ON SELF-IMPROVING FEED-FORWARD SFM: A GENERALIZABLE MULTI-VIEW POSE ESTIMATOR

Anonymous authors
Paper under double-blind review

ABSTRACT

Classical Structure-from-Motion (SfM) is an optimization method: it jointly recovers camera poses, verified image correspondences, and depth (i.e., a point cloud) from an image collection without labels. Modern feed-forward SfM models instead regress these quantities in a single pass. Yet they must first be trained on labeled data, part of which is itself produced by classical SfM pipelines. Like other learned models, they fail on inputs beyond the training distribution. We close this gap by making feed-forward SfM produce its own labels: an optional bundle adjustment (BA) stage refines its predictions, and the refined output fine-tunes the network, forming a *self-improving* loop. We realize this loop for camera intrinsic and extrinsic estimation, formulating SfM as three foundation models: a frozen depth and correspondence estimator, and a tunable feed-forward pose estimator (the *SfM pose estimator*). We propose a global BA that depends entirely on model outputs: initialized with the feed-forward camera estimates, it optimizes for projective consistency between the predicted depth and correspondence maps, then extracts a subset of accurately registered images whose refined poses serve as pseudo ground truth. The pipeline scales end-to-end: from feed-forward initialization to BA, and from small scenes to $\sim$15,000-image unconstrained internet collections. Retraining on this self-generated supervision adapts the SfM pose estimator to the new domain. Across small-scene (IMC2021, ETH3D), large-scene (FastMap, T&T), and unconstrained internet (GLDv2, 1DSfM) datasets, our BA-refined poses approach the accuracy of SoTA SfM, making them reliable pseudo labels: fine-tuning on them substantially improves the feed-forward pose estimator. These results point to an emerging self-improving learning paradigm for SfM that generalizes across unseen domains.

1 INTRODUCTION

3D reconstruction is classically an optimization problem: Structure-from-Motion (SfM) (Schönberger & Frahm, 2016) first recovers the camera poses, and Multi-View Stereo (MVS) (Schönberger et al., 2016) then produces dense image correspondences and depth (a point cloud) from a collection of unlabeled images. Recently, feed-forward SfM models (Wang et al., 2025a; 2024b) learn to regress these quantities in a single pass, after training on labeled data. Unlike their classical counterparts, they are bound to their training distribution and degrade on inputs beyond it. We seek a model that improves itself, adapting to a new domain from nothing but its unlabeled images.

This work realizes that goal with a *self-improving* pipeline: bundle adjustment refines the model's predictions, and the refined output fine-tunes the model. Following the classical SfM-MVS split, we address the first stage, camera pose estimation, leaving the MVS component, i.e., depth and correspondence, frozen. Concretely, as Fig. 1 illustrates, we formulate SfM as three foundation models: a frozen depth estimator, a frozen correspondence estimator, and a tunable feed-forward pose estimator, which we call the *SfM pose estimator*. A single feed-forward pass regresses camera intrinsics and extrinsics; an optional bundle adjustment (BA) stage then refines these predictions and generates pseudo ground truth that improves the SfM pose estimator, closing the loop.

Our bundle adjustment pseudo-labeling process is guided by three design principles: it is *learning-primitive-anchored*, *scalable*, and *open-set*. First, it is learning-primitive-anchored: for a self-improving pipeline, we seek the optimal camera poses explained by the model's own depth and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

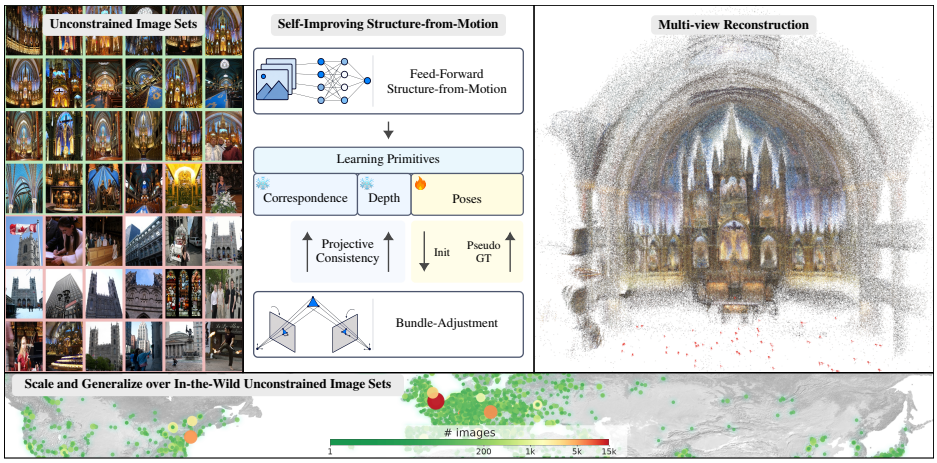


Figure 1: Overview. We study a self-improving feed-forward Structure-from-Motion (SfM): the model improves its extrinsic and intrinsic estimation by applying optimization over its own output. We assemble an SfM from a frozen foundation depth model, a frozen foundation correspondence model, and a tunable feed-forward pose estimator (the SfM pose estimator). Relying only on the frozen foundation models' depth and correspondence and the pose estimator's cameras, our bundle adjustment (BA) recovers accurate camera intrinsics and extrinsics on a subset of accurately registered images, which in turn serve as pseudo ground truth to fine-tune the SfM pose estimator. The pipeline scales to large unconstrained collections such as GLDv2 (Weyand et al., 2020) and 1DSfM (Wilson & Snavely, 2014) (bottom).

correspondence outputs, without relying on off-the-shelf SfM algorithms. Second, it is scalable: jointly registering more overlapping views lengthens feature tracks and strengthens triangulation and bundle adjustment (Schönberger & Frahm, 2016; Resch et al., 2015). Third, it is open-set: as in classical pipelines such as COLMAP (Schönberger & Frahm, 2016), not all images are registered, and we retain only the accurately registered subset to serve as pseudo ground truth. These properties let our system adapt to unseen static image collections, and they move the supervision ceiling from pose annotation to the frozen depth and correspondence models, which keep improving independently.

To register poses that make the predicted depth and correspondence projectively consistent, we build on the recent Marginalized Bundle Adjustment (MBA) (Zhu et al., 2026), which optimizes camera intrinsics, extrinsics, and a low-degree-of-freedom per-view depth affine adjustment. We further extend MBA with free-point refinement: whereas MBA can only scale and shift each depth map as a whole, this stage lets every 3D point move freely to best fit the multi-view geometry. We also improve MBA's first-order optimizer to a second-order one: reading its CDF objective as an M-estimator yields a reweighted least-squares surrogate, enabling Gauss–Newton for improved convergence efficiency.

Our pipeline runs end-to-end at large scale, enabling single-pass inference over up to $\sim$15,000 images on a single 8$\times$A100-80GB node. Prior works approximate the global attention via chunking or sparsification (Deng et al., 2026c; Pan et al., 2026; Deng et al., 2026b; Shen et al., 2026); we instead compute it exactly. Adapting context-parallel flash attention to image-sharded multi-view inference, we divide the images evenly across GPUs. Total memory then grows only linearly with the number of images, and capacity scales with the device count, across nodes as well as within one.

In practice, BA cannot register every input image accurately: some views have insufficient overlap, unreliable depth or correspondence, or poorly initialized poses, and are thus best excluded. We therefore generalize MBA from closed-set inputs to open-set image registration, combining a false-positive-aware correspondence estimator (Doppelgangers++ (Cai et al., 2023)) with a pose-graph filtering strategy to isolate a sub-collection of jointly registered images. Their poses reach the accuracy of SoTA SfM, allowing our method to operate on genuinely unconstrained internet collections such as GLDv2 (Weyand et al., 2020) and 1DSfM (Wilson & Snavely, 2014), which may contain images entirely unrelated to the scene being reconstructed.

Our self-improving pipeline is distinct from self-supervised learning: it cannot be reduced to a single differentiable loss over a few adjacent frames, as in self-supervised monocular depth estimation (Zhou et al., 2017; Godard et al., 2019). Preparing pseudo ground truth outside backpropagation, through a

full BA pipeline, yields a cleaner supervision signal than a local graph window with a single loss update, letting the SfM pose estimator adapt to in-the-wild domains for which no annotation exists.

Our pipeline is itself an SfM that reflects the robustness of the generalizable depth and correspondence predictions: users may stop after the fast feed-forward pass or invoke BA for higher accuracy, and after fine-tuning on self-generated labels the pose estimator closes much of the gap to BA-refined quality. Across small-scene (Jin et al., 2021; Schöps et al., 2017), large-scene (Li et al., 2026; Knapitsch et al., 2017), and unconstrained internet (Weyand et al., 2020; Wilson & Snavely, 2014) datasets, our BA-refined poses approach SoTA SfM accuracy, making them reliable pseudo labels that substantially improve our feed-forward pose accuracy.

Our contributions are summarized as follows:

1. A complete self-improving pipeline that adapts a feed-forward SfM pose estimator to unseen domains, spanning video and unordered internet collections, from small to very large scale.
2. A learning-primitive-anchored SfM whose BA optimizes for projective consistency across the model's own outputs, without relying on off-the-shelf SfM.
3. Multi-view pose estimation that outperforms prior feed-forward methods across diverse benchmarks and scales.

2 Related Work

Global Structure-from-Motion. Incremental SfM registers images, triangulates, and bundle-adjusts in sequence (Snavely et al., 2006; Schönberger & Frahm, 2016); global SfM solves all cameras at once, trading robustness for speed through approximations such as motion averaging (Wilson & Snavely, 2014; Pan et al., 2024; Li et al., 2026; Pan et al., 2026). Feed-forward networks (Wang et al., 2025a; 2024a) are conceptually global too, reasoning over all views in one pass, but scale poorly. Recent work therefore sparsifies their global attention, compresses its state, or splits the view graph (Shen et al., 2026; Elflein et al., 2026; Deng et al., 2026b), at the cost of fidelity or a sequential bottleneck. Ours stays global in both inference and bundle adjustment, using full attention and exact projective geometry, and scales to thousands of images across GPUs.

Feed-Forward Structure-from-Motion. DUSt3R (Wang et al., 2024b) introduced feed-forward stereo reconstruction by pointmap regression, with an optimization-based global alignment lifting its pairwise predictions to multi-view, later made scalable and integrated into SfM (Leroy et al., 2024; Duisterhof et al., 2025) and extended to online video by spatial-memory and streaming variants (Wang & Agapito, 2025; Zhuo et al., 2026; Chen et al., 2026a). Single-stage multi-view networks bypass that alignment, exchanging information across views in one pass (Tang et al., 2025), though for tens of views rather than the thousands we target. VGGT (Wang et al., 2025a) regresses cameras, depth, and tracks for hundreds of views in one pass, and the line has since scaled along architecture, prediction target, data, and modality (Yang et al., 2025b; Wang et al., 2026b; Lin et al., 2026; Wang et al., 2026a; Keetha et al., 2026; Elflein et al., 2025). Out of distribution these models still fail (Deng et al., 2026b), and repairing them by fine-tuning has historically needed a classical pipeline for the training poses. Ours derives pseudo ground truth from the model's own predictions through an initialize–refine–filter–retrain loop, widening the training data.

Self-Supervised and Weakly-Supervised SfM. Because pose labels are costly, self- and weakly-supervised methods draw supervision from model and pipeline outputs directly. Zhou et al. (2017) and Godard et al. (2019) couple a depth and a pose network through photometric reprojection over nearby frames, a signal that feed-forward successors carry to unlabeled video (Wimbauer et al., 2025; Jiang et al., 2025; Hariat et al., 2026; Smith et al., 2025). Chen et al. (2026b) supervise with monocular depth and sparse correspondences from pretrained predictors, Wang et al. (2026a) with a momentum-averaged teacher, and Izquierdo & Civera (2023) and Brachmann et al. (2024) with an existing reconstruction. Our method instead produces higher-quality supervision by refining those outputs with a global bundle adjustment over a covisibility pose graph spanning the whole collection.

Optimization on Top of Feed-Forward Predictions. Geometric optimization reliably improves a feed-forward prediction, and our loop builds on that. VGGT (Wang et al., 2025a) gains substantially from bundle adjustment over its own point maps, and later systems couple the two more tightly through virtual tracks, dense warps, or direct optimization against monocular depth (Pan et al., 2026; Deng et al., 2026a; Zhu et al., 2026; Wang & Agapito, 2026); others optimize cameras jointly with



Figure 2: Self-improving SfM. Frozen depth and correspondence networks complement a trainable feed-forward pose estimator; bundle adjustment and outlier view filtering turn their predictions into pseudo ground truth for fine-tuning. The mosaic marks inlier (green) and outlier (red) views; the retained cameras yield the reconstruction.

a rendered scene (Lin et al., 2021; Bian et al., 2023; Fu et al., 2024). In each, the refinement is the deliverable: consumed at test time, then discarded. Selfi (Deng et al., 2026d) comes closest to reusing it, but trains a feature adapter for view synthesis, so its refined poses never supervise the pose estimator. We close the loop, training the pose estimator on the optimized poses so a collection is refined once rather than every run.

