MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion

Zador Pataki[1] Paul-Edouard Sarlin[2] Johannes L. Schönberger[1,3] Marc Pollefeys[1,3]
[1] ETH Zurich [2] Google [3] Microsoft Spatial AI Lab

Abstract

While Structure-from-Motion (SfM) has seen much progress over the years, state-of-the-art systems are prone to failure when facing extreme viewpoint changes in low-overlap, low-parallax or high-symmetry scenarios. Because capturing images that avoid these pitfalls is challenging, this severely limits the wider use of SfM, especially by non-expert users. We overcome these limitations by augmenting the classical SfM paradigm with monocular depth and normal priors inferred by deep neural networks. Thanks to a tight integration of monocular and multi-view constraints, our approach significantly outperforms existing ones under extreme viewpoint changes, while maintaining strong performance in standard conditions. We also show that monocular priors can help reject faulty associations due to symmetries, which is a long-standing problem for SfM. This makes our approach the first capable of reliably reconstructing challenging indoor environments from few images. Through principled uncertainty propagation, it is robust to errors in the priors, can handle priors inferred by different models with little tuning, and will thus easily benefit from future progress in monocular depth and normal estimation. Our code is publicly available at `github.com/cvg/mpsfm`.

1. Introduction

Structure-from-Motion (SfM) is a prevalent problem in computer vision involving the estimation of 3D structure and camera motion from a collection of 2D images. The tremendous progress in the field has culminated in a variety of state-of-the-art SfM pipelines (*e.g.*, Bundler [57], VisualSfM [68], COLMAP [50], GLOMAP [43]). Today, these systems are successfully applied in a wide range of scenarios with high relevance for tasks such as (simultaneous) localization and mapping [46], multi-view stereo [51], or novel-view synthesis [32, 40]. Despite this feat, many challenges and failure cases remain to be solved, including but not limited to extreme viewpoint and illumination changes [13, 17, 45], repetitive structure [9, 72], scalability to large scenes [1, 21, 26], or privacy concerns [22, 58].

One of the most frequent failure cases for SfM is the scenario of extreme viewpoint changes. These can manifest as either extremely wide-baseline and low-overlap or as low-parallax image pairs. As such, these scenarios present challenges to different stages in the reconstruction process. Many recent efforts in the community have centered around solving the matching problem using the immense progress in machine learning [13, 17, 45, 65], and we are now able to match across extreme viewpoint and illumination conditions or even detect symmetry issues from a single pair of images [9]. However, when we feed the established matches into the subsequent reconstruction algorithms, we still face multiple fundamental limitations leading to unstable and inaccurate results or even outright reconstruction failures.

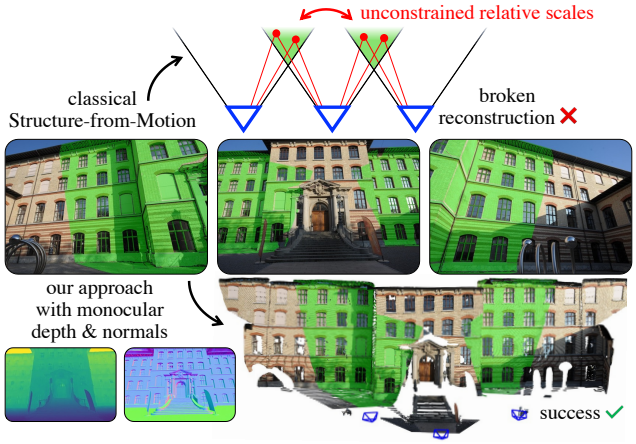


Figure 1. **A typical failure case for SfM.** Existing approaches cannot handle low-overlap image pairs because they require three-view tracks to ensure a consistent scale across the scene. We bridge this limitation by augmenting SfM with monocular depth and normal priors from off-the-shelf deep networks. This makes SfM significantly more robust for data captured by non-expert users.

In particular, the current state-of-the-art systems [16, 43, 50, 64] inherently require three-view overlap with sufficient baseline and parallax or otherwise cannot perform multi-view consistent 3D reconstruction. In practice, this turns out as one of the main difficulties for the non-expert user and oftentimes also for experts for a variety of reasons. First, it is intrinsically hard to capture large and complex scenes while ensuring sufficient viewpoint overlap and variation. Even for small scenes and structured setups, it is far from trivial with

the many constraints to satisfy and it often requires careful prior planning or repeated trials to capture a scene with the desired completeness and accuracy. The naive approach of capturing overly redundant viewpoints is typically also not a solution, as it stands in opposition to other important considerations like capture and processing time or storage and compute costs. Furthermore, the general purpose reconstruction systems frequently show diminishing returns when fed with too many redundant views.

In this paper, we make an important step towards overcoming several of the remaining limitations by leveraging the recent advances in monocular depth estimation. In particular, we integrate monocular depth and normal cues into the classical incremental SfM paradigm to lift the requirement for three-view tracks. Our proposed pipeline is able to perform accurate multi-view 3D reconstruction from two-view tracks only and thus works in extremely challenging low-overlap scenarios, while retaining state-of-the-art performance in higher-overlap conditions (Fig. 1). As a consequence, our system can also directly take advantage of dense pairwise matches [19, 59, 65]. Compared to the classical approach of only using sparse feature matches, we thus achieve higher reconstruction completeness in scenes with little texture. In addition, the use of single-view depth priors for regularization of scene geometry leads to significant reliability improvements under low-parallax conditions. We further introduce a dense depth consistency check to identify incorrect posing of images, especially to prevent symmetry issues.

By tight integration of single- and multi-view optimization and principled uncertainty propagation, we handle large errors in the monocular priors, which in turn enables us to build upon off-the-shelf deep models. Future progress in monocular depth estimation will benefit our approach with little to no tuning required. To maintain the scalability characteristics of classical SfM, we formulate the global reconstruction objective as an alternating optimization of single- and multi-view sub-problems. In extensive experiments on challenging datasets, we show significant improvements in terms of accuracy and completeness as compared to state-of-the-art classical and recent learned SfM pipelines.

2. Related work

Traditional SfM: In the early days, the field of computer vision largely focused on SfM from ordered video sequences [5, 44, 60] while later works shifted to unordered image collections [1, 21, 26, 49, 50, 57, 68]. The literature has traditionally categorized methods into the incremental and global paradigms. Over the years, several software packages have become available [42, 43, 50, 57, 68] with COLMAP as the, arguably, most widely adopted SfM pipeline.

While achieving reliable results for a wide range of inputs, each pipeline exhibits its own unique weaknesses and failure modes. Common to all approaches, however, is the fundamental requirement for three-view overlap and tracks, which we address specifically in this work. Prior works [10, 30] have identified this issue as well and proposed methods for camera pose estimation from both hybrid 2D-3D and 2D-2D correspondences. Sinha *et al*. [55] presented an approach for pure two-view SfM from silhouettes. However, it relied on the extraction of silhouettes in outside-in capture scenarios, while we intend to solve more general scenarios. Furthermore, it requires known epipolar geometry to at least two previously registered views and thus still needs three-view overlap. Zheng and Wu [73] also tackled the issue and proposed an approach for structure-less resectioning from 2D-2D correspondences. While this does away with needing three-view tracks, equivalent to Sinha *et al*. [55], it still requires three-view overlap (*i.e*., three images are all pairwise matchable but no two-view matches form three-view tracks) to constrain the scale. It also suffers from common degenerate viewpoint configurations, such as sideward motion. In contrast, we require neither three-view tracks nor three-view overlap and also do not suffer from the same degeneracies due to integration of strong monocular priors.

