GRF-Recon: Global Ray-Field Optimization for Long-Sequence Feed-forward Reconstruction

Enpeng Li, Yunzhou Zhang*, Zhiyao Zhang, Dexuan Lyu, Chenyu Wang, Chiyuan Cui, and Cheng Cheng

College of Information Science and Engineering, Northeastern University, China
zhangyunzhou@mail.neu.edu.cn

Abstract. Feed-forward 3D reconstruction provides an efficient paradigm for scene modeling from image sequences. Scaling these models to large monocular scenarios are constrained by excessive GPU memory footprint, degraded local geometry, and long-term trajectory drift. Existing chunk-based optimization strategies provide limited geometric constraints and fail to maintain global consistency over extended trajectories. We present a unified framework for stable and scalable feed-forward 3D reconstruction from long monocular sequences. Our approach builds on coarse-to-fine trajectory alignment augmented by lightweight geometric prior injection. Distilling monocular geometric cues into the feed-forward backbone via LoRA adaptation improves depth accuracy on fine structures while preserving inference efficiency. We introduce a hybrid-weight sparse ray-field optimization that leverages high-frequency geometric features to guide local point-cloud refinement and enforce consistent inter-frame ray constraints. Unlike prior chunk-based methods, this establishes strong cross-frame geometric coupling while maintaining scalability. Finally, an efficient trajectory stitching strategy with joint ray-error optimization explicitly reduces accumulated drift. Extensive experiments show that our approach achieves competitive trajectory accuracy compared with representative SLAM systems, while maintaining globally consistent 3D reconstruction in large-scale scenarios.

Keywords: Long-sequence Reconstruction · Ray-field Optimization · Feed-forward Mapping · Global Consistency

1 Introduction

Achieving accurate camera pose estimation and dense 3D scene reconstruction from monocular RGB video streams remains a core challenge in 3D vision and autonomous driving [8, 20]. Due to the inherent inability of monocular vision to perceive absolute scale, the problem of accumulated drift in long-sequence tracking is particularly severe, often leading to significant global scale inconsistencies. Traditional systems for processing large-scale monocular sequences

* Corresponding author.

Fig. 1: Large-scale 3D reconstruction using GRF-Recon. Our proposed framework successfully reconstructs a globally consistent, low-drift point cloud map over a kilometer-scale monocular video sequence (KITTI [11] Seq. 00, 3724m). The left panel shows the accurately closed global trajectory, while the right panel highlights the high-fidelity geometric details recovered at a complex urban intersection.

heavily rely on complex, multi-module pipelines consisting of camera calibration, visual odometry, multi-sensor fusion, and heavy backend graph optimization [5,25,33]. In loosely coupled architectures, sensor noise and data association errors inevitably amplify and accumulate over large-scale movements.

Recently, Transformer-based [36] end-to-end foundation models have emerged in 3D vision, enabling efficient pose inference and high-quality dense point cloud reconstruction. Ranging from DUSt3R [42] and MASt3R [15] to the latest VGGT [38] and Depth Anything 3 (DA3) [16], these models are pre-trained on massive datasets, integrating camera pose estimation, intrinsic regression, and 3D scene representation into a tightly coupled framework. Their core advantage lies in enabling the end-to-end backpropagation of errors throughout the entire system, thereby establishing a universal 3D reconstruction foundation model capable of directly processing uncalibrated, raw RGB images.

While these models perform well on benchmark datasets and short sequences, their scalability to complex, large-scale real-world scenarios remains limited. Their computational and memory requirements grow quadratically with the number of input tokens due to the Transformer attention mechanism, resulting in substantial GPU memory consumption during inference [22, 27]. In practice, this often restricts processing to only a few dozen frames on a consumer-grade GPU. Moreover, the lack of explicit global constraints makes long sequences susceptible to trajectory drift. Local estimation errors accumulate over time, leading to Sim(3) drift that degrades multi-view consistency and metric accuracy [9, 19]. Robustness further deteriorates in large, unbounded scenes. Since these models are strongly influenced by their pre-training distributions, drastic viewpoint changes or unseen geometric configurations can destabilize local reconstruction, particularly around fine structures and depth discontinuities.

In long-sequence settings, scaling feed-forward foundation models requires reduced memory consumption and sensitivity to fine-grained geometric structures. We therefore incorporate feature redundancy filtering and lightweight geometric adaptation [13, 23] to construct an efficient local perception front-end. Improving the local front-end alone does not address the challenge of maintaining global consistency over kilometer-scale trajectories, where loosely coupled chunk-stitching strategies often fail to enforce long-range geometric constraints. To overcome this limitation, we introduce **GRF-Recon**, a monocular reconstruction system for large-scale scenes (Fig. 1). It formulates a global joint optimization scheme based on a hybrid-weight sparse ray-field to explicitly couple cross-view geometric constraints. Experiments on large-scale outdoor autonomous driving benchmarks, including KITTI [11], show that our method achieves accurate reconstruction of kilometer-scale environments without requiring camera calibration or depth supervision. It outperforms prior feed-forward models and achieves competitive performance compared to recent SLAM-based approaches in trajectory accuracy and global map consistency [30, 45].

2 Related Work

Monocular 3D Reconstruction and SLAM. Classical Structure-from-Motion (SfM) and visual SLAM systems estimate camera poses and sparse 3D structures either implicitly or explicitly via multi-view geometry [12]. Traditional pipelines like COLMAP [25] typically follow an incremental paradigm, relying heavily on feature point detection [10, 34], cross-view matching [17, 24], and heavy bundle adjustment [33]. While these methods demonstrate high robustness in ideal environments, they exhibit extreme vulnerability in textureless regions and accumulate irreversible scale drift over long sequences. Recently, fully differentiable systems like VGGSfM [39] have demonstrated the immense potential of end-to-end learning frameworks through pixel-level 2D point tracking and global simultaneous recovery mechanisms. In terms of dense matching and tracking, DROID-SLAM [30] innovatively integrates deep learning features with dense bundle adjustment. Its backend essentially inherits the optimization logic of sparse SLAM. Lacking explicit global geometric constraints, it remains prone to 3D geometric inconsistencies in long sequences [18]. Recent efforts such as MASt3R-SLAM [19] attempt to embed feed-forward priors into real-time SLAM but still require complex traditional backend machinery.

