

Relative Pose Estimation through Affine Corrections of Monocular Depth Priors

Yifan Yu^{1,2*} Shaohui Liu¹ Rémi Pautrat³ Marc Pollefeys^{1,3} Viktor Larsson⁴
¹ETH Zurich ²PICO ³Microsoft Spatial AI Lab ⁴Lund University

Abstract

Monocular depth estimation (MDE) models have undergone significant advancements over recent years. Many MDE models aim to predict affine-invariant relative depth from monocular images, while recent developments in large-scale training and vision foundation models enable reasonable estimation of metric (absolute) depth. However, effectively leveraging these predictions for geometric vision tasks, in particular relative pose estimation, remains relatively under explored. While depths provide rich constraints for cross-view image alignment, the intrinsic noise and ambiguity from the monocular depth priors present practical challenges to improving upon classic keypoint-based solutions. In this paper, we develop three solvers for relative pose estimation that explicitly account for independent affine (scale and shift) ambiguities, covering both calibrated and uncalibrated conditions. We further propose a hybrid estimation pipeline that combines our proposed solvers with classic point-based solvers and epipolar constraints. We find that the affine correction modeling is beneficial to not only the relative depth priors but also, surprisingly, the “metric” ones. Results across multiple datasets demonstrate large improvements of our approach over classic keypoint-based baselines and PnP-based solutions, under both calibrated and uncalibrated setups. We also show that our method improves consistently with different feature matchers and MDE models, and can further benefit from very recent advances on both modules. Code is available at <https://github.com/MarkYu98/madpose>.

1. Introduction

Recently, deep learning based monocular depth estimation (MDE) has made remarkable progress [9, 31, 32, 69, 74, 75], offering increasingly accurate 3D information from single 2D images. These advancements have opened up new possibilities for enriching traditional geometric computer vision tasks with 3D priors. The ability to infer depth information from a single image has been shown to have useful implications in various applications [14, 28, 61].

*This work was primarily done as Yifan’s master thesis at ETH Zurich.

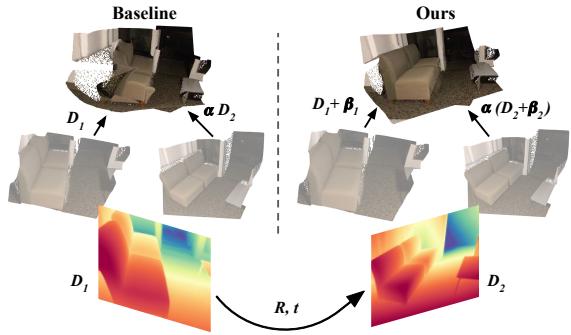


Figure 1. Our method jointly estimates affine corrections of monocular depth maps $D_1 + \beta_1$ and $\alpha(D_2 + \beta_2)$ together with relative pose R, t (Right), whereas the classic way of aligning the depth maps with only scale modeling (α) leads to wrong and distorted alignments (Left).

Despite the significant progress in monocular depth estimation, the integration of these depth priors in fundamental geometric computer vision tasks, particularly camera pose estimation, remains under-studied. While incorporating depth information would be beneficial intuitively, effectively leveraging these priors to find geometric relationships between multiple views presents unique challenges that have not been fully addressed by existing works [4, 17, 41].

A critical limitation of existing approaches is the common assumption that the predicted depth maps from different views can be related by a single scale factor. This assumption, however, fails to account for the intrinsic properties of existing monocular depth estimation models, as state-of-the-art MDE models [32, 69, 75] are typically trained to predict relative depth or disparity (inverse-depth) that are invariant to both scale and shift (affine) transformations. While there have been recent progress in developing models for metric depth estimation [8, 48, 75, 76], we find that surprisingly, they also benefit from modeling the affine corrections as they are not perfectly consistent with the actual metric depth (See Figure 6, Table 1, and Sec. 5.3).

In this paper, we propose three solvers specialized for solving relative pose under calibrated or uncalibrated camera setups. These solvers use pixel correspondences along with their depth priors from MDE models as input, and incorporate explicit modeling of both scale and shift varia-

tions in the depth predictions. Specifically, the solvers introduced are as follows (with only the calibrated solver being minimal and the two others being over-constrained):

- **Calibrated 3-point solver:** for calibrated image pairs.
- **Shared-focal 4-point solver:** for uncalibrated image pairs with a common unknown focal length.
- **Two-focal 4-point solver:** for uncalibrated image pairs.

In addition, we integrate these new solvers into a flexible hybrid robust estimation pipeline that combines our depth-aware solvers with classic point-based solvers [29, 46, 63]. We also develop hybrid solutions for scoring and local optimization, where we optimize both the classic Sampson error [53] and depth-induced reprojection error using depth and the solved affine corrections. This enables combining the advantages of both worlds and results in a robust relative pose estimator that achieves consistent improvement across datasets. Our key contributions are summarized as follows:

- We propose to solve relative pose with explicit affine (scale and shift) corrections on monocular depth predictions, addressing a limitation in existing methods.
- We develop 3 new solvers tailored to different calibration setups: calibrated, uncalibrated with common focal length, and fully uncalibrated image pairs.
- Our hybrid estimation pipeline integrates depth-aware and classic point-based solvers, scoring and local optimization, leading to largely improved accuracy and robustness in relative pose estimation.
- Our framework is compatible with a wide range of image matchers and MDE models with consistent improvement, making it easy to integrate in existing pipelines.

2. Related Work

Monocular Depth Estimation: Recovering depth information from a single image is a fundamental problem in 3D computer vision. Early methods [56, 57] resort to probabilistic modeling, and the advent of deep learning enables a variety of robust monocular depth estimators [23, 25, 27, 38, 51, 52, 69, 74, 75, 80] in the past decade. While most of them learn monocular depth with an error that is invariant to scale or affine transformation, thanks to large-scale data and vision foundation models in recent years, researchers have also developed models [7, 8, 10, 39, 48, 74–76] that aim to predict absolute (metric) depth from images. Recently, there have also been works [26, 32, 47, 58, 78] that demonstrate the benefits of repurposing generative models for monocular depth estimation. With such rapid and active developments, the latest state-of-the-art models [69, 75] have started to show foundation model level performance on open-domain images. This raises increasing attention and demand on studying how the off-the-shelf monocular depth models can best benefit geometric vision tasks. Our proposed method works with any monocular depth models and achieves improvement on relative pose estimation with

both relative and metric depth priors.

