VidMap: Exploiting Temporal Structure for Video-Based Structure-from-Motion

Zador Pataki[1] Paul-Edouard Sarlin[2] Marc Pollefeys[1,3]

[1]ETH Zurich [2]Google [3]Microsoft Spatial AI Lab

Abstract. Accurately recovering the camera's calibration and metric poses for any unconstrained video would unlock large-scale training data for navigation and scene understanding. The dominant approaches to this problem are severely limited: Simultaneous Localization and Mapping (SLAM) is sensitive to initialization and transient failures due to its causal, incremental nature; it is often over-optimized for real-time operation and generally requires known camera calibration; while Structure-from-Motion (SfM) typically forgoes any image ordering, enabling optimal initialization and global optimization, but lacks robustness to visual symmetries and extreme motions. To bridge this gap, we introduce a system that combines the strong sequential constraints of SLAM with the flexibility and global optimization of offline SfM, enabling the metric reconstruction of arbitrary, long, uncalibrated videos. This system leverages recent advances in wide-baseline dense image matching, treats temporal ordering as a first-class citizen for reliable loop closure, and augments global optimization with metric monocular depth priors. As a result, thorough evaluations on diverse, challenging datasets that exhibit extreme motion and visual symmetries reveal that our approach is significantly more robust and accurate than both state-of-the-art SLAM and SfM, classical or learned, with given or unknown camera calibration.

Keywords: Structure-from-Motion · SLAM · Video · Calibration

1 Introduction

Video is now the dominant medium generated and consumed in the world, with billions of videos available on online platforms. Robotics and wearable systems also perceive the world through uninterrupted video streams. To best make use of this data, many computer vision applications require camera poses and calibration. These cannot be reliably estimated by existing systems, which have been primarily engineered for data captured by expert users and thus cannot handle the challenges of arbitrary videos: degenerate and fast motions, visual symmetries, texture-less environments, or unknown calibration. These systems typically fall into the categories of SLAM or SfM.

SLAM [5, 9, 16, 21, 26, 31, 45] is designed for real-time operation and thus registers each new image as it becomes available, in a causal and incremental

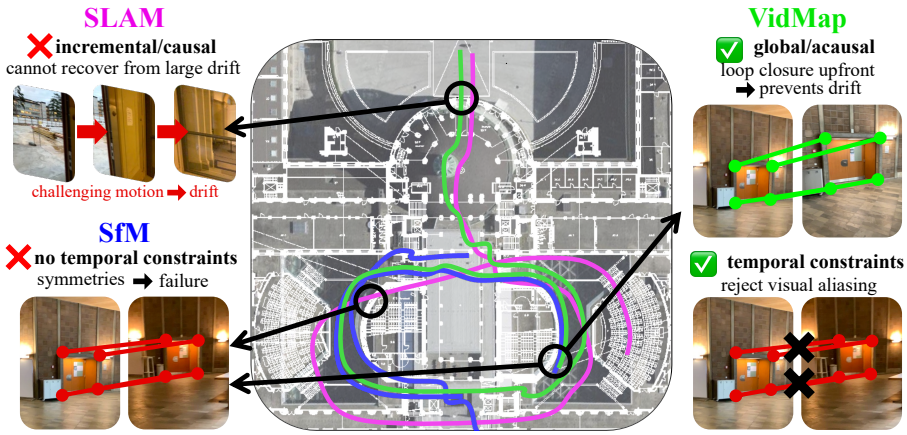

Fig. 1: VidMap reconstructs long, unconstrained videos with complex motion and visual ambiguity. SLAM cannot recover from critical tracking failures and drift because it is causal. SfM treats videos as unordered image sets and therefore cannot leverage temporal cues to disambiguate visual symmetries. VidMap uses global cues, such as loop closure, upfront to prevent drift, while preserving temporal provenance to reject visual aliasing. Moreover, by injecting metric depth constraints into global positioning, VidMap remains robust against scale drift and degenerate motion.

manner. The system commits to a single solution at each point in time using only past information, even under partial observability of the environment or motion. SLAM is thus sensitive to weak initialization, as errors introduced early on are propagated throughout the sequence and compound over time. Most systems thus require accurate estimates of the camera intrinsics [5,31,45]. Transient failures, due to degenerate motions, result in a tracking loss that is often unrecoverable. In practice, academic SLAM systems are optimized for speed and tightly couple all components, making it difficult to explore alternative designs and leverage advances in computer vision.

In contrast, SfM is designed for offline operation and considers all available information before committing to a solution, thus maximizing accuracy. This process is either global [30, 34, 54], by solving for all poses at once, or incremental [27, 37, 40, 43], by registering images in the optimal order, and can jointly optimize the camera calibration. SfM ignores the sequential structure of videos and can thus likewise handle unordered sets of images [34, 40, 43], making it more widely applicable. This however means that SfM cannot exploit the temporal continuity of videos: it treats all image pairs identically and is thus less robust to visual symmetries. Additionally, because it performs all data association before motion estimation, SfM is unable to adapt its computation and keyframing to motion dynamics.

In general, classical SfM and SLAM typically lack robustness to textureless environments, poor lighting, blur, and degenerate motion. Data-driven priors for image matching and tracking [19,25,38,44,59], calibration and metric depth [18, 36,47,51], 3D surface estimation [13,23,32,35], and symmetry disambiguation [4,

55] make SfM and SLAM more robust but do not address their above-mentioned structural limitations. Learned end-to-end models [48, 52] are more robust and efficient but lack precision, scalability, and long-term consistency.

To bridge this gap, we introduce *VidMap*, a new system that combines the complementary strengths of SLAM and SfM. VidMap leverages the intrinsic temporal ordering of videos to handle extreme motions while avoiding incorrect data association due to visual symmetries. This ordering makes it easier to leverage dense, pixel-wise matching, which has been historically incompatible with the discrete tracks required by SfM. Inspired by SLAM, we separate local tracking and long-term loop closure in the global optimization of SfM, thereby maximizing local consistency while reducing long-term drift. We empower this optimization with metric monocular depth priors to mitigate scale drift and handle degenerate configurations where multi-view constraints alone are insufficient.

Our system is highly efficient, leveraging motion-based keyframing of SLAM, while being trivially parallelizable over long sequences. We evaluate VidMap on new benchmarks [2, 39] that include very long videos captured by robots and non-expert users, thus exhibiting extreme motions, viewpoints, and symmetries. Our experiments reveal that VidMap outperforms all existing approaches by a large margin, both SLAM and SfM, from classical to end-to-end learned, when camera intrinsics are given or unknown, while retaining similar performance on established saturated benchmarks [3, 41]. The code is publicly available at `https://github.com/cvg/vidmap` with a permissive license.

2 Related Work

Visual SLAM leverages the sequential structure of video by tracking features across consecutive frames, selecting keyframes based on motion, and closing loops to reduce drift. Classical systems [5, 15] perform well when intrinsics are known and scenes are well-textured but degrade under repetitive structure and require careful calibration. Learned SLAM methods [26, 45, 46] replace handcrafted components, like tracking, with learned alternatives, improving accuracy in challenging conditions while retaining the same incremental architecture. Later research [18, 23, 29, 32] integrates dense geometry predictions and monocular depth into SLAM, yet the underlying processing remains causal, estimating the geometry before observing the entire sequence. This causal commitment is the structural bottleneck: pose and calibration errors from degenerate configurations (forward motion, low parallax, local symmetries) are locked in before well-conditioned observations are used. Loop closure corrects pose drift but cannot undo decisions already embedded in the map, and tracking failures are unrecoverable. In contrast, our approach leverages SfM's global optimization and self-calibration to maximize robustness and best use multi-view constraints.

