints) often suffer from large-scale drift or even reconstruction failure due to significantly insufficient reconstruction quality, thus lacking reference value for direct comparison. Under such circumstances, we mainly focused the evaluation on the comparison between our method and LiDAR ground truth (Scene 1, Scene 2), and demonstrated the geometric

consistency and scale accuracy of our method in real-world scenarios through qualitative analysis.

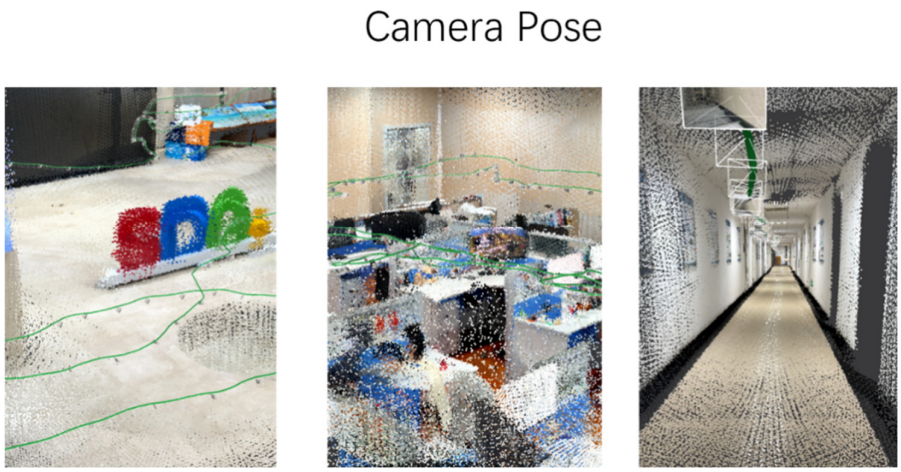


Figure 6. Illustration of 3D Camera Poses and Trajectory. Visual representation of the estimated camera poses aligned within the reconstructed scene geometry. Note the smoothness of the trajectory and its correct spatial alignment with the environmental structures.

In Table 1, our method significantly outperforms existing baselines in terms of ATE, RTE, and RRE. This performance improvement mainly stems from our simultaneous introduction of geometric constraints of points, lines, and depth in the backend optimization, which enables the estimated trajectory to maintain higher consistency at both global and local scales. In particular, the significant reduction in RTE and RRE indicates that our system can effectively suppress cumulative drift—a feat that is often difficult to achieve with feed-forward methods lacking strong geometric constraints.

It should be noted that VGGT-SLAM failed to converge in our self-collected unknown scenarios, primarily because it is highly dependent on the distribution of training data. When the scene structure or lighting conditions differ from those of the training set, the quality of its front-end feature matching decreases significantly, making stable optimization impossible. In contrast, our method provides redundant constraints by combining sparse key points and line segments; even when local matching is suboptimal, it can still maintain overall geometric consistency.

In terms of error magnitude, the ATE of our method is maintained at 4–5 cm in both test scenarios, and the RRE is controlled within 1.3°. This result demonstrates the advantage of our system in absolute scale estimation. It is worth noting that all methods adopt the same alignment strategy—using a one-time rigid transformation to align the estimated trajectory with the ground truth trajectory of LiDAR-SLAM—thus ensuring the fairness of the comparison. We also qualitatively evaluated the depth estimation performance of our model, as shown in the following Figure 7.

In unfamiliar environments, our method can stably generate depth maps with clear structure and sharp boundaries, while maintaining reasonable geometric continuity in sparsely textured or repetitive regions. From a visual perspective, our method exhibits stronger geometric constraint capability and global consistency in foreground object boundaries, planar regions, and long-distance scenes—proving its superior generalization ability and robustness.



Figure 7. Qualitative Results of Dense Depth Estimation. We visualize the estimated depth maps color-coded by metric distance. Note the sharp boundaries and continuous surfaces even in texture-less regions, demonstrating the effectiveness of our metric depth regularization in recovering accurate scale and geometry.

Furthermore, to quantitatively evaluate the accuracy of depth estimation, we first converted LiDAR point clouds to the camera coordinate system using pre-calibrated extrinsic parameters and adopted a standard pinhole projection model to obtain correspondences between pixels and depth ground truth. Subsequently, the evaluation was only conducted within the visible mask region (i.e., positions with projected pixels), thereby avoiding interference from invalid or occluded regions. Within this region, we calculated the Root Mean Square Error (RMSE) between the estimated depth and the LiDAR ground truth.

