A. More Results

Additional speed benchmarking with different hardware configurations is reported in Tab. 6. We show per-scene pose accuracy metrics for all datasets from Tab. 7 to Tab. 11. The per-scene NeRF and Gaussian Splatting evaluation on Tanks and Temples is shown in Tab. 12.

B. Technical Details

B.1. Distortion Estimation

Hierarchical search In order to accelerate the interval search in distortion estimation, which scales linearly with the number of candidates, we employ a hierarchical search strategy that iteratively shrinks the interval. At each level of the hierarchy, after finding the solution, we set the left and right candidates as the endpoints of the new interval for the next level. The solution at the last level is the final estimate.

Multiple cameras If the two images in a pair do not share intrinsics, but the distortion parameter of one of the images is known, we can use a similar 1D search method to determine the distortion parameter of the other image. With this, we can extend the distortion estimation algorithm to deal with multiple different cameras, each of which corresponds to a known subset of images.

We say that an image pair is *ready* for a camera if either

1. both images correspond to that camera; or
2. only one of the images corresponds to that camera, but the distortion parameter of the other image is already estimated.

We estimate the distortion parameters for these cameras one by one. Each time, among the cameras whose distortion have not been estimated, we pick the one with the largest number of ready image pairs. The distortion parameter for this camera is then estimated using these image pairs. To do that, the only modification to the original algorithm (originally for a single image pair) is that for each candidate of α we compute the average epipolar error over all the point pairs in all the ready image pairs. The next camera is picked likewise, until the distortion parameters for all the cameras are estimated.

Importance of undistortion The most direct impact of incorrect distortion is on focal length estimation as illustrated in Figure 3, which visualizes focal length validity (smoothed over discrete samples) on two of the scenes with and without distortion estimation. After keypoints are undistorted (green), the validity score peaks at an accurate FoV estimation. Without distortion estimation (red), the total validity score decreases drastically, and the peak deviates from the correct FoV.

B.2. Focal Length Estimation

We adopt a similar strategy as distortion estimation Sec. B.1 to deal with multiple cameras. However we do not use hierarchical sampling because processing each candidate is relatively cheap, and we can afford to densely sample the interval.

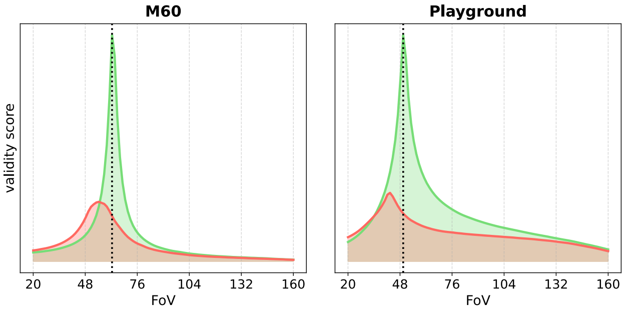

Figure 3. Effect of distortion on focal length estimation. Curves in green are with undistortion, and curves in red without. The dotted lines indicate the ground-truth FoVs.

B.3. Global Rotation

B.3.1 Initialization

In this section we discuss how to initialize the global rotation matrices for optimization. It is a modified version of Martinec and Pajdla [38].

Denote the set of images as $\mathcal{I}$ and the set of image pairs as $\mathcal{P} = \{(i,j)\}_{(i,j)\in\mathcal{I}\times\mathcal{I}}$, where each element has relative rotation matrix $\mathbf{R}^{i\to j}$. The goal is to construct a solution of global world to camera rotation matrices $\{\mathbf{R}_i\}_{i\in\mathcal{I}}$ to minimize the objective (note that in contrast to the final objective used in iterative optimization, this one uses L2 loss)

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{P}} \left\| \mathbf{R}^{(j)} - \mathbf{R}^{i\to j}\mathbf{R}^{(i)} \right\|^2 \tag{10}$$

Unfortunately, this problem cannot be directly solved using simple least square techniques since $\mathbf{R}^{(i)} \in \mathrm{SO}(3)$ are 3×3 matrices with orthogonality constraints. However, we can decompose it into several sub-problems to circumvent the constraints. In the following we use $\mathbf{A}_{*,k}$ to denote the k^{th} solumn of a matrix $\mathbf{A}$. Note that each term inside the summation in Eqn. 10 can be splitted into three parts

$$\mathcal{L} = \sum_{(i,j)\in\mathcal{P}} \sum_{k=1,2,3} \left\| \mathbf{R}^{(j)}_{*,k} - \mathbf{R}^{i\to j}\mathbf{R}^{(i)}_{*,k} \right\|^2 \tag{11}$$

Since $\mathbf{R}^{(i)}$ is an orthogonal matrix, the column vectors $\mathbf{R}^{(i)}_{*,k}$ have unit length and are mutually orthogonal. These orthogonality constraints are difficult to deal with, but if we only look at one particular column, say $k=1$, and ignore the unit length constraint, the objective becomes an unconstrained least squares problem

$$\mathcal{L}^{(1)} = \sum_{(i,j)\in\mathcal{P}} \left\| \mathbf{R}^{(j)}_{*,1} - \mathbf{R}^{i\to j}\mathbf{R}^{(i)}_{*,1} \right\|^2 \tag{12}$$