Relative Pose Estimation: Classic relative pose estimators mainly rely on cross-image point correspondences, which either comes from the detection [16, 18, 43, 67, 81] and matching [5, 40, 54] of salient keypoints, or direct dense / semi-dense matching [19, 20, 64, 66, 68]. While minimal solvers for point correspondences have been well studied in the two-view case under calibrated [46] and uncalibrated conditions [29, 63], researchers have been making continuous efforts on improving relative pose estimation with the help of data-driven machine learning, including differentiable robust estimation [50, 71], relative pose regression [82], and advanced proposal scoring [6]. However, they have not yet achieved the same generalization ability as classic methods in practice, with the LO-RANSAC [13, 36] in PoseLib [33] remaining the best practical solutions. Very recently, DUS3R [70] introduced a 3D foundation model for two-view geometry that benefits from large-scale training, which is further improved by MASt3R [37] with better matching. In this paper, we present a relative pose estimator that consistently improves over PoseLib [33] with off-the-shelf monocular depth models, and further show that the improvements are orthogonal and can be combined with the recent advances from MASt3R [37].

Monocular Depth Priors for Geometric Tasks: As monocular depth estimation become increasingly robust, there has been growing interests on using the dense depth predictions to benefit downstream geometric tasks, including visual odometry/SLAM [14, 41, 42, 65], dense reconstruction [44, 49, 72, 77], and novel view synthesis [62, 73]. However, only a few works have attempted employing monocular depth priors to enhance relative pose estimation. Barath *et al.* [4] propose a 2pt+D solver and also attempt to incorporate affine correspondence [22]. However, the 2-point solver is intrinsically degenerate for rigid alignment due to rank deficiency. In [41], PnP-RANSAC is directly applied on top of monocular depths for small-baseline image pairs. Most recently, Ding *et al.* [17] develop fundamental matrix solvers for relative depths, and fCOP [79] relates monocular depth estimation with canonical object priors for focal length estimation. While all prior works assume the depth priors are up to scale, we show that modeling additional shifts are extremely beneficial in practice, even for off-the-shelf metric depth models.

3. Problem Formulation

Given two images I_1 and I_2 , relative pose estimation aims to estimate the rotation \mathbf{R} and translation \mathbf{t} transforming the coordinate frame of the first camera to the second one. The classic approach to solve this problem is to establish a set of corresponding pixel matches between the two images through either sparse feature matching (*e.g.*, Super-

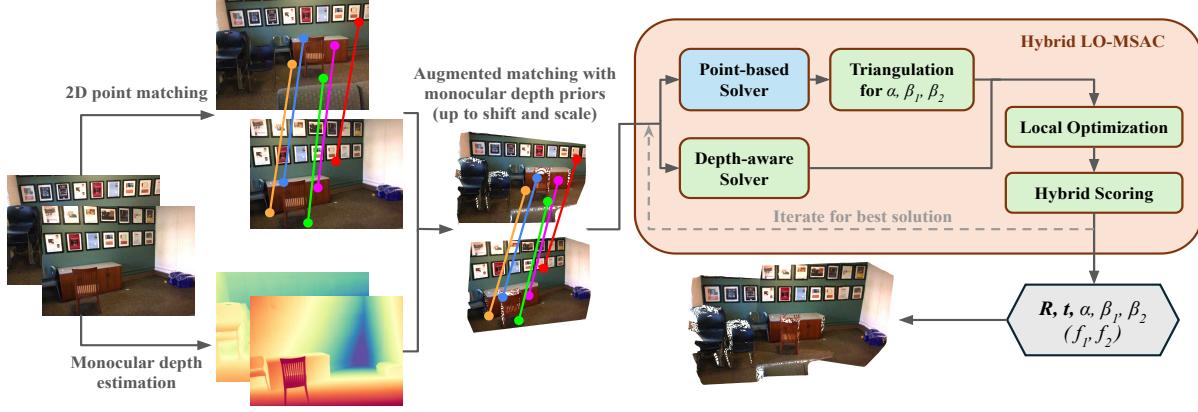


Figure 2. Pipeline overview: Our method takes a pair of images as input, runs off-the-shelf feature matching and monocular depth estimation, then jointly estimates the relative pose, scale and shift parameters of the two depth maps, and optionally the focal lengths. Our method incorporates monocular depth priors in all stages (in green) of hybrid LO-MSAC [12, 35], including 3 new depth-aware solvers, while still being able to leverage traditional point-based solvers [29, 46, 63] (in blue).

Point [16] + LightGlue [40]) or dense correspondence techniques (e.g., RoMa [20]), then solve for the essential matrix relating these point matches.

In addition to the 2D correspondences, we aim to leverage depth maps D_1 and D_2 predicted by some off-the-shelf monocular depth estimation (MDE) model for each image. These depth maps provide valuable 3D priors but come with inherent ambiguities due to the nature of monocular depth estimation. State-of-the-art MDE models [9, 21, 32, 75] are often trained with scale- and shift-invariant loss [51]. They predict either disparity [9, 51, 75] or relative depth [21, 32, 75] that are related to the actual value with an unknown linear (affine) transformation. In our formulation, we focus on affine-invariant relative depth and model this transformation as:

$$\hat{D}_i = a_i D_i + b_i, \quad i \in \{1, 2\} \quad (1)$$

where a_i and b_i are unknown scalar values of the scale and shift of both depth maps.

For a pair of image correspondence $(\mathbf{p}_{1j}, \mathbf{p}_{2j})$, and their corresponding predicted depth values after the unknown linear transformation \hat{d}_{1j} and \hat{d}_{2j} , we can lift the 2D points to 3D from both camera views:

$$\mathbf{P}_{ij} = \mathbf{K}_i^{-1} \begin{bmatrix} \mathbf{p}_{ij} \\ 1 \end{bmatrix} \cdot \hat{d}_{ij}, \quad i \in \{1, 2\} \quad (2)$$

where \mathbf{K}_i is the camera intrinsic matrix of the i -th image.

The two 3D points \mathbf{P}_{1j} and \mathbf{P}_{2j} are related by the rotation \mathbf{R} and translation \mathbf{t} of the relative pose:

$$\mathbf{P}_{2j} = \mathbf{R}\mathbf{P}_{1j} + \mathbf{t}. \quad (3)$$

Thus, the monocular depth maps could provide additional constraints on the relative pose. However, explicitly modeling the scales and shifts for both views is necessary to obtain

such constraint. In relative pose estimation, the actual scale of the scene is inherently ambiguous, so estimating the relative scale ratio $\alpha = a_2/a_1$ is sufficient. Therefore, in addition to the relative rotation \mathbf{R} and translation \mathbf{t} , we also explicitly estimate two scalar shifts $\beta_1 = b_1/a_1$, $\beta_2 = b_2/a_2$, and one scale α , which transforms the two depth maps:

$$\hat{D}_1 = D_1 + \beta_1, \quad \hat{D}_2 = \alpha(D_2 + \beta_2). \quad (4)$$

While this formulation assumes the depth priors are affine-invariant relative depth, using our method with metric depth predictions can also lead to more accurate and robust relative pose as we will show empirically in the experiments. This formulation is also illustrated in Fig. 1.