Feed-forward 3D Foundation Models and Long-Sequence Scaling. The most significant recent paradigm shift in 3D vision is the emergence of end-to-end foundation models based on the Transformer architecture [15, 36, 42]. These methods are capable of directly generating remarkably stable local 3D maps from uncalibrated overlapping images. The immense computational and memory overhead induced by the self-attention mechanism strictly limits their application to relatively short sequences. To break the bottleneck of processing long sequences, Spann3R [37] and LONG3R [7] explore streaming reconstruction based on spatio-temporal memory, whereas MUSt3R [4] and Fast3R [44]

focus on multi-view parallelization to eliminate pairwise matching errors. Despite some progress, streaming processing is susceptible to long-term feature drift, and parallel networks still hit physical memory ceilings when facing ultra-long trajectories. VGGT-Long [9] explores a minimalist "chunk-and-align" strategy. While the chunking paradigm initially alleviates the memory crisis, existing chunk stitching methods lack deep geometric constraints [43].

Efficient 3D Perception and Feature Redundancy Filtering. As feed-forward foundation models process increasing viewpoints and resolutions, the proliferation of tokens in Transformers leads to high inference costs. In real-world 3D reconstruction, the density of geometric information across different spatial regions is highly uneven, with substantial textureless backgrounds and overlapping viewpoint redundancies. Recently, dynamic token pruning [22] and token merging [3] techniques in visual architectures have been studied to accelerate inference. In the realm of 3D foundation models, Fast-VGGT [27] proposes an efficient feature filtering paradigm. By identifying and suppressing redundant features in overlapping views and stable regions, the model significantly reduces computational burden and memory consumption with almost no loss in local reconstruction accuracy. To recover high-frequency geometric details while obtaining accurate poses, the research community (e.g., Fin3R [23]) has introduced Parameter-Efficient Fine-Tuning and monocular knowledge distillation techniques to enhance the model's perceptual acuity [2, 41].

In this work, we propose a hybrid-weight sparse ray-field that constructs rigorous explicit geometric constraints. We leverage LoRA-based [13] lightweight prior injection to accurately restore local geometric edges and ultimately establish strict global physical constraints. With minimal system overhead, this framework successfully and stably extends the boundaries of feed-forward monocular 3D reconstruction to kilometer-scale unbounded scenes.

3 Method

3.1 Local Depth Refinement and Initial Trajectory Stitching

The proposed GRF-Recon system, an overview of which is illustrated in Fig. 2, utilizes DA3 [16] as the perception backbone, introducing enhancements to enable low-drift 3D reconstruction over long sequences. Although DA3 is pre-trained on large-scale datasets, applying it to sequential reconstruction often suffers from geometric over-smoothing and degradation near occlusion boundaries. Inspired by Fin3R [23], we adopt a re-normalized Low-Rank Adaptation [13] strategy to fine-tune the DA3 encoder. This operation injects high-frequency monocular geometric priors while preserving the model's multi-view matching generalization ability, thereby improving local reconstruction fidelity.

To implement this fine-tuning without degrading DA3's cross-view correlation capability, we freeze the decoder and insert low-rank adapters into the Query–Key–Value projections of the DINOv2 [21] encoder attention layers. However, naive monocular distillation causes a feature norm drift, leading to a mis-

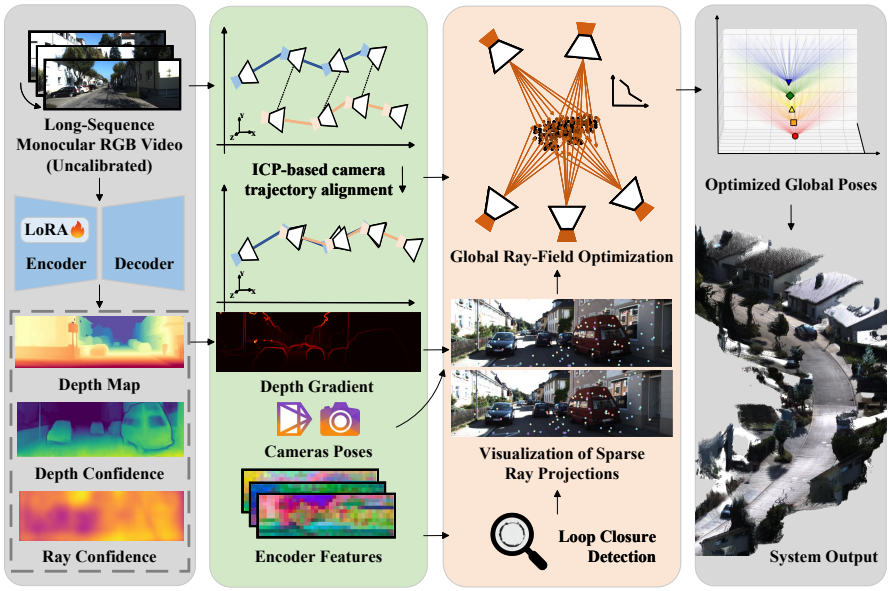

Fig. 2: Overview of the GRF-Recon System Framework. GRF-Recon reconstructs 3D geometry from uncalibrated long-sequence monocular RGB videos, building on a pretrained model with LoRA fine-tuning to enhance geometric perception. The system adopts a coarse-to-fine chunk-based parallel pipeline, performing confidence-guided sparse ray sampling and hybrid-weighted ray matching to enforce cross-view consistency. The orange module visualizes the projection of ray maps from neighboring frames onto the 2D image plane. Global optimization is carried out in a factor graph that jointly incorporates ray-field and loop-closure constraints, producing refined camera trajectories and a globally consistent, high-fidelity 3D point cloud.

match with the frozen decoder's expected input distribution. We address this by constraining the updated effective weight $\mathbf{W}'$ via a re-normalization term:

$$\mathbf{W}' = \frac{(\mathbf{W}_0 + \mathbf{BA}) \cdot \|\mathbf{W}_0\|_F}{\|\mathbf{W}_0 + \mathbf{BA}\|_F} \tag{1}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. This strategy introduces less than 1% additional trainable parameters while preserving the original weight norm, effectively stabilizing the feature activation distribution.

