

MegaVGGT: Extending Vision Geometry Models to Large-Scale Complex Scenes with Vision Place Recognition

Ming Rui Yu¹ Danki Ho Chi-Han Peng¹ ¹National Yang Ming Chiao Tung University*

June 8, 2026

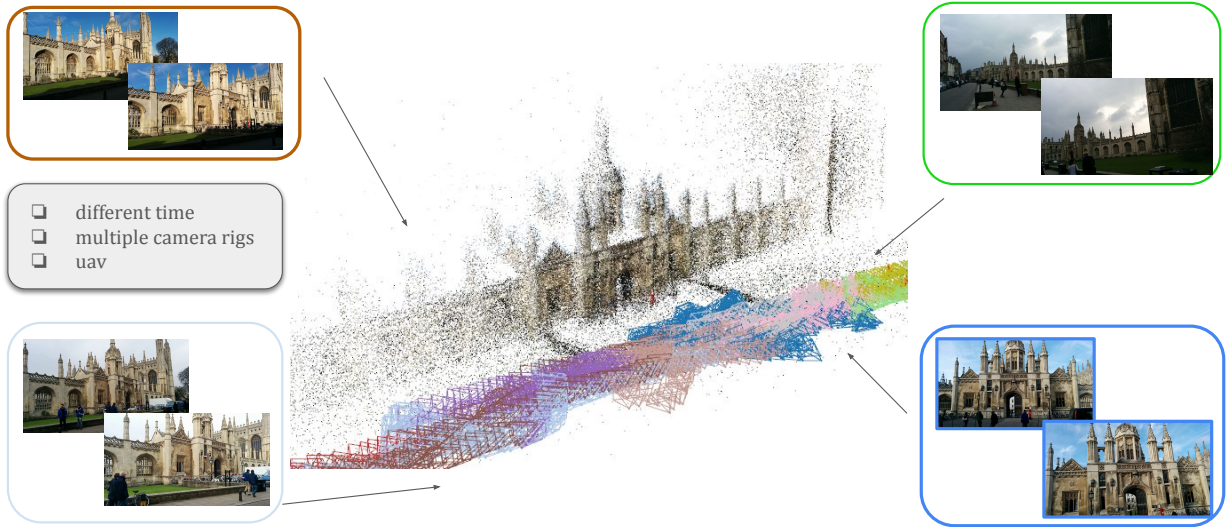


Figure 1: Our method reconstructs large-scale 3D scenes from 1000+ image collection from different camera rigs and capture sessions, within a 24 GB memory budget.

Abstract

Visual Geometry Transformer models offer fast feed-forward 3D reconstruction without scene-specific training, yet their all-pairs attention mechanism imposes a hard memory boundary that prevents direct application to large-scale scenes with thousands of images. We introduce **MegaVGGT**, a framework that uses visual place recognition (VPR) model to extends foundation models to large-scale, unordered, and heterogeneous image collections, including mixed drone, ground-level, and multi-camera captures beyond sequential assumptions. Experiments on standard indoor and outdoor benchmarks demonstrate that our method not only scales to scenes that exceed the memory budget of single-pass foundation models, but also achieves superior pose accuracy compared to existing scalable alternatives, with particular gains on unordered and structurally complex scenes.

1 Introduction

Reconstructing 3D scenes from a large set of images with diverse camera poses is a key problem in autonomous navigation, augmented reality, and large-scale mapping. Classical Structure-from-Motion pipelines, while robust, require solving expensive feature matching and bundle adjustment problems that scale poorly to large unordered collections. Recent transformer-based models such as DUST3R [34], VGGT [33], and π^3 [36] offer a compelling alternative, jointly inferring camera parameters and dense point clouds from raw images in a single forward pass without scene-specific training.

Despite this progress, these models share a critical scalability bottleneck: the self-attention mechanism grows quadratically in both computation and memory with the number of input images, making it infeasible to process large collections in a single pass. Existing attempts to address this either apply a sliding window over sequen-

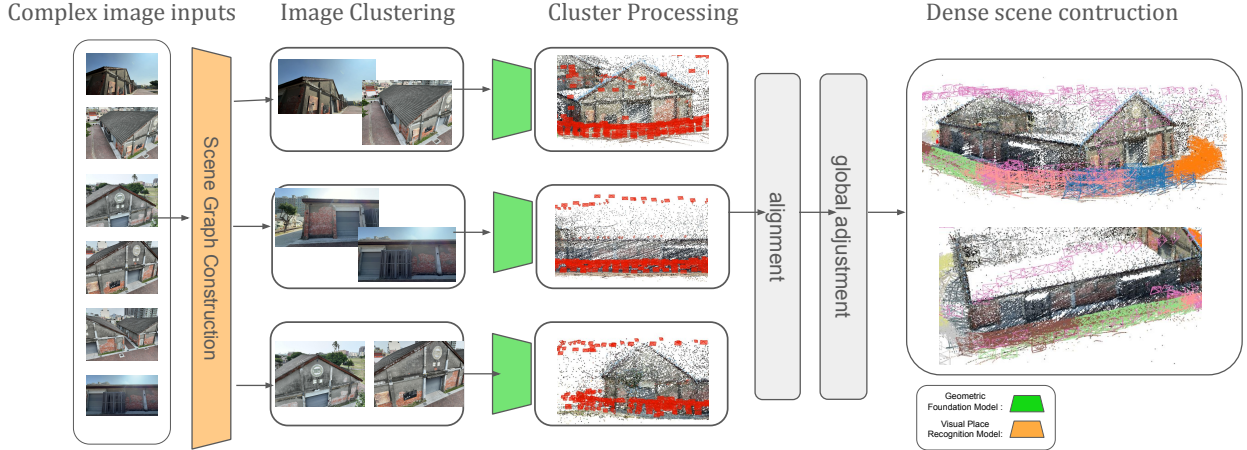


Figure 2: Overview of our pipeline. Given an unordered image collection from heterogeneous sources, we extract visual model embeddings and construct a scene graph. The graph is reduced to a Minimum Spanning Tree and recursively bisected into spatially coherent clusters. Each cluster is independently processed by a geometric foundation model to recover local poses and point maps, which are then aligned via overlap-based Sim(3) estimation and refined through global bundle adjustment to produce a dense, globally consistent 3D reconstruction.

tial inputs, partitioning the image stream into overlapping chunks to extend coverage—or reduce model capacity in ways that degrade geometric accuracy under wide viewpoint variation.