4. Method

We propose three solvers that solve for the relative pose \mathbf{R}, \mathbf{t} as well as the shifts β_1, β_2 and scale α using the feature correspondences and depth priors:

1. Calibrated solver when all camera intrinsics are known;
2. Shared-focal solver when both images are captured by camera with a shared unknown focal length;
3. Two-focal solver when focal lengths are unknown for both cameras.

The solvers work in a 2-step manner. First, find all possible values for the depth scale and shifts α, β_1, β_2 , as well as the focal lengths if unknown. Then, solve for the rotation and translation for each set of the possible scales and shifts.

4.1. Solving Depth Scale, Shifts, and Focal Lengths

Given a sampled set of M pixel correspondences $\{(\mathbf{p}_{11}, \mathbf{p}_{21}), \dots, (\mathbf{p}_{1M}, \mathbf{p}_{2M})\}$, and the corresponding depth priors from the predicted depth maps $d_{ij} = D_i(\mathbf{p}_{ij})$, the solvers rely on the assumption that the corresponding

pair-wise distances δ among the lifted 3D points should be equal in the two camera frames. Formally, let us define

$$\delta_{jk}^{(i)} = \mathbf{P}_{ij} - \mathbf{P}_{ik}, \text{ where } i \in \{1, 2\} \text{ and } j \neq k.$$

Because the rotation and translation transformation in (3) is length-preserving, the following equality constraint must hold for all pairs (j, k) of the 3D points:

$$\left\| \delta_{jk}^{(1)} \right\|^2 = \left\| \delta_{jk}^{(2)} \right\|^2. \quad (5)$$

We can thus obtain a system of $\binom{M}{2}$ equations where the rotation and translation have been eliminated. The three solvers require different numbers of correspondences to solve these equations due to the different numbers of unknowns. In addition to the three values α, β_1, β_2 , the shared-focal solver also solves for the shared focal f , and the two-focal solver solves for the two focals f_1, f_2 .

Calibrated Solver: The calibrated solver addresses the scenario where both cameras have known intrinsic parameters. In this case the problem is minimal with $M = 3$ points (along with their associated depth priors), as we get $\binom{3}{2} = 3$ equations in (5) in the three unknowns α, β_1 and β_2 .

Since the intrinsics \mathbf{K}_1 and \mathbf{K}_2 are known, let $\tilde{\mathbf{P}}_{ij} = \mathbf{K}_i^{-1} \begin{bmatrix} \mathbf{p}_{ij} \\ 1 \end{bmatrix} = (\tilde{x}_{ij}, \tilde{y}_{ij}, 1)^\top$. We have:

$$\begin{aligned} \left\| \delta_{jk}^{(i)} \right\|^2 &= \left\| \hat{d}_{ij} \tilde{\mathbf{P}}_{ij} - \hat{d}_{ik} \tilde{\mathbf{P}}_{ik} \right\|^2 \\ &= \left\| \hat{d}_{ij}(\tilde{x}_{ij}, \tilde{y}_{ij}, 1)^\top - \hat{d}_{ik}(\tilde{x}_{ik}, \tilde{y}_{ik}, 1)^\top \right\|^2 \\ &= (\hat{d}_{ij}\tilde{x}_{ij} - \hat{d}_{ik}\tilde{x}_{ik})^2 + (\hat{d}_{ij}\tilde{y}_{ij} - \hat{d}_{ik}\tilde{y}_{ik})^2 + (\hat{d}_{ij} - \hat{d}_{ik})^2. \end{aligned}$$

This yields three quartic equations of α, β_1 and β_2 . In the equations, α only appears as squares, so we re-parameterize with $\gamma = \alpha^2$ to make the equations cubic and reduce the number of solutions. Using the method from Larsson *et al.* [34] we find that the system of polynomials have at most 4 solutions and create a Gröbner basis solver. The solver performs linear elimination on a 12x12 matrix followed by a 4x4 eigenvalue problem.

Shared-Focal Solver: The shared-focal solver is designed for scenarios where both images are captured by camera with an unknown shared focal length f . This situation is common in applications such as visual odometry or structure-from-motion with a single moving camera. We assume that the focal length is the only unknown intrinsic parameter, and thus we can assume without loss of generality that $\mathbf{K} = \text{diag}(f, f, 1)$. We then have

$$\tilde{\mathbf{P}}_{ij} = \mathbf{K}^{-1} \begin{bmatrix} \mathbf{p}_{ij} \\ 1 \end{bmatrix} = \left(\frac{x_{ij}}{f}, \frac{y_{ij}}{f}, 1 \right)^\top. \quad (6)$$

As we now have an additional unknown, one more correspondence is needed to solve for α, β_1, β_2 , and f . However,

taking all equations (5) from $M = 4$ matches will give an overdetermined system. We therefore only select 4 out of the 6 possible equations from inserting (6) into (5). Similar to the α reparametrization in the calibrated case, we additionally perform substitution $\omega = 1/f^2$. After these two substitutions we get four equations of degree 4. Again applying the method from Larsson *et al.* [34] we find that the problem has 8 solutions and we generate a solver with template size 36x36.

Two-Focal Solver: Last but not least, the two-focal solver addresses the most general case where the two images are captured by cameras with potentially different, unknown focal lengths. This scenario is particularly relevant for applications involving crowd-sourced imagery.

This case can also be solved with $M = 4$ correspondences, but now requires using 5 out of the 6 equations they generate. The derivation is similar to the shared focal length case, except that we now have independent calibration matrices $\mathbf{K}_1 = \text{diag}(f_1, f_1, 1)$ and $\mathbf{K}_2 = \text{diag}(f_2, f_2, 1)$ for each camera. In the equations we now introduce $\omega_1 = 1/f_1^2$ and $\omega_2 = 1/f_2^2$. Applying the method from Larsson *et al.* [34] we surprisingly find that the problem only has 4 solutions, fewer than in the shared focal length case. The solver however requires linear elimination on a 40x40 matrix followed by solving a 4x4 eigenvalue problem.

4.2. Finding Rotation and Translation

Only real solutions of α, β_1, β_2 that result in positive depth values in (4) are kept as valid solutions (and similarly for the focal lengths). For each possible solution, we estimate a rigid transformation (a rotation \mathbf{R} and a translation \mathbf{t}), which can be easily retrieved using the SVD method [60] aligning the back-projected 3D points from both views which satisfy (3). Since relative poses are inherently up to scale, \mathbf{R} and \mathbf{t} are the final outputs of our proposed solvers.

