Light3R-SfM: Towards Feed-forward Structure-from-Motion

Sven Elflein[1,2,3] Qunjie Zhou[1] Laura Leal-Taixé[1]

[1]NVIDIA [2]Vector Institute [3]University of Toronto

Abstract

We present Light3R-SfM, a feed-forward, end-to-end learnable framework for efficient large-scale Structure-from-Motion (SfM) from unconstrained image collections. Unlike existing SfM solutions that rely on costly matching and global optimization to achieve accurate 3D reconstructions, Light3R-SfM addresses this limitation through a novel latent global alignment module. This module replaces traditional global optimization with a learnable attention mechanism, effectively capturing multi-view constraints across images for robust and precise camera pose estimation. Light3R-SfM constructs a sparse scene graph via retrieval-score-guided shortest path tree to dramatically reduce memory usage and computational overhead compared to the naive approach. Extensive experiments demonstrate that Light3R-SfM achieves competitive accuracy while significantly reducing runtime, making it ideal for 3D reconstruction tasks in real-world applications with a runtime constraint.

1. Introduction

Structure-from-Motion (SfM) is the task of jointly recovering camera poses and reconstructing the 3D scene structure from a set of unconstrained images. This long-standing problem is essential to many computer vision applications, including novel view synthesis via NeRFs [3, 26] and 3DGS [17], multi-view stereo (MVS) reconstruction [28, 43], and visual localization [31, 33]. Traditional SfM methods generally follow two main approaches: incremental [34, 37, 48] and global [7, 27, 47] SfM. Both paradigms rely on key components such as feature detection and matching for correspondence search, 3D triangulation to reconstruct geometry from 2D correspondences, and joint optimization of camera poses and scene geometry through bundle adjustment. A major research direction has been to replace these components with learning-based modules, progressing towards fully end-to-end SfM [6, 36, 44]. Recently, the seminal work DUSt3R [45] proposed to train an unconstrained stereo 3D reconstruction model through pointmap regression, *i.e.*, by directly predicting 3D points in a common reference system for every pixel. Learning from large-scale annotated data, it shows impressive performance in handling images with extreme viewpoint changes. To perform SfM from an image collection, DUSt3R works [19, 45] first compute stereo reconstruction exhaustively for all image pairs and then obtain globally aligned pointmaps for all cameras through joint optimization of pairwise rigid transformations and local pointmaps. This baseline has been significantly improved by the concurrent work MASt3R-SfM [10] that leverages image retrieval to drastically reduce the computation overhead, boosts optimization efficiency by optimizing only over the sparse pixel correspondences, and appends a global bundle adjustment stage for accuracy refinement. While optimization-based alignment has been proven to be the key to accurate 3D reconstruction by DUSt3R, MASt3R-SfM and classical SfM methods [22, 27, 34], this comes at the cost of slow runtime and extensive memory footprint even for moderately-sized image collections.

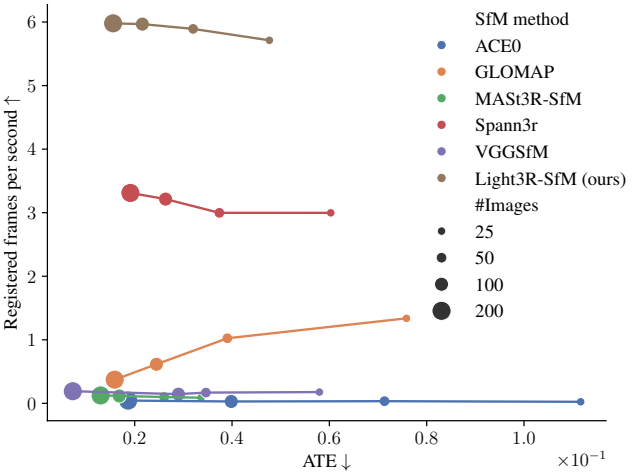

Figure 1. **Processing speed vs. accuracy for various SfM methods.** Our work significantly decreases the runtime across various sizes of image collections compared to traditional pipelines while obtaining comparable accuracy. Results are measured on the Tanks&Temples dataset.

To this end, we propose Light3R-SfM, a fully learnable feed-forward SfM model that directly obtains globally aligned camera poses from an unordered image collection,

without expensive optimization-based global alignment. Instead, we perform implicit global alignment in the latent space with a scalable attention module between the image encoding and 3D decoding stages, which enables global information sharing between features before solving the pairwise 3D reconstruction. This enables exploiting multi-view information across images, which is crucial for learning globally consist pointmaps.

Concurrent work Spann3R [42] tackles online reconstruction from videos by directly regressing pointmaps in a global coordinate system leveraging an explicit memory bank to store information from all previous frames and the current frame. The price paid for being an online model is that the memory bank is constrained by its fixed capacity and prone to drifting due to the propagation of errors over time. In contrast, our work focuses on offline reconstruction from unordered image sets. We exploit multi-view constraints via latent attention while minimizing the redundant processing supported by intelligent graph construction, delivering significantly more accurate camera poses with lower runtime than Spann3R.

We summarize the key contributions of this work as follows: (i) We propose Light3R-SfM, a novel feed-forward SfM approach that replaces classical global optimization with a learnable latent alignment module, leveraging a scalable attention mechanism. (ii) Through extensive experiments, we demonstrate that Light3R-SfM achieves more accurate globally aligned camera poses compared to the concurrent Spann3R method. Its performance rivals state-of-the-art optimization-based SfM techniques while offering significant improvements in efficiency and scalability. Specifically, Light3R-SfM reconstructs a scene of 200 images in just 33 seconds, whereas the comparable MASt3R-SfM takes approximately 27 minutes, resulting in a $>$**49** $\times$ speedup. We highlight the potential of fully feed-forward SfM and aim to inspire future research toward developing more reliable and accurate feed-forward methods for large-scale 3D reconstruction in real-world settings.