Learning for SfM: Driven by the overwhelming success of machine learning, many works have been devised to integrate data-driven methods into individual components of the traditional pipeline with a focus on addressing the challenges in the feature representation [13, 17, 41, 52] and matching [18, 36, 45, 59, 65, 65] stages. Relatively fewer works tackled other components like RANSAC [7, 66], bundle adjustment [35, 67], or camera calibration [62]. Notably, the recent works of DuSt3R [65] and MASt3R [34] demonstrate impressive two-view matching and reconstruction results. Based on this feat, the MASt3R-SfM [16] pipeline performs multi-view reconstruction using a paradigm akin to traditional global SfM. Furthermore, several methods take advantage of an end-to-end learning objective to jointly train multiple SfM components [8, 56, 64]. Despite achieving state-of-the-art results in specific scenarios, these pipelines still do not serve as universal replacement for traditional SfM in more general, unstructured, or large-scale settings [8, 56, 64]. In contrast, we integrate depth priors into traditional incremental SfM to solve its failure modes while retaining the generality and scalability of the classical approaches.

Monocular depth priors: Since the early works on monocular depth estimation [20, 27, 48], astounding progress has been made. The latest models [4, 6, 31, 70, 71] are able to estimate depths and normals with zero-shot generalization on in-the-wild images. Typically, the resulting depth maps are visually pleasing with a sharp delineation of occlusion boundaries and good relative depth accuracy. When it comes to directly using the depth estimates for metric, multi-view 3D reconstruction, the absolute accuracy is typically insuffi-

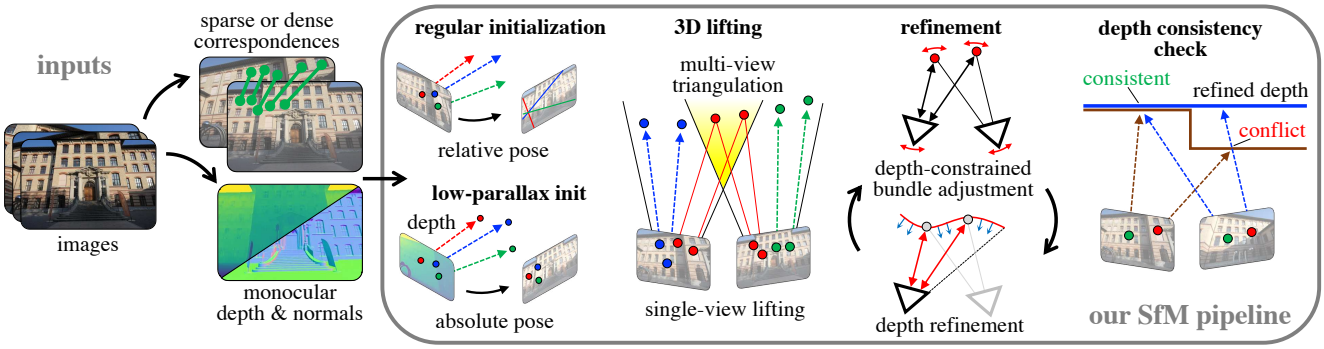


Figure 2. **Overview of our approach.** Given image correspondences, depth, and surface normals, we first initialize the reconstruction by estimating a relative pose or, if the parallax is low, an absolute pose from points lifted to 3D by depth. While SfM can generally estimate 3D only for points observed in multiple views, we leverage single-view observations with depth. This helps registering images with lower visual overlap. Camera poses, 3D points, and depth maps are refined by alternating between bundle adjustment and normal integration with depth constraints. Finally, we reject incorrect registrations, *e.g.*, due to symmetries, by checking that the depth is consistent across views.

cient and inconsistent across posed overlapping views.

In our work, we build on the tremendous progress to improve upon the limitations of SfM. Our method leverages monocular depth to reconstruct previously unsolved two-view overlap scenarios as well as to improve the overall robustness of the system. Through a carefully designed optimization strategy and principled uncertainty propagation, we can robustly deal with noisy monocular depth input and jointly refine them with the multi-view triangulated structure. We use monocular surface normals, which are easier to predict than depth and often more accurate, by integrating them into depth following Cao *et al*. [11]. COMO [14] also leverages normal integration and monocular depth for real-time monocular mapping and odometry with a tightly coupled system, similar to ours. However, their approach is designed for real-time operation and is thus limited to GPU and to a few images. Furthermore, it is not amenable to the more general problem of SfM from unstructured images.

Closer to our approach, StudioSfM [37] leverages monocular depth in SfM but only for the scenario of low-parallax videos by modifying COLMAP's initialization and bundle adjustment stages. Ours is more general and tackles multiple of the limitations of incremental SfM, including low-parallax, to handle unstructured image collections.

3. Method

We first formulate the problem and provide an overview of our system – see also Fig. 2.

Inputs: Our system takes as input a set of $i = 1...n$ unordered images $\mathcal{I}=\{I_i \in \mathbb{R}^{H_i \times W_i}\}$ and their intrinsic pinhole camera parameters $\mathcal{K}=\{K_i \in \mathbb{R}^{3\times3}\}$. For each of them, we estimate monocular depth maps $\mathcal{D}=\{D_i \in \mathbb{R}_0^{+H_i \times W_i}\}$ and normal maps $\mathcal{N}=\{N_i \in \mathcal{S}^{2^{H_i \times W_i}}\}$ with their respective confidence maps $\Sigma_{D_i}, \Sigma_{N_i}$.

Following standard practice, we extract sparse local image features $\mathcal{F}=\{\mathcal{F}_i\}$ for each image, where features $\mathcal{F}_i$ in image i are defined by their pixel keypoint positions $\{x_j \in [0, W_i]\times[0, H_i] \mid j = 1...f_i\}$ with assumed Gaussian noise Σ_{x_j} and descriptors $\{\boldsymbol{d}_j \in \mathbb{R}^d\}$. Feature matching and geometric verification produce correspondences $\mathcal{M}=\{\{\mathcal{M}_{a,b}\} \subset [1...f_a]\times[1...f_b]\}$ with associated scores $\{q_{a,b} \in \mathbb{R}^{|\mathcal{M}_{a,b}|}\}$, between sparse features for each image pair, which later form feature tracks across multiple images. We optionally compute dense pixel-wise matches, sub-sample them using either sparse keypoints or non-maximum suppression, and apply geometric verification, resulting in two-view correspondences $\mathcal{M}^*=\{\{\mathcal{M}^*_{a,b}\} \subset [1...H_aW_a]\times[1...H_bW_b]\}$ with scores $\{q^*_{a,b} \in \mathbb{R}^{|\mathcal{M}^*_{a,b}|}\}$.

Outputs: Given the intrinsics $\mathcal{K}$, depth maps $\mathcal{D}$, normal maps $\mathcal{N}$, sparse features $\mathcal{F}$ as well as correspondences $\mathcal{M}$ and $\mathcal{M}^*$, our incremental reconstruction algorithm estimates the camera poses $\mathcal{P}=\{P_i \in \mathrm{SE}(3) \mid i \in \mathcal{R}\}$ for a subset $\mathcal{R}$ of images that could be confidently registered. It also estimates $k = 1...m$ scene points $\mathcal{X}=\{X_k \in \mathbb{R}^3\}$ and refined, globally consistent depth maps $\mathcal{D}^*$.