High-quality 3D scene reconstruction demands comprehensive viewpoint coverage, which in practice often requires images collected across multiple sessions. A single capture pass rarely suffices: operators return on different days to fill missing viewpoints, lighting and weather conditions vary between sessions, and distances to the subject differ depending on the capture device or vantage point. The resulting collection has no consistent temporal ordering and exhibits large appearance and viewpoint variation across images. We refer to this practically important but underexplored setting as **complex scene reconstruction**.

We present **MegaVGGT**, a training-free divide-and-conquer framework that enables feed-forward reconstruction models to operate on large, unordered, and heterogeneous image collections far beyond their native memory limits, without sacrificing global geometric accuracy. By modeling visual connectivity through a VPR-based scene graph rather than imposing temporal order, our method naturally handles complex scene reconstruction where sequential assumptions break down.

Our key contributions are:

- A VPR-based scene graph construction that captures visual and spatial relationships across heterogeneous unordered collections without sequential assumptions.
- A recursive MST-based graph clustering strategy that

partitions large scenes into memory-feasible sub-scenes for independent foundation model inference.

- A dataset comprising unordered images from heterogeneous capture sources—including drone aerial photography, ground-level GoPro captures, and multi-camera rigs. We define this setting as *complex scene reconstruction* and demonstrate that our method consistently outperforms existing approaches under these conditions while remaining competitive on standard benchmarks.

2 Related Work

2.1 Structure-from-Motion.

Classical SfM [28] detects keypoints, matches features, and refines poses via bundle adjustment, but scales poorly to large unordered collections. Recent methods improve scalability via global scene graphs [25], distributed motion averaging [44, 7], first-order pose optimization [22], and retrieval-based candidate pairing [27, 43], yet still depend on handcrafted feature matching and expensive geometric verification. Our work adopts a similar scene graph abstraction but replaces these steps with VPR-based graph construction and a foundation model for reconstruction.

2.2 Visual Place Recognition.

VPR retrieves images of the same place from a large database, serving as a scalable proxy for spatial proximity without explicit 3D reasoning. Methods have pro-

gressed from VLAD-based aggregation [2, 26, 11] to classification-inspired [4] and optimal-transport-based descriptors [16], with recent work leveraging foundation model features for improved cross-domain generalization [17, 5, 1]. MegaLoc [3] combines a DINOv2 backbone with SALAD aggregation to achieve state-of-the-art retrieval across diverse scenes; we adopt its descriptors to construct our VPR-based scene graph as a lightweight alternative to classical feature matching and geometric verification.

2.3 Foundation Models for 3D Vision.

A recent trend replaces multi-stage SfM with end-to-end Transformer-based models. DUST3R [34] pioneers joint estimation of camera parameters and dense point maps in a single forward pass; subsequent work extends this to feature matching [21], dynamic scenes [41], and incremental processing [32, 35, 38], and sparse uncalibrated views [42]. VGGT [33] further advances the paradigm by processing entire image sets through alternating attention blocks, with follow-on work improving its robustness and efficiency [31, 14, 18]. π^3 [36] eliminates the fixed reference-view assumption via a fully permutation-equivariant architecture. Despite these advances, the quadratic memory cost of self-attention limits all these models to short sequences, making direct application to large unordered collections infeasible.

2.4 Scalable Reconstruction with Foundation Models.

While foundation models such as VGGT excel at reconstructing scenes from a few hundred images, their full attention mechanism creates a significant scalability bottleneck for large collections.

Sequential chunking methods process ordered frames in segments—VGGT-Long [9] aligns overlapping chunks while InfiniteVGGT [39] extends this to endless streams—but both assume temporally ordered input unsuited for unordered collections. FastVGGT and LiteVGGT [29, 30] reduce cost within a single pass by merging redundant tokens, while CUT3R and TTT3R [35, 6] avoid full self-attention via per-image depth prediction at the cost of global geometric consistency. MERG3R [8] takes a divide-and-conquer approach but sorts images into a pseudo-video before clustering, which fails for complex scenes.

3 Method

We present a scalable pipeline for large-scale 3D reconstruction from unordered, heterogeneous image collec-

tions. As illustrated in Figure 2, the pipeline decomposes the scene into memory-feasible sub-scenes, reconstructs each independently with a foundation model, and merges the results into a globally consistent map.

3.1 VPR-Based Scene Graph Construction

We extract a global descriptor for each image using MegaLoc [3], which combines a DINOv2 backbone with SALAD aggregation [16] to enable reliable retrieval across large viewpoint and appearance changes—unlike raw DINOv2 features, which largely restrict connectivity to adjacent frames.

A scene graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is constructed over the image collection, where each node $v_i \in \mathcal{V}$ represents an image and each edge $(i, j) \in \mathcal{E}$ is weighted by cosine similarity between VPR descriptors. Since all image pairs receive a finite score, we reduce $\mathcal{G}$ to its **Maximum Spanning Tree (MST)**, retaining exactly $|\mathcal{V}| - 1$ edges through the most visually similar pairs and eliminating spurious connections.

