Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Robust Global Structure-from-Motion via View Graph Pruning

Jiamin Xu
Hangzhou Dianzi University
Lixing Yao
Hangzhou Dianzi University
Weichen Dai
Hangzhou Dianzi University
Renshu Gu
Hangzhou Dianzi University
Zunjie Zhu
Hangzhou Dianzi University
Weiwei Xu
Zhejiang University
Gang Xu
gxu@hdu.edu.cn

Hangzhou Dianzi University

Research Article

Keywords:

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11182723/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Robust Global Structure-from-Motion via View Graph Pruning

Jiamin Xu[1], Lixing Yao[1], Weichen Dai[1], Renshu Gu[1], Zunjie Zhu[1], Weiwei Xu[2], Gang Xu[1]

[1]Hangzhou Dianzi University, Hangzhou, 310018, China.
[2]State Key Lab of CAD&CG, Zhejiang University, Hangzhou, 310058, China.

Abstract

Structure-from-Motion (SfM) aims to estimate camera poses and reconstruct 3D structures from a collection of unordered images. Compared with incremental SfM, global SfM achieves better scalability by jointly estimating camera poses based on a view graph constructed from pairwise correspondences. However, its performance is highly sensitive to erroneous edges caused by visually ambiguous matches, which may lead to incorrect camera registration and reconstruction artifacts. In this work, we propose a subgraph-guided view graph pruning framework for robust global SfM. Our key idea is to exploit the internal consistency of reliable subgraphs to identify and remove unreliable connections. Specifically, we first partition the view graph into locally consistent subgraphs and perform global SfM within each subgraph to obtain reliable camera poses. We then apply RANSAC-based edge pruning across subgraphs to remove inconsistent edges, and finally perform global SfM on the refined view graph. Extensive experiments on ambiguous, sequential, and unordered image datasets demonstrate that our method improves the robustness of global SfM under challenging conditions. Further evaluation with neural rendering shows that the improved camera estimation leads to higher-quality novel view synthesis results.

Keywords: Structure-from-Motion, Global SfM, View Graph, Graph Pruning, 3D Reconstruction

1 Introduction

Structure-from-Motion (SfM) is a fundamental computer vision technique that simultaneously estimates camera poses and reconstructs sparse 3D structures from multiple images. Over the past decade, SfM has found widespread applications across various domains, including indoor–outdoor 3D reconstruction [1, 2], cultural heritage digitization [3], and, more recently, neural rendering [4].

SfM methods can be broadly categorized into incremental and global approaches. Incremental methods start from an initial image pair and sequentially expand the reconstruction by adding new cameras and triangulated points, interleaving pose estimation, triangulation, and bundle adjustment [7]. While they typically achieve high accuracy and robustness, their scalability is limited by repeated optimization. In contrast, global methods construct a view graph with pairwise correspondences and relative poses, and estimate all camera parameters simultaneously through rotation and translation averaging, offering much higher efficiency and scalability [5].

However, because global SfM relies on the view graph as the foundation for averaging, its performance is highly sensitive to erroneous edges—incorrect pairwise relations caused by false feature matches. These edges are particularly common in scenes with repetitive structures, such as building facades or decorative patterns, where

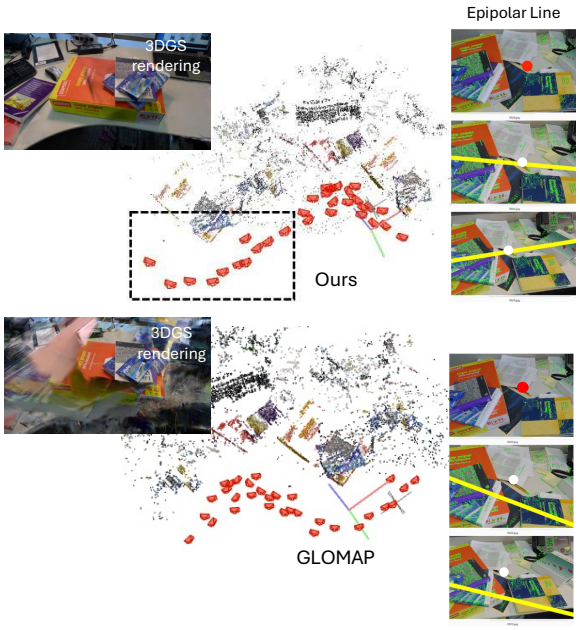


Fig. 1 Our view graph pruning strategy enhances the GLOMAP [5] method for scenes with visually ambiguous regions. *Right*: We evaluate epipolar geometry errors by marking a red point in the reference image and checking whether the epipolar line passes through the corresponding point (marked in white). *Left*: We also assess novel view synthesis results using 3DGS [6]. The results show that our method produces higher-quality novel view images, demonstrating improved SfM consistency.

visually similar regions lead to ambiguous correspondences. Consequently, a key challenge in global SfM is to reliably identify and remove such spurious edges. One approach is to use RANdom SAmple Consensus (RANSAC) [8] to filter out spurious edges. For instance, if the relative pose of an edge $a \leftrightarrow b$ does not align with other paths from a through different nodes to b, it could be considered spurious. However, we found this method ineffective, as false consensus can occur. Specifically, the edge $a \leftrightarrow b$ may be correct but still mistakenly filtered out, while the path through other nodes to b may contain a spurious edge, causing all paths to reach a false consensus.

As a result, we propose an enhanced strategy for identifying spurious edges. Our method first extracts a consensus subgraph and detects outlier edges across subgraphs that deviate from the dominant consensus. Specifically, we iteratively partition the cameras into subgraphs and perform global SfM within each subgraph. If intra-subgraph pose consistency—evaluated through loop consistency or by comparing two-view relative poses with the subgraph's global estimates—is not satisfied, the subgraph is further subdivided. RANSAC is then applied across subgraphs to remove edges that violate inter-subgraph consistency. To resolve scale ambiguities between subgraphs, we employ the Perspective-n-Point (PnP) algorithm to estimate their relative poses. Finally, we re-run global SfM on the refined view graph to obtain the final reconstruction.

Our subgraph-based strategy differs fundamentally from the scene component reconstruction schemes used in GLOMAP [5], COLMAP [7], and commercial software RealityScan [9]. These methods reconstruct individual scene components independently and subsequently merge them into a complete model. In contrast, our subgraph partitioning is designed to filter out spurious edges from the view graph prior to global optimization. By eliminating unreliable connections early, we effectively prevent error propagation during rotation and translation averaging, leading to more accurate and geometrically consistent 3D reconstructions (Fig. 1).

Overall, our contributions can be summarized as:

- We propose a subgraph-guided view graph pruning framework that improves the robustness of global SfM under visually ambiguous conditions.
- We introduce an iterative subgraph partitioning strategy based on intra-subgraph pose consistency, which identifies reliable local geometric structures from corrupted view graphs.
- We develop an inter-subgraph edge pruning strategy based on RANSAC and relative pose consistency to remove unreliable connections while preserving valid geometric relationships.

We evaluate our SfM system on both sequential and unordered image datasets. Our method achieves competitive or superior camera-pose accuracy across diverse datasets, with particularly strong improvements over GLOMAP[5] in visually ambiguous scenes. We also assess novel view synthesis results using 3DGS [6] (Fig. 1), demonstrating improved reconstruction accuracy. Regarding reconstruction efficiency, our method achieves comparable performance to existing global SfM techniques while being significantly faster than most incremental SfM approaches.