2. Related Work

Classical SfM. Conventional structure from motion (SfM) methods can be divided into two main categories: incremental and global SfM. Incremental SfM [1, 14, 34, 37, 48] approaches gradually reconstruct a 3D scene from a collection of images starting from a carefully selected two-view initialization. Its main building blocks involve correspondence searching via feature detection and matching, pairwise pose estimation and 3D triangulation followed by bundle adjustment. Compared to incremental SfM, global SfM methods [2, 7, 8, 27, 47] start with a similar correspondence search and pairwise pose estimation stage, but then jointly align all cameras through rotation and translation averaging followed by 3D triangulation and bundle adjustment. Global methods usually have faster runtime yet being less accurate and robust. A recent work GLOMAP [27] reduces its accuracy gap to incremental methods by combining the estimation of camera positions and 3D structure in a single global positioning step. Our method, similar to the hybrid SfM method [24], divide images into subsets and incremental reconstruct the whole scene by accumulating locally aligned subsets. However, we fundamentally differ from them by estimating camera poses for each local subset in a feed-forward manner through our learned deep network without requiring a further step of global bundle adjustment.

Optimization-based deep SfM. Recently, deep learning has been leveraged to improve core building blocks in SfM pipelines such as sparse feature detection [9, 30] and matching [23, 32]. DFSfM [15] adapts traditional keypoint-based SfM for leveraging dense feature matchers [11, 12, 38, 51]. PixSfM [22] introduces features-metric alignment to keypoints refinement and bundle adjustment, leading to improved accuracy and robustness under challenging conditions. VGGSfM [44] adapts individual SfM components to their learned version forming a fully differentiable SfM framework. Instead of optimizing camera parameters and scene geometries, ACEZero [6] proposes a new learning-based SfM pipeline to incrementally optimizing scene coordinate regression and camera refinement networks that output camera parameters and geometries through reprojection errors. Similarly, FlowMap [36] presents another end-to-end differentiable SfM pipeline where a depth estimation network is optimized per-scene through offline optical flow and point tracking supervisions. Yet, those methods [6, 36] struggle with image pairs with low visual overlapping.

Recently, the emerging unconstrained stereo 3D reconstruction model, DUSt3R [45], opened up a new paradigm for tackling 3D reconstruction tasks such as SfM and multi-view stereo (MVS). Essentially, it proposed a radically novel approach for a two-view reconstruction via direct pointmap regression from a pair of RGB images. Different from monocular scene coordinate regression [4–6, 35], such stereo pointmap regression formulation can benefit from large-scale training achieving strong generalization capability to new scenes. To perform SfM from an image collection, DUSt3R [45] and its improved version MASt3R [19] merge pairwise pointmap predictions via optimization-based global alignment, however, exhaustive pairwise pointmaps in a brute-force manner limits their application to a small set of images [6]. To tackle this, MASt3R-SfM [10] incorporates image retrieval exploiting the MASt3R encoder feature embedding to build a sparse scene graph, leading to significantly reduced runtime. Additionally, it leverages sparse correspondences to boost optimization efficiency and accuracy. Another concurrent work MonST3R [50] extends MASt3R to handle dynamic scenes by leveraging an offline optical flow estimation. However,

those DUSt3R-based SfM methods still rely on expensive iterative optimization for accurate globally aligned poses.

Feed-forward SfM. Instead of performing optimization-based global alignment, Spann3R [42], leverages explicit spatial memory to implicitly align pointmaps w.r.t. the first frame, which requires to maintain spatial information for all following frames in a sequence of arbitrary length overtime. Compared to Spann3R, our proposed feed-forward SfM efficiently exploits multi-view constraints from a collection of unconstrained images at the same time, utilizing a scalable latent alignment module together with efficient graph construction, leading to more efficient and accurate global alignment of pointmaps than Spann3R.

3. Light3R-SfM

In this section, we present Light3R-SfM, a novel feed-forward SfM model that enables robust, accurate and efficient structure-from-motion in the wild for large-scale real-world applications. The key component is an attention mechanism that allows optimization-free globally aligned pose estimation for the entire image set.

Given an unordered image collection or a sequence of images, denoted as $\{\mathcal{I}_i\}_{i=1}^N$ with $\mathcal{I}_i \in \mathbb{R}^{H\times W\times 3}$, our pipeline reconstructs per image camera extrinsics $P \in \mathbb{R}^{4\times 4}$, intrinsics $K_i \in \mathbb{R}^{3\times 3}$ and dense 3D pointmap at image resolution $X \in \mathbb{R}^{H\times W\times 3}$, which represents the globally aligned scene geometry observed by individual images. As shown in Fig. 2, we start with the (i) *encoding*, where an image encoder extract per-image feature tokens. After that we have the (ii) *latent global alignment*, in which information is exchanged between all image tokens via a scalable attention mechanism to globally align image tokens in the feature space (Sec. 3.1). Next, the (iii) *scene graph construction* constructs a scene graph maximizing pairwise image similarities via running the shortest path tree (SPT) algorithm. The (iv) *decoding* step converts image pairs connected by an edge to pointmaps using a stereo reconstruction decoder (Sec. 3.2). Finally, we run the (iiv) *global optimization-free reconstruction*, which accumulates the pairwise pointmaps by traversing the scene graph (Sec. 3.3) to obtain the globally aligned pointmaps.

3.1. Latent Global Alignment

We start by encoding each image $\mathcal{I}_i$ to image tokens

$$F_i^{(0)} = \texttt{Enc}(\mathcal{I}_i), \quad F_i^{(0)} \in \mathbb{R}^{\lfloor H/p\rfloor \times \lfloor W/p\rfloor \times d}, \tag{1}$$

where p is the patch size of the encoder and d is the token dimensionality. To allow information sharing between all images without running into memory constraints, we take inspiration from Karaev et al. [16], who apply a similar principle to point tracks, and factorize the attention operation between all frames via a smaller set of tokens. Specifically, for each set of image tokens $F_i^{(0)}$ we compute its global token $g_i^{(0)} \in R^d$ via averaging along its spatial dimensions. We then use L blocks of our latent global alignment block to achieve global information sharing across all image tokens. For each level $l \in (0, L)$, we first share information across all global image tokens $\{g_i^{(l)}\}_{i=1}^N$ using self-attention defined as

$$\{g_i^{(l+1)}\}_{i=1}^N = \texttt{Self}(\{g_i^{(l)}\}_{i=1}^N). \tag{2}$$

We then propagate the updated global information to dense image tokens $\{F_i^{(l)}\}_{i=1}^N$ for each image independently via cross-attention:

$$F_i^{(l+1)} = \texttt{Cross}(F_i^{(l)}, \{g_i^{(l+1)}\}_{i=1}^N). \tag{3}$$

Finally, we obtain the globally aligned image tokens F_i via a residual connection, *i.e.*, $F_i := F_i^{(0)} + F_i^{(L)}$.