To further improve edge quality, we replace raw cosine similarity with the GNN-based scoring model from [37], which propagates contextual information over the graph to predict a globally consistent matchability score. Each edge is first initialized as:

$$e_{ij}^0 = \text{ReLU}(f_1[d_i, d_j, \langle d_i, d_j \rangle]),$$

and refined through two iterations of edge-node message passing. The final edge rank is predicted as:

$$\hat{r}_{ij} = f_{\text{MLP}}(e_{ij}^2),$$

which we use as the edge weight s_{ij} for scene graph construction. Unlike the original work, which uses MST-based graphs for SfM pose initialization, we repurpose this scoring to guide scene decomposition for foundation model inference.

3.2 Recursive Graph Clustering

Large scenes yield graphs with thousands of nodes that exceed the memory budget of any single foundation model inference. We recursively partition $\mathcal{G}$ into sub-graphs by cutting the MST. At each recursion, if the number of images in a sub-graph exceeds the memory upper bound N_{max} , we bisect it into two equally-sized halves by selecting the MST edge e^* that minimises the size imbalance between the two resulting sub-graphs:

$$e^* = \arg \max_{e \in \mathcal{T}} \left(\sum_{(i,j) \in \mathcal{E}_{\text{left}} \cup \mathcal{E}_{\text{right}}} s_{ij} - \lambda ||\mathcal{V}_{\text{left}}| - |\mathcal{V}_{\text{right}}|| \right) \quad (1)$$

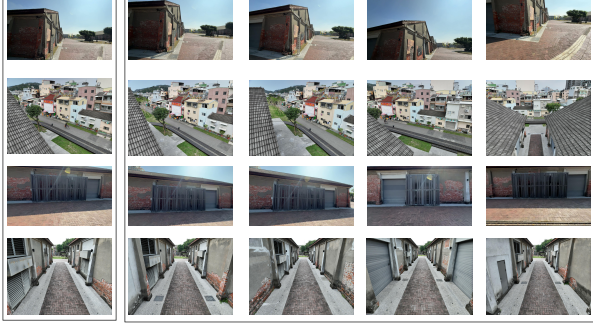


Figure 3: Qualitative results of visual place recognition. For each query image (left), the top-4 retrieved images (right) are shown.

where s_{ij} denotes the VPR similarity score between images i and j , $\mathcal{E}_{\text{left}}$ and $\mathcal{E}_{\text{right}}$ are the edge sets of the two resulting sub-graphs, and λ controls the trade-off between internal cohesion and balanced partitioning. Both partitions are recursively bisected until all sub-graphs satisfy $|\mathcal{V}_k| \leq N_{\text{max}}$, yielding K spatially coherent clusters $\{\mathcal{G}_k\}_{k=1}^K$.

To ensure adjacent clusters share sufficient context for subsequent point map alignment, we augment each cluster with a set of *overlap images* derived from the cut edges. Let $\mathcal{E}_{\text{cut}} = \mathcal{E} \setminus \bigcup_k \mathcal{E}_k$ denote the set of edges removed by the graph cut. For each cut edge $(i, j) \in \mathcal{E}_{\text{cut}}$, we define the h -hop neighbourhood of a node v within sub-graph $\mathcal{G}_k$ as:

$$\mathcal{N}_k^h(v) = \{u \in \mathcal{V}_k \mid d_k(v, u) \leq h\}, \quad (2)$$

where $d_k(v, u)$ is the shortest-path distance between v and u in $\mathcal{G}_k$. The augmented node set of each cluster is then:

$$\tilde{\mathcal{V}}_k = \mathcal{V}_k \cup \bigcup_{\substack{(i,j) \in \mathcal{E}_{\text{cut}} \\ j \in \mathcal{V}_k}} \mathcal{N}_k^h(j), \quad (3)$$

where the union collects the h -hop neighbourhoods of all cut-edge endpoints that belong to $\mathcal{V}_k$. The depth h controls the trade-off between alignment quality and redundancy: a larger h provides a broader shared context at the cost of increased overlap between clusters, while a smaller h keeps clusters compact. In practice, we find $h=2$ to strike a good balance, adding enough shared structure for robust point map alignment without causing excessive overlap.

3.3 Foundation Model Inference

Each cluster $\mathcal{G}_k$ is processed independently by a foundation model $\mathcal{F}$, which produces camera parameters and a dense point map in a single feed-forward pass:

$$\{\hat{\mathbf{P}}_i, \hat{\mathbf{K}}_i, \hat{\mathbf{C}}_i\}_{i \in \mathcal{G}_k}, \mathcal{M}_k = \mathcal{F}(\{I_i\}_{i \in \mathcal{G}_k}), \quad (4)$$

where $\hat{\mathbf{P}}_i \in SE(3)$, $\hat{\mathbf{K}}_i \in \mathbb{R}^{3 \times 3}$, and $\mathcal{M}_k$ is the local point map reconstructed in its own coordinate frame. Decomposing the full collection into K clusters of size T reduces the self-attention cost from $O(N^2)$ to $O(KT^2)$, lowering peak GPU memory and enabling parallel inference across multiple GPUs.

3.4 Overlap-Based Point Map Alignment

To merge sub-scene reconstructions into a common coordinate frame, we exploit the overlap images shared between adjacent clusters. Each overlap image appears in two clusters $\mathcal{G}_k$ and $\mathcal{G}_{k'}$, yielding two point maps $\mathcal{M}_k^{\text{ov}}$ and $\mathcal{M}_{k'}^{\text{ov}}$ defined in different local coordinate frames, along with their associated confidence maps $\mathbf{C}_k^{\text{ov}}$ and $\mathbf{C}_{k'}^{\text{ov}}$.