3 Method

We close the generalization gap of feed-forward SfM identified in Sec. 1 with a *self-improving* loop that adapts the SfM pose estimator to unconstrained target domains. A complete feed-forward SfM comprises three components: camera pose, structure (i.e., depth), and correspondence estimation. We self-improve only the pose component, leaving the depth and correspondence estimators frozen.

3.1 System Overview

Our system takes a collection of N RGB images $\mathcal{I} = \{\mathbf{I}_i\}_{i=1}^{N}$ and outputs their camera parameters together with the depth adjustments that place the monocular depth maps in a common coordinate frame. Fig. 2 shows the three stages of the self-improving loop. The SfM pose estimator $\mathcal{F}_\theta$ regresses all cameras in a single pass as an estimate $\mathcal{X}$; bundle adjustment refines $\mathcal{X}$ into pseudo ground truth $\mathcal{X}^*$, and fine-tuning then trains on $\mathcal{X}^*$ together with the annotated data,

$$\mathcal{X}^* = \mathrm{BA}\big(\mathcal{F}_{\theta^\circ}(\mathcal{I})\big), \qquad \theta^\dagger = \arg\min_\theta \; \mathcal{L}\big(\theta;\; \mathcal{D}_{\text{ann}} \cup \{(\mathcal{I},\, \mathcal{X}^*)\}\big), \tag{1}$$

where $\mathcal{D}_{\text{ann}}$ is the annotated training set, θ° the weights trained on it alone, and $\theta^\dagger$ those obtained by training jointly on $\mathcal{D}_{\text{ann}}$ and the self-generated labels. The union lists the two sources of supervision; Eq. (5) sets their relative weight. BA always labels with the strongest checkpoint available, which here is θ°. Sec. 3.2 develops the pose estimator, which reads the images with their monocular depth maps and predicts each camera in one pass. Starting from those predictions, the bundle adjustment of Sec. 3.3 jointly refines the cameras and the depth alignment and discards frames it cannot register reliably, leaving a refined estimate $\mathcal{X}^*$ over a retained subset $\mathcal{V}^*$. This matters most when the estimator generalizes poorly, e.g., on internet collections with many views unrelated to the scene being reconstructed. Sec. 3.4 treats $\mathcal{X}^*$ as pseudo ground truth for fine-tuning.

3.2 Feed-Forward SfM

Architecture. We build $\mathcal{F}_\theta$ on the π^3 architecture (Wang et al., 2026b), diagrammed in Fig. 3, which additionally consumes the monocular depth map $\mathbf{D}_i = \mathcal{N}_\mathbf{D}(\mathbf{I}_i)$ from an off-the-shelf estimator $\mathcal{N}_\mathbf{D}$, so it can reason about metric structure and predict the depth scale used above. We attach a camera head that regresses, per frame, the extrinsics $\mathbf{P}_i$, the field-of-view fixing the intrinsics $\mathbf{K}_i$, and the scale s_i resolving the depth map's scale ambiguity; App. E.1 gives the parameterization. We omit both dense heads: the point-map head, since per-frame depth with camera poses already recovers a consistent point cloud (Lin et al., 2026), and the depth head, whose role the frozen monocular estimator assumes. $\mathcal{F}_\theta$ thus regresses only low-DoF camera parameters, allowing a lower input resolution that aids scaling.

Context-Parallel Multi-View Inference. The π^3 aggregator alternates between frame-local attention and global attention over all $N{\cdot}P$ tokens. Prior methods scale this global attention by approximating it,

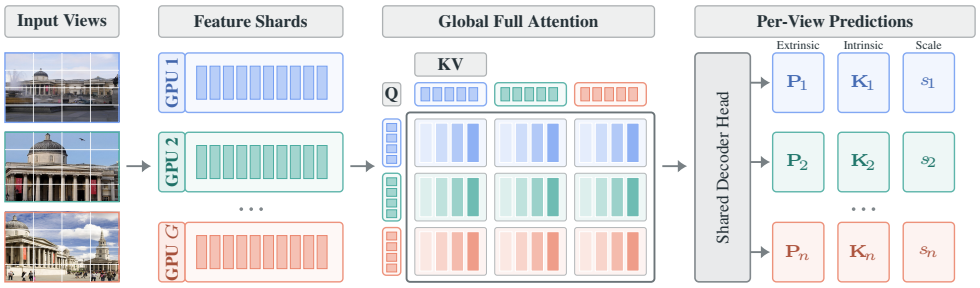


Figure 3: Context-parallel multi-view inference. Images are partitioned evenly across GPUs and encoded into feature shards. Each GPU keeps its local queries and streams all key–value shards, evaluating exact global attention block by block without materializing the dense score matrix. A shared decoder predicts per-view extrinsics, intrinsics, and scale.

through view chunking or sparsification (Deng et al., 2026c; Pan et al., 2026; Deng et al., 2026b; Shen et al., 2026); we keep it exact and distribute it across GPUs, as Fig. 3 illustrates. The online-softmax statistics of FlashAttention (Dao et al., 2022; Dao, 2024) merge associatively and commutatively over disjoint key–value blocks. We therefore adapt context-parallel attention (Liu et al., 2024; Li et al., 2024; Yang et al., 2025a) to non-causal, image-sharded multi-view inference, derived in full in App. E.3. The N images are split into G equal shards, one per GPU, and each device evaluates its local query rows by streaming the peer key–value blocks and merging each into its running state. In exact arithmetic the result equals dense full attention, with no materialized $NP \times NP$ score matrix. Each GPU holds only $\mathcal{O}(NP/G)$ activations, linear in the number of images rather than $\mathcal{O}(N^2P^2)$. Capacity in image count therefore grows with the device count and is not bounded by one node. Omitting both dense heads further lets us run at a lower input resolution. Together, these let us train a 3B-parameter model on up to 256 views and, at test time, run a single joint pass over ∼15,000 images within one 8×A100-80GB node, a scale that previously required chunking or sparsification.

3.3 Bundle Adjustment

Besides the camera predictions, BA consumes monocular depth maps $\mathcal{D} = \{\mathbf{D}_i\}$ and pairwise correspondences $\mathcal{C} = \{\mathbf{C}_{i,j}\}$ from a pretrained matcher $\mathcal{N}_\mathbf{C}$. It refines all variables by maximizing a score over the edges of a sparse covisibility pose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, built as in Fig. 4 (left), $\mathcal{X}^* = \arg\max_\mathcal{X} \sum_{(i,j)\in\mathcal{G}} \mathcal{H}(\mathbf{X}_i, \mathbf{X}_j \mid \mathbf{D}_i, \mathbf{C}_{i,j})$. Here $\mathbf{X}_i = (\mathbf{P}_i, \mathbf{K}_i, s_i, b_i)$ collects frame i's parameters, and $\mathcal{H}$ is the score of Eq. (2) restricted to that edge's correspondences. We construct the graph as described in App. E.2. We adopt the MBA objective (Zhu et al., 2026) as our optimization core. Backprojecting each correspondence with the affine-adjusted monocular depth and reprojecting it into the other view gives a projective residual, and the area under the residual CDF scores a solution,

$$\mathcal{Q}(\mathcal{X}) = \frac{1}{\bar{\tau}_{\max}} \int_0^{\bar{\tau}_{\max}} F(\bar{r})\, d\bar{r} \;\in [0,1], \tag{2}$$

$\mathcal{Q}(\mathcal{X}) \approx 1$ when most correspondences are well explained by the poses and structure, and its complement $\bar{\mathcal{Q}} = 1 - \mathcal{Q}$ measures the fraction that are not. App. E.4 details the residuals and the surrogate loss. The components below make this objective more robust on unconstrained collections, more accurate, and more efficient at scale.

Fast Scale Initialization. The per-frame depth scale $\{s_i\}$ sets the metric extent of each depth map, and the projective objective is highly sensitive to it, so we re-initialize it rather than trusting the network's coarse estimate. Conditioned on the current relative pose, each correspondence triangulates to a depth in closed form, and its ratio to the monocular depth at the same pixel is a candidate scale $\hat{s}_{i,k}$, derived in App. E.6. We draw M candidates, score each by Eq. (2) over all correspondences of frame i, and keep the best,

$$s_i^* = \hat{s}_{i,m^*}, \quad m^* = \arg\min_{m\in[M]} \bar{\mathcal{Q}}_i(\hat{s}_{i,m}). \tag{3}$$

Every candidate is closed-form, so the initialization parallelizes across frames. Global scale is a free gauge, so we anchor it to the metric depth maps: dividing the depth scales, biases, and translations by the median scale leaves $\mathrm{median}_i\, s_i = 1$, keeping each depth map near its metric prediction.

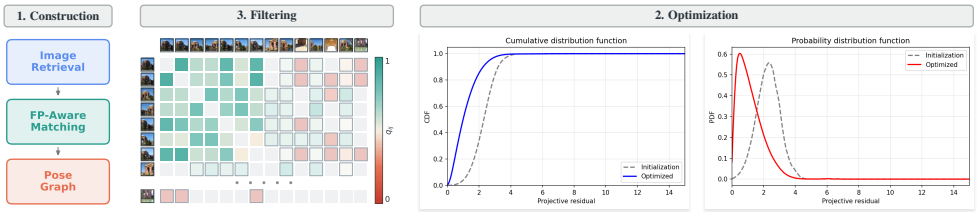


Figure 4: The three bundle adjustment stages, numbered in procedural order. Construction (left): retrieval proposes candidate pairs and false-positive-aware matching keeps the verified correspondences whose edges form the pose graph. Optimization (right): the projective-residual CDF and PDF, before and after optimization. Filtering (centre), which runs afterwards: edges are pruned by quality q_{ij} and the largest reliable component is retained. Views are ordered inliers first: preserved pairs are filled by q_{ij}, pruned pairs carry a faint fill, and gray pairs never entered the optimization.

Outlier View Filtering. Classical pipelines filter inconsistent geometry and observations; we apply the same principle at the view level, removing frames whose incident edges are unreliable while preserving a connected reconstruction (Schönberger & Frahm, 2016; Pan et al., 2024). The matcher $\mathcal{N}_{\mathbf{C}}$ is Doppelgangers++ (Cai et al., 2023), a MASt3R-based (Leroy et al., 2024) model whose scene-level verification rejects pairs that look alike but image distinct surfaces. Because the correspondences can then be trusted, a large projective residual indicates poor geometry rather than a bad match, so an edge's residuals measure how well its two views are registered. For each edge we define an edge score $q_{i,j} = \mathcal{Q}(\{\bar{r}_{i,j,k}\}_{k\in[\kappa]})$, the CDF-area score restricted to it, and call the edge an *inlier* if $q_{i,j} > \tau$. Letting $\mathcal{E}_n$ be the edges induced by the surviving nodes $\mathcal{V}_n$, we greedily remove the worst-connected node,

$$v_n = \arg\min_{v\in\mathcal{V}_n} \min_{(v,j)\in\mathcal{E}_n} q_{v,j}, \qquad \mathcal{V}_{n+1} = \mathcal{V}_n \setminus \{v_n\}, \tag{4}$$

recompute the affected edge scores, and iterate until every surviving edge is an inlier, returning the largest connected component as the retained subset $\mathcal{V}^*$. Fig. 4 (centre) shows the resulting connectivity, and its right panel the residual distributions that the edge scores summarize.

Free-Point Refinement. The stages above are *depth-anchored*: each 3D point is tied to its source pixel through a per-frame depth affine map, so it slides along the viewing ray via s_i and b_i but cannot leave it. Free-point refinement removes this constraint, bringing the pipeline closer to a classical BA formulation. Because it helps only once the poses are accurate, we apply it *after* filtering, over $\mathcal{V}^*$. Every retained correspondence is promoted to a free world-frame variable $\mathbf{Y}_k \in \mathbb{R}^3$, distinct from the camera-frame point $\mathbf{X}_k$ of Eq. (E.2), initialized at its depth-anchored value and optimized jointly with the cameras. We minimize the reprojection CDF loss in native pixels on the symmetric residual $r_k = \frac{1}{2}(\|\pi_i(\mathbf{Y}_k) - p_k\|_2 + \|\pi_j(\mathbf{Y}_k) - q_k\|_2)$, dropping the depth-prior term so each point moves to best explain the multi-view correspondences. Freeing the points can degrade a well-registered frame, so we re-apply the pruning of Eq. (4) to the refined result.

Second-Order Optimization. Eq. (2) is not a sum of squares, so Gauss–Newton does not apply directly. Viewing its surrogate loss as an M-estimator does: with the CDF derivative supplying a per-correspondence weight w_k, the reweighted least squares $\frac{1}{2}\sum_k w_k \|\mathbf{e}_k\|^2$ has, at fixed weights, exactly the gradient of that surrogate, so it shares its stationary points while additionally exposing curvature, as derived in App. E.5. We therefore solve each stage with Levenberg–Marquardt, optimizing rotations in the $\mathfrak{se}(3)$ tangent. Free-point refinement introduces a point block, which we eliminate by Schur complement.