		FastMap (1G+2C)		GLOMAP			COLMAP	
	n_imgs	w/ cuda	w/o cuda	1G+48C	1G+12C	48C	1G+48C	48C
z_alameda	1734	134 : ×1.0	917 : ×6.8	848 : ×6.3	934 : ×6.9	3805 : ×28.2	4541 : ×33.6	24641 : ×182.5
z_berlin	1511	152 : ×1.0	545 : ×3.6	893 : ×5.9	1015 : ×6.7	1802 : ×11.8	6478 : ×42.4	24648 : ×161.4
z_london	1874	102 : ×1.0	556 : ×5.4	566 : ×5.5	669 : ×6.5	2092 : ×20.3	2643 : ×25.7	19238 : ×187.1
z_nyc	990	88 : ×1.0	338 : ×3.8	451 : ×5.1	487 : ×5.5	921 : ×10.5	1618 : ×18.4	1988 : ×22.6
mill19_building	1920	258 : ×1.0	1366 : ×5.3	6289 : ×24.3	7792 : ×30.1	38428 : ×148.4	27080 : ×104.6	152839 : ×590.1
mill19_rubble	1657	240 : ×1.0	789 : ×3.3	2849 : ×11.8	2466 : ×10.2	11571 : ×48.0	12153 : ×50.4	64987 : ×269.8
urbn_Campus	5871	740 : ×1.0	3009 : ×4.1	3869 : ×5.2	4175 : ×5.6	21916 : ×29.6	106055 : ×143.2	349490 : ×472.0
urbn_Sci-Art	3019	445 : ×1.0	1760 : ×4.0	4601 : ×10.3	5712 : ×12.8	28824 : ×64.7	42032 : ×94.4	286454 : ×643.4
eft_apartment	3804	549 : ×1.0	1003 : ×1.8	5905 : ×10.8	8341 : ×15.2	124310 : ×226.3	185361 : ×337.5	timeout
eft_kitchen	6042	2202 : ×1.0	6796 : ×3.1	22884 : ×10.4	34287 : ×15.6	timeout	timeout	timeout

Table 6. Detailed system runtime comparisons (seconds:speed_ratio) with different GPU (G) and CPU threads (C) configurations. Despite the cuda-accelerated ceres solver for bundle adjustment, a significant part of GLOMAP and COLMAP pipeline workload is still CPU-bound, and having at least 12 threads is necessary for higher speed. FastMap performs all data structure marshaling on GPU with non-trivial tensor indexing, and consumes less CPU resource.

		time (sec)			ATE↓			RTA@3↑			RRA@3↑			RTA@1↑			RRA@1↑			AUC-R&T @ 3 ↑			AUC-R&T @ 1 ↑		
	n_imgs	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP
m360_bicycle	194	20	74	151	7.5e-5	5.4e-5	5.7e-5	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	99.9	100.0	100.0	100.0	96.9	97.6	97.5	90.8	92.7	92.6
m360_bonsai	292	47	306	1043	3.9e-5	2.1e-5	1.7e-4	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	97.9	100.0	100.0	100.0	96.4	98.6	92.8	89.2	95.7	79.7
m360_counter	240	35	201	443	1.3e-5	2.5e-6	2.3e-6	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.1	99.8	99.8	97.2	99.5	99.5
m360_flowers	173	19	54	120	6.5e-5	6.0e-5	1.6e-4	100.0	100.0	100.0	100.0	100.0	100.0	99.8	99.9	99.6	100.0	100.0	100.0	96.1	96.4	92.0	88.5	89.3	76.3
m360_garden	185	28	152	490	7.7e-6	1.5e-5	1.5e-5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	99.6	99.0	99.1	98.8	97.1	97.3
m360_kitchen	279	50	376	1308	4.4e-5	4.0e-5	3.9e-5	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	97.2	97.2	97.2	91.5	91.5	91.7
m360_room	311	61	218	691	3.4e-3	1.1e-5	9.3e-6	99.3	100.0	100.0	99.4	100.0	100.0	98.8	100.0	100.0	99.4	100.0	100.0	96.6	98.9	99.0	91.5	96.7	97.0
m360_stump	125	15	40	74	3.5e-5	1.9e-5	2.1e-5	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	98.8	99.4	99.3	96.5	98.1	98.0
m360_treehill	141	18	68	202	8.7e-5	7.7e-5	4.6e-5	100.0	100.0	100.0	100.0	100.0	100.0	99.7	99.8	99.9	100.0	100.0	100.0	95.8	96.8	98.3	87.6	90.6	94.9

Table 7. Per scene camera pose metrics on the MipNeRF360 Dataset.

Using techniques like SVD we can find a non-trivial solution that is not all zero. And since the relative rotation matrices $\mathbf{R}^{i\to j}$ are also orthogonal matrices, left multiplying it with a vector preserves the vector length. This means that in the solution, the three-dimensional vectors $\mathbf{R}^{(i)}_{*,1}$ should have similar lengths. We can normalize them to unit length and get a solution of the actual first columns of global rotation matrices $\{\mathbf{R}^{(i)}\}_{i\in\mathcal{I}}$.

Given already estimated first columns $\{\mathbf{R}^{(i)}_{*,1}\}_{i\in\mathcal{I}}$ of the global rotation matrices, we can estimate the second columns by minimizing the same objective but adding additional constraints enforcing orthogonality to the first columns

$$\begin{aligned}\mathcal{L}^{(2)} =& \frac{1}{|\mathcal{P}|}\sum_{(i,j)\in\mathcal{P}}\left\|\mathbf{R}^{(j)}_{*,2} - \mathbf{R}^{i\to j}\mathbf{R}^{(i)}_{*,2}\right\|^2 \\ &+ \frac{1}{|\mathcal{I}|}\sum_{i\in\mathcal{I}}\left\|{\mathbf{R}^{(i)}_{*,1}}^\top\mathbf{R}^{(i)}_{*,2}\right\|^2,\end{aligned} \tag{13}$$

where $|\mathcal{P}|$ is the number of image pairs and $|\mathcal{I}|$ is the number of images, and they are used heuristically for controlling the relative weighting of the two terms. Note that in the above $\mathbf{R}^{(i)}_{*,1}$ are already fixed and the only free variables are $\{\mathbf{R}^{(i)}_{*,2}\}$. So this is still a least squares problem, and can be solved by applying SVD and then normalizing each 3-dimensional vector to unit length. A simple Gram-Schmidt process can be used for enforcing the orthogonality between the first and second columns.