Correspondence Construction. For each overlap image shared by clusters $\mathcal{G}_k$ and $\mathcal{G}_{k'}$, we obtain 3D point correspondences $\{(\mathbf{p}_k^i, \mathbf{p}_{k'}^i)\}$ by pairing pixels at the same image coordinate across the two point maps. A correspondence is retained only if both of its confidence scores exceed the 50th percentile of their respective cluster:

$$c_k^i > \tau_k \text{ and } c_{k'}^i > \tau_{k'}, \quad \tau_k = \text{P}_{50}(\mathbf{C}_k^{\text{ov}}), \quad (5)$$

Robust Sim(3) Estimation. For each cluster overlap edge $(\mathcal{G}_k, \mathcal{G}_{k'})$, we estimate a similarity transformation $\mathbf{S}_{k \rightarrow k'} \in \text{Sim}(3)$ that aligns $\mathcal{M}_{k'}^{\text{ov}}$ to $\mathcal{M}_k^{\text{ov}}$ by minimising a robust weighted cost:

$$\mathbf{S}^* = \arg \min_{\mathbf{S} \in \text{Sim}(3)} \sum_i w_i \|\mathbf{p}_k^i - \mathbf{S} \mathbf{p}_{k'}^i\|_2^2, \quad (6)$$

solved via Iteratively Reweighted Least Squares (IRLS), where at each iteration t the per-correspondence weights are updated as:

$$w_i^{(t)} = c_i \cdot \frac{\rho'(r_i^{(t)})}{r_i^{(t)}}, \quad (7)$$

where c_i is the confidence score of correspondence i , $r_i^{(t)} = \|\mathbf{p}_k^i - \mathbf{S}^{(t)} \mathbf{p}_{k'}^i\|_2$ is the residual from the previous iteration, and $\rho(\cdot)$ is the Huber loss. Each weighted least-squares subproblem is solved in closed form via the weighted Umeyama algorithm. This formulation naturally down-weights low-confidence regions and concentrates alignment on stable, high-confidence structures such as buildings and static surfaces.

Global Merging. The pairwise $\text{Sim}(3)$ transformations estimated over all cluster overlap edges form a spanning tree over the cluster graph. We designate one cluster as the global reference frame and compose transformations

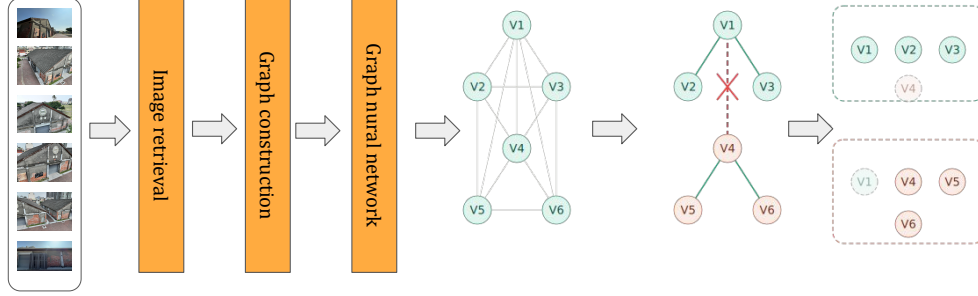


Figure 4: Detailed illustration of the scene graph construction and clustering pipeline.

along tree paths to obtain the absolute similarity $\mathbf{S}_k \in \text{Sim}(3)$ for each cluster k :

$$\mathbf{S}_k = \mathbf{S}_{k_{m-1} \rightarrow k} \circ \dots \circ \mathbf{S}_{k_1 \rightarrow k_2} \circ \mathbf{S}_{\text{ref} \rightarrow k_1}, \quad (8)$$

where $\text{ref} \rightarrow k_1 \rightarrow \dots \rightarrow k_{m-1} \rightarrow k$ is the unique path from the reference cluster to cluster k in the spanning tree. Each local point map $\mathcal{M}_k$ is then transformed as $\tilde{\mathcal{M}}_k = \mathbf{S}_k(\mathcal{M}_k)$, and all transformed maps are merged into a single global point cloud ready for bundle adjustment.

3.5 Feature Tracking

Accurate pixel correspondences are critical for global bundle adjustment. To avoid exhaustive all-pairs matching of quadratic complexity, we restrict matching to image pairs connected in the VPR scene graph, reducing the number of candidate pairs to $O(kN)$ where k is the mean node degree of the graph.

For each retained pair (i, j) , we extract keypoints with SuperPoint [10] and compute matches with LightGlue [23]. Pairs with fewer than 50 inlier matches are discarded as unreliable edges.

The remaining correspondences $\{(\mathbf{x}_m^i, \mathbf{x}_n^j)\}$ are aggregated into multi-view tracks using a Union-Find structure, which efficiently merges observations of the same scene point across multiple views. Each track $\mathcal{T}_l$ is associated with a confidence weight derived from the LightGlue [23] match confidence scores, which are used to down-weight uncertain correspondences during bundle adjustment.

3.6 Global Bundle Adjustment

After aligning all sub-scenes into a common coordinate frame, we perform global bundle adjustment to correct accumulated drift and enforce consistency across cluster boundaries. Following [8], we jointly optimize camera intrinsics, extrinsics, and 3D point positions by minimizing the confidence-weighted reprojection error over all multi-

view tracks $\mathcal{T}$:

$$\min_{\mathbf{R}, \mathbf{t}, \mathbf{K}, \mathbf{P}} \sum_{(\mathcal{T}_l, \mathbf{x}_l, C_l) \in \mathcal{T}} C_l \sum_{\mathbf{y}^{l,i} \in \mathcal{T}_l} \|\mathbf{y}^{l,i} - \pi_i(\mathbf{x}_l)\|_2, \quad (9)$$

where $\pi_i : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ denotes the projection onto the image plane of camera i , and C_l is the confidence weight derived from Model Point Head.

