Relative Pose Estimation through Affine Corrections of Monocular Depth Priors

Supplementary Material

A. Additional Details on Experimental Setup

Generating 2D-3D-S [2] image pairs: As mentioned in Sec. 5.1 in the main paper, we hereby provide the details of how the 1064 image pairs from the Stanford 2D-3D-S [2] panoramic dataset are generated.

The dataset includes panoramic scans captured in different rooms across 7 different areas. We generate 4 image pairs from each pair of panoramas captured in a same room by:

- First, generate a random roll ($\pm 30°$), pitch ($\pm 30°$), yaw ($\pm 180°$), and FOV (60°- 105°, effectively a random focal length);
- Then, project the first image using the generated roll, pitch, yaw, and FOV using equirectangular projection, and crop the image to a randomly chosen crop size among 1080x540, 960x540, 1024x768, 640x480 (WxH);
- Next, sample a random pixel within the middle $(W/2, H/2)$ of the first image (thus ensuring reasonable covisibility), and lift-project it into the second panorama using the ground truth depth;
- Finally, project the second image using equirectangular projection by centering on this projected point, with again a random roll ($\pm 30°$), a random FOV (60°- 105°), and a random crop size.

Among all generated image pairs, we randomly sample a maximum of 152 image pairs per each of the 7 areas, resulting in a total of 1064 image pairs.

Hyperparameters: The RANSAC thresholds are tuned for best performance for our method as well as the baselines on each dataset. The reprojection error threshold τ_r for our method is set to 8px for ScanNet [15] images (resized to 640x480) and ETH3D [59] images (resized to 720x480), and 16px for MegaDepth [38] and Stanford 2D-3D-S [2] sampled images. The Sampson error threshold τ_s for our method is set to 2px on ScanNet and 1px on other datasets. The epipolar error threshold for the PoseLib [33] baselines is set to 1px on ETH3D and 2px on other datasets. The threshold of GC-RANSAC [3] for the solvers from [17] is set to 0.75px uniformally.

As mentioned in the main paper, we empirically fix the Sampson error weight λ_s to 1.0 in our experiments to demonstrate the effectiveness of the proposed hybrid estimation. This however can be tunable to adjust to different reliability of the depth priors and feature matchers on different dataset to make the estimator focus more on the depth-augmented correspondences or pure point correspondences.

Task	Method / MD Model	Med. Err. ↓ $\varepsilon_R(°)$	$\varepsilon_t(°)$	$\varepsilon_f(\%)$	Pose Err. AUC (%) ↑ @5°	@10°	@20°
ScanNet-1500 Calibrated	PoseLib-5pt	2.08	5.44	-	19.48	38.09	56.08
	Omnidata	1.76	5.42	-	21.24	39.74	57.75
	Marigold	1.72	5.28	-	21.18	40.38	58.13
	DA-met.	<u>1.68</u>	<u>4.97</u>	-	<u>22.41</u>	<u>42.18</u>	<u>59.96</u>
	DAv2 (inv)	1.90	5.72	-	19.60	38.10	56.18
	DAv2-met.	1.72	5.25	-	21.90	41.01	59.16
	MoGe	**1.57**	**4.77**	-	**23.36**	**43.39**	**61.08**
	GT Depth	1.54	4.57	-	25.03	45.06	61.85
ETH3D Shared-focal	PoseLib-6pt	0.90	1.81	8.79	46.29	56.79	65.43
	Omnidata	1.09	2.42	9.78	41.90	54.75	65.38
	Marigold	<u>0.94</u>	<u>1.79</u>	<u>8.06</u>	<u>45.55</u>	<u>58.91</u>	<u>68.71</u>
	DA-met.	1.06	2.37	10.47	41.81	53.33	62.94
	DAv2 (inv)	1.26	2.84	9.85	38.50	52.74	64.85
	DAv2-met.	**0.86**	**1.78**	**7.71**	**47.60**	**59.60**	**69.41**
	GT Depth	0.36	0.81	2.07	62.26	70.52	75.90
MegaDepth-1500 Two-focal	PoseLib-7pt	1.97	5.72	23.64	21.23	36.80	54.89
	Omnidata	<u>1.53</u>	<u>5.41</u>	<u>18.60</u>	<u>22.67</u>	<u>39.94</u>	<u>59.15</u>
	Marigold	2.03	6.72	23.61	18.56	34.01	53.46
	DA-met.	**1.25**	**4.81**	**15.99**	**25.70**	**42.90**	**61.79**
	DAv2 (inv)	1.62	5.69	18.93	21.50	38.46	57.04
	DAv2-met.	2.06	7.45	24.53	18.05	32.44	50.33
	GT Depth	0.48	3.32	6.43	38.04	54.85	70.08

Table 8. Results with different MDE models on three tasks on three different datasets. All results are with SP+LG matches. Best results among the different models on each task are **bolded**, and second best <u>underlined</u>.

