Global Structure-from-Motion Meets Feedforward Reconstruction

Linfei Pan
ETH Zurich

Johannes Schönberger
Meta Reality Labs

Marc Pollefeys
ETH Zurich, Microsoft

Abstract

Structure-from-Motion – the process of simultaneously estimating camera poses and 3D scene structure from a collection of images – remains a central challenge in computer vision, with many open problems yet to be solved. Recent advances in feedforward 3D reconstruction have made significant strides in overcoming persistent failure cases of classical SfM methods, particularly in scenarios characterized by low texture, limited overlap, and symmetries. However, while feedforward approaches excel in these challenging conditions, they often face limitations regarding scalability, accuracy, or robustness, and typically fall short of classical methods in standard reconstruction settings. In this work, we systematically analyze these limitations and propose a new Structure-from-Motion pipeline by combining the respective strengths of classical and feedforward methods. Extensive experiments across multiple datasets show the benefits of our approach, achieving state-of-the-art results across a wide range of scenarios. We share our system as an open-source implementation at https://github.com/colmap/gluemap.

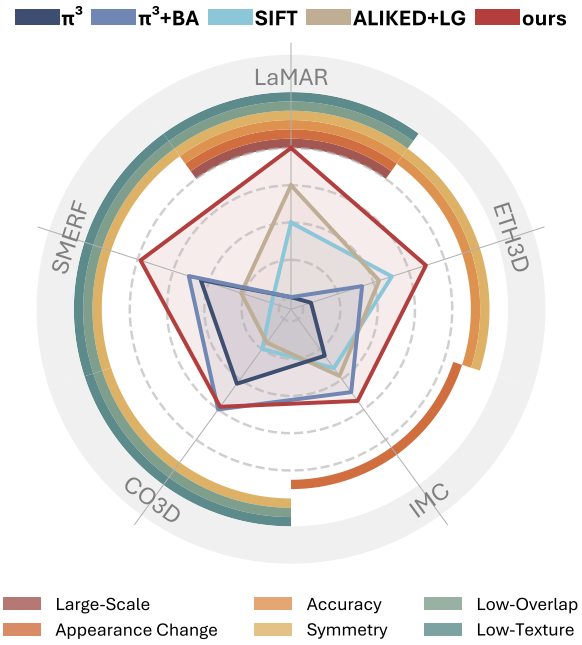


Figure 1. We evaluate on 5 datasets, featuring different challenges. The average rank on each dataset is reported. Classical and feedforward methods struggle under different settings, while ours consistently maintains best performance across the board.

1. Introduction

Structure-from-Motion (SfM) tackles the problem of reconstructing 3D scene structure and cameras given a set of images. It is a fundamental technique in computer vision and serves as a critical building block for numerous applications like localization [32], multi-view stereo [35], novel-view-synthesis [17], or 3D training data generation [44].

Throughout its long history [41], progress in the field has been primarily driven by optimization-based algorithms [21, 25, 27, 33, 38, 48], which we refer to as *classical* methods in the remainder of this text. While these approaches differ in their specific formulations, they generally share a common structure: correspondence search using pairwise (local) feature matching and (global or incremental) reconstruction using robust optimization. To this day, SIFT [22] remains the standard choice for correspondence search. The global reconstruction paradigm [27] sets itself apart from the incremental [33] one by estimating camera poses and scene structure in a single, unified global optimization step, rather than incrementally building up a partial reconstruction using repeated local and global optimizations. The global approach generally results in improved runtime scalability and better robustness against symmetry issues. The state-of-the-art classical SfM systems [27, 33] achieve high levels of reliability [28] for sufficiently overlapping input images with enough parallax and discriminative scene texture. In contrast, these methods frequently struggle when the input images are sparse, have limited parallax, or do not contain enough distinctive features to match.

To address the limitations of classical approaches, the research community has increasingly focused on replacing hand-crafted modules in the 3D reconstruction pipeline with learned feedforward components, while still leverag-

ing classical optimization techniques. For example, learned features like ALIKED [53] and learned matchers like LightGlue [20] enable more reliable matching under challenging appearance changes. More recently, fueled by significant advances in large-scale model training, the community has achieved breakthrough results by solving the multi-view 3D reconstruction problem end-to-end with a single feedforward architecture, effectively eliminating the need for explicit geometric solvers and optimization. After a long period of incremental progress, these methods have enabled a substantial leap forward, resolving several persistent failure cases that challenged traditional approaches.

However, these methods are no silver bullet. In many cases, they still lag behind classical techniques in terms of scalability, accuracy, and robustness. The reliance on heavy transformer-based architectures leads to high memory requirements, limiting scalability to only several hundred input views at relatively low image resolutions, thus limiting accuracy. Recent efforts [8, 50] have sought to address these scalability issues, but they still struggle to process beyond a thousand images. Moreover, as we will demonstrate later, these approaches can exhibit counterintuitive behavior in which adding more input images does not necessarily improve the outputs. For scenarios where classical SfM methods succeed, feedforward approaches typically lag significantly behind in terms of accuracy. When feedforward methods do approach classical performance in these nominal cases, they often do so by incorporating optimization-based bundle adjustment (BA), which only partially bridges the gap. In terms of robustness, none of the existing methods can reliably handle multiple connected components, resolve symmetric structures, or systematically reject outliers such as irrelevant input images.

Our key contributions are an analysis of the limitations of classical and feedforward 3D reconstruction. We then use these insights to propose a novel pipeline by combining their respective strengths. Extensive experiments across multiple datasets show the benefits of our approach, achieving state-of-the-art results across a wide range of scenarios.

2. Related Work

2.1. Classical Methods

Classical SfM algorithms broadly categorize into incremental and global methods. Both types begin with correspondence search, which typically involves extracting local image features [9, 22, 53] and matching them across images [20]. To establish the view graph, images can be paired exhaustively or, for greater scalability, by employing image retrieval techniques [1, 34] to identify overlapping image pairs. Correspondence search concludes by estimating two-view geometries [13] using robust estimation techniques.

