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

VGGT-Long: Chunk it, Loop it, Align it – Pushing VGGT's Limits on Kilometer-scale Long RGB Sequences

Kai Deng[1], Zexin Ti[2], Jiawei Xu[1], Jian Yang[2], Jin Xie[2†]

***Abstract*— Foundation models for 3D vision have recently demonstrated remarkable capabilities in 3D perception. However, extending these models to large-scale RGB stream 3D reconstruction remains challenging due to memory limitations. In this work, we propose VGGT-Long, a simple yet effective system that pushes the limits of monocular 3D reconstruction to kilometer-scale, unbounded outdoor environments. Our approach addresses the scalability bottlenecks of existing models through a chunk-based processing strategy combined with overlapping alignment and lightweight loop closure optimization. Without requiring camera calibration, depth supervision or model retraining, VGGT-Long achieves trajectory and reconstruction performance comparable to traditional methods. We evaluate our method on KITTI, Waymo, and Virtual KITTI datasets. VGGT-Long not only runs successfully on long RGB sequences where foundation models typically fail, but also produces accurate and consistent geometry across various conditions. Our results highlight the potential of leveraging foundation models for scalable monocular 3D scene in real-world settings, especially for autonomous driving scenarios. Code is available at `https://github.com/DengKaiCQ/VGGT-Long`.**

I. Introduction

Perceiving 3D environments from monocular RGB streams is crucial for autonomous driving, yet existing methods struggle with kilometer-scale and uncalibrated sequences. Unlike the small-scale indoor 3D vision tasks, driving scenarios involve long trajectories with sparse frame correspondence, dynamic objects and challenging outdoor conditions. While some approaches [1], [2], [3] handle large-scale monocular scenes, they often depend on sophisticated multi-module pipelines or assume known camera intrinsics. Others leverage additional sensors (LiDAR [4], IMU [5] or stereo [6]), sidestepping the core challenge: scalable, calibration-free reconstruction from monocular RGB alone and it is a critical for autonomous systems.

A recent paradigm shift in 3D vision has witnessed the rise of end-to-end foundation models, largely based on the Transformer architecture [7]. A mainly of works from DUSt3R [8] and MASt3R [9] to CUT3R [10], Fast3R [11], and most recently VGGT [12], aim to replace complex, multi-component SfM and SLAM pipelines with a single and unified deep learning model. These models are trained on massive datasets to integrate camera pose estimation, intrinsic parameter regression, and 3D scene representation (typically as a point map) into one cohesive framework. A key goal is to enable backpropagation of errors through the entire system, creating a powerful and versatile foundation model for 3D reconstruction that operates on raw and uncalibrated RGB inputs. However, foundation models like CUT3R and Fast3R still struggle with severe drift in outdoor environments, even on short sequences of a few dozen frames, which limits their practical applicability. In contrast, VGGT delivers remarkably stable and accurate local reconstructions, establishing it as the state-of-the-art in terms of reconstruction quality. Its primary limitation is not performance, but its immense computational and memory footprint.

The computational and memory demands of Transformer based foundation models severely limit their scalability. Standard self-attention [7] scales quadratically with input size, and while techniques like Flash-Attention [13], [14] reduce compute complexity to linear. However, GPU memory requirement still remains prohibitive. For example, VGGT can just process 60 to 80 images on a 24 GiB RTX 4090 GPU scaling to a KITTI Seq 00 trajectory (about 4,600 frames) would require unreachable GPU memory requirement, far exceeding current hardware. This bottleneck confines such models to small-scale scenes, as both memory and drift accumulation become intractable over long sequences.

Our work is inspired by recent efforts to integrate foundation models into large-scale systems. A notable example is MASt3R-SLAM [15], which builds SLAM system on top of the MASt3R [9] model. To achieve global consistency, it employs pose graph optimization and bundle adjustment within its backend, which are standard components in modern SLAM systems.

This raises a fundamental question: must large-scale reconstruction always equate to system-level complexity? Our philosophy diverges significantly from this trend. We advocate for a minimalist approach that unlocks the inherent potential of the foundational model itself. We posit that VGGT is already a remarkably powerful engine for large-scale 3D perception, and the primary challenge is not a lack of capability, but a lack of scalability. Instead of building another intricate system around it, we ask: can we solve the problem with the minimal overhead?

To this end, we propose VGGT-Long, a framework that extends VGGT to long sequences through a simple yet effective framework that is processing the sequence in overlapping chunks, robustly aligning adjacent chunks, and correcting for drift using a high-quality loop closure module. This "chunk-

This work was supported by the National Key R&D Program of China No. 2024YFC3015801, National Science Fund of China under Grant Nos. U24A20330, 62361166670, and 62276144.

[1] College of Computer Science, Nankai University, China
[2] School of Intelligence Science and Technology, Nanjing University, China
[†] Corresponding author, email: csjxie@nju.edu.cn

Fig. 1. For large-scale outdoor scenarios, previous work suffers from: 1) severe drift (CUT3R and Fast3R); 2) unable to complete the entire long sequence (MASt3R-SLAM and VGGT). Our method VGGT-Long is able to reconstruct the kilometer-scale scene while maintaining the accuracy of the scene.

and-align" paradigm avoids the need for a graph-based optimization backend (such as bundle adjustment [16]). It is a testament to the power of the underlying VGGT model, demonstrating that with the right strategy, its exceptional local reconstruction capabilities can be seamlessly stitched together to form a globally consistent, kilometer-scale map. Our work champions the idea that, **a sufficiently powerful base model may not necessarily require a complex backend system to assist**.

In summary, our contributions are as follows:

1) We present the first system that successfully extends monocular 3D reconstruction models to kilometer-scale, unbounded outdoor scenes, without requiring camera calibration and depth supervision.
2) We introduce a simple yet effective chunk-and-align pipeline that resolves the memory limitations of foundation models like VGGT on long video sequences, while achieving accuracy comparable to traditional methods with calibrated cameras.
3) We address the accumulated Sim(3) drift problem inherent in processing long sequences with local models, demonstrating that VGGT can serve as a robust front-end for a large-scale reconstruction system without requiring a complex backend.