Structure-from-Motion optimizes over all observations jointly. Incremental approaches [27, 37, 40, 43] register images one at a time through repeated resectioning and bundle adjustment; global methods [8, 30, 34, 54] solve for all camera

parameters simultaneously. SfM's optimization is provenance-agnostic, however: it treats all correspondences identically regardless of whether they arise from sequential overlap or loop closure, leaving it vulnerable to visual aliasing from repetitive structure [4, 55]. Learned image matching [11, 17, 25] improves the front-end robustness but does not address this structural limitation. Learned representations have been integrated into SfM [13,49], replacing geometric primitives with dense pointmaps or learned tracks but trading geometric precision on long sequences. ParticleSfM [59] integrates dense point trajectories from optical flow into global SfM, improving coverage in textureless regions, but constructs correspondences from chained consecutive flow without loop closure, limiting drift correction to local smoothing. FlowMap [42] recovers poses and depth from video via gradient descent over optical flow but is restricted to short sequences without geometric verification. Methods that integrate monocular depth into classical SfM [28, 35] preserve geometric precision but inherit provenance-agnostic matching. Our approach instead augments global SfM with video-aware tracking and monocular depth priors. Distinguishing sequential from loop-closure correspondences preserves temporal trust while mitigating visual aliasing; metric depth with per-image scale estimation regularizes degenerate configurations where multi-view constraints alone are insufficient.

Feedforward learned algorithms [20,24,48,52,53,57] jointly predict camera poses, calibration, and dense 3D structure from multiple images in a feedforward manner, without any explicit geometric optimization. Through their learned priors, they achieve impressive robustness in scenarios with sparse views and low visual overlap. However, their precision remains lower when images have sufficient overlap, and they exhibit significantly more drift and inconsistency with long, high-frequency video sequences. Since memory and context windows are limited, scaling beyond the training distribution requires chunked alignment [10] or streaming fusion [29], re-introducing incremental drift between chunks without propagating uncertainties across them. Other approaches tailored to videos [7,50] process frames sequentially with compressed memory, extending effective context through recurrence or test-time training, but degrade beyond their training distribution of motions, environments, and video lengths. The core tradeoff is structural: learned models gain robustness and speed at the cost of geometric precision, scalability, and zero-shot generalization. In contrast, we use learned models for what they do well—image matching and metric depth estimation—but rely on explicit geometric optimization for precision and scalability.

3 Method

Problem formulation. Given a video $\mathbf{I} = \{I_1, \ldots, I_N\}$ with unknown camera intrinsics, we estimate camera poses $T_i \in SE(3)$ for a selected set of keyframes $\mathcal{K} \subset \{1, \ldots, N\}$, intrinsics K_i, and a sparse 3D point cloud $\mathbf{X} = \{X_k\}$.

Overview. First, *video-aware extraction* (Secs. 3.1 and 3.2) adapts sequential tracking and keyframing concepts from SLAM into an offline dense matching

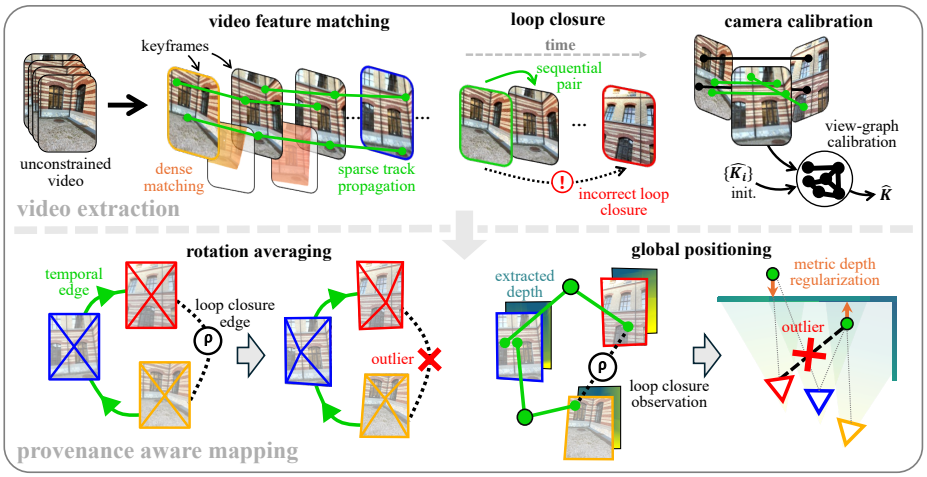

Fig. 2: System overview. We select keyframes using optical flow and extract monocular depth for each keyframe. We sample and propagate sparse tracks across keyframes using dense matching and across loop-closures detected using image retrieval. We estimates initial camera intrinsics using view-graph calibration and monocular calibration priors. Mapping preserves provenance for pairwise relative poses and track observations: we record which edges are sequential or loop closures to treat them differently. Robust optimization therefore treats incorrect loop-closure image pairs as outliers during rotation averaging and rejects loop-closure observations caused by visual aliasing during global positioning. By injecting metric depth priors, global positioning remains robust to degenerate motion and scale drift.

process: selecting keyframes from the continuous video stream, building long sparse tracks with covariance through dense warp chaining, separating loop closure (LC) observations from sequential ones, and extracting monocular depth and focal length priors per keyframe. Although extraction follows the temporal sequence to exploit video structure, it is non-causal: multiple passes revisit the data (low-resolution matching, keyframe detection, high-resolution refinement, track propagation), and no decision is committed until the entire sequence has been observed. Second, *global mapping* (Sec. 3.4) introduces new optimization machinery into global SfM [34]: monocular depth priors with per-image scale estimation in global positioning (GP) and provenance-dependent robust losses that distinguish sequential from LC edges.

3.1 Video Feature Matching

Dense matchers trained on large-scale data have become highly robust to appearance variation, viewpoint change, and textureless surfaces. We leverage this robustness to extract sparse feature tracks with sub-pixel precision and localization covariance, properties that sparse matching cannot provide in low-texture regions. Our approach adapts sequential tracking concepts from SLAM (keyframing, sliding-window propagation, flow chaining) to the dense matching setting, freed from real-time constraints. The key structural property is that consecutive

video frames are locally unambiguous: adjacent views share high overlap, and dense matching uses the entire image context to resolve the local ambiguities that confuse sparse or orderless matching. By chaining dense correspondences along time, this guarantee extends across the full video, producing long tracks that are robust to the perceptual aliasing that defeats image retrieval.

Tracking. We build on a coarse-to-fine dense matcher [14]. Given images I_a and I_b, the matcher yields a dense warp field $\mathcal{W}_{a\to b} : \mathbb{R}^2 \to \mathbb{R}^2$, per-pixel certainty $c_{a\to b} \in [0,1]$, and per-pixel localization covariance $\Sigma_{a\to b} \in \mathbb{R}^{2\times 2}$. Unlike classical SfM, which depends on repeatable keypoint detectors, dense matchers predict accurate warps for each pixel, from which correspondences can be sampled at any location, even in textureless regions. Given sparse keypoints in image I_a, each point $\mathbf{x}$ is propagated to I_b via the warp: $\hat{\mathbf{x}}_b = \mathcal{W}_{a\to b}(\mathbf{x}_a)$, with covariance $\Sigma_{a\to b}$ predicted by the matcher. Chaining this propagation across consecutive frames produces long tracks. Under a Gaussian error model, the localization covariance accumulates additively across sequential hops: $\Sigma_i^{\text{seq}} = \Sigma_{i-1}^{\text{seq}} + \Sigma_{(i-1)\to i}$, providing a principled measure of the spatial drift accumulated along each track.