Following the encoder adaptation, we mitigate the risk of dataset overfitting during direct supervision by employing a teacher–student dual-branch distillation strategy. Specifically, a depth branch leverages high-quality pseudo-labels generated by an advanced monocular geometry estimator [41], applying a log-uncertainty-weighted ℓ_2 loss to supervise the DA3 depth predictions. Concurrently, a point branch utilizes ground-truth 3D point clouds to provide regression supervision, ensuring fundamental multi-view geometric consistency. The overall

distillation loss is thus defined as $\mathcal{L} = \mathcal{L}_{\text{depth}} + \lambda \mathcal{L}_{\text{point}}$, where $\lambda = 0.4$ balances the two modalities.

Once the perception front-end is capable of producing high-fidelity depth and ray-field, we process the long sequences using a chunk-based alignment strategy. To merge isolated local chunks into a unified global coordinate system, we extract the poses of the overlapping frames from two neighboring chunks $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$. We compute the similarity transformation $\mathbf{M}_{\text{sim}} \in \text{Sim}(3)$ via a decoupled optimization process. The rotation covariance matrix $\mathbf{Q}$ is formulated from the overlapping frames, and the optimal rotation $\mathbf{R}^*$ is derived through Singular Value Decomposition (SVD):

$$\mathbf{U}, \boldsymbol{\Sigma}, \mathbf{V}^\top = \text{SVD}(\mathbf{Q}), \quad \mathbf{R}^* = \mathbf{U}\,\text{diag}(1, 1, \det(\mathbf{U}\mathbf{V}^\top))\,\mathbf{V}^\top \tag{2}$$

We solve for the scale s^* and translation $\mathbf{t}^*$ via a least-squares formulation:

$$s^*, \mathbf{t}^* = \arg\min_{s,\mathbf{t}} \sum_{i=1}^{m} \left\| \mathbf{t}_k^{(i)} - \left(s\mathbf{R}^* \mathbf{t}_{k+1}^{(i)} + \mathbf{t} \right) \right\|_2^2 \tag{3}$$

The resulting matrix $\mathbf{M}_{\text{sim}}$ is applied to align $\mathcal{C}_{k+1}$ with $\mathcal{C}_k$. By propagating this spatial alignment along the sequence, the system establishes a coarse but globally consistent initial trajectory $\mathcal{T}_{\text{global}}$.

3.2 Hybrid Sparse Ray-Field Construction

Unlike traditional methods that strictly rely on pinhole models for initial geometric reasoning, our perceptual front-end directly utilizes the DA3 ray decoding head to predict pixel-level features end-to-end. For a given pixel p_i in the source frame I_i, the network outputs a 7-channel ray vector. The first 6 channels formulate the 3D ray parameterized as $\mathbf{r}_i = [\mathbf{d}_i^\top, \mathbf{o}_i^\top]^\top \in \mathbb{R}^6$, where $\mathbf{d}_i$ is the normalized ray direction and $\mathbf{o}_i$ indicates the ray origin in the local coordinate system. The 7th channel yields the ray prediction confidence c_i^{ray}. Simultaneously, the depth branch outputs the metric depth value z_i along with its associated confidence c_i^{depth}. Leveraging these metric properties, the local 3D point corresponding to the pixel is precisely reconstructed as $\mathbf{P}_i = \mathbf{o}_i + z_i \mathbf{d}_i$.

To manage the matching complexity inherent in long sequences, we sample the predicted rays sparsely based on a dual confidence fusion mechanism:

$$c_i^{\text{joint}} = \beta \cdot \hat{c}_i^{\text{depth}} + (1 - \beta) \cdot \hat{c}_i^{\text{ray}} \tag{4}$$

where $\beta = 0.5$ acts as a trade-off parameter, and $\hat{c}_i^{\text{depth}}$, $\hat{c}_i^{\text{ray}}$ are the min-max normalized confidences. We then incorporate the depth gradient magnitude $\hat{g}_i = \|\nabla D(i)\|_2$ to construct a composite score $s_i = \gamma_{\text{conf}} \cdot c_i^{\text{joint}} + \gamma_{\text{grad}} \cdot \hat{g}_i$ ($\gamma_{\text{conf}} = 0.8$ and $\gamma_{\text{grad}} = 0.2$). Pixels satisfying a predefined confidence threshold are sampled according to the probability distribution $\mathcal{P}(i) \propto s_i$, followed by Non-Maximum Suppression to ensure a uniform spatial distribution across the image.

We establish multidimensional local constraints to associate observations across frames. For a reconstructed 3D point $\mathbf{P}_i$ originating from p_i, we determine its initial search anchor in the target frame I_j via perspective projection:

$\mathbf{p}_j^{\text{init}} = \pi(\mathbf{K}_j(\mathbf{R}_j\mathbf{P}_i + \mathbf{t}_j))$, where $\pi(\cdot)$ denotes homogeneous normalization and $\mathbf{K}_j$ represents the camera intrinsics. Using this anchor as the center, feature matching is performed within a local search window $\mathcal{W}$. The primary geometric criterion is the ray distance metric e_{ray}, defined as the perpendicular Euclidean distance from the target candidate ray $\mathbf{r}_j$ to the point $\mathbf{P}_i$:

$$e_{\text{ray}}(\mathbf{P}_i, \mathbf{r}_j) = \frac{\|(\mathbf{P}_i - \mathbf{o}_j) \times \mathbf{d}_j\|_2}{\|\mathbf{d}_j\|_2} \tag{5}$$

To further disambiguate structural and textural similarities, we introduce a gradient similarity term $e_{\text{grad}} = 1 - (\nabla I_i^\top \nabla I_j)/(\|\nabla I_i\|\|\nabla I_j\|)$ and a photometric error e_{photo}, computed via the Sum of Absolute Differences (SAD) over a local patch. The comprehensive matching score is thus constructed as:

$$\mathcal{S}(p_i, p_j) = w_r \cdot e_{\text{ray}} + w_g \cdot e_{\text{grad}} + w_p \cdot e_{\text{photo}} \tag{6}$$

where $w_r = 0.6$, $w_g = 0.2$, and $w_p = 0.2$ are the optimal weighting coefficients (evaluated in Sec. 4.5). The target match p_j^* is obtained by minimizing this composite score $\mathcal{S}$. To filter out spurious correspondences, bidirectional symmetry verification is conducted: the 3D point reconstructed from p_j^* is projected back to the source frame. The match is accepted only if the reprojection error satisfies $\sqrt{(u_i^{\text{back}} - u_i)^2 + (v_i^{\text{back}} - v_i)^2} < \tau_{\text{sym}}$, with $\tau_{\text{sym}} = 5.0$.