II. Related Work

Structure-from-Motion (SfM). SfM methods estimate camera poses and sparse 3D structure from multi-view images. Classical SfM pipelines [17], [18], [19] typically follow an incremental strategy: detecting keypoints [20], [21], matching features [22], [23] and refining poses through bundle adjustment [16]. While robust, these pipelines rely heavily on features extraction and are limited in textureless or ambiguous scenes. Recent deep learning methods aim to enhance or replace traditional modules. Hybrid frameworks such as PixSfM [24] and DFSfM [25] combine deep features with classical optimization to refine both tracks and structure. Fully differentiable SfM pipelines [26], [27], [28] further explore end-to-end learning of camera poses and depth but often suffer from scalability issues or poor generalization. VGGSfM [29] demonstrates that learned systems can surpass classical SfM on real-world datasets by integrating dense features and multi-view consistency into a unified framework.

SLAM and Visual Odometry. To overcome the scalability issue of foundation models, researchers have explored to integrate them into larger SLAM systems. Traditional SLAM systems [30], [31] rely on handcrafted features and optimization, while learning-based approaches [32], [33] integrate differentiable components into deep networks. However, these methods either scale poorly to long sequences or require pre-calibrated camera. MASt3R-SLAM [15] builds a sophisticated real-time SLAM framework around the MASt3R [9] without the calibration, employing complex backend machinery such as pose graph optimization and bundle adjustment to ensure global consistency. Concurrent work VGGT-SLAM [34] introduces a complete SLAM system that aligns submaps via SL(4)-based factor graph optimization for accurate indoor reconstruction. In contrast, our VGGT-Long targets large-scale outdoor scenes using a lightweight pipeline with Sim(3) transformations, prioritizing simplicity and scalability over a full SLAM framework.

Transformer based 3D Vision Method. A recent trend of 3D vision is the development of end-to-end foundation models, predominantly based on the Transformer architecture. These methods target dense geometry reconstruction from overlapping images, assuming unknown poses and unknown camera calibration. A pioneering line of work, including DUSt3R [8] and its successor MASt3R [9], demonstrates the feasibility of jointly estimating camera parameters and dense 3D geometry from uncalibrated image pairs. Subsequent models like CUT3R [10] and Fast3R [11] further refined this paradigm. Most recently, VGGT [12] obtained a new state-of-the-art in reconstruction quality, producing remarkably stable and accurate local 3D maps from raw RGB inputs. However, a common limitation of these models is their significant computational and memory cost, which restricts their application to short image sequences.

Our work, VGGT-Long, distinguishes itself by adopting a minimalist philosophy. Instead of building a complex system, we focus on unlocking the full potential of the powerful VGGT model itself, using a simple yet effective chunk-and-align framework to extend its capabilities to long-sequence,

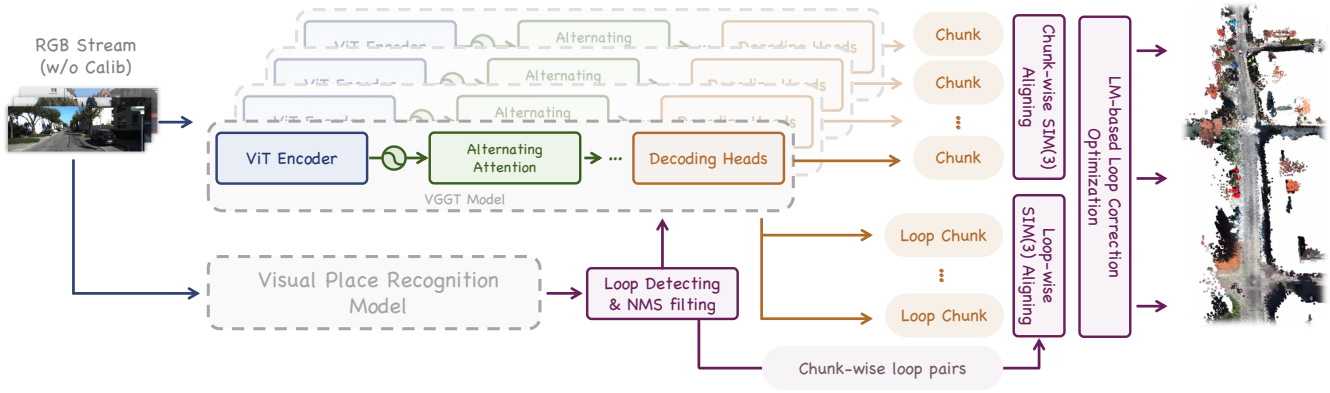


Fig. 2. Overview of VGGT-Long. VGGT-Long processes long sequences by dividing them into different chunks, thereby handling the input RGB stream in a sliding window manner. We fully utilize VGGT's pointmap and confidence to perform lightweight loop closure and alignment on the output chunks, thus extending VGGT to long-sequence datasets for autonomous driving.

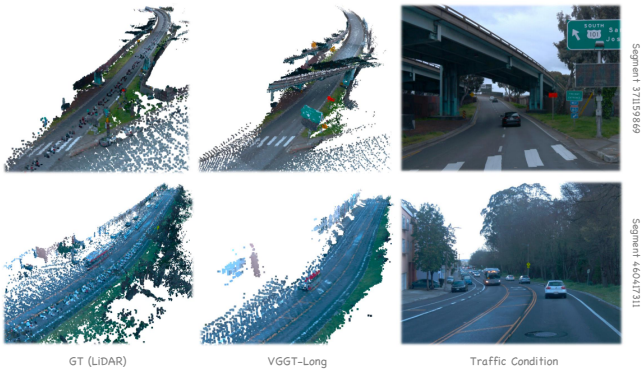


Fig. 3. Confidence-aware alignment suppress the influence of high-speed dynamic objects (such as vehicles) on alignment and reconstruction. It could be observed that higher-density vehicles cannot be effectively filtered out by the LiDAR, but VGGT-Long has the ability to handle this situation.

large-scale scenarios with the minimal overhead.

III. Method

Our proposed method, VGGT-Long, processes long monocular RGB sequences by decomposing the problem into three stages: chunking, chunk-wise alignment, loop closure and loop correction. Our method maintains the local accuracy of VGGT while ensuring global accuracy when handling the outdoor long sequences.

A. Sequence Chunking and Local Aligning with Confidence

Given a long image sequence $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_N\}$, we partition it into K overlapping chunks. Let L be the chunk size and O be the overlap size. The k-th chunk, for $k = 1, \ldots, K$, is defined as the subsequence of frames indexed from $(k-1)(L-O)$ to $(k-1)(L-O)+L$. Each chunk $\mathcal{C}_k$ is processed independently by the VGGT model [12], which outputs a set of locally consistent camera poses and a 3D point map $\mathbf{P}_k \in \mathbb{R}^{H\times W\times 3}$ with the corresponding confidence values $\mathbf{c}_k \in \mathbb{R}^{H\times W}$.