Depth extraction. For each selected keyframe i, we extract a monocular depth map [24], yielding a depth prior m_{ik} and uncertainty σ_{ik} for each observed pixel k. These priors are used during global positioning and bundle adjustment to regularize degenerate geometry and scale drift.

Keyframe Detection. Not all video frames are needed for reconstruction. We select keyframes adaptively based on observed motion, triggering on actual scene displacement rather than fixed temporal intervals. For each consecutive frame pair, the dense matcher produces a low-resolution warp field and certainty map (coarse stage only), from which we estimate the inter-frame motion. Starting from sparse keypoints sampled at each keyframe, we propagate them through subsequent frames via the warp field. We measure scene motion by the fraction of keypoints for which tracking was lost (*e.g.* due to occlusion) or that have drifted beyond a displacement threshold. A new keyframe is triggered when this fraction surpasses a target threshold.

Track Chaining. Given the selected keyframes $\mathcal{K}$, we build tracks by chaining dense correspondences across the keyframe sequence. For each keyframe $\mathcal{K}_i$, processed in sequence, we compute high-resolution dense matches from up to W preceding keyframes to $\mathcal{K}_i$ and propagate keypoints forward as described above.

Multi-flow drift correction. Because each keyframe $\mathcal{K}_i$ is matched against up to W=9 preceding keyframes, direct dense correspondences to earlier keyframes are already available, not just to the immediate predecessor. Following [33], we select the path with lowest total uncertainty among all such correspondences. For each $\mathcal{K}_{i-n}$ within the window ($n \in \{1,\dots,W\}$) that has a direct match to $\mathcal{K}_i$, the total covariance along the direct path is $\Sigma_i^{(i-n)} = \Sigma_{i-n}^{\text{seq}} + \Sigma_{(i-n)\to i}$. We select the anchor $n^* = \arg\min_n \operatorname{tr}(\Sigma_i^{(i-n)})$ and accept the direct prediction only if its Mahalanobis distance to the sequential estimate under $\Sigma_i^{(i-n^*)} + \Sigma_i^{\text{seq}}$

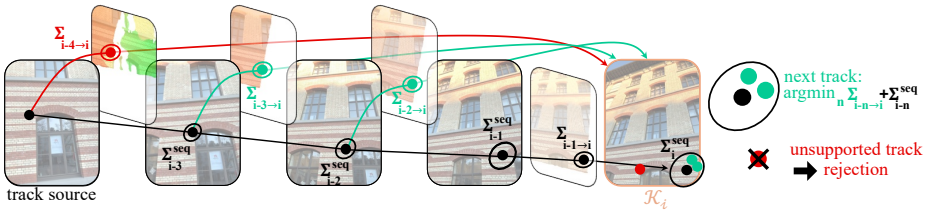

Fig. 3: Multi-flow drift correction. For each new keyframe $\mathcal{K}_i$, sequential tracks are propagated through the latest match $\mathcal{K}_{i-1} \to \mathcal{K}_i$, yielding accumulated covariance Σ_i^{seq}. To reduce drift, VidMap also evaluates direct predictions from earlier keyframes $\mathcal{K}_{i-n} \to \mathcal{K}_i$ with covariance $\Sigma_i^{(i-n)}$, rejects unsupported hops caused by visual aliasing, and keeps the supported prediction with minimum total covariance.

is below a threshold; otherwise the sequential prediction is kept. This rejects erroneous direct matches caused by repetitive texture or occlusion.

Track Filtering. We remove tracks that would harm optimization. Tracks are terminated when certainty drops below the visibility threshold. For each pair $(\mathcal{K}_{i-n}, \mathcal{K}_i)$ within the W-keyframe window, we estimate a fundamental matrix and remove tracks with large epipolar error, catching drift across depth boundaries and hallucinated correspondences through occlusions, which would otherwise introduce systematic errors in triangulation. Finally, we probabilistically thin spatially clustered tracks to maintain uniform coverage (see supplementary).

3.2 Loop Closure

Sequential tracking (Sec. 3.1) produces reliable tracks across consecutive keyframes but cannot recover constraints between temporally distant views of the same scene. Loop closure adds such constraints but may be incorrect and thus corrupt the reconstruction.

Retrieval and Matching. We extract global descriptors with MegaLoc [1] for each keyframe and retrieve the top-k candidates per query that exceed a retrieval similarity threshold, excluding pairs already covered by sequential overlap. Retrieved pairs are matched using the dense matcher at existing keypoint locations and verified through geometric verification.

Track Establishment. Standard global SfM merges all correspondences into tracks via transitivity, making sequential and LC observations indistinguishable. We instead build tracks exclusively from sequential edges. When an LC match links a keypoint that already belongs to a sequential track, we do not merge it via transitivity; instead, the correspondence is attached as a separate observation with provenance label $l_{ik} = \text{lc}$, while sequential observations carry $l_{ik} = \text{seq}$, preserving the chained sequential track. This separation enables provenance-dependent robust losses in global mapping (Sec. 3.4): sequential edges are trusted more tightly than LC edges, which may arise from visual aliasing (see Fig. 4).

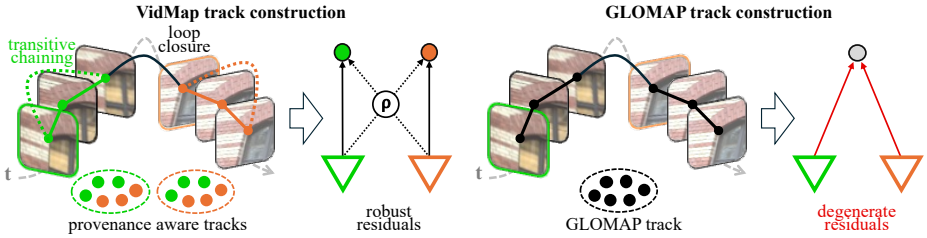

Fig. 4: Provenance Aware Track Establishment. Left: VidMap constructs tracks from sequential matching and treats loop closure (LC) as a soft link between sequential tracks, which remain reliable when the LC is incorrect. Right: GLOMAP merges LC matches by transitivity, so provenance is lost and an incorrect closure can combine separate sequential chains into an inseparable track.

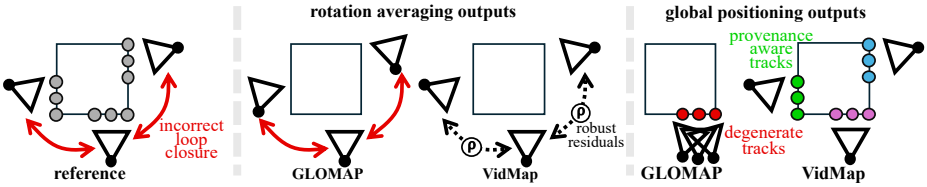

Fig. 5: Provenance Aware Rotation Averaging and Global Positioning. Under visual symmetry, GLOMAP treats incorrect loop closure (LC) constraints as ordinary edges and observations, corrupting rotation averaging and creating degenerate tracks in global positioning. VidMap preserves provenance and robustly downweights LC outliers, recovering the correct orientations and positions.