4 Experiments

4.1 Implementation Details

Environment. All experiments are conducted on a single NVIDIA RTX 6000 Pro GPU. To ensure reproducibility and fair comparison with hardware-constrained settings, we deliberately limit the memory budget to 24 GB, corresponding to the capacity of a consumer-grade RTX 3090/4090. The memory upper bound $N_{\max}$ is set accordingly, such that each cluster processed by VGGT fits within this 24 GB constraint. This design choice also demonstrates that our framework is accessible without high-end hardware.

Following [24], all inference runs in bfloat16 and activations are retained only at layers $\{4, 11, 17, 23\}$, reducing peak GPU memory with negligible impact on accuracy.

4.2 Metrics

We evaluate camera pose estimation using the EVO toolkit [13]. We report Absolute Pose Error (APE, m), Relative Rotation Error (RRE, $^\circ$), and Relative Pose Error (RPE, m), with predicted trajectories aligned to ground truth prior to error computation. We additionally report pairwise metrics: Relative Rotation Accuracy (RRA@30), Relative Translation Accuracy (RTA@30), and AUC@30, measuring the percentage of image pairs whose relative rotation and translation errors fall below 30° .

4.3 Datasets

4.3.1 Ours Dataset.

We collect an in-house dataset comprising images captured across multiple real-world outdoor scenes, including street intersections, warehouse facilities, open lawns, and outdoor installations. Images are gathered from heterogeneous and unordered sources such as drone aerial photography, GoPro ground-level captures, and multi-camera rigs, with no consistent temporal structure across devices. Ground-truth camera poses are obtained via COLMAP reconstruction.

Table 1 reports results on our dataset. Our method consistently outperforms all baselines across every scene. On Intersection (2 cams), our method achieves the best performance across all metrics where MERG3R fails entirely. On Office, we outperform even VGGT-Long, suggesting VPR-based connectivity is more robust in low-texture environments. On Warehouse and Lawn, where VGGT and FastVGGT run out of memory, our decomposition keeps GPU usage within 24 GB while substantially outperforming MERG3R (RRE = 34.1° on Lawn due to unreliable pseudo-video ordering over homogeneous terrain).

4.3.2 Tanks and Temples.

Tanks and Temples [20] provides unordered image collections of indoor and outdoor scenes with ground-truth point clouds.

Our method achieves the lowest RRE and RPE on most T&T scenes; the higher APE on Church (0.605) reflects limited pairwise overlap in this structurally complex scene.

4.3.3 7-Scenes.

7-Scenes [12] is an indoor RGB-D benchmark comprising seven small-scale scenes captured with a handheld depth camera, with ground-truth camera poses provided for each sequence. To simulate a large unordered multi-session setting while keeping the image count manageable, we merge all sequences into a single unordered pool and uniformly subsample the frames such that each scene contains approximately 1000 images.

4.3.4 Cambridge Landmarks.

Cambridge Landmarks [19] is an outdoor urban benchmark with multiple video sequences per scene. We merge all sequences into a single unordered image pool to simulate real-world multi-source capture conditions.

Table 3: Results on Cambridge Landmarks dataset. **Bold** denotes best result.

Scene	Method	APE↓	RRE↓	RPE↓
King’s College (1562)	MERG3R	0.094	1.142	0.103
	Ours	0.091	0.388	0.038
Old Hospital (1084)	MERG3R	0.766	1.359	0.854
	Ours	0.785	0.611	0.561
Shop (334)	MERG3R	0.400	0.829	0.511
	Ours	0.288	0.348	0.280
St Mary’s Church (2108)	MERG3R	8.816	12.597	8.333
	Ours	0.575	0.594	0.260
Great Court (2291)	MERG3R	33.297	48.416	8.435
	Ours	1.978	0.733	0.858

Our method consistently outperforms MERG3R on Cambridge Landmarks, with particularly large gains on St Mary’s Church and Great Court (APE reduced by >93%).

4.4 Overall Benchmark Comparison

Table 4 summarizes results across all benchmarks. Our method achieves the best or highly competitive performance in every setting. The largest gains appear on our dataset and Cambridge Landmarks, where heterogeneous unordered captures expose the limitations of pseudo-sequential clustering in MERG3R. On Tanks and Temples, our method achieves the best overall accuracy despite relatively coherent image ordering. On 7-Scenes, both methods perform comparably; our method obtains lower RRE and RPE while MERG3R achieves marginally better APE.

4.5 Computation Time and Memory Usage

Figure 6 reports runtime and peak GPU memory as a function of input size. VGGT and FastVGGT exhibit quadratic memory growth, exceeding the 24 GB budget at ~500 and ~900 images respectively, while our method and MERG3R maintain near-constant memory by bounding each cluster independently. At 1000 images, our method runs in ~120 s, comparable to MERG3R (117 s), while outperforming all baselines in reconstruction accuracy.

4.6 Ablation Study

Scene Graph Construction. We initially used COLMAP retrieval and geometric verification for scene graph construction, but its overhead scales poorly to large collections. Table 5 shows that MegaLoc descriptors achieve comparable accuracy at a fraction of the cost, motivating our VPR-based design.

Table 1: Results on Ours dataset.

Method	Intersection (1 cam)			Intersection (2 cams)			Office			Warehouse (loop)			Outdoor Art			Lawn (aerial)		
	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓
VGGT	0.529	2.336	0.393	0.513	2.307	0.427	0.381	12.397	0.383	OOM			0.659	2.666	0.281	OOM		
VGGT-Long	0.811	1.166	0.280	N/A			0.328	3.501	0.294	—			—			—		
FastVGGT	0.308	1.817	0.262	0.279	1.496	0.228	0.317	7.380	0.353	0.172	1.240	0.103	1.160	2.480	0.523	0.654	2.577	0.245
LiteVGGT	0.177	1.018	0.152	0.227	1.173	0.204	0.389	11.480	0.419	0.224	1.476	0.100	0.842	2.387	0.375	0.677	2.605	0.224
MERG3R	0.234	1.529	0.209	0.408	4.671	0.520	0.351	11.517	0.532	0.347	4.567	0.329	0.874	6.747	1.204	1.948	34.091	1.250
Ours	0.038	1.216	0.176	0.196	0.956	0.134	0.195	1.535	0.260	0.111	0.691	0.050	0.148	1.816	0.074	0.756	1.264	0.099

Table 2: Results on Tanks and Temples dataset. **Bold** denotes best result.