4.3. Evaluating Solutions

The solutions $\Theta = (\mathbf{R}, \mathbf{t}, \alpha, \beta_1, \beta_2)$ (optionally also include f or f_1, f_2) need to be evaluated on all correspondences with their depth priors. To do so, we compute the *depth-induced reprojection errors*:

$$\begin{aligned} E_{r(1 \rightarrow 2)}(\Theta) &= \left\| \mathbf{\Pi}_2(\mathbf{R}\mathbf{P}_1 + \mathbf{t}) - \mathbf{p}_2 \right\|^2 \\ E_{r(2 \rightarrow 1)}(\Theta) &= \left\| \mathbf{\Pi}_1(\mathbf{R}^{-1}\mathbf{P}_2 - R^{-1}\mathbf{t}) - \mathbf{p}_1 \right\|^2, \end{aligned} \quad (7)$$

where $\mathbf{P}_1, \mathbf{P}_2$ are the lifted 3D points of a pair of correspondences using the shift-and-scaled depth priors $\hat{d}_1 = d_1 + \beta_1$, $\hat{d}_2 = \alpha(d_2 + \beta_2)$, and $\mathbf{\Pi}_i(\mathbf{P})$ is the camera projection using \mathbf{K}_i . \mathbf{K}_i is either known with calibrated cameras, or assembled using the estimated focal lengths solved by solvers. Note that the errors are computed separately for

two directions. This is designed to be robust against inconsistencies between the depth maps when combining into a MSAC score with a squared inlier threshold τ_r :

$$\bar{E}_r(\Theta) = \min(E_{r(1 \rightarrow 2)}, \tau_r) + \min(E_{r(2 \rightarrow 1)}, \tau_r). \quad (8)$$

4.4. Hybrid Estimation

While monocular depth estimation models can provide useful geometric cues, solely relying on the depth priors can be prone to erroneous results when the priors are unreliable. We further propose to jointly use our depth-aware solvers and reprojection errors together with the classic point-based solvers and epipolar (Sampson) error E_s [45, 53] in a hybrid LO-MSAC [12, 35] framework. For the calibrated case, the Sampson error is rescaled with the focal lengths to ensure both errors are in pixels.

This approach enhances the estimation with good depth priors but can also utilize 2D correspondences when the priors are unsatisfactory. As shown in Fig. 2, each iteration of the hybrid LO-MSAC pipeline starts by selecting a solver between a depth-aware and a point-based solver depending on the setting:

- **Calibrated:** calibrated solver & 5-pt essential matrix solver [46];
- **Shared-focal:** shared-focal solver & 6-pt shared focal relative pose solver [63];
- **Two-focal:** two-focal solver & 7-pt fundamental matrix solver [29].

For point-based solvers, we also need to estimate α, β_1, β_2 consistent with the estimated pose and potentially focal lengths. This is done by triangulating 3D points from the M correspondences, and fitting the depth priors to the projected depth to find the scale and shift in least-squares manner. For the 7-pt fundamental matrix solver, the possible focal lengths are retrieved using Bougnoux formula[11].

The two solvers are initially randomly selected with equal probabilities, and are later chosen using the inlier ratios of the data types following [12]. Specifically, we view each correspondences as three different data types: $(\mathbf{p}_1, \mathbf{p}_2, d_1)$, $(\mathbf{p}_1, \mathbf{p}_2, d_2)$, and $(\mathbf{p}_1, \mathbf{p}_2)$. The first two are evaluated using $E_{r(1 \rightarrow 2)}$ and $E_{r(2 \rightarrow 1)}$ respectively, and the third using E_s . Inliers are defined as correspondences (in each data type) with error within the squared thresholds: $E_r < \tau_r$ or $E_s < \tau_s$. The local optimization runs on the inliers with least square optimizer that jointly optimizes all parameters of Θ to minimize the cost function

$$E(\Theta) = \sum_{i \in I_1} E_{r(1 \rightarrow 2)} + \sum_{i \in I_2} E_{r(2 \rightarrow 1)} + 2\lambda_s \frac{\tau_r}{\tau_s} \sum_{i \in I_3} E_s. \quad (9)$$

with I_1, I_2, I_3 being the inlier sets of the three data types, and λ_s is the Sampson error weight (tunable, empirically set to 1 in our experiments).

With $\bar{E}_s = \min(E_s, \tau_s)$, the combined hybrid MSAC score of all correspondences is then

$$\bar{E}(\Theta; \mathbf{p}_1, \mathbf{p}_2) = \bar{E}_r + 2\lambda_s \frac{\tau_r}{\tau_s} \bar{E}_s, \quad (10)$$

which is used to select the best solution with minimal \bar{E} summed over all correspondences.

5. Experiments

5.1. Experimental Setup

Datasets: We evaluate our method across multiple datasets: ScanNet-1500 [15] and MegaDepth-1500 [38] are used to evaluate relative pose estimation performance with calibrated cameras on indoor and outdoor images. ScanNet-1500 is also used for the unknown shared-focal case, and MegaDepth-1500 for the two unknown focal case. To also evaluate the relative pose estimation with unknown, different focal lengths on indoor images, we sample 1064 image pairs from the Stanford 2D-3D-S panoramic dataset [2] with random focal lengths using the procedure detailed in Appendix A. In addition, we also sampled 1451 image pairs from the 7 indoor scenes¹ of ETH3D [59] dataset. These include all covisible image pairs with a minimum of 50 covisible ground-truth (GT) 3D points. Having these GT 3D points allows us to analyze the performance of our framework especially in hard cases where the image pair has a low covisibility (Fig. 3).

Baselines: We compare against the three classic point-based methods correspond to the three settings. We use the PoseLib [33] RANSAC implementation of each point-based baseline. We also compare with previous methods in their corresponding settings: the Perspective-n-Points (PnP) method [24] used in [41] for the calibrated setting; the 3p3d solver [17] for the unknown shared-focal setting; and the 4p4d solver [17] for the two-focal setting. The two solvers from [17] are run within GC-RANSAC [3] as in the original paper. We also compare with two recent two-view dense reconstruction methods: DUST3R [70] and MAST3R [37], in the shared-focal and two-focal settings.