3.3 Uncalibrated Videos

When camera intrinsics are unknown, we estimate and refine focal lengths in two stages before global mapping.

Focal Length Estimation. GeoCalib [47] predicts dense perspective fields per image, from which focal length is recovered via Levenberg-Marquardt optimization. We first estimate focal length independently per keyframe, then select the top-k most confident keyframes and re-run with a shared-intrinsics constraint, jointly optimizing a single focal length across all k views.

Intrinsics Refinement. The shared intrinsics are refined via view graph calibration [34], which estimates focal lengths from fundamental matrices of well-conditioned pairs (large baselines and triangulation angles [40]). When insufficient well-conditioned pairs exist, we fall back to the estimate of GeoCalib. The refined intrinsics serve as initialization for bundle adjustment, which jointly optimizes intrinsics alongside poses and structure.

3.4 Global Mapping

We build on the GLOMAP [34] reconstruction framework, which estimates camera poses through four stages: view graph construction, rotation averaging, GP, and bundle adjustment. We extend this framework with two mechanisms that are new to global SfM. First, monocular depth priors with per-image scale estimation

in GP enable metric-scale reconstruction and regularize degeneracies inherent to video sequences where multi-view geometry alone is underconstrained. Second, provenance-dependent robust losses distinguish sequential from LC edges throughout rotation averaging and GP.

Rotation Averaging. For each keyframe pair, we estimate pairwise relative rotations $\tilde{R}_{ij}$ based on depth and correspondences [12]. Rotation averaging [34] then estimates global orientations $\{R_i\}$ by minimizing the geodesic distance $\|\log(R_j^\top \tilde{R}_{ij} R_i)\|$ over all edges. We assign provenance-dependent losses: sequential edges receive Huber loss, which trusts the tracked edges, while LC edges receive Cauchy loss, which downweights erroneous closures.

Global Positioning with depth priors. GLOMAP's GP estimates camera centers and 3D points from bearing constraints. For each observation (i,k), let m_{ik} be the monocular depth prior at the track location and σ_{ik} its uncertainty. With the rotations $\{R_i\}$ fixed from the previous step, we jointly optimize camera centers $\{c_i\}$, 3D points $\{X_k\}$, and per-image depth scale factors $\{s_i\}$, where s_i rescales the monocular depth prior for image i. The GP objective combines a bearing residual $\mathbf{e}_{ik}^{\mathrm{GP}}$, a depth residual r_{ik}^{depth} that compares the optimized depth to $s_i \cdot m_{ik}$, and a scale regularizer $r_i^{\mathrm{scale}} = \log s_i$ with scale uncertainty $\sigma_{s,i}$:

$$E_{\mathrm{GP}} = \sum_{(i,k)} \left[\rho_{l_{ik}}^{\mathrm{GP}} \left(\mathbf{e}_{ik}^{\mathrm{GP}\top} {\Sigma_{ik}^{\mathbf{v}}}^{-1} \mathbf{e}_{ik}^{\mathrm{GP}} \right) + \rho_{l_{ik}}^{\mathrm{depth}} \left(\frac{r_{ik}^{\mathrm{depth}}}{\sigma_{ik}} \right) \right] + \sum_i \rho^{\mathrm{scale}} \left(\frac{r_i^{\mathrm{scale}}}{\sigma_{s,i}} \right) \tag{1}$$

Each observation carries provenance label $l_{ik} \in \{\mathrm{seq}, \mathrm{lc}\}$ (Sec. 3.2), and the robust loss $\rho_{l_{ik}}$ is selected by provenance. We now define the residuals.

Bearing residual. This penalizes the angular discrepancy between the observed camera ray $\mathbf{v}_{ik}$ and the direction to the 3D point:

$$\mathbf{e}_{ik}^{\mathrm{GP}} = \mathbf{v}_{ik} - d_{ik}(X_k - c_i) \tag{2}$$

where $\mathbf{v}_{ik}$ is the globally rotated camera ray from c_i through observation $\mathbf{x}_{ik}$ and $d_{ik} \geq 0$ is a normalizing factor. In BA, where rotations are free, this is replaced by the standard reprojection error $\mathbf{e}_{ik}^{\mathrm{BA}} = \pi(K_i, T_i, X_k) - \mathbf{x}_{ik}$ in pixel space.

Covariance weighting. Each observation carries a 2D pixel covariance $\Sigma_{ik} \in \mathbb{R}^{2\times 2}$ from track extraction. In BA, this weights the reprojection error directly. In GP, the residual lives in 3D bearing space. We propagate Σ_{ik} through the unprojection and rotation into a 3D bearing covariance $\Sigma_{ik}^{\mathbf{v}} = J_i \, \Sigma_{ik} \, J_i^\top \in \mathbb{R}^{3\times 3}$, where J_i is the Jacobian of the camera-frame bearing with respect to pixel coordinates. We provide the full derivation in the supplementary.

Depth residual. For each observation (i,k), the camera-frame depth z_{ik} should agree with the monocular prediction m_{ik} after accounting for the per-image scale s_i. Points in front of the camera use a log-space residual; points behind the

Table 1: Evaluation on the LaMAR dataset [39]. We report the AUC of the translation error (in %, higher is better) for windows of different lengths. VidMap outperforms all existing approaches in both calibrated and uncalibrated settings.

	uncalibrated					calibrated				
W-AUC →	10 m	25 m	50 m	100 m	full	10 m	25 m	50 m	100 m	full
LoGeR	65.9	64.5	60.3	60.1	59.8	cannot use calibration				
VGGT-SLAM2	74.0	69.2	60.3	49.4	56.9	cannot use calibration				
Lingbot-Map	74.2	75.0	70.1	64.7	63.0	cannot use calibration				
DA3-Long	86.7	86.7	84.6	78.1	76.5	cannot use calibration				
DPV-SLAM	requires calibration					53.1	39.0	25.1	14.9	38.2
DROID-W	requires calibration					88.0	84.7	80.2	79.5	76.2
MP-SfM	requires calibration					88.6	83.7	74.5	60.2	60.6
GLOMAP-LG	76.4	69.4	60.1	53.6	62.3	77.3	70.4	61.6	48.4	63.0
GLOMAP-RoMA	71.0	66.0	58.4	46.7	59.4	75.7	69.0	59.2	42.1	60.1
MASt3R-SLAM	60.8	63.6	55.6	47.3	51.5	69.3	68.8	65.0	58.1	57.7
MegaSaM	77.5	68.4	59.9	51.4	52.7	78.1	69.2	58.3	52.1	55.6
ViPE	86.6	83.8	80.2	78.1	76.6	87.3	84.9	80.4	79.5	78.8
VidMap	**91.9**	**92.3**	**89.9**	**89.3**	**88.5**	**93.2**	**93.0**	**90.5**	**89.9**	**89.7**

camera fall back to linear:

$$r_{ik}^{\text{depth}} = \begin{cases} \log\left(\dfrac{z_{ik}}{s_i \cdot m_{ik}}\right) & z_{ik} > 0 \\ z_{ik} - s_i \cdot m_{ik} & z_{ik} \le 0 \end{cases} \quad \text{where} \quad z_{ik} = \mathbf{e}_3^\top R_i(X_k - c_i) \tag{3}$$