We do not need to solve a least square problem again for the third columns because it can be directly computed by taking the cross product of the first two columns

$$\mathbf{R}^{(i)}_{*,3} = \mathbf{R}^{(i)}_{*,1} \times \mathbf{R}^{(i)}_{*,2} \tag{14}$$

B.3.2 Filtering

To improve the robustness of global rotation alignment, we filter out some image pairs whose number of inlier point pairs does not exceed certain threshold. Determining this threshold can be tricky: a low threshold might introduce a lot of outlier image pairs, but a high threshold could reduce the number of connections and even disconnect the images into several clusters. To alleviate this problem, we start from a large threshold, and reduce it by half if it leads to disconnected clusters. We do this iteratively until either all the images are connected, or a minimal threshold is reached. This partially solves the problem by making the threshold adaptive to the data. However it is still common that even at the minimal threshold, a few images are disconnected from the others. In that case we just consider them as outliers and ignore them in rotation alignment and later stages.

		time (sec)			ATE↓			RTA@3↑			RRA@3↑			RTA@1↑			RRA@1↑			AUC-R&T @ 3 ↑			AUC-R&T @ 1 ↑		
	n_imgs	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP
tnt_advn_Auditorium	298	74	185	529	1.4e-2	3.3e-2	1.8e-3	29.6	94.3	95.8	55.2	94.0	93.4	11.0	93.7	92.8	17.0	94.0	89.6	12.0	90.8	88.0	1.8	84.8	83.8
tnt_advn_Ballroom	324	55	652	1615	1.8e-2	3.2e-2	5.1e-3	49.1	32.5	98.1	63.7	36.3	98.2	28.3	28.3	97.7	28.6	32.2	98.2	25.0	26.5	93.5	7.7	17.8	84.6
tnt_advn_Courtroom	301	45	296	882	9.3e-4	4.3e-5	2.3e-5	85.1	99.9	99.9	98.3	100.0	100.0	27.9	99.3	99.8	47.4	100.0	100.0	37.7	96.8	98.6	3.8	91.0	96.1
tnt_advn_Museum	301	44	275	752	8.5e-4	5.5e-5	3.3e-5	91.4	100.0	100.0	100.0	100.0	100.0	39.9	99.7	99.9	85.2	100.0	100.0	53.8	97.2	98.7	12.1	91.9	96.1
tnt_advn_Palace	501	113	547	1722	3.3e-3	6.4e-3	1.5e-4	74.5	47.9	97.1	84.3	46.4	98.4	51.6	44.5	92.6	36.1	45.9	96.5	41.2	42.2	91.6	13.3	37.9	84.4
tnt_advn_Temple	302	36	190	596	9.8e-4	5.0e-5	1.5e-4	98.8	99.9	99.8	100.0	100.0	100.0	95.1	99.6	99.4	99.1	100.0	100.0	85.8	98.4	98.1	61.3	95.6	95.0
tnt_intrmdt_Family	152	22	152	335	8.3e-5	1.1e-5	2.8e-6	100.0	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	95.4	99.4	99.9	86.5	98.3	99.7
tnt_intrmdt_Francis	302	29	269	789	4.2e-5	6.2e-6	2.4e-6	99.9	100.0	100.0	100.0	100.0	100.0	99.5	99.9	100.0	100.0	100.0	100.0	96.2	99.5	99.9	89.2	98.5	99.7
tnt_intrmdt_Horse	151	21	134	255	6.8e-5	1.4e-5	4.6e-6	100.0	100.0	100.0	100.0	100.0	100.0	99.9	100.0	100.0	100.0	100.0	100.0	97.2	99.4	99.8	91.5	98.3	99.4
tnt_intrmdt_Lighthouse	309	47	355	1270	1.0e-4	2.3e-5	9.4e-6	99.6	100.0	100.0	100.0	100.0	100.0	97.4	99.8	100.0	99.2	100.0	100.0	92.0	98.3	99.4	78.3	95.1	98.2
tnt_intrmdt_M60	313	37	329	873	4.4e-5	1.2e-5	6.2e-6	99.9	100.0	100.0	100.0	100.0	100.0	99.5	100.0	100.0	100.0	100.0	100.0	96.4	99.1	99.7	89.5	97.3	99.0
tnt_intrmdt_Panther	314	43	390	989	6.4e-5	6.4e-5	1.5e-4	99.9	100.0	99.9	100.0	100.0	100.0	99.0	99.7	97.4	100.0	100.0	100.0	94.1	97.3	94.8	82.9	92.1	85.8
tnt_intrmdt_Playground	307	41	473	1255	1.4e-4	1.8e-5	1.9e-3	100.0	100.0	98.7	100.0	100.0	98.7	99.6	99.9	98.7	97.3	100.0	98.7	87.8	99.2	98.3	63.9	97.5	97.5
tnt_intrmdt_Train	301	41	414	901	8.0e-5	7.8e-6	4.3e-6	99.9	100.0	100.0	100.0	100.0	100.0	99.5	99.9	100.0	100.0	100.0	100.0	94.1	99.4	99.8	82.7	98.2	99.3
tnt_trng_Barn	410	50	503	3126	1.0e-4	6.7e-6	4.6e-6	100.0	100.0	100.0	100.0	100.0	100.0	99.7	100.0	100.0	100.0	100.0	100.0	94.2	99.4	99.7	82.9	98.2	99.2
tnt_trng_Caterpillar	383	44	374	1367	5.1e-5	5.6e-6	4.1e-6	99.8	100.0	100.0	100.0	100.0	100.0	99.5	99.9	99.9	100.0	100.0	100.0	96.2	99.4	99.7	89.0	98.4	99.3
tnt_trng_Church	507	85	666	4589	8.8e-3	2.6e-2	8.0e-4	75.6	71.1	99.6	94.8	75.4	100.0	57.4	70.3	99.3	72.4	70.8	100.0	52.3	69.3	98.4	19.5	66.6	96.1
tnt_trng_Courthouse	1106	169	1297	8285	1.2e-2	1.7e-2	1.3e-3	40.9	97.3	99.8	71.4	97.3	99.8	35.9	97.3	99.8	67.3	97.3	99.8	33.7	96.5	99.4	24.0	95.0	98.6
tnt_trng_Ignatius	263	33	269	682	5.2e-5	7.6e-6	2.1e-6	100.0	100.0	100.0	100.0	100.0	100.0	99.8	100.0	100.0	100.0	100.0	100.0	96.9	99.5	99.9	90.9	98.6	99.8
tnt_trng_Meetingroom	371	32	209	575	3.1e-4	1.4e-5	8.3e-6	98.5	100.0	100.0	100.0	100.0	100.0	82.5	99.9	100.0	84.4	100.0	100.0	72.0	99.0	99.4	30.5	97.0	98.2
tnt_trng_Truck	251	31	289	635	6.9e-5	3.4e-2	2.5e-6	100.0	52.7	100.0	100.0	52.6	100.0	99.8	52.6	100.0	100.0	52.6	100.0	95.4	51.8	99.9	86.4	50.3	99.6