Feature Matchers and MDE Models: Our framework can work with off-the-shelf feature matchers and MDE models that produce affine-invariant depth priors. We test our method with sparse matchings like Superpoint [16]+SuperGlue [54] / LightGlue [40], as well as the state-of-the-art (SOTA) dense matcher RoMa [20]. Several MDE models predicting affine-invariant depths are used: Omnidata [31], Marigold [32], the metric depth model of Depth-Anything v1 [74] and v2 [75], and the recent MoGe

¹ETH3D indoor scenes: delivery-area, kicker, office, pipes, relief, relief-2, terrains

Matches	Method	MD Model	Med. Err. ↓		Pose Error AUC (%) ↑		
			ε_R (°)	ε_t (°)	@5°	@10°	@20°
SP+SG	PoseLib-5pt	-	2.08	5.44	19.48	38.09	56.08
	PoseLib-PnP	DA-met.	2.10	6.25	15.41	34.94	56.13
	PoseLib-PnP	MoGe	<u>1.72</u>	4.97	20.71	42.07	61.89
	Ours-calib	DA-met.	1.74	4.73	23.12	43.37	62.26
	Ours-calib	MoGe	1.68	<u>4.53</u>	<u>23.83</u>	44.10	63.09
SP+LG	PoseLib-5pt	-	1.86	5.53	21.55	39.11	55.60
	PoseLib-PnP	DA-met.	2.03	6.44	15.05	34.16	53.97
	PoseLib-PnP	MoGe	1.72	5.37	19.69	40.21	58.73
	Ours-calib	DA-met.	<u>1.66</u>	<u>4.94</u>	<u>22.74</u>	42.43	<u>59.87</u>
	Ours-calib	MoGe	1.60	<u>4.82</u>	<u>23.22</u>	43.17	60.88
RoMa	PoseLib-5pt	-	1.29	3.18	33.40	55.34	72.41
	PoseLib-PnP	DA-met.	1.63	4.46	22.83	47.04	68.12
	PoseLib-PnP	MoGe	1.38	3.75	28.38	52.62	71.61
	Ours-calib	DA-met.	1.26	3.14	34.21	<u>56.52</u>	<u>73.47</u>
	Ours-calib	MoGe	1.25	<u>3.12</u>	<u>34.05</u>	56.59	73.59

Table 1. Results of relative pose estimation with known camera intrinsics on ScanNet-1500 [15]. The reported metrics are median rotation and translation direction errors ε_R , ε_t , as well as pose error AUCs under 5°, 10°, and 20° thresholds. Best results with each type of matches are **bolded** and second best underlined.

Matches	Method	MD Model	Med. Err. ↓		Pose Error AUC (%) ↑		
			ε_R (°)	ε_t (°)	@5°	@10°	@20°
SP+LG	PoseLib-5pt	-	<u>0.46</u>	<u>1.23</u>	<u>60.16</u>	<u>74.25</u>	84.34
	PoseLib-PnP	DA-met.	0.99	3.29	34.79	54.55	72.32
	PoseLib-PnP	MoGe	0.56	1.80	52.46	71.05	83.80
	Ours-calib	DA-met.	0.51	1.37	57.93	73.81	<u>84.84</u>
	Ours-calib	MoGe	0.42	1.16	63.48	77.44	86.85

Table 2. Relative pose estimation with known intrinsics on MegaDepth-1500 [38].

[69]. Due to limited space, we report results with best performing models in the main paper and put the others in Appendix B. Note that while the Depth-Anything metric models are trained to predict metric depth, Tab. 6 show that modeling the scale and shift is still beneficial.

Implementation Details: We implement our solvers and hybrid estimation pipeline in C++. Our hybrid LO-MSAC is based on the hybrid-RANSAC implementation from RansacLib [55], with non-minimal solver and least square optimization implemented using automatic differentiation from Ceres solver [1]. We further export Python bindings using pybind11 [30] to conveniently use our framework in Python scripts. For all experiments, ScanNet images are resized to 640x480, ETH3D images to 720x480, and other images are used with their original resolution.

5.2. Results and Analysis

Calibrated Cameras: Tab. 1 shows the median rotation error ε_R and translation error ε_t , as well as the Area Under the Curve (AUC) of the pose error (max of ε_R and ε_t) for various thresholds on ScanNet-1500 [15]. With sparse feature matching (SP+SG or SP+LG) as well as dense matching (RoMa), our method consistently improves over the baselines across all metrics using priors from both the Depth-Anything-v1 [74] metric model (DA-met.) and the non-

Matches	Method	MD Model	Med. Err. ↓		Pose Err. AUC (%) ↑		
			ε_R (°)	ε_t (°)	ε_f (%)	@5°	@10°
SP+SG	PoseLib-6pt	-	-	-	3.01	8.12	11.47
	3p3d [17]	DAv2	-	-	3.94	9.87	15.56
	Ours-sf	DA-met.	2.00	5.70	6.84	18.35	37.54
SP+LG	PoseLib-6pt	-	-	-	2.71	8.05	10.94
	3p3d [17]	DAv2	-	-	3.27	9.37	13.66
	Ours-sf	DA-met.	1.93	5.78	6.99	17.74	36.71
RoMa	PoseLib-6pt	-	-	-	1.52	3.85	4.49
	3p3d [17]	DAv2	-	-	1.64	4.15	5.30
	Ours-sf	DA-met.	1.37	3.54	3.71	29.81	53.11
MAS3R	PoseLib-6pt	-	-	-	1.44	3.31	4.18
	3p3d [17]	DAv2	-	-	1.51	3.40	4.54
	Ours-sf	DA-met.	<u>1.35</u>	<u>3.09</u>	3.73	<u>31.87</u>	<u>56.20</u>
Reference entry - DUS3R	MAS3R	MAS3R	1.32	3.06	3.79	32.58	56.99
	Reference entry - MAST3R	-	-	-	-	-	-

Table 3. Uncalibrated shared-focal relative pose estimation results on ScanNet-1500 [15]. In addition to ε_R and ε_t and pose error AUCs, the median relative focal error ε_f is also reported.

metric affine-invariant depth model MoGe [69]. Compared to the PnP baseline, even when the MDE model is trained to predict metric depth, modeling the depth priors as affine-invariant in our method is still beneficial.

We also evaluate on the outdoor image pairs of MegaDepth-1500 [38]. Due to the larger and more varied scene scales, MDE models often perform poorly on outdoor data compared to indoor. However, as the results in Tab. 2 show, our method can still improve over the baselines on outdoor images with the latest MoGe [69] model. While DA-met. depth estimations becomes less reliable outdoor, our method still outperforms RANSAC PnP using the same depth priors and is reasonably close to the point-based baseline thanks to the hybrid estimation.