B. Additional Experiment Results

Results with different monocular depth models: Our method is designed to work with any off-the-shelf MDE models, with Depth-Anything variants [74, 75] and MoGe [69] giving the best results. The accuracy can further benefit from developments on more accurate MDE models.

We include here a comparison of using different monocular depth estimation (MDE) models with our method across three tasks and three datasets in Tab. 8. Our method can improve upon the baseline with both metric depth priors (Depth-Anything v1 [74] and v2 [75] metric models) and non-metric relative depth priors (Omnidata [31], Marigold [32], MoGe [69]). MoGe is only evaluated in the calibrated setting due to its ability to also produce a good estimation of focal lengths, and can therefore directly benefit from using the more accurate calibrated estimation in the uncalibrated cases. We include a row using the GT Depth as the "depth priors" for each task to show the potential of our method with potentially more advanced monocular depth models especially for the shared-focal and two-focal settings. It is worth noting that, while disparity priors in general do not align with our affine-invariant relative depth formulation and inverting those would break

	Method	MD Model	Med. Err. ↓ $\varepsilon_{\boldsymbol{R}}(^\circ)$	$\varepsilon_{\boldsymbol{t}}(^\circ)$	Pose Error AUC (%) ↑ @5°	@10°	@20°
From [4]	2PT+D & 4PT+D	DA-met.	5.31	17.65	6.42	16.32	29.90
	2PT+D & 4PT+D	MoGe	4.10	14.43	8.53	20.03	34.47
	2PT+D & 5pt	DA-met.	1.90	5.67	20.62	38.45	54.94
	2PT+D & 5pt	MoGe	1.88	5.66	20.74	38.53	54.98
	Sim. P3P & 5pt	DA-met.	1.87	5.64	20.92	38.48	54.63
	Sim. P3P & 5pt	MoGe	1.90	5.76	20.59	38.20	54.43
Ours	Ours-calib	DA-met.	<u>1.68</u>	<u>4.97</u>	<u>22.41</u>	<u>42.18</u>	<u>59.96</u>
	Ours-calib	MoGe	**1.57**	**4.77**	**23.36**	**43.39**	**61.08**

Table 9. Comparison with scale-only solvers from [4] with calibrated cameras on ScanNet-1500.

	Method	MD Model	Med. Err. ↓ $\varepsilon_{\boldsymbol{R}}(^\circ)$	$\varepsilon_{\boldsymbol{t}}(^\circ)$	Pose Error AUC (%) ↑ @5°	@10°	@20°
From [4]	2PT+D & 4PT+D	DA-met.	6.84	25.54	8.89	16.64	26.96
	2PT+D & 4PT+D	MoGe	3.54	14.56	13.72	23.74	36.20
	2PT+D & 5pt	DA-met.	0.52	1.37	57.83	72.85	83.73
	2PT+D & 5pt	MoGe	0.54	1.44	55.86	70.97	82.20
	Sim.P3P & 5pt	DA-met.	0.49	1.34	58.23	72.97	83.68
	Sim.P3P & 5pt	MoGe	0.57	1.52	55.99	71.14	82.26
Ours	Ours-calib	DA-met.	<u>0.47</u>	<u>1.26</u>	<u>59.80</u>	<u>74.77</u>	<u>85.47</u>
	Ours-calib	MoGe	**0.41**	**1.16**	**63.48**	**77.79**	**87.18**

Table 10. Comparison with scale-only solvers from [4] with calibrated cameras on MegaDepth-1500.

the affine-invariance of disparity values, we find that on outdoor images inverting the Depth-Anything-v2 [75] disparities could lead to better results than its metric depth sibling. We postulate that this is due to the disparity being able to encode a larger range of depths within the output range of the model, which is beneficial for outdoor scenes.