Table 8. Per scene camera pose metrics on the Tanks and Temples Dataset.

		time (sec)			ATE↓			RTA@3↑			RRA@3↑			RTA@1↑			RRA@1↑			AUC-R&T @ 3 ↑			AUC-R&T @ 1 ↑		
	n_imgs	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP
nosr_europa	309	33	239	3387	9.3e-4	9.0e-4	8.8e-4	88.7	88.8	89.4	100.0	100.0	100.0	59.2	60.1	60.2	98.5	97.2	98.3	63.2	63.5	63.9	33.7	33.7	34.1
nosr_lk2	199	32	117	570	1.7e-3	5.0e-4	4.8e-4	96.3	97.2	97.2	100.0	100.0	100.0	87.4	88.5	88.2	96.7	100.0	99.0	83.4	84.8	84.3	63.0	65.1	64.0
nosr_lwp	354	41	202	2020	5.9e-4	6.0e-4	5.9e-4	96.0	96.1	96.2	100.0	100.0	100.0	74.5	75.7	76.2	100.0	100.0	100.0	74.2	75.9	76.2	43.5	48.3	48.4
nosr_rathaus	515	76	593	5965	5.4e-4	5.1e-4	5.1e-4	88.3	88.1	88.6	99.9	100.0	100.0	54.9	55.1	56.5	99.5	100.0	100.0	59.9	60.5	61.2	26.9	28.4	28.8
nosr_schloss	379	44	320	2675	6.1e-4	6.1e-4	5.8e-4	92.2	92.1	92.2	100.0	100.0	100.0	72.7	73.2	73.8	99.9	100.0	100.0	71.7	72.3	72.6	43.3	44.8	45.1
nosr_st	397	40	262	1435	2.9e-3	3.0e-3	2.9e-3	96.5	96.2	96.0	98.0	98.5	98.0	86.9	86.4	83.4	96.8	95.6	96.7	80.1	80.3	78.6	53.3	54.8	51.5
nosr_stjacob	722	80	564	6267	4.9e-3	1.7e-3	3.3e-3	90.4	93.0	92.2	97.0	99.7	98.9	75.7	78.0	76.8	96.8	99.4	98.9	73.1	75.9	74.8	47.3	50.6	49.5
nosr_stjohann	347	50	296	2986	7.8e-4	7.9e-4	7.9e-4	84.9	84.7	85.1	100.0	100.0	100.0	57.8	57.9	58.3	99.3	99.4	99.4	61.6	61.9	62.1	34.7	36.1	35.8

Table 9. Per scene camera pose metrics on the NeRF-OSR Dataset.

B.4. Tracks

Consider the graph in which 2D keypoints are nodes and pairwise edges denote keypoint matches. When a 3D scene point is observed in m different images, the projected 2D keypoints should ideally form a complete subgraph with m vertices. In practice, keypoint matching has low recall and tends to miss many point-pairs. A subgraph of related keypoints is often far from fully connected. Since the number (and quality) of matches is critical to the accuracy of pose estimation, we make up for low matching recall by *track completion*. A *track* is a connected component in the 2D keypoint connectivity graph and implies the existence of a shared 3D point. Tracks are used heavily in SfM [44, 53] to impose extra constraints, e.g., bundle adjustment [60] initializes 3D points based on tracks and minimizes the reprojection error of each 3D point with its track members.

FastMap avoids bundle adjustment and data structures containing both 3D scene points and 2D keypoints. Instead we explicitly convert tracks to additional matches with pairwise combinations of all keypoints in each track. This way we still make use of the transitivity of matching and benefit from the extra constraints. These additional point pairs are only introduced after global rotation alignment, and are used in the global translation alignment and epipolar adjustment steps described below.

