MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion
Supplementary Material

Bardienus Pieter Duisterhof[1] Lojze Zust[2,3] Philippe Weinzaepfel[3]
Vincent Leroy[3] Yohann Cabon[3] Jerome Revaud[3]

[1]Carnegie Mellon University [2]University of Ljubljana [3]Naver Labs Europe

https://github.com/naver/mast3r/

1. Qualitative Results

We first present some qualitative reconstruction examples in Fig. 1. These are the raw outputs of the proposed SfM pipeline, without further refinement. We point out that our method produces relatively dense outputs, despite the fact that it only leverages sparse matches. This is because the inverse reprojection function $\pi^{-1}(\cdot)$ (Section 4.2 of the main paper) can be used to infer a 3D point for *every* pixel, *i.e.* not just those belonging to sparse matches. Since MASt3R is limited to image downscaled to 512 pixels in their largest dimension, we can typically produce about 200,000 3D points per image.

2. Other retrieval variants based on MASt3R features

In the main paper, we propose to use ASMK [10] on the token features output from the MASt3R encoder, after applying whitening. In this supplementary material, we compare this strategy to using a global descriptor representation per image with a cosine similarity between image representations. We also compare to a strategy where a small projector is learned on top of the frozen MASt3R encoder feature with ASMK, following an approach similar to HOW [11] and FIRe [13] for training it. Results are reported in Table 1.

For the global representation, we experimentally find that global average pooling performs slightly better than global max-pooling, and that applying PCA-whitening was beneficial and report this approach. However, the performance of such a method remains lower than applying ASMK on the token features (top row).

For learning a projector prior to applying ASMK, we follow the strategy of HOW and FIRe, which show that a model can be trained with a standard global representation obtained by a weighted sum of local features. As training dataset, we use the same training data as MASt3R, compute the overlap in terms of 3D points between these image pairs, and consider as positive pairs any pair with more than 10% overlap, and as negatives pairs any pair coming from two different sequences or datasets. While we observe an improvement in terms of the retrieval mean-average-precision metric on an held-out validation set, this does not yield significant gains when applied to visual localization (bottom row). We thus keep the training-free ASMK approach for MASt3R-SfM.

Retrieval	Aachen-Day-Night		InLoc	
	Day	Night	DUC1	DUC2
MASt3R-ASMK	**88.7/94.9/98.2**	**77.5/90.6/97.9**	58.1/**82.8/94.4**	69.5/90.8/92.4
MASt3R-global	86.7/93.7/97.6	68.6/84.8/93.2	**60.6**/81.8/91.9	66.4/87.8/90.8
MASt3R-proj-ASMK	88.0/94.8/**98.2**	70.2/88.0/94.2	60.1/80.8/91.4	**74.0/92.4/93.1**

Table 1. **Comparison of retrieval based on MASt3R features.** We compare the visual localization accuracy using top-20 retrieved images with ASMK (top row), a global feature representation obtained by averaging pooling the local features and applying whitening (middle row), and ASMK when first learning a projector on top of the MASt3R features (bottom row).

3. Robustness to pure rotations

We perform additional experiments regarding purely rotational cases, *i.e.* situations where all cameras share the same optical center. In such cases, the triangulation step from traditional SfM pipeline becomes ill-defined and notoriously fails. To that aim, we leverage mapping images from the InLoc dataset [9] which are conveniently generated as perspective crops (with a 60° field-of-view) of 360 panoramic images at three different pitch values, regularly sampled every 30°. This leads to bundles of 36 RGB images that exactly share a common optical center. Using regular sampling, we select 20 sequences from the DUC1 and DUC2 sets and use them to evaluate rotation estimation accuracy. Results in terms of RRA@5 in Tab. 2 clearly confirm that methods based on the traditional SfM pipeline

