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

CVPR 2025

Supplementary material

In the following pages, we present additional details on the experiments conducted in the main paper.

A. Qualitative results

In Fig. 5, we show reconstruction results of COLMAP [50], MASt3R-SfM [16], and our approach in low-overlap scenarios. COLMAP, which relies on three-view tracks, fails to register many of the images. MASt3R-SfM successfully registers all images but does so in an incorrect way. It struggles in larger scenes (rows 1 and 2) and in the presence of symmetries (rows 3 and 4). In contrast, our pipeline successfully reconstructs these scenes with high accuracy.

B. Depth refinement

To refine our depth maps, we use the normal integration approach introduced by Cao *et al.* [11]. We extend their cost function to account for uncertainties in the monocular surface normals. The residual of the normal component of our integration cost is $\mathbf{r}_i(u,v) = N_i(u,v) - \Delta D_i^*(u,v) \in \mathbb{R}^4$ for each image i and pixel (u,v). We drop the indexing for the remainder of this section. The residual is expressed as

$$\mathbf{r} = \begin{pmatrix} r_u^+ \\ r_u^- \\ r_v^+ \\ r_v^- \end{pmatrix} = \begin{pmatrix} \tilde{n}_{z,x}\, \partial_u^+ D^* + n_x \\ \tilde{n}_{z,x}\, \partial_u^- D^* + n_x \\ \tilde{n}_{z,y}\, \partial_v^+ D^* + n_y \\ \tilde{n}_{z,y}\, \partial_v^- D^* + n_y \end{pmatrix} \quad , \tag{7}$$

where the terms

$$\begin{aligned} \tilde{n}_{z,x} &= n_x(u - c_x) + n_y(v - c_y) + n_z \cdot f_x, \\ \tilde{n}_{z,y} &= n_x(u - c_x) + n_y(v - c_y) + n_z \cdot f_y, \end{aligned} \tag{8}$$

simplify the perspective case of the normal integration equations. n_x, n_y and n_z are the three component of each normal estimate, and f and $c \in \mathbb{R}^2$ are the focal length and principal point of the camera, while $\partial_{u/v}^{\pm} D^*$ are the discretized one-sided partial derivatives of the refined depth map D^* [11].

To minimize the integration cost jointly with other costs, they should be weighted by their uncertainties. As such, we propagate the normal uncertainties Σ_N into residual uncertainties $\Sigma_{\mathbf{r}} = \operatorname{diag}(\sigma^2_{N_u^+}, \sigma^2_{N_u^-}, \sigma^2_{N_v^+}, \sigma^2_{N_u^-})$. In the following, we derive and approximate $\sigma_{N^{\pm}_{u/v}} \approx \sigma_{N_{u/v}}$.

Monocular normal estimators [4, 28] estimate angular isotropic uncertainties, which are projected from the Spherical into the Cartesian coordinate system using

$$\Sigma_{xyz} = J_{xyz}\, \Sigma_{\theta,\varphi}\, J_{xyz}^\top, \tag{9}$$

where

$$\begin{aligned} \Sigma_{\theta\varphi} &= \begin{bmatrix} \sigma_\theta^2 & 0 \\ 0 & \sigma_\varphi^2 \end{bmatrix} \\ J_{xyz} &= \begin{bmatrix} \cos\theta\cos\varphi & -\sin\theta\sin\varphi \\ \cos\theta\sin\varphi & \sin\theta\cos\varphi \\ -\sin\phi & 0 \end{bmatrix} \quad . \end{aligned} \tag{10}$$

Then, we express the uncertainties in the residual space $\mathbf{r}$ as

$$\sigma^2_{N_{u/v}} = J_{u/v}\, \Sigma_{xyz}\, J_{u/v}^\top \quad , \tag{11}$$

where

$$\begin{aligned} J_u &= \begin{bmatrix} (u - c_u)\partial_u D^* + 1 & (v - c_v)\partial_u D^* & \partial_u D^*\, f_x \end{bmatrix}, \\ J_v &= \begin{bmatrix} (u - c_u)\partial_v D^* & (v - c_v)\partial_v D^* + 1 & \partial_v D^*\, f_y \end{bmatrix}. \end{aligned} \tag{12}$$

Here, to make the computation tractable, we approximate $\partial_u z = -\frac{n_x}{\tilde{n}_z} \approx \partial_u^{\pm} z$ and $\partial_u z = -\frac{n_y}{\tilde{n}_z} \approx \partial_v^{\pm} z$.

In addition to the normal uncertainty estimates, we approximate normal uncertainties using a flip consistency check between the normal estimates of the original and flipped images. In the spherical coordinates, we compute their mean $\bar{n}_{\theta,\varphi}$ and covariance

$$\Sigma_{\theta,\varphi} = \begin{bmatrix} \angle_{1,\theta}^2 + \angle_{2,\theta}^2 & \angle_{1,\theta}\angle_{1,\varphi} + \angle_{2,\theta}\angle_{2,\varphi} \\ \angle_{1,\theta}\angle_{1,\varphi} + \angle_{2,\theta}\angle_{2,\varphi} & \angle_{1,\varphi}^2 + \angle_{2,\varphi}^2, \end{bmatrix} \tag{13}$$

where $\angle_{1,\theta}$ is the angular difference between the θ component of the original image $n_{1,\theta}$, and $\bar{n}_\theta$. As we theorize that it is better to overestimate the uncertainties, we take the maxima between the estimate uncertainties and the uncertainties derived from flip consistency.