System overview: Our system builds upon the COLMAP incremental SfM framework [50]. COLMAP only takes prior calibrations $\mathcal{K}$, sparse features $\mathcal{F}$, and correspondences $\mathcal{M}$ to estimate the output camera poses $\mathcal{P}$ and scene points $\mathcal{X}$. To leverage the monocular priors $\mathcal{D}, \mathcal{N}$ and the dense two-view correspondences $\mathcal{M}^*$, we make significant changes to many of the underlying algorithms as well as the control logic in the pipeline. The correspondence graph is built from either sparse or dense correspondences. Then, similar to COLMAP, we start with an initial image pair and incrementally register more images, interleaved with local and global refinements.

3.1. Two-View Initialization

Initial pose: Following COLMAP, we rank image pairs by number of inlier correspondences and select the first pair

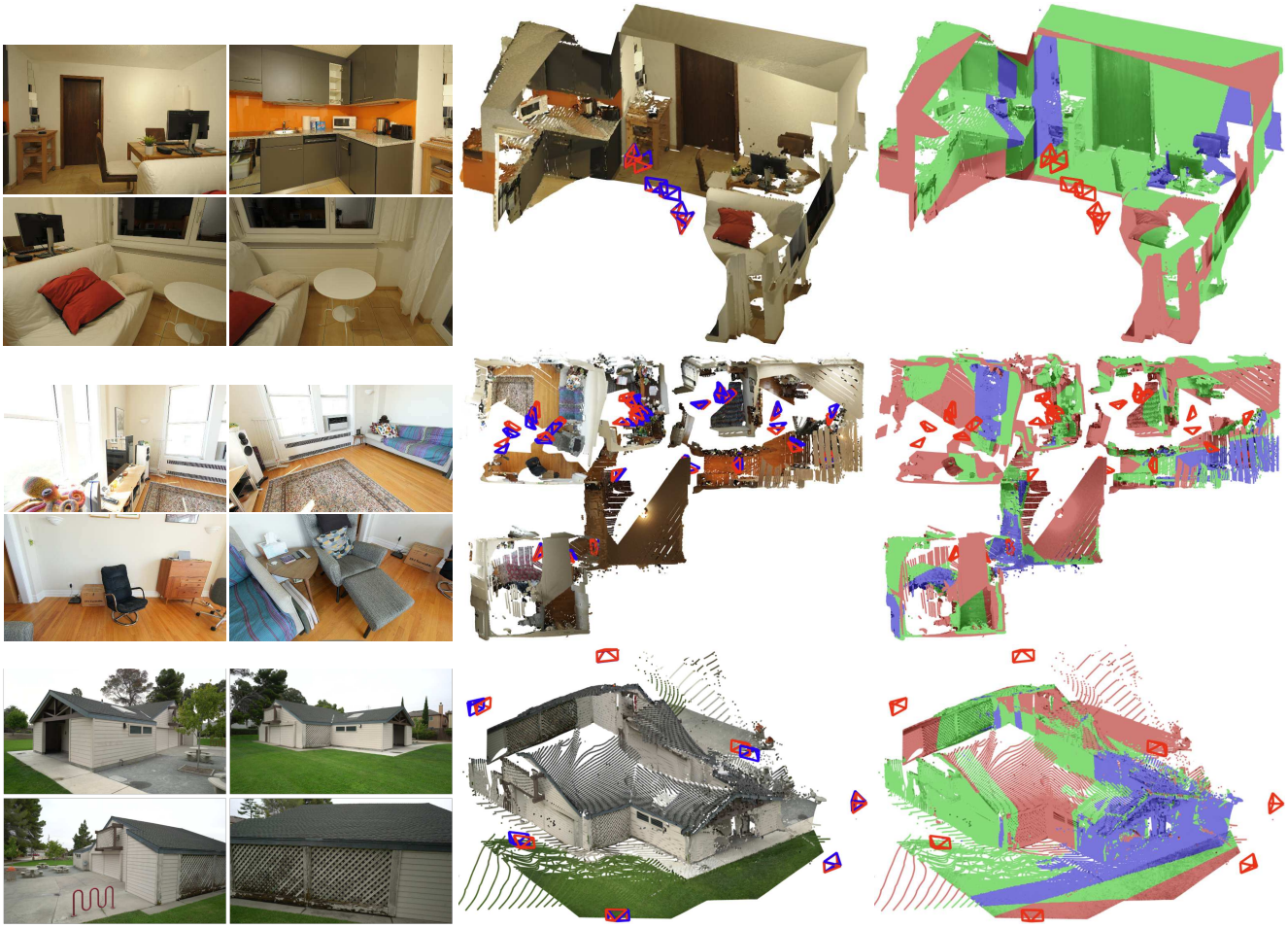


Figure 3. **Qualitative results for low overlap scenes.** Left: Input images with low overlap. Center: Estimated (red) and ground-truth (blue) camera poses with the monocular depth refined by our system. Right: Lifted refined depth of which points are colored differently whether they are visible in a single image (red), two (green), or at least three images.

(I_a, I_b) yielding a stable relative pose $T_{ba} \in \mathrm{SE}(3)$, *i.e.*, sufficient inliers and parallax. If no such pair exists, we leverage the monocular depth prior to constrain the cameras. We lift feature points from I_a using its depth D_a and calibration K_a, and form 2D-3D matches with corresponding points in I_b to estimate the absolute pose T_{ba} via PnP [12, 23]. We initialize the reconstruction using the estimated pose *i.e.*, $P_b = T_{ba}$ and P_a as the identity.

Initial 3D points: We initialize scene points $\mathcal{X}$ by lifting low-parallax inliers and triangulating the rest. Next, we scale each depth map D_i to be consistent with these 3D points by computing a scaling factor

$$D_i^* = D_i \cdot \underset{j,k}{\mathrm{median}} \left(\left\{ \hat{D}_i(X_k) / D_i(x_j) \right\} \right) \quad , \qquad (1)$$

where $\hat{D}_i(X_k) = (P_i \cdot X_k)_3$ denotes the depth of point X_k in camera frame i and $D_i(x_j)$ interpolates the depth map at coordinate x_j. We then refine the depth maps and the 3D points following the optimization described later in Sec. 3.3.

Next, we check whether the resulting depth maps are consistent, as described in Sec. 3.4, to reject the initial image pair if it is incorrectly posed, *e.g.*, due to symmetry. If they are consistent, we accept the pair and consider the images as registered, *i.e.*, $\mathcal{R} = \{a, b\}$, otherwise, we search for another pair. Finally, we augment the scene points $\mathcal{X}$ with image points not yet associated with a scene point by lifting them using their respective scaled and refined monocular depth.

3.2. Next View Registration

View selection: To register a next view, we consider images that have not been registered before, *i.e.*, $\{I_c \mid c \notin \mathcal{R}\}$. We rank these candidates by their maximum two-view correspondence scores to registered images, *i.e.*, $\arg\max_{i \in \mathcal{R}} \left(\sum q_{c,i} + \sum q^*_{c,i} \right)$. In ranked order, we attempt to register the candidates using 2D-3D correspondences and robust absolute pose estimation. Here, the 2D-3D correspondences contain both regular multi-view triangulated 3D

points but also single-view 3D points previously lifted using monocular depth. This enables us to register a next view without three-view overlap, whereas other SfM approaches can only consider points that have been triangulated from at least two registered views (see Fig. 3 and Appendix H.3).