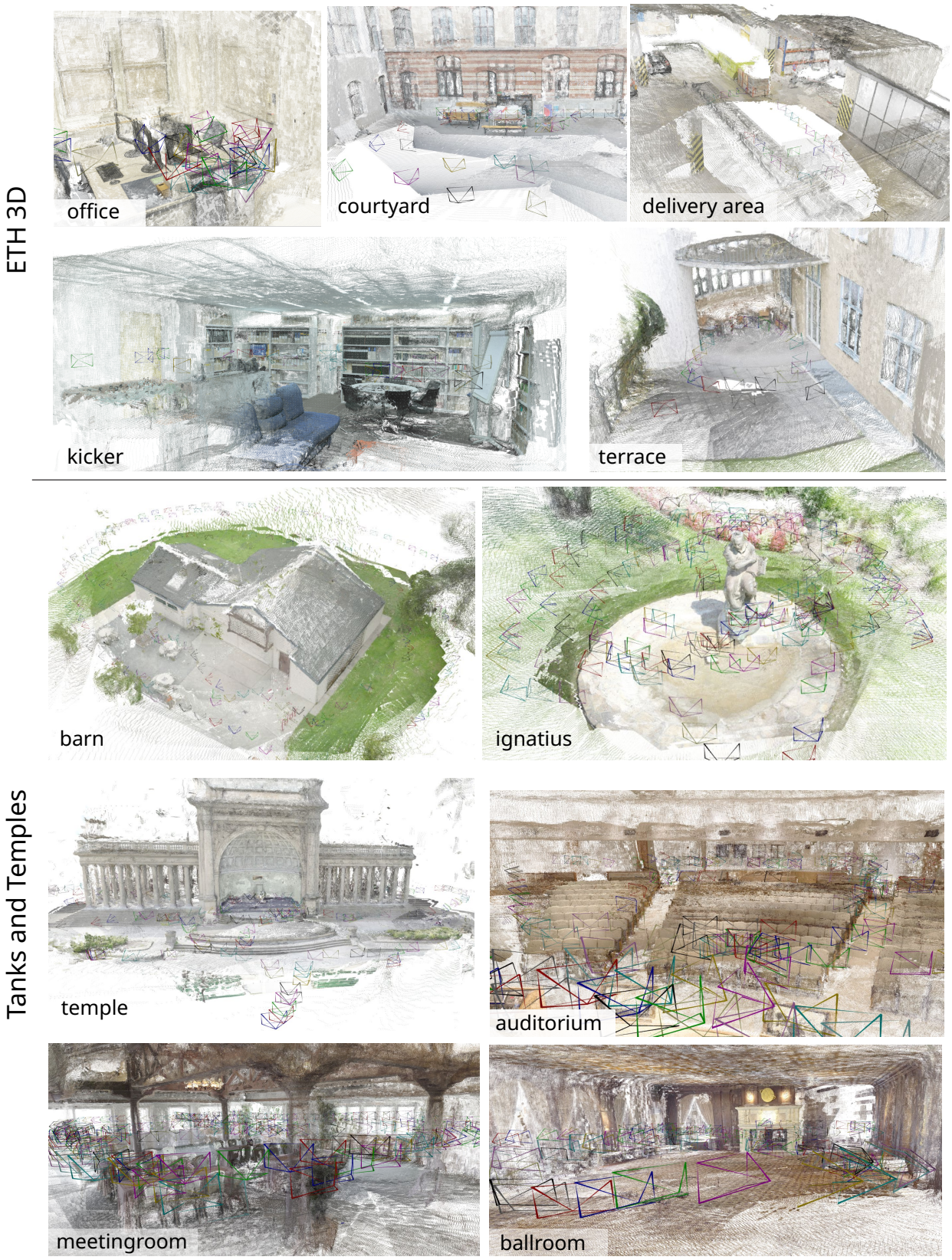

Figure 1. **Qualitative reconstruction results** for MASt3R-SfM on ETH-3D (top) and Tanks&Temples (bottom). These are the raw outputs of the proposed SfM pipeline, without further refinement.

such as COLMAP [7] or VGGSfM [12] do dramatically fail in such a situation. In contrast, MASt3R-SfM performs much better, achieving 100% accuracy on some scenes, even though it also fail in a few cases. Disabling the optimization of anchor depth values (*i.e.* fixing depth to the canonical depthmaps) slightly improves the performance.

Failure cases. After analyzing the results, we observe that failures are due to the presence of outlier (false) matches between similar-looking structures. A few examples of such wrong matching are given in Fig. 2. These are typically hard outliers that would pass geometric verification. In fact, the matching problem in such cases becomes ill-defined, since even for a human observer it can be challenging to notice that the two images show different parts of the scene.

4. Additional Results

More comparisons on CO3D and RealEstate10K. We provide comparisons with further baselines on the CO3D and RealEstate10K datasets for the cases of 3, 5 and 10 input images in Tab. 3. We observe that MASt3R-SfM largely outperforms all competing approaches, only neared by DUSt3R which is much less precise overall.

Detailed Tanks&Temple results. For completeness, we provide detailed results for every scene of the Tanks&Temples dataset [3] in Tab. 5. As mentioned in the main paper, some scenes from T&T are part of the MegaDepth dataset, and thus were used as training data for the MASt3R checkpoint. These scenes are individually marked with a † in Tab. 5 and listed in full in Tab. 4. Importantly, we do *not* observe any significant differences between seen and unseen scenes in terms of accuracies and comparison with the state of the art. For instance, performances on `Courtroom`, `Ignatius` or `Barn` are constantly better than state-of-the-art methods in all metrics for most numbers of input views.

5. Additional ablations

We study the effect of varying the hyperparameters for the construction of the sparse scene graph (Section 4.1 of the main paper) in Fig. 3. Generally increasing the number of key images (N_a) or nearest neighbors (k) leads to improvements in performance, which saturates above $N_a \geq 20$ or $k \geq 10$.

6. Parametrizations of Cameras

As noted by other authors [4], a clever parametrization of cameras can significantly accelerate convergence. In the main paper, we describe a camera $\mathcal{K}_n = (K_n, P_n)$ classically as intrinsic and extrinsic parameters, where

$$K_n = \begin{bmatrix} f_n & 0 & c_x \\ 0 & f_n & c_y \\ 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{3\times3}, \tag{1}$$

$$P_n = \left[\begin{array}{c|c} R_n & t_n \\ 0 & 1 \end{array}\right] \in \mathbb{R}^{4\times4}. \tag{2}$$

Here, $f_n > 0$ denotes the camera focal, $(c_x, c_y) = (W/2, H/2)$ is the optical center, $R_n \in \mathbb{R}^{3\times3}$ is a rotation matrix typically represented as a quaternion $q_n \in \mathbb{R}^4$ internally, and $t_n \in \mathbb{R}^3$ is a translation.

Camera parametrization. During optimization, 3D points are constructed using the inverse reprojection function $\pi^{-1}(\cdot)$ as a function of the camera intrinsics K_n, extrinsics P_n, pixel coordinates and depthmaps Z^n (see Section 4.2 of the main paper). One potential issue with this classical parametrization is that small changes in the extrinsics can typically induce a large change in the reconstructed 3D points. For instance, small noise on the rotation R_n could result in a potentially large absolute motion of 3D points, motion whose amplitude would be proportional to the points' distance to camera (*i.e.* their depth). It seems therefore natural to reparametrize cameras so as to better balance the variations between camera parameters and 3D points. To do so, we propose to switch the camera rotation center from the optical center to a point 'in the middle' of the 3D point-cloud generated by this camera, or more precisely, at the intersection of the $\overrightarrow{z}$ vector from the camera center and the median depth plane. In more details, we construct the extrinsics P_n using a fixed post-translation $\tilde{T}_n \in \mathbb{R}^4$ on the z-axis as as $P_n \overset{\text{def}}{=} T_n P'_n$, with

$$\tilde{T}_n = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & \tilde{m}_n^z \\ 0 & 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $\tilde{m}_n^z = \text{median}(\tilde{Z}^n) f_n / \tilde{f}_n$ is the median canonical depth for image I^n modulated by the ratio of the current focal length w.r.t. the canonical focal $\tilde{f}_n$, and P'_n is again parameterized as a quaternion and a translation. This way, rotation and translation noise in R_n are naturally compensated and have a lot less impact on the positions of the reconstructed 3D points, as illustrated in Tab. 6.