The primary distinction between incremental and global pipelines lies in how they estimate cameras and structure. Incremental SfM pipelines – such as Bundler [38] or COLMAP [33] – build the reconstruction by adding images one at a time. At each step, they perform robust BA to jointly refine cameras and 3D points, interleaving this optimization as new images are incorporated. Global SfM pipelines, in contrast, estimate all cameras and 3D points simultaneously from the entire set of images. Historically, incremental methods have been regarded as more robust, particularly in challenging scenarios. However, recent methods like GLOMAP [27] demonstrate that global SfM pipelines can now achieve comparable robustness and accuracy but with better efficiency and scalability characteristics. In our work, we build upon the global SfM paradigm.

In global SfM, camera intrinsics, rotations, and translations are usually recovered in different stages. For camera intrinsics, Sweeney *et al*. [39] proposed to perform view-graph calibration. The estimated intrinsics are then used to decompose two-view geometries into relative camera poses before estimating global camera rotations using rotation averaging [5, 12, 24]. Next, many systems use translation averaging [2, 26, 54] to solve for global camera translations, but since pairwise translations are only up-to-scale, the formulation is often ill-posed. GLOMAP [27] addresses this with global positioning to simultaneously recover camera poses and 3D points, but it can get stuck in local minima when tracks are insufficient. Similarity averaging [6] takes a different approach by using depth-derived scale constraints but remains sensitive to noise. All approaches finalize the reconstruction with global BA to improve accuracy.

Classical SfM methods face significant challenges in certain types of scenes, particularly those that are texture-less, have low image overlap, or exhibit low parallax and symmetries. In areas with **low texture**, local feature matching becomes unreliable or even impossible. Without sufficient correspondences, accurate two-view geometry estimation is not feasible. This, in turn, leaves BA under-constrained, often resulting in reconstruction failures. When images have **low overlap**, it becomes difficult to constrain the scale of the reconstruction. Since relative camera poses estimated from two views are only determined up to an unknown scale, at least three-view overlap is necessary to achieve consistent and accurate global scale. Furthermore, relative pose estimation becomes degenerate when image pairs exhibit **low parallax** (*i.e*., when the camera motion is mostly rotational or the scene is far away). This degeneracy can cause the entire pipeline to fail. Last but not least, the presence of visually similar or **symmetric scene structure** (*i.e*. Doppelgangers [4, 51]) leads to unresolved ambiguities and most often results in collapsed 3D reconstructions.

2.2. Feedforward Methods

Recent feedforward methods approach 3D reconstruction by learning to infer scene geometry and cameras in an end-to-end fashion, leveraging large-scale training datasets that typically include both synthetics and 3D models generated by classical techniques. This allows them to learn complex priors to solve some of the failure cases of classical methods. Depending on their network architecture, feedforward methods can be broadly categorized into three groups: diffusion, recurrent networks, and transformers.

CameraAsRays [52] and PoseDiffusion [42] are two examples from the first category. Initialized from random positions, they adopt diffusion processes to obtain the final pose estimation. CUT3R [45] and its follow-ups are recurrent methods. It maintains a scene state and incrementally reconstructs each incoming image. DUSt3R [46] and MASt3R [18] represent two-view transformer-based methods. Images are patchified into a set of tokens and then fed through a set of transformation layers. These networks then directly regress 3D points in both images, and estimate camera poses and calibrations via RANSAC. However, since it only receives two-view input, the scale between image pairs is estimated by non-metric depth, which is often unstable. Methods like VGGT [44], MV-DUSt3R [40], and MapAnything [16] took a step further to directly estimate multi-view reconstructions end-to-end. π^3 [47] currently represents the state of the art by improving upon VGGT using a permutation invariant loss formulation. However, these methods have their intrinsic limitations, as we show later.

In terms of **scalability**, diffusion- and transformer-based methods are inherently bound by GPU memory. Even though Fast3R [50], FastVGGT [37], StreamVGGT [55], and SAIL-Recon [7] propose mechanisms to reduce memory usage, they are still limited to several hundreds of images. For recurrent networks, though theoretically scalable, they maintain a fixed state or they suffer from the same limitation as incremental classical approaches, where a single bad decision can lead to an unrecoverable failure.

Transformer-based models achieve the best **accuracy** [44, 47] across all feedforward approaches. Yet, in scenes where classical methods work well, transformer-based models still lag significantly behind in terms of camera pose accuracy – a limitation that is often underreported in prior literature. To better understand these challenges, we conduct an extensive analysis on datasets that are difficult for both classical and feedforward methods.

In terms of **robustness**, diffusion- and transformer-based models for multi-view estimation treat all images equally, enabling every image to interact with another. This full connectivity works well for object-centric scenes. However, in complex, large-scale scenes – characterized by a large view graph radius – attending information globally becomes problematic. The resulting quadratic increase in possible connections makes it difficult to distinguish relevant from irrelevant information, leading to significant performance drops. This problem is exacerbated in the presence of symmetric scene structure. In contrast, recurrent network models depend on a specific, typically sequential, input order. If images are provided in random order, or if an image lacks visual overlap with previously processed images, pose estimation can fail entirely.

2.3. Hybrid Methods

There are some existing attempts in leveraging the merits of both categories to improve system performance.

Early feedforward methods primarily aimed to enhance individual components of the classical SfM pipeline, such as image retrieval, pair filtering, feature extraction, feature matching, and optimization. For instance, NetVLAD [1] and more recent methods like SALAD [14] or MegaLoc [3] showcased strong improvements over traditional bag-of-words approaches. Doppelgangers [4] and its successor Doppelgangers++ [49] developed feedforward networks to identify image pairs with symmetry issues. SuperPoint [9] and ALIKED [53] are notable learned feature extractors, while SuperGlue [30] and LightGlue [20] proposed feedforward methods for feature matching. These methods serve as an add-on to classical SfM, leaving the optimization formulation unchanged. Meanwhile, PixSfM [19] performs a joint refinement over learned features and structure. Liu *et al*. [21] developed a hybrid SfM system to jointly reconstruct points and lines. MP-SFM [28] addresses some limitations of classical SfM by incorporating learned monocular priors into an incremental pipeline. However, it struggles to scale to large problem instances, due to the incremental nature of the pipeline and the computational cost of depth and normal optimization.