Figure 6 compares monocular depth priors and refined depth maps to the ground truth. In contrast to the priors, our refined depth maps are aligned with the ground truth.

C. Prior uncertainties

We analyze the calibration of the uncertainties predicted by Metric3D-v2 [28] and MASt3R [34] for the depth priors. We consider the training split of the ETH3D [53], which has sparse ground truth depth maps obtained with laser scanners. Figure 7 shows calibration plots that compare each uncertainty with the actual depth error. We calibrate the uncertainties by scaling them down by a constant factor which was tuned on a different dataset given sparse SfM point clouds as

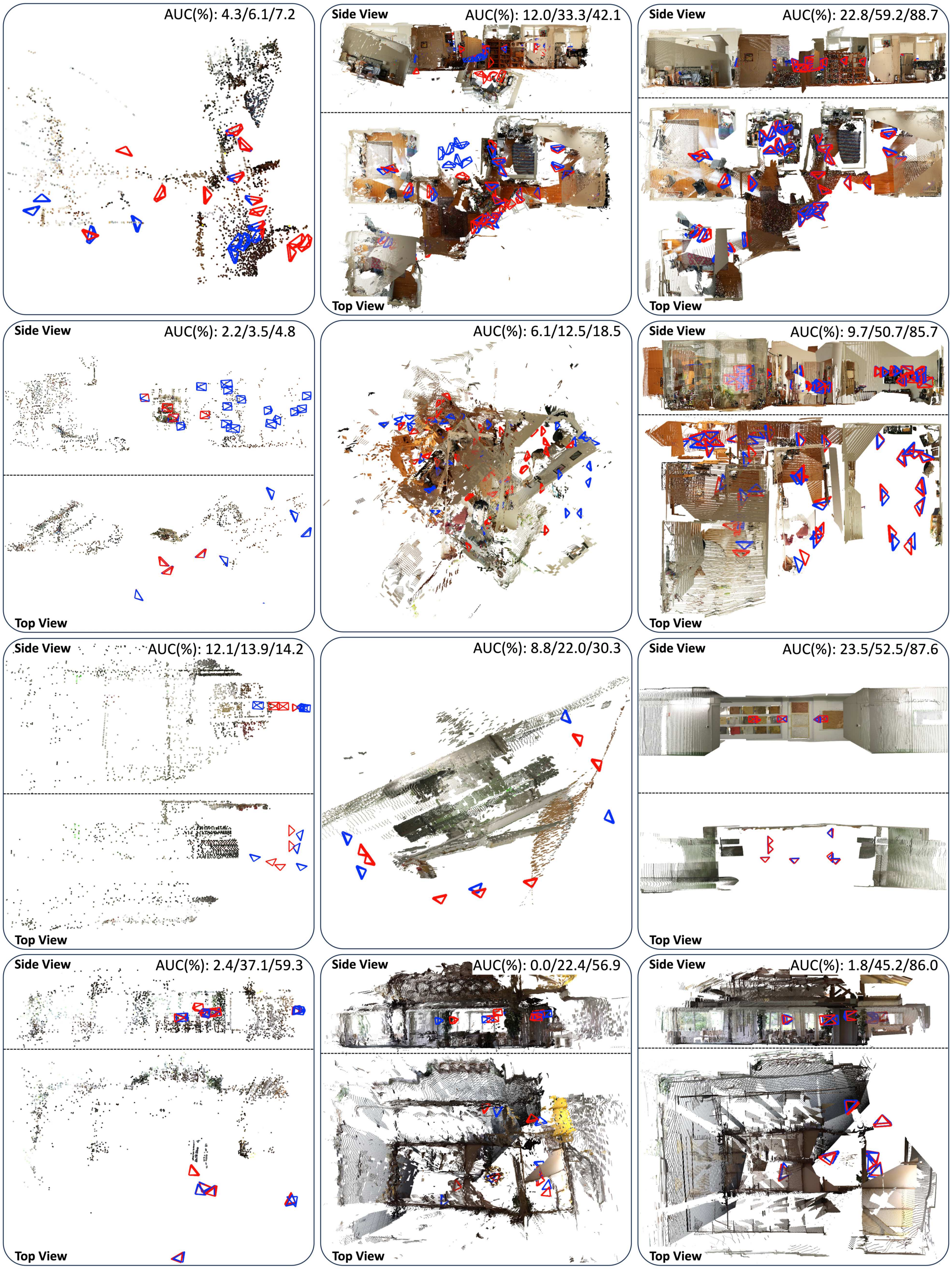

Figure 5. **Qualitative comparison of reconstructions for low-overlap scenes.** Estimated (red) and ground-truth (blue) camera poses, and AUC accuracies at $1^\circ/5^\circ/20^\circ$ error thresholds are presented. Left: COLMAP [50]. Center: MASt3R-SfM [16]. Right: Our method. Rows 1–2 show scenes from SMERF [15], while rows 3–4 are from ETH3D [53] and Tanks and Temples [33].