B.5. Global Translation

B.5.1 Relative Translation

Global translation alignment in our method relies heavily on relative translations between image pairs. Rather than using the translations determined via pose decomposition, we first re-estimate the relative translations. We do so for two reasons. First, since we have estimated the global rotations, we can go back and re-compute the relative rotation of any image pair. The relative rotation computed in this way is much more accurate than those from relative pose decomposition. In turn, a better estimate of the relative rotation enables us to more accurately estimate relative translation. Second, after generating new point pairs from tracks, some image pairs that originally had no matches might have some now, and we can use these new point pairs to estimate the relative translation. We use 2D grid search to re-estimate the unit relative translation vectors. We first sample a set of candidates on the surface of the unit sphere, evaluate the mean epipolar error of each sample, and choose the candidate with the lowest error as the final estimate.

B.5.2 Multiple Initializations

Although random initialization works surprisingly well for the objective in Eqn. 6, it occasionally produces a small number of outliers. To deal with the problem, we propose to do multiple independent runs from different random initializations, and merge the solutions as the initialization for the final optimization loop. Since the global rotations are the same, different solutions can be aligned by moving the

		time (sec)			ATE↓			RTA@3↑			RRA@3↑			RTA@1↑			RRA@1↑			AUC-R&T @ 3 ↑			AUC-R&T @ 1 ↑		
	n_imgs	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP
dploy_house1	220	34	138	419	7.5e-5	1.1e-4	1.1e-4	100.0	99.9	99.9	100.0	100.0	100.0	99.5	99.1	99.0	100.0	100.0	100.0	93.9	93.2	92.9	82.1	80.2	79.6
dploy_house2	725	124	592	5702	6.9e-5	7.2e-5	1.3e-4	99.9	99.9	99.8	100.0	100.0	100.0	99.2	99.3	95.6	99.6	99.9	87.3	91.6	91.7	82.0	75.4	75.6	51.1
dploy_house3	180	50	103	627	4.6e-3	1.2e-2	1.1e-3	95.1	95.9	97.2	94.5	95.6	97.9	80.7	85.8	78.7	70.5	71.2	59.1	66.7	68.0	58.6	24.7	26.4	21.3
dploy_house4	349	57	128	751	1.7e-2	1.9e-2	9.4e-3	94.8	98.3	98.8	95.6	97.7	98.9	90.8	97.8	97.9	86.2	97.7	98.3	81.1	89.9	91.0	62.0	74.5	76.7
dploy_pipes1	97	17	42	138	6.8e-5	6.2e-5	6.1e-5	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.9	99.9	100.0	100.0	100.0	95.8	95.8	95.8	87.5	87.3	87.4
dploy_ruins2	1171	156	907	8672	5.2e-5	4.1e-5	2.0e-4	100.0	100.0	98.3	100.0	100.0	100.0	99.0	99.4	82.3	98.0	100.0	81.6	88.8	92.1	71.9	67.2	76.5	30.6
dploy_ruins3	523	84	324	4986	1.2e-3	5.6e-3	3.5e-3	98.3	94.9	95.7	98.4	99.3	83.4	75.4	79.8	45.8	39.4	46.0	24.1	56.4	59.0	39.1	8.9	10.9	3.8
dploy_tower1	775	188	419	2809	2.0e-2	2.0e-3	2.5e-3	93.7	95.2	87.7	99.2	99.5	98.3	86.4	84.5	51.5	63.7	74.4	37.9	65.6	67.5	47.2	20.9	23.7	10.1
dploy_tower2	682	106	634	6060	1.7e-4	1.6e-4	1.3e-3	99.9	99.8	44.6	100.0	100.0	25.9	75.5	77.6	10.6	100.0	100.0	9.6	72.7	72.9	8.4	24.5	26.2	0.8

Table 10. Per scene camera pose metrics on the DroneDeploy Dataset.

		time (sec)			ATE↓			RTA@3↑			RRA@3↑			RTA@1↑			RRA@1↑			AUC-R&T @ 3 ↑			AUC-R&T @ 1 ↑		
	n_imgs	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP
z_alameda	1734	134	848	4541	1.1e-3	1.9e-4	6.7e-6	99.1	99.4	100.0	99.7	100.0	100.0	98.9	99.3	99.9	99.1	99.7	99.9	95.4	97.8	98.5	88.1	94.8	95.6
z_berlin	1511	152	893	6478	1.0e-3	1.5e-2	1.1e-3	97.7	93.0	98.9	96.8	92.9	98.9	91.8	92.8	98.8	94.7	92.8	98.3	83.6	90.9	96.6	61.0	86.9	92.4
z_london	1874	102	566	2643	6.6e-5	1.3e-2	2.8e-4	99.8	99.9	99.9	99.8	99.9	99.9	99.6	99.9	99.9	99.5	99.9	99.9	96.4	98.6	98.5	89.9	96.0	95.8
z_nyc	990	88	451	1618	9.7e-3	6.8e-6	5.3e-6	99.4	100.0	100.0	99.4	100.0	100.0	99.1	100.0	100.0	99.4	100.0	100.0	95.0	98.9	98.9	86.5	96.8	96.8
mill19_building	1920	258	6289	27080	3.0e-4	1.3e-2	1.9e-5	99.9	0.1	99.9	100.0	7.4	100.0	99.3	0.0	99.3	100.0	1.9	99.9	95.5	0.0	95.6	87.0	0.0	87.4
mill19_rubble	1657	240	2849	12153	3.6e-5	6.4e-5	3.4e-5	99.9	99.8	99.9	100.0	99.9	100.0	98.6	98.6	98.7	100.0	99.9	100.0	93.6	94.5	94.6	81.6	84.7	84.8
urbn_Campus	5871	740	3869	106055	1.1e-5	4.7e-6	5.0e-6	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.9	99.9	100.0	100.0	100.0	94.0	98.0	97.9	81.9	94.1	93.7
urbn_Residence	2582	359	2523	36778	2.8e-5	2.7e-5	2.6e-5	99.8	99.9	99.9	100.0	100.0	100.0	98.8	98.9	99.0	100.0	100.0	100.0	94.6	95.2	95.4	84.6	86.3	86.8
urbn_Sci-Art	3019	445	4601	42032	1.4e-5	1.0e-5	1.1e-5	100.0	100.0	100.0	100.0	100.0	100.0	99.9	99.9	99.9	100.0	100.0	100.0	97.4	97.7	97.8	92.2	93.1	93.5
eft_apartment	3804	549	5905	185361	2.8e-3	9.4e-3	2.2e-3	86.8	75.0	90.2	89.1	75.6	92.4	51.1	61.3	71.7	38.1	56.6	70.6	45.5	50.5	62.0	6.4	18.2	21.9
eft_kitchen	6042	2202	22884	timeout	3.1e-3	7.4e-3	-	85.0	59.9	-	85.1	62.3	-	46.7	51.7	-	26.4	44.5	-	38.1	41.2	-	4.6	14.4	-