3.4 Training

We realize the fine-tuning step of Eq. (1) as an *offline* pseudo-labeling process rather than an in-optimizer objective. For each target-domain collection we run the full BA pipeline ahead of training, caching the pseudo ground truth $\mathcal{X}^*$ over $\mathcal{V}^*$ as the self-generated set $\mathcal{D}_{\text{self}} = \{(\mathcal{I}, \mathcal{X}^*)\}$. Expanding the joint objective of Eq. (1), we weight the two sources against each other:

$$\theta^\dagger = \arg\min_{\theta} \; \mathbb{E}_{\mathcal{D}_{\text{ann}}}\big[\mathcal{L}_{\text{sup}}\big] + \beta\, \mathbb{E}_{\mathcal{D}_{\text{self}}}\big[\mathcal{L}_{\text{sup}}\big], \tag{5}$$

where β balances the two sources and the supervised loss $\mathcal{L}_{\text{sup}}$ scores predictions against pseudo targets exactly as against annotated ones. $\mathcal{L}_{\text{sup}}$ sums L_1 errors on pairwise relative rotation, relative

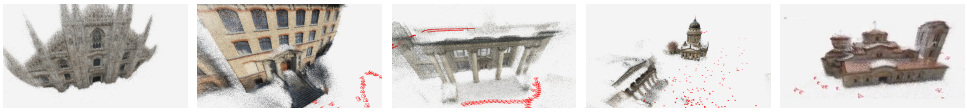


Figure 5: Qualitative multi-view reconstructions. One scene per benchmark. Each panel renders the point cloud recovered by our self-improving SfM together with the estimated camera frusta, drawn in red; per-dataset galleries covering all scenes are in App. F.

translation, field-of-view, and depth scale over the camera head's iterative predictions. Our training collections do not share a scale: some provide metric ground truth, while SfM-derived ones such as MegaDepth (Li & Snavely, 2018) are defined only up to scale, and neither convention need match the frozen monocular estimator. We therefore adopt the estimator's scale as the common reference, deriving a per-frame depth scale with the estimation of Eq. (3) and rescaling the target poses to match, as App. E.7 details. Preparing $\mathcal{X}^*$ outside of backpropagation lets the model benefit from the full BA pipeline, which is hard to differentiate through, while the annotated term guards against catastrophic forgetting. All hyperparameters, together with the network architecture and fine-tuning schedule, are in App. D.

4 Experiments

We compare across four SfM benchmarks covering two scene-scale regimes: small-scene (ETH3D (Schöps et al., 2017), IMC2021 (Jin et al., 2021)) and large-scene (Tanks & Temples (Knapitsch et al., 2017), FastMap (Li et al., 2026)); Fig. 5 shows qualitative reconstructions on all four. We compare against representative baselines spanning classical SfM, learning-based SfM, and feed-forward methods. On each benchmark we follow the evaluation protocol of prior work and quote its baseline numbers: Deng et al. (2026a) for ETH3D, Aung et al. (2026) for IMC2021, Deng et al. (2026b) for Tanks & Temples, and Li et al. (2026) for FastMap. Throughout the tables, Ours$^\circ$ denotes our feed-forward model before target-domain pseudo-labeled fine-tuning and Ours† the same model after it, carrying weights θ° and $\theta^\dagger$ of Eq. (1). Appending $+$BA, as in Ours$^\circ$+BA, applies our bundle adjustment at test time. Training datasets, configuration, and further implementation details are in App. D. Both feed-forward configurations register every image and are scored on all frames, whereas the $+$BA configurations are scored on the subset $\mathcal{V}^*$ retained after the view filtering of Sec. 3.3, as in classical SfM, which measures accuracy on registered images and reports completeness separately. Since this favors the filtered setting, Tab. C.1 reports every configuration under both, together with the retained fraction. Our fine-tuning pool contains the test images themselves, though never their ground-truth poses, so Ours† is a transductive result; App. C reports a leave-one-domain-out control (Ours$^\ominus$) that excludes each evaluation domain from the pool entirely. Training one model per domain is resource-heavy, so these controls use a reduced schedule and understate inductive accuracy; Ours† remains our main result.

Benchmark on ETH3D. ETH3D (Schöps et al., 2017) contains 13 unordered, high-resolution indoor and outdoor scenes with precisely calibrated ground-truth camera poses. We report relative rotation accuracy (RRA) and relative translation accuracy (RTA) at the 5° and 1° thresholds. Tab. 1 presents results, with AUC in Tab. A.1. Our pseudo-labeled fine-tuning improves generalization to unseen test scenes, raising average RRA from 87.3 (Ours$^\circ$) to 100.0 (Ours†) and RTA from 60.4 to 88.5. This makes Ours† the strongest purely feed-forward entry in Tab. 1, ahead of AMB3R (Wang & Agapito, 2026) (98.2 RRA, 81.9 RTA). Applying BA at test time further refines the predictions, achieving 100.0 RRA and 98.6 RTA on average over the retained subset (Ours$^\circ$+BA), demonstrating that BA consistently benefits feed-forward SfM models. The inductive control Ours$^\ominus$ also retains the feed-forward lead, as reported in Tab. C.1 of the appendix.

Benchmark on IMC2021. We evaluate pose accuracy on IMC2021 (Jin et al., 2021) using the AUC of pose error at 3°, 5°, 10°. Among purely feed-forward models, Ours† attains the highest AUC at every threshold (49.55/63.68/78.26) in Tab. 2, ahead of DA3 (Lin et al., 2026) (44.15/58.54/74.21), Pi3 (Wang et al., 2026b), and VGGT (Wang et al., 2025a). Our BA-refined model (Ours$^\circ$+BA) reaches 55.28/68.03/80.86 AUC@3/5/10, substantially improving over the feed-forward initialization. The inductive control Ours$^\ominus$ preserves the feed-forward ordering, as reported in Tab. C.1 of the appendix.

Table 1: SfM on ETH3D. RRA/RTA at the 5° (left) and 1° (right) thresholds. **Key:** method-with-optimization / method-with-feed-forward; best, 2nd, 3rd per metric.

	Threshold @5°																		Threshold @1°					
	COLMAP CVPR'16		VGGSfM CVPR'24		DF-SfM CVPR'24		MASt3R-SfM 3DV'25		MBA 3DV'26		AMB3R CVPR'26		Glob3R arXiv'26		Ours°		Ours†		Ours°+BA		AMB3R CVPR'26		Glob3R arXiv'26		Ours°+BA	
Scenes	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA	RRA	RTA
courtyard	56.3	60.0	50.5	51.2	80.7	74.8	89.8	64.4	94.7	94.7	100.0	96.5	100.0	97.0	100.0	84.9	100.0	97.2	100.0	99.6	86.6	50.3	100.0	73.9	100.0	63.1
delivery	34.0	28.1	22.0	19.6	82.5	82.0	83.1	81.8	83.1	83.0	91.0	76.6	100.0	81.9	100.0	75.8	100.0	95.8	100.0	100.0	69.6	15.7	83.4	66.5	100.0	92.4
electro	53.3	48.5	79.9	58.6	82.8	81.2	100.0	95.5	95.6	78.2	95.6	81.2	100.0	85.3	83.1	75.2	100.0	85.8	100.0	99.3	80.1	36.7	72.1	56.0	100.0	91.5
facade	92.2	90.0	57.5	48.7	80.9	82.6	74.3	75.3	100.0	99.2	100.0	95.4	100.0	96.8	92.3	83.5	100.0	91.9	100.0	99.9	86.8	50.1	94.3	81.9	100.0	98.6
kicker	87.3	86.2	100.0	97.8	93.5	91.0	100.0	100.0	100.0	98.9	100.0	99.2	100.0	95.7	100.0	95.1	100.0	99.0	100.0	99.1	95.4	80.5	94.4	86.5	100.0	84.3
meadow	0.9	0.9	100.0	96.2	56.2	58.1	58.1	58.1	100.0	58.1	100.0	95.2	100.0	92.0	81.0	32.4	100.0	98.1	100.0	100.0	80.4	40.0	78.7	68.4	100.0	93.1
office	36.9	32.3	64.9	42.1	71.1	54.5	100.0	98.5	100.0	86.2	100.0	53.9	100.0	63.9	100.0	54.6	100.0	52.6	100.0	87.6	92.3	18.8	77.8	33.4	91.5	41.2
pipes	30.8	28.6	100.0	97.8	72.5	61.5	100.0	100.0	100.0	96.7	100.0	87.9	100.0	92.9	100.0	59.9	100.0	92.3	100.0	100.0	67.4	30.6	100.0	73.5	100.0	86.7
playground	17.2	18.1	37.3	40.8	70.5	70.1	100.0	93.6	94.7	93.8	98.7	62.2	100.0	95.4	99.9	43.0	100.0	84.1	100.0	98.2	69.9	19.7	90.0	68.0	100.0	88.2
relief	16.8	16.8	59.6	57.9	32.9	32.9	34.2	40.2	100.0	98.9	100.0	90.1	100.0	96.6	36.1	32.7	100.0	88.2	100.0	98.5	68.2	30.7	100.0	86.5	100.0	58.5
relief 2	11.8	11.8	69.9	70.3	40.9	39.1	57.4	76.1	100.0	98.9	100.0	75.7	100.0	95.9	42.4	26.0	100.0	76.2	100.0	100.0	45.5	8.6	100.0	73.5	100.0	90.0
terrace	100.0	100.0	38.7	29.6	100.0	99.6	100.0	100.0	100.0	100.0	100.0	97.2	100.0	95.6	100.0	83.2	100.0	96.8	100.0	100.0	90.2	61.4	100.0	92.8	100.0	95.7
terrains	100.0	99.5	70.4	54.9	100.0	91.9	58.2	52.5	100.0	95.4	91.6	53.8	100.0	97.3	100.0	39.4	100.0	92.2	100.0	99.5	37.6	9.4	100.0	90.5	96.9	72.4
Average	49.0	47.8	65.4	58.9	74.2	70.7	81.2	79.7	97.5	90.9	98.2	81.9	100.0	91.3	87.3	60.4	100.0	88.5	100.0	98.6	74.6	34.8	91.6	73.2	99.1	81.2

Table 2: Pose estimation on IMC2021 in AUC of pose error (@3°/5°/10°). **Key:** method-with-optimization / method-with-feed-forward; best, 2nd, 3rd per metric.

	COLMAP CVPR'16	GLOMAP ECCV'24	PixSfM ICCV'21	MASt3R-SfM 3DV'25	DF-SfM CVPR'24	Dense-SfM CVPR'25	MV-RoMa CVPR'26	VGGSfM CVPR'24	VGGT CVPR'25	VGGT+BA CVPR'25	Pi3 ICLR'26	Pi3+BA ICLR'26	DA3 ICLR'26	DGSfM ACCV'26	MBA 3DV'26	Ours°	Ours†	Ours°+BA
AUC@3	44.58	45.91	25.54	31.77	47.43	48.48	51.31	45.23	41.15	44.96	43.34	55.30	44.15	50.66	47.29	30.88	49.55	55.28
AUC@5	56.18	58.29	34.80	46.36	59.84	60.79	62.92	58.89	55.52	58.08	57.57	66.43	58.54	63.03	60.37	45.99	63.68	68.03
AUC@10	68.73	71.81	46.73	64.37	73.19	73.90	75.92	73.92	71.73	73.52	73.35	78.67	74.21	76.30	74.59	64.98	78.26	80.86

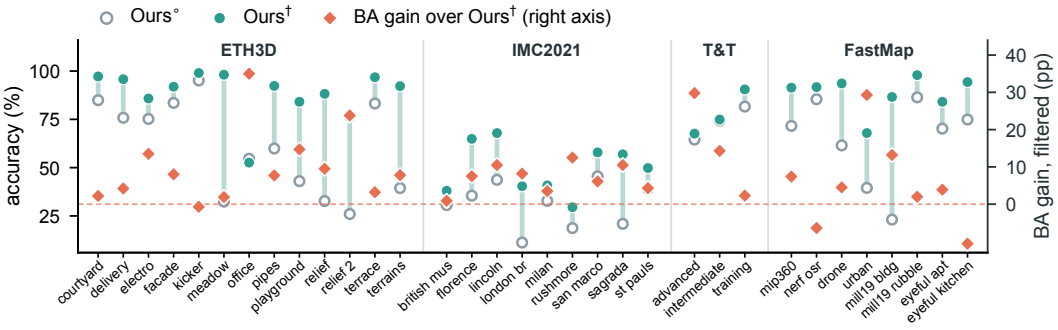


Figure 6: Self-improving SfM ablation. Left axis: the benefit of self-improving (Ours° → Ours†). Right axis: whether BA is still needed after fine-tuning (coral diamonds, dashed line at zero). Metrics are RTA@5 (ETH3D, T&T), AUC@3 (IMC2021), RTA@3 (FastMap).