We fully leverage the confidence VGGT outputs to perform chunk alignment. In VGGT, the output confidence reflects the model's certainty about the point cloud. More specifically, in outdoor environments, objects such as the sky, oncoming vehicles, fast-moving pedestrians or raindrops during rainy conditions can affect the SIM(3) alignment calculation to some extent. VGGT's final output assigns lower confidence to these points, while assigning higher confidence to static scenes (such as buildings or stationary vehicles). Our alignment strategy is designed to achieve robust alignment based on this principle.

For each pair of adjacent chunks, $\mathcal{C}_k$ and $\mathcal{C}_{k+1}$, we identify a set of 3D point correspondences $\{(\mathbf{p}_k^i, \mathbf{p}_{k+1}^i)\}$ and $\{(\mathbf{c}_k^i, \mathbf{c}_{k+1}^i)\}$ within their overlapping region. To robustly estimate the relative Sim(3) transformation $\mathbf{S}_{k,k+1} \in \mathrm{Sim}(3)$ that aligns $\mathcal{C}_{k+1}$ to $\mathcal{C}_k$, we employ an Iteratively Reweighted Least Squares (IRLS) optimization. The objective is to minimize the following robust cost function

$$\mathbf{S}_{k,k+1}^* = \arg\min_{\mathbf{S}\in \mathrm{Sim}(3)} \sum_i \rho\left(\|\mathbf{p}_k^i - \mathbf{S}\mathbf{p}_{k+1}^i\|_2\right), \quad (1)$$

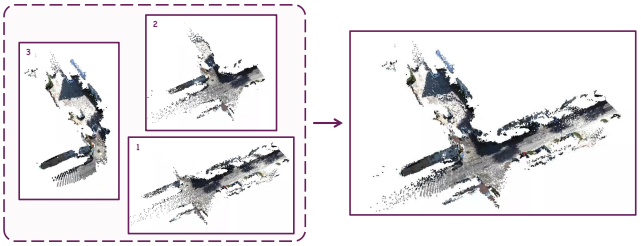
where $\rho(\cdot)$ is the Huber loss function, which down-weights the influence of outliers. The IRLS procedure solves this non-linear problem by iteratively minimizing a weighted sum of squared errors

$$\mathbf{S}^{(t+1)} = \arg\min_{\mathbf{S}\in \mathrm{Sim}(3)} \sum_i \mathbf{w}_i^{(t)} \|\mathbf{p}_k^i - \mathbf{S}\mathbf{p}_{k+1}^i\|_2^2. \quad (2)$$

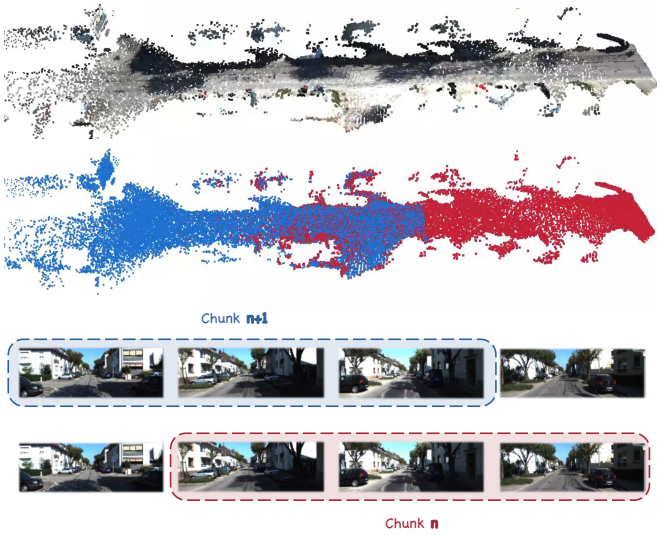
At each iteration t, the weight $\mathbf{w}_i^{(t)}$ for the i-th correspondence is a product of the model's confidence $\mathbf{c}_i$ and a robustness term derived from the Huber loss

$$\mathbf{w}_i^{(t)} = \mathbf{c}_i \cdot \frac{\rho'(r_i^{(t)})}{r_i^{(t)}}, \quad (3)$$

where $r_i^{(t)} = \|\mathbf{p}_k^i - \mathbf{S}^{(t)}\mathbf{p}_{k+1}^i\|_2$ is the residual from the previous iteration. Each weighted least-squares problem is solved efficiently in closed form using a weighted version of the Umeyama algorithm. Through this approach, we exclude low-confidence points (e.g., rapidly moving objects and sky



(a) The visualization of different chunks.



(b) The visualization of overlapping frames.

Fig. 4. (a) VGGT-long divides a kilometer-scale sequence into different chunks for processing. (b) The alignments are derived from the consistency of overlapping frames in 3D space.

regions) as outliers (see Fig. 3). For medium-confidence objects (e.g., slowly moving vehicles), we assign reduced alignment weights, while concentrating higher weights on high-confidence structures (e.g., buildings). In our implementation, we directly discard points with confidence values below $0.1\times$ the median confidence of the entire chunk. The remaining low-confidence points that survive this filtering contribute minimally to the alignment process due to their attenuated weights.

B. Loop Detection and Loop-wise SIM(3) Aligning

To correct the accumulated drift inherent in sequential estimation, we perform loop closure detection across the entire sequence. This process involves identifying non-adjacent chunks as the same scene and robustly estimating the Sim(3) transformation between them.

First, we use a pre-trained Visual Place Recognition (VPR) model [35], which leverages a DINOv2 backbone [36], to extract a compact and descriptive global feature vector $\mathbf{d}_i$ for each image $\mathbf{I}_i$ in the sequence. These descriptors capture the high-level semantic and geometric content of the images.