Additional comparison with solvers from [4]: The solvers proposed in [4] considers monocular depth priors in solving relative poses, thus they are highly related to our work. However, their modeling only considers the scale of the depth priors without the shift. We compare our method with the three minimal solver & non-minimal solver configurations mentioned in [4]: 2PT+D & 4PT+D, 2PT+D & 5pt, and Simulated P3P & 5pt. We use the implementations obtained from the author and plug them in the GC-RANSAC [3] framework. The results of the best performing 2PT+D & 5pt combination are reported in Tab. 1 and Tab. 2. The full comparison results (with calibrated cameras, SP+LG matches) are shown in Tab. 9 and Tab. 10. Our method consistently outperforms the scale-only methods from [4]. In addition, as we mentioned in Sec. 2, we find that the 2PT+D solver suffers from degeneration of using only 2 correspondences due to rank deficiency.

Additional visual results: We provide more visualization examples in addition to Fig. 4. In Fig. 7 and Fig. 8 we show examples on ETH3D [59] with the shared-focal setting, and on 2D-3D-S [2] images with the two-focal setting. By incorporating monocular depth priors, our method is able to find more accurate pose together with scale and shifts of the depth priors that lead to better and more correct alignment of the back-projected point clouds. In Fig. 9 we show examples on ScanNet [15] comparing to the scale-only ablated baseline as described in Sec. 5.3. Only modeling scale without the shift can lead to failure cases with incorrect alignment and distortion visible in the aligned point clouds.

C. Additional Discussion on Proposed Solvers

In Sec. 4.1 we mentioned the proposed calibrated solver is minimal with 3 point correspondences and related depth priors while the shared-focal and two-focal solvers are non-minimal. We provide in this section a simple reasoning of the minimality of the calibrated solver, and discuss about possible minimal versions for the shared-focal and two-focal solvers.

Minimality of Calibrated Solver: The calibrated solver takes 3 point correspondences and depth priors to solve for the relative pose $\boldsymbol{R}, \boldsymbol{t}$, depth scale α and shifts β_1, β_2. Conventionally, relative pose are solved by finding the essential matrix which has 5 degrees-of-freedom (DOFs) up to an unknown scale. In our setup, however, because the solved relative pose (translation) has a fixed scale consistent with the solved depth scale and shifts, the relative pose now has 6 DOFs. In total the problem has $6 + 1 + 2 = 9$ DOFs, and is minimally solvable with 3 point correspondences and depth priors since each pair of 2D correspondences gives 1 epipolar constraint, and with depth priors we can additionally have 2 projection constraints per pair.

Shared-focal and Two-focal Solvers: For the two solvers that we propose for uncalibrated cases, the problem is not minimal and our solvers ignore 1 or 2 of the 6 constraints we have. This means that the solutions we get might not exactly satisfy the correspondences in the sample set (which can be later taken care by the hybrid RANSAC pipeline). Another approach would be to drop some of the input data, instead of dropping equations. For example, one could take 3 pairs of point matches with depth and one pair without depth, or with partial depth (only in one view). One approach to formulate this would then be to parameterize the *missing depth* as extra unknowns. We briefly explored this option but applying [34] yielded solvers with elimination templates of size 360×374 (14 solutions) and 716×744 (28 solutions), which are too slow to be used in practice.

D. Limitation and Future Work

We discuss here some limitations of our affine modeling of the monocular depth priors and the proposed pipeline, improvements of which could lead to interesting and promising future works.

First, while our affine correction of the monocular depth priors is proven beneficial for estimating relative pose and outperforms previous methods that only model the scale, the affine modeling of depth maps is simple and limited with only two parameters (scale and shift). In practice, we have found that the estimated β might not uniformally agree with all pixels and their depth priors, but rather different groups of regions/surfaces in the image can be better fitted with different shift values. This is due to the fact that MDE models are better at inferring relative depth between pixels of the same object/surface than pixels across different surfaces due to the ambiguous scales among objects. We also observed that the same depth map could result in a few different groups of β values when estimating relative pose with different images (all with good estimated poses), depending on the different groups of regions/surfaces that are aligned by the inlier correspondences. Therefore, one interesting future work direction would be to enhance the affine modeling to more fine-grained region-based modeling, possibly with the help of the latest advances in image segmentation. At the same time, our method can also benefit from more advanced monocular depth models with better accuracy on outdoor images or inter-image consistencies as can be seen from Tab. 8.