Kinematic chain. A second source of undesirable correlations between camera parameters stems from the intricate relationship between overlapping viewpoints. Indeed, if two views overlap, then modifying the position or rotation of one camera will most likely also result in a similar modification of the second camera, since the modification will impact the 3D points shared by both cameras. Thus, instead of representing all cameras independently, we propose to express them relatively to each other using a kinematic chain. This naturally conveys the idea than modifying one camera will impact the other cameras *by design*. In practice, we define a *kinematic tree* $\mathcal{T} = (\mathcal{V}, \mathcal{D})$ over all cameras $\mathcal{V}$. $\mathcal{T}$ consists of a single root node $r \in \mathcal{V}$ and a set of directed edges $(n \to m) \in \mathcal{D}$, with $|\mathcal{D}| = N - 1$ since $\mathcal{T}$ is a tree. The pose of all cameras is then computed

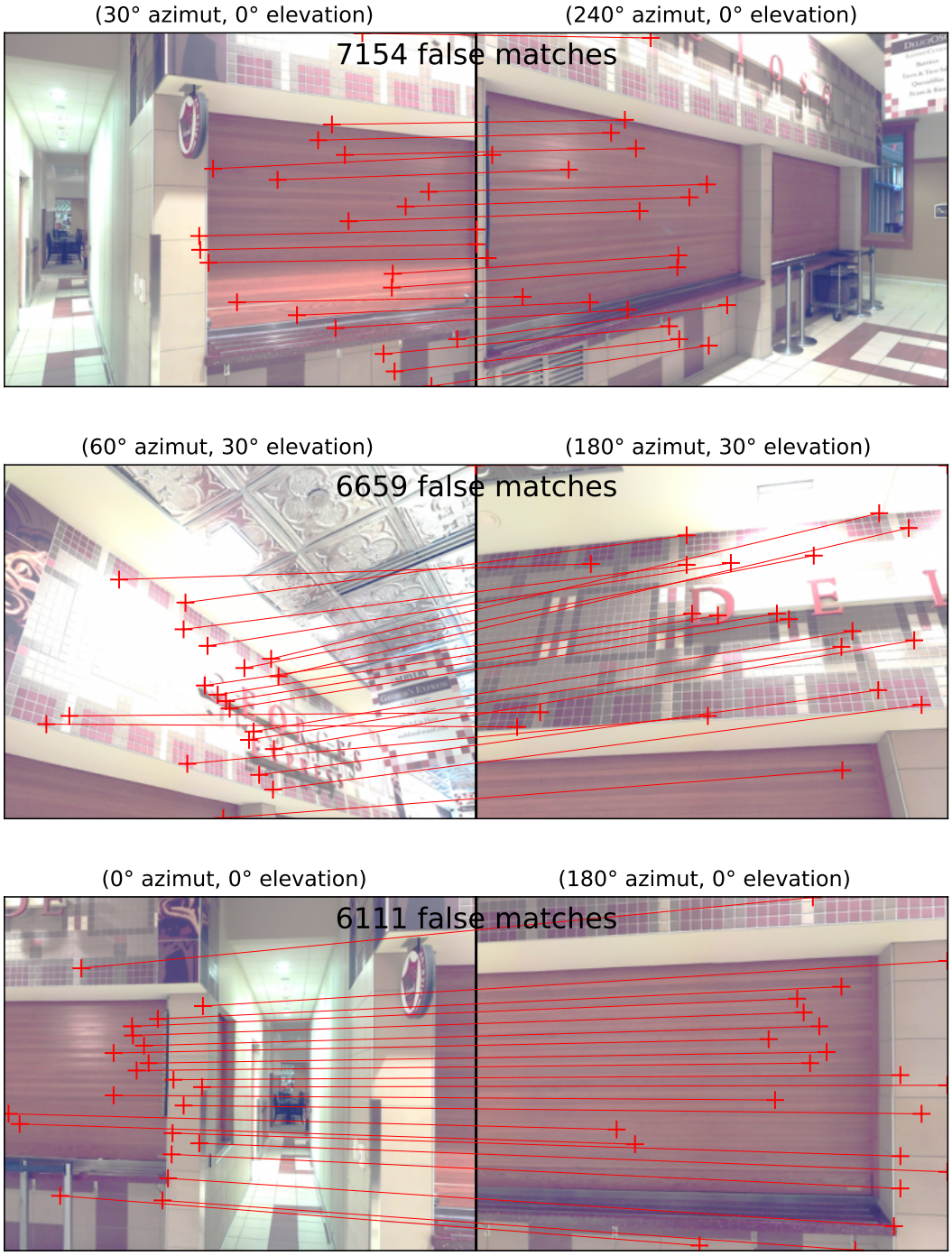

Figure 2. **Illustration of the typical failure case due to false matches.** In all failure cases that we have manually reviewed, the root cause of failure was the presence of wrong matches (outliers) between similar-looking parts of the same scene. Here, we show 3 such wrong pairs for the InLoc dataset (purely rotational case, specifically for the scene `DUC1/007`), each time printing the ground-truth cameras' azimuth and elevation and a small number of randomly-selected matches (showing all of them would impair readibility).

in sequence, starting from the root as

$$\forall (n \to m) \in \mathcal{D},\ P_m = P_{n\to m} P_n. \tag{4}$$

Internally, we thus only store as free variables the set of poses $\{P_r\} \cap \{P_{n\to m}\}_{(n\to m)\in\mathcal{D}}$, each one represented as mentioned above. In the end, this parametrization results in

Method	DUC1/000	DUC1/007	DUC1/014	DUC1/021	DUC1/070	DUC1/077	DUC1/084	DUC1/091	DUC2/033	DUC2/040	DUC2/047	DUC2/054	DUC2/061	DUC2/093	DUC2/100	DUC2/107	DUC2/115	DUC2/122	DUC2/129	DUC2/132	Mean
COLMAP [6]	1.0	6.0	4.4	0.5	12.4	0.5	4.4	1.0	1.0	0.5	1.0	2.4	14.4	5.7	7.8	8.4	5.7	0.5	1.3	3.7	4.1
FlowMap [8]	0.3	0.2	0.0	0.2	0.0	0.3	0.0	0.0	0.0	0.2	0.0	0.2	0.0	0.0	0.2	0.0	0.2	0.0	0.2	0.0	0.1
VGGSfM [12]	2.5	0.0	1.0	0.5	0.0	1.0	0.0	0.2	2.1	0.0	0.0	0.0	2.9	4.1	4.9	0.3	1.0	1.1	3.3	1.6	1.3
ACE-Zero [1]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	89.0	100.0	100.0	99.5
MASt3R-SfM	89.0	0.8	100.0	94.4	89.0	94.4	15.1	94.6	87.5	28.7	100.0	12.9	24.8	48.3	11.0	89.0	94.4	19.0	100.0	51.0	62.2
MASt3R-SfM†	94.4	15.2	99.5	100.0	89.0	94.4	84.0	94.4	94.4	25.1	94.4	23.0	29.7	100.0	30.5	94.4	22.2	23.5	89.0	37.1	66.7

Table 2. **Pure Rotation Case Evaluation.** RRA@5 (↑) on 20 randomly chosen scenes from the InLoc dataset. **MASt3R-SfM**† denotes our approach with disabled depth optimization for better optimization stability.