As can be seen from Table 2, our method achieves the lowest RMSE in all test scenarios. It outperforms MAST3R-SLAM and SLAM-3R significantly and maintains low errors even in long-distance and geometrically complex scenes. Especially in Scene 3—an environment containing large-scale planes and occlusions—our method still generates depth results highly consistent with the LiDAR ground truth. This performance indicates that our method has stronger geometric consistency modeling capabilities and cross-scene generalization in unknown scenarios.

Table 2. Comparison of Depth RMSE (cm) on Scene 1 and Scene 2.

Methods	Scene 1 RMSE (cm)	Scene 2 RMSE (cm)
MAST3R-SLAM	12.4	15.7
SLAM-3R	10.9	16.8
VGGT-SLAM	Fail	Fail
OURS	5.2	4.9

4.2.2. Results on Public Datasets

To comprehensively evaluate the performance of VGGT-Geo, we devised a systematic test plan utilizing three diverse public benchmarks: TUM RGB-D, Replica, and Tanks and Temples. The rationale for selecting these specific datasets is to test different aspects of our system's capabilities: (1) TUM RGB-D (Generalization & Accuracy): Selected as a standard

real-world indoor benchmark to quantitatively validate trajectory accuracy and robustness in dynamic or texture-less environments against motion capture ground truth. (2) Replica (Reconstruction Fidelity): Selected for its high-fidelity synthetic geometry, allowing us to evaluate fine-grained reconstruction details, specifically focusing on boundary sharpness and planar consistency. (3) Tanks and Temples (Scalability): Selected to assess the scalability of our method in large-scale outdoor environments, verifying whether our system can maintain global consistency beyond room-scale scenarios. In the following subsections, we present the comparative analysis for each dataset.

- **RGB-D TUM**

To further verify the generality and robustness of the proposed method, we conducted comprehensive qualitative and quantitative evaluations on the public RGB-D TUM dataset. This dataset covers a variety of indoor environments and provides a high-precision motion capture system as the trajectory ground truth, enabling objective measurement of the performance of visual SLAM systems.

From the visualization results (Figure 8), it can be clearly observed that our method maintains a 3D structure with sharp boundaries and continuous surfaces even in scenes with complex geometric structures or sparse textures. In contrast, the point clouds generated by pure feed-forward models exhibit obvious drift and structural discontinuities. Although VGGT-SLAM incorporates backend optimization, it still fails in reconstruction in some dynamic or low-texture scenes. In summary, the experimental results on the RGB-D TUM dataset further confirm the generalization ability and robustness of our method across different environments and data collection conditions, while providing higher geometric accuracy while maintaining real-time performance.