Registration: If we find enough inliers, we extend the reconstruction with the estimate of the camera pose P_c. After scaling its depth map following Eq. (1), we extend the structure by either continuing the tracks of existing points or initialize new ones from lifted single-view points. Next, we refine the reconstruction (Sec. 3.3) and check the depth consistency (Sec. 3.4). If successful, we label the image as registered, *i.e.*, $\mathcal{R} \cup \{c\} \rightarrow \mathcal{R}$. Otherwise, we discard the new image with its associated structure and we try a different image.

3.3. Local and Global Refinement

After two-view initialization and registration of new images, we jointly refine the camera poses and the scene structure. Following COLMAP's scheduling for local and global bundle adjustment, we perform either a global refinement over all registered images and scene points or over a local window around the last registered image. This strategy results in an amortized linear runtime for incremental SfM [69].

Optimization problem: Different from standard bundle adjustment, we condition the estimated scene structure and camera poses on the monocular depth and normal priors. To robustly deal with noisy priors, which are generally predicted by imperfect neural networks, we estimate a set of multi-view consistent depth maps $\mathcal{D}^*$. We model this objective by optimizing the following overall cost function

$$\arg\min_{\mathcal{P},\mathcal{X},\mathcal{D}^*} C_{\text{BA}} + C_{\text{reg}} + C_{\text{int}} \quad . \tag{2}$$

The first term defines the standard bundle adjustment cost function for each observation of multi-view 3D points as

$$C_{\text{BA}} = \sum_{i \in \mathcal{R}} \sum_{j,k} \rho_{\text{BA}} \left(\left\| \pi\left(K_i, P_i, X_k\right) - x_j \right\|^2_{\Sigma_{x_j}} \right) \quad , \tag{3}$$

where $\|\cdot\|_\Sigma$ is the Mahalanobis distance, ρ_{BA} is a truncated Smooth-$L1$ loss, and $\pi(K, P, X) \in \mathbb{R}^2$ projects scene points into the image plane in pixel units. The second term penalizes deviations between scene points and the refined depths:

$$C_{\text{reg}} = \sum_{i \in \mathcal{R}} \sum_{j,k} \rho_{\text{reg}} \left(\left\| \hat{D}_i(X_k) - D_i^*\left(x_j\right) \right\|^2 \right) \quad , \tag{4}$$

which is robustified with the Cauchy loss ρ_{reg}. Appendix H.2 shows supporting examples. The last term defines a depth integration cost that conditions the refined depth maps on the monocular depth and normal priors

$$\begin{aligned} C_{\text{int}} = \sum_{i \in \mathcal{R}} \sum_{u,v} \Big[& \rho_{\text{prior}} \left(\left\| D_i^*(u,v) - D_i(u,v) \right\|^2_{\Sigma_{D_i}(u,v)} \right) \\ & + \rho_{\text{int}} \left(\left\| N_i(u,v) - \Delta D_i^*(u,v) \right\|^2_{\Sigma_{N_i}(u,v)} \right) \Big] \quad , \end{aligned} \tag{5}$$

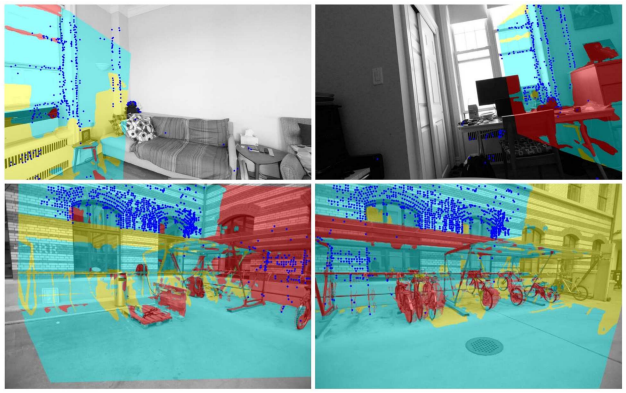


Figure 4. **Depth consistency check.** These two image pairs are incorrectly matched because of symmetries (blue points). Our approach successfully rejects them as a large ratio of pixels have an inconsistent depth (red), while ignoring occlusion (yellow) and areas with consistent depth (cyan).

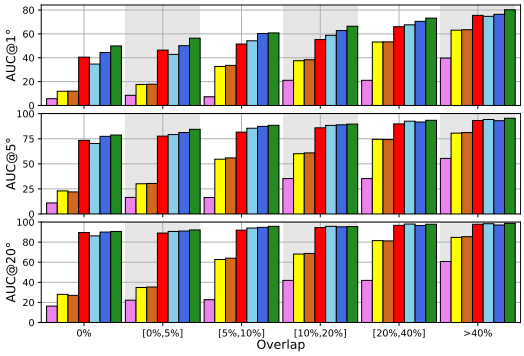
where $(u, v) \in [0, W_i] \times [0, H_i]$ are pixel coordinates, $\Delta D_i^*(u,v) \in \mathcal{S}^2$ defines the surface tangent plane normal using first-order differentiation of depth map D_i^* at (u, v), and $\rho_{\text{prior,int}}$ are truncated L_2 robust losses. In the depth integration, the first term conditions the refined on the prior depth, while the second term uses bilateral normal integration [11] with uncertainty weighting, for which we provide the exact formulation in Appendix B.

Efficient solving: The joint objective over $C_{\text{BA}} + C_{\text{reg}} + C_{\text{int}}$ leads to a sparsity structure in the Hessian, illustrated in Appendix G, not amenable to the Schur complement trick. To retain a tractable optimization, we therefore use an alternating block coordinate descent strategy, where we first minimize $C_{\text{reg}} + C_{\text{int}}$, for each image independently, using fixed scene points $\mathcal{X}$ and their propagated depth covariances $\Sigma_{\hat{D}_i(X_k)}$. Then, using fixed refined depth maps $\mathcal{D}^*$ and their propagated covariances $\Sigma_{D_i^*(x_j)}$, we jointly minimize $C_{\text{BA}} + C_{\text{reg}}$ over all views $\mathcal{P}$ and scene points $\mathcal{X}$. We alternate between these two steps for several rounds and interleave careful filtering of scene points using COLMAP's reconstruction filtering strategy.

3.4. Depth Consistency Check

Filtering the reconstruction based on sparse image observations alone, as performed by COLMAP and others, is important but limits the ability for identifying conflicting observations through occlusion and free-space reasoning [25, 39, 61]. As such, traditional sparse SfM often fails catastrophically when a single image is incorrectly registered due to symmetry issues, estimation failures, or other wrong decisions.

Our system uses the refined depths $\mathcal{D}^*$ to reason densely about occlusions and free space (Fig. 4). If a registered image I_c has too many conflicting observations with other registered images I_i, we de-register it and discard associated structure. To decide if an image causes inconsistencies, we