Discussion. A naive implementation through self-attention between all image tokens requiring $\mathcal{O}((N \times T)^2)$, while our latent global alignment module is able to achieve a time complexity of $\mathcal{O}(N^2 + N \times T)$, where $T = \lfloor H/p\rfloor \times \lfloor W/p\rfloor$ is the number of per-image tokens and N is the number of images. While the same asymptotic complexity class, we find reducing the constant factor for practical values of $N \approx T$ to be the key to scale to larger image collections.

3.2. Scene Graph Construction

Despite our global feature attention being lightweight by design, exhaustively decoding 3D pointmaps for all image pairs through a fully connected scene graph still leads to a computational bottleneck. We thus propose a more scalable approach to scene graph construction allowing us to decode pointmaps with just $N - 1$ edges. For that, we leverage the encoder embeddings to compute pairwise similarities similar to concurrent work [19], which allows us to filter out irrelevant image pairs, *e.g.*, pairs with low visual overlap, to avoid unnecessary computation [10, 49]. Specifically, we average pool the tokens of each image F_i to obtain one-dimensional embedding $\bar{F}_i$ and then compute the matrix S containing all pairwise cosine similarities as

$$S_{ij} = \langle \|\bar{F}_i\|_2, \|\bar{F}_j\|_2 \rangle \tag{4}$$

where $\langle \cdot, \cdot \rangle$ denotes the scalar product. Classical SfM methods [27, 34] build a scene graph as a minimum spanning tree (MST) which minimizes the sum of costs, *i.e.*, the negative similarities, of *all* edges. However, this often results in trees with high depth that result in drift when we accumulate pairwise pointmaps, as proposed in the next section.

Therefore, we propose to replace the MST with a shortest path tree (SPT) [13] to obtain a scene graph as a set of edges $E_{\texttt{SPT}} = \{(i, j)\}$ connecting all images, while minimizing the cost of the paths towards each node. Intuitively,

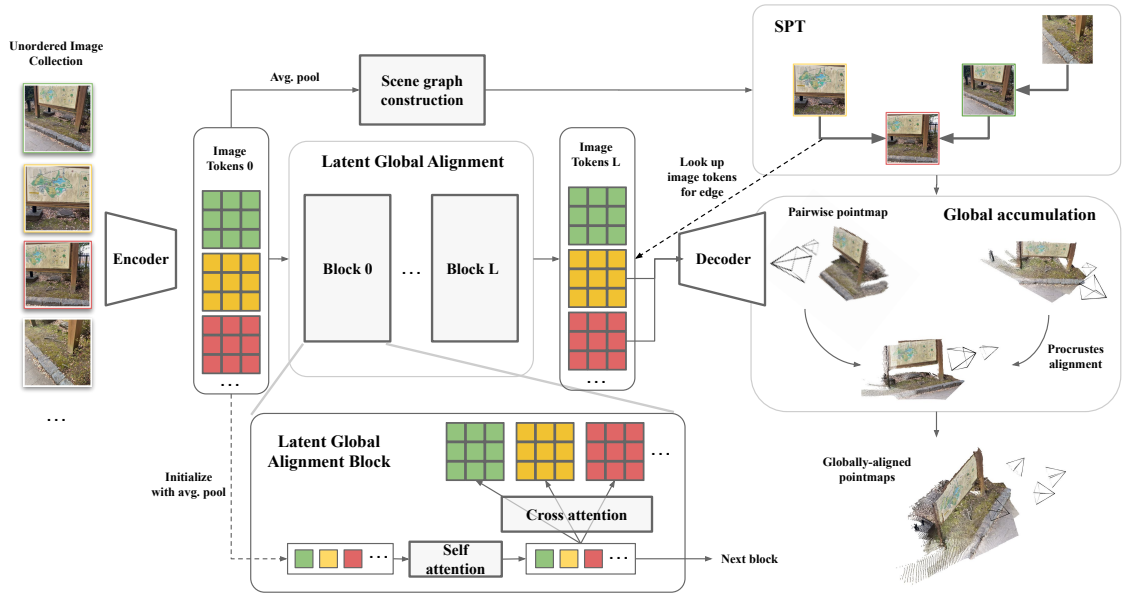

Figure 2. **Light3R-SfM Pipeline.** Given an unordered set of images, we first encode them to obtain image tokens from which we average pool global features for constructing a shortest path tree. We next feed image tokens into our attention-based latent global alignment to enable global context sharing. Afterwards, for each edge in the SPT, we decode pairwise pointmaps using the implicitly aligned feature tokens. Finally, we use global accumulation to obtain globally aligned pointmaps per image.

this leads to a flatter tree which only runs deep when it benefits the overall reconstruction. We set the root node for the SPT as the one with lowest total cost w.r.t. all other nodes, *i.e.*, $\text{argmin}_j \sum_i -S_{ij}$. By design, the number of edges in a tree is linear in the number of images N, *i.e.*, $|E_{\texttt{SPT}}| = N - 1$, leading to significantly better scalability than a fully-connected graph.

3.3. Global Optimization-free Reconstruction

To obtain the global reconstruction while still being end-to-end trainable, we first obtain per-edge local pointmap predictions and then merge local pointmaps into a global one.

Edge-wise pointmap decoding. For every edge in the scene graph $(i, j) \in E_{\texttt{SPT}}$, we run the decoder to output two pointmaps and associated confidence maps defined as:

$$(X^{i,i}, X^{j,i}), (C^{i,i}, C^{j,i}) = \texttt{Dec}\left(F_i, F_j\right). \tag{5}$$

Here, $X^{i,i} \in \mathbb{R}^{H\times W\times 3}$ is the pointmap of the i-th image and $X^{j,i}$ is the pointmap of the j-th image, both in the coordinate frame of the i-th image. $C^{j,i}, C^{i,i} \in \mathbb{R}^{H\times W}$ are the per-point confidence scores for each pointmap respectively. While this closely follows the setup in [45], note that the input features to the decoder are conditioned on all images which facilitates globally aligned pairwise pointmaps.