With global descriptors for all images, we identify potential loop closure candidates. For each descriptor, we perform an efficient nearest neighbor search to find other images with

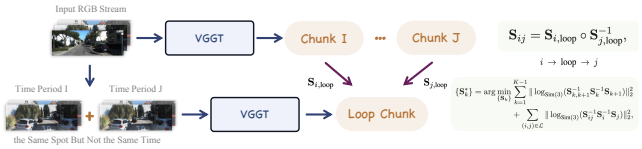


Fig. 5. Loop-wise Sim(3) Alignment

high cosine similarity. A pair of images $(\mathbf{I}_i, \mathbf{I}_j)$ is considered as a potential loop closure if their similarity score exceeds a threshold τ_s and their frame indices are sufficiently separated, i.e., $|i-j| > \Delta t_{min}$. To ensure that the detected loops are distinct and to avoid redundant matches in temporally close frames, we apply Non-Maximum Suppression (NMS). This filtering step selects the strongest match within a local time window, yielding a set of high-confidence image-level loop pairs.

For each validated loop pair $(\mathbf{I}_i, \mathbf{I}_j)$, we adopt a specialized strategy to generate a high-quality local reconstruction of the looped scene. We form a new, temporary image batch by concatenating sub-sequences of frames centered around the indices i and j. This batch, containing temporally disjoint views of the same location, is then processed by the VGGT model. Unlike the sequential, sliding-window processing described in Sec. III-A, this approach provides VGGT with a more diverse, time-dispersed perspective, enabling a more robust reconstruction of the scene's geometry by leveraging a wider baseline.

The resulting 3D point map, which we can call it the "loop-centric" chunk (see Figure 5), is then aligned with the original point maps of the corresponding chunks $\mathcal{C}_i$ and $\mathcal{C}_j$ (which were generated from temporally continuous frames). To compute the final loop-closing transformation $\mathbf{S}_{ij}$, we chain the alignments through this new chunk. Specifically, we compute the transformations from the original chunks to the loop-centric chunk and then compose them

$$\mathbf{S}_{ij} = \mathbf{S}_{i,\text{loop}} \circ \mathbf{S}_{j,\text{loop}}^{-1}, \tag{4}$$

where $\mathbf{S}_{i,\text{loop}}$ and $\mathbf{S}_{j,\text{loop}}$ are the transformations that align the loop-centric chunk to the coordinate frames of chunks $\mathcal{C}_i$ and $\mathcal{C}_j$ respectively (described in Sec. III-A). The composition $\mathbf{S}_{i,\text{loop}} \circ \mathbf{S}_{j,\text{loop}}^{-1}$ effectively chains the transformations to compute the direct alignment from chunk $\mathcal{C}_i$ to $\mathcal{C}_j$ (i.e., $i \to \text{loop} \to j$). This provides a robust geometric constraint for the global optimization by bridging the two distant chunks through a shared, high-quality local reconstruction.

C. Global SIM(3) LM-based Optimization

To achieve global consistency, We follow [37], [38], [3], performing a global optimization over the Sim(3) transformations of all chunks. Instead of constructing a complex factor graph, we directly minimizes a non-linear least-squares objective function composed of two types of geometric constraints: sequential constraints from adjacent chunks (Sec. III-A) and loop closure constraints from non-adjacent chunks

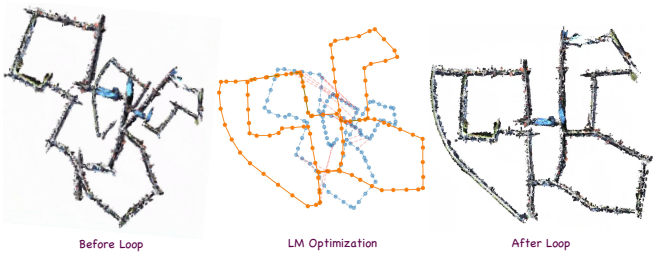


Fig. 6. Without loop constraints, errors will be accumulated continuously at the kilometer scale. The use of Global LM Optimization can alleviate this accumulated error.

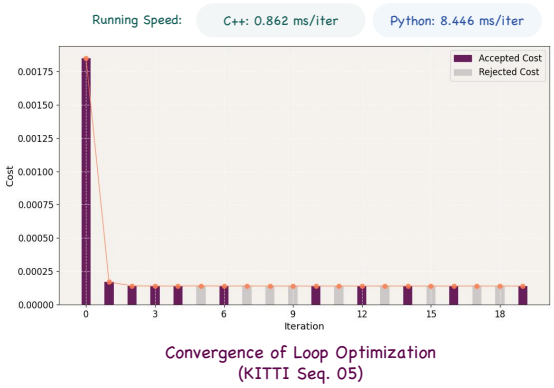


Fig. 7. The loop optimization converges in just 3 iterations and can achieve millisecond-level performance.

(Sec. III-B). The goal is to jointly optimize the transformations $\{\mathbf{S}_k\}_{k=1}^{K}$ for all chunks to be maximally consistent with these relative measurements. The optimization problem is formulated as

$$\{\mathbf{S}_k^*\} = \arg\min_{\{\mathbf{S}_k\}} \sum_{k=1}^{K-1} \| \log_{\mathrm{Sim}(3)}(\mathbf{S}_{k,k+1}^{-1}\mathbf{S}_k^{-1}\mathbf{S}_{k+1})\|_2^2 + \sum_{(i,j)\in\mathcal{L}} \| \log_{\mathrm{Sim}(3)}(\mathbf{S}_{ij}^{-1}\mathbf{S}_i^{-1}\mathbf{S}_j)\|_2^2, \tag{5}$$

where $\mathbf{S}_{k,k+1}$ is the relative transformation between adjacent chunks, $\mathbf{S}_{ij}$ is the loop closure transformation between chunks i and j, and $\mathcal{L}$ is the set of all loop closures. The $\log_{\mathrm{Sim}(3)}(\cdot)$ map converts a Sim(3) transformation into its 7-dimensional tangent space representation in the Lie algebra $\mathfrak{sim}(3)$, allowing for unconstrained optimization. We solve this non-linear least-squares problem efficiently using the Levenberg-Marquardt (LM) algorithm. The global optimization operates chunk-wise, maintaining a small set of SIM(3) variables (typically dozens even for KITTI, see Fig. 6). Convergence is achieved within few iterations, with processing times in milliseconds (see Fig. 7).

IV. Experiments

We evaluate VGGT-Long on three datasets, KITTI Dataset Odometry Track [41], Waymo Open Dataset (v1.4.1) [42], and Virtual KITTI Dataset (v1.3.1) [43] to assess its performance in large-scale monocular 3D reconstruction. Our experiments are conducted on the computer with Ubuntu 22.04, and equipped with 12 Intel Xeon Gold 6128 3.40 GHz CPUs, 67GiB of RAM. In this paper, most experiments were conducted using an NVIDIA RTX 4090 GPU with 24 GiB of VRAM, while for experiments with a chunk size of 90 or larger, we used an NVIDIA L20 GPU with 48 GiB of VRAM.