Method	Meeting Room			Museum			Barn			Caterpillar			Church		
	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓	APE↓	RRE↓	RPE↓
FastVGGT	0.070	0.456	0.056	0.148	0.932	0.123	0.059	0.580	0.056	0.059	0.470	0.051	2.909	3.759	0.189
VGGT	0.094	0.342	0.040	0.195	0.583	0.106	0.074	0.435	0.037	0.065	0.318	0.031	0.963	6.262	0.401
LiteVGGT	0.086	0.447	0.053	1.213	12.306	0.910	0.068	0.538	0.044	0.058	0.397	0.040	0.239	1.835	0.153
MERG3R	0.042	0.196	0.017	0.104	0.427	0.057	0.033	0.648	0.030	0.069	0.828	0.094	2.939	8.003	0.316
Ours	0.074	0.172	0.011	0.107	0.361	0.050	0.061	0.349	0.020	0.036	0.323	0.018	0.605	1.913	0.100

Table 4: Comparison on all benchmarks. **Bold** denotes best result.

Dataset	Method	APE↓	RRE↓	RPE↓	RRA@30↑	RTA@30↑	AUC@30↑
Ours	LiteVGGT	0.423	3.357	0.246	0.9459	0.9198	0.7196
	FastVGGT	0.317	7.380	0.353	0.7497	0.7909	0.5760
	MERG3R	0.507	30.000	0.548	0.7900	0.7335	0.5737
	Ours	0.236	1.219	0.150	1.0000	0.9803	0.8361
T&T	VGGT	0.278	1.588	0.123	0.9618	0.9563	0.8641
	FastVGGT	0.649	1.239	0.095	0.9176	0.9239	0.8472
	LiteVGGT	0.333	3.105	0.240	0.9447	0.9577	0.8109
	MERG3R	0.635	2.020	0.103	0.9176	0.9151	0.8487
	Ours	0.177	0.624	0.040	0.9734	0.9685	0.9030
Cambridge	MERG3R	4.455	6.870	4.218	0.9695	0.9560	0.8037
	Ours	0.333	0.491	0.149	0.9970	0.9952	0.9290
7-Scenes	LiteVGGT	0.837	1.215	0.627	0.9961	0.9651	0.7402
	MERG3R	0.635	1.253	0.654	0.9993	0.9859	0.8495
	Ours	0.618	0.399	0.510	0.9993	0.9896	0.8499

MegaLoc achieves the best APE and most consistent performance across all metrics. Although Clique [15] attains slightly lower RRE and RPE, MegaLoc avoids the failure modes of other methods—DINOv2 degrades on rotation errors and SALAD yields notably higher RRE—making it the more robust choice across diverse scene conditions (Figure 7).

GNN Edge Scoring. We evaluate GNN-based edge scoring on the Church scene (T&T), which features complex topology and limited inter-cluster overlap. As shown in Table 6, removing GNN scoring degrades APE from 0.605 m to 3.038 m and drops camera registration from 86.70% to 53.25%, confirming that reliable edge quality is critical for coherent scene decomposition.

Table 5: Comparison of scene graph construction methods on pose accuracy and preprocessing time.

Method	Time (min)↓	APE (m)↓	RRE (°)↓	RPE (m)↓
COLMAP	16.3	0.161	0.722	0.045
DINOv2 [40]	~1.0	0.315	1.202	0.113
SALAD [16]	~1.0	0.182	0.950	0.041
Clique [15]	~1.0	0.147	0.614	0.040
MegaLoc [3]	~1.0	0.111	0.691	0.050

Table 6: Ablation of GNN edge scoring on Church (T&T). **Bold** denotes best result.

Method	APE↓	RRE↓	RPE↓	APE↑	RRE↑	RPE↑
VGGT	0.963	6.262	0.401	0.809	0.796	0.657
MERG3R	2.939	8.003	0.316	0.588	0.582	0.456
Ours w/o GNN	3.038	3.541	0.182	0.533	0.581	0.474
Ours	0.605	1.913	0.100	0.867	0.847	0.741

5 Conclusion

We presented MegaVGGT, a training-free framework that extends feed-forward 3D reconstruction models to large-scale, unordered, and heterogeneous image collections. By replacing sequential ordering assumptions with a VPR-based scene graph, our method naturally handles complex scene reconstruction scenarios where images originate from mixed drone, ground-level, and multi-camera captures across different times and conditions. Recursive MST-based clustering partitions the scene into memory-feasible sub-scenes, and overlap-based Sim(3) alignment followed by global bundle adjustment produces a globally consistent reconstruction.

Experiments across indoor and outdoor benchmarks

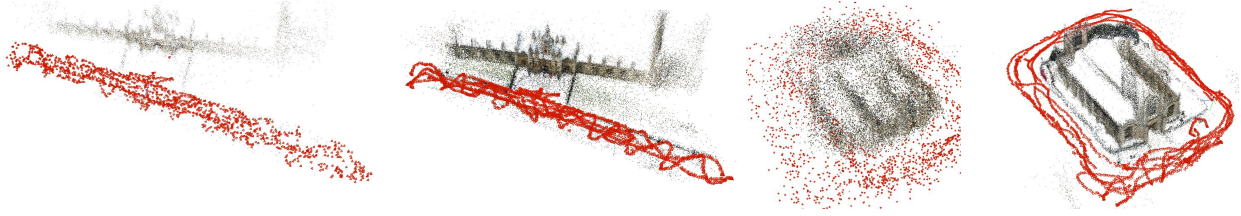


Figure 5: Qualitative comparison of camera trajectory estimation (red) and 3D point cloud reconstruction (grey) on two scenes. **Left pair:** MERG3R (1st) vs. ours (2nd, *KingsCollege*). **Right pair:** MERG3R (3rd) vs. ours (4th, *StMaryChurch*).