matching	approach	three-view visual overlap: minimal, 0%	three-view visual overlap: <5%	three-view visual overlap: <10%	three-view visual overlap: <30%	all images max overlap
SIFT	COLMAP	1.1/ 2.3/ 3.0	1.1/ 2.3/ 3.1	2.3/ 3.8/ 4.4	6.5/ 8.8/ 9.4	63.1/77.2/80.7
SP+LG	COLMAP	7.9/12.7/14.6	6.9/12.5/15.4	12.0/19.8/23.2	44.9/60.3/65.0	67.2/80.6/83.9
SP+LG	SLR	8.2/13.1/15.2	6.7/12.6/15.7	11.2/18.3/21.3	46.1/62.2/67.1	67.7/81.3/84.7
SP+LG	GLOMAP	8.4/15.8/22.5	5.5/15.3/25.7	12.1/25.3/35.7	50.1/66.7/71.8	67.5/78.5/82.3
SP+LG	**ours**	27.3/55.9/71.8	26.4/55.2/71.3	30.0/56.4/70.1	57.0/79.1/86.0	**74.3/88.3**/92.0
VGG-SfM		0.9/ 3.2/ 5.1	0.6/ 2.7/ 5.0	1.8/ 5.5/ 8.6	7.8/19.6/25.7	27.7/52.5/61.8
RoMa	COLMAP	7.3/11.9/14.0	5.3/9.6/12.4	11.5/18.9/22.6	41.4/56.4/61.8	63.6/76.4/79.7
LoFTR	DF-SfM	3.2/ 7.9/10.7	2.2/ 6.8/10.1	4.0/11.6/15.9	25.6/51.1/60.2	54.9/80.0/85.9
MASt3R	M-SfM	20.1/39.7/52.2	17.2/37.5/49.2	19.8/37.3/48.1	31.4/50.4/59.2	50.5/67.9/74.1
MASt3R	**ours**	**34.9/67.2/81.7**	**34.0/67.2/81.2**	37.7/**67.8/80.6**	55.5/79.3/**86.6**	70.3/88.2/**93.6**
RoMa	**ours**	33.4/60.6/74.4	32.8/60.7/74.2	39.7/65.8/77.9	**60.3/79.6**/85.6	71.6/87.0/91.3

Table 1. **SfM with low overlap on the ETH3D dataset.** Left: We construct a minimal dataset of triplets with reduced three-view overlap, as in Fig. 1. Right: We consider larger sets of images with increasing overlap. We report the AUC of the pose error up to 1/5/20°. Our approach yields the most accurate poses given either sparse or dense correspondences.

		SMERF dataset				Tanks & Temples dataset			
matching	overlap →	minimal	low	medium	high	minimal	low	medium	high
SIFT	COLMAP	.1/ 0.2/ 0.2	0.0/ 0.0/ 0.0	0.0/ 0.0/ 0.0	6.1/ 7.6/ 8.1	0.6/ 3.2/ 5.3	0.2/ 2.4/ 4.7	0.6/ 4.3/ 6.8	4.1/21.1/30.0
SP+LG	COLMAP	2.2/4.2/4.9	3.4/7.8/9.9	17.7/30.7/36.0	42.9/55.4/59.5	2.0/11.9/18.2	2.0/11.9/18.2	2.0/11.9/18.2	15.7/52.9/66.8
SP+LG	SLR	2.3/4.2/ 5.0	3.7/ 8.6/10.9	16.5/29.2/34.5	43.0/56.3/60.7	2.0/12.7/19.5	3.8/19.4/29.5	8.6/37.3/50.5	15.9/53.7/68.0
SP+LG	GLOMAP	3.5/8.0/13.3	3.4/9.6/15.9	20.9/35.3/42.8	48.1/60.9/65.4	1.3/9.1/17.8	2.7/15.7/25.0	9.0/30.9/39.9	16.1/45.7/55.9
SP+LG	**ours**	9.2/41.0/69.8	5.4/29.1/53.0	14.0/47.6/72.9	47.3/79.3/90.6	5.0/29.2/54.9	5.0/34.1/58.1	10.8/51.3/72.8	18.9/62.8/80.9
VGG-SfM		0.0/ 0.1/ 0.3	0.0/ 0.0/ 0.3	OOM	OOM	1.8/11.0/20.0	8.2/39.2/55.4	13.7/49.1/64.9	19.3/56.8/70.7
RoMa	COLMAP	2.5/4.3/5.2	6.9/13.5/16.3	23.0/36.2/40.7	40.5/50.2/53.1	3.2/13.6/19.6	7.4/28.5/38.7	12.8/43.0/54.1	17.4/46.4/56.9
LoFTR	DF-SfM	0.5/ 1.2/ 1.6	0.7/ 1.5/ 2.1	3.8/ 8.6/11.1	20.6/33.6/39.6	1.4/12.0/20.9	1.4/12.0/20.9	4.9/28.5/42.2	13.4/51.7/66.7
MASt3R	M-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	**18.0**/53.5/68.6	**26.6**/61.8/72.0	**27.0**/62.5/73.0	**28.5**/64.4/75.6
MASt3R	**ours**	**17.2/54.6/77.1**	**26.6/63.2/84.1**	**40.4/72.8/87.5**	**57.1/84.6/94.1**	14.8/**56.8/79.6**	21.7/**65.7/83.4**	22.8/**68.4/85.5**	25.9/**71.1/86.7**
RoMa	**ours**	10.6/41.0/61.8	10.1/37.2/59.2	21.1/48.4/66.7	41.4/69.3/79.4	7.2/41.5/65.3	10.3/44.0/63.2	14.2/56.2/75.9	22.1/67.4/86.1

Table 2. **SfM with low overlap on the SMERF and T&T datasets.** We report the AUC of the pose error up to 1/5/20° for approaches based on sparse (top) and dense matching (bottom). Ours is the only approach capable of reconstructing low-overlap scenes in the SMERF dataset.

reproject its depth map D_c^* into overlapping other images and accumulate a min-depth buffer $D_{i\leftarrow c}^*$ with associated reprojected depth uncertainty $\Sigma_{D_{i\leftarrow c}}$. Vice versa, we reproject other depth maps into image I_c to build $D_{c\leftarrow i}^*$ and $\Sigma_{D_{c\leftarrow i}}$.

We then compute the amount of forward-backward inconsistency as the ratio of inconsistent pixels with confidence γ:

$$\beta_i = \frac{1}{W_i H_i} \sum_{u,v} \mathbb{1}\left(\frac{D_i^*(u,v) - D_{i\leftarrow c}^*(u,v)}{\Sigma_{D_i}(u,v) + \Sigma_{D_{i\leftarrow c}}(u,v)} > \gamma \right) \quad (6)$$
$$+ \frac{1}{W_c H_c} \sum_{u,v} \mathbb{1}\left(\frac{D_c^*(u,v) - D_{c\leftarrow i}^*}{\Sigma_{D_c}(u,v) + \Sigma_{D_{c\leftarrow i}}(u,v)} > \gamma \right) ,$$

where $\mathbb{1}$ is the indicator function and $\beta_i \in [0, 2]$. We consider a view c as inconsistent if any of the overlapping views' β_i exceeds a ratio $\hat{\beta}$ of occluded pixels. We check the depth consistency for each newly registered view and once in the very end to discard potentially incorrect registrations.

3.5. Implementation Details

Correspondence search: We largely rely on COLMAP's correspondence search pipeline with the main difference that we use SuperPoint [13] and LightGlue [36] for sparse feature extraction and matching. Since our system does not require multi-view tracks, it can handle dense correspondences as well as is – we experiment with those of RoMA [19]. MASt3R [34], on the other hand, yields the best results when used to match sparse SuperPoint keypoints. For scalable matching, we compute global image features [3] to shortlist, per image, the n^* most similar other images as candidate pairs for feature matching.

Monocular Depth priors: Our most general configuration is based on monocular depth and normals predicted by Metric3D-v2 [28], which also estimates uncertainties. When matching images with MASt3R [34], we instead use its depth estimates along with normals from DSINE [4]. We also experiment with other depth models [6, 70] in Sec. 4.3.

Refinement optimization: The optimization of $C_{\text{reg}} + C_{\text{int}}$ is implemented on the GPU, while the term $C_{\text{reproj}} + C_{\text{reg}}$ is solved using Ceres [2] on the CPU. In our experiments, we did not experience a significant difference in end-to-end accuracy metrics between using the propagated covariance $\Sigma_{D_i^*(x_j)}$ as compared to the prior covariance $\Sigma_{D_i(x_j)}$. For faster runtimes, we therefore use $\Sigma_{D_i(x_j)}$ in our implementation of the local and global refinement optimization.