On the feedforward side, efforts have been made in improving scalability and accuracy. MASt3R-SfM [11] performs two-view inference on the view graph and obtains a multi-view reconstruction by minimizing the 3D point cloud alignment and reprojection errors. VGGT-Long [8] and VGGT-SLAM [23] propose to divide a long sequential input into small segments and apply factor graph optimization to align them while VGGT [44] and VGGSfM [43] produce tracks as part of outputs and inject these into bundle adjustment for improving the final pose accuracy. However, these methods are often still less accurate than the classical SfM systems. In contrast, our method achieves state-of-the-art performance under a wide range of input scenarios and scales well to tens of thousands of images.

3. Method

In this section, we introduce our approach for integrating classical and feedforward techniques into a unified, end-to-end reconstruction system. The overall pipeline consists of

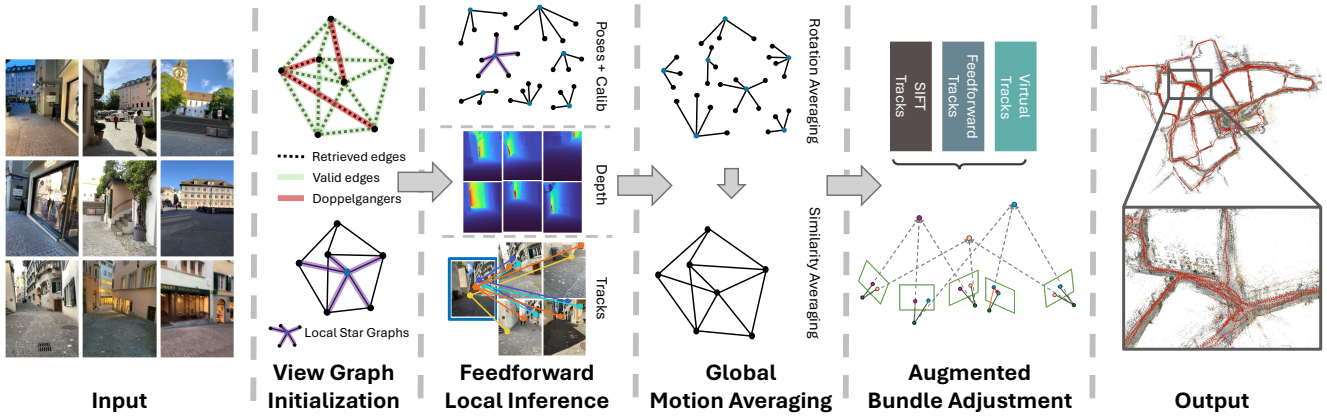


Figure 2. Illustration of our proposed GLUEMAP pipeline consisting of four major steps: view graph initialization, feedforward local inference, global motion averaging, and augmented bundle adjustment. The proposed method combines the advantages of both feedforward and classical methods and can efficiently scale to tens of thousands of images while remaining robust.

four key stages: (1) *view graph initialization* selects tentatively overlapping view pairs using scalable image retrieval and Doppelganger filtering, (2) from the resulting local star graphs centered on each input view, *feedforward local inference* then estimates local reconstructions, (3) *global motation averaging* performs rotation and similarity averaging to initialize the global reconstruction from the local stars, and (4) *augmented bundle adjustment* jointly refines camera poses and structure using a combination of classical SIFT and virtual feedforward tracks. An overview of our GLUEMAP system is shown in Figure 2.

Our system takes as input a set of $i = 1...n$ unordered images $\mathcal{I} = \{I_i \in \mathbb{R}^{H_i \times W_i \times 3}\}$ and estimates $k = 1...m$ scene points $\mathcal{X} = \{X_k \in \mathbb{R}^3\}$ as well as camera poses $\mathcal{P} = \{P_i = (R_i, t_i) \in \mathrm{SE}(3) \mid i \in \mathcal{I}^*\}$ and intrinsics $\mathcal{K} = \{\pi_i \in \mathbb{R}^3 \to \mathbb{R}^2 | \; i \in \mathcal{I}^*\}$ for a subset $\mathcal{I}^*$ of confidently registered images. Further technical details for each stage are provided in the following sections.

3.1. View Graph Initialization

In this stage, we begin by tentatively selecting overlapping image pairs using scalable image retrieval techniques combined with Doppelganger filtering. Instead of the standard feedforward approach of globally attending over all images, we use the resulting sparse view graph $G(\mathcal{I}, \mathcal{E})$ to only locally attend feedforward reconstruction. This allows our approach to more efficiently scale to an arbitrary number of input images, in particular by avoiding out-of-memory issues and by batch-reconstructing many local problems in parallel. Furthermore, by restricting attention to only the most relevant information and by explicit Doppelganger filtering, feedforward inference becomes significantly more accurate and robust against symmetry issues.

More specifically, for each input image I_i, we retrieve a fixed number of c candidate neighbors $\mathcal{C}_i$ using SALAD [14], resulting in a total of $O(c \cdot n)$ candidate pairs. Then, for each of the pairs $(i, j), j \in \mathcal{C}_i$, we identify potentially non-overlapping or Doppelganger edges as

$$\alpha_{ij} = \mathrm{DG}(I_i, I_j) \tag{1}$$

where $\alpha_{i,j}$ are Doppelgangers++ (DG) [49] scores. To ensure local connectivity, we apply dynamic thresholding on the scores as follows. Starting from an empty view graph $G_{t=0}$ with $\mathcal{E}_{t=0} = \emptyset$, we iteratively add edges between different connected components $\mathrm{CC}(i)$ as

$$\mathcal{E}_{t+1} = \mathcal{E}_t \cup \{(i, j) \mid \alpha_{ij} > \delta_t, \mathrm{CC}(i) \neq \mathrm{CC}(j)\} \tag{2}$$

with an initial filtering threshold $\delta_0 = 0.8$. If the updated graph G_{t+1} is connected, the iteration halts and the graph is accepted. Otherwise, the threshold is lowered as $\delta_{t+1} = \delta_t - 0.1$. The iteration halts if $\delta_t < 0.2$ and the largest connected component is kept. The final view graph G_T defines a local neighborhood for each image l as $\mathcal{N}_l = \{l\} \cup \{m \mid (l, m) \in \mathcal{E}_T\}$.