Methods	#Frames	Co3Dv2			RealEstate10K
		RRA@15	RTA@15	mAA(30)	mAA(30)
COLMAP+SPSG	3	∼22	∼14	∼15	∼23
PixSfM	3	∼18	∼8	∼10	∼17
Relpose	3	∼56	-	-	-
PoseDiffusion	3	∼75	∼75	∼61	- (∼77)
VGGSfM	3	58.7	51.2	45.4	-
DUSt3R	3	**95.3**	88.3	77.5	69.5
MASt3R-SfM	3	94.7	**92.1**	**85.7**	**84.3**
COLMAP+SPSG	5	∼21	∼17	∼17	∼34
PixSfM	5	∼21	∼16	∼15	∼30
Relpose	5	∼56	-	-	-
PoseDiffusion	5	∼77	∼76	∼63	- (∼78)
VGGSfM	5	80.4	75.0	69.0	-
DUSt3R	5	**95.5**	86.7	76.5	67.4
MASt3R-SfM	5	95.0	**91.9**	**86.4**	**85.3**
COLMAP+SPSG	10	31.6	27.3	25.3	45.2
PixSfM	10	33.7	32.9	30.1	49.4
Relpose	10	57.1	-	-	-
PoseDiffusion	10	80.5	79.8	66.5	48.0 (∼80)
VGGSfM	10	91.5	86.8	81.9	-
DUSt3R	10	**96.2**	86.8	76.7	67.7
MASt3R-SfM	10	96.0	**93.1**	**88.0**	**86.8**

Table 3. **Comparison with the state of the art for multi-view pose regression on the CO3Dv2 [5] and RealEstate10K [14] datasets with 3, 5 and 10 random frames.** (Parentheses) indicates results obtained after training on RealEstate10K. In contrast, we report results *without* training on RealEstate10K.

MegaDepth ID	T&T scene	MegaDepth ID	T&T scene
5000	Family	5007	Ballroom
5001	Auditorium	5008	Museum
5002	Courthouse	5009	Panther
5003	Horse	5010	Playground
5004	Francis	5011	Temple
5005	Lighthouse	5012	Train
5006	M60	5013	Palace

Table 4. **List of T&T scenes that are part of the training of the MASt3R checkpoint.**

exactly the same number of parameters as the classical one.

We experiment with different strategies to construct the kinematic tree $\mathcal{T}$ and report the results in Tab. 6: 'star' refers to a baseline where $N-1$ cameras are connected to the root camera, which performs even worse than a classical parametrization; 'MST' denotes a kinematic tree defined as maximum spanning tree over the similarity matrix S; and 'H. clust.' refers to a tree formed by hierarchical clustering using either raw similarities from image retrieval or actual number of correspondences after the pairwise forward with MASt3R. This latter strategy performs best and significantly improves over previous baselines, highlighting the importance of a balanced graph with approximately $\log_2(N)$ levels (in comparison, a star-tree has just 1 level, while a MST tree can potentially have $N/2$ levels at most). Note that the sparse scene graph $\mathcal{G}$ from Section 4.1 of the main paper and the kinematic tree $\mathcal{T}$ share no relation other than being defined over the same set of nodes.

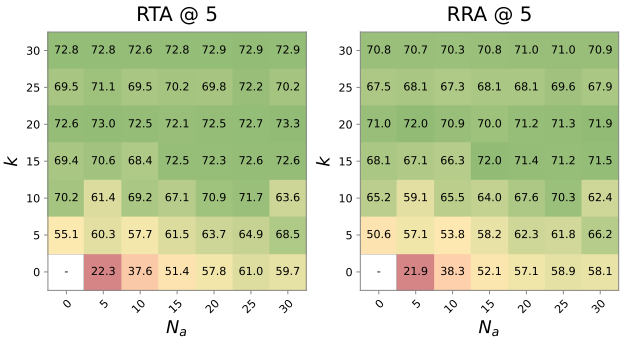

Figure 3. **Pose accuracy** (↑) on T&T-200 w.r.t. the number of key images N_a and number of nearest neighbors k

References

[1] 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. 5, 6, 7

[2] Xingyi He, Jiaming Sun, Yifan Wang, Sida Peng, Qixing Huang, Hujun Bao, and Xiaowei Zhou. Detector-free structure from motion. In *CVPR*, 2024. 6, 7

[3] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Trans. Graphics*, 2017. 3

[4] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T. Barron, and Ricardo Martin-Brualla. Camp: Camera preconditioning for neural radiance fields. *ACM Trans. Graphics*, 2023. 3

[5] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotný. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *ICCV*, 2021. 5