4. Experiments

We assess the performance of our system in both low-overlap and low-parallax scenarios.

4.1. Low-overlap reconstruction

Setup: We consider collections of images from several SfM datasets [15, 33, 53]. For each scene, we sample multiple groups of images with a pre-defined maximum amount of visual overlap across views. The overlap is computed as the ratio of covisible pixels using ground-truth depth maps, when available. Otherwise, it is defined as the ratio of co-visible feature points in the original, full SfM model. We assume calibrated cameras. Following standard practice [29], we evaluate the camera pose accuracy by comparing the relative poses to the ground truth. The error is defined as the maximum of angular errors in rotation and translation. We report the Area Under the recall Curve (AUC) up to 1/5/20°.

Approaches: Among those based on sparse correspondences, we consider COLMAP [50], with correspondences estimated by SIFT [38] and by SuperPoint [13] and LightGlue [36], as well as extensions for structure-less resectioning (SLR) [73] and global SfM with GLOMAP [43]. As for dense correspondences, existing systems cannot handle them easily, because they do not form multi-view tracks. Past works [47, 54] cluster them into tracks based on proximity. We thus run COLMAP with RoMa [19]. DF-SfM [24] combines COLMAP with LoFTR [59] and further refines the tracks. VGGSfM [64] estimates correspondences for a subset of images simultaneously. We also consider MASt3R-SfM [16]. It is based on two-view correspondences and point maps, which embed some monocular priors as well.

Triplet evaluation: To minimally capture the low-overlap scenario, we first consider triplets of images, as shown in Fig. 1. We sample them from 25 indoor and outdoor scenes of the ETH3D dataset [53] with different levels of three-view overlap, from zero to 50%. The results, in Tab. 1-left, confirm that COLMAP fails catastrophically for triplets with the least overlap. Learned matching (SP+LG) and structure-less resectioning both improve but are largely insufficient. Our approach is robust (high AUC@20°) and is increasingly accurate with the overlap (increasing AUC@1°). It significantly outperforms all existing approaches.

Full-scene SfM: We assemble larger images sets from multiple datasets. In addition to ETH3D, we consider 7 scenes from the Tanks & Temples dataset [33] and 4 scenes introduced in SMERF [15], with indoor/outdoor spaces, repeated structures, and texture-less views. The GT camera poses were estimated with COLMAP – achieving sufficient accuracy by using 10 to 100 times more images. For each dataset overlap bucket, we sample 5 image collections per scene.

The results, in Tab. 1-right and Tab. 2, show that our approach outperforms existing ones in the low-overlap scenarios, even when considering the same input correspondences. As the overlap increases, it maintains superiority in terms of robustness, even when the views are dense (*all images* on ETH3D). Using dense features like RoMa or MASt3R works better than SuperPoint+LightGlue. MASt3R's ability to reject negative pairs and handle extreme viewpoints yields the best overall performance. Our approach is, however, less accurate than MASt3R-SfM at AUC@1° in the T&T dataset, due to a lack of foreground matches in object-centric scenes. We show qualitative examples in Fig. 3.

		pose AUC at X°		
strategy	approach	1°	10°	30°
incremental	SP+LG + COLMAP (default)	18.3	49.2	55.1
incremental	SP+LG + COLMAP (tuned)	25.8	71.0	82.4
incremental	SP+LG + StudioSfM	19.9	72.3	86.5
global	SP+LG + GLOMAP	34.2	81.0	90.7
incremental	SP+LG + **ours**	30.9	79.5	90.4
global	MASt3R-SfM	33.4	80.8	91.2
incremental	MASt3R + **ours**	**35.5**	**81.9**	**91.5**

Table 3. **SfM with low parallax on the RealEstate10k dataset.** Our approach outperforms previous incremental SfM pipelines as well as MASt3R-SfM, even though global SfM does not suffer from the same fundamental low parallax limitations.

4.2. Low-parallax reconstruction

As incremental SfM relies on noisy 3D structure for registration, it is typically prone to failure in low-parallax scenarios.

Setup: Following past research [16, 63], we consider the RealEstate10k dataset [74]. It features indoor and outdoor scenes with low texture and challenging camera motion, *e.g.*, forward translation and in-place rotation. We evaluate COLMAP and GLOMAP [43] with SuperPoint and Lightglue [13, 36] and MASt3R-SfM [16]. We found that relaxing the minimum triangulation angle allowed in COLMAP is critical. We report results for this variant in addition to one using the default parameters. We also evaluate StudioSfM [37], which is designed for this low-parallax scenario. It also leverages monocular depth but only for initialization, and regularization, and it does not refine it or handle uncertainties.

Results: In Table 3, we distinguish between global and incremental SfM, as global methods do not rely on noisy 3D structures for registration. MP-SfM largely closes the gap between incremental and global SfM and outperforms COLMAP, StudioSfM and MASt3R-SfM, likely because they do not handle well uncertainties of monocular priors.

4.3. Ablation Study

We present ablations to justify the design of our system.

Impact of the monocular priors: In Tab. 4, we compare the SfM accuracy with depth and normals estimated by different approaches and we study the impact of the prior uncertainty.

Impact of each component: In Tab. 5, we evaluate the

#	depth		surface normals		ETH3D dataset		SMERF dataset	
	model	uncertainty	model	uncertainty	minimal overlap	all images	minimal overlap	high overlap
1	**Metric3D-v2** [28]	**yes**	**Metric3D-v2** [28]	**yes**	27.3/55.9/71.8	74.3/88.3/92.0	9.2/**41.0/69.8**	**47.3/79.3/90.6**
2	Metric3D-v2 [28]	no	Metric3D-v2 [28]	yes	27.3/55.4/71.2	72.8/86.8/90.5	8.7/38.7/65.8	40.0/67.2/82.1
3	Metric3D-v2 [28]	yes	Metric3D-v2 [28]	no	27.3/55.8/71.7	74.2/88.3/91.9	9.1/39.9/67.2	43.6/76.2/89.2
4	Metric3D-v2 [28]	yes	DSINE [4]	yes	26.3/54.2/70.4	74.3/88.3/91.9	8.9/39.0/66.7	44.5/78.3/89.8
5	Metric3D-v2 [28]	no	DSINE [4]	yes	26.3/53.7/69.5	72.5/86.8/90.5	8.2/36.2/61.9	41.2/68.3/79.4
6	Depth Pro [6]	no	DSINE [4]	yes	23.3/47.5/63.4	74.5/88.3/91.9	7.0/30.8/57.0	30.7/56.5/72.7
7	Depth Anything v2 [70]	no	DSINE [4]	yes	21.3/44.8/60.8	71.8/86.2/89.9	4.0/19.8/43.8	20.7/42.2/60.4
8	M3D-v2 Large [28]	yes	M3D-v2 Large [28]	yes	**28.2/56.5/72.2**	**75.0/88.6/92.1**	**9.2**/40.2/69.4	43.4/72.6/84.1
9	M3D-v2 Small [28]	yes	M3D-v2 Small [28]	yes	26.1/54.0/70.6	74.3/88.0/91.8	6.8/33.4/62.5	34.5/59.8/76.4