3.2. Feedforward Local Inference

Next, feedforward reconstruction proceeds from the established view graph. To this end, the view graph is decomposed into local star graphs $S_l = G(\mathcal{N}_l, \; \{(l, m) \mid m \in \mathcal{N}_l, l \neq m\})$ which we batch-reconstruct independently as

$$(\mathcal{P}_l, \mathcal{F}_l, \mathcal{D}_l, \mathcal{T}_l) = \mathrm{FF}(I_{\mathcal{N}_l}) \;, \tag{3}$$

where we use π^3 [47] (FF) to infer local poses $\mathcal{P}_l = \{P_i | \; i \in \mathcal{N}_l\}$, depth maps $\mathcal{D}_l = \{D_i \in \mathbb{R}_0^{H_i \times W_i} | \; i \in \mathcal{N}_l\}$, focal lengths $\mathcal{F}_l = \{f_i \in \mathbb{R}^+ | \; i \in \mathcal{N}_l\}$, and tracks $\mathcal{T}_l$ [43]. We keep the 25 frames with the highest DG scores if there are more neighbors than that.

Since each image is part of $|\mathcal{N}_l|$ local star graphs, we infer overlapping local reconstructions. To merge the overlapping tracks across stars, we snap track positions to SIFT [22] keypoints within a radius $\beta = 1$px and merge tracks snapping to the same keypoints. The global set of merged tracks across all stars is denoted as $\mathcal{T}$.

For each local star, we further apply a forward-backward depth consistency check to determine visual overlap. More specifically, we calculate the one-way reprojection error $\epsilon_{i\to j}$ from any image i to j in star S_l as

$$X_i = D_i(x,y)\cdot(u,v,1)^\top \tag{4}$$
$$(x',y')^\top = \Pi_j(R_{ij}X_i + t_{ij}) \tag{5}$$
$$X_j = D_j(x',y')\cdot(u',v',1)^\top \tag{6}$$
$$(x'',y'')^\top = \Pi_i(R_{ij}X_j + t_{ij}) \tag{7}$$
$$\epsilon_{i\to j} = \|(x,y)^\top - (x'',y'')^\top\|_2 \ , \tag{8}$$

where $(x,y) \in \mathbb{R}^2$ are pixel image coordinates and $(u,v,1) = \pi^{-1}(x,y)$ are normalized coordinates, respectively. For simplicity, l is omitted in the above equations. Using a reprojection threshold τ, the raw overlap ratio $\tilde{o}^l_{ij}$ between i and j in star S_l is defined as

$$\tilde{o}^l_{ij} = \frac{1}{W_i H_i}\cdot\sum_{(x,y)\in H_i\times W_i}\mathbb{1}(\epsilon_{i\to j}<\tau) \ . \tag{9}$$

To measure transitive co-visibility, we define

$$o^l_{ij} = \max_{\tilde{\mathcal{O}}\in\mathcal{O}_{i,j}}\prod_{(p,q)\in\tilde{\mathcal{O}}}\tilde{o}^l_{pq} \ , \tag{10}$$

where $\mathcal{O}_{ij}$ represents all the paths between image i and j in the fully connected graph on $\mathcal{N}_l$. Edges with small o^l_{ij} are filtered unless removing them disconnects the view graph.

3.3. Global Motion Averaging

Using global SfM techniques, this stage merges the n independent local reconstructions into a global one. More specifically, camera intrinsics, rotations, and centers are estimated with intrinsics averaging, rotation averaging, and similarity averaging, respectively.

First, for intrinsics averaging, we simply calculate the median of all inferred focal lengths per physical camera.

Next, rotation averaging, also known as rotation synchronization, estimates global camera rotations R_i from a set of relative rotations R_{ij} by optimizing

$$\min_R\sum_{(i,j)\in E}\rho\left(o^l_{ij}\cdot d(R^l_{ij}, R_jR_i^\top)\right) \ , \tag{11}$$

where d is the geodesic error function, ρ is the Huber loss as a robustifier, and R^l_{ij} is the relative rotation derived from the locally inferred camera poses in star S_l.

After global rotation averaging, we use similarity averaging [6] to infer camera centers $c_i \in \mathbb{R}^3$. Relative camera translations can be calculated as

$$t^l_{ij} = s^l\cdot R^l_{ij}(c_i - c_j) \ , \tag{12}$$

where t^l_{ij} is the relative translation from locally inferred camera poses in star S_l. Since relative translations within local star reconstruction are scale-consistent, only a single scale s^l is needed for each star. Note that this is different from the original formulation [6], where relative scales between individual edges are estimated from noisy triangulations, which is more error-prone than our formulation.

The camera centers c_i and star scales s_l are estimated by

$$\min_{c,s}\sum_{l,(i,j)\in S_l}o_{ij}\cdot d\left(R_{ij}^\top t_{ij} - s_l\cdot(c_i - c_j)\right), \ s_0 = 1 \tag{13}$$

We discuss an alternative formulation and its relation to translation averaging in the supplementary material. We initialize this optimization using the maximum spanning tree, where the weight of each edge is the overlap ratio o^l_{ij}.

Finally, globally scale-consistent depth maps can be computed as $\tilde{D}^l_i = \frac{1}{s_l}D^l_i$ with $\tilde{D}^l_i$ as the depth map for image i in S_l.

3.4. Augmented Bundle Adjustment

The accuracy of the final reconstruction is improved by a process referred to as augmented bundle adjustment (BA). As discussed in Section 2.1, standard BA formulations are only well-conditioned when there are sufficiently many tracks $\mathcal{T}$ with many-view overlap [28]. Due to low-overlap view configurations, it can be impossible to establish such tracks even with theoretically perfect matching algorithms. Furthermore, even with sufficient overlap, low texture may prevent tracks with enough overlap in practice. In contrast, multi-view feedforward models can overcome this drawback by leveraging sophisticated scene priors to infer accurate relative camera poses and consistent depth maps with two-view or sometimes with no view overlap at all. We encode these scene priors by augmenting the standard BA formulation with *virtual tracks* as follows.