Shared-focal Cameras: For the shared-focal case, we again use the ScanNet-1500 image pairs but ignore the GT intrinsics. We additionally report the median ε_f of relative focal errors which are computed as $|f - f_{gt}| / f_{gt}$. Tab. 3 shows the results. Note that the 3p3d solver [17] takes disparity values as input, so we use the disparity priors predicted by Depth-Anything-v2 [75] (as in [17]). Our results are reported with the best performing DA-met. model. MoGe is not used for shared-focal or uncalibrated case because it also estimates focal lengths and can therefore benefit directly from using calibrated hybrid estimation.

With all tested matches, our method improves over the baselines consistently and significantly. 3p3d solver only models the depth priors as scale-invariant but does not consider the shift, so it is sensitive to noise in the priors [17] and therefore fails to improve over point-based baseline. The recent works DUS3R [70] and MAS3R [37] produce good image matching, two-view depth maps, and relative pose by directly regressing aligned point maps from two views, so we report their relative pose estimation results

Matches	Method	MD Model	Med. Err. ↓			Pose Err. AUC (%) ↑			
			ε_R (°)	ε_t (°)	ε_f (%)	@2°	@5°	@10°	@20°
SP+LG	PoseLib-7pt	-	16.83	21.57	59.88	5.85	13.95	21.94	30.71
	4p4d [17]	DAv2	19.27	21.64	66.57	5.31	13.66	21.78	30.22
	Ours-tf	DAv2-met.	5.66	9.26	29.06	9.15	22.22	32.80	43.26
RoMa	PoseLib-7pt	-	7.99	8.77	37.72	8.73	20.31	30.45	41.48
	4p4d [17]	DAv2	10.94	11.35	44.94	7.23	16.50	25.31	36.26
	Ours-tf	DAv2-met.	3.33	4.32	17.29	13.50	29.19	42.18	54.42
MASt3R	PoseLib-7pt	-	3.21	3.34	19.95	12.58	30.27	45.57	59.85
	4p4d [17]	DAv2	3.21	3.28	20.07	12.13	29.82	45.30	60.05
	Ours-tf	DAv2-met.	1.98	2.22	11.27	18.05	39.92	56.64	70.86
Reference entry - DUS3R			2.65	4.65	7.41	6.43	24.47	42.39	58.36
Reference entry - MASt3R			1.71	2.49	2.91	13.39	38.41	57.92	71.91

Table 4. Uncalibrated relative pose estimation results on our generated image pairs with different focal lengths from Stanford 2D-3D-S [2] dataset. The median focal error ε_f is the maximum of the two relative focal errors per image pair.

Method	MD Model	Med. Err. ↓			Pose Err. AUC (%) ↑		
		ε_R (°)	ε_t (°)	ε_f (%)	@5°	@10°	@20°
PoseLib-7pt	-	1.97	5.72	23.64	21.23	36.80	54.89
4p4d [17]	DAv2	3.48	8.70	39.92	14.04	26.70	44.08
Ours-tf	DA-met.	1.25	4.81	15.99	25.70	42.90	61.79

Table 5. Uncalibrated relative pose estimation on MegaDepth-1500 [38].

for reference and also evaluate our method with MASt3R matches and depth. By using the shared-focal constraint which they do not model, our method performs better than these two-view methods with RoMa matches and DA-met. depth priors, and can be further improved with better matches and depth from MASt3R.

Unknown focal cameras: For the most general case with two uncalibrated cameras, we define the focal error as the max of the two relative focal errors of the two cameras and again report the median ε_f . Significant improvements can be observed in Tab. 4 on the uncalibrated image pairs from our sampled 2D-3D-S dataset. Our method achieves less than 50% in the median errors of the point-based baseline and significant improvements in pose error AUCs.

Similar to the 3p3d solver, the 4p4d solver [17] is tested with Depth Anything v2 [75] depth priors and again fails to improve over point-based baseline. Using better matches from MASt3R [37], our method is able to achieve on-par results using monocular depth priors (Depth-Anything-v2 metric model), with significantly better accuracy on small threshold AUCs. Our method further outperforms MASt3R if using the depth from MASt3R as priors.

We also evaluate with the uncalibrated camera setting on outdoor MegaDepth-1500 [38] images (Tab. 5). Compared with the calibrated results in Tab. 2, our method (with DA-met. model) now shows significant improvements over the point-based fundamental matrix estimator.

Limited Covisibility: We further analyze the pose errors

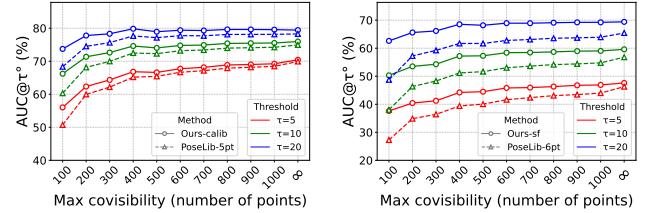


Figure 3. Pose error AUCs on sampled indoor ETH3D [59] image pairs with covisible GT points less than the thresholds on X-axis. Left: calibrated, SP+LG, and MoGe [69] priors; Right: shared-focal, SP+LG, and DAv2-met.[75] metric priors.



Figure 4. Visualization on ETH3D [59]. **Left:** back-projected GT depth with relative pose found by point-based method and translation rescaled to match GT translation; **Middle:** back-projected monocular 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 relative pose.

under different levels of covisibility between image pairs on the indoor ETH3D [59] pairs and the results are shown in Fig. 3. Comparing the gap between our method and point-based baselines, incorporating the depth priors can particularly improve the relative pose estimation results on difficult image pairs with limited covisibility.

Visual examples: In Fig. 4 we show visualizations of point clouds reconstructed with relative pose estimated by point-based baseline and ground-truth (GT) depth, and results from our method with scale-and-shifted monocular depth priors. The visualizations demonstrate that by incorporating depth priors and by explicit modeling of the affine transformation, our method is able to find a better pose or to avoid failure cases with better depth alignment.

Runtime: When running with SP+LG matches, 4 LO steps and at least 1000 iterations in the hybrid LO-MSAC (a good trade-off between performance and efficiency), the typical (median) solution time for an image pair of our hybrid estimation pipeline are: 0.36 seconds for calibrated case, 0.61 seconds for shared-focal case on ScanNet-1500; and 0.95 seconds for uncalibrated case on 2D-3D-S image pairs. The numbers are measured when running single-threaded on Intel Core i7-10700K CPU. The speed can be improved by implementing analytical jacobians to replace Ceres automatic differentiation, or with fewer LO steps at the cost of slightly decreased accuracy. When running on a RTX 3080

Dataset	MD Model	Hybrid	Shift	Med. Err. ↓		Pose Error AUC (%) ↑		
				$\varepsilon_R(\text{°})$	$\varepsilon_t(\text{°})$	@5°	@10°	@20°
ScanNet-1500	DA-met.	✓	✓	1.74	4.73	23.12	43.37	62.26
		✓		1.80	5.09	19.96	41.45	61.58
			✓	1.88	5.33	19.13	40.21	60.65
	Marigold			2.00	5.85	16.63	37.24	59.37
		✓	✓	1.87	5.05	20.68	40.71	59.92
		✓		2.85	9.17	12.34	26.64	44.53
MegaDepth-1500	MoGe	✓		2.19	6.41	14.70	33.73	54.75
			✓	6.24	20.44	1.09	4.84	17.15
		✓	✓	0.42	1.16	63.48	77.44	86.85
				0.63	1.90	51.32	70.09	83.02
			✓	0.55	1.56	56.87	73.80	85.11
				0.74	2.45	44.15	64.90	79.84

Table 6. Ablation of shift modeling with calibrated cameras. All results are with SP+LG correspondences.