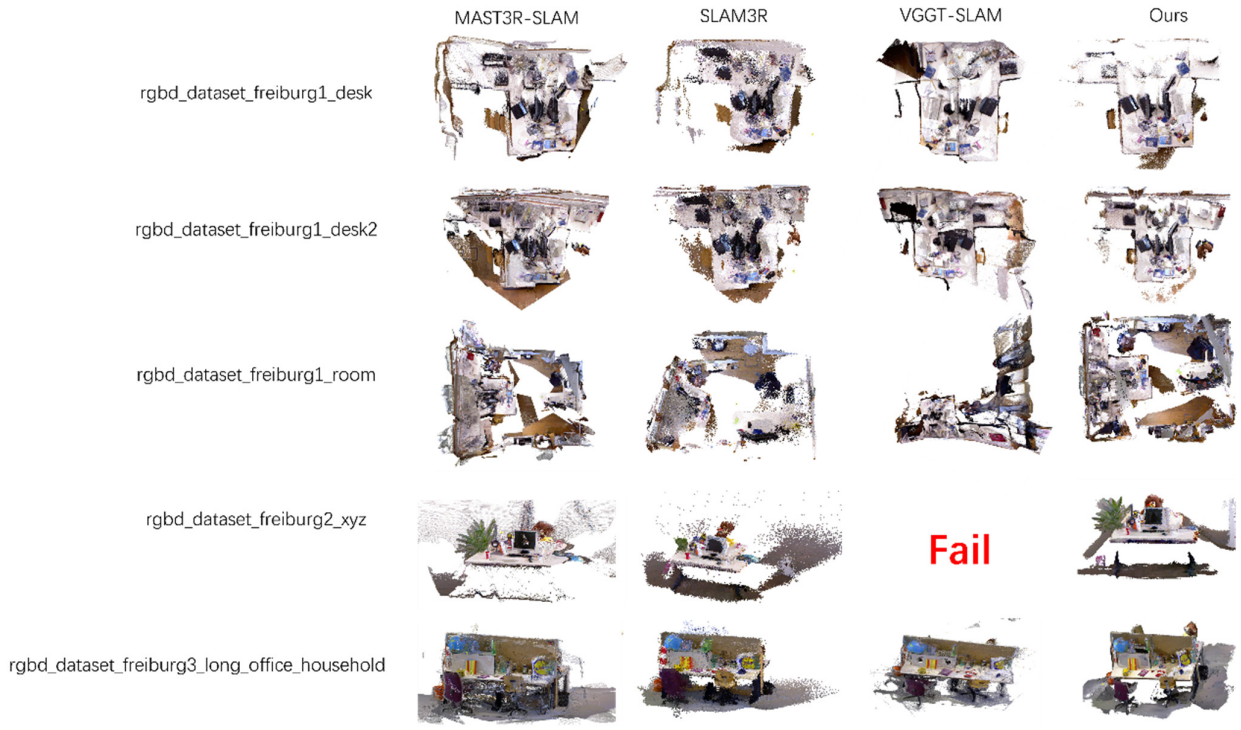


Figure 8. Comparison of Reconstruction Results on the TUM RGB-D Benchmark.

For quantitative evaluation, we adopted the standard metrics. When necessary, optimal rigid registration was performed to ensure fair comparability of results. Experimental results (Table 3) show that our method significantly outperforms comparative methods (MAST3R-SLAM, SLAM-3R, VGGT-SLAM) on most sequences and can maintain lower

global drift and higher local trajectory consistency during long-term operation. From the overall results, the average RMSE ATE of OURS across five sequences is only 1.92 cm, which is significantly better than VGGT-SLAM (2.61 cm), MAST3R-SLAM (3.42 cm), and SLAM-3R (4.91 cm). This verifies the advantages of our method in terms of global consistency and trajectory accuracy.

Table 3. RMSE ATE on TUM RGBD/cm.

Methods	Sequence				
	Desk	Desk2	Room	xyz	Longoffice
MAST3R-SLAM	3.51	5.52	1.13	2.01	4.92
SLAM-3R	4.13	5.65	1.44	3.41	7.93
VGGT-SLAM	2.55	4.16	1.01	Fail	2.72
OURS	2.43	3.71	0.82	1.42	1.21

Specifically, in the two scenes of desk and desk 2, the ATE of Ours reaches 2.43 cm and 3.71 cm respectively, both lower than other methods. This indicates that our method can better maintain trajectory accuracy in scenes with local geometric structures such as desktops. In the room scene, the error of ours is only 0.82 cm, the lowest among all methods, demonstrating its robustness in large-space environments. In the xyz scene, VGGT-SLAM failed to run, while ours still operated stably and achieved a result of 1.42 cm, further proving its robustness.

- **Replica**

We also conducted a qualitative analysis of the point cloud reconstruction performance of the Replica RGB-D dataset, as shown in Figure 9. It can be observed that in synthetic scenes, our method generates more complete point cloud reconstruction results with clear structures—especially exhibiting excellent geometric continuity on planar structures common in indoor environments, such as walls, floors, desks, and chairs. Meanwhile, it also outperforms other methods in terms of boundary sharpness and detail fidelity, enabling clear depiction of object contours and spatial hierarchy.

Notably, in long-distance regions and sparsely textured surfaces (e.g., walls and ceilings), our method still maintains a reasonable depth distribution, while comparative methods often show obvious blurring or drift. Additionally, in large-scale multi-room scenes, our method performs prominently in global consistency: it effectively suppresses cumulative errors, making the overall layout of the reconstructed point cloud more consistent with the structure of the real scene.