		Scene	ATE (↓)						RTA@5 (↑)						RRA@5 (↑)						Reg. (↑)					
			COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**
25 views	Train	Barn	0.0000	0.1128	0.1101	0.0898	0.1143	0.0011	0.3	2.3	1.0	53.3	46.7	100.	0.3	1.3	0.3	51.3	47.3	100.	8.0	100.	100.	96.0	100.	100.
		Caterpillar	0.0631	0.1125	0.1075	0.0301	0.0887	0.0299	15.3	2.3	1.0	92.0	46.7	94.3	17.0	3.0	0.0	92.0	47.0	92.0	60.0	100.	100.	100.	100.	100.
		Church	0.0868	0.1097	0.1071	0.0962	0.0936	0.0697	33.3	0.7	1.0	32.7	60.0	50.3	41.0	1.3	0.0	35.7	66.7	45.7	92.0	100.	100.	80.0	100.	100.
		Courthouse†	0.0000	0.1060	0.1119	0.1126	0.1119	0.1040	0.0	1.0	1.3	17.3	16.3	44.3	0.0	0.0	0.7	18.0	23.3	43.0	8.0	100.	100.	100.	96.0	100.
		Ignatius	0.0129	0.1129	0.1090	0.0005	0.0004	0.0002	92.0	1.3	1.0	100.	100.	100.	100.	2.0	0.7	100.	100.	100.	100.	100.	100.	100.	100.	100.
		Meetingroom	0.0000	0.1125	0.1046	0.0559	0.0996	0.0049	0.3	2.0	1.0	38.7	50.0	85.7	0.3	0.3	1.7	36.3	46.3	82.3	8.0	100.	100.	100.	100.	100.
		Truck	0.0916	0.1145	0.1072	0.0012	0.0981	0.0010	27.7	2.3	0.3	99.3	42.0	99.7	27.0	1.7	0.3	100.	40.7	100.	80.0	100.	100.	100.	100.	100.
	Intermediate	Family†	0.0023	0.1099	0.1090	0.0043	0.0045	0.0042	17.0	1.0	4.3	98.3	98.3	95.0	18.3	1.7	0.3	73.7	75.3	78.0	44.0	100.	100.	100.	100.	100.
		Francis†	0.0001	0.1084	0.1138	0.0024	0.0898	0.0176	15.0	0.3	3.0	98.0	42.3	76.3	15.0	1.0	0.3	92.0	43.7	75.3	40.0	100.	100.	100.	100.	100.
		Horse†	0.0055	0.1120	0.1056	0.0058	0.0072	0.0052	16.7	1.3	1.7	89.3	88.7	74.3	14.7	1.7	0.0	65.7	67.0	65.0	52.0	100.	100.	100.	100.	100.
		Lighthouse†	0.0411	0.1146	0.1128	0.0034	0.0853	0.0007	0.3	0.7	1.3	97.0	61.7	100.	0.7	0.3	0.7	100.	64.0	100.	40.0	100.	100.	100.	100.	100.
		M60†	0.0407	0.1118	0.1120	0.0970	0.0461	0.0005	2.0	2.0	2.7	73.7	83.3	99.7	2.0	2.0	2.3	77.3	84.3	100.	20.0	100.	100.	100.	100.	100.
		Panther†	0.0000	0.1147	0.1125	0.0016	0.1122	0.0005	2.0	0.7	2.0	99.3	48.0	99.7	2.0	0.0	2.0	100.	48.7	100.	16.0	100.	100.	100.	100.	100.
		Playground†	0.0000	0.1101	0.1065	0.0017	0.0009	0.0004	0.3	0.7	3.0	99.7	100.	100.	0.3	2.0	2.0	100.	100.	100.	8.0	100.	100.	100.	100.	100.
		Train†	0.0807	0.1116	0.1091	0.0777	0.1152	0.0770	5.7	1.3	0.7	61.0	28.7	65.0	12.3	0.0	0.0	64.3	28.7	64.3	68.0	100.	100.	100.	100.	100.
	Advanced	Auditorium†	0.0630	0.1071	0.1087	0.1063	0.1066	0.1067	0.0	1.3	2.0	2.3	3.3	2.0	0.0	0.3	0.3	0.3	0.3	0.3	44.0	100.	100.	100.	100.	100.
		Ballroom†	0.0912	0.1114	0.1129	0.1108	0.0955	0.0618	11.3	2.0	1.3	12.3	31.0	20.7	11.7	4.0	2.7	24.7	32.3	24.7	64.0	100.	100.	100.	100.	100.
		Courtroom	0.0865	0.1107	0.1102	0.1057	0.1048	0.0847	15.3	4.0	1.7	1.7	12.0	44.3	15.3	2.0	0.0	0.3	23.0	42.0	48.0	100.	100.	84.0	92.0	100.
		Museum†	0.1012	0.1130	0.1059	0.0994	0.1077	0.0969	2.0	0.3	0.3	2.0	7.0	11.0	2.7	0.0	0.0	2.0	12.3	12.0	84.0	100.	100.	76.0	100.	100.
		Palace†	0.0321	0.1136	0.0684	0.1057	0.1126	0.0273	5.0	0.7	1.0	13.7	22.3	38.3	5.0	0.0	0.7	12.3	13.0	34.0	28.0	100.	100.	88.0	100.	100.
		Temple†	0.0069	0.1147	0.1030	0.1090	0.1089	0.0122	2.3	0.7	0.7	24.3	33.7	75.0	2.0	0.0	0.3	24.7	33.7	69.7	20.0	100.	100.	96.0	100.	100.
50 views	Train	Barn	0.0003	0.0786	0.0793	0.0641	0.0007	0.0005	20.7	1.7	3.4	33.1	99.7	99.9	20.7	1.5	0.7	24.3	100.	100.	46.0	100.	100.	96.0	100.	100.
		Caterpillar	0.0313	0.0802	0.0795	0.0162	0.0161	0.0161	55.0	4.5	3.5	95.2	96.9	96.7	67.2	4.4	2.3	96.0	96.0	96.0	92.0	100.	100.	100.	100.	100.
		Church	0.0389	0.0681	0.0799	0.0436	0.0443	0.0707	59.6	9.8	1.2	64.2	71.8	49.4	60.5	16.2	0.7	70.3	85.0	47.5	96.0	100.	100.	98.0	100.	100.
		Courthouse†	0.0001	0.0784	0.0799	0.0694	0.0752	0.0738	2.3	1.4	1.8	35.1	32.6	25.9	2.3	0.1	0.5	34.7	33.2	25.8	16.0	100.	100.	100.	100.	100.
		Ignatius	0.0008	0.0118	0.0808	0.0004	0.0004	0.0001	91.9	95.8	1.4	99.9	100.	100.	92.1	100.	0.5	100.	100.	100.	96.0	100.	100.	100.	100.	100.
		Meetingroom	0.0175	0.0770	0.0694	0.0159	0.0767	0.0141	8.2	7.0	2.1	81.7	43.3	83.7	8.2	5.6	1.3	83.1	37.6	86.4	32.0	100.	100.	100.	100.	100.
		Truck	0.0729	0.0734	0.0773	0.0009	0.0008	0.0005	38.0	8.5	3.0	99.5	99.8	99.8	38.4	5.7	2.0	100.	100.	100.	86.0	100.	100.	100.	100.	100.
	Intermediate	Family†	0.0071	0.0035	0.0176	0.0030	0.0029	0.0028	53.6	91.6	30.9	98.3	95.8	96.7	46.4	86.4	17.8	77.6	81.0	81.1	96.0	100.	100.	100.	100.	100.
		Francis†	0.0451	0.0796	0.0784	0.0013	0.0134	0.0201	37.4	1.6	2.4	98.4	96.0	38.4	37.3	6.2	3.3	100.	96.0	36.4	78.0	100.	100.	100.	100.	100.
		Horse†	0.0103	0.0742	0.0737	0.0039	0.0036	0.0036	66.3	4.7	5.5	90.3	73.7	75.8	61.5	8.7	1.8	66.4	67.0	65.9	100.	100.	100.	100.	100.	100.
		Lighthouse†	0.0009	0.0795	0.0762	0.0017	0.0659	0.0003	24.5	0.8	0.5	98.7	65.3	100.	24.5	0.0	0.0	100.	67.2	100.	50.0	100.	100.	100.	100.	100.
		M60†	0.0002	0.0784	0.0800	0.0018	0.0006	0.0003	9.7	2.8	1.5	98.4	99.8	100.	9.8	3.0	0.8	100.	100.	100.	32.0	100.	100.	100.	100.	100.
		Panther†	0.0001	0.0762	0.0779	0.0041	0.0734	0.0004	2.9	0.7	2.0	96.1	51.6	99.8	2.9	0.3	1.1	96.0	51.9	100.	18.0	100.	100.	100.	100.	100.
		Playground†	0.0092	0.0807	0.0653	0.0010	0.0003	0.0003	0.2	1.4	3.0	99.7	100.	100.	0.2	0.7	1.2	100.	100.	100.	10.0	100.	100.	100.	100.	100.
		Train†	0.0663	0.0810	0.0789	0.0545	0.0736	0.0530	11.6	1.3	1.1	55.8	28.7	64.7	25.8	0.3	1.1	58.7	29.9	64.6	70.0	100.	100.	98.0	100.	100.
	Advanced	Auditorium†	0.0789	0.0802	0.0790	0.0760	0.0756	0.0756	0.1	0.8	0.3	1.2	1.8	1.5	0.1	0.9	0.7	1.0	1.0	1.1	22.0	100.	100.	96.0	100.	100.
		Ballroom†	0.0656	0.0775	0.0777	0.0545	0.0732	0.0677	15.6	1.6	3.8	37.1	25.6	19.1	19.4	5.2	2.2	47.8	31.3	23.4	68.0	100.	100.	100.	100.	100.
		Courtroom	0.0794	0.0793	0.0754	0.0819	0.0649	0.0531	17.1	3.6	0.4	25.3	59.7	77.3	18.5	3.1	0.1	27.4	68.4	78.4	68.0	100.	100.	98.0	100.	100.
		Museum†	0.0636	0.0788	0.0723	0.0804	0.0767	0.0675	9.4	0.8	0.7	1.1	9.5	11.0	9.5	1.8	0.4	1.1	15.7	11.0	78.0	100.	100.	94.0	100.	100.
		Palace†	0.0199	0.0807	0.0607	0.0803	0.0547	0.0238	35.3	0.4	1.6	5.3	13.6	44.8	33.3	0.1	1.1	9.2	11.5	49.3	70.0	100.	100.	96.0	100.	100.