Benchmark on Tanks & Temples. Tanks & Temples (Knapitsch et al., 2017) is a large-scale SfM benchmark with 21 scenes of approximately 300 images each, organized into Training (7), Intermediate (8), and Advanced (6) splits. Tab. 3 presents overall results averaged across all 21 scenes. Our feed-forward Ours° is competitive with SAIL (Deng et al., 2026b) on RRA/RTA, while already yielding the lowest ATE among prior feed-forward methods. Pseudo-labeled fine-tuning improves it further to 71.7 RRA@5 and 78.0 RTA@5 with an ATE of 0.005, the best of the feed-forward methods compared here, and BA refinement brings the poses to 92.6/90.6 with an ATE of 0.004.

Benchmark on FastMap. The FastMap benchmark (Li et al., 2026) covers eight large-scale datasets of roughly 200 to 6,000 images, on which Tab. 4 compares us against FastMap, GLOMAP (Pan et al., 2024), and COLMAP (Schönberger & Frahm, 2016) in ATE, AUC@3, and RTA@3. Every baseline here is optimization-based; to our knowledge Ours† is the first feed-forward method reported at this scale. Most splits take a COLMAP reconstruction as ground truth, which may mildly favor these three pipelines as they share components. On the three splits that do not, Ours† attains a lower ATE than every classical baseline and leads on all three metrics on the Eyeful Tower kitchen scene, suggesting part of the gap elsewhere reflects the ground-truth source. Applying BA to our feed-forward predictions improves substantially over Ours°, with the largest gains on the more challenging large-scale scenes.

Table 3: SfM on Tanks & Temples. RRA/RTA (@5°) and ATE averaged across all 21 scenes, except **DF-SfM** and **FlowMap**, which **SAIL** reports over the 76.2% and 66.7% of scenes they register. **Key:** method-with-optimization / method-with-feed-forward; best, 2nd, 3rd per metric.

	DROID NeurIPS'21	GLOMAP ECCV'24	ACE-0 ECCV'24	Spann3R 3DV'25	MASt3R-SfM 3DV'25	DF-SfM CVPR'24	Cut3R CVPR'25	Light3R CVPR'25	SLAM3R CVPR'25	VGGT-SLAM NeurIPS'25	FlowMap 3DV'25	SAIL 3DV'26	SAIL-OPT 3DV'26	MBA 3DV'26	Ours°	Ours†	Ours°+BA
RRA@5↑	31.3	75.8	56.9	22.1	49.2	69.6	18.8	52.0	20.3	57.3	31.7	70.4	71.5	71.7	70.0	71.7	92.6
RTA@5↑	40.3	76.7	57.9	30.7	54.0	69.3	25.8	52.8	24.7	67.9	35.7	74.7	77.7	77.0	74.0	78.0	90.6
ATE↓	0.021	0.010	0.015	0.016	0.011	0.014	0.017	0.011	0.015	0.008	0.017	0.008	0.008	0.009	0.006	0.005	0.004

Table 4: Pose accuracy on the FastMap benchmark. Parenthesised counts give the number of scenes averaged in each split; rows without one are single scenes. Splits above the rule are scored against COLMAP-pipeline ground truth (PixSfM for urban scene and Mill-19); those below use poses from commercial software. **Key:** method-with-optimization / method-with-feed-forward; best, 2nd, 3rd per metric. failure, timeout.

		ATE↓						AUC@3↑						RTA@3↑					
	avg. n_imgs	COLMAP	FastMap	GLOMAP	Ours°	Ours†	Ours°+BA	COLMAP	FastMap	GLOMAP	Ours°	Ours†	Ours°+BA	COLMAP	FastMap	GLOMAP	Ours°	Ours†	Ours°+BA
mipnerf360 (9)	215.6	5.8e-5	4.2e-4	3.3e-5	6.1e-3	8.0e-4	1.2e-3	97.2	97.4	98.2	32.0	55.6	71.9	100.0	99.9	100.0	71.6	91.4	99.3
T&T advanced (6)	337.8	1.2e-3	6.4e-3	1.2e-2	1.02e-2	1.04e-2	1.1e-3	94.8	42.6	75.3	35.4	41.9	60.0	98.5	71.4	79.1	64.8	69.7	89.1
T&T intermed. (8)	268.6	2.6e-4	7.8e-5	1.9e-5	5.0e-4	3.0e-4	1.0e-3	98.9	94.1	99.0	55.7	67.3	67.9	99.8	99.9	100.0	95.1	97.4	93.3
T&T training (7)	470.1	3.0e-4	3.0e-3	1.1e-2	3.1e-3	2.2e-3	3.2e-3	99.5	77.2	87.9	44.7	59.1	68.5	99.9	87.8	88.7	78.0	91.7	91.9
nerf osr (8)	402.8	1.3e-3	1.6e-3	1.1e-3	1.2e-3	1.0e-3	1.2e-3	71.7	70.9	71.9	45.7	51.4	53.9	92.1	91.7	92.0	85.4	91.7	87.9
urban scene (3)	3824	1.4e-5	1.7e-5	1.4e-5	2.9e-3	1.6e-3	4.0e-4	97.0	95.3	97.0	8.9	37.4	61.5	100.0	99.9	99.9	39.5	67.9	96.5
mill19 bldg	1920	1.9e-5	3.0e-4	1.3e-2	3.5e-3	5.0e-4	2.0e-4	95.6	95.5	0.0	3.1	41.4	88.4	99.9	99.9	0.1	23.1	86.6	99.9
mill19 rubble	1657	3.4e-5	3.6e-5	6.4e-5	2.1e-3	2.0e-4	1.0e-4	94.6	93.6	94.5	33.0	69.8	90.9	99.9	99.9	99.8	86.4	97.9	99.9
drone deploy (9)	524.7	2.0e-3	4.9e-3	4.3e-3	2.5e-3	7.0e-4	8.0e-4	65.2	79.2	81.1	21.9	51.0	74.3	91.3	97.9	98.2	61.5	93.6	98.8
eyeful apt	3804	2.2e-3	2.8e-3	9.4e-3	2.5e-3	1.3e-3	2.0e-4	62.0	45.5	50.5	26.4	44.4	44.2	90.2	86.8	75.0	70.2	84.1	91.1
eyeful kitchen	6042	–	3.1e-3	7.4e-3	7.0e-4	2.0e-4	2.0e-4	–	38.1	41.2	27.3	56.2	53.6	–	85.0	59.9	74.9	94.3	95.4

Self-improving gain. The left axis of Fig. 6 shows the loop improving 32 of 33 units, most where feed-forward inference generalizes poorly: the weaker half by initial accuracy gains $+32.8$ points against $+13.5$ for the stronger (e.g. Mill 19 building, $23.1 \rightarrow 86.6$). The leave-one-domain-out control (Ours$^{\ominus}$) of Tab. C.1 still improves substantially over Ours°. We run the loop once: Ours†+BA matches Ours°+BA on most columns there, so a second round would retrain on nearly the same labels.

BA is still needed. BA remains the more accurate configuration, but pseudo-labeling transfers most of its advantage: its average gain shrinks from $+31.5$ points over Ours° to $+8.6$ over Ours†, and it still improves 30 of 33 units (right axis of Fig. 6).

Free-point refinement. The Ours°+BA rows with and without free-point refinement in Tab. C.1 isolate the final stage: releasing each point from its viewing ray improves most entries.

Runtime. Tab. B.1 reports per-collection inference timings: despite computing exact full attention, our feed-forward pass is faster than COLMAP on every split COLMAP completes. Adding global BA dominates our runtime. Pseudo-label generation and fine-tuning run offline, once per domain, costing one BA pass over the collection and one training run.

5 Conclusion

We presented a self-improving feed-forward SfM: bundle adjustment refines the model's own predictions, and the refined poses fine-tune the model that produced them. Our BA depends only on the outputs of frozen learning primitives and keeps only the views it registers reliably. Exact context-parallel attention and a sparse pose graph let the pipeline run end-to-end on unconstrained collections of $\sim$15,000 images. Across small-scene, large-scene, and internet datasets its refined poses approach SoTA SfM accuracy, making them reliable pseudo labels that substantially improve feed-forward pose accuracy.

Limitations. The loop's reach is bounded by what BA can already register: pseudo labels come only from the views BA retains, and re-filtering can only shrink that subset, so the images BA fails on never become training signal. BA therefore stays in the pipeline, both as the runtime bottleneck and as the only stage that filters views and recovers the depth bias and radial distortion. The loop also tunes only the SfM pose estimator, leaving the depth and correspondence models frozen, and we evaluate a single round: the first pass helps, but we do not show that repeating it helps further.

Reproducibility Statement

Our bundle adjustment requires no training and consumes only model outputs, so it can be reproduced independently of our network; its frozen inputs are public, namely MoGe-2 (Wang et al., 2025c) depth and RoMa (Edstedt et al., 2024) correspondences with Doppelgangers++ (Cai et al., 2023) verification. Sec. 3.3 defines the objective, Apps. E.4 to E.6 derive the residuals, solver, and scale initialization, and App. D lists every hyperparameter. The pose estimator starts from the public DINOv2 (Oquab et al., 2024) and VGGT (Wang et al., 2025a) checkpoints, and App. D also gives its architecture, the training and fine-tuning datasets, all of which are public, and the optimization settings; App. E.7 describes how pseudo labels are prepared. All six benchmarks are public, and Sec. 4 states whose protocol and baseline numbers we adopt on each.

AI Use Statement

We used generative AI assistants to improve the clarity and readability of text written by the authors, search for and summarize relevant literature, check bibliographic information, and verify that baseline results quoted in the manuscript are consistent with their cited sources. We also used these tools to provide feedback on manuscript organization, experimental presentation, ablation design, and the clarity and internal consistency of our interpretation of the results. AI-assisted tools were additionally used to revise LaTeX source and identify inconsistencies in terminology, equations, tables, and cross-references.

All research ideas, methodological decisions, experimental implementations, training configurations, data collection procedures, and final experimental choices were determined and approved by the authors. Generative AI tools did not autonomously conduct experiments or generate, modify, select, or discard experimental measurements. Every numerical result reported in the paper was produced by experiments conducted by the authors and was checked against the corresponding experimental record. Suggestions and text produced with AI assistance were treated as unverified drafts and were reviewed, corrected, and approved by the authors. Citations and externally reported baseline values were independently checked against their original sources. The authors take full responsibility for the final content of this paper, including all text, claims, experimental results, citations, and artifacts produced or revised with the assistance of generative AI.

References

Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *ECCV*, 2022.

Sithu Aung, Viktor Kocur, Yaqing Ding, Torsten Sattler, and Zuzana Kukelova. DGSfM: Depth-guided scale-aware global structure-from-motion. In *ACCV*, 2026.

Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. NoPe-NeRF: Optimising neural radiance field with no pose prior. In *CVPR*, 2023.

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.

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.

Xingyu Chen, Yue Chen, Yuliang Xiu, Andreas Geiger, and Anpei Chen. TTT3R: 3d reconstruction as test-time training. In *ICLR*, 2026a.

Youyu Chen, Junjun Jiang, Yueru Luo, Kui Jiang, Xianming Liu, Xu Yan, and Dave Zhenyu Chen. Reliev3R: Relieving feed-forward 3d reconstruction from multi-view geometric annotations. In *CVPR*, 2026b.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017.

Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *ICLR*, 2024.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *NeurIPS*, 2022.

Junyuan Deng, Heng Li, Kejie Qiu, Lingteng Qiu, Rui Peng, Weichao Shen, Weihao Yuan, Siyu Zhu, Zilong Dong, and Ping Tan. Glob3R: Global structure-from-motion with 3d foundation models. *arXiv preprint arXiv:2607.09225*, 2026a.

Junyuan Deng, Heng Li, Tao Xie, Weiqiang Ren, Qian Zhang, Ping Tan, and Xiaoyang Guo. SAIL-Recon: Large SfM by augmenting scene regression with localization. In *3DV*, 2026b.

Kai Deng, Zexin Ti, Jiawei Xu, Jian Yang, and Jin Xie. VGGT-Long: Chunk it, loop it, align it, pushing VGGT's limits on kilometer-scale long RGB sequences. In *ICRA*, 2026c.

Youming Deng, Songyou Peng, Junyi Zhang, Kathryn Heal, Tiancheng Sun, John Flynn, Steve Marschner, and Lucy Chai. Selfi: Self-improving reconstruction engine via 3d geometric feature alignment. In *CVPR*, 2026d.

Bardienus Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jérôme Revaud. MASt3R-SfM: A fully-integrated solution for unconstrained structure-from-motion. In *3DV*, 2025.

Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. RoMa: Robust dense feature matching. In *CVPR*, 2024.

Sven Elflein, Qunjie Zhou, Sérgio Agostinho, and Laura Leal-Taixé. Light3R-SfM: Towards feed-forward structure-from-motion. In *CVPR*, 2025.

Sven Elflein, Ruilong Li, Sérgio Agostinho, Zan Gojcic, Laura Leal-Taixé, Qunjie Zhou, and Aljosa Osep. VGG-T^3: Offline feed-forward 3d reconstruction at scale. In *CVPR*, 2026.