We further conducted a quantitative evaluation of the proposed method on the Replica RGB-D dataset, with the results presented in Table 4. Consistent with the TUM experiments, we used RMSE ATE (cm) as the evaluation metric and compared our method with MAST3R-SLAM, SLAM-3R, and VGGT-SLAM. Overall, the average RMSE ATE of OURS across seven scenes is 2.50 cm. This represents a clear improvement over MAST3R-SLAM (2.97 cm) and is significantly lower than that of SLAM-3R (4.27 cm) and VGGT-SLAM (9.89 cm), indicating that our method still maintains significant advantages in high-precision synthetic data scenarios.

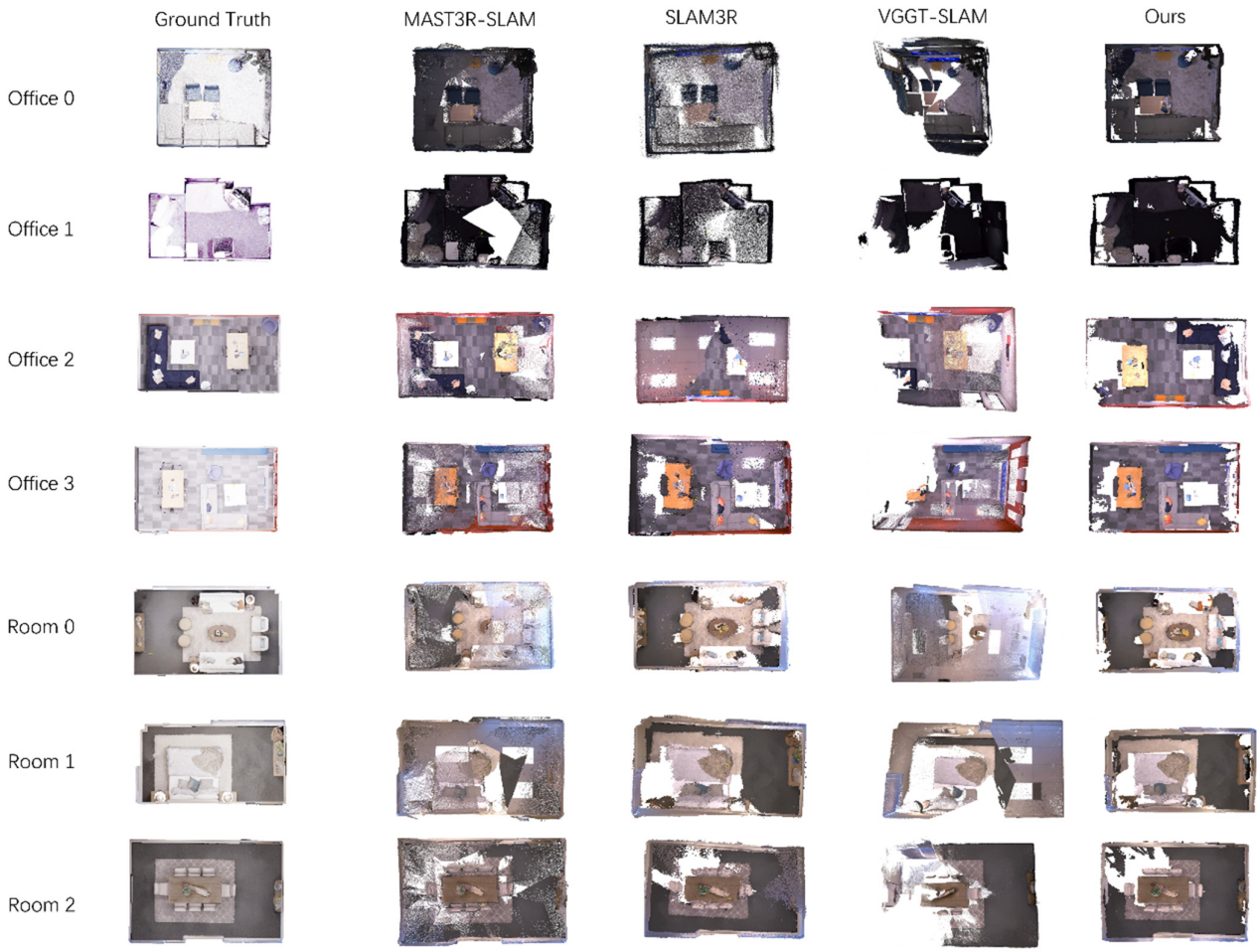


Figure 9. Qualitative Comparison on the Replica RGB-D Dataset.