Bundle Adjustment. This refines poses $\{T_i\}$, 3D points $\{X_k\}$, intrinsics $\{K_i\}$, and per-image depth scales $\{s_i\}$. In contrast to MP-SfM [35], which fixes the depth scales after incremental reconstruction, we also optimize them:

$$E_{\text{BA}} = \sum_{(i,k)} \left[\rho^{\text{BA}}\left(\mathbf{e}_{ik}^{\text{BA}\top} \Sigma_{ik}^{-1} \mathbf{e}_{ik}^{\text{BA}}\right) + \rho^{\text{depth}}\left(\frac{r_{ik}^{\text{depth}}}{\sigma_{ik}}\right) \right] + \sum_i \rho^{\text{scale}}\left(\frac{r_i^{\text{scale}}}{\sigma_{s,i}}\right) \tag{4}$$

where $\mathbf{e}_{ik}^{\text{BA}} = \pi(K_i, T_i, X_k) - \mathbf{x}_{ik}$ is the reprojection error in pixels, weighted by the 2D pixel covariance Σ_{ik} from track extraction. By the BA stage, the reconstruction is near the correct solution and geometric verification has already filtered inconsistent loop closure observations, so we apply uniform robust losses to all observations regardless of provenance.

Robust Optimization. Inspired by graduated non-convexity [56], we progressively tighten the robust losses across iterations in GP and BA. Following MP-SfM [35], the Cauchy scale for depth residuals is annealed as the geometry stabilizes. We additionally flag depth-match inconsistencies by projecting monocular depth through the pairwise relative pose and switch the corresponding depth residuals to Cauchy loss when the depth ratio exceeds a threshold.

4 Experiments

4.1 Datasets

We evaluate the accuracy of video camera pose estimation using four datasets. All hyperparameters used throughout the system are fixed across all datasets and were tuned on separate validation sequences.

LaMAR [39] is a collection of videos recorded by phones and AR devices in three large indoor scenes (CAB, HGE, LIN), with ground truth poses and calibration. We select 50 phone videos from those in the mapping split, whose ground truth poses are publicly available. This dataset exhibits long trajectories with complex motion, visual aliasing, textureless surfaces, and degenerate viewpoints, which all stress both data association and trajectory optimization.

CroCoDL [2] is a collection of sequences recorded by phones (62) and robots (26) in 4 disaster-site buildings, with images showing unsual viewpoints and unstructured built environments and motion. This dataset is suitable for evaluating the generalization of learned approaches outside of their training distribution.

ETH3D-SLAM [41] contains 55 indoor and outdoor sequences recorded with a hand-held camera, featuring high-accuracy ground truth poses and calibration and varying scene complexity and camera motion.

EuRoC [3] contains 11 sequences captured by a drone, featuring grayscale stereo fisheye cameras and fast 6-DOF motion with blur, which are both challenging for learned computer vision models.

4.2 Baselines and Metrics

Baselines. We evaluate existing approaches that belong to three categories. *SLAM and visual odometry:* DPV-SLAM [26] is based on learned patch tracking and refinement; MegaSaM [23] and ViPE [18] combine learned optical flow with off-the-shelf monocular depth and calibration priors; DROID-W [22] also masks out dynamic objects; MASt3R-SLAM [32] relies on two-view prediction of dense pointmaps; *End-to-end deep models:* DA3-Long [10, 24] aligns overlapping subsequences reconstructed using Depth Anything 3 [24]. LoGeR [58] and LingBot-Map [6] leverage long-context mechanisms. *Global SfM:* We evaluate GLOMAP [34] with SuperPoint+LightGlue [11, 25] and SuperPoint+RoMA v2 [11, 14]. We evaluate two calibration settings: *calibrated* uses ground-truth intrinsics; *uncalibrated* estimates intrinsics from the video without prior.

Metrics. For ETH3D and EuRoC, we report the area under the recall curve (AUC, %) of the translation error after global alignment. Since the alignment is easily dominated by drift and misses fine-grained local accuracy, we proceed differently for LaMAR and CroCoDL, which exhibit longer sequences: we compute a windowed AUC (W-AUC, %) over local windows of each trajectory, in which we adapt the error threshold to 5% of the window size – see details in the supplementary material.

Table 2: Evaluation on the CroCoDL dataset [2]. We report the AUC of the translation error (in %, higher is better) for windows of different lengths. VidMap outperforms all existing approaches by a large margin, especially on the challenging robot sequences, in both calibrated and uncalibrated settings.

		Phone					Robot			
	Window-AUC →	10 m	25 m	50 m	100 m	full	10 m	25 m	50 m	full
uncalib.	DA3-Long	86.2	88.1	88.5	88.8	88.9	79.5	82.4	82.0	78.3
	VGGT-SLAM2	62.8	64.3	66.0	62.5	74.9	34.1	31.9	33.1	52.8
	Lingbot-Map	55.0	60.5	64.0	63.8	69.6	43.4	44.1	47.4	57.4
	LoGeR	50.0	56.7	57.9	63.7	63.8	47.4	55.8	62.3	65.5
	GLOMAP-LG	77.8	66.6	60.0	65.9	76.4	10.2	7.4	10.3	19.2
	GLOMAP-RoMA	66.2	57.8	53.5	56.3	74.8	12.7	10.0	10.3	20.3
	MASt3R-SLAM	51.8	57.3	60.2	65.2	70.0	25.8	30.4	32.2	46.1
	MegaSaM	50.2	50.2	43.5	35.2	50.1	41.9	39.5	32.4	45.5
	ViPE	66.7	68.2	68.6	70.7	71.4	59.5	61.4	66.2	71.0
	VidMap	**94.0**	**95.0**	**95.3**	**93.0**	**95.5**	**91.4**	**85.7**	**82.5**	**80.3**
calib.	DPV-SLAM	19.8	15.6	14.9	19.7	31.8	66.1	68.3	73.0	77.0
	DROID-W	71.8	70.4	67.9	70.6	72.7	71.9	72.7	71.0	67.8
	GLOMAP-LG	73.9	66.8	62.0	57.9	72.2	42.7	38.9	34.1	42.4
	GLOMAP-RoMA	77.3	65.2	57.2	60.6	75.3	45.9	42.5	37.7	45.0
	MASt3R-SLAM	50.1	52.6	43.8	38.6	63.6	29.8	29.5	27.1	45.4
	MegaSaM	68.1	59.8	51.0	50.3	52.3	74.5	59.8	43.0	45.7
	ViPE	70.0	71.6	71.0	71.1	74.3	62.1	64.5	68.9	72.2
	VidMap	**94.7**	**95.8**	**96.3**	**95.6**	**96.6**	**91.5**	**85.2**	**80.6**	**78.2**

4.3 Results

LaMAR – Tab. 1, Fig. 6. Long sequences exhibit geometric degeneracies, such as narrow baselines, forward motion, or textureless surfaces, that are challenging for classical approaches: GLOMAP and DPV-SLAM, which rely on multi-view geometry alone, drift by an order of magnitude more than VidMap. Approaches that leverage metric depth priors, such as ViPE and MegaSaM, are more robust to these degeneracies and maintain limited drift, but still fall short on precision. VidMap also outperforms all existing approaches when given ground-truth calibration, confirming that no single component but rather the interplay of video-aware extraction and depth-augmented global optimization closes the gap. The accuracy of VidMap is nearly identical in the uncalibrated and GT-calibrated settings, confirming that VidMap estimates accurate calibration.