Table 11. Per scene camera pose metrics on several large-scale datasets including ZipNeRF, Mill-19, Urbanscene3D and Eyeful Tower.

		Instant-NGP			Gaussian Splatting		
		FastMap	GLOMAP	COLMAP	FastMap	GLOMAP	COLMAP
training	Barn	23.37	23.68	23.69	26.17	27.81	27.79
	Caterpillar	20.17	20.20	20.23	23.30	23.47	23.59
	Courthouse	19.96	14.73	20.27	21.12	12.23	22.25
	Ignatius	18.11	18.14	18.26	21.42	22.04	21.86
	Meetingroom	21.61	22.59	22.41	23.71	25.35	25.17
	Truck	21.19	16.86	21.45	23.58	18.34	24.50
intermediate	Family	22.10	21.99	21.45	23.67	24.54	24.75
	Francis	23.68	23.73	23.48	26.94	27.30	27.59
	Horse	21.07	21.04	21.13	22.96	24.05	23.89
	Lighthouse	20.84	20.99	20.91	22.00	22.19	22.12
	M60	24.80	25.11	25.11	26.44	28.07	27.95
	Panther	25.25	26.01	25.74	27.13	28.27	27.97
	Playground	21.48	21.96	21.99	24.12	26.00	26.03
	Train	19.14	19.26	19.23	20.66	21.79	21.65
advanced	Auditorium	17.05	19.41	19.76	17.38	16.67	24.20
	Ballroom	17.94	13.92	19.12	19.17	11.91	23.64
	Courtroom	17.39	18.95	17.80	20.87	23.11	22.77
	Museum	14.58	14.73	14.53	20.00	20.97	20.96
	Palace	16.94	17.57	17.19	16.41	18.99	20.05
	Temple	15.65	17.10	16.87	19.08	20.80	20.73

Table 12. Per-scene novel view synthesis results on Tanks and Temples.

centroid to the origin and rescaling uniformly to have unit average norm. Then for each image the solution with the lowest average loss is chosen to be in the merged result.

B.6. Epipolar Adjustment

In this section we derive the following equivalent form of the L2 epipolar adjustment objective (the re-weighting objective is similar)

$$\mathcal{L} = \frac{1}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}(\tilde{\boldsymbol{x}}_{nm}^{(2)\top}\mathbf{E}_n\tilde{\boldsymbol{x}}_{nm}^{(1)})^2 = \frac{2}{Z}\sum_{n=1}^{|\mathcal{P}|}\boldsymbol{e}_n^\top\mathbf{W}_n\boldsymbol{e}_n \tag{15}$$

Note that each error term is linear in the essential matrix, so we can re-write it as a dot product of the a weight vector and a flattened version of the essential matrix

$$\begin{aligned}
\mathcal{L} &= \frac{1}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}(\tilde{\boldsymbol{x}}_{nm}^{(2)\top}\mathbf{E}_n\tilde{\boldsymbol{x}}_{nm}^{(1)})^2 && \text{(16a)}\\
&= \frac{1}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}((\tilde{\boldsymbol{x}}_{nm}^{(2)}\tilde{\boldsymbol{x}}_{nm}^{(1)\top})\otimes\mathbf{E}_n)^2 && \text{(16b)}\\
&= \frac{1}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}(\boldsymbol{w}_{nm}^\top\boldsymbol{e}_n)^2, && \text{(16c)}
\end{aligned}$$

where $\otimes$ is the element-wise multiplication operator, $\boldsymbol{w}_{nm} = \text{flatten}(\tilde{\boldsymbol{x}}_{nm}^{(2)}\tilde{\boldsymbol{x}}_{nm}^{(1)\top}) \in \mathbb{R}^9$, and $\boldsymbol{e}_n = \text{flatten}(\mathbf{E}_n) \in \mathbb{R}^9$. Now re-arrange the terms to get the summation of a set of quadratic forms

$$\begin{aligned}
\mathcal{L} &= \frac{1}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}(\boldsymbol{w}_{nm}^\top\boldsymbol{e}_n)^2 && \text{(17a)}\\
&= \frac{1}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}(\boldsymbol{w}_{nm}^\top\boldsymbol{e}_n)^\top(\boldsymbol{w}_{nm}^\top\boldsymbol{e}_n) && \text{(17b)}\\
&= \frac{2}{Z}\sum_{n=1}^{|\mathcal{P}|}\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}\boldsymbol{e}_n^\top\boldsymbol{w}_{nm}\boldsymbol{w}_{nm}^\top\boldsymbol{e}_n && \text{(17c)}\\
&= \frac{2}{Z}\sum_{n=1}^{|\mathcal{P}|}\boldsymbol{e}_n^\top\left(\sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}\boldsymbol{w}_{nm}\boldsymbol{w}_{nm}^\top\right)\boldsymbol{e}_n && \text{(17d)}\\
&= \frac{2}{Z}\sum_{n=1}^{|\mathcal{P}|}\boldsymbol{e}_n^\top\mathbf{W}_n\boldsymbol{e}_n && \text{(17e)}
\end{aligned}$$

where $\mathbf{W}_n = \sum_{m=1}^{|\tilde{\mathcal{Q}}_n|}\boldsymbol{w}_{nm}\boldsymbol{w}_{nm}^\top \in \mathbb{R}^{9\times 9}$