A. Metrics

We report four different metrics. For tracking performance, following the recent literature, we use the Absolute Trajectory Error (ATE) metric [44] to evaluate the model's long-term consistency over extended sequences. For reconstruction metrics, we adopt the settings from VGGT [12] and employ: 1) Accuracy: Euclidean distance from each predicted point to its nearest Ground Truth (GT) point; 2) Completeness: Euclidean distance from each GT point to its nearest predicted point; 3) Chamfer Distance: The average of the above two metrics. Due to the scale ambiguity inherent in monocular 3D methods, we first perform coarse alignment between the predicted poses and GT poses before calculating the reconstruction metrics. Subsequently, we refine the alignment using the Point-to-Point Iterative Closest Point (ICP) method [45] and then compute these metrics.

B. Experiments Settings

For the calculation of reconstruction performance metrics, to avoid interference from outlier point clouds, we uniformly retain points with confidence values greater than $0.75\times$ the average confidence for Transformer-based 3D methods. As for models like DROID-SLAM, which do not output confidence values, we retain points with depth values larger than $0.75\times$ the average inverse depth as the filtered point cloud.

Following the VGGT's configuration, we downsample input images to 518-pixel width while the hight maintaining aspect ratio during inference. Our inference pipeline first processes all images through VPR, then releases the VPR model's memory occupancy to allocate sufficient GPU memory resources for VGGT.

C. CPU Memory Management Strategy

To handle large-scale scenarios where CPU memory cannot accommodate all VGGT outputs, we implement a memory-efficient strategy: after processing each chunk, we store the results on disk. During subsequent alignment phases, we selectively load only relevant chunk pairs into CPU memory for SIM(3) computation, and immediately freeing the memory after calculation. This design offloads CPU memory pressure to disk storage, and will effectively prevents the operating system crashes, freezes, or other catastrophic failures caused by excessive CPU memory consumption. Upon completing computations, the model purges all intermediate results to minimize the storage overhead. The final outputs (colored point clouds and camera poses) employ stream writing techniques to bypass CPU memory constraints during large-scale reconstruction, effectively

TABLE I

CAMERA TRACKING RESULTS (ATE RMSE [M] ↓) ON THE KITTI DATASET. COLOR DEFINITION: **[FIRST]**, **(SECOND)**, THIRD.

	Methods	LC	Calibration	Recon.	Avg.	Avg.*	00	01	02	03	04	05	06	07	08	09	10
	seq. frames	-	-	-	*2109*	*2210*	*4542*	*1101*	*4661*	*801*	*271*	*2761*	*1101*	*1101*	*4071*	*1591*	*1201*
	seq. length (m)	-	-	-	*2012.243*	*1968.147*	*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 / frame)	-	-	-	*0.95*	*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	-	-	-	-	-	✓	✗	✓	✗	✗	✓	✓	✓	✗	✓	✗
Classic	ORB-SLAM2 (w/o LC) [31]	✗	*Required*	*Sparse*	69.727	26.480	40.65	502.20	47.82	**[0.94]**	1.30	29.95	40.82	16.04	**(43.09)**	38.77	**[5.42]**
Classic	ORB-SLAM2 (w/ LC) [31]	✓	*Required*	*Sparse*	54.816	**[9.464]**	**[6.03]**	508.34	**[14.76]**	**(1.02)**	1.57	**[4.04]**	11.16	**(2.19)**	**[38.85]**	**[8.39]**	**(6.63)**
Classic	LDSO [39]	✓	*Required*	*Sparse*	**[22.425]**	23.500	9.32	**(11.68)**	**(31.98)**	2.85	1.22	**(5.10)**	13.55	2.96	129.02	21.64	17.36
Learning Based	DROID-VO [33]	✗	*Required*	*Dense*	54.188	51.187	98.43	84.20	108.80	2.58	0.93	59.27	64.40	24.20	64.55	71.80	16.91
Learning Based	DPVO [40]	✗	*Required*	*Sparse*	53.609	57.701	113.21	12.69	123.40	2.09	**[0.68]**	58.96	54.78	19.26	115.90	75.10	13.63
Learning Based	DROID-SLAM [33]	-	*Required*	*Dense*	100.278	75.846	92.10	344.60	107.61	2.38	1.00	118.50	62.47	21.78	161.60	72.32	118.70
Learning Based	DPV-SLAM [38]	✓	*Required*	*Sparse*	53.034	57.187	112.80	**[11.50]**	123.53	2.50	0.81	57.80	54.86	18.77	110.49	76.66	13.65
Learning Based	DPV-SLAM++ [38]	✓	*Required*	*Sparse*	25.749	27.138	8.30	11.86	39.64	2.50	**(0.78)**	5.74	11.60	**[1.52]**	110.90	76.70	13.70
Learning Based	MASt3R-SLAM [15]	✓	*No Need*	*Dense*	/	/	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*
Learning Based	CUT3R [10]	✗	*No Need*	*Dense*	/	/	*OOM*	*OOM*	*OOM*	148.07	22.31	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
Learning Based	Fast3R [11]	✗	*No Need*	*Dense*	/	/	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
Learning Based	VGGT [12]	✗	*No Need*	*Dense*	/	/	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
	VGGT-Long (Chunk Size=30)	✓	*No Need*	*Dense*	44.713	39.564	9.06	96.20	99.95	19.76	11.36	10.20	10.04	4.00	139.79	50.53	40.94
	VGGT-Long (Chunk Size=60)	✓	*No Need*	*Dense*	26.358	**(19.298)**	**(8.06)**	96.96	34.16	6.83	4.16	9.15	**[4.68]**	**(2.68)**	63.15	32.24	27.87
	VGGT-Long (Chunk Size=90)	✓	*No Need*	*Dense*	**(22.718)**	20.938	11.97	40.51	49.85	5.41	2.86	9.88	**(6.07)**	3.47	66.27	32.27	21.33
	VGGT-Long (Chunk Size=120)	✓	*No Need*	*Dense*	25.597	22.814	16.13	53.43	51.98	4.37	2.15	12.69	11.33	3.603	70.29	34.55	21.05

TABLE II

CAMERA TRACKING RESULTS (ATE RMSE [M] ↓) ON THE WAYMO OPEN DATASET. COLOR DEFINITION: **[FIRST]**, **(SECOND)**.