Yang Fu, Sifei Liu, Amey Kulkarni, Jan Kautz, Alexei A. Efros, and Xiaolong Wang. COLMAP-Free 3d gaussian splatting. In *CVPR*, 2024.

Clément Godard, Oisín Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, 2019.

Marwane Hariat, Gianni Franchi, David Filliat, and Antoine Manzanera. SS3D: End2end self-supervised 3d from web videos. *arXiv preprint arXiv:2604.22686*, 2026.

Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. DF-SfM: Detector-free structure-from-motion. In *CVPR*, 2024.

Sergio Izquierdo and Javier Civera. SfM-TTR: Using structure from motion for test-time refinement of single-view depth networks. In *CVPR*, 2023.

Hanwen Jiang, Hao Tan, Peng Wang, Haian Jin, Yue Zhao, Sai Bi, Kai Zhang, Fujun Luan, Kalyan Sunkavalli, Qixing Huang, and Georgios Pavlakos. RayZer: A self-supervised large view synthesis model. In *ICCV*, 2025.

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.

Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, Jonathon Luiten, Manuel Lopez-Antequera, Samuel Rota Bulò, Christian Richardt, Deva Ramanan, Sebastian Scherer, and Peter Kontschieder. MapAnything: Universal feed-forward metric 3d reconstruction. In *3DV*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM TOG*, 2017.

JongMin Lee and Sungjoo Yoo. Dense-SfM: Structure from motion with dense consistent matching. In *CVPR*, 2025.

Jongmin Lee, Seungyeop Kang, and Sungjoo Yoo. MV-RoMa: From pairwise matching into multi-view track reconstruction. In *CVPR*, 2026.

Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with MASt3R. In *ECCV*, 2024.

Dacheng Li, Rulin Shao, Anze Xie, Eric P. Xing, Xuezhe Ma, Ion Stoica, Joseph E. Gonzalez, and Hao Zhang. DISTFLASHATTN: Distributed memory-efficient attention for long-context LLMs training. In *COLM*, 2024.

Jiahao Li, Haochen Wang, Muhammad Zubair Irshad, Igor Vasiljevic, Matthew R. Walter, Vitor Campagnolo Guizilini, and Greg Shakhnarovich. FastMap: Revisiting structure from motion through first-order optimization. In *3DV*, 2026.

Zhengqi Li and Noah Snavely. MegaDepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018.

Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *ICCV*, 2021.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *ICLR*, 2026.

Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10K: A large-scale scene dataset for deep learning-based 3d vision. In *CVPR*, 2024.

Hao Liu, Matei Zaharia, and Pieter Abbeel. Ring attention with blockwise transformers for near-infinite context. In *ICLR*, 2024.

Yuzheng Liu, Siyan Dong, Shuzhe Wang, Yingda Yin, Yanchao Yang, Qingnan Fan, and Baoquan Chen. SLAM3R: Real-time dense scene reconstruction from monocular RGB videos. In *CVPR*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

Dominic Maggio, Hyungtae Lim, and Luca Carlone. VGGT-SLAM: Dense RGB SLAM optimized on the SL(4) manifold. In *NeurIPS*, 2025.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *TMLR*, 2024.

Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes L. Schönberger. GLOMAP: Global structure-from-motion revisited. In *ECCV*, 2024.

Linfei Pan, Johannes L. Schönberger, and Marc Pollefeys. Global structure-from-motion meets feedforward reconstruction. In *CVPR*, 2026.

Benjamin Resch, Hendrik P. A. Lensch, Oliver Wang, Marc Pollefeys, and Alexander Sorkine-Hornung. Scalable structure from motion for densely sampled videos. In *CVPR*, 2015.

Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.

Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016.

Thomas Schöps, Johannes L. Schönberger, 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.

You Shen, Zhipeng Zhang, Yansong Qu, Xiawu Zheng, Jiayi Ji, Shengchuan Zhang, and Liujuan Cao. FastVGGT: Fast visual geometry transformer. In *ICLR*, 2026.

Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. FlowMap: High-quality camera poses, intrinsics, and depth via gradient descent. In *3DV*, 2025.

Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. In *ACM SIGGRAPH*, 2006.

Zhenggang Tang, Yuchen Fan, Dilin Wang, Hongyu Xu, Rakesh Ranjan, Alexander Schwing, and Zhicheng Yan. MV-DUSt3R+: Single-stage scene reconstruction from sparse views in 2 seconds. In *CVPR*, 2025.

Zachary Teed and Jia Deng. DROID-SLAM: Deep visual slam for monocular, stereo, and rgb-d cameras. In *NeurIPS*, 2021.

Hengyi Wang and Lourdes Agapito. Spann3R: 3d reconstruction with spatial memory. In *3DV*, 2025.

Hengyi Wang and Lourdes Agapito. AMB3R: Accurate feed-forward metric-scale 3d reconstruction with backend. In *CVPR*, 2026.

Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. VGGSfM: Visual geometry grounded deep structure from motion. In *CVPR*, 2024a.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *CVPR*, 2025a.

Jianyuan Wang, Minghao Chen, Shangzhan Zhang, Nikita Karaev, Johannes Schönberger, Patrick Labatut, Piotr Bojanowski, David Novotny, Andrea Vedaldi, and Christian Rupprecht. VGGT-Ω: Scaling feed-forward 3d reconstruction. In *CVPR*, 2026a.

Qianqian Wang, Yifei Zhang, Aleksander Holynski, Alexei A. Efros, and Angjoo Kanazawa. Continuous 3d perception model with persistent state. In *CVPR*, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: Accurate monocular geometry with metric scale and sharp details. In *NeurIPS*, 2025c.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jérôme Revaud. DUSt3R: Geometric 3d vision made easy. In *CVPR*, 2024b.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. TartanAir: A dataset to push the limits of visual slam. In *IROS*, 2020.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3: Permutation-equivariant visual geometry learning. In *ICLR*, 2026b.

Tobias Weyand, Andre Araujo, Bingyi Cao, and Jack Sim. Google landmarks dataset v2 – a large-scale benchmark for instance-level recognition and retrieval. In *CVPR*, 2020.

Kyle Wilson and Noah Snavely. Robust global translations with 1DSfM. In *ECCV*, 2014.

Felix Wimbauer, Weirong Chen, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. AnyCam: Learning to recover camera poses and intrinsics from casual videos. In *CVPR*, 2025.

Amy Yang, Jingyi Yang, Aya Ibrahim, Xinfeng Xie, Bangsheng Tang, Grigory Sizov, Jeremy Reizenstein, Jongsoo Park, and Jianyu Huang. Context parallelism for scalable million-token inference. In *MLSys*, 2025a.

Jianing Yang, Alexander Sax, Kevin J. Liang, Mikael Henaff, Hao Tang, Ang Cao, Joyce Chai, Franziska Meier, and Matt Feiszli. Fast3R: Towards 3d reconstruction of 1000+ images in one forward pass. In *CVPR*, 2025b.

Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. BlendedMVS: A large-scale dataset for generalized multi-view stereo networks. In *CVPR*, 2020.

Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. ScanNet++: A high-fidelity dataset of 3d indoor scenes. In *ICCV*, 2023.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G. Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 2017.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019.

Shengjie Zhu, Ahmed Abdelkader, Mark J. Matthews, Xiaoming Liu, and Wen-Sheng Chu. Marginalized bundle adjustment: Multi-view camera pose from monocular depth estimates. In *3DV*, 2026.

Dong Zhuo, Wenzhao Zheng, Jiahe Guo, Yuqi Wu, Jie Zhou, and Jiwen Lu. Streaming 4d visual geometry transformer. In *ICLR*, 2026.

A ETH3D AUC RESULTS

Tab. A.1 presents pose estimation results on ETH3D (Schöps et al., 2017) in AUC of pose error, complementing the RRA/RTA results in Tab. 1. Baseline numbers are quoted from MBA (Zhu et al., 2026) and DF-SfM (He et al., 2024); runtimes in Tab. B.1 are from (Li et al., 2026).

Table A.1: Pose estimation on ETH3D in AUC of pose error at $1°/3°/5°$, following **MBA** on the 13-scene training set (as in Tab. 1); the pose error is the larger of the rotation and translation errors, so the metric differs from the separate RRA/RTA of Tab. 1. **Key:** method-with-optimization / method-with-feed-forward; best, 2nd, 3rd per metric.

Type	Method	AUC@1°	AUC@3°	AUC@5°
Detector-Based	SIFT+NN + COLMAP	26.71	38.86	42.14
	SIFT+NN + PixSfM	26.94	39.01	42.19
	D2Net+NN + PixSfM	34.50	49.77	53.58
	R2D2+NN + PixSfM	43.58	62.09	66.89
	SP+SG + PixSfM	50.82	68.52	72.86
Detector-Free	LoFTR + PixSfM	54.35	73.97	78.86
	LoFTR + DF-SfM	59.12	75.59	79.53
	AspanTrans. + DF-SfM	57.23	73.71	77.70
	MatchFormer + DF-SfM	56.70	73.00	76.84
Dense Matching	DKM + Dense-SfM	59.04	77.73	82.20
	RoMa + Dense-SfM	60.92	78.41	82.63
Point-Based	MASt3R-SfM	35.85	58.46	65.03
	Dense-SfM + MASt3R-SfM	37.50	59.18	65.48
MDE	MBA	27.72	63.35	74.02
	Ours$^{\circ}$	2.93	20.59	33.49
	Ours†	10.80	45.64	61.33
	Ours$^{\circ}$+BA	33.26	70.96	81.58

B RUNTIME ON THE FASTMAP BENCHMARK

Tab. B.1 reports per-dataset wall-clock runtime on the FastMap benchmark (Li et al., 2026), complementing the pose accuracy results in Tab. 4. Baseline timings are quoted from Li et al. (2026), measured on a single A6000 GPU with a 48-thread AMD EPYC CPU and excluding feature extraction and matching, which is shared across those three methods. We measure our own timings on a single 8×A100-80GB node.

Table B.1: Runtime on the FastMap benchmark (wall-clock seconds). **Key:** method-with-optimization / method-with-feed-forward; `t/o` timeout, as reported by Li et al. (2026).

Method avg. n_imgs	mipnerf360 215.6	T&T advn 337.8	T&T intm 268.6	T&T trng 470.1	nerf osr 402.8	drone dep 524.7	urban sc 3824	mill19 bldg 1920	mill19 rbl 1657	eyeful apt 3804	eyeful ktc 6042
GLOMAP	165	357	314	515	324	365	3664	6289	2849	5905	22884
FastMap	33	61	35	63	50	91	515	258	240	549	2202
COLMAP	503	1016	833	2751	3163	3352	61622	27080	12153	185361	t/o
Ours†	162	274	242	348	308	242	6541	1426	1074	5861	12833
Ours$^{\circ}$+BA	1624	2548	2179	3439	3806	2859	6659	4662	4980	9312	17262

C EXTENDED ABLATIONS

Tab. C.1 is the full table underlying the ablation discussion in Sec. 4, covering pseudo-labeled fine-tuning (†) and test-time bundle adjustment (+BA) across all four benchmarks. Following the protocol of Sec. 4, each configuration is reported twice: over *all* input frames, and over the BA-retained subset $\mathcal{V}^*$ (*retained*), with the registration rate *% Reg.* listed below. For the +BA rows, the *all* row scores BA's own refined pose for every input frame: view filtering selects which frames are trustworthy enough to serve as pseudo labels, and does not remove frames from this evaluation or substitute the feed-forward pose for them. Comparisons are drawn *within* a single block, so the configurations being compared are always scored on the same frames. Because our feed-forward configurations register every frame by construction, their *retained* rows repeat their *all* rows and serve only as a reference for the adjacent BA rows.

Leave-one-domain-out control. Ours† is transductive: its fine-tuning pool contains the evaluation images, though never their poses. The Ours$^{\ominus}$ row controls for this: for each evaluation domain we

Table C.1: Extended ablations across benchmarks. Accuracy over *all* frames and over the BA-retained subset $\mathcal{V}^*$ (*retained*), for pseudo-labeled fine-tuning (†) and test-time bundle adjustment (+BA). **Ours**$^\ominus$ is the leave-one-domain-out control described below, one checkpoint per column group. *w/o Free* ablates the free-point refinement of Sec. 3.3; *% Reg.* is the fraction of frames registered after view filtering. **Key:** method-with-optimization / method-with-feed-forward.