We form two types of virtual tracks by reprojection of sampled pixels (x,y) in each star's center image l as

$$\mathcal{V}^l_{l\to m} = \Pi_m\left(R^l_{lm}\left(\tilde{D}^l_l(x,y)\cdot\Pi_l^{-1}(x,y) - c^l_{lm}\right)\right) \tag{14}$$

$$\tilde{\mathcal{V}}^l_{l\to m} = \Pi_m\left(R_m\left({R_l}^\top\tilde{D}^l_l\cdot\Pi_l^{-1}(x,y) + c_l\right) - c_m\right) \ . \tag{15}$$

The resulting $|\mathcal{N}_l|$-view tracks $\mathcal{V} = \{\mathcal{V}^l_{l\to m}|\ m\in\mathcal{N}_l, l\neq m\}$ and equivalently $\tilde{\mathcal{V}}$ are conditioned on the respective poses from the feedforward local inference and the global motion averaging results. In contrast to standard feature tracks, we allow virtual tracks to project outside of neighboring images, and the virtual 3D points may be observed behind the neighboring cameras. For numerical stability, we ignore observations coinciding with the imaging plane during the optimization. Empirically, we chose to sample ≈ 100 virtual tracks with a ratio of 10% of tracks being conditioned on global camera poses. Intuitively, a higher ratio of tracks conditioned on global poses leads to a BA result closer to the output of global motion averaging.

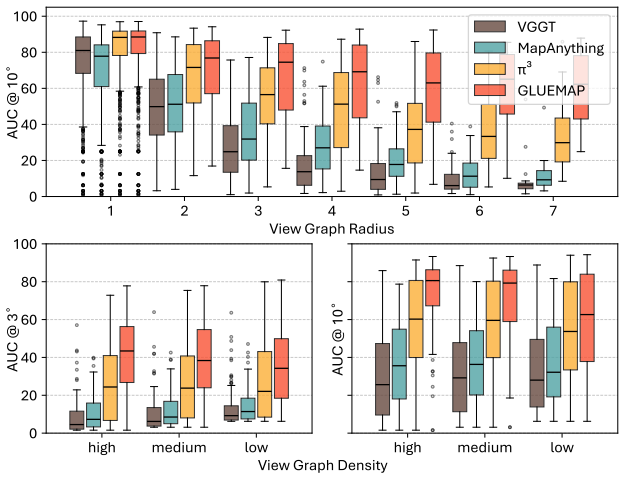


Figure 3. Impact of view graph radius and density. The performance of feedforward methods drops with increasing radius and stays relatively stable with different densities. Our method is more robust to increasing radius and benefits from higher density.

For the final augmented BA problem, we use three types of tracks: tracks $\mathcal{T}$ from the feedforward network, virtual tracks $\mathcal{V}$ and $\tilde{\mathcal{V}}$ as described above, as well as classical SIFT tracks. We use a standard reprojection cost function Π over all images with Huber for SIFT and feedforward tracks, and with Arctan for virtual tracks as robustifiers. For the snapping of $\mathcal{T}$, we obtain SIFT features as a free side product and found them to improve reconstruction accuracy. 3D positions of virtual tracks are known by construction, while those of the other tracks are obtained by triangulation.

4. Experiments

In this section, we first empirically analyze the behavior of feedforward methods in terms of the most important structural properties of a 3D reconstruction problem. We then compare the performance of classical, feedforward, and our proposed method in extensive experiments on datasets covering a wide range of 3D reconstruction challenges. Figure 1 summarizes all results.

4.1. Metrics

Throughout all experiments, we follow standard practice [27, 44, 47] and measure AUC@X (Area Under the recall Curve) scores calculated based on pose errors, which is defined as the maximum of relative rotation and translation errors between every possible image pair at different angular error thresholds X. For tighter thresholds, the scores reflect accuracy of the reconstruction. For looser thresholds, they reflect the completeness of the reconstruction.

4.2. Feedforward Graph Analysis

In this experiment, we evaluate the performance of recent state-of-the-art feedforward methods with VGGT [44], MapAnything [16] (*MA*), and π^3 [47], as well as the proposed motion averaging result, in terms of structural scene complexity. To this end, we sample view graphs with varying radius $r = \min_{(i,j)\in\mathcal{E}} \max_{(i,j)\in\mathcal{E}} \delta(i,j)$ and density $\rho = \max_{(i,j)\in\mathcal{E}} \min_{(i,j)\in\mathcal{E}} \delta(i,j)$ with δ defining the graph distance. For this evaluation, we chose LaMAR [31] as it contains a large number of sequences covering both indoor and outdoor scenes with a large variety in terms of view graph properties. The details of the sampling can be found in the supplementary material.

The view graph radius – defined as the minimum eccentricity among all vertices in a connected graph – is used as a measure of scene complexity. It determines the minimum number of passes required to exchange information between every pair of images in the graph, reflecting the intrinsic difficulty of the reconstruction problem. We determine the groundtruth view graph for each subsequence by rendering depth maps from the provided mesh. The results are summarized in the upper part in Figure 3, where we observe a consistent performance drop with increasing radius, while π^3 consistently outperforms VGGT and MapAnything. The proposed method achieves the best accuracy, especially for view graphs with a large radius. This is because the proposed method inherits the classical SfM pipeline, thus it is more stable with increasing graph radius.

To analyze the impact of view densities, we compare the performance on the same segments with different frame sampling rates. A higher sampling rate leads to sparser inputs. Results are summarized in the lower part in Figure 3. For lower densities, counterintuitively, the accuracy of pose estimation improves. This difference is especially observable for VGGT [44] and MapAnything [16]. For higher densities, the performance stays at similar levels, with a slightly increased variance. This behavior is further underlined by the results in Tables 2, where performance of feedforward methods degrades with more input views of the same scene. Inversely, the proposed method behaves like other optimization-based pipelines, and the performance increases with the density of input, thanks to the additional redundancy of observations.

4.3. Comparative Analysis on Real-World Datasets

We conduct experiments on multiple datasets covering a wide range of challenges: ETH3D [36], IMC2021 [15], CO3Dv2 [29], SMERF [10] and LaMAR [31].

We evaluate state-of-the-art methods from both the classical and feedforward categories. For classical ones, we chose GLOMAP + SIFT (*SIFT*), GLOMAP + ALIKED + LightGlue (*AL+LG*) as baselines. To rule out the impact of image pairs with symmetry, for the classical baselines, we use the same input view graph using Doppelganger++ [49] filtering. This removes the potential performance differences caused by symmetry. For feedforward methods, we first compare their performance on ETH3D [36] and then

Table 1. Results on ETH3D [36]. Results with * are with groundtruth calibration. GLUEMAP achieves best overall results while π^3 is the best feed-forward method.