Segment ID	Calib.	Avg.	163453191	183829460	315615587	346181117	371159869	405841035	460417311	520018670	610454533
Frame num.	-	*198*	*198*	*199*	*199*	*199*	*196*	*199*	*198*	*199*	*198*
Segment length	-	*172.533*	*159.963*	*42.301*	*165.149*	*351.213*	*272.661*	*85.743*	*265.906*	*134.552*	*62.739*
Segment speed	-	*0.871*	*0.808*	*0.213*	*0.830*	*1.765*	*1.391*	*0.431*	*1.343*	*0.676*	*0.317*
Traffic	-	-	*Low*	*High*	*Low*	*Low*	*Medium*	*Low*	*Medium*	*Low*	*High*
DROID SLAM [33]	*Required*	**(4.396)**	**(3.705)**	**[0.301]**	**[0.447]**	**(8.653)**	9.320	7.621	**(4.170)**	*TL*	**[0.264]**
MASt3R-SLAM [15]	*No Need*	5.560	4.500	**(0.556)**	1.833	12.544	**(8.601)**	**[1.412]**	5.428	**(7.910)**	1.195
CUT3R [10]	*No Need*	9.872	8.781	3.810	5.790	24.015	13.070	7.261	13.206	8.597	3.229
Fast3R [11]	*No Need*	/	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
VGGT [12]	*No Need*	/	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
VGGT-Long (Ours)	*No Need*	**[1.996]**	**[1.753]**	2.629	**(0.559)**	**[3.452]**	**[3.343]**	**(1.444)**	**[1.541]**	**[2.547]**	**(0.455)**

TABLE III

POINT MAP ESTIMATION RESULTS ON THE WAYMO OPEN DATASET. THE GT POINT CLOUD IS COLLECTED BY LIDAR. COLOR DEFINITION: **[FIRST]**.

Segment ID	Metric	Calib.	Avg.	163453191	183829460	315615587	346181117	371159869	405841035	460417311	520018670	610454533
DROID-SLAM [33]	Accuracy ↓	*Required*	*1.201*	**[0.781]**	1.136	2.247	2.393	**[1.090]**	**[0.539]**	**[0.740]**	*TL*	**[0.677]**
	Completeness ↓		*8.540*	4.610	10.245	5.540	8.669	8.592	11.144	5.320	*TL*	14.201
	Chamfer ↓		*4.870*	2.696	5.691	3.893	5.531	4.841	5.842	3.030	*TL*	7.439
MASt3R-SLAM [15]	Accuracy ↓	*No Need*	3.772	3.189	2.988	3.787	4.689	4.436	1.166	4.637	6.417	2.637
	Completeness ↓		3.177	**[1.715]**	**[3.284]**	2.047	**[2.981]**	**[2.679]**	**[2.895]**	2.002	4.429	6.560
	Chamfer ↓		3.474	2.452	3.136	2.917	3.835	3.558	2.031	3.319	5.423	4.599
CUT3R [10]	Accuracy ↓	*No Need*	3.884	3.580	1.144	2.418	3.712	3.679	4.346	2.012	12.320	1.744
	Completeness ↓		6.801	8.251	9.352	8.748	8.537	5.467	3.393	6.164	**[2.302]**	8.999
	Chamfer ↓		5.343	5.916	5.248	5.583	6.125	4.573	3.869	4.088	7.311	5.371
VGGT-Long (Ours)	Accuracy ↓	*No Need*	**[1.182]**	1.002	**[0.395]**	**[0.925]**	**[1.668]**	2.580	0.679	0.784	**[1.358]**	1.246
	Completeness ↓		**[2.860]**	2.762	3.417	**[1.738]**	3.261	2.791	3.216	**[1.840]**	4.694	**[2.022]**
	Chamfer ↓		**[2.021]**	**[1.882]**	**[1.906]**	**[1.331]**	**[2.465]**	**[2.685]**	**[1.948]**	**[1.312]**	**[3.026]**	**[1.634]**

TABLE IV

CAMERA TRACKING RESULTS (ATE RMSE [M] ↓) ON THE VIRTUAL KITTI DATASET. COLOR DEFINITION: **[FIRST]**, **(SECOND)**.

	Calib.	01 Avg.	02 Avg.	06 Avg.	18 Avg.	20 Avg.	All Avg.
DROID-SLAM [33]	*Required*	**(1.1366)**	**[0.0640]**	**[0.0377]**	**(2.2046)**	**[4.1572]**	**[1.5200]**
MASt3R-SLAM [15]	*No Need*	*TL*	*TL*	*TL*	*TL*	*TL*	*TL*
CUT3R [10]	*No Need*	48.3618	20.1191	0.7724	17.1494	103.6918	38.0189
Fast3R [11]	*No Need*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
VGGT [12]	*No Need*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*	*OOM*
VGGT-Long (Ours)	*No Need*	**[1.0490]**	**(0.7026)**	**(0.4377)**	**[1.3966]**	**(6.6830)**	**(2.0538)**

TABLE V

RUNTIME ANALYSIS OF THE COMPONENTS.

Chunk Size = 75		VPR Model	Chunk Process	Chunk Align	LM Opt. (C++)	LM Opt. (Python)
Seq.	Seq. Frames	*Time / Frame*	*Time / Chunk*	*Time / Iter*	*Time / Iter*	*Time / Iter*
00	*4542*	21.264 ms	2.811 s	0.284 s	1.249 ms	13.394 ms
05	*2761*	17.023 ms	2.614 s	0.273 s	0.862 ms	8.446 ms
06	*1101*	18.523 ms	2.728 s	0.278 s	0.436 ms	3.592 ms

eliminating RAM pressure from handling massive point cloud datasets.

D. KITTI, Waymo & Virtual KITTI

We begin with the KITTI odometry dataset, a classic benchmark for SLAM evaluation. In Table I, we test different chunk size settings, and the overlap was set to half of the chunk size. Due to the long sequence lengths and outdoor settings, KITTI presents significant challenges for monocular reconstruction. Table I reports the ATE across 11 sequences. In Table I, *[LC]* denotes *loop closure*. Since the Seq. 01 is a high-speed sequence, its movement pattern is significantly different from those of other sequences. *[Avg.*]* shows the mean ATE excluding Seq 01. *VGGT-Long* achieves relatively good tracking accuracy without the input of camera intrinsic matrix. *[OOM]* is short for *CUDA Out-Of-Memory* on a single RTX 4090. *[TL]* is short for *Tracking Lost*.