2 Related Work

Structure-from-Motion mainly follows two paradigms: incremental and global.

Incremental and global SfM:

Incremental methods [7, 10–13] reconstruct scenes progressively, starting from an initial image pair and sequentially adding new images and 3D points while alternating between pose estimation, triangulation, and local or global bundle adjustment (BA). This strategy typically yields high accuracy and robustness. However, it often suffers from drift accumulation in large-scale reconstructions and is computationally expensive due to repeated BA operations. To alleviate these issues, several approaches perform distributed camera registration followed by global merging [14–16].

Global SfM methods [5, 17–20] estimate all camera poses simultaneously through rotation and translation averaging [19], followed by triangulation and a single global bundle adjustment (BA). Recent advancements, such as GLOMAP [5], enhance efficiency and scalability by jointly optimizing camera poses and 3D structures within a unified global optimization framework. Nevertheless, despite their efficiency, global SfM approaches remain sensitive to outliers, including incorrect correspondences and ambiguities arising from visually indistinguishable structures.

Outlier Rejection in SfM:

A critical challenge in SfM is handling incorrect correspondences and erroneous camera pose estimates. Traditional approaches employ geometric verification using RANSAC [8] with fundamental or essential matrix [21] constraints to prune outliers. In incremental SfM, outliers can also be identified during triangulation [10] and bundle adjustment [22] by measuring reprojection errors, removing points or images that deviate significantly from the reconstructed model [7, 23]. Global SfM methods often incorporate robust optimization objectives (e.g., L1 norms or robust loss functions) [24] in rotation and translation averaging to reduce the impact of inconsistent pairwise estimates [20, 25]. These mechanisms are crucial for improving reconstruction stability; however, they rely heavily on the correctness of the underlying view graph. Consequently, a line of work focuses on improving the reliability and structure of the view graph.

View Graph Partition and Refinement:

View-graph partitioning provides another line of work for improving the scalability and robustness of SfM. Bhowmick et al. [26] applied normalized cuts to the image match graph, reconstructed the resulting components independently, and aligned them using epipolar relationships between connecting images. Toldo et al. [27] constructed a balanced hierarchical dendrogram over unordered images, enabling partial models to be created and merged in a bottom-up manner to reduce bundle-adjustment complexity and limit drift. Chen et al. [16] proposed a graph-based parallel SfM framework that clusters images into strongly connected overlapping groups and merges local reconstructions through graph-based tree structures. Locher et al. [28] introduced progressive SfM, which updates reconstructions as images arrive and avoids binding decisions made from incomplete view graphs. While these methods primarily use partitioning to decompose SfM into local reconstruction and merging stages, our approach employs partitioning to identify subgraphs and refine ambiguous inter-cluster edges, allowing the final model to be recovered directly by global SfM.

One approach to view-graph refinement is to directly remove incorrect edges. In a view-graph, nodes represent images and edges correspond to pairwise epipolar geometries, forming the structural backbone of global SfM. Early refinement methods focused on eliminating unreliable edges: Zach et al. [29, 30] and Cui and Tan [19] exploited missing correspondences and loop-consistency checks to detect false geometric relations, while Roberts et al. [31] introduced an expectation–maximization framework incorporating temporal cues. Wilson and Snavely [32] proposed a visibility-based local clustering measure to assess edge reliability, and learning-based approaches [33, 34] formulated disambiguation as a binary classification task on image pairs. More recently, Wang et al. [35] leveraged high-level spatial context through a novel track-community structure to segment the scene, detect and correct ambiguous regions, and perform partial reconstructions and merging, thereby improving reconstruction accuracy in visually ambiguous scenes.

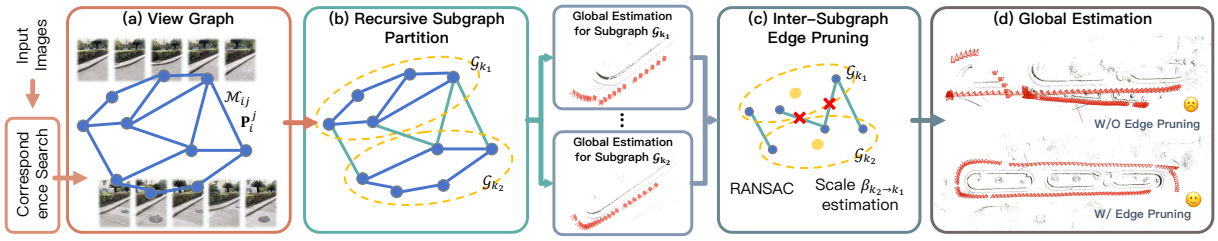


Fig. 2 Our proposed pipeline. Given a set of input images, our method: (a) performs correspondence search to construct a view graph; (b) recursively partitions the view graph into a set of internally consistent subgraphs and applies global SfM to each subgraph using GLOMAP[5]; (c) performs inter-subgraph edge pruning to eliminate visually similar but geometrically inconsistent edges; and (d) applies global SfM to the pruned view graph to obtain the final reconstruction.

Our method also focuses on view-graph refinement in visually ambiguous scenes. Unlike previous approaches, we enhance global SfM by applying RANSAC-based edge refinement between subgraphs (clusters), producing reconstruction results directly from global SfM rather than by merging partial reconstructions. Leveraging global SfM in this way makes our method achieve a favorable trade-off between efficiency and robustness.

3 Proposed Method

3.1 Overview

Our approach builds upon the GLOMAP framework [5]. As shown in Fig. 2, we first reconstruct the view graph using the original GLOMAP pipeline (Sec. 3.2). The view graph is then iteratively partitioned into subgraphs until intra-camera pose consistency is achieved across all subgraphs (Sec. 3.3). For each resulting subgraph, we perform global estimation, including rotation averaging, global positioning, and bundle adjustment. Next, inter-subgraph edge pruning is applied to remove inconsistent connections from the graph (Sec. 3.4). Finally, global estimation is re-applied on the refined view graph to produce the final reconstruction.

In the following sections, we briefly review the key components of GLOMAP as preliminaries (Sec. 3.2), and then introduce the subgraph partitioning and inter-subgraph edge pruning components of our method.

3.2 Preliminaries

View graph reconstruction.

In the first step, GLOMAP [5] performs feature extraction and matching, followed by verification through two-view geometry estimation. The remaining correspondences are considered inliers. Based on these inliers, a view graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ is constructed, where $\mathcal{V}$ denotes the set of images and $\mathcal{E}$ denotes the set of edges.

An edge $\langle i, j \rangle \in \mathcal{E}$ indicates that the corresponding image pair contains a sufficient number of matching inlier points $\mathcal{M}_{ij} = \{\langle \mathbf{x}_n^i, \mathbf{x}_n^j \rangle | n = 1, \dots, M_{ij}\}$ where $\mathbf{x}_n^i, \mathbf{x}_n^j \in \mathbb{R}^2$ denote the 2D pixel coordinates of the n-th matching point in images i and j, respectively. A reliable relative pose $\mathbf{P}_i^j = (\mathbf{R}_i^j, \mathbf{t}_i^j)$ can then be estimated from their two-view geometry [36], where $\mathbf{P}_i^j$ denotes the pose of image j expressed in the coordinate of image i.