		ETH3D			IMC2021			Tanks & Temples			FastMap										
	Config	RRA	RTA	@3	@3	@5	@10	RRA	RTA	ATE	mip360	t_advn	t_intm	t_trng	n_osr	drone	urban	m_bl	m_rb	e_ap	e_ki
Accuracy (all)	**Ours**$^\circ$	87.3	60.4	20.7	30.9	46.0	65.0	70.0	74.0	0.006	71.6	64.8	95.1	78.0	85.4	61.5	39.5	23.1	86.4	70.2	74.9
	Ours†	100.0	88.5	45.6	49.6	63.7	78.3	71.7	78.0	0.005	91.4	69.7	97.4	91.7	91.7	93.6	67.9	86.6	97.9	84.1	94.3
	Ours$^\ominus$	98.9	84.3	39.8	46.4	60.9	75.8	72.4	76.1	0.005	85.7	67.3	96.1	86.7	86.5	69.3	48.2	48.8	92.9	71.2	84.1
	Ours$^\circ$**+BA**	100.0	97.0	66.6	54.8	67.7	80.7	71.2	79.2	0.005	99.1	71.8	98.0	89.9	90.7	95.5	69.4	51.8	99.8	82.2	95.1
	Ours†**+BA**	100.0	97.3	67.1	54.8	67.6	80.6	70.9	79.3	0.005	99.2	70.8	98.0	90.8	90.7	98.6	76.8	99.8	99.9	83.1	95.3
Accuracy (retained)	**Ours**$^\circ$	87.3	60.4	20.7	30.9	46.0	65.0	70.0	74.0	0.006	71.6	64.8	95.1	78.0	85.4	61.5	39.5	23.1	86.4	70.2	74.9
	Ours$^\circ$**+BA** *w/o Free*	100.0	98.3	71.1	54.9	67.8	80.7	91.4	90.3	0.004	97.7	88.6	93.1	91.7	87.5	98.6	96.5	99.9	99.9	87.4	83.9
	Ours$^\circ$**+BA**	100.0	98.6	73.3	56.7	68.9	81.3	92.6	90.6	0.004	99.3	89.1	93.3	91.9	87.9	98.8	96.5	99.9	99.9	91.1	95.4
	Ours†	100.0	88.5	45.6	49.6	63.7	78.3	71.7	78.0	0.005	91.4	69.7	97.4	91.7	93.6	67.9	86.6	97.9	84.1	94.3
	Ours†**+BA**	100.0	98.6	73.7	56.3	68.6	81.1	94.1	92.7	0.002	98.8	93.4	92.9	92.3	85.3	98.1	97.2	99.8	99.9	88.0	83.7
% Reg.	**Ours**$^\circ$**+BA**	59.5			98.0			71.0			78.8	72.1	70.7	70.3	57.9	77.3	83.5	87.1	74.0	57.4	79.3
	Ours†**+BA**	59.6			98.0			71.4			78.7	73.5	70.9	70.1	60.1	77.3	82.8	78.9	74.0	58.7	79.7

fine-tune a separate model whose pool excludes that entire domain, matched by scene-name prefix, and evaluate it *only* on the domain it never saw. Each arm drops the full domain: 25 ETH3D scenes, 21 Tanks & Temples, 9 MipNeRF-360, 9 DroneDeploy, 8 NeRF-OSR, and the entire IMC2021 pool. Every one of the 20 numbers we report improves under Ours$^\ominus$ relative to Ours$^\circ$, including the Tanks & Temples ATE. Per benchmark, Ours$^\ominus$ still tops the feed-forward baselines of Sec. 4: AMB3R (Wang & Agapito, 2026) on ETH3D, DA3 (Lin et al., 2026) on IMC2021, and SAIL (Deng et al., 2026b) on Tanks & Temples.

Where fine-tuning helps most. Sec. 4 reports that the loop helps most where Ours$^\circ$ is weakest; this expands that comparison over the 33 units of Fig. 6. Splitting the units at the median Ours$^\circ$ accuracy leaves 16 units in each half, the median unit itself falling in neither; the weaker half gains $+32.8$ points on average against $+13.5$ for the stronger. The ordering holds within each benchmark separately, where the metric is fixed, so it does not follow from pooling benchmarks whose metrics differ. Part of the gap is headroom, since a unit starting at 32 has more room to improve than one starting at 77, so we claim only the absolute gains that Fig. 6 plots.

Why one round. Sec. 4 states that we stop after a single round of the loop. The pseudo-labels of a second round would come from Ours†+BA, and those of the first from Ours$^\circ$+BA. The two rows of Tab. C.1 over the retained subset therefore bound what a second round could supply. They agree within one point on 13 of the 19 accuracy columns, with a mean change of -0.6 points, and the retained pool is nearly unchanged. The exceptions are the Eyeful Tower kitchen scene at $95.4 \to 83.7$, a *regression*, and Tanks & Temples advanced at $89.1 \to 93.4$. BA over the fine-tuned model thus recovers essentially the labels it already produced, so a second round would retrain on its own supervision. We report one round and leave the multi-round regime, together with the drift these outliers suggest, to future work.

D Implementation Details

Training Datasets. We train $\mathcal{F}_\theta$ on a mixture of annotated datasets with ground-truth or SfM-derived camera poses: MegaDepth (Li & Snavely, 2018), ScanNet (Dai et al., 2017), Map-Free Relocalization (Arnold et al., 2022), BlendedMVS (Yao et al., 2020), TartanAir (Wang et al., 2020), DL3DV (Ling et al., 2024), and ScanNet++ (Yeshwanth et al., 2023). For pseudo-labeled fine-tuning, we additionally use unannotated image collections from Google Landmarks v2 (Weyand et al., 2020) and 1DSfM (Wilson & Snavely, 2014), as well as ETH3D (Schöps et al., 2017), IMC2021 (Jin et al., 2021), and FastMap (Li et al., 2026) scenes where ground-truth poses are held out. Only the poses are withheld: the test images themselves are part of this unannotated pool, so Ours† is a transductive result, which the leave-one-domain-out control of App. C quantifies.

Monocular Depth. The monocular depth maps $\mathbf{D}_i = \mathcal{N}_\mathbf{D}(\mathbf{I}_i)$ consumed by both the pose estimator and bundle adjustment are produced by MoGe-2 (Wang et al., 2025c), using the publicly released `moge-2-vitl` checkpoint run locally with frozen weights. We take its predicted depth channel

only and discard the metric scale, since the per-frame scale s_i and bias b_i are re-estimated as part of our pipeline, as described in App. E.6.

Covisibility and Correspondence. We use the covisibility threshold $\nu = 0.5$ and confidence threshold $\chi = 0.1$ for correspondence filtering. For each edge, $\kappa = 5{,}000$ correspondences are sampled with replacement.

Bundle Adjustment. The optimized variables are the per-frame rotation and translation, the depth scale s_i and bias b_i, per-frame depth-intrinsic focal offsets $(\Delta f_{x,i}, \Delta f_{y,i})$, a per-camera focal length parameterized as $f_{\text{norm}} = f/\sqrt{H^2 + W^2}$, and optionally a single-parameter radial distortion coefficient. The two focal terms act at different points of Eq. (E.2): the depth-intrinsic offsets enter $\mathbf{K}_i^{-1}$, setting the ray directions along which the monocular depth is backprojected, while the per-camera focal enters $\mathbf{K}_j$ and is shared across the frames of one physical camera. The histogram-based KDE uses $T = 250$ bins over the residual range $[0, \bar{\tau}_{\text{max}}]$ with $\tau_{\text{max}} = 100$ pixels. RANSAC scale initialization draws $M = 1{,}000$ candidates per frame. Optimization proceeds in three coarse-to-fine stages: (i) per-node CDF with log-transformed residuals, (ii) global CDF with log-transformed residuals, and (iii) global CDF with Euclidean residuals, each solved by the Levenberg–Marquardt procedure of App. E.5. Each stage uses early stopping: the optimizer advances to the next stage if $\bar{Q}$ does not improve by more than $\delta = 10^{-4}$ for a prescribed number of consecutive iterations. After the global stage converges, we greedily prune frames whose retained edges fall below an inlier CCDF threshold $\tau = 0.65$ and keep the largest connected component of the surviving pose graph. Both endpoints of the worst edge attain the minimum in Eq. (4), so we break that tie by mean incident edge score and drop the node whose remaining edges are worse. Free-point refinement then re-optimizes the retained subset, jointly refining the per-correspondence world points and cameras against the reprojection CDF loss in native pixels (600 histogram bins over $[0, 50]$ pixels); the per-frame depth adjustments serve only to initialize the points and drop out of the residual.

SfM Pose Estimator. We initialize the image encoder from DINOv2 (Oquab et al., 2024) and the alternating frame-local/global attention decoder from VGGT (Wang et al., 2025a). We then follow π^3 (Wang et al., 2026b) in making the decoder permutation-equivariant, removing the anchor-frame encoding so that no view is privileged as a reference. The production network is deepened to 72 alternating frame-local/global attention blocks by replicating the pretrained blocks twice. We adopt the 6D rotation representation (Zhou et al., 2019) by default. During fine-tuning we freeze most of the pretrained backbone and train primarily the camera head.

Training. We train with multi-scale sampling: each mini-batch randomly selects a scene and draws $N \in \{5, 25, 64, 128\}$ views at probabilities $\{0.2, 0.4, 0.25, 0.15\}$. In the first stage, the model is trained exclusively on annotated datasets with ground-truth poses; this yields the **Ours**$^{\circ}$ configuration ($^{\circ}$ marks *no* target-domain fine-tuning). In the second stage, we apply bundle adjustment over the unannotated datasets to produce pseudo ground truth, then fine-tune the model using these pseudo labels as additional supervised targets mixed with the annotated data at weight $\beta = 1$; this yields the **Ours**† configuration († marks pseudo-labeled fine-tuning on the target domain). Training uses AdamW (Kingma & Ba, 2015; Loshchilov & Hutter, 2019) with a cosine learning rate schedule. The camera head makes $L = 4$ iterative predictions. The supervised loss uses decay factor $\gamma = 0.6$ and weights $\lambda_R = 2$, $\lambda_t = 2$, $\lambda_f = 1$, $\lambda_s = 0.5$.

Inference. The pose estimator predicts all camera parameters in a single pass, after which the bundle adjustment of Sec. 3.3 optionally refines them.

E Deferred Method Details

This appendix expands the method components of Sec. 3 that were deferred for space.

E.1 Camera Head Parameterization

The camera head of Sec. 3.2 regresses, per frame, the extrinsics $\mathbf{P}_i$, the intrinsics $\mathbf{K}_i$, and the depth scale s_i from the images and monocular depths, $\{\mathbf{P}_i, \mathbf{K}_i, s_i\} = \mathcal{F}_\theta(\{\mathbf{I}_i, \mathbf{D}_i\})$. The extrinsics use a 6D rotation representation (Zhou et al., 2019) together with a translation: the head emits $\mathbf{r}_i \in \mathbb{R}^6$ mapped to $\mathbf{R}_i \in \mathrm{SO}(3)$ via Gram–Schmidt, and $\mathbf{t}_i \in \mathbb{R}^3$, giving $\mathbf{P}_i = [\mathbf{R}_i \mid \mathbf{t}_i]$. The intrinsics are

parameterized by the per-axis field-of-view $(\phi_{x,i}, \phi_{y,i})$ with the principal point fixed at the image center, so $\mathbf{K}_i$ is determined by $(\phi_{x,i}, \phi_{y,i})$ and the image size. BA additionally recovers a per-camera radial distortion coefficient and a per-frame depth bias b_i, neither of which the head predicts.

E.2 Covisibility Graph and Correspondence Sampling

Covisibility Graph. We construct the sparse pose graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ of Sec. 3.3 with one node per frame and one edge per co-visible pair. We use RoMa (Edstedt et al., 2024) to measure covisibility: for each image, we extract feature maps $\mathbf{f}_i$ of $H_p \times W_p$ patches from RoMa's DINOv2 (Oquab et al., 2024) backbone at the coarse (stride-16) level and project them via a linear layer. The covisibility score between frames i and j is the mean best-match similarity $v_{i,j} = \frac{1}{H_p W_p} \sum_u \max_{u'} \mathrm{sim}(\mathbf{f}_i[u], \mathbf{f}_j[u'])$, where u and u' index the $H_p \times W_p$ patch locations of the two feature maps and sim is cosine similarity, and an edge (i, j) is included if $v_{i,j} > \nu$. The sparsity of $\mathcal{G}$ is essential for scalability: BA memory and computation scale with $|\mathcal{E}|$, kept far below $\binom{N}{2}$ by the covisibility threshold.

Correspondence Sampling. For each edge $(i, j) \in \mathcal{E}$ we run the matcher $\mathcal{N}_\mathbf{C}$ to obtain a dense correspondence map $\mathbf{C}_{i,j} = \{(p_k, q_k, \omega_k)\}$ with confidence scores ω_k. For each frame i, we aggregate the correspondences from all its covisible frames into a single pool, retaining only those with confidence $\omega_k \geq \chi$. From this pool we sample κ correspondences with replacement, together with the source depth $\mathbf{D}_i[p_k]$, yielding a data matrix $\mathbf{M} \in \mathbb{R}^{|\mathcal{E}| \times \kappa \times 5}$ with rows $\mathbf{M}_{e,k} = [p_k^x, p_k^y, q_k^x, q_k^y, \mathbf{D}_i[p_k]]$. Each row carries exactly the inputs of one projective residual (Eq. (E.2)), so this fixed-shape tensor lets BA evaluate all residuals in parallel on GPU.