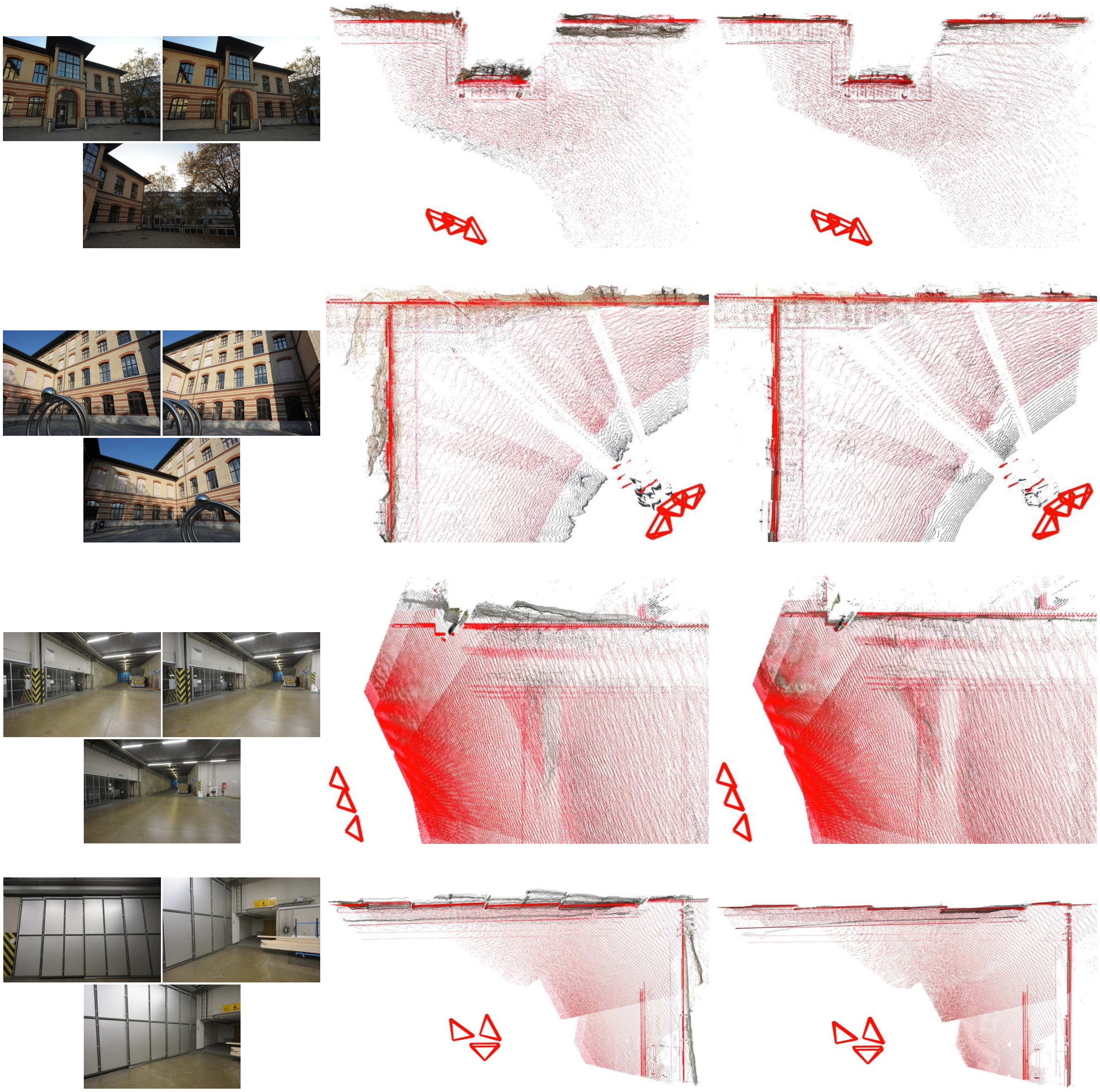

Figure 6. **Visualizations of prior and refined depth map.** For four reconstructions of the ETH3D datasets, we show the input images (left) and the colored point clouds obtained by unprojecting the monocular prior depth maps (center) and the refined depth maps (right). We overlay the points obtained using the ground-truth maps in red. The refined depth maps are closer to the ground truth and more consistent across views.