Global accumulation. To combine the pairwise pointmap predictions into a global reconstruction $\mathbf{X}$ with per-point confidences $\mathbf{C}$, we traverse the SPT $E_{\texttt{SPT}}$ in breadth-first order, starting from the root of the tree. For the first edge, we initialize the global point cloud as $\mathbf{X} = \{X^i, X^j\}$ and $\mathbf{C} = \{C^i, C^j\}$ where $X^i := X^{i,i}$ and $X^j := X^{j,i}$ are the pointmap predictions for the edge in the coordinate system of the i-th image and C^i, C^j their corresponding confidences. The i-th camera is thus implicitly defined as the canonical frame for the global reconstruction. We next register the remaining local reconstructions predicted from the consecutive edges to this initial global reconstruction.

Based on the traversal order, node k of the next edge (k, l) has already obtained its global registered pointmap $X^k \in \mathbf{X}$ in the previous step. We first update the global confidence map of this node to $C^k := C^k \odot C^{k,k}$, where $\odot$ denotes the element-wise geometric mean, to take into consideration the confidence of the pointmap prediction $C^{k,k}$ given the current pair. To register the l-th node to the global reconstruction, we then estimate the optimial rigid body transformation between the two pointmaps, X^k (in the global coordinate) and $X^{k,k}$ (in the same coordinate system of the l-th node) via Procrustes alignment [40]

$$P_k = \texttt{Procrustes}(X^k, X^{k,k}, \log C^k) \tag{6}$$

where $\log C^k \in [0, \infty]^{H\times W}$ serves as a per-point weight. Finally, we transform the pointmap of node l into the global coordinate frame

$$X^{l,} = P_k^{-1} X^{k,l} \tag{7}$$

and add it the global reconstruction $\mathbf{X} := \mathbf{X} \cup \{X^l\}$. Repeat it for all edges in $E_{\texttt{SPT}}$, we obtain per-image globally registered pointmaps X^i with associated confidences C^i .

Discussion. While our method still involves Procrustes alignment for each node, it is a significantly simpler problem compared to jointly optimizing a large number of 3D points and camera parameters among all images, which is much more sensitive to noisy pointmap and confidence predictions and limited to small number of images. Furthermore, compared to iterative solvers used in bundle adjustment, Procrustes alignment can be efficiently solved in closed form, and thus its computation overhead, linear in the number of images, is negligible.

3.4. Supervision

We jointly supervise pairwise local pointmaps and the globally aligned pointmaps, with the focus of the latter to enforce accurate and consistent global alignment learning.

Pairwise supervision. Given a set of ground-truth pointmaps in the world coordinate frame $\bar{\mathbf{X}} = \{\bar{X}^i\}_{i=1}^N$, the corresponding valid pixels $\{\mathcal{D}^i\}_{i=1}^N$, and the ground truth camera poses $P^i{}_{i=1}^N$, we compute $\mathcal{L}_{\texttt{pair}}$ that supervises the pairwise local pointmaps per-edge in the coordinate frame of the first camera following DUSt3R [45]:

$$\mathcal{L}_{\texttt{pair}} = \sum_{(i,j)\in E_{\texttt{SPT}}} (\mathcal{L}_{\texttt{conf}}(P_i\bar{X}^i, X^{i,i}, C^{i,i}, \mathcal{D}^i) + \mathcal{L}_{\texttt{conf}}(P_i\bar{X}^j, X^{j,i}, C^{j,i}, \mathcal{D}^j)), \quad (8)$$

$$\mathcal{L}_{\texttt{conf}}(\bar{X}, X, C, \mathcal{D}) \coloneqq \sum_{p\in\mathcal{D}} C_p \left\|X_p - \bar{X}_p\right\| - \alpha C_p. \quad (9)$$

Here $X, C, \bar{X}$ are the predicted pointmap, confidence map and the ground-truth pointmap, $D \subseteq \{1\dots W\}\times\{1\dots H\}$ defines the valid pixels with ground-truth, and $\alpha > 0$ regularizes the confidences to not be pushed to 0 [41].

Global supervision. We first align the global pointmaps $\mathbf{X} = \{X^{1,}, \dots, X^{N,}\}$ that are defined w.r.t. the root node of the SPT, to the ground truth pointmaps, by estimating the optimal rigid body transformation P_{align} using Procrustes alignment. We then supervise the transformed global pointmap prediction for each image as

$$\mathcal{L}_{\texttt{global}} = \sum_{i\in\{1,\dots,N\}} \mathcal{L}_{\texttt{conf}}(\bar{X}^i, P_{\texttt{align}}X^i, C^i, \mathcal{D}^i) \quad (10)$$

In practice, we do not compute the loss for samples with less than 100 valid pixels due to inaccurately estimated rigid body transformation. This loss implicitly supervises the accuracy of poses extracted from these pointmaps since inaccurate poses from the pairwise Procrustes alignment leads to higher global loss. We optimize $\mathcal{L} = \mathcal{L}_{\texttt{pair}} + \lambda\mathcal{L}_{\texttt{global}}$, empirically setting $\lambda = 0.1$.

4. Experiments

In this section, we conduct extensive evaluation across diverse datasets and scenes, covering a wide range of typical SfM settings, and thorough ablations to understand our model.

We train our model on four datasets: Waymo Open Dataset [39], CO3Dv2 [29], MegaDepth [20], and TartanAir [46]. For training, we sample graphs of $N = 8$ images and random crop images with different aspect ratios and apply color jitter augmentation. We initialize our model encoder and decoder using MASt3R pretrained weights and train for 100,000 iterations with batch size 8 (each batch element corresponds to one graph of images) using AdamW [25] with learning rate 10^{-5}. Like Wang et al. [45], we extract the camera focal length and pose of each view from the pointmap prediction in global reference frame X^i using PnP. We provide more details for training and inference in the supplementary material.

4.1. Scene-level Multi-view Pose Estimation

Dataset. We first evaluate our method on multi-view pose estimation using Tanks&Temples [18] covering 21 indoor and outdoor scenes, where each scene contains 150-1100 images with uncalibrated cameras [6].

Baselines. We categorize our SfM baselines into two main categories, *i.e.*, *optimization-based* (OPT) and *feedforward-based* (FFD) methods, according to their global alignment methodology. For *optimization-based* methods, we consider the classical SfM pipelines Colmap [34] (with SuperPoint [9] and SuperGlue [32]), DF-SfM [15], Glomap [27], PixelSfM [22] and VGGSfM [44], the end-to-end SfM including ACE-Zero [6] and FlowMap [36], as well as the recent state-of-the-art MASt3R-SfM [10]. For *feedforward-based* methods, we compare to our only baseline, the concurrent method Spann3R, where we evaluate both its online and offline version whenever possible.