GPU, the typical runtime of matchers and MDE models are: SP+LG 16ms/pair; DAv2-met. (ViT-L) 0.16s/image; MoGe (ViT-L) 3ms/image. All models are run with default settings. While the added considerations for depth priors increase the computation complexity, our method is still able to run efficiently.

5.3. Ablation Studies

Modeling scale and shift: To further demonstrate the effectiveness of our affine correction modeling, we compare with an ablated scale-only baseline that removes the shift modeling in the calibrated solver and scoring. A visual example is shown in teaser Fig. 1. Without the unknown β_1, β_2 , the scale and pose can be together solved by finding the rigid transformation using the 3 point matches and depth priors. For scoring, the reprojection error also removes β_1, β_2 and only lift the 2D points with α .

Tab. 6 shows the comparison with the ablated baselines. The non-hybrid version uses only the depth-aware calibrated solver, and only reprojection errors in LO and scoring. The results show that the shift modeling especially makes a difference for non-metric monocular depth models like Marigold [32] and for outdoor images due to the larger and varied scale. With depth priors targeting metric depth, modeling the shift is again proven beneficial.

We further analyze when modeling the shift starts making a difference with a synthetic experiment. We add a shift value ranging from 0 to the median of GT depth for each image and use this shifted GT depth as “depth priors”. We compare the median rotation and translation errors on ScanNet-1500 [15] with SP+SG matches. As shown in Fig. 5, modeling the shift is beneficial as long as the magnitude of β is greater than 10% of the median depth priors, and our method is not influenced by the increased shift magnitude while the errors of the PnP and scale-only baselines increase. In Fig. 6 we show the computed shift values by fitting depth priors to GT depth with metric [74] and non-metric [32] models on indoor [15] and outdoor [38] images. Shift values larger than 10% of the median of the priors are

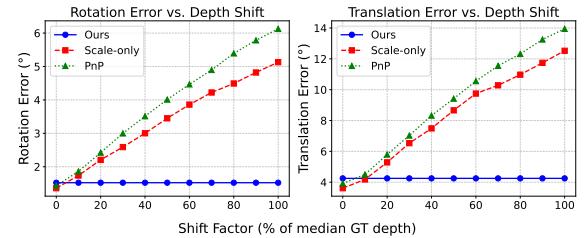


Figure 5. Rotation and translation error by adding shift values to GT depth as “depth priors” on ScanNet-1500 [15].

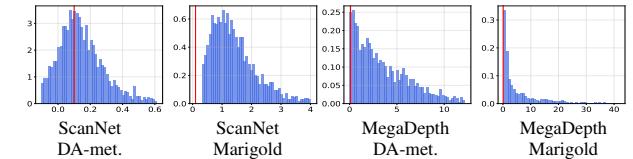


Figure 6. Distribution (probabilistic density) of GT shift values (β_1, β_2) after fitting MDE depth priors to the GT depth. Red lines indicate $\beta = 10\% = 0.1$ (refer to Fig. 5). Even with metric depth priors on indoor data, modeling the shift is crucial.

Solver	LO	Score	Med. Err. ↓		Pose Error AUC (%) ↑		
			$\varepsilon_R(\text{°})$	$\varepsilon_t(\text{°})$	$\varepsilon_f(\%)$	@5°	@10°
H	H	H	2.00	5.70	6.84	18.35	37.54
D	H	H	2.20	5.89	7.42	17.17	36.17
P	H	H	2.20	6.10	7.36	16.60	35.33
D	D	H	2.22	6.33	7.95	15.42	34.39
P	P	H	2.42	6.68	9.09	15.06	32.32
D	D	D	2.21	6.83	8.18	13.85	32.11
P	P	P	2.84	7.45	11.30	14.06	29.86
							48.13

Table 7. Ablation of different components of hybrid LO-MSAC on ScanNet-1500 [15] with shared-focal setting, SP+SG matches and DA-met. depth priors. H - Hybrid; D - Depth solver and reprojection score; P - Point-based solver and Sampson score.

common even with metric depth models on indoor data, justifying the gains of our method.

Ablation of hybrid estimation: Tab. 7 shows ablation on different components of our hybrid estimation. While the hybrid scoring brings the largest improvements, all components contribute to the success of our hybrid estimation.

Applicability to different MDE models: Our method can 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 (refer to Appendix B).

6. Conclusion

In this paper, we present a relative pose estimator that benefits from explicit modeling of affine variants of monocular depth priors. We propose solvers for calibrated and uncalibrated camera setups, and combine them with classic methods to achieve consistent improvement across multiple datasets. The proposed method can further benefit from latest advances on image matching and monocular depth estimation.

Acknowledgements. We sincerely thank Philipp Lindenberger for insightful ideas, as well as Lei Li, Chong Bao, Yiming Zhao, and Zihan Zhu for their support and helpful discussions. Viktor Larsson was supported by ELLIIT and the Swedish Research Council (Grant No. 2023-05424).

Appendices

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 ($\pm 30^\circ$), pitch ($\pm 30^\circ$), yaw ($\pm 180^\circ$), and FOV ($60^\circ - 105^\circ$, 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 ($\pm 30^\circ$), a random FOV ($60^\circ - 105^\circ$), 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 uniformly.

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

Task	Method / MD Model	Med. Err. ↓			Pose Err. AUC (%) ↑		
		ε_R (°)	ε_t (°)	ε_f (%)	@5°	@10°	@20°
ScanNet-1500 Calibrated	PoseLib-5pt	2.08	5.44	-	19.48	38.09	56.08
	Omnidata	<u>1.76</u>	5.42	-	21.24	39.74	57.75
	Marigold	1.72	5.28	-	21.18	40.38	58.13
	DA-met.	<u>1.66</u>	4.94	-	<u>22.74</u>	<u>42.43</u>	<u>59.87</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	1.60	4.82	-	23.22	43.17	60.88
ETH3D Shared-focal	GT Depth	1.54	4.57	-	25.03	45.06	61.85
	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
MegaDepth-1500 Two-focal	GT Depth	0.36	0.81	2.07	62.26	70.52	75.90
	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	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 underlined.