Next, using the view graph as input, the global camera pose estimation step simultaneously estimates absolute camera poses $\hat{\mathbf{P}}_i \in \mathrm{SE}(3)$. The hat notation $\hat{\cdot}$ is used to distinguish the globally estimated poses from the relative poses in the view graph obtained via two-view geometry. This procedure involves rotation averaging and global positioning to estimate camera orientations and positions, respectively, followed by a global bundle adjustment to refine the entire reconstruction.

Rotation averaging.

Rotation averaging aims to estimate the absolute rotations $\{\hat{\mathbf{R}}_1, ..., \hat{\mathbf{R}}_N\}$ from a set of relative rotations. GLOMAP [5] formulates this task as a

nonlinear optimization problem:

$$\underset{\hat{\mathbf{R}}_1,\ldots,\hat{\mathbf{R}}_N}{\operatorname{argmin}} \sum_{\langle i,j\rangle\in\mathcal{E}} \mathrm{d}(\mathbf{R}_i^j, \hat{\mathbf{R}}_j^\top \hat{\mathbf{R}}_i), \tag{1}$$

where $\mathrm{d}(\cdot)$ denotes a distance metric on SO(3). Since this optimization is inherently nonlinear and difficult to solve directly, GLOMAP adopts a linearization strategy [37].

Global positioning.

Next, instead of performing translation averaging followed by global triangulation, GLOMAP [5] estimates camera positions and 3D points jointly through a unified optimization. The formulation discards relative translation constraints and relies solely on camera-ray consistency:

$$\underset{\mathbf{X},\,\mathbf{c},\,s_{ik}}{\operatorname{argmin}} \sum_{i,k} \rho\left(\|\mathbf{v}_{ik} - s_{ik}(\mathbf{X}_k - \mathbf{c}_i)\|_2^2\right), \quad \text{s.t. } d_{ik} \geq 0, \tag{2}$$

where $\mathbf{v}_{ik}$ is the camera ray that observes the 3D point $\mathbf{X}_k$ from the camera center $\mathbf{c}_i$, and s_{ik} is a scale factor. The function $\rho(\cdot)$ denotes the Huber loss [38], which serves as a robustifier.

3.3 Subgraph Partition

To identify spurious edges in $\mathcal{G}$, we first partition it into a set of subgraphs $\mathcal{G}_1, \ldots, \mathcal{G}_K$, as shown in Fig. 3. This partitioning is performed using the Louvain community detection algorithm, which greedily optimizes modularity to identify subgraphs with dense intra-community connections and sparse inter-community connections [39] and is applied recursively until intra-subgraph camera pose consistency is achieved. After each partition, we perform GLOMAP's global estimation step for each subgraph to obtain the camera poses $\hat{\mathbf{P}}_i = (\hat{\mathbf{R}}_i, \hat{\mathbf{t}}_i) \in \mathrm{SE}(3)$, which are then used to determine whether further subdivision is necessary and will also be used in subsequent edge pruning step.

Specifically, for a subgraph $\mathcal{G}_k$ with n images (initially $\mathcal{G}$), we compute the similarity matrix $\mathbf{W} \in \mathbb{R}^{n\times n}$ for all image pairs. Each similarity is defined by the number of inlier correspondences $|\mathcal{M}_{ij}|$. We regard $\mathbf{W}$ as the weighted adjacency matrix of $\mathcal{G}_k$, where $\mathbf{A}_{ij} = \mathbf{W}_{ij}$. The subgraph is then partitioned using the Louvain community detection algorithm [39], which greedily optimizes modularity to discover densely connected communities.

For a weighted view graph, the modularity of a partition is defined as

$$\mathbf{Q} = \frac{1}{2m} \sum_{i,j} \left[\mathbf{A}_{ij} - \frac{k_i k_j}{2m}\right] \delta(c_i, c_j), \tag{3}$$

where $\mathbf{A}_{ij}$ denotes the edge weight between images i and j, $k_i = \sum_j \mathbf{A}_{ij}$ is the weighted degree of node i, $m = \frac{1}{2}\sum_{i,j} \mathbf{A}_{ij}$ is the total edge weight of the graph, c_i is the community assignment of node i, and $\delta(c_i, c_j)$ equals 1 if i and j belong to the same community and 0 otherwise.

The Louvain algorithm starts by assigning each image to an individual community. For each node i, it evaluates the modularity gain obtained by moving i into the community of one of its neighboring nodes. The modularity gain is computed as

$$\begin{aligned}\Delta\mathbf{Q} = &\left[\frac{\Sigma_{\text{in}} + k_{i,\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}} + k_i}{2m}\right)^2\right] \\ &- \left[\frac{\Sigma_{\text{in}}}{2m} - \left(\frac{\Sigma_{\text{tot}}}{2m}\right)^2 - \left(\frac{k_i}{2m}\right)^2\right],\end{aligned} \tag{4}$$

where Σ_{in} is the sum of edge weights inside the target community, Σ_{tot} is the sum of edge weights incident to nodes in the target community, k_i is the weighted degree of node i, and $k_{i,\text{in}}$ is the sum of edge weights from node i to nodes in the target community.

Node i is moved to the neighboring community that yields the largest positive $\Delta\mathbf{Q}$; otherwise, it remains in its current community. This local optimization is repeated until no individual node movement can further increase modularity. After that, each discovered community is collapsed into a super-node, and the edge weights between super-nodes are obtained by summing the weights of edges between the corresponding communities. The local optimization and community aggregation steps are then repeated on the coarsened graph until the modularity no longer increases. The final communities are used as the partitioned subgraphs for subsequent inter-cluster edge refinement.

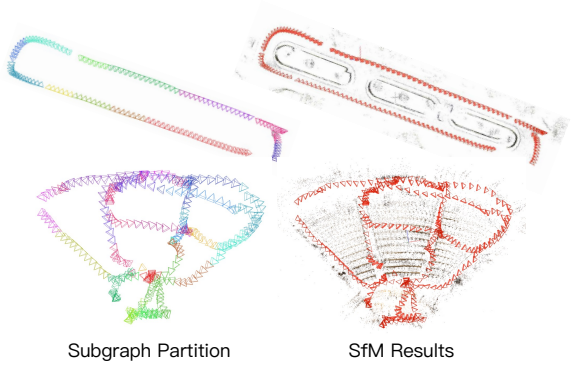


Fig. 3 Visualization of subgraph partition. Each subgraph in the view graph is shown in a different color, demonstrating that our partitioning algorithm can separate visually similar views into distinct subgraphs.

Intra-subgraph camera pose consistency. As a termination criterion for the subdivision, we assess the consistency of camera poses within each subgraph using two metrics. The first is loop consistency, and the second is relative-to-absolute camera pose consistency, which measures the agreement between relative poses estimated from two-view geometry and absolute camera poses obtained from the subgraph's global estimation.

For **loop consistency**, a subgraph is considered inconsistent if any loop—for example, $\mathbf{R}_p^i \mathbf{R}_j^p \mathbf{R}_i^j$—deviates significantly from the identity rotation. Since the relative translations do not share a consistent scale, we only evaluate the loop rotation error:

$$e_{\mathrm{l}} = \|\mathrm{Log}(\mathbf{R}_j^i \dots \mathbf{R}_i^j)\|_2 \tag{5}$$

where $\mathrm{Log}(\cdot)$ maps a rotation matrix in SO(3) directly to its corresponding axis-angle vector. For efficiency, we only consider loops of length l that contain fewer than eight cameras. Since consistency is evaluated within each subgraph, this simplification is considered acceptable. If any loop exhibits a loop-consistency error e_{l} greater than $5^\circ \cdot \left\lfloor \frac{l-1}{2} \right\rfloor$, the subgraph is further subdivided.