3.3 Loop Detection and Global Graph Optimization

To balance geometric constraints and computational overhead, we dynamically extract keyframes and detect loop closures using a unified feature similarity metric. For any two frames I_t and I_s, we extract DINOv2 global descriptors f_t and f_s, and compute their visual cosine similarity [32]:

$$\text{sim}(I_t, I_s) = \frac{f_t^\top f_s}{\|f_t\|\|f_s\|} \tag{7}$$

Based on this metric, a new keyframe is triggered when the similarity between the current frame and the last keyframe drops below a threshold $\tau_{kf} = 0.70$. This adaptive strategy ensures dense sampling during complex viewpoint transitions and sparse sampling during smooth linear motion, empirically yielding an average interval of 3 to 5 frames on the KITTI [11] sequences. To recognize revisited areas and close the trajectory loops, a valid loop candidate is identified when the similarity between the current keyframe and a historical keyframe exceeds a strict threshold τ_{sim} (set to 0.85), provided that the temporal interval ($\Delta t > 100$ frames) is sufficient to avoid redundant local constraints.

Upon detecting a loop candidate, we expand the temporal windows to include adjacent frames and leverage DA3 to extract high-confidence relative pose transformations $\hat{\mathbf{T}}_{ij}$. To eliminate the accumulated drift and achieve global consistency, we jointly optimize the entire set of camera poses $\{\mathbf{T}_k \in SE(3)\}$ [28]

using a unified factor graph. The global objective function balances the hybrid ray-field constraints against the multi-frame loop pose priors:

$$E(\{\mathbf{T}_k\}) = \sum_{i,j,p} \rho\left(e_{\text{ray}}(\mathbf{T}_j^{-1}\mathbf{T}_i\mathbf{P}_i, \mathbf{r}_j)\right) + \lambda_{\text{pose}} \sum_{i,j} \rho_{\text{pose}}(\mathbf{T}_i, \mathbf{T}_j, \hat{\mathbf{T}}_{ij}) \quad (8)$$

where the first term ensures geometric consistency across sparse ray-fields by transforming local points $\mathbf{P}_i$ into the target frame j, and the second term anchors the trajectory using the multi-frame predictions, weighted by λ_{pose} (set to 10.0 to balance the point-level scale). The function $\rho(\cdot)$ denotes the robust Huber loss. With the metric-scale initialization provided by the piece-wise trajectory alignment (Sec. 3.1), the Levenberg-Marquardt algorithm converges, yielding a globally consistent reconstruction with reduced drift.

4 Experiments

We evaluate GRF-Recon on large-scale, long-sequence monocular 3D reconstruction using outdoor driving benchmarks and standard depth datasets. Trajectory estimation and scene reconstruction are evaluated on KITTI [11] and the Waymo Open Dataset [29], while depth estimation is evaluated on ETH3D [26], SINTEL [1], and DIODE [35]. All experiments are conducted on a local workstation running Ubuntu 22.04, equipped with an Intel Core i9-13900K CPU, 128 GB RAM, and a single NVIDIA GeForce RTX 3090 GPU (24 GB).

4.1 Experimental Setup

Evaluation Metrics. We adopt standard protocols to assess performance across all tasks. For long-sequence camera tracking, we report the absolute trajectory error as RMSE (ATE RMSE [m] ↓) to measure global consistency and accumulated drift. For monocular depth estimation, we evaluate relative error (Rel ↓) and threshold accuracy (δ_1 ↑), capturing both metric fidelity and scale stability. For large-scale point cloud reconstruction, we measure point-level geometric accuracy using Accuracy (↓), Completeness (↓), and Chamfer Distance (↓). Given the scale ambiguity inherent in monocular predictions, all estimated trajectories and point clouds are aligned to LiDAR ground truth via a global Sim(3) transformation before metric computation. To assess downstream novel-view synthesis quality, we report PSNR and SSIM to measure rendering fidelity.

Data Preprocessing and Ground Truth. During inference, we resize each input image to a width of 504 pixels while preserving the aspect ratio, rounding the other dimension to a multiple of 14 to match the patch partitioning of the ViT-based front end. To ensure fair comparison, we apply a unified point filtering rule: for methods predicting confidence, we retain 3D points with a confidence score greater than 0.7 times the global mean.

In autonomous driving datasets like the Waymo Open Dataset [29], the vehicle-mounted LiDAR ground truth is physically constrained by sensor height and vertical field of view, leaving upper structures such as building tops and

Fig. 3: Qualitative comparison of monocular depth estimation. We compare the baseline DA3 with the LoRA-enhanced model (Ours) on the KITTI dataset. The original DA3, using the DA3 NESTED-GIANT-LARGE-1.1 weights, exhibits over-smoothing and fails to capture thin, high-frequency structures. We fine-tune this model, and experiments show that it can recover sharp geometric details and cleanly separate slender utility poles, foliage, and complex vehicle structures from the background.

tree canopies unobserved. Vision-based feed-forward reconstruction successfully recovers these regions. When evaluated against sparse LiDAR references, these valid geometric reconstructions are often heavily penalized under distance-based metrics such as Accuracy (Fig. 4). Consequently, qualitative visual comparison serves as an essential complement to quantitative evaluation when assessing the true geometric reconstruction capabilities.

4.2 Monocular Depth Estimation

High-fidelity local geometric priors underpin the hybrid-weight ray-field and enable accurate correspondence search. Since LiDAR scans in the original KITTI odometry split are highly sparse, we evaluate depth on the semi-dense ground truth provided by the KITTI Depth [11] (Fig. 3). Zero-shot generalization across diverse scene geometries is assessed on ETH3D [26], SINTEL [1], and DIODE [35].

The perception backbone is initialized from a pre-trained DA3 [16] model. Following a monocular knowledge-distillation paradigm, the decoder responsible for cross-view feature matching remains frozen. Only the image encoder and depth head are updated. To prevent feature distribution shift, we insert a re-normalized low-rank adaptation (Re-normalized LoRA, $r = 8$) layer into the encoder. Fine-tuning runs on a single NVIDIA RTX 3090 with automatic mixed precision (AMP), using a batch size of 16 for 15 epochs ($\approx$ 14 hours). The total parameter increase remains strictly below 3%.