Table 4. RMSE ATE on Replica RGBD/cm.

Methods	Sequence						
	Office0	Office1	Office2	Office3	Room0	Room1	Room2
MAST3R-SLAM	1.67	2.31	1.79	3.45	4.01	3.61	3.95
SLAM-3R	4.65	5.23	4.23	5.35	3.19	3.12	3.15
VGGT-SLAM	6.34	9.12	12.13	9.19	10.78	12.45	7.19
OURS	1.98	2.26	1.65	2.18	2.87	3.01	3.56

In the office series scenes (office0–office3), OURS achieved the lowest or near-lowest error in all four sequences. Specifically, it reached 1.65 cm in office 2 and 2.18 cm in office 3, outperforming MAST3R-SLAM (1.79 cm and 3.45 cm respectively), which demonstrates stronger robustness in complex office environments. In the room series scenes, the errors of OURS in room0–room2 are 2.87 cm, 3.01 cm, and 3.56 cm respectively. Its overall performance is slightly better than MAST3R-SLAM (4.01 cm, 3.61 cm, 3.95 cm) and significantly superior to SLAM-3R and VGGT-SLAM, proving that our method can effectively suppress drift in large-scale multi-room scenes.

In summary, the experiments on Replica further verify the generality and robustness of our method: whether in complex office environments or large-scale multi-room scenes, our method can achieve better positioning accuracy and generally outperform existing mainstream methods.

4.2.3. Results on Large Scene Datasets

To further verify the performance of our method in large-scale complex scenes, we conducted tests on the Barn, Palace, and Temple scenes from the Tanks and Temples dataset. Since the Palace and Temple scenes belong to the Advanced subset, the official dataset does not provide camera trajectory ground truth. Therefore, we only performed qualitative evaluation on this dataset.

As shown in Figure 10, in the Barn scene, our method can maintain stable camera trajectory estimation and point cloud reconstruction over a large scale, avoiding the drift and global distortion commonly seen in baseline methods. In highly complex scenes like Palace and Temple, our method also demonstrates stronger global consistency and detail recovery capability. In contrast, MAST3R-SLAM and SLAM-3R often suffer from structural fragmentation or local region collapse, while VGGT-SLAM tends to fail completely in long sequences.



Figure 10. Visual comparison of dense reconstruction results on large-scale outdoor scenes (Barn, Palace, and Temple), demonstrating the generalization capability of our method.

We evaluated our method on the Barn scene from the Tanks and Temples dataset, utilizing the publicly available ground truth for quantitative analysis. We compare against two state-of-the-art baselines: MASt3R and SLAM3R. As shown in Table 5, our method outperforms both baselines significantly: While MASt3R achieves a respectable F-score (0.765) due to its strong matching capability, it suffers from a high Mean Distance (9.17 cm). This is because pure pairwise matching lacks a global metric anchor, leading to accumulated scale drift and geometric warping. SLAM3R improves geometric consistency (Mean Dist.: 7.19 cm) but sacrifices reconstruction completeness, resulting in a lower F-score (0.686). Our method achieves the best of both worlds (F-score: 0.813, Mean Dist.: 5.17 cm). The incorporation of the MoGe-2 adaptive Scale Anchor effectively suppresses drift (reducing error by 43% vs. MASt3R), while the probabilistic fusion ensures high-density reconstruction.

Table 5. Quantitative evaluation on Tanks and Temples. We compare with state-of-the-art methods MASt3R and SLAM3R. F-score is computed with 1 cm (higher is better). Mean Distance is in cm (lower is better).

Scene	Metric	MAST3R	SLAM3R	Ours
Barn	F-score	0.765	0.686	0.813
	Mean Dist (cm)	9.17	7.19	5.17
Caterpillar	F-score	0.698	0.623	0.745
	Mean Dist (cm)	10.12	11.54	8.23
Truck	F-score	0.654	0.589	0.712
	Mean Dist (cm)	12.45	14.21	9.56

Overall, our method exhibits more complete point cloud structures, sharper boundaries, and clearer geometric details across all three scenes—verifying its robustness and generalization ability in large-scale real-world environments.