Metrics. Given a set of images, we follow previous work [10, 27, 45] to compute the relative camera pose errors for all image pairs and measure the percentage of pairs with angular rotation/translation error below a certain threshold τ, denoted as relative rotation accuracy (RRA@τ) and relative translation accuracy (RTA@τ). We report its accuracy score average over all data samples. We further report the percentage of successfully registered images (Reg.) where we count failed scenes with a registration rate of 0, and average translation errors (ATE) where we align estimated camera positions to the ground-truth (estimated using Colmap using all frames provided and provided by [10]) with Procrustes [40] and report an average normalized error. We also report the runtime for a subset of methods on a system with a NVIDIA V100-32GB.

Comparison to state-of-the-art methods. We follow previous work [10] compare across 5 different view settings

	Method	Align.	RRA@5 ↑	RTA@5 ↑	ATE ↓	Reg. ↑	Time [s] ↓
25	COLMAP	OPT	13.7	12.6	0.038	44.4	-
	GLOMAP	OPT	58.4	53.6	0.076	86.1	16.1
	ACE0	OPT	1.2	1.4	0.112	100.0	1042.7
	DF-SfM	OPT	47.5	48.7	0.081	99.4	-
	FlowMap	OPT	0.7	1.5	0.107	100.0	-
	VGGSfM	OPT	55.7	57.4	0.058	96.2	135.7
	MASt3R-SfM	OPT	68.0	70.3	0.034	100.0	283.2
	Spann3r	FFD	19.6	30.7	0.060	100.0	8.3
	Light3R-SfM	FFD	50.9	54.2	0.048	100.0	4.4
50	COLMAP	OPT	28.2	27.4	0.029	60.5	-
	GLOMAP	OPT	69.3	70.3	0.039	97.3	47.5
	ACE0	OPT	11.9	11.5	0.071	100.0	1530.0
	DF-SfM	OPT	63.0	62.7	0.041	100.0	-
	FlowMap	OPT	1.9	3.4	0.073	100.0	-
	VGGSfM	OPT	63.1	64.2	0.035	98.7	291.3
	MASt3R-SfM	OPT	69.1	70.1	0.026	100.0	503.0
	Spann3r	FFD	21.1	31.4	0.037	100.0	16.7
	Light3R-SfM	FFD	52.5	55.2	0.032	100.0	8.5
200	COLMAP	OPT	64.7	57.7	0.019	97.0	-
	GLOMAP	OPT	73.5	74.8	0.016	100.0	536.7
	ACE0	OPT	55.7	57.4	0.019	100.0	4604.4
	DF-SfM	OPT	66.8	69.3	0.016	33.3	-
	FlowMap	OPT	22.2	25.8	0.024	100.0	-
	VGGSfM	OPT	84.5	86.3	0.007	47.6	1511.6
	MASt3R-SfM	OPT	68.2	68.4	0.013	100.0	1609.0
	Spann3r	FFD	22.8	28.6	0.019	100.0	60.4
	Light3R-SfM	FFD	52.4	53.1	0.016	100.0	33.4
full	COLMAP	OPT	GT	GT	GT	GT	-
	GLOMAP	OPT	75.8	76.7	0.010	100.0	1977.7
	ACE0	OPT	56.9	57.9	0.015	100.0	5499.5
	DF-SfM	OPT	69.6	69.3	0.014	76.2	-
	FlowMap	OPT	31.7	35.7	0.017	66.7	-
	VGGSfM	OPT	-	-	-	0.0	2134.2
	MASt3R-SfM	OPT	49.2	54.0	0.011	100.0	2723.1
	Spann3r	FFD	20.3	24.7	0.016	100.0	116.2
	Light3R-SfM	FFD	52.0	52.8	0.011	100.0	63.4

Table 1. **Multi-view pose estimation on Tanks&Temples [18].** We adopt the benchmark by [10] and consider 25/50/200 view subsets and using the full sequence. We report relative pose accuracy RRA@5 and RTA@5, absolute translation error (ATE) and registration rate (Reg.). For clarity, we color-code results with a linear gradient between the worst and best result for a given scene. '-' results indicate that all scenes did not converge or that we did not obtain runtime measurements. We specify the type of alignment used by each methods, 'OPT' stands for *optimization-based* and 'FFD' stands for *feedforward-based*.

including sparsely sampled 25/50/200 frame subsets and the original full sequence. As shown in Tab. 1, our method is competitive with other learning-based methods including VGGSfM, ACE-Zero, and FlowMap. Our method is less accurate than Glomap, Colmap and the concurrent work MASt3R-SfM, particularly for the dense view setting with more than 200 images. Those methods rely on classical optimization techniques such as bundle adjustment [10, 15, 27, 34] or 3D global alignment [10, 45] to achieve better accuracy, but they suffer from limited scalability. For example, Glomap and MASt3R-SfM require 30× and 43× more runtime than our method in full-sequence. The only method that has a runtime in the same magnitude as ours is Spann3R, which also proposes to replace the computationally expensive global alignment of DUSt3R with an implicit alignment implemented via a memory bank.

	Model	Images	RRA@5 ↑	RTA@5 ↑	ATE ↓	Time [s] ↓
25	Spann3R	sorted	19.6	30.7	0.060	8.3
		unordered	10.6	20.1	0.070	9.3
		all pairs	20.5	31.8	0.057	77.7
	Light3R-SfM 224	SPT	29.9	33.0	0.066	**2.2**
	Light3R-SfM	SPT	**50.9**	**54.2**	**0.048**	4.4
50	Spann3R	sorted	21.1	31.4	0.037	16.7
		unordered	12.4	19.0	0.050	18.3
		all pairs	25.8	33.5	0.043	306.0
	Light3R-SfM 224	SPT	34.8	36.3	0.044	**4.3**
	Light3R-SfM	SPT	**52.5**	**55.2**	**0.032**	8.5
full	Spann3R	sorted	20.3	24.7	0.016	116.2
		unordered	12.8	19.8	0.018	125.6
		all pairs		OOM		
	Light3R-SfM 224	SPT	32.9	34.5	0.017	**29.4**
	Light3R-SfM	SPT	**52.0**	**52.8**	**0.011**	63.4

Table 2. **Detailed comparison to Spann3R.** We compare on Tanks&Temples using the 25 and 50 image subsets as well as the full sequences.