Tab. 1 summarizes the quantitative results. After fine-tuning (DA3+Ours), front-end depth accuracy improves significantly: on KITTI, Rel decreases from 5.14 to 4.50, and δ_1 increases from 92.4 to 97.1. On ETH3D and dynamic high-illumination scenes, the model maintains high stability, reducing average Rel to 6.95 and pushing average δ_1 to 91.4, tightly approaching the teacher model (MoGe [41]). Compared to full fine-tuning, our LoRA-based scheme effectively mitigates overfitting while producing sharp, high-fidelity local geometry. These

Table 1: Quantitative results for monocular depth estimation.

Method	KITTI [11]		ETH3D [26]		SINTEL [1]		DIODE [35]		Avg	
	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑	Rel ↓	δ_1 ↑
DA3 [16]	5.14	92.4	3.25	97.6	16.20	73.5	6.31	94.2	7.73	89.4
DA3+Ours	4.50	97.1	2.48	99.2	15.21	74.3	5.62	94.9	6.95	91.4
MoGe (Teacher) [41]	**4.39**	**97.2**	**2.10**	**99.8**	**11.50**	**81.4**	**4.50**	**96.6**	**5.62**	**93.8**

Table 2: Camera Tracking on KITTI [11] (ATE RMSE [m] ↓). To ensure a fair comparison, the Average (Avg.) is computed by excluding the high-speed Seq. 01, which typically causes catastrophic scale drift in monocular methods. Methods are categorized based on their requirement for camera intrinsic calibration. *TL* and *OOM* denote Tracking Lost and Out-of-Memory, respectively. The top three results are highlighted in **red**, orange, and yellow.

	Method	Avg.	00	01	02	03	04	05	06	07	08	09	10
Metadata	*seq. frames*	2210	4542	1101	4661	801	271	2761	1201	1101	4071	1591	1201
	seq. length (m)	1968.15	3724.19	2453.20	5067.23	560.89	393.65	2205.58	1232.88	649.70	3222.80	1705.05	919.52
	seq. speed (m/f)	0.89	0.82	2.23	1.09	0.70	1.45	0.80	1.12	0.59	0.79	1.07	0.77
	contains loop	-	✓	×	✓	×	×	✓	✓	✓	×	✓	×
Calibrated	ORB-SLAM3 [5]	7.93	5.12	**7.23**	11.50	**1.35**	2.15	4.60	13.41	**1.55**	**25.30**	8.50	**5.85**
	DROID-SLAM [30]	74.88	95.10	334.20	117.31	4.38	2.20	128.50	54.47	15.58	151.60	67.33	112.32
	DPVO [31]	54.79	123.34	8.29	113.34	4.09	1.68	48.56	44.78	20.36	104.21	64.10	23.43
	DPV-SLAM [18]	59.04	113.20	9.50	132.53	2.31	**0.79**	54.80	52.34	16.37	132.44	74.26	11.35
	DPV-SLAM++ [18]	24.25	7.67	10.36	**10.36**	3.45	1.28	5.34	13.57	2.01	124.46	63.64	10.67
Uncalibrated	MASt3R-SLAM [19]	-	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*
	VGGT [38]	-	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
	DA3 [16]	-	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
	VGGT-Long [9]	19.78	11.16	111.36	34.16	6.83	5.16	9.35	6.68	5.23	56.15	42.24	20.87
	Pi-Long [43]	19.01	10.82	170.92	77.59	5.67	0.92	6.13	4.82	3.24	36.84	20.27	23.82
	DA3-Streaming [16]	12.72	9.34	83.64	37.84	4.32	3.92	5.87	3.73	4.83	30.82	10.93	15.57
	GRF-Recon (Ours)	**7.18**	**4.35**	76.52	16.50	1.65	0.85	**3.25**	**2.35**	1.64	26.40	**7.25**	6.55

precise priors constrain the search space for subsequent ray-field optimization, laying a solid foundation for stable long-sequence reconstruction.

4.3 Camera Tracking on Long Sequences

Camera pose estimation is evaluated on KITTI [11] and the Waymo Open Dataset [29]. All compared methods operate under a strictly monocular setting, producing scale-ambiguous trajectories. Prior to evaluation, each predicted trajectory undergoes a similarity transformation to align with the ground truth.

As reported in Tab. 2 and Tab. 3, existing approaches struggle to maintain stability over extended sequences. DROID-SLAM [30] accumulates severe drift on several trajectories. MASt3R-SLAM [19] frequently fails when sudden scene changes enlarge the keyframe baseline, resulting in tracking loss (*TL*). Feed-forward foundation models like VGGT [38] and the original DA3 encounter fatal out-of-memory (*OOM*) errors on long sequences.

As shown in Fig. 4, among long-sequence feed-forward variants, DA3-Streaming maintains reasonable accuracy on most runs but exhibits clear pose discontinuities at street junctions. This limitation stems from inadequate long-range

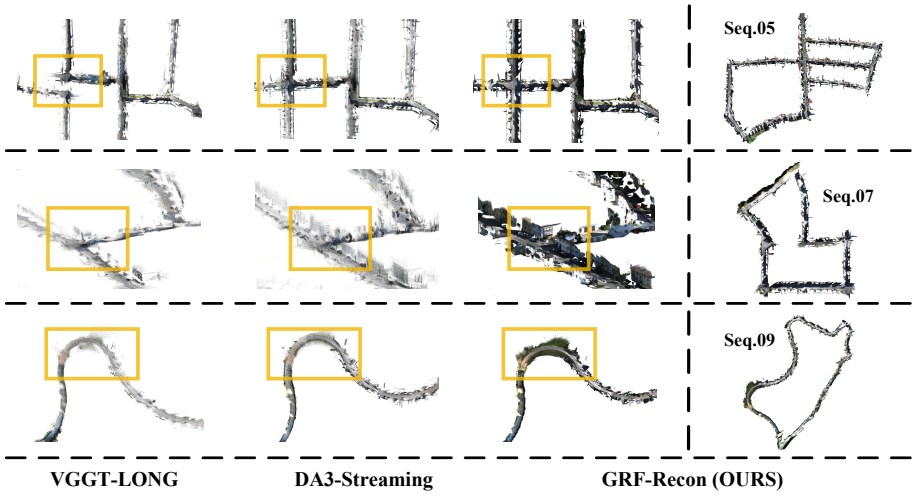

Fig. 4: Qualitative comparison of dense local reconstruction on KITTI.