VGGT-Long outperforms the learning-based methods such as DROID-SLAM and DPVO. Unlike ORB-SLAM2 or LDSO, our method does not rely on calibrated camera

intrinsics and still maintains competitive accuracy. Notably, VGGT-Long runs successfully on all sequences, while foundation models like Fast3R, CUT3R, and VGGT fail due to memory overflow. This verifies that our chunk-and-align framework effectively extends VGGT to long sequences without sacrificing scalability. When running MASt3R-SLAM on KITTI Odometry, we observe that tracking stalled after about 100 frames. New frames cannot be selected as keyframes, causing the mapping part to stop updating even though the process continues running. This phenomenon indicates that MASt3R-SLAM undergoes tracking lost. Tracking lost is defined when the system fails to register new frames for a prolonged period, leading to no further updates in the map and trajectory. Although Figure 6 shows a better reconstruction correction and Table I demonstrates better tracking performance, we can still observe that there are still some parts of the central intersection that are not aligned. This is because there is no explicit loop closure constraint for the input image in this area.

To validate generalization, we test VGGT-Long on the Waymo Open Dataset, which features urban driving with high variability in scene appearance and traffic conditions. VGGT-Long achieves an average ATE of 1.996m across ten 200-frame segments. Compared to methods like CUT3R and MASt3R-SLAM, VGGT-Long yields significantly lower errors, as shown in Table II. On segments with strong viewpoint diversity, VGGT-Long consistently produces more accurate and complete 3D reconstructions.

We further assess robustness under synthetic domain shifts using the Virtual KITTI dataset, which offers multiple weather and lighting conditions. As Table IV shows, VGGT-Long maintains stable ATE across all tested conditions, including fog, rain, and sunset. Unlike DROID-SLAM, which occasionally fails to track, and CUT3R, which exhibits large drifts, our method remains robust without requiring retraining or domain adaptation.

In the experiments, the previous methods achieve lower tracking accuracies. The reasons for DROID-SLAM's poor performance on long outdoor sequences have been thoroughly discussed in [3], [38]. MASt3R-SLAM works in short sequences but often fails in longer ones, a limitation tied to its reliance on MASt3R [9] for feature matching. In autonomous driving scenarios with less scene variation than that in indoor scenarios, like on a long straight road, the system may not generate new keyframes for extended periods. When it finally does, the significant distance from the last keyframe causes feature matching to fail, leading to tracking loss. As for CUT3R, it employs continuous state tokens to encode the entire 3D scene, which is analogous to NeRF [46] in novel view synthesis. Consequently, CUT3R faces similar challenges as NeRF in large-scale outdoor scenes. That is, a compact representation struggles to capture the vast geometric details of expansive large-scale outdoor environments.

E. Loop Optimization Analysis & Ablation Study

Table V reports the runtime of VGGT-Long on several KITTI sequences. All measurements exclude disk I/O time, as it can vary depending on OS background programs, disk bandwidth, and memory throughput. On our computer, each chunk takes approximately 25ms to load and 95ms to write, which is negligible given that each KITTI sequence contains only a few dozen chunks. Therefore, disk latency does not significantly impact overall runtime. Our system achieves efficient chunk-wise processing (about 2.6–2.8s per chunk) and fast Sim(3) alignment (about 0.2s).

We evaluate the effectiveness and efficiency of our loop optimization module in Fig. 7. The proposed Sim(3)-based Levenberg-Marquardt solver converges within **3 iterations** on average and takes less than **15ms per step** (taking 0.4-1.3ms/iter in C++ or 3.5–13ms/iter in Python). These results demonstrate the practicality of our system in real-time or near real-time scenarios even on kilometer-scale sequences. Table VI presents an ablation study on KITTI sequences. "LC" indicates whether loop closure is enabled. "IRLS" denotes the use of iterative reweighted least squares in chunk alignment; disabling it results in a single-pass confidence-weighted alignment. "Weight" refers to whether confidence-based weighting is applied during alignment. To disable it, we normalize the VGGT confidence map to a uniform value across all points. Removing either the loop closure module or the IRLS weighting results in noticeable accuracy degradation. Specifically, removing loop closure increases the ATE to 58.69m on Seq. 00, while disabling IRLS leads to a 13% drop in performance. The best performance is achieved when all components, loop closure, IRLS and confidence-weighted alignment are enabled.

TABLE VI
ABLATION STUDY (ATE RMSE).

LC	IRLS	Weight	00	05
✗	✓	✓	58.69	36.01
✓	✗	✓	12.29	10.98
✓	✓	✗	11.28	10.13
✓	✓	✓	8.67	8.31

V. CONCLUSION

In this work, we presented VGGT-Long, a simple yet effective framework that extends monocular RGB-only 3D reconstruction to long, unbounded video sequences using foundation models. Our method overcomes the GPU memory limitations of existing 3D vision models without requiring camera calibration. Through extensive experiments on KITTI, Waymo, and Virtual KITTI, we demonstrated that VGGT-Long achieves accurate and scalable 3D reconstruction across diverse real-world and synthetic environments. In the future, we will continue researching ways to improve the accuracy and consistency of 3D foundation models for long outdoor sequences.

REFERENCES

[1] H. Zhan, C. S. Weerasekera, J.-W. Bian, R. Garg, and I. Reid, "Df-vo: What should be learnt for visual odometry?" *arXiv preprint arXiv:2103.00933*, 2021.

[2] N. Yang, R. Wang, J. Stuckler, and D. Cremers, "Deep virtual stereo odometry: Leveraging deep depth prediction for monocular direct sparse odometry," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 817–833.

[3] K. Deng, Y. Zhang, J. Yang, and J. Xie, "Gigaslam: Large-scale monocular slam with hierarchical gaussian splats," *arXiv preprint arXiv:2503.08071*, 2025.

[4] J. Zhang, S. Singh *et al.*, "Loam: Lidar odometry and mapping in real-time." in *Robotics: Science and systems*, vol. 2, no. 9. Berkeley, CA, 2014, pp. 1–9.

[5] J. Zhang and S. Singh, "Visual-lidar odometry and mapping: Low-drift, robust, and fast," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 2174–2181.