AUC@	SIFT	AL+LG	π^3	π^3 + BA	GLUEMAP†	GLUEMAP	GLUEMAP*
1	45.6	42.9	13.2	30.6	20.3	53.0	74.0
3	62.2	62.1	36.1	55.1	49.0	76.9	85.9
5	66.7	67.4	48.9	65.1	61.9	83.6	89.0

AUC@	CUT3R	VGGT	MA	MASt3R-SfM	MP-SFM* (s)	MP-SFM* (d)
1	5.0	8.6	5.1	39.2	74.3	70.3
3	11.4	24.0	11.1	55.6	-	-
5	18.8	35.0	18.3	60.5	88.3	88.2

select π^3 as the best performing method for all other experiments. We also include results for π^3 with bundle adjustment (π^3 + *BA*) and our results after global motion averaging (*GLUEMAP*†) as additional baselines. We additionally compare to MP-SFM with both sparse (*MP-SFM (s)*) and dense (*MP-SFM (d)*) tracks on ETH3D and SMERF.

Experiments are conducted on GH200 GPU with 96GB memory, but our method can fit on the RTX 4090 with 24GB memory.

ETH3D [36] consists of an unordered collection of high-resolution images of both outdoor and indoor scenes with millimeter-accuracy groundtruth. We use this dataset to analyze the performance of different methods with a focus on **accuracy**. We adopt the same thresholds used in [27] and the results are summarized in Table 1. To fairly compare with MP-SFM [28], we also report results with ground truth intrinsics (marked with *). Our method achieves the highest accuracy in both the calibrated and uncalibrated settings, with a large performance gap compared to feed-forward methods. The performance difference to classical methods is less pronounced. We attribute the slight performance gap to the local robustness provided by the feed-forward backbone, as well as some scenes exhibiting very low overlap. Meanwhile, π^3 demonstrates a clear performance advantage over CUT3R [45], VGGT [8], and MapAnything [16], supporting the choice of our backbone.

IMC2021 (Image Matching Challenge 2021) [15] features unordered internet photo collections of outdoor landmarks captured by heterogeneous devices under drastic **appearance changes**. The biggest challenge are extreme illumination and geometric changes. We used the same thresholds used in [44] and we report the results for different collections separately in Table 2. From the table, when the number of images is small, with large appearance changes, classical methods struggle to establish enough matches, resulting in comparatively low accuracy. However, when more images from the collection are included, low matching efficiency is compensated for by high density. As a result, GLOMAP + SIFT achieves the best performance. Our method remains competitive across different input sizes, inheriting the benefits from both categories.

CO3Dv2 [29] features a large set of object-centric

Table 2. Results on IMC2021 [15]. The relative performances of classical and feedforward methods changes with the number input views, while our method is competitive across all settings.

	bag 5			bag 10			bag 25			Full		
AUC@	3	5	10	3	5	10	3	5	10	3	5	10
SIFT	39.6	47.3	57.0	50.1	60.9	72.5	64.4	74.3	84.2	76.9	83.8	90.1
AL+LG	48.4	58.5	70.6	50.9	62.2	74.3	54.7	64.9	75.6	62.8	72.4	82.5
π^3	46.2	58.4	73.0	39.7	53.4	69.7	36.6	51.1	68.4	35.2	49.2	66.2
π^3 + BA	54.0	64.9	77.6	54.1	65.3	77.5	57.8	69.1	80.9	45.8	54.8	65.9
GLUEMAP†	46.2	58.4	72.9	39.8	53.5	69.9	36.7	51.2	68.5	36.7	51.7	69.1
GLUEMAP	54.3	65.3	77.8	58.0	69.3	81.2	63.5	74.0	84.4	73.0	81.3	89.1

Table 3. Results on CO3Dv2 [29]. Ours is on-par with feedforward approaches while classical ones fall far behind.

	10 images			20 images			40 images			Average		
AUC@	3	10	30	3	10	30	3	10	30	3	10	30
SIFT	25.8	36.3	42.1	35.1	47.4	53.9	50.0	65.9	73.3	37.0	49.9	56.4
AL+LG	20.6	29.2	35.0	40.8	56.1	64.0	52.9	71.5	80.7	38.1	52.2	59.9
π^3	48.2	77.1	89.9	46.4	76.2	89.2	47.1	76.5	89.3	47.3	76.6	89.5
π^3 + BA	55.3	78.8	90.1	59.4	80.9	90.9	60.5	81.3	91.2	58.4	80.3	90.7
GLUEMAP†	47.0	76.6	89.5	47.1	76.6	89.3	48.2	77.1	89.6	47.4	76.8	89.5
GLUEMAP	54.8	79.3	90.3	56.7	79.8	90.3	58.7	80.7	90.7	56.7	79.9	90.4

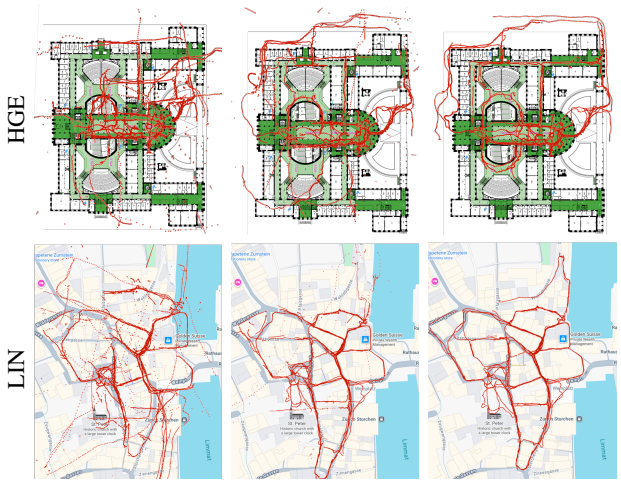
scenes presenting the methods with challenges in terms of **low-texture** and **low-overlap**. Following the practice in [44, 47], we randomly sample 10 images from each test sequence, and report AUC scores there. We also sample at most 10 sequences from each category with 20 and 40 randomly sampled images. Results are summarized in Table 3. On **sparse** sequences, a clear performance gap between classical and feedforward methods can be seen. For denser sequences with more images, the gap diminishes. Our method maintains high performance in both settings.