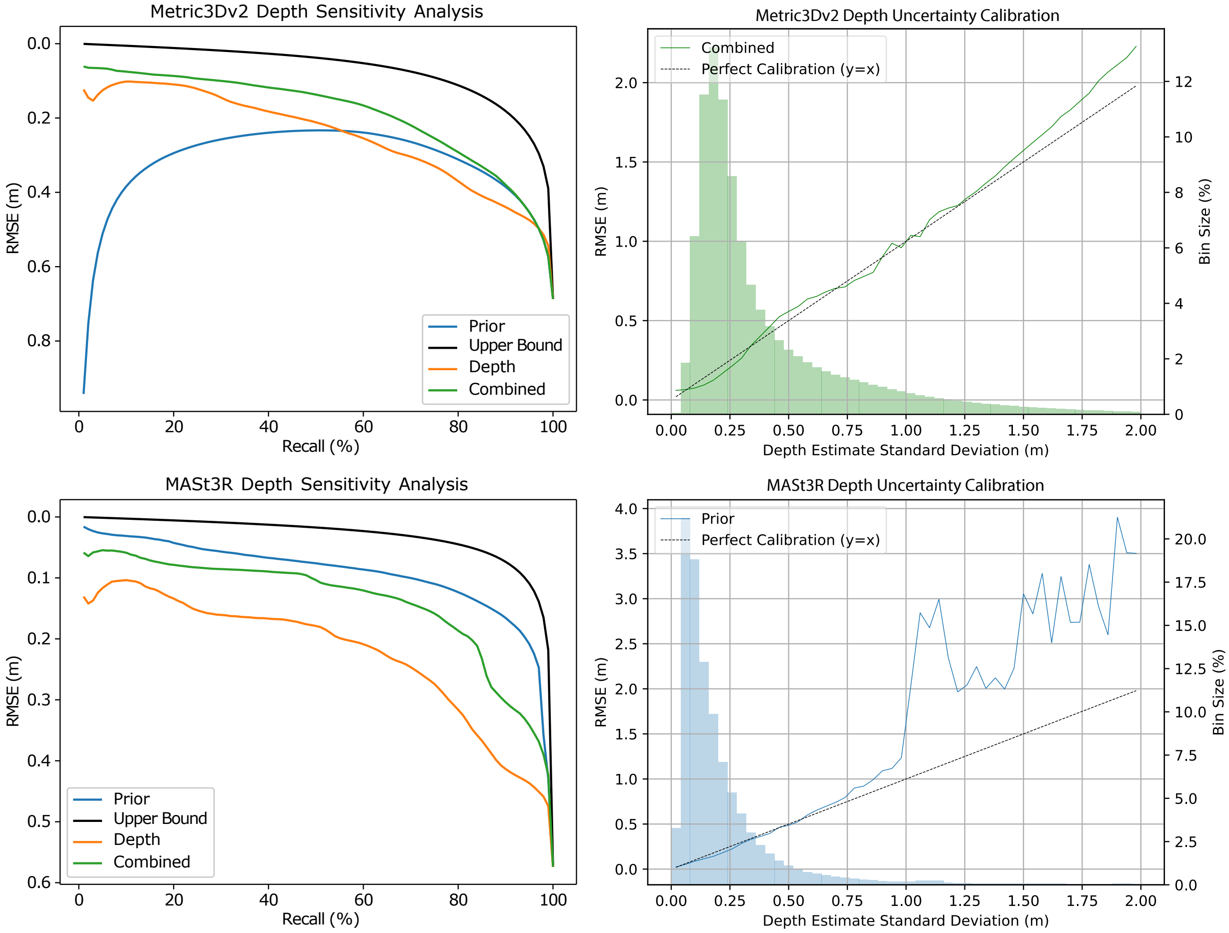

Figure 7. **Analysis of the prior depth uncertainties on the ETH3D [53] dataset.** Left: The sensitivity analysis of Metric3D-v2 [28] and MASt3R [16] depth estiamtes shows the total RMSE for the X% of pixels (recall) with the lowest uncertainty. The uncertainties include the monocular prior prediction, a depth proportional uncertainty, and the combination of the two, combined via per pixel maxima. The *Upper Bound* is based on ground truth RMSE. While we found that using the combination yileded more reliable uncertainties for Metric3D-v2, the prior uncertainties predicted by MASt3R alone was the most reliable. Right: The calibration plot for the selected combination of depth uncertainties per model, optimized using a constant scaling factor. Histograms show the amount of depth estimates belonging to the estimated uncertainty bins.

pseudo ground truth. This improves the calibration overall, except for the 10% most confident pixels in Metric3D-v2 depth estimates.

To handle these unreliable uncertainties, we clip the standard deviations at a minimum of 2 cm and we augment them with an uncertainty proportional to the depth estimate. The total uncertainty is the maximum of the scaled uncertainty and the depth-proportional uncertainty. We select the scaling factor of the depth-proportional uncertainty by maximizing the AUC of the sensitivity analysis plots (left). We also apply robust loss functions during bundle adjustment and depth refinement.

In the case of MASt3R depth, the predicted depth uncertainty alone was the most reliable. To calibrate the other monocular depth estimators explored in Tab. 4, we followed a similar approach. Although we used the same setup for the different sizes of the Metric3D-v2 models, Depth Anything V2 [70] and Depth Pro [6] do not predict uncertainties. We explored using a flip consistency check to estimate uncertainties; however, the improvements over the depth-proportional uncertainties were marginal.

D. Leveraging two view correspondences

In Fig. 8, we illustrate our approach for leveraging dense matches in non-salient regions. Building long tracks in salient regions is crucial for high accuracy in high-overlap scenarios. However, in low-overlap scenarios, leveraging matches in non-salient regions improves performance during