Table 3: Camera Tracking Results (ATE RMSE [m] ↓) on the Waymo Dataset [29].

Method	Avg.	Seq.163	Seq.183	Seq.315	Seq.346	Seq.371	Seq.405	Seq.460	Seq.520	Seq.610
seq. frames	198	198	199	199	199	196	199	198	199	198
seq. length (m)	172.53	159.96	42.30	165.15	351.21	272.66	85.74	265.91	134.55	62.74
seq. speed (m/f)	0.87	0.81	0.21	0.83	1.77	1.39	0.43	1.34	0.68	0.32
DROID-SLAM [30]	5.43	3.82	**0.31**	0.47	8.82	9.45	7.82	4.25	13.65	**0.28**
MASt3R-SLAM [19]	5.01	4.61	0.59	1.89	12.85	8.75	1.46	5.54	8.12	1.25
CUT3R [40]	9.99	8.95	3.92	5.92	24.52	13.45	7.42	13.55	8.85	3.31
VGGT-Long [9]	2.03	1.83	2.72	0.59	3.55	3.43	1.49	1.59	2.62	0.47
DA3-Streaming [16]	1.76	1.72	1.86	0.55	3.25	2.92	1.39	1.47	2.22	0.42
GRF-Recon (Ours)	**1.39**	**1.49**	1.55	**0.45**	**2.52**	**2.15**	**1.15**	**1.12**	**1.78**	0.29

geometric consistency modeling: DA3-Streaming strictly optimizes inter-chunk pose constraints, leaving intra-chunk poses entirely reliant on feed-forward predictions. GRF-Recon overcomes this by introducing an explicit ray-field representation into the global optimization, directly encoding 3D geometry and coupling it with depth priors to enforce rigid scale constraints. This design allows our method to achieve trajectory accuracy comparable to traditional calibrated SLAM systems [5], entirely without requiring known camera intrinsics.

4.4 Optimization and Reconstruction Quality Analysis

We analyze system efficiency by decomposing the runtime of the frontend pipeline and backend optimization (Tab. 4). A lightweight trajectory-based chunking strategy replaces dense point cloud registration.

By exploiting pipeline parallelism, the computational cost of inter-chunk ray matching ($\approx 0.16\,$s per keyframe) and chunk alignment ($\approx 0.15\,$s per chunk)

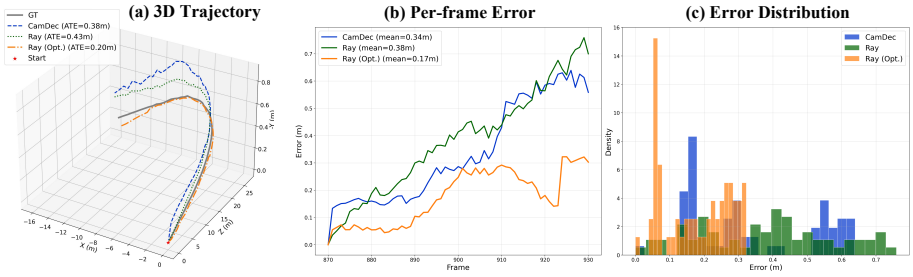

Fig. 5: Evaluation of intra-chunk pose estimation on KITTI sequence 07. Under loop-free conditions, our proposed ray optimization method (Ray (Opt.)) significantly mitigates local drift and improves intra-chunk pose accuracy compared to the DA3 camera decoder (CamDec) and raw ray predictions (Ray).

Table 4: Runtime of GRF-Recon on Key KITTI [11] Sequences. End-to-end runtime excludes model loading and disk I/O. Our parallel pipeline masks the overhead of chunk alignment and ray matching behind feed-forward inference, while a sparse C++ backend minimizes global optimization cost for long sequences.

Seq.	Frames	Frontend Pipeline			Backend & Total		
Chunk Size = 60		VPR Model (s / chunk)	Ray Match (s / keyframe)	Chunk Align (s / chunk)	Loop Det. (ms / q)	LM Opt. (s / seq)	Total Time
KITTI 00	4542	8.1	0.16	0.142	32.5	10.23	11min 45s
KITTI 07	1101	8.3	0.17	0.164	30.7	2.72	2min 41s
KITTI 09	1591	8.4	0.16	0.157	27.8	4.08	4min 03s

is effectively hidden by the dominant feed-forward inference (≈ 8 s per chunk), introducing negligible additional latency to the system.

In the backend, our system filters noise to retain only high-confidence sparse ray constraints, resulting in a highly sparse factor graph. Even for sequences exceeding a thousand frames, the full Levenberg–Marquardt optimization converges within approximately 10 seconds. This synergy between lightweight frontend alignment and sparse backend refinement achieves an optimal balance between trajectory accuracy and computational efficiency under a strict 24 GB memory budget. As illustrated in Fig. 5, we compare the raw ray decoding from DA3 [16] with our refined ray estimation under loop-free conditions. The results demonstrate that frontend optimization effectively regularizes the trajectory and mitigates accumulated drift, confirming that high-quality ray matching inherently provides strong geometric constraints for monocular reconstruction.

Tab. 5 reports the 3D reconstruction results on Waymo [29]. Operating strictly under the uncalibrated monocular setting, GRF-Recon consistently achieves Accuracy, Completeness, and Chamfer Distance across most sequences. The global constraints imposed by our ray-field representation actively mitigate point

Table 5: 3D Reconstruction on Waymo Open Dataset [29]. Performance is evaluated using three geometric metrics: Accuracy ↓, Completeness ↓, and Chamfer Distance ↓. All methods operate under the uncalibrated monocular setting. The best result for each metric is highlighted with **bold and underline**.