		Temple†	0.0041	0.0809	0.0753	0.0724	0.0727	0.0029	16.7	0.7	0.9	33.5	51.8	87.3	14.2	0.1	0.5	31.6	50.5	84.7	46.0	100.	100.	96.0	100.	100.
100 views	Train	Barn	0.0301	0.0555	0.0316	0.0557	0.0004	0.0019	72.9	12.0	1.6	12.9	99.9	97.9	72.5	9.4	0.6	11.3	100.	98.0	99.0	100.	100.	92.0	100.	100.
		Caterpillar	0.0289	0.0119	0.0455	0.0111	0.0111	0.0112	56.7	77.5	20.9	95.4	96.9	95.3	61.8	54.8	20.2	96.0	96.0	94.1	98.0	100.	100.	100.	100.	100.
		Church	0.0298	0.0368	0.0516	0.0296	0.0348	0.0353	65.9	67.5	1.1	61.1	76.0	63.1	66.6	77.3	0.9	72.1	86.7	63.2	99.0	100.	100.	97.0	99.0	100.
		Courthouse†	0.0516	0.0572	0.0548	0.0561	0.0564	0.0465	3.8	0.6	0.5	18.9	51.5	24.4	3.8	0.1	0.8	20.0	50.6	20.4	27.0	100.	100.	96.0	100.	100.
		Ignatius	0.0100	0.0007	0.0469	0.0002	0.0002	0.0001	96.0	99.9	16.4	100.	100.	100.	100.	100.	9.6	100.	100.	100.	100.	100.	100.	100.	100.	100.
		Meetingroom	0.0097	0.0525	0.0457	0.0411	0.0135	0.0089	58.2	7.1	8.3	79.7	83.8	85.1	50.9	6.7	5.1	78.9	82.3	86.1	84.0	100.	100.	100.	100.	100.
		Truck	0.0208	0.0008	0.0170	0.0005	0.0005	0.0003	92.1	99.6	32.8	99.7	99.8	99.7	92.1	100.	15.2	100.	100.	100.	100.	100.	100.	100.	100.	100.
	Intermediate	Family†	0.0047	0.0034	0.0040	0.0446	0.0021	0.0019	58.8	83.7	70.4	48.9	96.3	96.9	50.0	71.5	50.0	43.6	80.8	81.3	100.	100.	100.	100.	100.	100.
		Francis†	0.0400	0.0077	0.0547	0.0009	0.0002	0.0027	51.8	45.0	0.4	98.7	99.9	88.1	51.1	22.6	0.2	100.	100.	72.6	100.	100.	100.	100.	100.	100.
		Horse†	0.0039	0.0054	0.0142	0.0026	0.0025	0.0026	70.2	67.2	36.3	91.7	73.9	75.9	68.6	42.1	14.6	68.6	68.0	68.0	100.	100.	100.	100.	100.	100.
		Lighthouse†	0.0090	0.0571	0.0536	0.0014	0.0479	0.0003	83.7	1.4	1.3	97.2	66.9	99.8	90.2	0.9	0.6	100.	68.9	100.	99.0	100.	100.	100.	100.	100.
		M60†	0.0019	0.0547	0.0573	0.0057	0.0003	0.0002	48.9	28.7	8.1	93.1	99.9	100.	50.2	31.0	6.8	92.1	100.	100.	71.0	100.	100.	99.0	100.	100.
		Panther†	0.0244	0.0521	0.0561	0.0004	0.0521	0.0002	28.2	18.0	1.6	99.2	51.9	99.5	27.2	15.2	1.6	100.	52.4	100.	70.0	100.	100.	100.	100.	100.
		Playground†	0.0002	0.0570	0.0527	0.0004	0.0002	0.0002	14.2	0.7	8.6	99.9	100.	100.	14.2	0.3	6.2	100.	100.	100.	38.0	100.	100.	100.	100.	100.
		Train†	0.0533	0.0564	0.0392	0.0373	0.0554	0.0372	24.6	0.9	4.1	66.4	31.6	65.6	42.7	1.5	3.9	65.0	37.2	65.0	99.0	100.	100.	100.	100.	100.
	Advanced	Auditorium†	0.0481	0.0555	0.0550	0.0536	0.0532	0.0532	1.5	1.2	0.5	2.1	1.5	1.6	1.2	2.1	0.3	1.1	1.2	1.2	94.0	100.	100.	100.	100.	100.
		Ballroom†	0.0477	0.0531	0.0531	0.0431	0.0491	0.0377	27.6	6.6	2.3	36.0	32.9	20.2	38.2	9.1	2.9	47.9	41.8	25.0	97.0	100.	100.	98.0	100.	100.
		Courtroom	0.0526	0.0525	0.0529	0.0576	0.0461	0.0398	41.2	15.2	0.6	45.2	64.6	72.3	42.3	18.6	0.4	47.4	66.3	71.8	88.0	100.	100.	100.	100.	100.
		Museum†	0.0502	0.0517	0.0526	0.0554	0.0481	0.0502	8.1	7.8	0.4	8.7	13.4	11.7	8.1	8.7	0.4	9.2	14.9	12.5	99.0	100.	100.	98.0	100.	100.
		Palace†	0.0167	0.0572	0.0480	0.0590	0.0419	0.0205	38.1	1.7	1.1	7.1	21.2	38.1	34.2	1.2	1.4	9.2	16.6	30.0	79.0	100.	100.	90.0	99.0	100.
		Temple†	0.0210	0.0575	0.0499	0.0526	0.0528	0.0025	14.8	0.5	2.2	35.7	54.5	83.9	9.0	0.0	0.9	32.9	52.3	82.9	58.0	100.	100.	99.0	100.	100.
200 views	Train	Barn	0.0222	0.0317	0.0199	-	-	0.0010	73.4	44.1	29.8	-	-	93.0	73.0	32.0	22.4	-	-	89.7	100.	100.	100.	-	-	100.
		Caterpillar	0.0160	0.0076	0.0100	0.0076	-	0.0075	73.3	93.8	53.0	94.4	-	95.9	85.4	82.6	43.1	96.0	-	96.0	100.	100.	100.	100.	-	100.
		Church	0.0274	0.0218	0.0320	-	-	0.0212	65.7	71.6	2.4	-	-	65.5	82.6	84.6	1.9	-	-	66.4	100.	100.	100.	-	-	100.
		Courthouse†	0.0406	0.0407	0.0389	-	0.0404	0.0303	28.0	0.7	3.2	-	35.1	30.1	27.9	0.1	1.8	-	33.4	23.6	75.0	100.	100.	-	100.	100.
		Ignatius	0.0218	0.0004	0.0124	0.0004	0.0002	0.0001	68.7	100.	33.1	98.8	99.9	100.	100.	100.	22.2	99.0	100.	100.	100.	100.	100.	99.5	100.	100.
		Meetingroom	0.0145	0.0060	0.0329	-	0.0063	0.0063	68.9	81.7	13.6	-	88.1	88.5	69.5	88.0	8.9	-	84.4	90.9	100.	100.	100.	-	100.	100.
		Truck	0.0002	0.0007	0.0034	0.0004	-	0.0003	99.9	99.7	78.1	99.8	-	99.8	100.	100.	85.8	100.	-	100.	100.	100.	100.	100.	-	100.
	Intermediate	Family†	0.0032	0.0014	0.0028	0.0015	-	0.0274	62.4	98.3	69.2	98.2	-	25.0	50.4	82.0	50.4	79.0	-	23.2	100.	100.	100.	100.	-	100.
		Francis†	0.0001	0.0101	0.0062	0.0097	-	0.0040	100.	80.3	56.4	90.8	-	77.6	100.	51.2	54.7	92.1	-	60.7	100.	100.	100.	100.	-	100.
		Horse†	0.0023	0.0019	0.0021	0.0018	0.0018	0.0019	74.0	81.5	77.7	92.1	74.2	73.8	69.3	67.8	57.4	69.0	68.5	63.0	100.	100.	100.	100.	100.	100.
		Lighthouse†	0.0001	0.0192	0.0377	0.0010	-	0.0013	98.0	31.1	1.9	96.7	-	98.8	98.0	28.0	1.9	100.	-	98.9	99.0	100.	100.	100.	-	100.
		M60†	0.0062	0.0004	0.0351	-	-	0.0002	73.7	99.9	21.4	-	-	100.	100.	100.	20.9	-	-	100.	100.	100.	100.	-	-	100.
		Panther†	0.0011	0.0004	0.0177	0.0236	0.0003	0.0002	92.8	99.5	32.3	55.9	99.3	99.5	100.	100.	28.8	55.6	100.	100.	100.	100.	100.	100.	100.	100.
		Playground†	0.0371	0.0071	0.0174	0.0003	-	0.0001	26.2	60.0	38.2	99.7	-	100.	61.5	62.6	39.4	100.	-	100.	100.	100.	100.	100.	-	100.
		Train†	0.0362	0.0270	0.0297	-	-	0.0264	27.2	59.4	18.0	-	-	59.4	47.8	61.1	12.4	-	-	64.5	100.	100.	100.	-	-	100.
	Advanced	Auditorium†	0.0374	0.0389	0.0395	-	-	0.0378	1.4	1.2	0.8	-	-	1.3	1.5	1.7	1.2	-	-	1.6	100.	100.	100.	-	-	100.
		Ballroom†	0.0365	0.0243	0.0347	0.0265	-	0.0264	26.8	35.7	7.1	36.9	-	25.7	46.5	50.1	7.0	54.6	-	42.4	98.5	100.	100.	100.	-	100.
		Courtroom	0.0353	0.0353	0.0367	-	-	0.0291	61.4	51.7	1.4	-	-	65.8	62.0	62.5	1.3	-	-	67.1	99.5	100.	100.	-	-	100.
		Museum†	0.0341	0.0360	0.0383	-	-	0.0351	15.7	11.3	0.6	-	-	11.9	15.6	13.1	0.5	-	-	11.4	99.0	100.	100.	-	-	100.
		Palace†	0.0124	0.0405	0.0233	-	0.0302	0.0142	45.9	2.7	3.3	-	31.7	42.1	44.1	1.7	3.1	-	26.2	49.2	87.0	100.	100.	-	100.	100.
		Temple†	0.0091	0.0404	0.0373	-	0.0358	0.0021	29.0	0.6	0.6	-	57.0	83.1	23.8	0.1	0.5	-	55.2	83.9	79.5	100.	100.	-	100.	100.