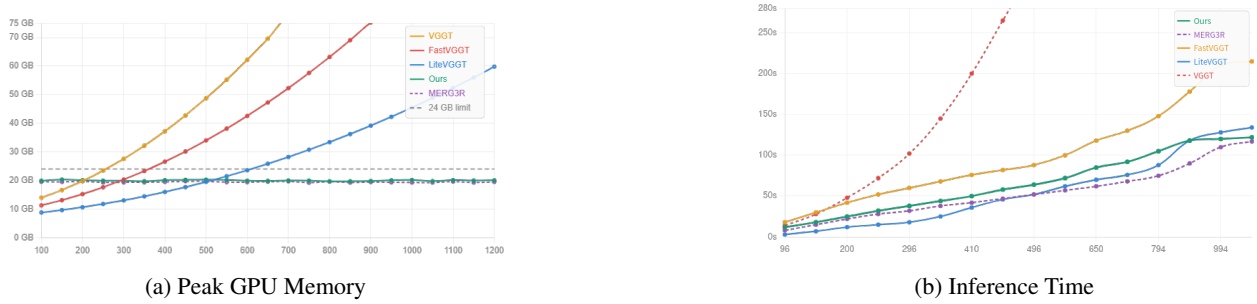


Figure 6: Scalability comparison across methods. Our cluster-based decomposition maintains bounded memory usage and competitive runtime regardless of collection size, while single-pass methods either run out of memory or exhibit rapidly growing inference time as the number of images increases.

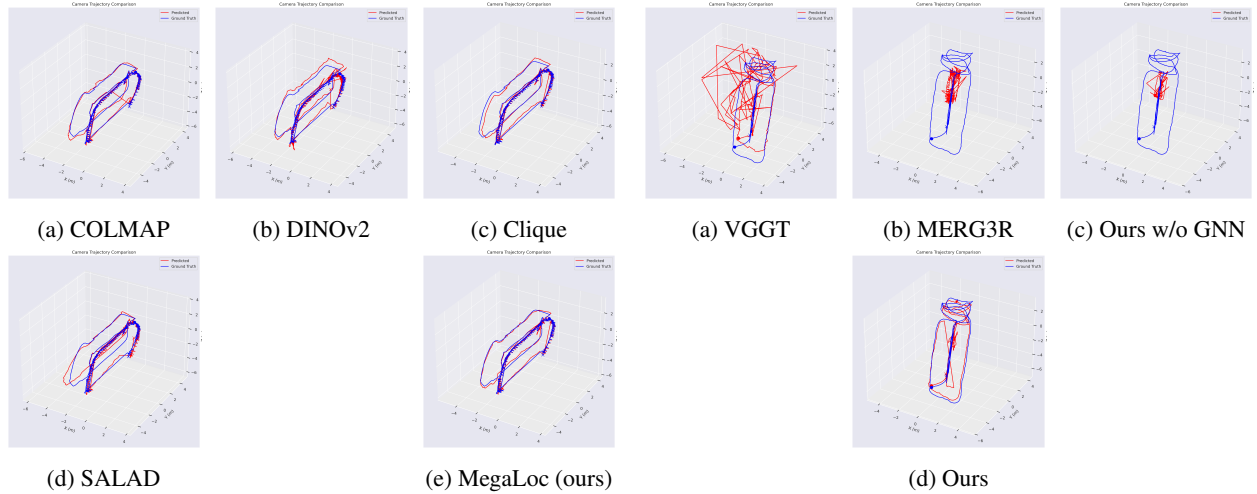


Figure 7: Trajectory comparison across scene graph construction methods.

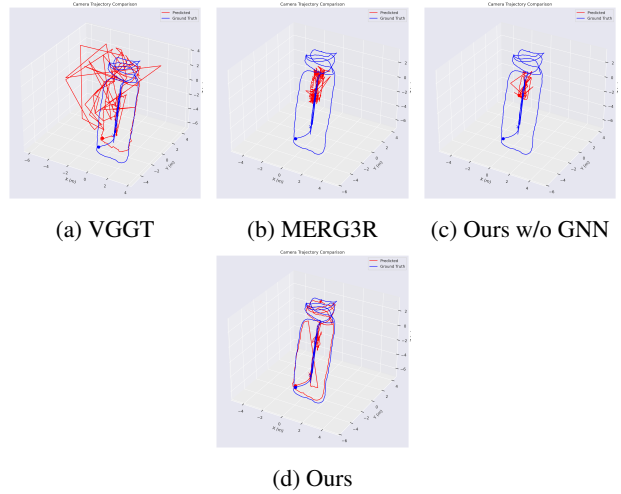


Figure 8: Trajectory comparison on the Church scene.

References

demonstrate that MegaVGGT consistently outperforms existing scalable methods in pose accuracy, particularly on unordered multi-source collections where sequential methods fail outright. On standard benchmarks, our method remains competitive with single-pass foundation models while lifting the memory constraint entirely.