The **relative-to-absolute camera pose consistency** metric evaluates the agreement between the relative pose of an edge in the view graph and the corresponding pose derived from the subgraph's global estimation in GLOMAP. Since these poses may have different scales, only the rotation error and the scale-invariant translation error are considered:

$$e_{\mathrm{r}} = \|\mathrm{Log}(\hat{\mathbf{R}}_j \mathbf{R}_i^j \hat{\mathbf{R}}_i^\top)\|_2, \tag{6}$$

$$e_{\mathrm{t}} = \mathrm{acos}\left(\frac{(\mathbf{t}_i^j)^\top \hat{\mathbf{R}}_j^\top (\hat{\mathbf{t}}_i - \hat{\mathbf{t}}_j)}{\|\mathbf{t}_i^j\| \left\| \hat{\mathbf{R}}_j^\top (\hat{\mathbf{t}}_i - \hat{\mathbf{t}}_j) \right\|} \right). \tag{7}$$

If, for any edge in the subgraph, either e_{r} or e_{t} exceeds 10°, the subgraph is further subdivided.

3.4 Inter-Subgraph Edge Pruning

After partitioning the entire view graph into a set of subgraphs, we perform inter-subgraph edge pruning to remove inconsistent edges from $\mathcal{G}$ (Fig. 2). The input to this step consists of the complete view graph $\mathcal{G}$ and the set of subgraphs $\{\mathcal{G}_1, \dots, \mathcal{G}_K\}$. For the complete view graph, no global estimation is performed; consequently, it does not contain absolute camera poses or a global point cloud. In contrast, after the subgraph partitioning step, each subgraph $\mathcal{G}_k$ maintains its own set of absolute camera poses $\hat{\mathbf{P}}_i^k = (\hat{\mathbf{R}}_i^k, \hat{\mathbf{t}}_i^k) \in \mathrm{SE}(3)$ and a corresponding local point cloud $\mathbf{X}^k$. Here, the subscript k indicates that both the camera poses and the point cloud are represented in the local coordinate frame.

For any two subgraphs $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$, we define the set of inter-subgraph edges as those in the view graph whose endpoints belong to different subgraphs: $\{\langle i, j\rangle \in \mathcal{E}' | i \in \mathcal{G}_{k_1}, j \in \mathcal{G}_{k_2}\}$. Since the absolute poses of different subgraphs are estimated in their own local coordinate systems and may differ in scale, for each image j on an inter-subgraph edge, we use the Perspective-n-Point (PnP) method to estimate its pose in the coordinate system of image i's subgraph $\mathcal{G}_{k_1}$. We then align the two subgraphs by estimating a relative scale factor $\beta_{k_2 \to k_1}$ based on all edges connecting the two subgraphs.

For an image i in subgraph $\mathcal{G}_{k_1}$, we obtain its 2D correspondences with image j as $\langle \mathbf{x}_i, \mathbf{x}_j \rangle$. To simplify notation, we slightly overload the symbol $\mathbf{x}_i$ to represent all matching points in image i for the image pair $\langle i, j \rangle$. The corresponding 3D points can be retrieved from the point clouds of $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$, denoted as $\mathbf{X}_i^{k_1}$ and $\mathbf{X}_j^{k_2}$, respectively. The camera poses $\hat{\mathbf{P}}_i^{k_1}$ and $\hat{\mathbf{P}}_j^{k_2}$ are also assumed to be known.

The camera pose of image j in the local coordinate system of $\mathcal{G}_{k_1}$, denoted as $\hat{\mathbf{P}}_j^{k_1}$, can be estimated using the PnP formulation:

$$s \begin{bmatrix} \mathbf{x}_j & 1 \end{bmatrix}^\top = \mathbf{K}_j (\hat{\mathbf{P}}_j^{k_1})^{-1} \begin{bmatrix} \mathbf{X}_j^{k_1} & 1 \end{bmatrix}^\top, \tag{8}$$

where $\mathbf{K}_j$ denotes the camera intrinsic matrix obtained in the previous step. To solve this problem, we employ the Efficient PnP (EPnP) algorithm [40], which introduces a set of virtual control points to linearize the otherwise nonlinear PnP problem, thereby significantly improving computational efficiency.

For inter-subgraph edges, we apply RANSAC [8] along the path connecting $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$. In each RANSAC iteration, a scale factor $\beta_{k_2 \to k_1}$ is estimated to align the scales between the two subgraphs, as their reconstructions are generally inconsistent in scale. First, we define the pose of each subgraph, as the average pose of all images in subgraph as $\mathbf{P}_{c_1}^{k_1}$ and $\mathbf{P}_{c_2}^{k_2}$, in each local coordinate. Then we can compute the relative pose between c_1 and c_2 in the $\mathcal{G}_{k_2}$'s local coordinate:

$$\dot{\mathbf{P}}_{c_1 \to i \to j \to c_2}^{k_2} = (\mathrm{S}_{k_1}(\hat{\mathbf{P}}_{c_2}^{k_2}))^{-1} \hat{\mathbf{P}}_j^{k_2} (\hat{\mathbf{P}}_j^{k_1})^{-1} \hat{\mathbf{P}}_{c_1}^{k_1}. \tag{9}$$

Here, $\dot{\mathbf{P}}^{k_2}$ denotes the pose in the coordinate system of $\mathcal{G}k_2$ after scaling to align with $\mathcal{G}k_1$, and $\mathrm{S}_{k_1}(\cdot)$ represents the operation of scaling a subgraph to match $\mathcal{G}_{k_1}$:

$$\mathrm{S}_{k_1}(\hat{\mathbf{P}}_{c_2}^{k_2}) = \begin{bmatrix} \hat{\mathbf{R}}_{c_2}^{k_2} & \beta_{k_2 \to k_1} \hat{\mathbf{t}}_{c_2}^{k_2} \\ 0 & 1 \end{bmatrix}, \tag{10}$$

To determine the scale factor, we minimize the discrepancies among the translation differences of all relative poses $\mathbf{P}_{c_1 \to j \to c_2}^{k_2}$ for edges connecting subgraphs $\mathcal{G}_{k_1}$ and $\mathcal{G}_{k_2}$:

$$\beta^* = \underset{\beta}{\operatorname{argmin}} \sum_{\langle i,j \rangle \in \mathcal{E}''} \| \dot{\mathbf{t}}_{c_1 \to i \to j \to c_2}^{k_2} - \bar{\mathbf{t}}_{c_1 \to c_2}^{k_2} \|_2^2, \tag{11}$$

$$\bar{\mathbf{t}}_{c_1 \to c_2}^{k_2} = \frac{1}{|\mathcal{E}''|} \sum_{\langle i,j \rangle \in \mathcal{E}''} \dot{\mathbf{t}}_{c_1 \to i \to j \to c_2}^{k_2}, \tag{12}$$

where $\dot{\mathbf{t}}^{k_2}$ denotes the translation component of $\dot{\mathbf{P}}^{k_2}$, and $\mathcal{E}''$ represents the subset of edges sampled from the original set $\mathcal{E}'$ connecting the two subgraphs during the RANSAC procedure. Since this is a linear least-squares problem, it can be solved directly.