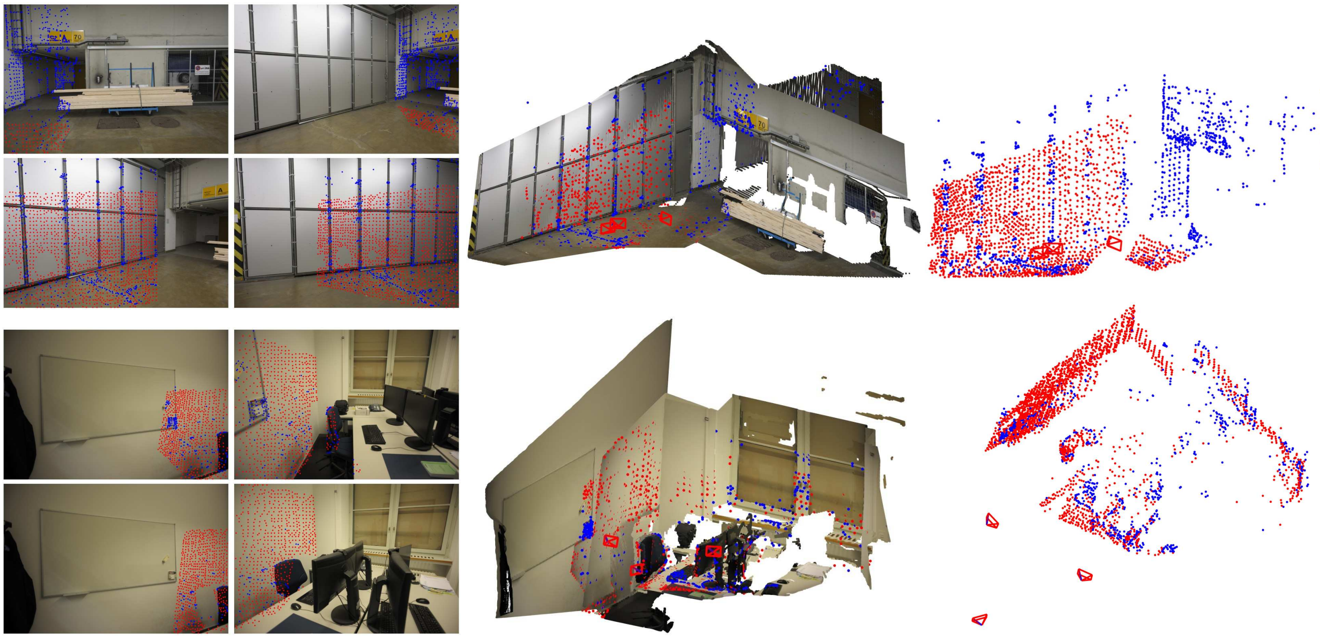

Figure 8. **Leveraging dense matching to build long tracks in salient regions and two-view tracks in featureless areas.** Left: sparse matches sampled at SuperPoint [13] keypoints, and in textureless areas. Middle: dense reconstruciotn of the scene with the colmap points overlayed. Right: Visualization of only the triangulated points.

both registration and bundle adjustment. For this reason, we use dense matchers not only to match sparse salient features, but also to sample matches in non-salient regions that later form two-view-only tracks. However, we found that with the MASt3R [34] matcher, using these sampled two-view tracks degrades performance due to low match precision.

E. Efficiency Analysis

Table 7 presents the efficiency analysis of our pipeline for three scenes with 25, 44, 110 and 505 images. Depth refinement constitutes the most significant overhead compared to COLMAP [50]. During post-registration refinement, the depth map of a newly registered image is refined first. If the image passes the depth consistency check, its depth map undergoes another refinement during local refinement. Notably, this second refinement converges faster as the depth map is already close to the optimization minima, evidenced by the reduced processing time.

The global refinement is performed periodically during the incremental reconstruction pipeline and refines all depth maps. Since reconstructions may not change significantly between consecutive global refinements, many depth maps remain near their optimization minima. To avoid unnecessary computation, we compare the total refinement costs of the previous and current calls. If the costs differ insignificantly, the refinement is skipped. In reconstructions with 25, 44, 110 and 505 images, 309, 724, 2186 and 10935 refinements were skipped out of 520, 1071, 2860 and 13566 calls, respectively.

(seconds)	# images:	25	44	110	505
Global Refinement	Bundle Adjustment	2.76	18.8	45.4	610
	Depth Refinement	4.39	19.6	44.3	268
	3D Point Covariances	0.96	6.49	11.5	64.4
Local Refinement	Bundle Adjustment	0.47	1.91	2.34	19.1
	Depth Refinement	0.21	0.73	1.19	9.94
	3D Point Covariances	0.80	2.57	4.66	55.5
Post Registration Refinement	3D Point Refinement	0.20	0.63	0.85	7.50
	Depth Refinement	2.33	2.71	15.9	40.8
	3D Point Covariances	0.50	2.00	3.78	33.9
	Depth check	0.44	0.73	2.01	13.9

Table 7. **Efficiency analysis of the significant components of our pipeline.** Results, cumulated over all function calls, are presented for reconstructions of scenes from ETH3D [53] of sizes 25, 44, 110 and 505 images. The components we add to COLMAP are highlighted in blue.

In addition to depth refinement, 3D point covariance computation introduces notable overhead. However, the computational cost of the depth consistency check is negligible.