Second, while we are able to get good results by empirically setting λ_s to 1.0 in the experiments, a more mathematically sound way of balancing between the depth-induced reprojection errors and Sampson error could be developed. This can be especially beneficial when the depth priors are less reliable (*e.g.* on outdoor images), and could utilize information such as uncertainty modeling of the depth priors and inlier ratios of the different types of correspondences.

Third, our proposed pipeline is dependent on pixel correspondences produced by off-the-shelf matchers, and therefore only limited part of the estimated depth priors are utilized. It would be interesting to explore whether depth priors of other unmatched pixels could provide additional geometric constraints.

Lastly, a natural extension of our pipeline is to extend our affine modeling to multi-view, possibly through bundle-adjustment to solve multi-view problems like structure-from-motion.

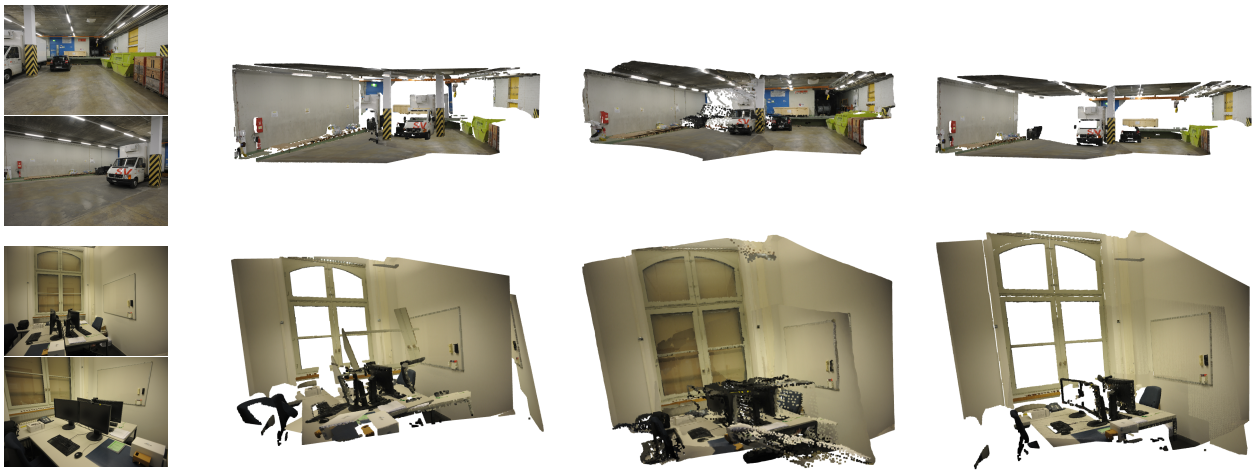

Figure 7. Additional visualizations on ETH3D [59] with shared-focal setting. **Left:** back-projected GT depth with relative pose found by PoseLib-6pt [33] and translation rescaled to match scale with GT translation; **Middle:** back-projected depth priors from Marigold [32] aligned using the output scale, shifts, relative pose, and focal length from our method; **Right:** Aligned GT depth with GT pose.