Table 4. **Comparing monocular priors.** Only Metric3D-v2 estimates depth uncertainties – critical for fusing with feature correspondences (1 vs 2, 4 vs 5). Models without depth uncertainties are less reliable (1 vs 6/7). Factoring out uncertainties, Metric3D-v2 outperforms DepthPro and DepthAnything-v2 overall (5 vs 6 vs 7). Having uncertainties for surface normals is also important (1 vs 3). Both Metric3D-v2 and DSINE estimate normal uncertainties and work well with our pipeline (1 vs 4). While the smaller and more efficient variants of Metric3D-v2 (Large and Small) struggle on SMERF, they perform similarly as the largest model (Giant2) on ETH3D (1 vs 8/9).

variant	ETH3D dataset		SMERF dataset	
	min. overlap	all images	min. overlap	high overlap
SP+LG + **ours**	27.3/55.9/71.8	74.3/88.3/92.0	9.2/41.0/69.8	47.3/79.3/90.6
no depth refinement	26.8/55.2/69.9	71.9/87.6/91.7	8.4/37.6/66.7	32.2/63.6/82.0
no depth reg.	23.6/49.6/65.9	75.1/88.7/92.2	5.7/21.0/45.9	45.5/64.1/73.0
ground-truth depth	42.9/68.0/77.1	73.8/87.3/90.8	-	-
no lifting	10.6/16.1/18.7	74.1/87.2/90.6	1.0/ 1.7/ 2.1	51.9/69.6/75.2
ROMA + **ours**	33.4/60.6/74.4	71.6/87.0/91.3	10.6/41.0/61.8	41.4/69.3/79.4
no depth refinement	30.9/60.1/74.5	66.6/85.3/90.8	8.8/35.7/57.4	29.3/61.1/77.5
no depth reg.	29.1/50.5/65.3	72.1/87.6/91.8	7.2/24.8/45.4	54.0/73.6/80.6
no lifting	13.5/19.2/22.4	70.4/84.6/88.1	1.4/ 2.3/ 2.9	55.0/72.9/77.8

Table 5. **Ablation of our pipeline.** Refining the prior depth with surface normals is consistently effective as they are complementary. The depth regularization is useful in low overlap. It can impair the accuracy for under-constrained 2-view tracks because monocular priors provide noisier constraints than reprojection errors. Using the ground-truth LiDAR depth, which is sparser but not biased, largely improves the accuracy at 1°. While lifting greatly improves low overlap reconstruction, it introduces noise in well-posed scenarios.

variant	ETH3D dataset		SMERF dataset	
	min. overlap	all images	min. overlap	high overlap
no check	26.9/55.7/71.5	74.7/88.4/92.0	8.7/39.1/66.9	32.8/58.8/72.8
Doppelgangers (filter)	21.7/39.6/50.0	70.2/83.5/86.9	0.2/1.0/1.4	8.3/15.9/22.6
Doppelgangers (rank)	26.9/**56.0/72.1**	**77.0/91.4/94.8**	8.5/38.9/66.5	32.9/57.0/67.5
our check	**27.3**/55.9/71.8	74.3/88.3/92.0	**9.2/41.0/69.8**	**47.3/79.3/90.6**

Table 6. **Evaluation of the depth consistency check.** Our check significantly improves the camera poses on scenes with repeated structures (SMERF). It is more effective than the learning-based approach Doppelganger, which is affected by domain shift.

impact of the depth refinement, regularization, and lifting. All components are beneficial in different scenarios.

Robustness to symmetry: In Tab. 6, we show that our depth consistency check is effective in scenes with symmetries. We compare it to Doppelgangers [9] as an approach that also rejects non-matching image pairs but based on a neural network. We use it to pre-filter pairs or to rerank images in the incremental process.

Efficiency: Extracting depth and normals takes 0.65 s per image. Sky masks are also extracted to handle overly confident depth estimates, adding 0.12 s per image. Refining the depth takes on average 70 ms per image, is done for only few images, and is faster as the poses and 3D points are less updated throughout the incremental process. The depth check takes 14 ms per newly registered image.

5. Limitations

Our system depends on reliable uncertainties for the monocular priors. State-of-the-art depth models rarely estimate uncertainties and those that do are often over-confident. Estimating accurate normals is also difficult for some types of surfaces, such as vegetation, limiting the performance outdoors.

Furthermore, while our approach pushes SfM beyond previously tackled scenarios, it still relies on standard components that often fail in the harsh environments we address.

For example, image retrieval struggles with strong perspective changes and symmetries, while image matching often cannot handle extremely low visual overlap. We explore solutions to these limitations, such as the depth consistency check, but our approach will also seamlessly benefit from future improvements with little adaptation. Lastly, our approach increases the run time of SfM, which is often only the first step of complex applications.

6. Conclusion

This paper introduces an approach that significantly improves the robustness of SfM in low-overlap, low-parallax and high symmetry scenarios, which are common failure cases in data captured by non-expert users. To achieve this, we augment the classical SfM paradigm with monocular depth and surface normals estimated by deep neural networks. Our system leverages these priors in multiple ways: at initialization, when registering new views, when optimizing the poses and geometry, and to reject incorrect registrations, thereby making SfM more robust to repeated patterns. This results in improvements across different image correspondences, environments, and levels of difficulty.

Acknowledgements

The authors thank Shaohui Liu for useful discussions. This work was supported by the Swiss National Science Foundation Advanced Grant 216260: Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World.

References

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

[2] Sameer Agarwal, Keir Mierle, and Others. Ceres Solver. http://ceres-solver.org, 2024. 6

[3] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. NetVLAD: CNN architecture for weakly supervised place recognition. In *CVPR*, 2016. 6

[4] Gwangbin Bae and Andrew J. Davison. Rethinking Inductive Biases for Surface Normal Estimation. In *CVPR*, 2024. 2, 6, 8, 1

[5] Paul A Beardsley, Andrew Zisserman, and David William Murray. Sequential Updating of Projective and Affine Structure from Motion. *IJCV*, 1997. 2

[6] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: Sharp Monocular Metric Depth in Less Than a Second. *arXiv:2410.02073*, 2024. 2, 6, 8, 4

[7] Eric Brachmann and Carsten Rother. Visual Camera Relocalization from RGB and RGB-D Images Using DSAC. *IEEE TPAMI*, 2021. 2

[8] 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 *ECCV*, 2024. 2

[9] Ruojin Cai, Joseph Tung, Qianqian Wang, Hadar Averbuch-Elor, Bharath Hariharan, and Noah Snavely. Doppelgangers: Learning to Disambiguate Images of Similar Structures. In *ICCV*, 2023. 1, 8

[10] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid camera pose estimation. In *CVPR*, 2018. 2

[11] Xu Cao, Hiroaki Santo, Boxin Shi, Fumio Okura, and Yasuyuki Matsushita. Bilateral normal integration. In *European Conference on Computer Vision*, pages 552–567. Springer, 2022. 3, 5, 1

[12] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally Optimized RANSAC. In *GCPR*, 2003. 4

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. In *CVPR Workshops*, 2018. 1, 2, 6, 7, 5, 8

[14] Eric Dexheimer and Andrew J. Davison. COMO: Compact Mapping and Odometry. In *ECCV*, 2024. 3

[15] 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. *arXiv:2312.07541*, 2023. 7, 2, 6

[16] 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:2409.19152*, 2024. 1, 2, 7, 4, 6

[17] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *CVPR*, 2019. 1, 2

[18] Mihai Dusmanu, Johannes L. Schönberger, and Marc Pollefeys. Multi-View Optimization of Local Feature Geometry. In *ECCV*, 2020. 2