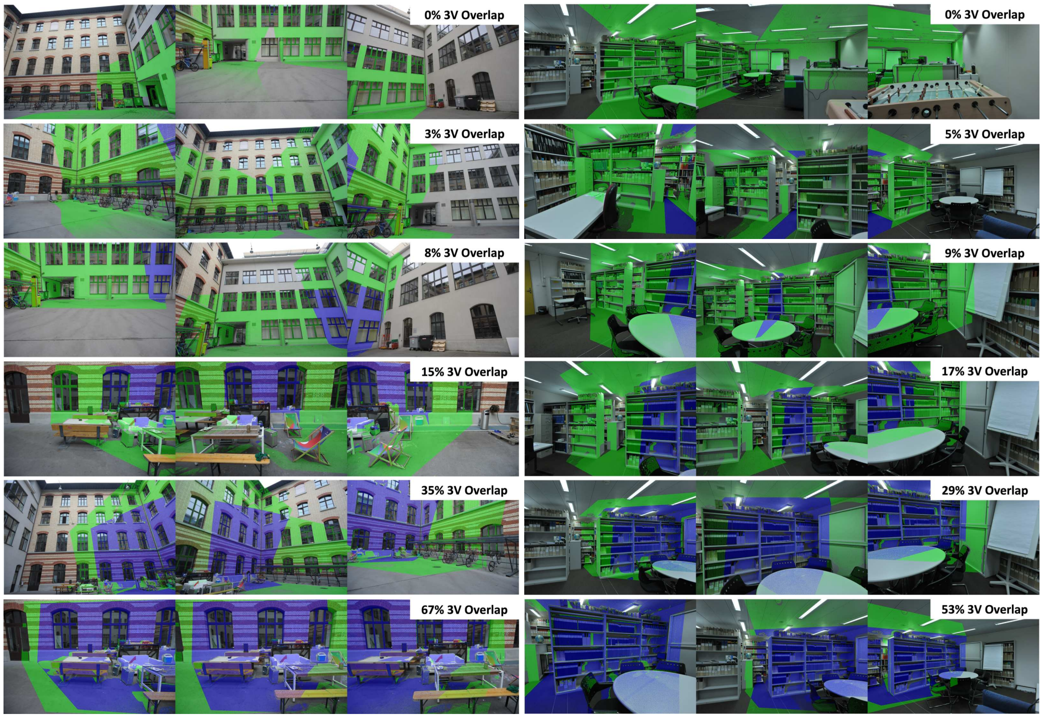

Figure 9. **Visual examples of our triplet test set.** Each row corresponds to a triplet from one of our triplet test set categories: 0%, $[0\%, 5\%]$, $[5\%, 10\%]$, $[10\%, 20\%]$, $[20\%, 40\%]$, and $> 40\%$, respectively. Triplets are colored by two-view overlap and three-view overlap. All triplets are from the ETH3D [53] dataset. Left: examples from the Courtyard scene. Right: examples from the Kicker scene.

F. Implementation details

F.1. Low-overlap evaluation

To evaluate the performance of our pipeline under varying levels of image overlap, we constructed test sets by incrementally sampling images based on their overlap with existing images in the scene. This approach allows us to simulate low- to high-overlap scenarios.

Sampling Criteria: Images are added to the test set if they satisfy two conditions: 1) The three-view (3V) overlap with existing test set images does not exceed the target threshold. 2) There is sufficient two-view (2V) overlap with at least one image that has been already selected.

Dataset-Specific Overlap Determination: For each dataset, the overlap is determined using different criteria:

- ETH3D [53]: for the training set we use the ground truth depth maps; for the test set we use depth maps estimated by multi-view stereo with COLMAP [51].
- SMERF [15] and Tanks and Temples [33]: the overlap is based on the ratio of detected sparse keypoints in an image with the number of common 3D points in a retriangulated SfM point cloud.

Minimal test sets: For the minimal (zero 3V overlap) test sets, it is often infeasible to sample images spanning the entire scene. In such cases, images with some 3V overlap are sampled, provided they result in at least one other image with zero 3V overlap.

Test test construction and statistics: For the "minimal" test sets, we sample 10 sets of images per scene, while all other test sets contain 5 sets per scene. Exceptions exist for ETH3D, where some scenes lack sufficient images to meet the overlap criteria. To ensure diversity, each image set includes a minimum of 5 images, and each image set differs from others within a test set by at least two images.

The average number of images per test set, computed across scenes, is presented below:

- ETH3D: **minimal:** 6.7, **<5%:** 6.9, **<10%:** 8.8, **<30%:** 14.0, **all:** 36.9.
- SMERF: **minimal:** 36.4, **low-overlap:** 78.6, **medium:** 100.6, **high:** 170.7.
- Tanks and Temples: **minimal:** 7.8, **low-overlap:** 17.9, **medium:** 26.6, **high:** 45.0.

Triplet Test set: Visualizations of the triplet test set are presented in Fig. 9, ranging from zero to $> 40\%$ 3V overlap.

F.2. Low-parallax evaluation

For the low-parallax evaluation, we reconstruct scenes from the RealEstate10K dataset [74]. Following the sampling strategy of MASt3R-SfM [16], we randomly sample 10 images per scene from 1.8k randomly selected videos.

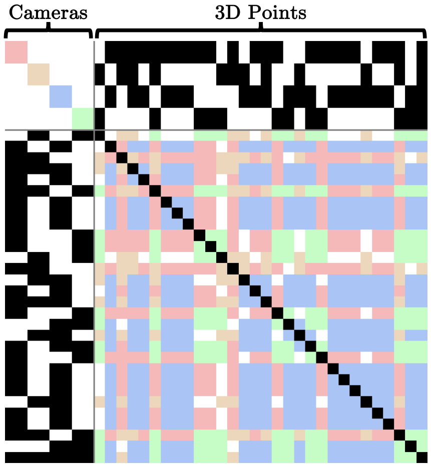

Figure 10. **Sparsity structure in the Hessian of the cost function.** Off-diagonals in the 3D point block are color-coded according to the observing camera, reflecting the per-image normal constraints.

F.3. Baselines

For all baselines, except MASt3R-SfM [16] and VGG-SfM [64], and our approach, we select image pairs using NetVLAD top 20 retrieval. We run both MASt3R-SfM and VGG-SfM with their default approach.

For COLMAP-based baselines, including COLMAP and its variants (Structureless Resectioning and Detector-free SfM), we use the default hyperparameters provided by COLMAP. Additionally, we fix the intrinsics to their ground truth values in all experiments.

To adapt COLMAP [50] for low-parallax scenarios, we tune its triangulation angle hyperparameters. Specifically, the minimum triangulation angle is set to 0.001 for both initialization and during the main loop of the incremental mapper. Additionally, the same minimum triangulation angle of 0.001 is applied during 3D point filtering.

For MASt3R-SfM and VGG-SfM, we similarly use ground truth intrinsics, and fix them during optimization, which yielded the best results. To reconstruct a subset of the high-overlap scenes in ETH3D and Tanks & Temples, as well as the low-overlap scenes in the SMERF dataset, we reduced the number of keypoints in VGG-SfM. This was necessary to avoid running out of memory, but likely impacts reconstruction quality.