After scaling, we calculate the rotation error for each sampled edge in $\mathcal{E}''$:

$$E_\mathrm{r} = \| \mathrm{Log}(\dot{\mathbf{R}}_{c_1 \to i \to j \to c_2}^{k_2}) - \bar{\mathbf{r}}_{c_1 \to c_2}^{k_2} \|_2, \tag{13}$$

$$\bar{\mathbf{r}}_{c_1 \to c_2}^{k_2} = \frac{1}{|\mathcal{E}''|} \sum_{\langle i,j \rangle \in \mathcal{E}''} \mathrm{Log}(\dot{\mathbf{R}}_{c_1 \to i \to j \to c_2}^{k_2}), \tag{14}$$

and the translation error:

$$E_\mathrm{t} = \| \dot{\mathbf{t}}_{c_1 \to i \to j \to c_2}^{k_2} - \bar{\mathbf{t}}_{c_1 \to c_2}^{k_2} \|_2. \tag{15}$$

During each RANSAC iteration, we sample $0.4 \cdot |\mathcal{E}'|$ edges between each pair of subgraphs. The thresholds for E_r and E_t are set to 5° and $0.05 \cdot \left\| \bar{\mathbf{t}}_{c_1 \to c_2}^{k_2} \right\|_2$, respectively. The RANSAC procedure runs for 30 iterations, during which we retain the largest subset $\mathcal{E}''$. The remaining edges, i.e., $\mathcal{E}' \setminus \mathcal{E}''$, are then discarded. After pruning the inter-subgraph edges, global estimation is reapplied on the updated view graph to obtain the final reconstruction, as shown in Fig. 2.

4 Experiments

We compare our method with five SfM techniques, evaluating performance both quantitatively and qualitatively. Among these methods, GLOMAP [5] is a global SfM approach and serves as our baseline, as our method is based on it. COLMAP [7] is the most widely used incremental SfM method and is often employed as the first step in neural rendering pipelines such as 3DGS [6]. GraphSfM [16] is an SfM method designed to handle large-scale scenes. Recent learning-based SfM methods Doppelgangers [33] and Doppelgangers++ [34] are designed to handle scenes with ambiguous structures.

Datasets.

We evaluate our method on four types of datasets: ambiguous image datasets [31, 41], sequential image datasets [42], unordered Internet image datasets [43], and our captured dataset. For all evaluated scenes, the ground-truth camera poses are reconstructed using RealityScan [9] with manually annotated control points. To ensure the reliability of the reconstructed poses, we further

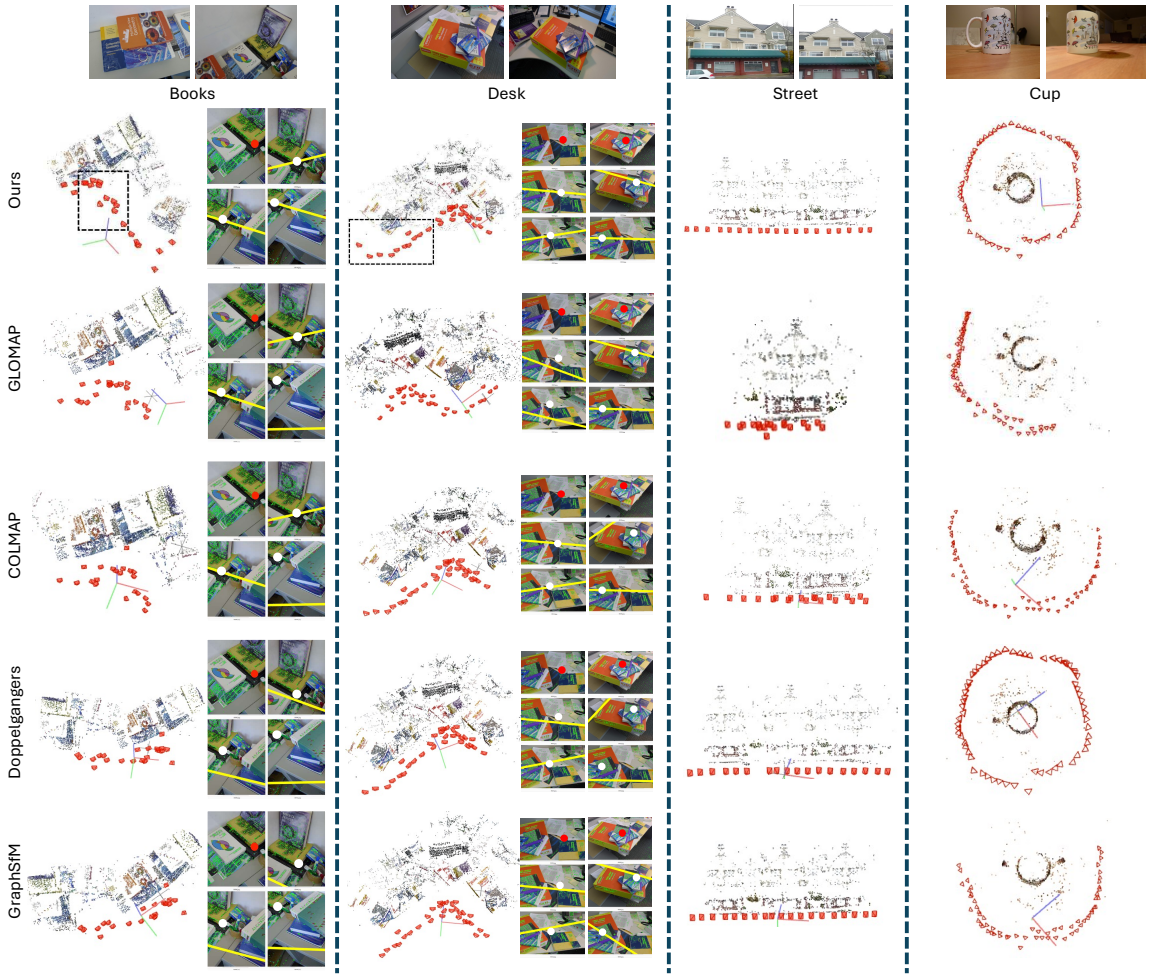


Fig. 4 Qualitative evaluation on ambiguous image datasets. *The first row* shows the input images of selected views. *The second row* shows our results and the rest rows shows competing methods. In each unit, we provide the SfM results (point cloud and camera trajectory) as well as epipolar line verification. The top left corner of the book is selected (marked in red) and epipolar lines are drawn on the other three views; ideally the epipolar line should pass through the corresponding point (marked in white) in every view.

conduct epipolar line verification on the resulting camera geometry.

4.1 Evaluation on Ambiguous Image Datasets

We test our method on the datasets from [31], specifically the "books", "desk", "street" and "cup" benchmark datasets. The visualization results are shown in Fig. 4. For every unit, we show the reconstructed results as well as epipolar line verification on the right. We select a reference point on the first image (marked in red), and draw the epipolar lines in the other 3 views, which should ideally pass through the corresponding point (marked in white). Our method shows robust performance on all four datasets.