		ATE (↓)						RTA@5 (↑)						RRA@5 (↑)						Reg. (↑)					
	Scene	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**	COLMAP [6]	ACE-Zero [1]	FlowMap [8]	VGGSfM [12]	DF-SfM [2]	**MASt3R-SfM**
full Train	Barn	GT	0.0216	-	-	0.0002	0.0020	GT	55.6	-	-	99.8	85.6	GT	56.1	-	-	100.	52.6	GT	100.	-	-	100.	100.
	Caterpillar	GT	0.0053	-	-	-	0.0053	GT	95.6	-	-	-	92.3	GT	87.3	-	-	-	84.2	GT	100.	-	-	-	100.
	Church	GT	0.0128	-	-	-	0.0139	GT	76.3	-	-	-	16.8	GT	90.5	-	-	-	11.6	GT	100.	-	-	-	100.
	Courthouse†	GT	0.0155	-	-	-	0.0130	GT	45.0	-	-	-	9.9	GT	44.1	-	-	-	8.8	GT	100.	-	-	-	100.
	Ignatius	GT	0.0003	0.0033	-	0.0001	0.0045	GT	99.9	70.0	-	99.9	60.1	GT	100.	62.5	-	100.	43.6	GT	100.	100.	-	100.	100.
	Meetingroom	GT	0.0286	0.0087	-	0.0046	0.0046	GT	38.5	39.8	-	89.0	89.9	GT	39.3	26.3	-	84.1	92.6	GT	100.	100.	-	100.	100.
	Truck	GT	0.0006	0.0039	-	0.0003	0.0002	GT	99.7	69.6	-	99.8	99.7	GT	100.	53.4	-	100.	100.	GT	100.	100.	-	100.	100.
full Intermediate	Family†	GT	0.0162	-	-	-	0.0094	GT	44.6	-	-	-	25.9	GT	38.9	-	-	-	22.3	GT	100.	-	-	-	100.
	Francis†	GT	0.0115	0.0039	-	0.0002	0.0051	GT	79.0	67.7	-	99.7	41.0	GT	57.4	57.6	-	100.	17.0	GT	100.	100.	-	100.	100.
	Horse†	GT	0.0012	-	-	-	0.0148	GT	81.8	-	-	-	6.3	GT	68.2	-	-	-	6.4	GT	100.	-	-	-	100.
	Lighthouse†	GT	0.0111	0.0260	-	0.0282	0.0038	GT	38.8	9.5	-	66.0	72.1	GT	30.6	4.8	-	66.3	50.8	GT	100.	100.	-	100.	100.
	M60†	GT	0.0003	0.0258	-	0.0004	0.0003	GT	99.9	48.3	-	99.8	100.	GT	100.	50.4	-	100.	100.	GT	100.	100.	-	100.	100.
	Panther†	GT	0.0003	0.0026	-	0.0003	0.0002	GT	99.5	77.6	-	99.1	99.5	GT	100.	100.	-	100.	100.	GT	100.	100.	-	100.	100.
	Playground†	GT	0.0017	0.0042	-	0.0003	0.0006	GT	85.5	63.8	-	99.9	99.3	GT	82.7	49.1	-	100.	99.3	GT	100.	100.	-	100.	100.
	Train†	GT	0.0216	0.0233	-	0.0293	0.0230	GT	62.5	29.2	-	41.8	15.8	GT	62.6	18.4	-	42.8	10.6	GT	100.	100.	-	100.	100.
full Advanced	Auditorium†	GT	0.0335	0.0341	-	0.0326	0.0326	GT	1.1	1.4	-	1.7	1.5	GT	1.6	1.3	-	1.7	1.7	GT	100.	100.	-	100.	100.
	Ballroom†	GT	0.0196	0.0199	-	0.0199	0.0201	GT	43.2	16.7	-	44.4	29.6	GT	56.4	14.1	-	56.0	43.8	GT	100.	100.	-	100.	100.
	Courtroom	GT	0.0280	0.0308	-	0.0276	0.0265	GT	54.1	3.6	-	66.3	69.1	GT	62.5	5.3	-	66.8	67.2	GT	100.	100.	-	100.	100.
	Museum†	GT	0.0287	0.0275	-	0.0281	0.0290	GT	11.1	1.2	-	13.5	11.0	GT	13.5	0.8	-	14.8	12.3	GT	100.	100.	-	100.	100.
	Palace†	GT	0.0276	-	-	0.0198	0.0102	GT	3.9	-	-	27.7	35.7	GT	3.1	-	-	25.6	27.0	GT	100.	-	-	100.	100.
	Temple†	GT	0.0334	0.0271	-	0.0289	0.0030	GT	0.9	1.2	-	60.7	72.2	GT	0.4	0.5	-	55.5	80.7	GT	100.	100.	-	100.	100.