To enable Spann3R online evaluation, we sort the multi-view image frames based on their timestamps. Compared to Spann3R, we show that our latent global alignment module is significantly superior for the SfM setting in both accuracy and runtime, leading to an average of 145% and 84% increase in RRA and RTA scores across 5 view settings with approximately half of its runtime.

Detailed comparison to Spann3R. To fully demonstrate the advantage of our proposed approach for feed-forward SfM, we present a more thorough comparison to the other feed-forward method Spann3R. As shown in Tab. 2, Spann3R degrades significantly if the input is an unordered image set due to the lack temporal coherence between frames. To process an unordered image set, Spann3R proposes an offline version that relies on estimating the optimal order of frames via exhaustively evaluating all images pairs. We find that this leads to a significant runtime increase, *e.g.*, ×36 higher runtime than our Light3R-SfM in the 50-view setting, while still being significantly less accurate. When moving to the full video sequences with up to 1100 frames, this setup even encounters out-of-memory errors. Finally, even when operating on the full, sorted video sequence, which Spann3R is specifically designed to handle, performance is still subpar. In comparison, our method is able to provide more accurate poses, even when considering the same image resolution, fully validating the superiority of our method for the SfM setting.

4.2. Object-centric Multi-view Pose Estimation

Dataset. Next, we evaluate our method on object-centric scenes, for which we consider CO3Dv2 [29], containing 37k turntable-like videos of objects from 51 MS-COCO categories and camera poses annotated using Colmap.

Baselines. In addition to the previously mentioned baselines, we further compare to the state-of-the-art multi-view pose regression methods including PoseDiff [43], Pos-

	Method	Global Align.	Co3Dv2↑ RRA@15	RTA@15	mAA@30
10	Colmap [34]	OPT	31.6	27.3	25.3
	Glomap [27]	OPT	45.9	40.3	37.3
	PixSfM [22]	OPT	33.7	32.9	30.1
	VGGSfM [44]	OPT	92.1	88.3	74.0
	DUSt3R-GA [45]	OPT	**96.2**	86.8	76.7
	MASt3R-SfM [10]	OPT	96.0	**93.1**	**88.0**
	PoseDiff [43]	FFD	80.5	79.8	66.5
	PosReg [43]	FFD	53.2	49.1	45.0
	RelPose++ [21]	FFD	82.3	77.2	65.1
	Spann3R [42]	FFD	89.5	83.2	70.3
	MASt3R * [19]	FFD	94.5	80.9	68.7
	Light3R-SfM	FFD	94.7	85.8	72.8
2	DUSt3R [45]	FFD	94.3	88.4	77.2
	MASt3R [19]	FFD	94.6	91.9	**81.8**
	Spann3R [42]	FFD	91.9	89.9	77.6
	Light3R-SfM	FFD	**95.5**	**93.2**	81.6

Table 3. **Wide-baseline, multi-view camera pose estimation on CO3Dv2 [29].** We vary the number of input images by randomly sampling from the original sequence.

Reg [43], and RelPose++ [21], which also predict aligned camera poses in a feed-forward manner. We also compare to a MASt3R * baseline where we employ the off-the-shelf MASt3R model without any additional training inside our proposed framework. **Metrics.** We report relative rotation accuracy and relative translation accuracy with threshold 15, *e.g.*, RRA@15 and RTA@15. We further calculate the mean Average Accuracy (mAA)@30, defined as the area under the curve accuracy of the angular differences at $\min$(RRA@30, RTA@30). **Results.** Similar to [10], we evaluate on 2-view and 10-view settings. As shown in the *upper* part of Tab. 3, for the 10-view setting, we are significantly more accurate than traditional *optimization-based* SfM methods such as Colmap, Glomap and PixSfM. Our performance is on-par with VGGSfM which trains one model per dataset for evaluation, while our method generalizes across datasets. We are less accurate than DUSt3R and concurrent MASt3R-SfM that benefit from *optimization-based* global alignment at the cost of scalability and efficiency as discussed in Sec. 4.1, however, we note that our latent global alignment almost closes the gap towards DUSt3R-GA. Compared to other *feed-forward* methods, our method achieves the best performance across all metrics. Comparing to MASt3R *, we demonstrate the benefit coming from our contributions leading to a 6.1% increase in RTA@15. Among the feed-forward methods, only ours and Spann3R evaluate on a large number of images, while the other works typically focus on object-centric scenes. In the *bottom* part of Tab. 3, we demonstrate superior performance compared to other *feed-forward* methods on pairwise pose estimation. This setting decouples the need for global alignment, indicating that our proposed global supervision on the accumulated pairwise predictions helps to predict more accurate pointmaps for pose estimation in general.

Method	Waymo [39] Val. Split RRA@5↑	RTA@5↑	ATE↓	Runtime(s)↓
MASt3R-SfM [19]	75.7	**63.7**	**0.005**	1662.0
Spann3R [42]	55.1	14.5	0.025	53.8
Light3R-SfM	**78.3**	57.7	0.019	**8.5**

Table 4. **Camera pose estimation on driving scenes.**

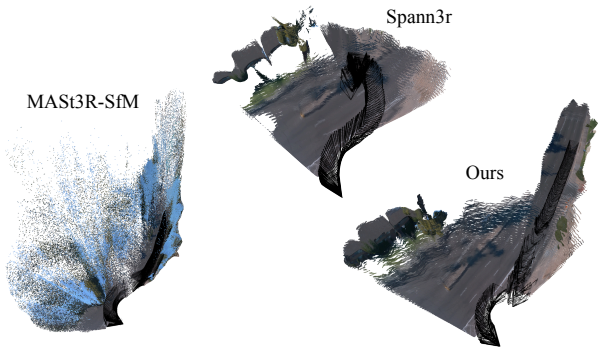

Figure 3. **Qualitative comparison on a Waymo scene.** Note how the MASt3R-SfM reconstruction does not truthfully reconstruct the 90° turn, while Spann3R predictions degrade after tens of frames.