GLOMAP [5] occasionally gives reasonable results for ambiguous datasets such as "Books". However, in some cases, such as "desk", GLOMAP incorrectly registers cameras and points of duplicated structures, resulting in incorrect reconstruction. Additionally, GLOMAP performs poorly on the "cup" dataset, where it mistakenly connects the front and back faces of the cup in the view graph, due to their visual similarity, resulting in suboptimal performance. COLMAP [7] produces incomplete structures for the "cup" dataset, although slightly better than GLOMAP. For

"books" and "desk", the epipolar line verification shows that COLMAP is inaccurate as the epipolar lines do not always pass through the top left corner of the book (marked in white). GraphSfM [16] cannot reconstruct the complete camera trajectory for "cup". It also does not satisfy the epipolar constraints well. Doppelgangers [33] performs well on the "cup" dataset but still struggles with registration accuracy in other cases. In contrast, all four scenes are successfully recovered using our proposed method.

Table 1 reports the quantitative comparison on four ambiguous image datasets, i.e., Books, Desk, Street, and Cup. For all evaluations, we report the AUC (Area Under the recall Curve) scores calculated from the maximum of relative rotation and translation error between every image pair, similar to [5], such an error formulation considers the deviation between every possible camera pair. We evaluate all methods using the AUC thresholds of 3°, 5°, and 10°. On Books, our method achieves the best results across all thresholds, reaching 98.7, 99.3, and 99.4, which significantly outperforms the second-best Doppelgangers++ [34] by 54.2, 51.3, and 48.7 points, respectively. On Desk, our method also consistently ranks first, improving over the second-best COLMAP by 9.9, 8.4, and 7.1 points. For Street, our method obtains 96.0, 97.6, and 98.8, again achieving the best performance and surpassing Doppelgangers++ [34] at all thresholds. These results demonstrate that our method is particularly effective in resolving ambiguous epipolar relationships and producing more reliable image associations.

On Cup, Doppelgangers++ [34] achieves the best performance, while our method ranks second with 66.4, 78.4, and 87.6. Although our method is not the best on this dataset, it still clearly outperforms COLMAP, GLOMAP, Doppelgangers [33], and GraphSfM under all three thresholds. Overall, across the four datasets, our method achieves the highest average AUC values of 90.1, 93.7, and 96.4 at 3°, 5°, and 10°, respectively. This indicates that our method provides strong and stable performance on ambiguous image datasets, especially in challenging cases where repetitive structures can easily lead to incorrect geometric verification.

Since metrics such as AUC cannot fully capture the accuracy of camera registration, we also evaluate our method for novel view synthesis using 3DGS [6] in scenes without ground truth camera poses. Table 2 presents the novel view synthesis evaluation results. We use 80% of the images for training and 20% for testing, and all models are trained for 30,000 iterations. We compare the Peak Signal-to-Noise Ratio (PSNR) on the testing images, and the results show that our method achieves competitive performance and consistently improves over traditional SfM approaches, demonstrating its ability to produce globally consistent camera poses.

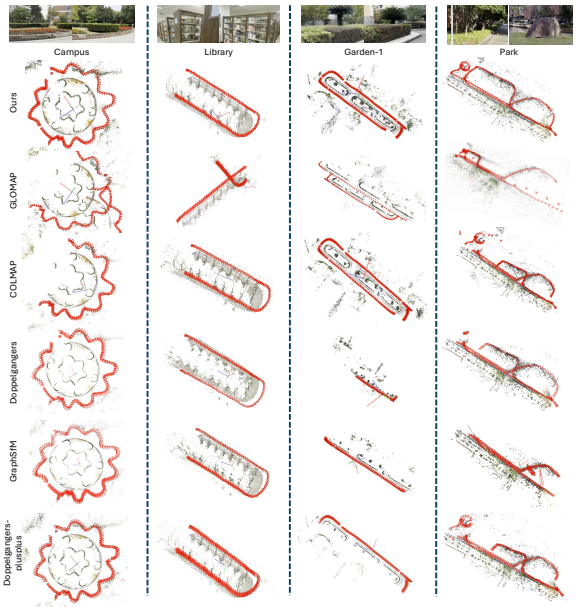


Fig. 5 Qualitative evaluation on sequential datasets.

4.2 Evaluation on Sequential Datasets

We also evaluate our method on sequential image datasets from [42], covering outdoor and indoor scenes. Specifically, we provide results on the "Family", "Auditorium", "Courthouse" and "Meetingroom" datasets. Again, we compare our method with GLOMAP [5], COLMAP [7], Doppelgangers [33], Doppelgangers++ [34] and GraphSfM [16]. The quantitative evaluation results are presented in Table 1. Our method achieves substantially higher AUC scores than COLMAP[7] on both Family and Auditorium,

Method	COLMAP[7]			GLOMAP[5]			Doppelgangers[33]			Doppelgangers++[34]			GraphSfM[16]			Ours		
	auc@3° ↑	*auc@5°* ↑	*auc@10°* ↑	*auc@3°* ↑	*auc@5°* ↑	*auc@10°* ↑	*auc@3°* ↑	*auc@5°* ↑	*auc@10°* ↑	*auc@3°* ↑	*auc@5°* ↑	*auc@10°* ↑	*auc@3°* ↑	*auc@5°* ↑	*auc@10°* ↑	*auc@3°* ↑	*auc@5°* ↑	*auc@10°* ↑
Books [31]	18.2	27.4	38.8	17.3	27.0	40.0	43.5	47.4	50.4	44.5	48.0	50.7	20.2	28.8	39.5	**98.7**	**99.3**	**99.4**
Desk [31]	89.2	91.0	92.6	51.8	53.0	53.9	88.9	90.8	92.2	89.0	90.8	92.2	48.1	60.0	70.8	**99.1**	**99.4**	**99.7**
Street [31]	28.3	33.3	40.5	17.5	19.5	27.2	28.0	43.5	67.9	87.4	92.4	96.2	20.2	28.5	39.1	**96.0**	**97.6**	**98.8**
Cup [31]	29.8	37.3	43.7	23.6	27.2	29.9	28.0	43.5	67.9	**93.6**	**96.2**	**98.1**	20.4	30.0	37.5	66.4	78.4	87.6
Indoor [41]	25.7	35.7	43.9	23.6	32.6	40.0	55.9	73.0	86.5	65.2	78.8	**89.4**	25.6	35.6	44.0	**68.0**	**79.9**	89.3
Brandenburg gate [43]	29.0	46.5	67.3	28.7	45.4	65.2	**56.1**	**62.0**	**67.5**	23.1	36.5	52.6	41.2	45.6	50.0	27.0	42.8	61.6
Arc de triomphe [43]	14.6	22.8	34.2	14.2	22.9	35.5	57.0	61.6	65.8	17.2	30.6	52.4	20.1	22.3	24.4	**73.9**	**80.4**	**86.3**
Family [42]	20.3	33.9	58.5	81.5	88.9	94.4	**87.5**	**92.5**	**96.2**	19.1	32.5	56.6	87.1	92.2	96.1	80.4	88.2	94.1
Auditorium [42]	16.5	32.9	60.0	18.2	35.5	60.8	**75.9**	**85.2**	**92.5**	16.7	33.0	60.1	14.2	15.9	17.4	40.6	64.7	79.4
exhibition_hall [44]	65.4	77.5	87.2	33.6	38.8	42.9	49.0	59.6	69.6	61.0	72.6	82.0	44.3	55.6	68.2	**70.7**	**82.1**	**91.6**
office [44]	0.0	0.0	0.0	45.1	52.7	61.2	19.5	23.4	29.1	0.0	0.0	0.2	15.1	20.3	25.7	**47.8**	**54.0**	**61.9**
old_computer [44]	6.8	9.6	13.7	42.7	50.2	57.0	**65.6**	**76.7**	**85.9**	0.3	0.6	1.7	8.5	10.7	12.1	64.3	75.7	85.5
Campus (Ours)	42.1	45.9	48.8	34.5	42.5	49.3	47.3	67.2	83.2	27.9	41.6	63.8	41.7	59.8	79.1	**80.7**	**87.8**	**93.2**
Library (Ours)	78.5	86.4	93.2	43.2	45.7	47.7	75.3	84.1	92.0	**79.8**	**87.2**	**93.6**	74.1	83.6	91.8	78.8	86.7	93.3
Garden-1 (Ours)	14.4	29.5	54.0	0.8	1.7	5.4	7.7	11.5	14.4	17.9	23.5	27.9	0.9	1.9	5.2	**64.4**	**74.8**	**84.8**
Park (Ours)	23.4	37.1	53.5	15.6	24.8	39.4	30.5	40.9	52.5	29.4	44.1	59.4	17.7	21.6	25.1	**76.2**	**83.9**	**90.8**