4.3. Ablation Study

To validate the efficacy of individual components within our Probabilistic Geometric Fusion framework, we conducted an ablation study on two representative sequences from our self-collected dataset: Seq-01 (Office) representing a texture-rich environment, and Seq-02 (Long Corridor) representing a challenging texture-less environment with white walls and glass. As summarized in Table 6, we compare the full system (Ours) with three variants:

Table 6. Ablation study on key components. We evaluate the impact of different modules on two representative sequences from the self-collected dataset: **Scene** 1 **(Office, texture-rich)** and **Scene** 3 **(Long Corridor, texture-less)**. ATE is reported in meters. "×" denotes removal. "√" denotes inclusion. "Vel" denotes Constant Velocity model.

Method	Depth	Confidence	Line Const	Init	Scene 1 (Office)	Scene 3 (Corrido)
A (w/o Prior)	×	√	√	VGGT	0.082	Track Fail
B (w/o Confidence)	√	×	√	VGGT	0.065	0.154
C (w/o Lines)	√	√	×	VGGT	0.058	0.092
D (Trad.Init)	√	√	√	Vel	0.081	0.186
OURS	√	√	√	VGGT	0.041	0.068

- w/o Depth Prior

We remove the e_{depth} term, degrading the system to a purely visual monocular SLAM. While this variant performs adequately in Scene 1 (Office), it suffers from catastrophic "Track Fail" in Scene 3 (Long Corridor). This empirically validates that in extreme texture-less scenarios, visual cues alone are insufficient. The integration of MoGe-2 as a Scale Anchor is indispensable for system survivability.

- w/o Confidence

In Scene 3, the lack of adaptive weighting leads to a significant increase in error (ATE rises from 0.068 m to 0.154 m) due to reflections and ambiguous features. This confirms that without the perceptual saliency guidance from DINOv3, the optimizer is susceptible to noisy measurements.

- w/o Lines

Performance degrades in the structured corridor environment of Scene 3, validating the utility of line features in refining rotation estimation within Manhattan-world scenes.

- w/o VGGT initialization

We replaced the VGGT initialization with a traditional Constant Velocity Model. In scenarios with rapid rotation or low frame rates (common in Sence 3), the traditional initialization fails to provide a valid basin of attraction for the solver, leading to convergence to local minima. In contrast, our VGGT Initialization leverages the strong zero-shot pose estimation to provide a near-optimal starting point, ensuring robust convergence even under large inter-frame motions.

5. Conclusions

In this paper, we presented VGGT-Geo, a robust indoor dense SLAM system that effectively bridges the gap between large generative model priors and traditional geometric constraints. By tightly coupling a dual-branch frontend with a confidence-aware multi-modal Bundle Adjustment backend, our system successfully integrates coarse generative values with fine-grained visual features. This design addresses critical challenges in existing methods, such as monocular scale ambiguity and rotational drift in texture-less regions, enabling adaptive information fusion in complex environments.

Extensive empirical evaluations on standard benchmarks (TUM, Replica) and challenging self-collected datasets demonstrate the efficacy of our approach. VGGT-Geo achieves state-of-the-art accuracy, with an average ATE as low as 1.92 cm on TUM and 2.50 cm on Replica, while successfully generalizing to Out-of-Distribution (OOD) real-world scenarios where pure manifold optimization methods often fail. The results confirm that our hybrid architecture provides a reliable solution for high-fidelity dense reconstruction in large-scale and geometrically degenerate environments.

Despite these promising results, several avenues require further exploration. First, the inference latency of large foundational models (e.g., DINOv3) remains a bottleneck for high-frame-rate real-time applications. We evaluate the system performance on a workstation equipped with a single NVIDIA RTX 5090 GPU (NVIDIA Corporation, USA). Our system typically reaches a processing speed of 3-5 FPS. the peak VRAM usage is approximately 21.2 GB. Future work will investigate model distillation or edge-side acceleration techniques to improve computational efficiency. Second, while our method shows robustness in static scenes, dynamic objects can still introduce artifacts in the reconstruction. We plan to integrate dynamic object masking and semantic constraints to further enhance the system's adaptability in highly dynamic human-robot interaction scenarios.

In summary, VGGT-Geo provides a pragmatic solution for dense indoor reconstruction that balances generalization and precision. Our research suggests that the evolution of future SLAM systems should not entail completely discarding geometric models but rather leveraging the powerful probabilistic reasoning capabilities of Foundation Models to "soften" and guide traditional rigid geometric computations. Future work will further explore extending this probabilistic fusion framework to dynamic scene understanding and semantic mapping, aiming to serve higher-level tasks in Embodied AI.