Method	Metric	Seq.163	Seq.183	Seq.315	Seq.346	Seq.371	Seq.405	Seq.460	Seq.520	Seq.610
MASt3R-SLAM [19]	Accuracy ↓	3.420	3.250	4.050	4.950	4.850	1.250	5.020	6.850	2.850
	Completeness ↓	**1.850**	3.550	2.220	3.250	2.920	3.150	2.180	4.820	7.150
	Chamfer ↓	2.650	3.320	3.150	4.150	3.880	2.220	3.550	5.850	4.950
VGGT-Long [9]	Accuracy ↓	1.120	0.420	1.020	1.820	2.750	0.740	0.850	1.480	1.350
	Completeness ↓	2.950	3.680	1.880	3.550	3.050	3.480	2.010	5.050	2.220
	Chamfer ↓	2.080	2.050	1.450	2.680	2.920	2.120	**1.420**	3.250	1.780
DA3-Streaming [16]	Accuracy ↓	1.050	**0.380**	0.960	1.680	2.250	0.710	0.820	1.360	1.180
	Completeness ↓	2.680	3.350	1.680	3.150	2.780	3.050	1.820	4.450	2.020
	Chamfer ↓	1.820	1.900	1.350	2.450	2.650	1.920	1.580	2.980	1.580
GRF-Recon (Ours)	Accuracy ↓	**0.910**	0.410	**0.820**	**1.350**	**1.980**	**0.620**	**0.740**	**1.150**	**1.020**
	Completeness ↓	1.980	**3.050**	**1.450**	**2.850**	**2.450**	**2.620**	**1.600**	**4.150**	**1.800**
	Chamfer ↓	**1.560**	**1.680**	**1.250**	**2.120**	**2.320**	**1.680**	1.480	**2.550**	**1.350**

Table 6: Ablation study of GRF-Recon (ATE RMSE [m] ↓). The entry *(0,0,0)* under the Ray Comp. column indicates that the ray-matching module is completely removed. w/o Opt. disables both loop closure and ray matching, and the trajectory is obtained by directly concatenating sequential camera poses. The Baseline uses the default configuration with $C/O = 60/10$.

Dataset	Baseline (Full)	w/o LoRA	Sparse KF	Ray Comp. (w_r, w_g, w_p)				w/o Loop	w/o Opt.	Config (C/O)		
				(1,1,0)	*(1,0,1)*	*(0,1,1)*	*(0,0,0)*			*30/10*	*60/10*	*60/20*
KITTI 00	**4.35**	6.80	5.45	5.62	6.95	6.42	7.35	8.12	12.35	4.41	4.35	**4.32**
KITTI 07	**1.64**	2.65	2.15	2.45	2.92	3.12	4.85	5.85	6.65	**1.61**	1.64	1.66
Seq. 346	**2.52**	3.10	3.15	3.52	3.25	3.55	6.60	6.85	7.40	2.58	2.52	**2.49**
Seq. 371	**2.15**	2.80	2.74	3.25	3.85	3.15	4.25	6.10	8.20	2.18	**2.15**	2.17

cloud layering and ghosting artifacts over long trajectories, yielding structurally accurate and geometrically consistent scene reconstructions.

4.5 Ablation Study and Applications

Tab. 6 details our ablation study, quantifying the contribution of each core module. Front-end depth fine-tuning (w/o LoRA) and the hybrid-weight ray-matching module prove essential. Completely removing the ray matching module (*(0,0,0)*) or disabling the global backend optimization (w/o Opt.) forces the system to degrade into pure sequential camera concatenation, triggering a drastic increase in ATE RMSE. This confirms that both components are indispensable for suppressing drift and preserving geometric consistency.

The spatio-temporal configuration ablation (C/O) demonstrates that, equipped with precise geometric constraints and a robust backend optimizer, the system exhibits robustness to varying chunk sizes and overlap ratios. To balance memory

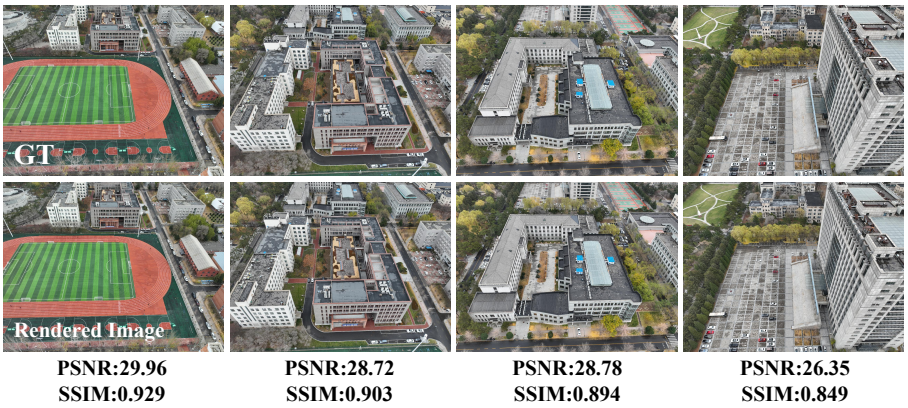

Fig. 6: Novel view synthesis on large-scale uncalibrated UAV sequences. We use the point clouds reconstructed by GRF-Recon to initialize 3D Gaussian Splatting, eliminating the need for a separate COLMAP-based structure-from-motion stage. With this geometry-aware initialization, 3DGS converges within approximately 1,000 iterations while maintaining accurate structural consistency.

footprint and computation speed, we select $C/O = 60/10$ as the default baseline, securing a practical trade-off between performance and resource consumption.

We deploy GRF-Recon for UAV-based urban building reconstruction (Fig. 6). Operating in GPS-denied environments, the system achieves robust trajectory closure and generates high-fidelity dense point clouds. Utilizing these point clouds to initialize 3D Gaussian Splatting [14] accelerates convergence compared to traditional SfM-based initialization (e.g., COLMAP [25]) and can effectively mitigate artifacts in edges and fine structures, demonstrating its considerable potential for downstream dense neural SLAM [6] and rendering tasks.

5 Conclusion

We present GRF-Recon, a scalable framework for uncalibrated monocular 3D reconstruction over long sequences. By distilling high-fidelity geometric priors through re-normalized LoRA adaptation, we mitigate local depth degradation. To address long-term drift under strict memory constraints, we introduce a hybrid-weight sparse ray-field optimization. Experiments on autonomous driving datasets show that GRF-Recon outperforms existing feed-forward models, achieving trajectory accuracy that surpasses traditional calibrated SLAM systems in selected scenarios. The method offers a practical, geometry-aware initialization for downstream applications, including UAV-based urban mapping, virtual reality, and rapid digital twin generation via 3D Gaussian Splatting. Although robust in rigid environments, it remains challenged by highly dynamic objects. Future work will focus on improving reconstruction accuracy and real-time consistency in complex scenes.