		$m = 1$	$m = 2$	$m = 3$	$m = 4$	$m = 8$
zipnerf_nyc	RTE@30	**0.59**	**0.09**	0.09	0.09	0.10
	RTA@3	98.59	99.45	99.45	99.45	99.45
zipnerf_alameda	RTE@30	**0.32**	**0.14**	0.13	0.09	0.09
	RTA@3	97.90	98.28	98.29	98.40	98.39
tnt_Train	RTE@30	**1.40**	**0.02**	0.02	0.02	0.29
	RTA@3	97.15	99.62	99.64	99.61	99.08
tnt_Lighthouse	RTE@30	**1.98**	**0.01**	0.01	0.02	0.01
	RTA@3	96.82	99.26	99.23	99.34	99.33
dploy_ruins3	RTE@30	**1.23**	**0.66**	0.67	0.92	0.69
	RTA@3	92.70	94.56	94.44	94.06	93.77
dploy_house4	RTE@30	**3.82**	**0.83**	1.00	0.83	0.83
	RTA@3	93.00	96.92	95.69	97.48	95.48

Table 13. Ablation of multiple translation initializations on selected scenes. The results are obtained right after translation alignment and before epipolar adjustment. We bold the RTE@30 entries for $m = 1$ and $m = 2$ initializations to highlight its effect.

B.7. Sparse Reconstruction

After pose estimation, we do a sparse reconstruction of the scene by triangulating the matched keypoint pairs from track completion. The 3D points corresponding to the same track are merged by averaging. To eliminate outliers, after merging the 3D points, we compute the re-projection error for each 2D keypoint, and mark those with large errors to be outlier keypoints. A 3D point is dropped if the number of inlier keypoints in the track is smaller than 3. We also filter out a 3D point if the maximal triangulation angle is smaller than some threshold.

B.8. Data Ground Truth

With the exception of Tanks and Temples, each of the datasets we evaluate on includes author-provided reference camera poses. These reference poses are obtained through different means, including COLMAP (for MipNeRF360, ZipNeRF, NeRF-OSR), PixSfM [35] (for Mill-19 and Urbanscene3D), and commercial software (for DroneDeploy and Eyeful Tower). In the case of Urbanscene3D, we use the poses provided by Turki et al. [61]. For Tanks and Temples, we use COLMAP poses provided by Kulhanek and Sattler [31]. On one of the scenes from Tanks and Temples (Courthouse), we found that the reference poses are obviously inconsistent with the images, but decided to still treat them as ground-truth.

B.9. Additional Ablation Study

B.9.1 Track completion

In Table 14 we show the final performance of the FastMap with and without augmented point pairs from track completion (Sec. B.4). Track completion significantly improves performance for MipNeRF360 scenes and some but not all ZipNeRF scenes.

		AUC@3	AUC@10	RTA@1	RTA@5	RRA@3
m360 (9)	FastMap	**97.2**	**99.1**	**99.8**	**100.0**	**100.0**
	w/o epipolar adjustment	75.0	90.8	85.5	99.5	94.5
	w/o track completion	80.4	86.4	83.3	91.4	83.6
alameda	FastMap	**95.2**	**98.1**	**99.0**	**99.3**	**99.9**
	w/o epipolar adjustment	86.7	95.5	94.9	99.2	99.8
	w/o track completion	94.8	**98.1**	**99.0**	99.4	**99.9**
berlin	FastMap	**81.6**	**93.2**	**92.8**	**99.2**	**97.5**
	w/o epipolar adjustment	70.4	89.5	82.4	98.8	95.7
	w/o track completion	60.4	81.3	70.4	90.8	90.3
london	FastMap	**96.6**	**98.8**	**99.6**	**99.9**	99.7
	w/o epipolar adjustment	90.1	96.6	97.8	99.7	99.3
	w/o track completion	96.1	98.7	**99.6**	**99.9**	**99.8**
nyc	FastMap	**94.6**	**98.1**	99.2	99.6	99.6
	w/o epipolar adjustment	89.7	96.7	97.0	99.7	99.6
	w/o track completion	93.9	98.0	**99.4**	**99.8**	**99.8**

Table 14. Epipolar adjustment and track completion ablation on the MipNeRF360 [4] and ZipNeRF [5] datasets. Results for MipNeRF360 are averaged over all the scenes, and for ZipNeRF each scene is listed separately.

B.9.2 Multiple translation initializations

As shown in Table 13, while a single initialization is prone to large-error outliers (see RTE@30 defined as 100.0 - RTA@30), increasing the number of initalizations improves performance. However, the effect plateaus with increased initializations and does not completely fix the outlier problem.

B.9.3 Epipolar adjustment

Table 14 also presents the performance of FastMap with and without the final epipolar adjustment step. On all metrics, epipolar adjustment consistently improves over the poses from global translation alignment. The improvement is more prominent for stricter metrics (RTA@1), but less so for more tolerant metrics like RTA@5. This suggests that after translation alignment the cameras are already roughly in place, and epipolar adjustment continues to squeeze as much precision as it can.

C. Limitations

Sparse Views. Our method assumes that the input images densely cover the 3D scene. Many components in the pipeline implicitly assume that the coverage is dense so that the negative effect of outlier image or point pairs will be averaged out. If the coverage is sparse, the pipeline will be sensitive to outliers and likely break down. We tested our method on the ETH3D MVS (DSLR) [54], where each scene only contains a small number of images, and show the results in Tab. 15. While FastMap still succeeds on many scenes, it is less robust than GLOMAP.