Table 5. **Detailed per-scene results on Tanks & Temples** in terms of ATE, pose accuracy (RTA@5 and RRA@5) and registration rate (Reg.). For easier readability, we color-code the results as a linear gradient between worst and best per-row result for that metric. Reg. is color-coded with linear gradient between 0% and 100%. We mark missing results with - (not converged / runtime errors / ground truth). Scenes that are part of the training set of MegaDepth (*i.e.* used to train the MASt3R checkpoint) are marked with a †.

	ATE↓	RTA@5↑	RRA@5↑
Camera reparametrization			
No	0.01445	56.0	52.5
Yes	**0.01243**	**70.9**	**67.6**
Kinematic chain			
No	0.01675	52.2	50.0
Star	0.02013	42.0	39.2
MST	0.01600	64.4	62.1
H. clust. (sim)	0.01517	64.2	62.6
H. clust (#corr)	**0.01243**	**70.9**	**67.6**

Table 6. Effects of camera reparametrization and kinematic chain on T&T-200.

[6] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 5, 6, 7

[7] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 3

[8] Cameron Smith, David Charatan, Ayush Tewari, and Vincent Sitzmann. FlowMap: High-Quality Camera Poses, Intrinsics, and Depth via Gradient Descent. In *ECCV*, 2024. 5, 6, 7

[9] Hajime Taira, Masatoshi Okutomi, Torsten Sattler, Mircea Cimpoi, Marc Pollefeys, Josef Sivic, Tomas Pajdla, and Akihiko Torii. InLoc: Indoor visual localization with dense matching and view synthesis. In *CVPR*, 2018. 1

[10] Giorgos Tolias, Yannis Avrithis, and Hervé Jégou. To aggregate or not to aggregate: Selective match kernels for image search. In *ICCV*, 2013. 1

[11] Giorgos Tolias, Tomas Jenicek, and Ondřej Chum. Learning and aggregating deep local descriptors for instance-level recognition. In *ECCV*, 2020. 1

[12] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Visual Geometry Grounded Deep Structure From Motion. In *CVPR*, 2024. 3, 5, 6, 7

[13] Philippe Weinzaepfel, Thomas Lucas, Diane Larlus, and Yannis Kalantidis. Learning super-features for image retrieval. In *ICLR*, 2022. 1

[14] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *SIGGRAPH*, 2018. 5