CroCoDL – Tab. 2, Fig. 6. This evaluates the generalization to out-of-domain videos captured in disaster-site buildings. DA3-Long, whose depth model is trained on standard indoor/outdoor scenes, degrades substantially on this out-of-distribution data. GLOMAP's pure geometry is limited by the lack of video-aware extraction. VidMap outperforms all baselines by a large margin on both phone and robot sequences, without retraining or domain-specific adaptation.

ETH3D-SLAM and EuRoC – Tab. 3. Because these sequences are shorter and well-conditioned, multi-view geometry alone is sufficient and geometric pri-

Table 3: Evaluation on the ETH3D-SLAM and EuRoC datasets. We report AUC of the translation error (in %, higher is better) for different error thresholds.

		ETH3D-SLAM					EuRoC				
	AUC@Xm →	5 cm	10 cm	50 cm	1 m	10 m	5 cm	10 cm	50 cm	1 m	10 m
uncalibrated	GLOMAP-LG	20.4	30.9	58.7	67.4	77.8	0.4	1.8	16.1	29.7	70.7
	GLOMAP-RoMA	27.7	39.3	56.8	60.8	65.5	0.0	0.2	12.3	27.9	64.8
	LoGeR	6.4	16.9	57.7	73.8	95.9	0.1	0.5	18.2	45.1	92.6
	Lingbot-Map	9.7	22.2	63.1	77.8	97.0	0.1	0.5	14.2	33.5	83.4
	DA3-Long	19.0	37.2	73.6	84.4	98.0	0.1	0.8	31.4	57.6	95.4
	VGGT-SLAM2	26.4	44.1	75.7	84.3	97.1	1.3	7.8	55.4	74.2	96.9
	MASt3R-SLAM	8.9	21.6	65.3	77.0	96.4	1.6	5.6	51.0	71.9	97.0
	MegaSaM	38.4	53.9	79.2	86.7	98.0	0.8	5.7	50.3	73.2	97.3
	ViPE	38.7	54.9	78.3	85.7	97.6	1.3	6.2	55.4	75.0	94.0
	VidMap	**51.4**	**68.3**	**90.1**	**94.0**	**99.0**	**11.5**	**31.6**	**79.1**	**89.4**	**99.0**
calibrated	GLOMAP-LG	27.9	38.6	64.6	72.3	80.9	12.8	35.3	72.4	79.8	91.8
	GLOMAP-RoMA	37.8	47.2	59.6	63.0	67.0	18.0	43.2	79.0	85.1	90.9
	MASt3R-SLAM	22.7	37.5	70.2	78.5	96.8	5.6	18.3	68.1	83.2	98.3
	MegaSaM	53.8	62.9	80.4	87.4	98.0	15.6	35.4	84.1	92.1	99.2
	ViPE	54.4	64.9	82.2	88.4	98.2	5.5	26.4	79.9	89.9	99.0
	DROID-W	63.8	72.4	86.2	91.0	98.6	20.9	51.0	90.1	95.0	99.5
	DPV-SLAM	62.3	72.5	87.4	91.8	98.4	22.5	53.5	90.6	95.3	99.5
	VidMap	**69.4**	**79.2**	**92.2**	**95.0**	**99.1**	**23.0**	**54.6**	**90.9**	**95.4**	**99.6**

ors primarily add noise, making DPV-SLAM the strongest baseline. VidMap outperforms it by a significant margin across all thresholds on ETH3D and performs on par on EuRoC, which features grayscale fisheye cameras and motion blur, making learned depth and calibration priors are less reliable and helpful.

In the uncalibrated setting, VidMap significantly outperforms all baselines, classical and end-to-end learned, which a larger gap to the closest baselines ViPE and MegaSaM. The 17-point gap between VidMap's calibrated and uncalibrated results confirms that calibration remains the primary bottleneck for precision when the motion is easier.

4.4 Detailed Analysis

Ablation study. Table 4 measures the impact of key design decisions and shows that all components of VidMap are critical to its high robustness and accuracy.

Removing depth priors in either GP or BA incurs a drop of accuracy because these priors add geometric constraints that prevent degenerate estimates under complex motion. *No scale optimization* relies on the monocular metric scale and prevents per-image refinement. *No metric scale* discards the metric prior, optimizing the per-image scale solely based on multi-view constraints. Both variants impair the robustness on LaMAR over longer horizons. VidMap thus uses metric monocular depth as a soft prior without rigid constraint. Overall, the benefits of depth and metric scale are smaller on ETH3D-SLAM, where sequences are

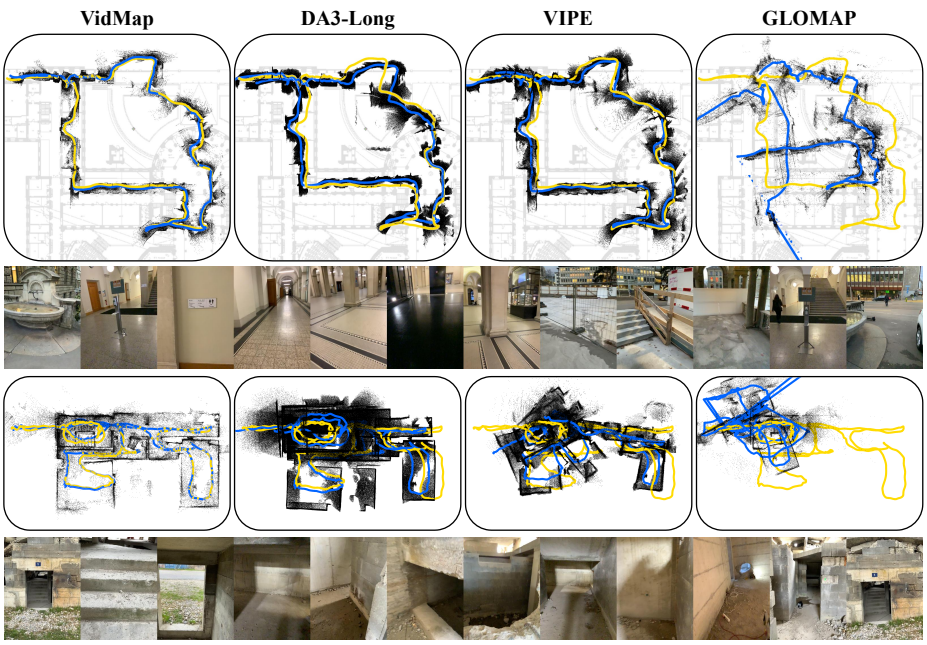

Fig. 6: Qualitative comparison to baselines on LaMAR (top) and CroCoDL (bottom). Each row compares VidMap with DA3-Long, ViPE, and GLOMAP on one scene and shows representative input images. VidMap reconstructs the scenes without drift, while baselines suffer from drift or lose track.