4.3. Evaluation on Driving Scenes

Dataset. We further evaluate our model on driving scenes where we validate our generalization capability to translating camera motion. We use the validation split of Waymo Open Dataset [39], which contains a collection of 200 20-second clips recorded at 10Hz from an autonomous vehicle. For each sequence, we use the 200 input frames from the forward-looking camera.

Baselines. We compare against our concurrent works Spann3R and MASt3R-SfM, which also build on top of DUSt3R and have also seen Waymo training split as us.

Results. As shown in Tab. 4, Light3R-SfM achieves comparable accuracy to *optimization-based* MASt3R-SfM at $\sim 195\times$ lower runtime. Compared to the concurrent *feed-forward* Spann3R, we significantly outperform Spann3R achieving $\sim 4\times$ better accuracy in RTA@5 at $>6\times$ lower runtime. This confirms that our latent alignment design leads to a more accurate and efficient modeling of global alignment compared to a memory-based architecture in Spann3R. We visualize pointmaps predicted by Light3R-SfM and our baselines in Fig. 3.

4.4. Ablation Studies

In this section, we perform in-depth ablation study to fully understand each component of our method. We perform all ablation experiments on Tanks&Temple [18] dataset using its 200 view subset and report the same multi-view pose estimation metrics as defined in Sec. 4.1.

Model components. In the part **(a)** of Tab. 5, we ablate

	Backbone Init.	Global Sup.	Latent Align.	Graph Const.	Tanks&Temple [18] - 200 Views RRA@5↑	RTA@5↑	ATE↓
(a)	MASt3R	✗	✗	SPT	47.5	48.3	0.019
	MASt3R	✗	✓	SPT	50.8	48.7	0.016
	DUSt3R	✓	✓	SPT	48.8	48.8	0.016
(b)	MASt3R	✓	✓	*Oracle*	*52.8*	*53.8*	*0.016*
	MASt3R	✓	✓	MST	44.4	39.5	0.017
	MASt3R	✓	✓	SPT	**52.4**	**53.1**	**0.016**

Table 5. **Model ablation.** We study the impact of backbone initialization, global supervision, latent alignment as well as different ways for graph construction on pose estimation performance.

	Conf. thr.	Reg. (↑)	RRA@5 (↑)	RTA@5 (↑)	RRA@15 (↑)	RTA@15 (↑)
MASt3R-SfM	N/A	**100.0**	**68.0**	**70.3**	73.8	77.3
Light3R-SfM	3	84.8	56.5	58.9	77.6	76.4
	5	83.2	63.0	62.2	80.0	78.7
	7	75.8	65.2	63.7	**81.3**	**80.2**

Table 6. **Pointmap confidence analysis on Tanks&Temples [18].** Our learned confidence maps effectively filter *outlier* points, leading to increased rejected frames yet overall more accurate poses.

the influence of backbone initialization, global supervision defined in Eq. (10) and latent alignment defined in Sec. 3.1 when using the same graph construction process and pose accumulation process described in Sec. 3.3. We show that adding latent alignment leads to 6.95% increase in RRA@5 and 15.78% decrease in ATE over the baseline. By further adding global supervision, we obtain our final model (*last row*), which brings another 3.14% and 9.03% increase in RRA@5 and RTA@5. Switching the backbone initialization of our full model from MASt3R to DUSt3R leads to 6.87% and 8.1% drop in RRA@5 and RTA@5.

Scene graph construction. In part **(b)** of Tab. 5, we analyze the impact of different options in building the scene graph introduced in Sec. 3.2. Specifically, we present an graph construction Oracle, where we use ground truth overlapping score to obtain an optimal spanning tree, indicating the upper bound performance our method can achieve by improving the retrieval step. We show that our encoder retrieval is able to construct a high-quality scene graph that leads to performance very close to Oracle. We further compare to another baseline where we obtain a minimal spanning tree instead. This leads to 15.26% and 25.61% accuracy decrease in RRA@5 and RTA@5, demonstrating the importance modulating the depth of the tree for our method.

Pointmap confidence analysis. During inference, we compute global camera poses for each frame from the predicted pointmaps where we filter out 3D points if their confidence score is lower than a certain confidence threshold. In Tab. 6, we study the impact of different confident thresholds, *e.g.*, from 3 to 7, on pose estimation performance. We show that the learned confidence maps effectively identify the confident points from images, leading to decreasing registration rate and improved pose accuracy when we increase the threshold, which enables a flexible control over trading-off accurateness and completeness depending on the downstream applications.

Image Resol.	Image Encoding	Latent. Alignment	Graph Const.	Pointmap Decoding	Global Accum.	Total Runtime(s)	Max. GPU VRAM (GB)
224	3.6	3.4	0.1	23.2	0.9	52.3	8.0
512	12.3	7.5	1.4	68.4	1.0	135.8	25.6

Table 7. **Runtime Analysis.** We evaluate Light3R-SfM on the Courthouse scene with 1106 images using a NVIDIA V100-32GB.

Generalization. We showed that our method trained on a specific 8-view graph structure generalizes well to all kinds of view settings, *e.g.*, (minimal) 2-view, (sparse) 10/25/50/200-view and full-view settings in Tab. 1. In addition, we confirm solid generalization of our method by consistently outperforming our concurrent baseline Spann3R, across different types scenes and datasets including object-centric scenes from Co3Dv2 [29], driving scenes from Waymo [39] and *unseen* natural indoor and outdoor scenes from Tanks and Temples [18].

Runtime analysis. We analyze the detailed runtime required by individual components and the memory footprint of our method in Tab. 7 by evaluating it on the Courthouse scene at two image resolutions using a NVIDIA V100-32GB. We split the batch into chunks of 32 for both the encoder and decoder which can be adapted to fit smaller or larger GPU memory budgets, trading off runtime.

5. Conclusion

We presented Light3R-SfM, a novel pipeline to perform SfM without traditional components such as matching or global optimization. For this, we build upon 3D foundation models operating on image pairs and scale these to large image collections via a scalable global latent alignment module, effectively aligning pairwise predictions in latent space, replacing global optimization. Further, we leverage a sparse scene graph keeping memory requirements low. We show that such an approach allows to significantly reduce runtime while providing competitive accuracy, opening up exciting new research opportunities towards data-driven approaches for a field that is traditionally dominated by optimization-based methods.