Table 1 Comparison with COLMAP [7], GLOMAP [5], Doppelgangers [33], Doppelgangers++ [34], GraphSfM [16], and Ours. The best results are shown in bold, and the second-best results are underlined.

although Doppelgangers[33] achieves the best performance on these two sequences. In this dataset, COLMAP [7] recovers structures robustly; however, its computational cost is significantly higher than ours, as shown in Table 3. Please refer to the supplementary material for additional qualitative comparisons.

We additionally evaluate three sequences from the benchmark of[44]: "Exhibition Hall," "Office," and "Old Computer."Our method achieves the best AUC scores on Exhibition Hall and Office and ranks second to Doppelgangers on Old Computer.

We also recorded our own datasets, namely "Campus", "Library", "Garden-1", and "Park", as shown in Fig. 5. For GLOMAP [5] in Row 2, a significant portion of each scene is missing, primarily because our captured scenes contain large areas with similar appearances. GLOMAP [5] is relatively sensitive to matching errors caused by these ambiguities. Doppelgangers [33] uses a network to address this issue, resulting in good performance on the "Campus", "Library", "Park" datasets. However, it misses a large portion of the structures in "Garden-1". COLMAP [7] recovers the complete structures in most scenes but terminates midway on the "Campus" scene, possibly due to false positive edges being selected as the next view. GraphSfM [16] also produces unsatisfactory results for the "Garden-1" and "Park' datasets. The quantitative results provided in Table 1 show that the AUC scores of our reconstruction is generally higher than GLOMAP [5].

4.3 Evaluation on Unordered Internet Image Datasets

Moreover, we evaluated our method on unordered Internet image datasets [43]. We tested on the "indoor" dataset from [41], and the "Brandenburg Gate" and "Arc de Triomphe" from [43]. Comparisons with GLOMAP [5], COLMAP [7], Doppelgangers[33], Doppelgangers++ [34] and GraphSfM [16] are also included in Table 1. On Indoor, our method achieves the best AUC@3° and AUC@5° and ranks second at AUC@10°, with 89.3 compared with 89.4 for Doppelgangers++. On Brandenburg Gate, Doppelgangers performs best, whereas on Arc de Triomphe, our method achieves the best results across all three thresholds. We include a qualitative comparison in the supplementary materials.

Method	COLMAP PSNR↑	GLOMAP PSNR↑	Doppelgangers PSNR↑	GraphSfM PSNR↑	Ours PSNR↑
Books	17.38	17.32	16.77	17.92	**22.63**
Cup	17.61	14.59	29.78	15.85	**31.20**
Desk	17.78	22.20	24.54	24.50	**26.31**
Street	18.51	16.89	17.74	17.69	**21.93**

Table 2 Evaluation using 3DGS. Our estimated camera poses produce improved novel view synthesis results, demonstrating that our method can produce globally consistent camera poses.

Method	#Image	COLMAP Time	GLOMAP Time	Doppelgangers Time	GraphSfM Time	Ours Time
Meetingroom	371	7 min	3.5 min	120 min	13 min	5.5 min
Courthouse	1106	210 min	30 min	300+ min	83 min	66 min

Table 3 Computational time comparison. The runtime of our method remains substantially lower than COLMAP[7] and Doppelgangers[33], while introducing additional computational overhead compared with GLOMAP[5].

Dataset	#Image	Full AUC@10°↑	W/O Recursive AUC@10°↑	W/O e_{l} AUC@10°↑
Garden-1	140	**84.8**	14.6	34.7

W/O e_{r} AUC@10°↑	W/O e_{t} AUC@10°↑	#RC=20 AUC@10°↑	#RC=40 AUC@10°↑
25.9	31.2	37.3	84.8

Table 4 Ablation studies. AUC@10° is reported on the Garden-1 dataset. "W/O Recursive" denotes performing only a single Louvain partition without recursively subdividing inconsistent subgraphs. W/O e_{l}, W/O e_{r}, and W/O e_{t} denote removing the loop consistency, relative rotation consistency, and relative translation consistency terms, respectively. #RC denotes the number of RANSAC iterations, while the full model uses 30 iterations.

4.4 Ablation study

To validate the effectiveness of the key components in our method, we conduct ablation studies on the recursive subgraph partitioning strategy, the consistency criteria, and the number of RANSAC iterations. The results are reported in Table 4.

We first evaluate the effect of recursive subgraph partitioning. Removing the recursive partitioning strategy and performing only a single Louvain partition significantly decreases the AUC@10° from 84.8 to 14.6. This result demonstrates that a single graph partition is insufficient to reliably isolate geometrically inconsistent regions. Recursively subdividing subgraphs according to their internal pose consistency is therefore important for obtaining reliable local geometric structures before inter-subgraph edge pruning.

We further evaluate the contribution of the three consistency criteria. Removing the loop consistency term e_{l} , the relative rotation consistency term e_{r} , or the relative translation consistency term e_{t} decreases AUC@10° from 84.8 to 34.7, 25.9, and 31.2, respectively. The largest degradation occurs when e is removed, while all three terms contribute substantially to robust view graph refinement.

Finally, we study the effect of the number of RANSAC iterations in the inter-subgraph edge pruning stage. Reducing the number of iterations from 30 to 20 decreases the AUC@10° from 84.8 to 37.3, indicating that an insufficient number of iterations may fail to identify a reliable consensus among inter-subgraph edges. Increasing the number of iterations to 40 yields the same AUC@10° of 84.8 as the full setting. This suggests that 30 iterations are sufficient on Garden-1 under this evaluation metric, while increasing the number to 40 provides no further improvement in AUC@10°. Therefore, we use 30 RANSAC iterations in all experiments to balance robustness and computational efficiency.