SMERF [10] contains four indoor captures covering multiple rooms. Following the setup in [28], we use it to evaluate **low-overlap** scenarios. Results are summarized in Table 4 and indicate that classical methods largely fail with low overlap. π^3 also does not achieve high scores because the view-graph radius and symmetry is high for scenes in this dataset, leading to multiple rooms collapsing in the reconstructions. By using Doppelganger++ filtering, our method can successfully distinguish different rooms, achieving satisfying results after motion averaging. And with the help of virtual tracks, the proposed method improves on the tightest threshold from bundle adjustment while avoiding significant degradation of reconstruction completeness, which is captured by AUC@20. Notably, when compared with MP-SFM [28], which was specifically targeting at low-overlap scenes, we achieve better accuracy than the version with sparse tracks. For the dense version of MP-SFM, it benefits from the dense correspondences and achieves better scores on the tightest threshold, while we achieve a similar level of completeness.

LaMAR [31] contains 3 large-scale indoor-outdoor scenes. Each scene has many egocentric sequences with abundant causal movement and motion blur. It presents the main challenges of **scalability** and **symmetries** while also exhibiting **low-overlap** and **low-texture** scenarios. The

Table 4. Results on SMERF [10] benchmark established in MP-SFM [28]. Results marked with * are with groundtruth calibration. While both classical and feedforward methods fail, our proposed method is able to reconstruct scenes with low overlap.

	minimal			low			medium			high		
AUC@	1	5	20	1	5	20	1	5	20	1	5	20
SIFT	3.5	4.4	5.4	1.4	1.6	1.8	1.5	2.3	3.1	13.0	19.2	23.2
AL+LG	4.3	6.9	9.8	2.4	6.1	9.8	8.6	19.1	28.0	28.6	46.8	57.0
π^3	3.2	18.0	51.7	1.5	14.3	49.8	1.3	15.7	52.1	0.9	15.5	51.4
π^3 + BA	3.1	18.5	54.1	1.5	14.3	53.4	1.2	14.7	52.2	0.7	12.5	43.4
MASt3R-SfM	3.9	10.4	18.0	4.3	11.7	23.0	5.9	15.8	28.1	10.4	22.9	39.9
MP-SFM* (s)	9.2	41.0	69.8	5.4	29.1	53.0	14.0	47.6	72.9	47.3	79.3	90.6
MP-SFM* (d)	17.2	54.6	77.1	26.6	63.2	84.1	40.4	72.8	87.5	57.1	84.6	94.1
GLUEMAP†	9.8	55.5	82.4	12.9	70.2	92.1	20.3	76.3	93.9	30.9	82.9	95.7
GLUEMAP	10.1	54.9	82.0	14.6	71.4	92.4	27.7	79.1	94.6	47.4	88.1	97.0
GLUEMAP*	10.5	54.8	81.5	14.7	71.8	92.5	28.1	79.5	94.8	47.8	88.3	97.1



(a) SIFT (b) AL + LG (c) GLUEMAP

Figure 4. Qualitative reconstruction resuls of different methods on LaMAR [31]. Our reconstructions are close to the real-world annotation, indicating its high accuracy on this challenging dataset.

view graph radii for CAB, HGE, and LIN are 49, 61, and 59, which underlines the difficulty of the problem, where feedforward methods already exhibit a drastic performance drop even for much smaller radii (*cf*. Figure 3).

Results can be found in Table 5. In this experiment, we only use images captured by phones, as none of the evaluated methods natively support modeling rig constraints. For this dataset, since scenes contain several thousand input views, feedforward methods fail due to out-of-memory issues. Classical methods, especially GLOMAP with ALIKED and LightGlue, perform reasonable on LIN, which is an outdoor-only scene. However, for the HGE and CAB scenes, classical methods struggle to estimate a good reconstructions. In contrast, feedforward models provide good local reconstructions, which serve as a solid foundation for global motion averaging to obtain accurate global reconstructions by our method. The accuracy of the proposed method on these datasets was further improved by BA, highlighting the robustness of the proposed method. From the qualitative results shown in Figure 4, the reconstructions from our pipeline align closely with the real-world annotation, indicating its high accuracy.

Table 5. Results on LaMAR [31]. Our method outperforms classical methods by a large margin while others entirely fail due to out-of-memory (OOM) issues.

	CAB (6587)			HGE (7553)			LIN (9319)			Average		
AUC@	3	10	30	3	10	30	3	10	30	3	10	30
SIFT	0.6	1.4	2.8	2.6	12.4	33.3	4.6	23.7	48.8	2.6	12.4	28.3
AL+LG	1.1	2.9	6.2	8.0	31.7	58.9	23.7	55.7	72.8	10.9	30.1	46.0
MASt3R-SfM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
π^3	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
π^3 + BA	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM	OOM
GLUEMAP†	2.6	21.0	61.2	22.1	70.1	89.3	30.2	69.9	83.9	18.3	53.7	78.1
GLUEMAP	4.5	27.1	64.8	37.3	78.1	92.3	37.3	72.1	84.6	26.4	59.1	80.6

4.4. Limitations & Future Work

Our method's performance critically depends on the quality of local reconstructions by feedforward methods. For example, we can currently not handle fisheye images since the feedforward models are only trained on images taken by pinhole camera models, while later stages in our pipeline can theoretically handle them. As newer feedforward models improve robustness and generality, our method will benefit as well. Furthermore, the formulation currently does handle purely rotational motion due to our augmented bundle adjustment formulation. Future work on incorporating (soft) depth priors from the local reconstructions can improve robustness in these situations. Last but not least, our method currently requires the combination of different feedforward methods. An interesting direction for future work will be in developing a network that can solve the problem with a shared feedforward architecture.

5. Conclusion

In this work, we systematically analyze the strengths and weaknesses of both classical and feedforward approaches to 3D reconstruction. We then present a novel end-to-end reconstruction system that integrates the advantages of both categories: Starting with establishing a tentative view graph and leverages feedforward methods to perform local reconstructions. Then, global motion averaging merges them to initialize an augmented bundle adjustment stage to improve the final accuracy. We extensively evaluate our approach on diverse datasets encompassing a wide range of real-world challenges. By combining the local robustness of feedforward methods with the scalability, accuracy, and global consistency of classical techniques, our system achieves state-of-the-art performance on a wide range of scenarios. We note that further evaluation on unordered image collections with severe appearance changes and transient objects remains an important direction for future work.