[19] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust Dense Feature Matching. *CVPR*, 2024. 2, 6, 7, 8

[20] David Eigen, Christian Puhrsch, and Rob Fergus. Depth Map Prediction from a Single Image using a Multi-Scale Deep Network. *NeurIPS*, 2014. 2

[21] 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 *ECCV*, 2010. 1, 2

[22] Marcel Geppert, Viktor Larsson, Pablo Speciale, Johannes L Schönberger, and Marc Pollefeys. Privacy Preserving Structure-from-Motion. In *ECCV*, 2020. 1

[23] Bert M. Haralick, Chung-Nan Lee, Karsten Ottenberg, and Michael Nölle. Review and Analysis of Solutions of the Three Point Perspective Pose Estimation Problem. *IJCV*, 1994. 4

[24] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-Free Structure from Motion. In *CVPR*, 2024. 7

[25] Jared Heinly, Enrique Dunn, and Jan-Michael Frahm. Correcting for Duplicate Scene Structure in Sparse 3D Reconstruction. In *ECCV*, 2014. 5

[26] Jared Heinly, Johannes L Schonberger, Enrique Dunn, and Jan-Michael Frahm. Reconstructing the World* in Six Days *(as Captured by the Yahoo 100 Million Image Dataset). In *CVPR*, 2015. 1, 2

[27] Derek Hoiem, Alexei A Efros, and Martial Hebert. Geometric Context from a Single Image. In *ICCV*, 2005. 2

[28] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: A Versatile Monocular Geometric Foundation Model for Zero-Shot Metric Depth and Surface Normal Estimation. *IEEE TPAMI*, 2024. 6, 8, 1, 4

[29] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiří Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *IJCV*, 2021. 7

[30] Klas Josephson, Martin Byrod, Fredrik Kahl, and Kalle Astrom. Image-based localization using hybrid feature correspondences. In *CVPR*, 2007. 2

[31] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. In *CVPR*, 2024. 2

[32] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *TOG*, 2023. 1

[33] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *TOG*, 2017. 7, 2, 6

[34] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding Image Matching in 3D with MASt3R. In *ECCV*, 2024. 2, 6, 1, 5, 8

[35] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In *ICCV*, 2021. 2

[36] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 2, 6, 7, 8

[37] Sheng Liu, Xiaohan Nie, and Raffay Hamid. Depth-Guided Sparse Structure-from-Motion for Movies and TV Shows. In *CVPR*, 2022. 3, 7

[38] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *IJCV*, 60(2):91–110, 2004. 7

[39] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-Time Visibility-Based Fusion of Depth Maps. In *ICCV*, 2007. 5

[40] 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. In *ECCV*, 2020. 1

[41] Anastasiia Mishchuk, Dmytro Mishkin, Filip Radenovic, and Jiri Matas. Working hard to know your neighbor's margins: Local descriptor learning loss. *NeurIPS*, 30, 2017. 2

[42] 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. 2

[43] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *ECCV*, 2024. 1, 2, 7

[44] Marc Pollefeys, Luc Van Gool, Maarten Vergauwen, Frank Verbiest, Kurt Cornelis, Jan Tops, and Reinhard Koch. Visual Modeling with a Hand-held Camera. *IJCV*, 2004. 2

[45] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching with Graph Neural Networks. In *CVPR*, 2020. 1, 2

[46] 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 *ECCV*, 2022. 1

[47] Paul-Edouard Sarlin, Philipp Lindenberger, Viktor Larsson, and Marc Pollefeys. Pixel-Perfect Structure-From-Motion With Featuremetric Refinement. *IEEE TPAMI*, 2023. 7

[48] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning Depth from Single Monocular Images. *NeurIPS*, 2005. 2

[49] Frederik Schaffalitzky and Andrew Zisserman. Multi-view Matching for Unordered Image Sets, or "How Do I Organize My Holiday Snaps?". In *ECCV*, 2002. 2

[50] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 1, 2, 3, 7, 5

[51] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*, 2016. 1, 6

[52] Johannes Lutz Schönberger, Hans Hardmeier, Torsten Sattler, and Marc Pollefeys. Comparative Evaluation of Hand-Crafted and Learned Local Features. In *CVPR*, 2017. 2

[53] 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 *CVPR*, 2017. 7, 1, 2, 4, 5, 6

[54] Zehong Shen, Jiaming Sun, Yuang Wang, Xingyi He, Hujun Bao, and Xiaowei Zhou. Semi-Dense Feature Matching With Transformers and its Applications in Multiple-View Geometry. *IEEE TPAMI*, 2023. 7

[55] Sudipta Sinha, Marc Pollefeys, and Leonard McMillan. Camera Network Calibration from Dynamic Silhouettes. In *CVPR*, 2004. 2

[56] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent. *ECCV*, 2024. 2

[57] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo Tourism: exploring photo collections in 3D. In *TOG*, 2006. 1, 2

[58] Pablo Speciale, Johannes L Schonberger, Sing Bing Kang, Sudipta N Sinha, and Marc Pollefeys. Privacy preserving image-based localization. In *CVPR*, 2019. 1

[59] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-Free Local Feature Matching with Transformers. In *CVPR*, 2021. 2, 7

[60] Richard Szeliski and Sing Bing Kang. Recovering 3D Shape and Motion from Image Streams Using Non-Linear Least Squares. *Journal of Visual Communication and Image Representation*, 1994. 2

[61] Hajime Taira, Ignacio Rocco, Jiri Sedlar, Masatoshi Okutomi, Josef Sivic, Tomas Pajdla, Torsten Sattler, and Akihiko Torii. Is this the right place? geometric-semantic pose verification for indoor visual localization. In *CVPR*, 2019. 5

[62] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image Calibration with Geometric Optimization. In *ECCV*, 2024. 2

[63] Jianyuan Wang, Christian Rupprecht, and David Novotny. PoseDiffusion: Solving Pose Estimation via Diffusion-aided Bundle Adjustment. In *ICCV*, 2023. 7

[64] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual Geometry Grounded Deep Structure From Motion. In *CVPR*, 2024. 1, 2, 7

[65] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. DUSt3R: Geometric 3D Vision Made Easy. In *CVPR*, 2024. 1, 2

[66] Tong Wei, Yash Patel, Alexander Shekhovtsov, Jiri Matas, and Daniel Barath. Generalized differentiable RANSAC. In *ICCV*, 2023. 2

[67] Xingkui Wei, Yinda Zhang, Zhuwen Li, Yanwei Fu, and Xiangyang Xue. DeepSFM: Structure From Motion Via Deep Bundle Adjustment. In *ECCV*, 2020. 2

[68] Changchang Wu. VisualSFM : A Visual Structure from Motion System. http://www.cs.washington.edu/homes/ccwu/vsfm, 2011. 1, 2

[69] Changchang Wu. Towards Linear-time Incremental Structure from Motion. In *3DV*, 2013. 5

[70] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. *arXiv:2406.09414*, 2024. 2, 6, 8, 4, 9

[71] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: Towards Zero-shot Metric 3D Prediction from A Single Image. In *ICCV*, 2023. 2

[72] Christopher Zach, Manfred Klopschitz, and Marc Pollefeys. Disambiguating Visual Relations Using Loop Constraints. In *CVPR*, 2010. 1

[73] Enliang Zheng and Changchang Wu. Structure From Motion Using Structure-Less Resection. In *ICCV*, 2015. 2, 7

[74] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo Magnification: Learning View Synthesis using Multiplane Images. In *TOG*, 2018. 7, 6