[6] I. Cvišić, I. Marković, and I. Petrović, "Soft2: Stereo visual odometry for road vehicles based on a point-to-epipolar-line metric," *IEEE Transactions on Robotics*, vol. 39, no. 1, pp. 273–288, 2022.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[8] S. Wang, V. Leroy, Y. Cabon, B. Chidlovskii, and J. Revaud, "Dust3r: Geometric 3d vision made easy," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 20 697–20 709.

[9] V. Leroy, Y. Cabon, and J. Revaud, "Grounding image matching in 3d with mast3r," in *European Conference on Computer Vision*. Springer, 2024, pp. 71–91.

[10] Q. Wang, Y. Zhang, A. Holynski, A. A. Efros, and A. Kanazawa, "Continuous 3d perception model with persistent state," *arXiv preprint arXiv:2501.12387*, 2025.

[11] J. Yang, A. Sax, K. J. Liang, M. Henaff, H. Tang, A. Cao, J. Chai, F. Meier, and M. Feiszli, "Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass," *arXiv preprint arXiv:2501.13928*, 2025.

[12] J. Wang, M. Chen, N. Karaev, A. Vedaldi, C. Rupprecht, and D. Novotny, "Vggt: Visual geometry grounded transformer," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 5294–5306.

[13] T. Dao, D. Y. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[14] T. Dao, "FlashAttention-2: Faster attention with better parallelism and work partitioning," in *International Conference on Learning Representations (ICLR)*, 2024.

[15] R. Murai, E. Dexheimer, and A. J. Davison, "Mast3r-slam: Real-time dense slam with 3d reconstruction priors," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 16 695–16 705.

[16] B. Triggs, P. F. McLauchlan, R. I. Hartley, and A. W. Fitzgibbon, "Bundle adjustment—a modern synthesis," in *International workshop on vision algorithms*. Springer, 1999, pp. 298–372.

[17] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski, "Building rome in a day," *Communications of the ACM*, vol. 54, no. 10, pp. 105–112, 2011.

[18] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik *et al.*, "Building rome on a cloudless day," in *European conference on computer vision*. Springer, 2010, pp. 368–381.

[19] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4104–4113.

[20] D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superpoint: Self-supervised interest point detection and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 224–236.

[21] M. Tyszkiewicz, P. Fua, and E. Trulls, "Disk: Learning local features with policy gradient," *Advances in Neural Information Processing Systems*, vol. 33, pp. 14 254–14 265, 2020.

[22] H. Chen, Z. Luo, J. Zhang, L. Zhou, X. Bai, Z. Hu, C.-L. Tai, and L. Quan, "Learning to match features with seeded graph matching network," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6301–6310.

[23] P. Lindenberger, P.-E. Sarlin, and M. Pollefeys, "Lightglue: Local feature matching at light speed," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 627–17 638.

[24] P. Lindenberger, P.-E. Sarlin, V. Larsson, and M. Pollefeys, "Pixel-Perfect Structure-from-Motion with Featuremetric Refinement," in *ICCV*, 2021.

[25] X. He, J. Sun, Y. Wang, S. Peng, Q. Huang, H. Bao, and X. Zhou, "Detector-free structure from motion," *CVPR*, 2024.

[26] C. Smith, D. Charatan, A. Tewari, and V. Sitzmann, "Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent," *arXiv preprint arXiv:2404.15259*, 2024.

[27] C. Tang and P. Tan, "Ba-net: Dense bundle adjustment network," *arXiv preprint arXiv:1806.04807*, 2018.

[28] Z. Teed and J. Deng, "Deepv2d: Video to depth with differentiable structure from motion," *arXiv preprint arXiv:1812.04605*, 2018.

[29] J. Wang, N. Karaev, C. Rupprecht, and D. Novotny, "Vggsfm: Visual geometry grounded deep structure from motion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 686–21 697.

[30] R. Mur-Artal, J. M. M. Montiel, and J. D. Tardos, "ORB-SLAM: A versatile and accurate monocular SLAM system," *IEEE Transactions on Robotics*, vol. 31, no. 5, pp. 1147–1163, oct 2015.

[31] R. Mur-Artal and J. D. Tardos, "ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-d cameras," *IEEE Transactions on Robotics*, vol. 33, no. 5, pp. 1255–1262, oct 2017.

[32] K. M. Jatavallabhula, S. Saryazdi, G. Iyer, and L. Paull, "gradslam: Automagically differentiable slam," *arXiv preprint arXiv:1910.10672*, 2019.

[33] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[34] D. Maggio, H. Lim, and L. Carlone, "Vggt-slam: Dense rgb slam optimized on the sl (4) manifold," *arXiv preprint arXiv:2505.12549*, 2025.

[35] S. Izquierdo and J. Civera, "Optimal transport aggregation for visual place recognition," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2024, pp. 17 658–17 668.

[36] M. Oquab, T. Darcet, T. Moutakanni, H. V. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, R. Howes, P.-Y. Huang, H. Xu, V. Sharma, S.-W. Li, W. Galuba, M. Rabbat, M. Assran, N. Ballas, G. Synnaeve, I. Misra, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2023.

[37] H. Strasdat, J. Montiel, and A. J. Davison, "Scale drift-aware large scale monocular slam," *Robotics: science and Systems VI*, vol. 2, no. 3, p. 7, 2010.

[38] L. Lipson, Z. Teed, and J. Deng, "Deep patch visual slam," in *European Conference on Computer Vision*. Springer, 2025, pp. 424–440.

[39] X. Gao, R. Wang, N. Demmel, and D. Cremers, "Ldso: Direct sparse odometry with loop closure," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 2198–2204.

[40] Z. Teed, L. Lipson, and J. Deng, "Deep patch visual odometry," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2012, pp. 3354–3361.

[42] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[43] A. Gaidon, Q. Wang, Y. Cabon, and E. Vig, "Virtual worlds as proxy for multi-object tracking analysis," in *CVPR*, 2016.

[44] J. Sturm, N. Engelhard, F. Endres, W. Burgard, and D. Cremers, "A benchmark for the evaluation of rgb-d slam systems," in *2012 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2012, pp. 573–580.

[45] S. Rusinkiewicz and M. Levoy, "Efficient variants of the icp algorithm," in *Proceedings third international conference on 3-D digital imaging and modeling*. IEEE, 2001, pp. 145–152.

[46] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.