Acknowledgement This work was supported under project ID a144 as part of the Swiss AI Initiative, through a grant from the ETH Domain and computational resources provided by the Swiss National Supercomputing Centre (CSCS) under the Alps infrastructure.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016. 2, 3

[2] Federica Arrigoni and Andrea Fusiello. Bearing-based network localizability: A unifying view. *IEEE transactions on pattern analysis and machine intelligence*, 41(9):2049–2069, 2018. 2

[3] Gabriele Berton and Carlo Masone. Megaloc: One retrieval to place them all. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 2861–2867, 2025. 3

[4] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to disambiguate images of similar structures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 34–44, 2023. 2, 3

[5] Avishek Chatterjee and Venu Madhav Govindu. Efficient and robust large-scale rotation averaging. In *Proceedings of the IEEE international conference on computer vision*, pages 521–528, 2013. 2

[6] Zhaopeng Cui and Ping Tan. Global structure-from-motion by similarity averaging. In *Proceedings of the IEEE international conference on computer vision*, pages 864–872, 2015. 2, 5

[7] Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. Sail-recon: Large sfm by augmenting scene regression with localization. *arXiv preprint arXiv:2508.17972*, 2025. 3

[8] 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. *arXiv preprint arXiv:2507.16443*, 2025. 2, 3, 7

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2, 3

[10] Daniel Duckworth, Peter Hedman, Christian Reiser, Peter Zhizhin, Jean-François Thibert, Mario Lučić, Richard Szeliski, and Jonathan T Barron. Smerf: Streamable memory efficient radiance fields for real-time large-scene exploration. *ACM Transactions on Graphics (TOG)*, 43(4):1–13, 2024. 6, 7, 8

[11] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2025. 3

[12] Venu Madhav Govindu. Combining two-view constraints for motion estimation. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, pages II–II. IEEE, 2001. 2

[13] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 2

[14] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 17658–17668, 2024. 3, 4

[15] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 6, 7

[16] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3, 6, 7

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[18] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, pages 71–91. Springer, 2024. 3

[19] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. 3

[20] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local feature matching at light speed. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 17627–17638, 2023. 2, 3

[21] Shaohui Liu, Yidan Gao, Tianyi Zhang, Remi Pautrat, Johannes Lutz Schönberger, Viktor Larsson, and Marc Pollefeys. Robust Incremental Structure-from-Motion with Hybrid Features. In *ECCV*, 2024. 1, 3

[22] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 1, 2, 4

[23] Dominic Maggio, Hyungtae Lim, and Luca Carlone. Vggt-slam: Dense rgb slam optimized on the sl (4) manifold. *arXiv preprint arXiv:2505.12549*, 2025. 3

[24] Daniel Martinec and Tomas Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *2007 IEEE conference on computer vision and pattern recognition*, pages 1–8. IEEE, 2007. 2

[25] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. Openmvg: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016. 1

[26] Onur Ozyesil and Amit Singer. Robust camera location estimation by convex programming. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2674–2683, 2015. 2

[27] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, pages 58–77. Springer, 2024. 1, 2, 6, 7

[28] Zador Pataki, Paul-Edouard Sarlin, Johannes L Schönberger, and Marc Pollefeys. Mp-sfm: Monocular surface priors for robust structure-from-motion. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 21891–21901, 2025. 1, 3, 5, 7, 8

[29] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021. 6, 7

[30] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 3

[31] Paul-Edouard Sarlin, Mihai Dusmanu, Johannes L Schönberger, Pablo Speciale, Lukas Gruber, Viktor Larsson, Ondrej Miksik, and Marc Pollefeys. Lamar: Benchmarking localization and mapping for augmented reality. In *European Conference on Computer Vision*, pages 686–704. Springer, 2022. 6, 7, 8

[32] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Fast image-based localization using direct 2d-to-3d matching. In *2011 International Conference on Computer Vision*, pages 667–674. IEEE, 2011. 1

[33] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2

[34] 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 *Asian Conference on Computer Vision (ACCV)*, 2016. 2

[35] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016. 1

[36] Thomas Schops, Johannes L Schonberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3260–3269, 2017. 6, 7

[37] You Shen, Zhipeng Zhang, Yansong Qu, and Liujuan Cao. Fastvggt: Training-free acceleration of visual geometry transformer. *arXiv preprint arXiv:2509.02560*, 2025. 3

[38] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *ACM siggraph 2006 papers*, pages 835–846. 2006. 1, 2

[39] Chris Sweeney, Torsten Sattler, Tobias Hollerer, Matthew Turk, and Marc Pollefeys. Optimizing the viewing graph for structure-from-motion. In *Proceedings of the IEEE international conference on computer vision*, pages 801–809, 2015. 2

[40] Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. Mv-dust3r+: Single-stage scene reconstruction from sparse views in 2 seconds. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5283–5293, 2025. 3

[41] Shimon Ullman. The interpretation of structure from motion. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, 203(1153):405–426, 1979. 1

[42] Jianyuan Wang, Christian Rupprecht, and David Novotny. Posediffusion: Solving pose estimation via diffusion-aided bundle adjustment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9773–9783, 2023. 3

[43] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024. 3, 4

[44] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 3, 6, 7

[45] Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 10510–10522, 2025. 3, 7

[46] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 3

[47] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. pi3: Scalable permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025. 3, 4, 6, 7

[48] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 1

[49] Yuanbo Xiangli, Ruojin Cai, Hanyu Chen, Jeffrey Byrne, and Noah Snavely. Doppelgangers++: Improved visual disambiguation with geometric 3d features. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27166–27175, 2025. 3, 4, 6

[50] 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 Computer Vision and Pattern Recognition Conference*, pages 21924–21935, 2025. 2, 3

[51] Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating visual relations using loop constraints. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1426–1433. IEEE, 2010. 2

[52] Jason Y Zhang, Amy Lin, Moneish Kumar, Tzu-Hsuan Yang, Deva Ramanan, and Shubham Tulsiani. Cameras

as rays: Pose estimation via ray diffusion. *arXiv preprint arXiv:2402.14817*, 2024. 3

[53] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. 2, 3

[54] Bingbing Zhuang, Loong-Fah Cheong, and Gim Hee Lee. Baseline desensitizing in translation averaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4539–4547, 2018. 2

[55] Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. *arXiv preprint arXiv:2507.11539*, 2025. 3