Intrinsics Estimation. The intrinsics estimation algorithms in our method can fail under certain cases. Since the interval search used in both distortion and focal length estimation requires at least one image pair of images with shared

		ATE↓		RTA@3↑		AUC-R&T @ 3 ↑		AUC-R&T @ 1 ↑	
	n_imgs	FastMap	GLOMAP	FastMap	GLOMAP	FastMap	GLOMAP	FastMap	GLOMAP
botanical_garden	30	8.2e-3	4.3e-4	86.9	100.0	68.3	94.3	52.0	83.8
boulders	26	6.7e-4	1.4e-4	99.1	100.0	91.2	97.0	76.2	91.0
bridge	110	1.3e-2	2.0e-5	92.9	100.0	85.3	97.7	73.3	93.1
courtyard	38	3.8e-2	1.8e-4	18.9	100.0	6.9	96.0	2.2	88.2
delivery_area	44	8.4e-2	8.1e-5	23.6	100.0	13.9	97.8	6.1	93.3
door	7	-	1.2e-4	-	100.0	-	98.0	-	94.1
electro	45	7.5e-2	3.0e-2	86.3	95.2	76.9	91.1	61.6	84.1
exhibition_hall	68	7.0e-2	6.9e-2	2.8	45.1	0.9	40.9	0.1	34.3
facade	76	6.5e-2	9.7e-5	71.0	100.0	66.8	97.4	60.8	92.4
kicker	31	5.9e-4	1.6e-2	98.5	93.8	86.6	91.7	65.0	88.1
lecture_room	23	3.0e-2	2.5e-4	84.2	100.0	71.7	95.0	55.3	85.7
living_room	65	1.3e-4	8.4e-5	99.7	99.8	95.3	96.2	86.8	89.2
lounge	10	9.6e-2	9.5e-2	33.3	33.3	32.3	32.7	30.2	31.4
meadow	15	1.4e-1	1.4e-1	13.3	86.7	7.7	80.2	4.9	68.2
observatory	27	6.5e-3	5.8e-4	94.9	99.1	76.5	86.5	48.5	63.9
office	26	9.7e-3	7.6e-4	54.8	95.7	43.9	82.7	34.5	61.2
old_computer	54	6.8e-2	5.6e-2	21.7	65.3	16.0	60.9	9.8	53.5
pipes	14	5.8e-4	2.6e-4	98.9	100.0	92.5	97.4	79.8	92.3
playground	38	8.3e-4	1.1e-4	99.4	99.9	89.4	97.1	70.8	91.7
relief	31	6.1e-3	7.2e-5	77.8	100.0	62.1	98.4	48.9	95.2
relief_2	31	3.7e-4	7.9e-5	99.8	100.0	94.5	98.4	84.3	95.1
statue	11	5.5e-5	2.3e-5	100.0	100.0	99.5	99.7	98.5	99.0
terrace	23	2.1e-4	1.2e-4	100.0	100.0	97.5	97.7	92.5	93.1
terrace_2	13	2.6e-4	2.2e-4	100.0	100.0	96.6	96.9	89.9	90.8
terrains	42	1.3e-3	2.1e-4	94.4	99.8	70.9	94.6	39.3	84.6

Table 15. Per scene camera pose metrics on ETH3D.

	Recall@1m↑		AUC@1m↑		AUC@5m↑	
	FastMap	GLOMAP	FastMap	GLOMAP	FastMap	GLOMAP
CAB	4.77	11.6	4.32	4.7	4.74	16.9
HGE	5.94	48.4	5.26	22.2	5.70	50.3
LIN	7.16	87.3	3.95	46.7	7.96	85.6

Table 16. Per scene camera pose metrics on LaMAR.

intrinsics to begin with, it will not work if all the images have different intrinsics. It is also not robust if the number of images for each distinct camera is small. In addition, the focal length extraction method relies entirely on fundamental matrices, and is unreliable when the scene is dominated by homographies.

Homography. Apart from the impact on focal length estimation, too many homography image pairs can also jeopardize relative pose decomposition. Both essential and homography decomposition produce four different solutions, and they are usually disambiguated with a cheirality check. But for some homography and keypoint pairs, cheirality check is not enough for determining a unique solution. Our current strategy is to simply pick the solution with the lowest index if there is a tie. This has potential issues if there are too many homography image pairs.

Repetitive Patterns and Symmetric Structures Non-learning based keypoint features and matching are not robust in the cases of repetitive patterns and symmetric structures in the scene. These wrong matches are hard to filter because they can have a lot of inlier point pairs with a very consistent two-view geometric model. Most traditional SfM methods are more or less impacted by these erroneous matches, and so is ours. In our experiments this problem is most prominent in the advanced split of the Tanks and Temples dataset (Tab. 8).

Degenerate Motions One important reason why bundle adjustment is popular in previous SfM methods is that it can utilize 3D points to resolve some ambiguities in global translation estimation. For example, when all the cameras are aligned in the same line, optimization methods that solely rely on relative motions or epipolar errors might fail because there is no way to uniquely (up to scale) determine the distance of any pair of cameras. Bundle adjustment uses tracks to impose extra constraints to solve this problem. This scenario is commonly seen in SLAM-like datasets. We tested our method on the large-scale LaMAR [52] dataset and show the results in the Tab. 16. Each scene in LaMAR consists of multiple trajectories of a moving VR headset or hand-held phone. These trajectories contain many straight and forward motions, and different trajectories only overlap sparsely. Our method does not work well compared to GLOMAP.