G. Structure of the Hessian

In Fig. 10, we present the Hessian structure of our overall cost function proposed in Sec. 3.3. Normal and depth constraints couple 3D points and off-diagonal terms in the point block of the Hessian matrix. This breaks the block diagonal assumption required for the Schur complement.

variant	ETH3D dataset		SMERF dataset	
	min. overlap	all images	min. overlap	high overlap
SP+LG + **ours**	27.3/55.9/71.8	74.3/88.3/92.0	9.2/41.0/69.8	47.3/79.3/90.6
num. of vis. points	25.4/48.9/62.6	71.8/85.2/88.5	9.4/40.8/68.0	44.2/67.9/76.6
num. inlier corresp.	26.7/54.9/70.4	74.8/88.6/92.0	8.9/40.2/67.6	45.8/77.2/90.5
ROMA + **ours**	33.4/60.6/74.4	71.6/87.0/91.3	10.6/41.0/61.8	41.4/69.3/79.4
num. vis. points	33.5/60.1/73.4	68.7/83.8/88.0	9.7/35.7/51.2	41.1/62.2/70.6
num. inlier corresp.	33.5/61.1/75.5	70.1/85.5/89.9	10.9/41.5/61.9	39.3/69.7/81.1
MASt3R + **ours**	34.9/67.2/81.7	70.3/88.2/93.6	17.2/54.6/77.1	56.5/84.4/94.0
num. vis. 3D points	32.9/63.7/77.4	64.9/81.7/86.7	14.2/46.4/65.4	51.6/71.9/78.4
num. inlier corresp.	33.3/64.0/77.8	64.0/80.2/85.2	14.1/46.9/68.9	54.5/78.8/86.2

Table 8. **Ablation of the next view selection.** We compare three approaches applicable in our pipeline. In contrast to prior works [50], *number of visible 3D points*, includes 3D points with a track length one. *Num. inlier corresp.* selects views by counting the maximum number inlier correspondences between query and registered images. This allows our pipeline to effectively reconstruct low-overlap environments. Counting the sum of feature matcher scores between these image pairs instead leads to the best overall performance.

H. Additional Ablations

H.1. Low- to high-overlap dense reconstruction

In Fig. 11, we compare sparse and dense view reconstruction. Our proposed bundle adjustment jointly optimizes camera poses, 3D points, and depth maps. As a result, the point clouds derived from the refined depth maps are well aligned—particularly evident in the second row when observing the building walls.

H.2. Robust reconstruction despite noisy priors

To demonstrate the importance of the robust loss in our objective (Sec. 3.3), we compare reconstructions with and without applying it to the depth term. Despite inaccurate depth priors, our method achieves accurate results. During reconstruction, depth maps are refined and residuals between depths and 3D points are reduced. We observe that using a smaller robust loss scale in the final global bundle adjustment leads to the best accuracy.

H.3. Next view selection

In Tab. 8, we ablate the impact of different next view selection approaches in our pipeline. *COLMAP*'s [50] next view selection maximizes the robustness of registration in traditional SfM pipelines. A natural adaption of this approach that adheres to our registration method is to count the *number of visible 3D points*, including those with track length one. As a result, however, the next view selection score scales with the number of registered images in local bundles, leading to frequent incorrect selection in the case of symmetries.

To handle all levels of overlap, we, instead rely on two-view information. While selecting the next view with the maximum *number of inlier correspondences* to any registered image would be unstable in traditional SfM, lifting 3D points via monocular depth makes the registration robust. However, the performance of this greedy approach still