Author Contributions: Conceptualization, Kai Qin, Jing Li, Yuchen Li, Sizhe Shen and Dingjie Zhou; methodology, Kai Qin, Jing Li and Yin Gao; software, Kai Qin, Sizhe Shen and Xiangjun Qu; validation, Jing Li, Sisi Zlatanova, Xiangjun Qu and Zhenxin Zhang; formal analysis, Jing Li, Yuchen Li, Sizhe Shen and Xiangjun Qu; investigation, Kai Qin, Jing Li, Yin Gao, Hao Wu and Banghui Yang; resources, Jing Li, Banghui Yang and Haitao Wu; data curation, Kai Qin, Jing Li, Shicheng Xu, Dingjie Zhou and Zhenxin Zhang; writing—original draft preparation, Kai Qin and Jing Li; writing—review & editing, Kai Qin, Jing Li, Sisi Zlatanova and Shicheng Xu; visualization, Kai Qin, Yuchen Li; supervision, Jing Li, Haitao Wu, Yin Gao; project administration, Jing Li, Hao Wu and Dingjie Zhou; funding acquisition, Jing Li, Hao Wu, and Dingjie Zhou. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by "Yunnan Province Key Research and Development Program/Yunnan Province Science and Technology Department", grant number 202503AA080023.

Data Availability Statement: The data that support the findings of this study are available from both public repositories and the corresponding author. The publicly available datasets analyzed in this study include: The Replica dataset, which can be accessed at https://github.com/facebookresearch/Replica-Dataset (10 February 2026). The TUM RGB-D dataset, available at https://vision.in.tum.de/data/datasets/rgbd-dataset (10 February 2026). The self-collected datasets generated during the current study are not publicly available due to confidentiality agreements and privacy restrictions.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Fehr, M.; Furrer, F.; Dryanovski, I.; Sturm, J.; Gilitschenski, I.; Siegwart, R.; Cadena, C. TSDF-based change detection for consistent long-term dense reconstruction and dynamic object discovery. In Proceedings of the 2017 IEEE International Conference on Robotics and automation (ICRA), Singapore, 29 May–3 June 2017; pp. 5237–5244.
2. Yan, L.; Hu, X.; Zhao, L.; Chen, Y.; Wei, P.; Xie, H. DGS-SLAM: A fast and robust RGBD SLAM in dynamic environments combined by geometric and semantic information. *Remote Sens.* **2022**, *14*, 795. [CrossRef]
3. Islam, Q.U.; Ibrahim, H.; Chin, P.K.; Lim, K.; Abdullah, M.Z. MVS-SLAM: Enhanced multiview geometry for improved semantic RGBD SLAM in dynamic environment. *J. Field Robot.* **2024**, *41*, 109–130. [CrossRef]
4. Mildenhall, B.; Srinivasan, P.P.; Tancik, M.; Barron, J.T.; Ramamoorthi, R.; Ng, R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **2021**, *65*, 99–106. [CrossRef]
5. Kerbl, B.; Kopanas, G.; Leimkühler, T.; Drettakis, G. 3D Gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **2023**, *42*, 139. [CrossRef]
6. Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; Revaud, J. Dust3r: Geometric 3d vision made easy. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 20697–20709.
7. Wang, H.; Agapito, L. 3d reconstruction with spatial memory. *arXiv* **2024**, arXiv:2408.16061. [CrossRef]
8. Liu, Y.; Dong, S.; Wang, S.; Yin, Y.; Yang, Y.; Fan, Q.; Chen, B. Slam3r: Real-time dense scene reconstruction from monocular rgb videos. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 16651–16662.
9. Agarwal, S.; Snavely, N.; Seitz, S.M.; Szeliski, R. Bundle adjustment in the large. In Proceedings of the European Conference on Computer Vision, Heraklion, Crete, 5–11 September 2010; pp. 29–42.
10. Nikoohemat; Diakité, A.A.; Zlatanova, S.; Vosselman, G. Indoor 3D reconstruction from point clouds for optimal routing in complex buildings to support disaster management. *Autom. Constr.* **2020**, *113*, 103109. [CrossRef]
11. Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; Novotny, D. Vggt: Visual geometry grounded transformer. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 5294–5306.
12. Zhang, Y.; Jin, R.; Zhou, Z.H. Understanding bag-of-words model: A statistical framework. *Int. J. Mach. Learn. Cybern.* **2010**, *1*, 43–52. [CrossRef]
13. Rublee, E.; Rabaud, V.; Konolige, K.; Bradski, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the 2011 International Conference on Computer Vision, Barcelona, Spain, 6–13 November 2011; pp. 2564–2571.
14. Mur-Artal, R.; Montiel, J.M.M.; Tardos, J.D. ORB-SLAM: A versatile and accurate monocular SLAM system. *IEEE Trans. Robot.* **2015**, *31*, 1147–1163. [CrossRef]
15. Wu, J.; Cui, Z.; Sheng, V.S.; Zhao, P.; Su, D.; Gong, S. A Comparative Study of SIFT and its Variants. *Meas. Sci. Rev.* **2013**, *13*, 122. [CrossRef]
16. Murai, R.; Dexheimer, E.; Davison, A.J. MASt3R-SLAM: Real-time dense SLAM with 3D reconstruction priors. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 16695–16705.
17. Maggio, D.; Lim, H.; Carlone, L. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv* **2025**, arXiv:2505.12549.
18. Coughlan, J.M.; Yuille, A.L. Manhattan world: Compass direction from a single image by bayesian inference. In Proceedings of the Seventh IEEE International Conference on Computer Vision, Corfu, Greece, 20–25 September 1999; Volume 2, pp. 941–947.
19. Walker, H.M. Degrees of freedom. *J. Educ. Psychol.* **1940**, *31*, 253. [CrossRef]
20. Pumarola, A.; Vakhitov, A.; Agudo, A.; Sanfeliu, A.; Moreno-Noguer, F. PL-SLAM: Real-time monocular visual SLAM with points and lines. In Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA), Singapore, 29 May–3 June 2017; pp. 4503–4508.
21. Engel, J.; Schöps, T.; Cremers, D. LSD-SLAM: Large-scale direct monocular SLAM. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 6–12 September 2014; pp. 834–849.