References

1. Baker, S., Scharstein, D., Lewis, J.P., Roth, S., Black, M.J., Szeliski, R.: A database and evaluation methodology for optical flow. Int. J. Comput. Vis. **92**(1), 1–31 (2011)
2. Birkl, R., Wofk, D., Müller, M.: MiDaS v3.1 – a model zoo for robust monocular relative depth estimation (2023), `https://arxiv.org/abs/2307.14460`, Accessed: 29 June 2026
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token Merging: Your vit but faster. In: Int. Conf. Learn. Represent. (2023)
4. Cabon, Y., Stoffl, L., Antsfeld, L., Csurka, G., Chidlovskii, B., Revaud, J., Leroy, V.: Must3r: Multi-view network for stereo 3D reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1050–1060 (2025)
5. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE Trans. Robotics **37**(6), 1874–1890 (2021)
6. Chen, X., Zhang, Y., Zhang, Z., Wang, G., Zhao, B., Wang, X.: VSS-SLAM: Voxelized surfel splatting for geometally accurate slam. In: IEEE Int. Conf. Robot. Autom. pp. 139–145 (2025)
7. Chen, Z., Qin, M., Yuan, T., Liu, Z., Zhao, H.: Long3r: Long sequence streaming 3D reconstruction. In: Int. Conf. Comput. Vis. pp. 5273–5284 (2025)
8. Czarnowski, J., Laidlow, T., Clark, R., Davison, A.J.: Deepfactors: Real-time probabilistic dense monocular slam. IEEE Robotics and Automation Letters **5**(2), 721–728 (2020)
9. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: VGGT-Long: Chunk it, loop it, align it – pushing VGGT's limits on kilometer-scale long RGB sequences. In: IEEE Int. Conf. Robot. Autom. (2026)
10. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 224–236 (2018)
11. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3354–3361 (2012)
12. Hartley, R., Zisserman, A.: Multiple view geometry in computer vision. Cambridge university press (2003)
13. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: Int. Conf. Learn. Represent. (2022)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3D gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139:1–139:14 (2023)
15. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: Eur. Conf. Comput. Vis. pp. 71–91 (2024)
16. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Zhao, Y., Peng, S., Guo, H., Zhou, X., Shi, G., Feng, J., Kang, B.: Depth Anything 3: Recovering the visual space from any views. In: Int. Conf. Learn. Represent. (2026)
17. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: Lightglue: Local feature matching at light speed. In: Int. Conf. Comput. Vis. pp. 17581–17592 (2023)
18. Lipson, L., Teed, Z., Deng, J.: Deep patch visual slam. In: Eur. Conf. Comput. Vis. pp. 424–440 (2024)

19. Murai, R., Dexheimer, E., Davison, A.J.: Mast3r-slam: Real-time dense slam with 3D reconstruction priors. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16695–16705 (2025)
20. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: DTAM: Dense tracking and mapping in real-time. In: Int. Conf. Comput. Vis. pp. 2320–2327 (2011)
21. Oquab, M., Darcet, T., Moutakanni, T., Vo, H.V., Szafraniec, M., Khalidov, V., Fernandez, P., HAZIZA, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P.Y., Li, S.W., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jegou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P.: DINOv2: Learning robust visual features without supervision. Trans. Mach. Learn Res. (2024), featured Certification
22. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.: DynamicViT: Efficient vision transformers with dynamic token sparsification. In: Adv. Neural Inform. Process. Syst. pp. 13937–13949 (2021)
23. Ren, W., Wang, H., Tan, X., Han, K.: Fin3R: Fine-tuning feed-forward 3D reconstruction models via monocular knowledge distillation. In: Adv. Neural Inform. Process. Syst. (2025)
24. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: Superglue: Learning feature matching with graph neural networks. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4937–4946 (2020)
25. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4104–4113 (2016)
26. Schops, T., Schonberger, J.L., Galliani, S., Sattler, T., Schindler, K., Pollefeys, M., Geiger, A.: A multi-view stereo benchmark with high-resolution images and multi-camera videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2538–2547 (2017)
27. Shen, Y., Zhang, Z., Qu, Y., Zheng, X., Ji, J., Zhang, S., Cao, L.: FastVGGT: Fast visual geometry transformer. In: Int. Conf. Learn. Represent. (2026)
28. Solà, J., Deray, J., Atchuthan, D.: A micro lie theory for state estimation in robotics (2021), https://arxiv.org/abs/1812.01537, Accessed: 29 June 2026
29. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., Anguelov, D.: Scalability in perception for autonomous driving: Waymo open dataset. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2443–2451 (2020)
30. Teed, Z., Deng, J.: DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In: Adv. Neural Inform. Process. Syst. pp. 16558–16569 (2021)
31. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. In: Adv. Neural Inform. Process. Syst. (2023)
32. Tolias, G., Avrithis, Y., Jégou, H.: To aggregate or not to aggregate: Selective match kernels for image search. In: Int. Conf. Comput. Vis. pp. 1401–1408 (2013)
33. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment - A modern synthesis. In: Workshop on Vision Algorithms. Lecture Notes in Computer Science, vol. 1883, pp. 298–372. Springer (1999)
34. Tyszkiewicz, M.J., Fua, P., Trulls, E.: DISK: Learning local features with policy gradient. In: Adv. Neural Inform. Process. Syst. (2020)
35. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., Shakhnarovich, G.: DIODE: A dense indoor and outdoor depth dataset (2019), https://arxiv.org/abs/1908.00463, Accessed: 29 June 2026

36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Adv. Neural Inform. Process. Syst. pp. 5998–6008 (2017)
37. Wang, H., Agapito, L.: 3D reconstruction with spatial memory. In: 2025 International Conference on 3D Vision (3DV). pp. 78–89 (2025)
38. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5294–5306 (2025)
39. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: Vggsfm: Visual geometry grounded deep structure from motion. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21686–21697 (2024)
40. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3D perception model with persistent state. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10510–10522 (2025)
41. 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: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5261–5271 (2025)
42. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3D vision made easy. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20697–20709 (2024)
43. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. In: Int. Conf. Learn. Represent. (2026)
44. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3r: Towards 3D reconstruction of 1000+ images in one forward pass. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21924–21935 (2025)
45. Zhang, Y., Tosi, F., Mattoccia, S., Poggi, M.: Go-slam: Global optimization for consistent 3D instant reconstruction. In: Int. Conf. Comput. Vis. pp. 3704–3714 (2023)