E.3 Context-Parallel Attention

The image-sharded multi-view inference of Sec. 3.2 rests on the context-parallel decomposition of attention (Liu et al., 2024; Li et al., 2024; Yang et al., 2025a), summarized here. The encoder maps the N images to NP tokens, P per image, over which global attention would form an $NP \times NP$ score matrix. FlashAttention (Dao et al., 2022; Dao, 2024) never materializes it: it streams the keys and values in blocks while maintaining, per query row, a running maximum, a denominator, and an unnormalized output, and normalizes once at the end. Two such partial states, computed over *disjoint* blocks of key–value indices, merge exactly by rescaling both to their common maximum and adding; write $\mathbf{z}_i$ for the partial state of query row i, $\oplus$ for that merge, and $\mathrm{block}(\cdot)$ for the partial state of a single block. Because $\oplus$ is associative and commutative, partial states can be reduced over *any* partition of the key–value indices, in any order, without changing the result.

We exploit this by partitioning the N images into G equal shards, one per GPU and possibly spanning multiple nodes, so each device holds NP/G tokens. Shard g contributes the token block $\mathcal{B}_g$, and GPU g stores only its own attention queries, keys, and values $(\mathbf{Q}, \mathbf{K}, \mathbf{V})$, not to be confused with the camera intrinsics $\mathbf{K}_i$. It evaluates its local query rows against every shard by streaming in the peer keys and values:

$$\mathbf{z}_i^{(g)} \;=\; \bigoplus_{g'=1}^{G} \mathrm{block}\big(\mathbf{Q}_i^{(g)},\, \mathbf{K}^{(g')},\, \mathbf{V}^{(g')}\big), \qquad i \in \mathcal{B}_g. \tag{E.1}$$

Since $\oplus$ is partition-independent, Eq. (E.1) computes the same dense full-attention operator, with no sparsification and no view chunking: the sharded output equals the single-device output in exact arithmetic, differing only through floating-point reduction order. At any instant GPU g holds its own shard and one streamed peer block, so its memory is $\mathcal{O}(NP/G)$, linear in the number of images, against the $\mathcal{O}(N^2P^2)$ of a materialized score matrix.

E.4 MBA Residuals and Surrogate Loss

Projective Residuals. Each correspondence $(p_k, q_k) \in \mathbf{I}_i \times \mathbf{I}_j$ is scored by backprojecting p_k with the affine-adjusted monocular depth and reprojecting into frame j (with perspective division):

$$\mathbf{X}_k = \big(s_i\, \mathbf{D}_i[p_k] + b_i\big)\, \mathbf{K}_i^{-1}[p_k^x, p_k^y, 1]^\top, \qquad \hat{q}_k = \mathbf{K}_j\big(\mathbf{R}_j(\mathbf{R}_i^\top(\mathbf{X}_k - \mathbf{t}_i)) + \mathbf{t}_j\big), \tag{E.2}$$

where the per-frame scale s_i and bias b_i map the monocular depth to metric depth. The residual $r_{i,j,k} = \|\hat{q}_k - q_k\|_2$ is log-transformed to $\bar{r}_{i,j,k} = \log(1 + \frac{1}{2} r_{i,j,k}^2)$ for robustness, and computed at the *original image resolution*, so the maximum-residual threshold $\bar{\tau}_{\max} = \log(1 + \frac{1}{2}\tau_{\max}^2)$, obtained

by applying the same transform to $\tau_{\max}$, is expressed in pixels and consistent across image sizes. Eq. (E.2) shows the pinhole map; the optional radial distortion coefficient, when enabled, warps the normalized coordinates before the final multiplication by $\mathbf{K}_j$.

Quality Score. Following (Zhu et al., 2026), we collect all valid residuals $R = \{\bar{r}_{i,j,k} \mid (i,j) \in \mathcal{G},\ k \in [\kappa]\}$, excluding invalid projections such as negative depth or excessively large reprojection error. We model them as an empirical distribution $\bar{r} \sim \mathcal{R} = \mathrm{KDE}(R)$ with CDF $F(\bar{r}) = \Pr[\bar{r}' < \bar{r}]$, estimated by a histogram-based KDE with T bins over $[0, \bar{\tau}_{\max}]$, so the CDF-area score of Eq. (2) is computed as

$$\mathcal{Q}(\mathcal{X}) \ \approx \ \frac{1}{T+1} \sum_{t=0}^{T} F\big(\tfrac{t}{T}\bar{\tau}_{\max}\big). \tag{E.3}$$

We maximize $\mathcal{Q}$ over the poses, depth adjustments, and intrinsics by annealing from robust log-residuals to raw pixel residuals; the stage schedule is in App. D and the solver in App. E.5.

Surrogate Loss. Maximizing $\mathcal{Q}$ directly is impractical, so we minimize the surrogate loss of (Zhu et al., 2026), written in terms of the residual CDF F and its density F':

$$\mathcal{L}_{\mathrm{MBA}} = \frac{1}{|R|} \sum_{i,j,k} F(\bar{r}_{i,j,k}) \cdot \mathbb{1}\big[\bar{r}_{i,j,k} < \bar{\tau}_{\max}\big], \tag{E.4}$$

where $\mathbb{1}[\cdot]$ is the indicator function, so residuals above the threshold contribute neither value nor gradient. Because F is non-decreasing, lowering $\mathcal{L}_{\mathrm{MBA}}$ drives each residual toward zero, which is what raises $\mathcal{Q}$. Its gradient is

$$\frac{\partial \mathcal{L}_{\mathrm{MBA}}}{\partial \bar{r}_{i,j,k}} = \frac{1}{|R|} F'(\bar{r}_{i,j,k}) \cdot \mathbb{1}\big[\bar{r}_{i,j,k} < \bar{\tau}_{\max}\big], \tag{E.5}$$

which weights each residual by the local density of the residual population, so residuals in the sparse outlier tail exert little influence.

E.5 Second-Order Solver

Let $\mathbf{e}_k \in \mathbb{R}^2$ be the pixel reprojection error of correspondence k and $r_k = \|\mathbf{e}_k\|_2$. Since $\mathcal{L}_{\mathrm{MBA}}$ depends on θ only through the r_k, Eq. (E.5) gives a per-correspondence sensitivity $g_k = \partial \mathcal{L}_{\mathrm{MBA}} / \partial r_k \geq 0$, and setting $w_k = g_k / r_k \geq 0$ makes the reweighted least-squares surrogate

$$\mathcal{Q}_{\mathrm{IRLS}}(\theta) = \tfrac{1}{2} \sum_k w_k \, \|\mathbf{e}_k(\theta)\|^2 \tag{E.6}$$

share its gradient at the current iterate, since $\partial_\theta[\frac{1}{2} w_k \|\mathbf{e}_k\|^2] = g_k \, \partial_\theta r_k$. The two objectives therefore share a gradient field and stationary points, so replacing the first-order optimizer changes the convergence path but not the fixed points. This matches the gradient of the surrogate rather than of Eq. (2) itself, whose gradient is uniform over inliers; we optimize the surrogate because its density weighting supplies the curvature exploited below. Because $w_k \geq 0$, the Gauss–Newton matrix $\mathbf{H} = \sum_k w_k \mathbf{J}_k^\top \mathbf{J}_k$ built from the Jacobians $\mathbf{J}_k = \partial \mathbf{e}_k / \partial \theta$ is positive semi-definite, so the damped system $(\mathbf{H} + \lambda \mathbf{I})\, \Delta = -\mathbf{g}$ is a descent step and λ regularizes the curvature rather than opposing it. We form $\mathbf{H}$ and $\mathbf{g} = \sum_k w_k \mathbf{J}_k^\top \mathbf{e}_k$, rebuilding the weights each iteration, and take Levenberg–Marquardt steps.

Several details matter in practice. Rotations are stored as the 6D representation but optimized in the $\mathfrak{se}(3)$ tangent, re-anchored after each accepted step, since the redundant storage would otherwise contribute spurious null directions to $\mathbf{H}$. Increments are applied in the camera frame, $\mathbf{T} \leftarrow \exp(\xi)\, \mathbf{T}$ for the camera-to-world pose $\mathbf{T}$ and tangent increment $\xi \in \mathfrak{se}(3)$, which keeps the tangent step approximately linear in the measured pixel displacement. The unknowns are 10 per frame, the $\mathfrak{se}(3)$ tangent plus the four depth-affine and depth-intrinsic terms, together with shared focal and distortion blocks. Stages (i)–(iii) marginalize structure, so no Schur complement is required. We exploit the covisibility graph's sparsity, storing $\mathbf{H}$ in block-sparse form and factoring it with a sparse Cholesky. Free-point refinement instead adds a point block, but each point is observed in exactly two views, so that block is 3×3 block-diagonal and is eliminated in closed form, trading track length for the closed-form elimination. Finally, $\mathbf{H}$ spans many orders of magnitude across parameter

blocks, so we solve in double precision with symmetric Jacobi preconditioning and a floored damping term, the gauge freedom being fixed by a per-block anchor. Because $\mathbf{H}$ and $\mathbf{g}$ are plain sums over correspondences, sharding pairs across GPUs and reducing the two accumulators reproduces the single-device system exactly.

E.6 Scale Initialization

Closed-Form Geometric Depth. For each correspondence (p_k, q_k) between frames i and j, the camera projection equation in normalized coordinates reads:

$$d'_k \, \tilde{q}_k = d_k^{\text{geo}} \, \mathbf{R}_{j|i} \, \tilde{p}_k + \mathbf{t}_{j|i}, \tag{E.7}$$

where $\tilde{p}_k = \mathbf{K}_i^{-1}[p_k^x, p_k^y, 1]^\top$ and $\tilde{q}_k = \mathbf{K}_j^{-1}[q_k^x, q_k^y, 1]^\top$ are normalized coordinates, and d_k^{geo}, d'_k are the depths in frames i and j. Solving for d_k^{geo} in closed form via the epipolar constraint gives the geometric depth, and dividing it by the monocular depth at the same pixel gives the candidate scale $\hat{s}_{i,k} = d_k^{\text{geo}} / \mathbf{D}_i[p_k]$ used in Eq. (3).

Scale Normalization. The reconstruction is determined only up to a global scale, and we fix that freedom so the depth maps stay close to their metric predictions rather than drifting with the cameras. Let $\bar{s} = \exp\big(\text{median}_{j\in[N]} \log s_j\big)$ be the median scale; we divide the per-frame scales, the depth biases, and the translations by it, $s_i \leftarrow s_i/\bar{s}$, $b_i \leftarrow b_i/\bar{s}$, and $\mathbf{t}_i \leftarrow \mathbf{t}_i/\bar{s}$, leaving $\text{median}_i \, s_i = 1$. The bias must be rescaled alongside the scale: by Eq. (E.2) a 3D point is $(s_i \mathbf{D}_i[p_k] + b_i)\, \mathbf{K}_i^{-1} \tilde{p}_k$, so dividing s_i and b_i together maps every point to $\mathbf{X}_k/\bar{s}$, matching the translations and leaving the geometry unchanged up to the intended similarity.

E.7 Label Preparation

Both supervised and pseudo-labeled training require target camera parameters, and no collection supplies the depth scale target directly: annotated ones are either metric or, when SfM-derived, defined only up to scale, and neither convention matches the frozen monocular estimator. We therefore adopt the estimator's scale as the common reference for every collection. Given ground-truth poses $\{\mathbf{P}_i^{\text{gt}}, \mathbf{K}_i^{\text{gt}}\}$, we recover the scale targets by running the RANSAC scale initialization (Eq. (3)) with those poses held fixed. For collections without ground truth, we instead run the full BA pipeline of Sec. 3.3 to produce pseudo ground truth $\{\hat{\mathbf{P}}_i, \hat{\mathbf{K}}_i, \hat{s}_i\}$ over the retained subset $\mathcal{V}^*$. In both cases we normalize the scale targets to unit median, $s_i^{\text{gt}} \leftarrow s_i^*/\bar{s}^*$ with $\bar{s}^* = \text{median}_j \, s_j^*$, and rescale the target translations by the same factor, $\mathbf{t}_i^{\text{gt}} \leftarrow \mathbf{t}_i^{\text{gt}}/\bar{s}^*$, so that $\text{median}_i \, s_i^{\text{gt}} = 1$ and the pose and depth targets stay mutually consistent in the monocular estimator's scale. The supervised loss aggregates its per-component terms over the camera head's L iterative predictions with a decay $\gamma < 1$, so that the final iteration receives unit weight:

$$\mathcal{L}_{\text{sup}} = \sum_{\ell=1}^{L} \gamma^{L-\ell} \sum_{c\in\{R,t,f,s\}} \lambda_c \, \mathcal{L}_c^{(\ell)}, \tag{E.8}$$

where $\mathcal{L}_c^{(\ell)}$ are the L_1 errors on relative rotation $\mathbf{R}_{j|i} = \mathbf{R}_j \mathbf{R}_i^\top$, relative translation $\mathbf{t}_{j|i} = \mathbf{t}_j - \mathbf{R}_{j|i}\mathbf{t}_i$, field-of-view, and depth scale.