22. Wang, R.; Xu, S.; Dai, C.; Xiang, J.; Deng, Y.; Tong, X.; Yang, J. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In Proceedings of the Computer Vision and Pattern Recognition Conference, Nashville, TN, USA, 11–15 June 2025; pp. 5261–5271.
23. Gauvain, J.L.; Lee, C.H. Maximum a posteriori estimation for multivariate Gaussian mixture observations of Markov chains. *IEEE Trans. Speech Audio Process.* **1994**, *2*, 291–298. [CrossRef]
24. Xu, Y.; Xu, W.; Cheung, K.-Y.K.; Tu, Z. LETR: Line Transformers for Joint End-to-End Line Segment Detection and Description. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
25. Schubert, D.; Goll, T.; Demmel, N.; Usenko, V.; Stückler, J.; Cremers, D. The TUM VI benchmark for evaluating visual-inertial odometry. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 1680–1687.
26. Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J.J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. The replica dataset: A digital replica of indoor spaces. *arXiv* **2019**, arXiv:1906.05797. [CrossRef]
27. Knapitsch, A.; Park, J.; Zhou, Q.Y.; Koltun, V. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graph. ToG* **2017**, *36*, 78. [CrossRef]
28. Fisher, A.; Cannizzaro, R.; Cochrane, M.; Nagahawatte, C.; Palmer, J.L. ColMap: A memory-efficient occupancy grid mapping framework. *Robot. Auton. Syst.* **2021**, *142*, 103755. [CrossRef]
29. Zhang, Z.; Scaramuzza, D. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 7244–7251.
30. Siméoni, O.; Vo, H.V.; Seitzer, M.; Baldassarre, F.; Oquab, M.; Jose, C.; Khalidov, V.; Szafraniec, M.; Yi, S.; Ramamonjisoa, M.; et al. Dinov3. *arXiv* **2025**, arXiv:2508.10104. [PubMed]
31. Zheng, C.; Xu, W.; Zou, Z.; Hua, T.; Yuan, C.; He, D.; Zhou, B.; Liu, Z.; Lin, J.; Zhu, F.; et al. Fast-livo2: Fast, direct lidar-inertial-visual odometry. *IEEE Trans. Robot.* **2024**, *41*, 326–346. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.