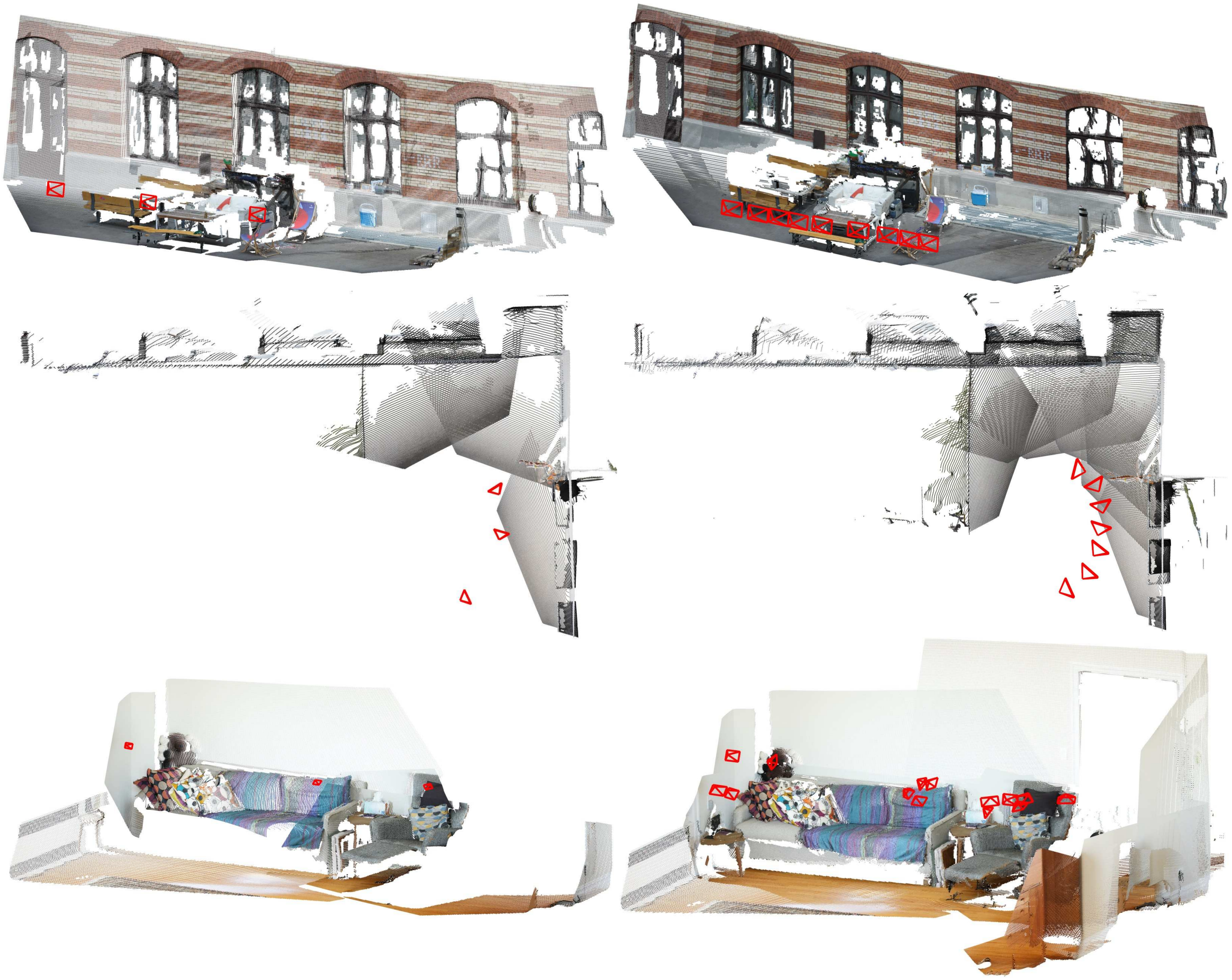

Figure 11. **Comparison between low and high overlap dense reconstruction.** Left: Sparse view reconstruction. Right: dense view reconstruction of the same scene. Multiple views constraining the depth refinements yield consistent depth maps.

largely depends on the quality of the matches.

Instead of counting the number of correspondences between query and registered image pairs, we sum their scores (as predicted by the feature matcher). This leads to better performance in our SP+LG-based [13, 36] pipeline. For dense matchers, selecting next views based on inlier correspondences can cause severe failure cases, especially with *MASt3R* [34], which often hallucinates inlier matches through surfaces. Leveraging matcher scores instead leads to drastic improvements. In the case of *RoMa* [19], the performances of the two approaches are similar.

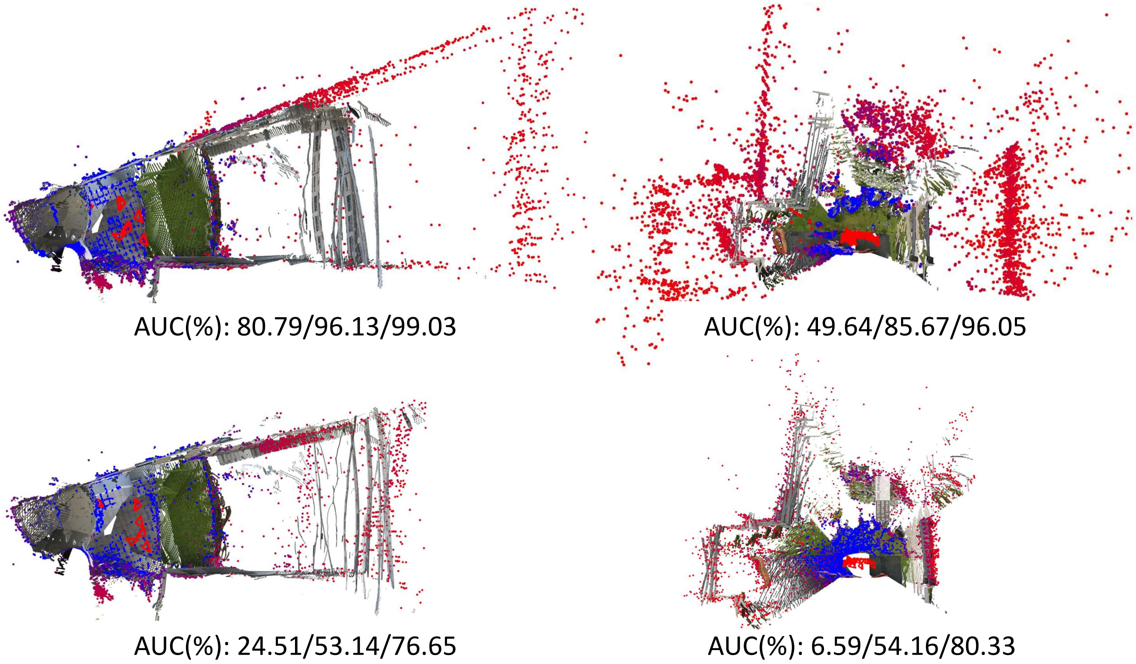
AUC(%): 80.79/96.13/99.03
AUC(%): 49.64/85.67/96.05
AUC(%): 24.51/53.14/76.65
AUC(%): 6.59/54.16/80.33

Figure 12. **Comparing reconstruction quality with and without robust loss.** Depth Anything V2 [70] struggles to estimate depth at large distances. We visualize 3D points with low and high covariance, overlaid on the lifted, refined depth maps. **Top:** Reconstruction using a robust loss in the depth term. **Bottom:** Same scene without the robust loss. Without it, high-covariance points converge to the noisy depth prior, leading to lower reconstruction precision. In contrast, our method achieves accurate reconstruction despite unreliable depth estimates.