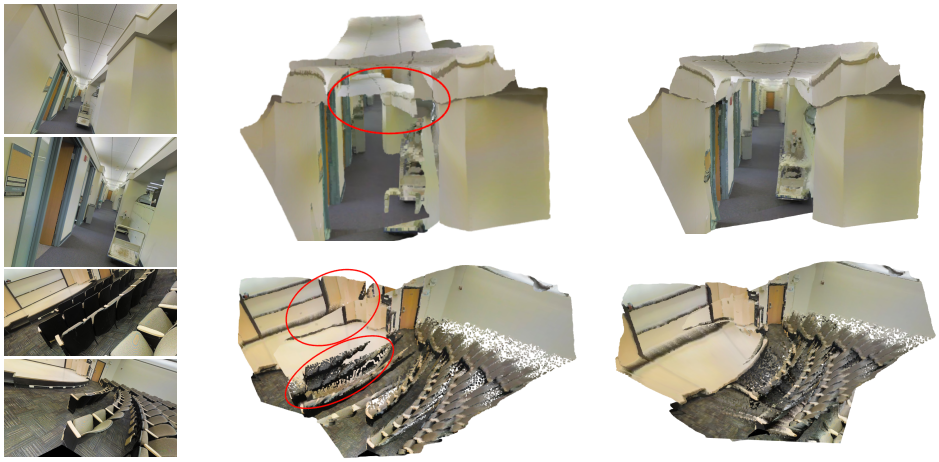

Figure 8. Additional visualizations on Stanford 2D-3D-S [2] image pairs with two-focal setting. **Left:** back-projected depth priors aligned using the relative pose found by PoseLib-7pt [33] baseline; **Right:** back-projected depth priors aligned using the relative pose found by our two-focal estimator. Both point clouds are aligned using the scale and shifts from our method, but focal lengths from each method, with translation found by the point-based baseline is rescaled to match the length of translation found by our method. (Currently no GT depth available for the sampled image pairs.)

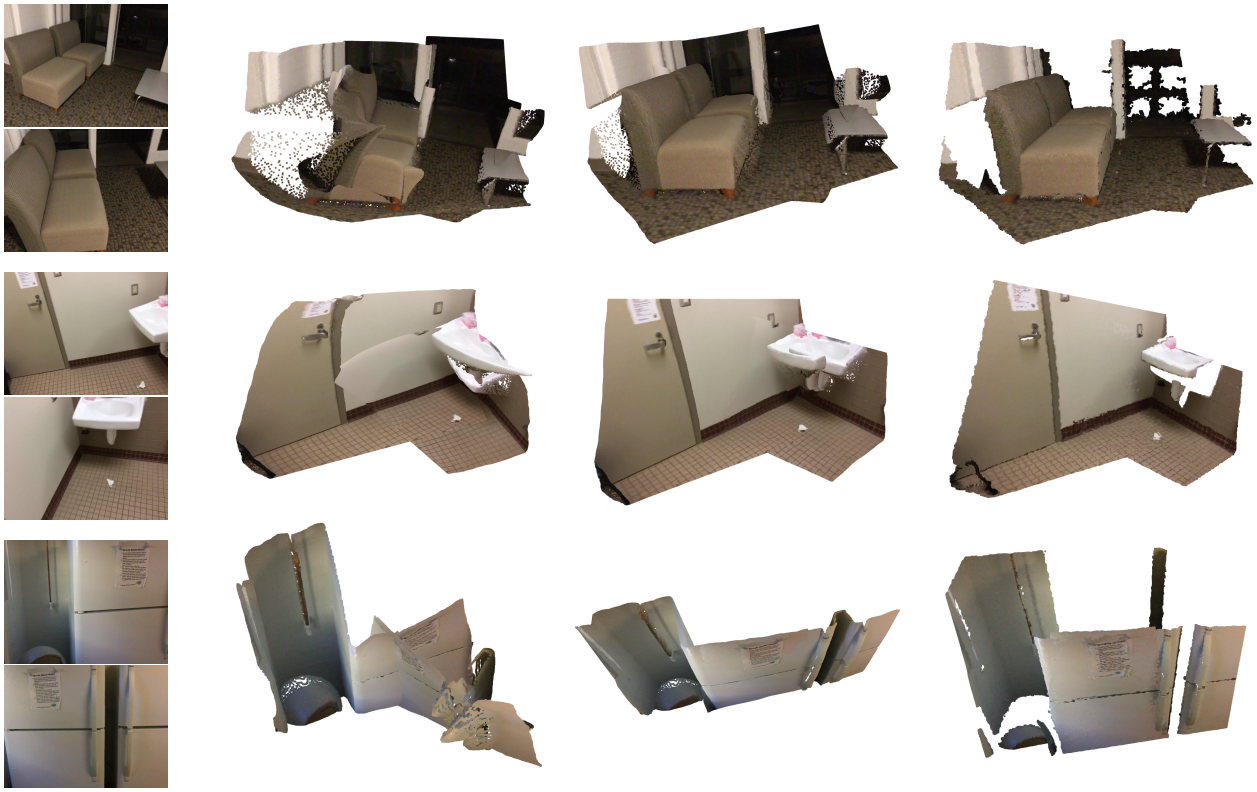

Figure 9. Visualization of aligned point clouds on image pairs from ScanNet-1500[15] using: **Left:** the scale-only ablated baseline (Sec. 5.3); **Middle:** our method (calibrated setting); **Right:** GT depth.