Computational time. Table 3 provides a comparison of the computational time for all methods. As our approach follows the global SfM paradigm, it is highly time-efficient, particularly for longer sequences. Although our method incurs slightly more time compared to the vanilla GLOMAP [5], it still runs over three times faster than COLMAP [7] and approximately five times faster than Doppelgangers [33] on the "Courthouse" dataset. For detailed information on our hardware, please refer to the supplementary materials.

5 Conclusion

This paper introduces a robust global SfM method capable of handling large scenes with repetitive or duplicate structures. The main components include a subgraph partitioning strategy to maintain intra-subgraph pose consistency and a RANSAC-based strategy to filter inconsistent edges across different subgraphs.

Limitation and future work. Our method may still encounter errors when repetitive regions are too large or densely distributed. Additionally, our approach relies on several hyperparameters that may not be optimal. In the future, we plan to incorporate learning-based techniques or advanced matching strategies to overcome these limitations.

References

[1] Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.-M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)

[2] Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. European Conference on Computer Vision (ECCV) (2018)

[3] Fuhrmann, S., Langguth, F., Goesele, M.: Mve-a multi-view reconstruction environment. GCH **3**(4), 2 (2014)

[4] Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)

[5] Pan, L., Baráth, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: European Conference on Computer Vision (ECCV) (2024)

[6] Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)

[7] Schonberger, J.L., Frahm, J.-M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)

[8] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981)

[9] CapturingReality: Reality capture, http://capturingreality.com (2016)

[10] Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)

[11] Frahm, J.-M., Fite-Georgel, P., Gallup, D., Johnson, T., Raguram, R., Wu, C., Jen, Y.-H., Dunn, E., Clipp, B., Lazebnik, S., *et al.*: Building rome on a cloudless day. In: European Conference on Computer Vision, pp. 368–381 (2010). Springer

[12] Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3d. In: ACM Siggraph 2006 Papers, pp. 835–846 (2006)

[13] Wu, C.: Towards linear-time incremental structure from motion. In: 2013 International Conference on 3D Vision-3DV 2013, pp. 127–134 (2013). IEEE

[14] Cui, H., Shen, S., Gao, W., Hu, Z.: Efficient large-scale structure from motion by fusing auxiliary imaging information. IEEE Transactions on Image Processing **24**(11), 3561–3573 (2015)

[15] Zhu, S., Zhang, R., Zhou, L., Shen, T., Fang, T., Tan, P., Quan, L.: Very large-scale global sfm by distributed motion averaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4568–4577 (2018)

[16] Chen, Y., Shen, S., Chen, Y., Wang, G.: Graph-based parallel large scale structure from motion. Pattern Recognition **107**, 107537 (2020)

[17] Arie-Nachimson, M., Kovalsky, S.Z., Kemelmacher-Shlizerman, I., Singer, A., Basri, R.: Global motion estimation from point matches. In: 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization & Transmission, pp. 81–88 (2012). IEEE

[18] Cai, Q., Zhang, L., Wu, Y., Yu, W., Hu, D.: A pose-only solution to visual reconstruction and navigation. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 73–86 (2021)

[19] Cui, Z., Tan, P.: Global structure-from-motion by similarity averaging. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 864–872 (2015)

[20] Wilson, K., Snavely, N.: Robust global translations with 1dsfm. In: European Conference on Computer Vision, pp. 61–75 (2014). Springer

[21] Andrew, A.M.: Multiple view geometry in computer vision. Kybernetes **30**(9/10), 1333–1341 (2001)

[22] Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment—a

modern synthesis. In: International Workshop on Vision Algorithms, pp. 298–372 (1999). Springer

[23] Kang, L., Wu, L., Yang, Y.-H.: Robust multi-view l2 triangulation via optimal inlier selection and 3d structure refinement. Pattern Recognition **47**(9), 2974–2992 (2014)

[24] Barron, J.T.: A general and adaptive robust loss function. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4331–4339 (2019)

[25] Goldstein, T., Hand, P., Lee, C., Voroninski, V., Soatto, S.: Shapefit and shapekick for robust, scalable structure from motion. In: European Conference on Computer Vision, pp. 289–304 (2016). Springer

[26] Bhowmick, B., Patra, S., Chatterjee, A., Govindu, V., Banerjee, S.: Divide and Conquer: Efficient Large-Scale Structure from Motion Using Graph Partitioning, vol. 9004, pp. 273–287 (2015). https://doi.org/10.1007/978-3-319-16808-1_19

[27] Toldo, R., Gherardi, R., Farenzena, M., Fusiello, A.: Hierarchical structure-and-motion recovery from uncalibrated images. Computer Vision and Image Understanding **140**, 127–143 (2015) https://doi.org/10.1016/j.cviu.2015.05.011

[28] Locher, A., Havlena, M., Van Gool, L.: Progressive structure from motion. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)

[29] Zach, C., Klopschitz, M., Pollefeys, M.: Disambiguating visual relations using loop constraints. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 1426–1433 (2010). IEEE

[30] Zach, C., Irschara, A., Bischof, H.: What can missing correspondences tell us about 3d structure and motion? In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). IEEE

[31] Roberts, R., Sinha, S.N., Szeliski, R., Steedly, D.: Structure from motion for scenes with large duplicate structures. In: CVPR 2011, pp. 3137–3144 (2011). IEEE

[32] Wilson, K., Snavely, N.: Network principles for sfm: Disambiguating repeated structures with local context. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 513–520 (2013)

[33] Cai, R., Tung, J., Wang, Q., Averbuch-Elor, H., Hariharan, B., Snavely, N.: Doppelgangers: Learning to disambiguate images of similar structures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 34–44 (2023)

[34] Xiangli, Y., Cai, R., Chen, H., Byrne, J., Snavely, N.: Doppelgangers++: Improved visual disambiguation with geometric 3d features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 27166–27175 (2025)

[35] Wang, L., Ge, L., Luo, S., Yan, Z., Cui, Z., Feng, J.: Tc-sfm: Robust track-community-based structure-from-motion. IEEE Transactions on Image Processing **33**, 1534–1548 (2024)

[36] Larsson, V., contributors: PoseLib - Minimal Solvers for Camera Pose Estimation (2020). https://github.com/vlarsson/PoseLib

[37] Chatterjee, A., Govindu, V.M.: Efficient and robust large-scale rotation averaging. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 521–528 (2013)

[38] Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in Statistics: Methodology and Distribution, pp. 492–518. Springer, ??? (1992)

[39] Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E.: Fast unfolding of communities in large networks. Journal of statistical mechanics: theory and experiment **2008**(10), 10008 (2008)

[40] Lepetit, V., Moreno-Noguer, F., Fua, P.: Ep n p: An accurate o (n) solution to the p n p problem. International journal of computer vision **81**(2), 155–166 (2009)

[41] Jiang, N., Tan, P., Cheong, L.-F.: Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1458–1465 (2012). IEEE

[42] Knapitsch, A., Park, J., Zhou, Q.-Y., Koltun, V.: Tanks and temples: Benchmarking large-scale scene reconstruction. ACM Transactions on Graphics (ToG) **36**(4), 1–13 (2017)

[43] Heinly, J., Dunn, E., Frahm, J.-M.: Correcting for duplicate scene structure in sparse 3d reconstruction. In: European Conference on Computer Vision, pp. 780–795 (2014). Springer

[44] Schöps, T., Schönberger, 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: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2538–2547 (2017). https://doi.org/10.1109/CVPR.2017.272