- [1] Amar Ali-bey, Brahim Chaib-draa, and Philippe Giguère. Boq: A place is worth a bag of learnable queries. In *CVPR*, 2024.
- [2] Relja Arandjelović, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn ar-

- chitecture for weakly supervised place recognition, 2016.
- [3] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *CVPRW*, pages 2886–2892, June 2025.
 - [4] Gabriele Berton, Carlo Masone, and Barbara Caputo. Rethinking visual geo-localization for large-scale applications. In *CVPR*, pages 4878–4888, June 2022.
 - [5] Gabriele Berton, Gabriele Trivigno, Carlo Masone, and Barbara Caputo. Eigenplaces: Training view-point robust models for visual place recognition. In *ICCV*, 2023.
 - [6] Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. Ttt3r: 3d reconstruction as test-time training, 2026.
 - [7] Yu Chen, Shuhan Shen, Yisong Chen, and Guoping Wang. Graph-based parallel large scale structure from motion. *Pattern Recognition*, page 107537, 07 2020.
 - [8] Leo Kaixuan Cheng, Abdus Shaikh, Ruofan Liang, Zhijie Wu, Yushi Guan, and Nandita Vijaykumar. MERG3R: A divide-and-conquer approach to large-scale neural visual geometry. *arXiv preprint arXiv:2603.02351*, 2026.
 - [9] Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. Vggt-long: Chunk it, loop it, align it – pushing vggt’s limits on kilometer-scale long rgb sequences, 2025.
 - [10] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description, 2018.
 - [11] Yixiao Ge, Haibo Wang, Feng Zhu, Rui Zhao, and Hongsheng Li. Self-supervising fine-grained region similarities for large-scale image localization. In *European Conference on Computer Vision*, 2020.
 - [12] Ben Glocker, Shahram Izadi, Jamie Shotton, and Antonio Criminisi. Real-time rgb-d camera relocation. In *ISMAR*, 2013.
 - [13] Michael Grupp. evo: Python package for the evaluation of odometry and slam. <https://github.com/MichaelGrupp/evo>, 2017.
 - [14] Jisang Han, Sunghwan Hong, Jaewoo Jung, Wooseok Jang, Honggyu An, Qianqian Wang, Seungryong Kim, and Chen Feng. Emergent outlier view rejection in visual geometry grounded transformers. *arXiv preprint arXiv:2512.04012*, 2025.
 - [15] Sergio Izquierdo and Javier Civera. Close, but not there: Boosting geographic distance sensitivity in visual place recognition, 2024.
 - [16] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *CVPR*, June 2024.
 - [17] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition, 2023.
 - [18] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kotschieder. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025.
 - [19] Alex Kendall, Matthew Grimes, and Roberto Cipolla. PoseNet: A convolutional network for real-time 6-DoF camera relocation. In *ICCV*, 2015.
 - [20] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.
 - [21] Vincent Leroy, Yohann Cabon, and Jerome Revaud. Grounding image matching in 3d with mast3r, 2024.
 - [22] Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R. Walter, Victor Campagnolo Guizilini, and Greg Shakhnarovich. Fastmap: Revisiting structure from motion through first-order optimization, 2025.
 - [23] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023.
 - [24] Yang Liu, Chuanchen Luo, Zimo Tang, Junran Peng, and Zhaoxiang Zhang. Vggt-x: When vggt meets dense novel view synthesis, 2025.
 - [25] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *ECCV*, 2024.
 - [26] Jerome Revaud, Jon Almazan, Rafael S. Rezende, and Cesar de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *ICCV*, 2019.

- [27] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019.
- [28] Johannes Lutz Schönberger, True Price, Torsten Sattler, Jan-Michael Frahm, and Marc Pollefeys. A vote-and-verify strategy for fast spatial verification in image retrieval. In *ACCV*, 2016.
- [29] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025.
- [30] Zhijian Shu, Cheng Lin, Tao Xie, Wei Yin, Ben Li, Zhiyuan Pu, Weize Li, Yao Yao, Xun Cao, Xiaoyang Guo, et al. Litevggt: Boosting vanilla vgggt via geometry-aware cached token merging. *arXiv preprint arXiv:2512.04939*, 2025.
- [31] Xianbing Sun, Zhikai Zhu, Zhengyu Lou, Bo Yang, Jinyang Tang, Liqing Zhang, He Wang, and Jianfu Zhang. Avggt: Rethinking global attention for accelerating vgggt, 2026.
- [32] Hengyi Wang and Lourdes Wang. Spann3r: 3d reconstruction with spatial memory. In *NeurIPS*, 2024.
- [33] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vgggt: Visual geometry grounded transformer. In *CVPR*, 2025.
- [34] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024.
- [35] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Cut3r: Continuous updating transformer for scene reconstruction. In *CVPR*, 2025.
- [36] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3 : Permutation-equivariant visual geometry learning, 2026.
- [37] Tong Wei, Giorgos Tolias, Jiri Matas, and Daniel Barath. Global-aware edge prioritization for pose graph initialization. In *CVPR*, 2026.
- [38] Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3r: Towards 3d reconstruction of 1000+ images in one forward pass. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2025.
- [39] Shuai Yuan, Yantai Yang, Xiaotian Yang, Xupeng Zhang, Zhonghao Zhao, Lingming Zhang, and Zhipeng Zhang. Infinitevggt: Visual geometry grounded transformer for endless streams, 2026.
- [40] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.
- [41] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. In *ICLR*, 2025.
- [42] Shangzhan Zhang, Jianyuan Wang, Yinghao Xu, Nan Xue, Christian Rupprecht, Xiaowei Zhou, Yujun Shen, and Gordon Wetzstein. Flare: Feed-forward geometry, appearance and camera estimation from uncalibrated sparse views, 2025.
- [43] Jiankun Zhong, Zitong Zhan, Quankai Gao, Ziyu Chen, Haozhe Lou, Jiageng Mao, Ulrich Neumann, Chen Wang, and Yue Wang. Instantsfm: Towards gpu-native sfm for the deep learning era. *arXiv preprint arXiv:2510.13310*, 2025.
- [44] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global SfM by distributed motion averaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4568–4577, 2018.