F Additional Qualitative Results

We provide per-dataset galleries of reconstructions across all evaluated scenes, complementing the representative results in Fig. 5. Figs. F.1 to F.3 show one fly-through view per scene for 1DSfM, ETH3D, and Tanks & Temples respectively. Fig. F.4 adds GLDv2, where the collections are unconstrained internet photos and no ground-truth poses exist to score against. Fig. F.5 shows 37 panels: 33 scenes drawn from six of the FastMap benchmark's seven collections, spanning object-centric captures, indoor rooms, ground-level facades, and large aerial surveys, together with four Zip-NeRF scenes reconstructed by the same model but not among the splits scored in Tab. 4. The benchmark's remaining collection is Tanks & Temples, covered separately by Fig. F.3.

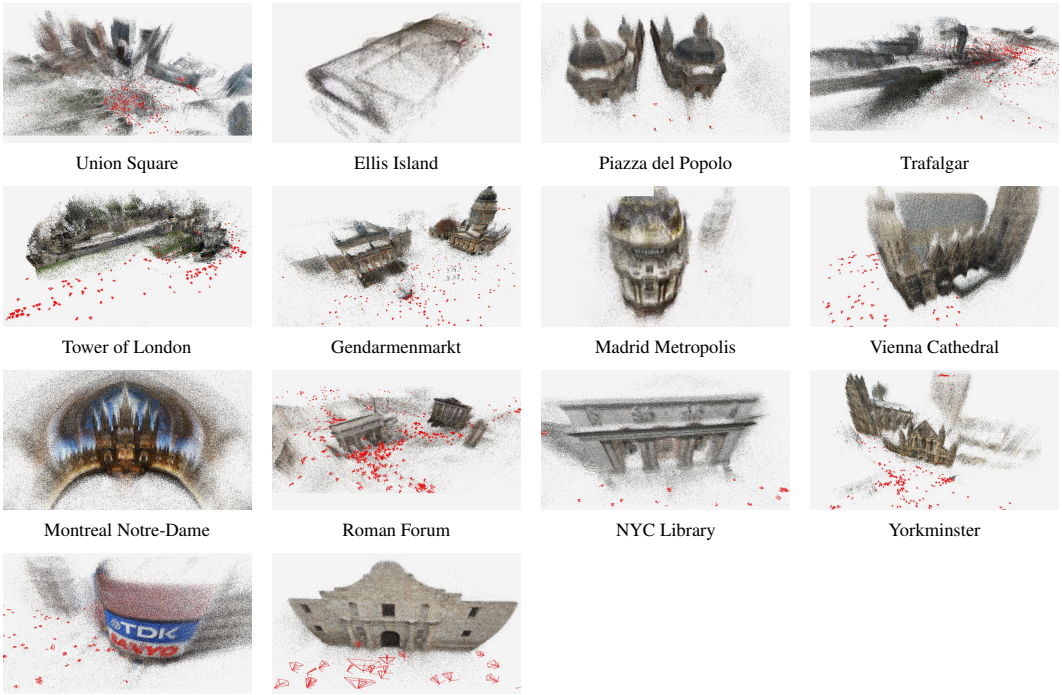


Figure F.1: Qualitative reconstructions on 1DSfM (Wilson & Snavely, 2014). All scenes, one fly-through view each. Each panel renders the reconstructed point cloud together with the recovered camera frusta, drawn in red.

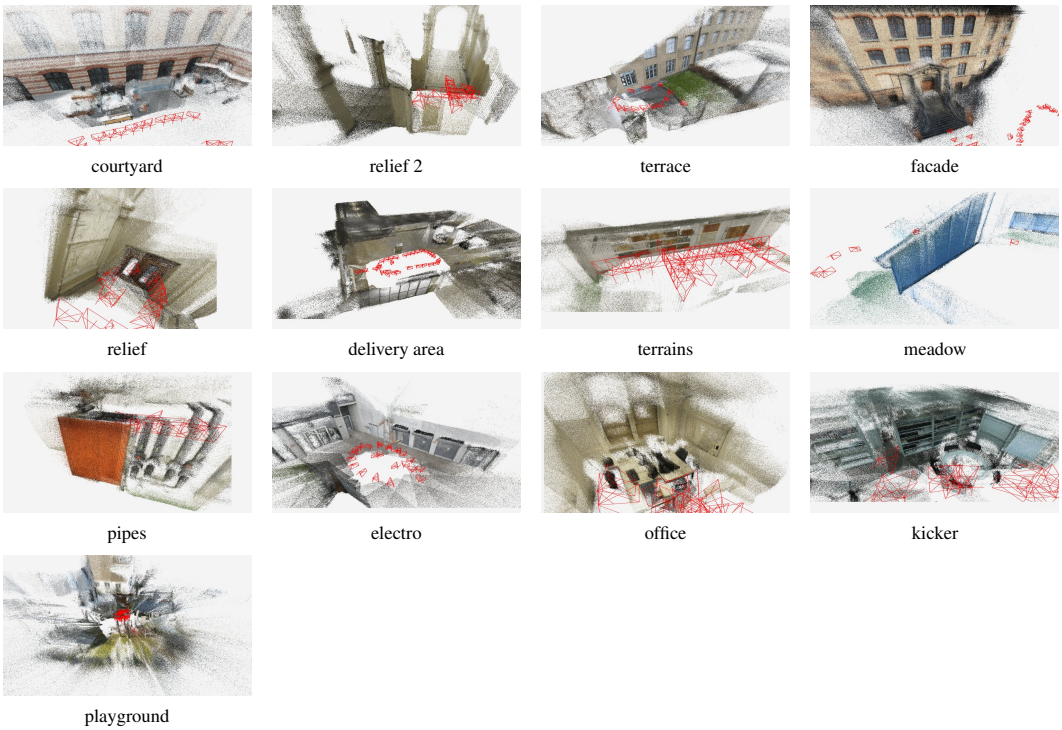


Figure F.2: Qualitative reconstructions on ETH3D (Schöps et al., 2017). All scenes, one fly-through view each. Each panel renders the reconstructed point cloud together with the recovered camera frusta, drawn in red.

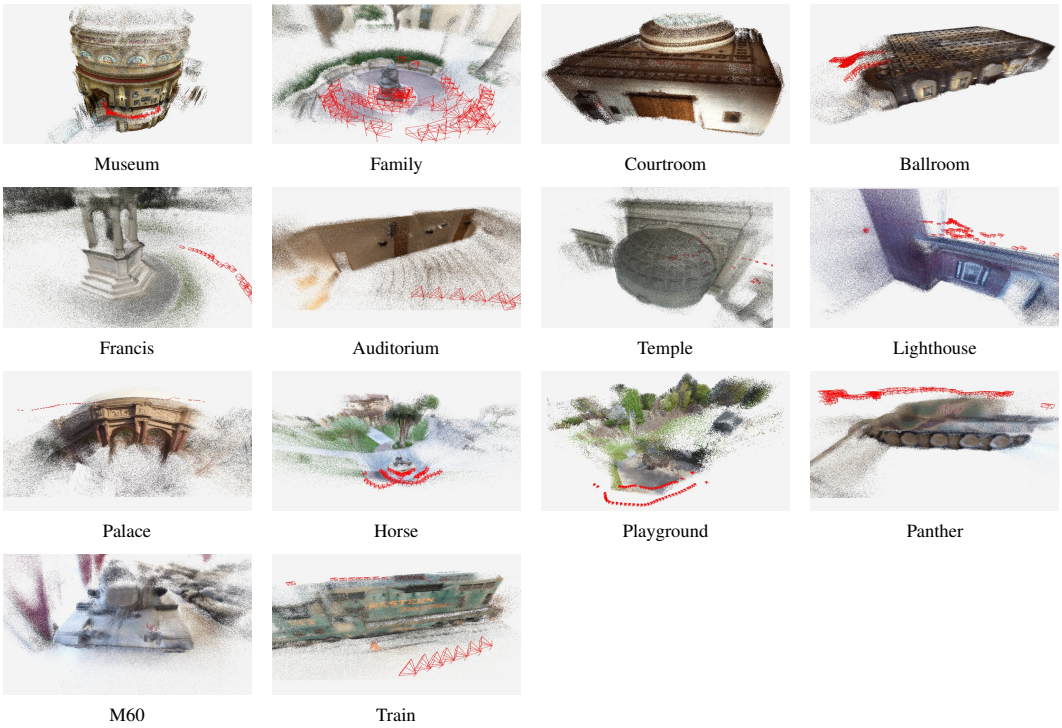


Figure F.3: Qualitative reconstructions on Tanks & Temples (Knapitsch et al., 2017). The 14 Intermediate and Advanced scenes, one fly-through view each. Each panel renders the reconstructed point cloud together with the recovered camera frusta, drawn in red.



Figure F.4: Qualitative reconstructions on GLDv2 (Weyand et al., 2020). One fly-through view per landmark, labeled by Google Landmarks v2 landmark id. Each panel renders the reconstructed point cloud together with the recovered camera frusta, drawn in red. These are unconstrained internet photo collections, with no ground-truth poses to supervise or score against.

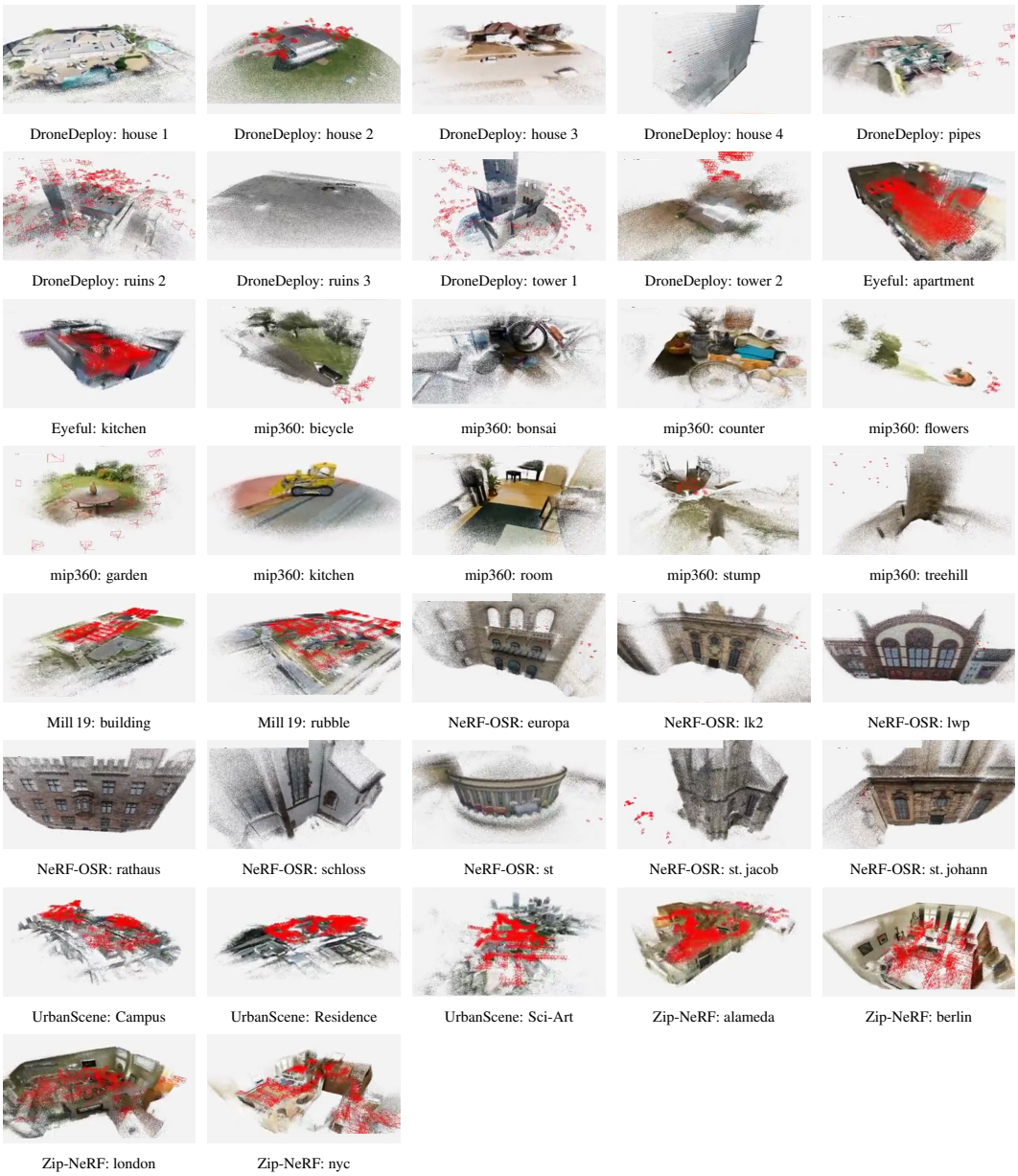


Figure F.5: Qualitative reconstructions on the FastMap benchmark (Li et al., 2026). Each panel renders the reconstructed point cloud together with the recovered camera frusta, drawn in red.