Limitations. We acknowledge that our current model does not scale to all SfM settings, for example for collections of tens of thousands of images. Furthermore, the accuracy of poses at tight thresholds still lacks behind SOTA optimization-based methods, most likely due to the low image resolution processed by learned methods. We also observed drift in long-running camera trajectories due to the compositional nature of the reconstruction. Finally, when extreme visual similiarity is present the retrieval-based graph connectivity might contain errors that degrade the reconstruction.

6. Acknowledgments.

We thank Sérgio Agostinho for all the helpful discussions and comments during the conception of this work. We also thank the authors of [42] and [43] for providing details on the evaluation protocol as well as the authors of [10] for providing their Tanks&Temples evaluation dataset split.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *Communications of the ACM*, 54 (10):105–112, 2011. 2

[2] Mica Arie-Nachimson, Shahar Z Kovalsky, Ira Kemelmacher-Shlizerman, Amit Singer, and Ronen Basri. Global motion estimation from point matches. In *2012 Second international conference on 3D imaging, modeling, processing, visualization & transmission*, pages 81–88. IEEE, 2012. 2

[3] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021. 1

[4] Eric Brachmann, Alexander Krull, Sebastian Nowozin, Jamie Shotton, Frank Michel, Stefan Gumhold, and Carsten Rother. Dsac-differentiable ransac for camera localization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6684–6692, 2017. 2

[5] Eric Brachmann, Tommaso Cavallari, and Victor Adrian Prisacariu. Accelerated coordinate encoding: Learning to relocalize in minutes using rgb and poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5044–5053, 2023.

[6] Eric Brachmann, Jamie Wynn, Shuai Chen, Tommaso Cavallari, Áron Monszpart, Daniyar Turmukhambetov, and Victor Adrian Prisacariu. Scene coordinate reconstruction: Posing of image collections via incremental learning of a relocalizer. In *European Conference on Computer Vision*, 2024. 1, 2, 5

[7] Qi Cai, Lilian Zhang, Yuanxin Wu, Wenxian Yu, and Dewen Hu. A pose-only solution to visual reconstruction and navigation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):73–86, 2021. 1, 2

[8] 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

[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, 5

[10] Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. *arXiv preprint arXiv:2409.19152*, 2024. 1, 2, 3, 5, 6, 7, 9

[11] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17765–17775, 2023. 2

[12] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19800, 2024. 2

[13] L. R. Ford. *Network Flow Theory.* RAND Corporation, Santa Monica, CA, 1956. 3

[14] Jan-Michael Frahm, Pierre Fite-Georgel, David Gallup, Tim Johnson, Rahul Raguram, Changchang Wu, Yi-Hung Jen, Enrique Dunn, Brian Clipp, Svetlana Lazebnik, et al. Building rome on a cloudless day. In *European Conference on Computer Vision*, pages 368–381. Springer, 2010. 2

[15] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21594–21603, 2024. 2, 5, 6

[16] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker: It is Better to Track Together. In *European Conference on Computer Vision (ECCV)*, 2024. 3

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[18] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 5, 6, 7, 8

[19] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *European Conference on Computer Vision*, 2024. 1, 2, 3, 7

[20] Zhengqi Li and Noah Snavely. MegaDepth: Learning Single-View Depth Prediction From Internet Photos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2041–2050, 2018. 5

[21] Amy Lin, Jason Y Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose++: Recovering 6d poses from sparse-view observations. *arXiv preprint arXiv:2305.04926*, 2023. 7

[22] 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. 1, 2, 5, 7

[23] 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

[24] Zhendong Liu, Wenhu Qv, Haolin Cai, Hongliang Guan, and Shuaizhe Zhang. An efficient and robust hybrid sfm method for large-scale scenes. *Remote Sensing*, 15(3):769, 2023. 2

[25] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization, 2019. 5

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1

[27] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision (ECCV)*, 2024. 1, 2, 3, 5, 6, 7

[28] Rui Peng, Rongjie Wang, Zhenyu Wang, Yawen Lai, and Ronggang Wang. Rethinking depth estimation for multi-view stereo: A unified representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8645–8654, 2022. 1

[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. 5, 6, 7, 8

[30] Jerome Revaud, Cesar De Souza, Martin Humenberger, and Philippe Weinzaepfel. R2d2: Reliable and repeatable detector and descriptor. *Advances in neural information processing systems*, 32, 2019. 2

[31] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12716–12725, 2019. 1

[32] 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. 2, 5

[33] Torsten Sattler, Bastian Leibe, and Leif Kobbelt. Improving image-based localization by active correspondence search. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12*, pages 752–765. Springer, 2012. 1

[34] 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, 3, 5, 6, 7

[35] Jamie Shotton, Ben Glocker, Christopher Zach, Shahram Izadi, Antonio Criminisi, and Andrew Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2930–2937, 2013. 2

[36] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. Flowmap: High-quality camera poses, intrinsics, and depth via gradient descent. *arXiv preprint arXiv:2404.15259*, 2024. 1, 2, 5

[37] 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

[38] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8922–8931, 2021. 2

[39] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 5, 7, 8

[40] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (4):376–380, 1991. 4, 5

[41] Sheng Wan, Tung-Yu Wu, Wing H. Wong, and Chen-Yi Lee. Confnet: Predict with Confidence. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2921–2925, 2018. 5

[42] Hengyi Wang and Lourdes Agapito. 3d reconstruction with spatial memory. *arXiv preprint arXiv:2408.16061*, 2024. 2, 3, 7, 9

[43] 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. 1, 6, 7, 9

[44] 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. 1, 2, 5, 7

[45] 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. 1, 2, 4, 5, 6, 7

[46] Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A Dataset to Push the Limits of Visual SLAM. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4909–4916, 2020. 5

[47] Kyle Wilson and Noah Snavely. Robust global translations with 1dsfm. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part III 13*, pages 61–75. Springer, 2014. 1, 2

[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, 2

[49] Shen Yan, Maojun Zhang, Shiming Lai, Yu Liu, and Yang Peng. Image retrieval for structure-from-motion via graph

convolutional network. *Information Sciences*, 573:20–36, 2021. 3

[50] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arXiv:2410.03825*, 2024. 2

[51] Qunjie Zhou, Torsten Sattler, and Laura Leal-Taixe. Patch2pix: Epipolar-guided pixel-level correspondences. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4669–4678, 2021. 2