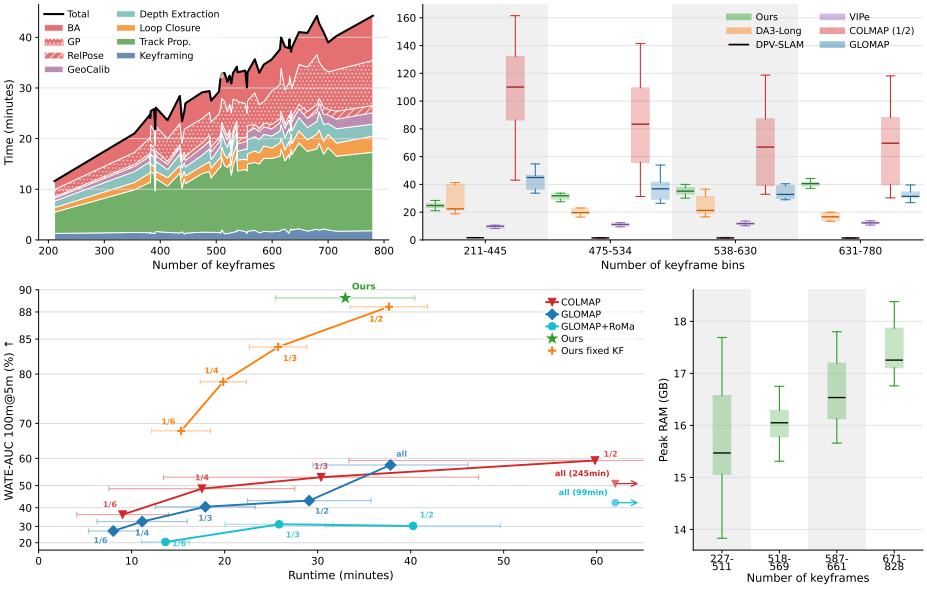

Fig. 7: Analysis of the runtime. Top left: Duration of each component with increasing number of keyframes. Matching is the most expensive step. Top right: Comparison of the total runtime against existing approaches. Bottom left: Impact of keyframing on the speed and accuracy. Motion-aware keyframing retains the accuracy of the dense keyframing but at lower cost, while aggressive keyframing reduces runtime at the cost of accuracy. Bottom right: peak RAM usage remains practical with longer videos.

Table 4: Ablation study on the LaMAR and ETH3D-SLAM datasets. Metric depth priors are most critical for long-range accuracy, while provenance-aware losses and loop closure improve robustness to incorrect multi-view observations.

	LaMAR (Window-AUC ↑)					ETH3D-SLAM (AUC ↑)				
	10 m	25 m	50 m	100 m	full	5 cm	10 cm	50 cm	1 m	10 m
VidMap	**91.9**	**92.3**	**89.9**	**89.3**	**88.5**	51.4	68.3	**90.1**	**94.0**	**99.0**
Depth priors										
No depth in GP	67.6	42.8	24.9	13.6	37.7	50.5	66.5	87.3	91.9	98.7
No depth in BA	91.5	91.1	88.0	86.5	85.5	44.5	60.6	86.0	91.9	98.8
No scale optimization	69.3	44.7	25.6	13.9	39.7	50.2	65.9	87.4	92.4	98.8
No metric scale	90.2	88.0	81.9	71.7	74.1	**52.2**	**68.8**	**90.1**	**94.0**	**99.0**
Temporal structure										
No loop closure edges	90.8	90.3	87.6	86.6	85.5	32.9	47.4	77.5	85.9	97.9
No provenance losses	88.4	85.7	80.2	77.0	79.9	50.6	66.1	87.1	91.5	98.4

shorter and motion is better constrained. Removing loop closure edges, *i.e.* relying only on sequential constraints, incurs a large drift on ETH3D-SLAM. Treating sequential and loop closure edges equally (*no provenance losses*) is prone to visual aliasing and results in catastrophic failure on LaMAR.

Runtime. We analyze the speed and memory requirements of VidMap in Fig. 7. VidMap is an offline system but can efficiently handle long videos, as time grows linearly with keyframes and the required memory remains practical. VidMap is comparably as efficient as GLOMAP and substantially faster than COLMAP.

5 Conclusion

In this paper, we introduced VidMap, a reconstruction framework that bridges the fundamental divide between the temporal awareness of causal SLAM and the robust global optimization of offline SfM. By explicitly distinguishing sequential tracking from loop-closure observations and integrating metric depth priors into global optimization, we formulate a non-causal paradigm that preserves temporal trust without prematurely committing to an incremental geometric solution. Our approach consequently estimates significantly more accurate trajectory, outperforming existing approaches by a large margin on challenging, uncalibrated datasets characterized by complex motions and severe visual aliasing.

While VidMap represents a distinct leap forward for monocular reconstruction, its geometric precision does not yet match the strict tolerances of visual-inertial systems. Mitigating long-range drift across increasingly longer sequences remains a formidable challenge, opening up exciting prospects for converting the world's abundant video streams into highly accurate 3D models for spatial computing and generative scene modeling.

Acknowledgments: This work was supported by DSO National Laboratories with grant DSOCO24035. We thank Xudong Jiang, Tobias Fischer, Shaohui Liu, and Haofei Xu for contributing to the early exploration of this topic.

References

1. Berton, G., Masone, C.: MegaLoc: One Retrieval to Place Them All. In: CVPRW (2025) 7
2. Blum, H., Mercurio, A., O'Reilly, J., Engelbracht, T., Dusmanu, M., Pollefeys, M., Bauer, Z.: CroCoDL: Cross-device Collaborative Dataset for Localization. In: CVPR. pp. 27424–27434 (June 2025) 3, 11, 12
3. Burri, M., Nikolic, J., Gohl, P., Schneider, T., Rehder, J., Omari, S., Achtelik, M.W., Siegwart, R.: The EuRoC micro aerial vehicle datasets. IJRR (2016) 3, 11
4. Cai, R., Tung, J., Wang, Q., Averbuch-Elor, H., Hariharan, B., Snavely, N.: Doppelgangers: Learning to disambiguate images of similar structures. In: ICCV (2023) 2, 4
5. Campos, C., Elvira, R., Rodríguez, J.J.G., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM3: An accurate open-source library for visual, visual-inertial, and multimap SLAM. IEEE TRO (2021) 1, 2, 3
6. Chen, L.Z., Gao, J., Chen, Y., Cheng, K.L., Sun, Y., Hu, L., Xue, N., Zhu, X., Shen, Y., Yao, Y., Xu, Y.: Geometric context transformer for streaming 3D reconstruction. arXiv preprint arXiv:2604.14141 (2026) 11
7. Chen, X., Chen, Y., Xiu, Y., Geiger, A., Chen, A.: TTT3R: 3D Reconstruction as Test-Time Training. In: ICLR (2026) 4
8. Crandall, D., Owens, A., Snavely, N., Huttenlocher, D.: Discrete-continuous optimization for large-scale structure from motion. In: CVPR (2011) 3
9. Davison, A.J., Reid, I.D., Molton, N.D., Stasse, O.: MonoSLAM: Real-time single camera SLAM. IEEE TPAMI (2007) 1
10. Deng, K., Ti, Z., Xu, J., Yang, J., Xie, J.: VGGT-Long: Chunk it, loop it, align it. arXiv:2506.00001 (2025) 4, 11
11. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. In: CVPRW (2018) 4, 11
12. Ding, Y., Kocur, V., Vávra, V., Berger Haladóvá, Z., Yang, J., Sattler, T., Kukelova, Z.: RePoseD: Efficient relative pose estimation with known depth information. In: ICCV (2025) 9
13. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In: 3DV (2025) 2, 4
14. Edstedt, J., Nordström, D., Zhang, Y., Bökman, G., Astermark, J., Larsson, V., Heyden, A., Kahl, F., Wadenbäck, M., Felsberg, M.: Roma v2: Harder better faster denser feature matching. arXiv:2511.15706 (2025) 6, 11
15. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE TPAMI (2018) 3
16. Engel, J., Schöps, T., Cremers, D.: LSD-SLAM: Large-scale direct monocular SLAM. In: ECCV (2014) 1
17. He, X., Sun, J., Wang, Y., Peng, S., Huang, Q., Bao, H., Zhou, X.: Detector-free structure from motion. In: CVPR. pp. 21594–21603 (2024) 4
18. Huang, J., Zhou, Q., Rabeti, H., Korovko, A., Ling, H., Ren, X., Shen, T., Gao, J., Slepichev, D., Lin, C.H., Ren, J., Xie, K., Biswas, J., Leal-Taixé, L., Fidler, S.: Vipe: Video pose engine for 3d geometric perception. arXiv:2508.10934 (2025) 2, 3, 11
19. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker: It is better to track together. In: ECCV (2024) 2

20. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3D with MASt3R. In: ECCV (2024) 4
21. Leutenegger, S.: OKVIS2: Realtime scalable visual-inertial SLAM with loop closure. arXiv preprint arXiv:2202.09199 (2022) 1
22. Li, M., Zhu, Z., Pollefeys, M., Barath, D.: DROID-SLAM in the wild. In: CVPR (2026) 11
23. Li, Z., Tucker, R., Cole, F., Wang, Q., Jin, L., Ye, V., Kanazawa, A., Holynski, A., Snavely, N.: MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos. In: CVPR (2025) 2, 3, 11
24. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025) 4, 6, 11
25. Lindenberger, P., Sarlin, P.E., Pollefeys, M.: LightGlue: Local Feature Matching at Light Speed. In: ICCV (2023) 2, 4, 11
26. Lipson, L., Teed, Z., Deng, J.: Deep patch visual SLAM. In: ECCV (2024) 1, 3, 11
27. Liu, S., Gao, Y., Zhang, T., Pautrat, R., Schönberger, J.L., Larsson, V., Pollefeys, M.: Robust incremental structure-from-motion with hybrid features. In: ECCV (2024) 2, 3
28. Liu, S., Chang, J., Zhu, Y., Luo, X., Chen, Y.: Depth-guided sparse structure-from-motion for movies and TV shows. In: CVPR (2022) 4
29. Maggio, D., Lim, H., Carlone, L.: VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. In: NeurIPS (2025) 3, 4
30. Moulon, P., Monasse, P., Marlet, R.: Global fusion of relative motions for robust, accurate and scalable structure from motion. In: ICCV (2013) 2, 3
31. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE TRO (2015) 1, 2
32. Murai, R., Dexheimer, E., Davison, A.J.: MASt3R-SLAM: Real-Time Dense SLAM with 3D Reconstruction Priors. In: CVPR (2025) 2, 3, 11
33. Neoral, M., Šerých, J., Matas, J.: MFT: Long-term tracking of every pixel. In: WACV (2024) 6
34. Pan, L., Barath, D., Pollefeys, M., Schönberger, J.L.: Global structure-from-motion revisited. In: ECCV (2024) 2, 3, 5, 8, 9, 11
35. Pataki, Z., Sarlin, P.E., Schönberger, J.L., Pollefeys, M.: MP-SfM: Monocular Surface Priors for Robust Structure-from-Motion. In: CVPR (2025) 2, 4, 10
36. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: UniDepth: Universal monocular metric depth estimation. In: CVPR (2024) 2
37. Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R.: Visual modeling with a hand-held camera. IJCV (2004) 2, 3
38. Sarlin, P.E., DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperGlue: Learning Feature Matching with Graph Neural Networks. In: CVPR (2020) 2
39. Sarlin, P.E., Dusmanu, M., Schönberger, J.L., Speciale, P., Gruber, L., Larsson, V., Miksik, O., Pollefeys, M.: LaMAR: Benchmarking Localization and Mapping for Augmented Reality. In: ECCV (2022) 3, 10, 11
40. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: CVPR (2016) 2, 3, 8
41. Schöps, T., Sattler, T., Pollefeys, M.: BAD SLAM: Bundle adjusted direct visual SLAM. In: CVPR (2019) 3, 11
42. Smith, C., Charatan, D., Tewari, A., Sitzmann, V.: FlowMap: High-quality camera poses, intrinsics, and depth via gradient descent. In: 3DV (2025) 4
43. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3D. In: ACM SIGGRAPH (2006) 2, 3

44. Teed, Z., Deng, J.: RAFT: Recurrent all-pairs field transforms for optical flow. In: ECCV (2020) 2
45. Teed, Z., Deng, J.: DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In: NeurIPS (2021) 1, 2, 3
46. Teed, Z., Lipson, L., Deng, J.: Deep patch visual odometry. In: NeurIPS (2023) 3
47. Veicht, A., Sarlin, P.E., Lindenberger, P., Pollefeys, M.: GeoCalib: Single-image Calibration with Geometric Optimization. In: ECCV (2024) 2, 8
48. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: VGGT: Visual geometry grounded transformer. In: CVPR (2025) 3, 4
49. Wang, J., Karaev, N., Rupprecht, C., Novotny, D.: VGGSfM: Visual geometry grounded deep structure from motion. In: CVPR (2024) 4
50. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3D perception model with persistent state. In: CVPR (2025) 4
51. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: MoGe: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: CVPR. pp. 5261–5271 (2025) 2
52. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3D vision made easy. In: CVPR (2024) 3, 4
53. Wang, Y., Zhou, J., Zhu, H., Chang, W., Zhou, Y., Li, Z., Chen, J., Pang, J., Shen, C., He, T.: π^3: Permutation-equivariant visual geometry learning. arXiv preprint arXiv:2507.13347 (2025) 4
54. Wilson, K., Snavely, N.: Robust global translations with 1DSfM. In: ECCV (2014) 2, 3
55. Xiangli, Y., Cai, R., Chen, H., Byrne, J., Snavely, N.: Doppelgangers++: Improved visual disambiguation with geometric 3D features. In: CVPR (2025) 2, 4
56. Yang, H., Antonante, P., Tzoumas, V., Carlone, L.: Graduated non-convexity for robust spatial perception: From non-minimal solvers to global outlier rejection. IEEE RAL **5**(2), 1127–1134 (2020) 10
57. Yang, J., Sax, A., Liang, K.J., Henaff, M., Tang, H., Cao, A., Chai, J., Meier, F., Feiszli, M.: Fast3R: Towards 3D reconstruction of 1000+ images in one forward pass. In: CVPR (2025) 4
58. Zhang, J., Herrmann, C., Hur, J., Sun, C., Yang, M.H., Cole, F., Darrell, T., Sun, D.: LoGeR: Long-context geometric reconstruction with hybrid memory. arXiv preprint arXiv:2603.03269 (2026) 11
59. Zhao, W., Liu, S., Guo, H., Wang, W., Liu, Y.J.: ParticleSfM: Exploiting dense point trajectories for localizing moving cameras in the wild. In: ECCV (2022) 2, 4