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.

B. Additional Experiment Results

Results with different monocular depth models: As mentioned in Sec. 5 in the main paper, 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 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 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 \mathbf{R}, \mathbf{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 uniformly 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.

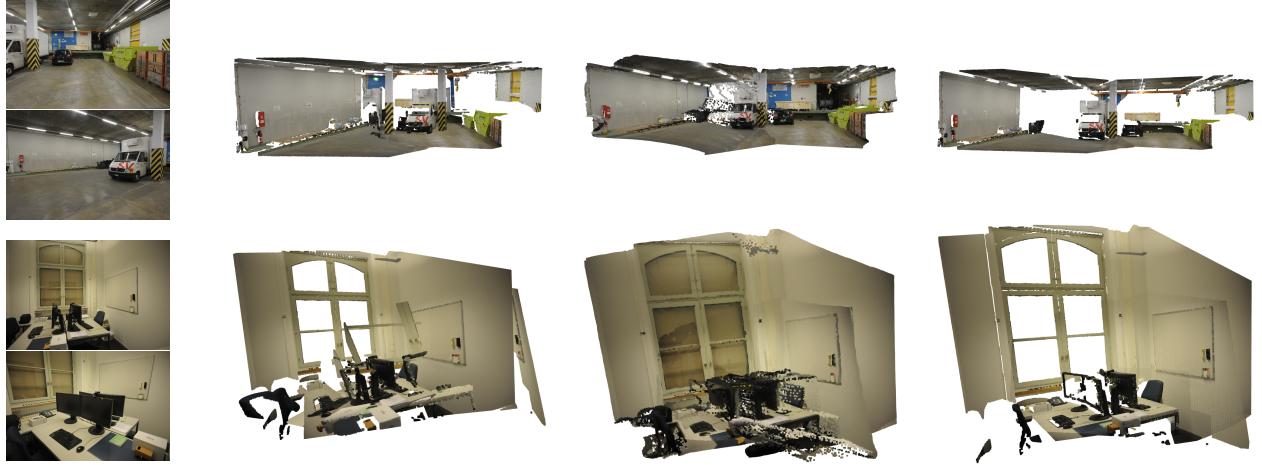


Figure 7. Additional visualizations on ETH3D [59] with shared-focal setting. **Left:** back-projected GT depth with relative pose found by PoseLib-6pt [33] and translation rescaled to match scale with GT translation; **Middle:** back-projected depth priors from Marigold [32] aligned using the output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

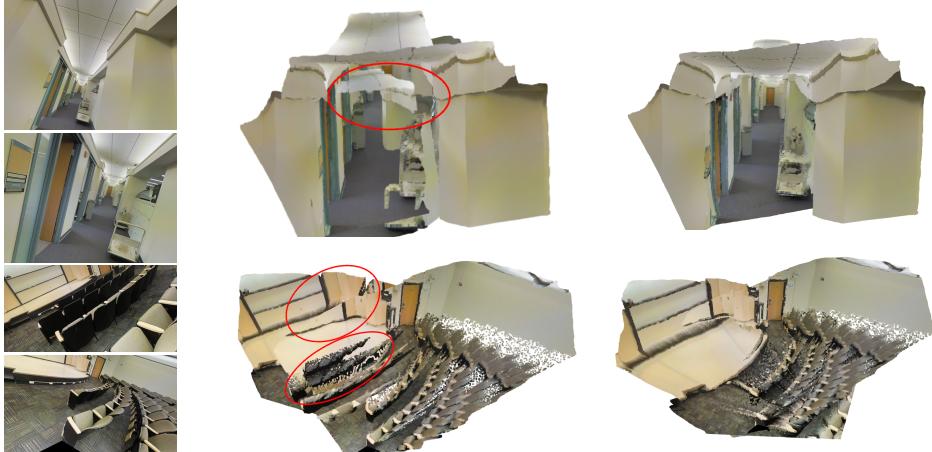


Figure 8. Additional visualizations on Stanford 2D-3D-S [2] image pairs with two-focal setting. **Left:** back-projected depth priors aligned using the relative pose found by PoseLib-7pt [33] baseline; **Right:** back-projected depth priors aligned using the relative pose found by our two-focal estimator. 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.

References

- [1] Sameer Agarwal, Keir Mierle, and The Ceres Solver Team. Ceres Solver, 2023. 6
- [2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv:1702.01105*, 2017. 5, 7, 9, 10, 11
- [3] Daniel Barath and Jiri Matas. Graph-cut RANSAC. In *CVPR*, 2018. 5, 9
- [4] Daniel Barath and Chris Sweeney. Relative pose solvers using monocular depth. In *ICPR*, 2022. 1, 2
- [5] Daniel Barath, Dmytro Mishkin, Luca Cavalli, Paul-Edouard Sarlin, Petr Hruba, and Marc Pollefeys. Stereoglu: Robust estimation with single-point solvers. In *ECCV*, 2024. 2
- [6] Axel Barroso-Laguna, Eric Brachmann, Victor Adrian Prisacariu, Gabriel J Brostow, and Daniyar Turmukhambov. Two-view geometry scoring without correspondences. In *CVPR*, 2023. 2
- [7] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, 2021. 2
- [8] Shariq Farooq Bhat, Reiner Birk, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv:2302.12288*, 2023. 1, 2
- [9] Reiner Birk, Diana Wofk, and Matthias Müller. Midas v3.1 – a model zoo for robust monocular relative depth estimation. *arXiv:2307.14460*, 2023. 1, 3
- [10] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv:2410.02073*, 2024. 2
- [11] Sylvain Bougnoux. From projective to euclidean space under any practical situation, a criticism of self-calibration. In *ICCV*, 1998. 5
- [12] Federico Camposeco, Andrea Cohen, Marc Pollefeys, and Torsten Sattler. Hybrid Camera Pose Estimation. In *CVPR*, 2018. 3, 5
- [13] Ondřej Chum, Jiří Matas, and Josef Kittler. Locally optimized ransac. In *Pattern Recognition: 25th DAGM Symposium, Magdeburg, Germany, September 10-12, 2003. Proceedings 25*, 2003. 2
- [14] Jan Czarnowski, Tristan Laidlow, Ronald Clark, and Andrew J Davison. Deepfactors: Real-time probabilistic dense monocular slam. *IEEE Robotics and Automation Letters*, 5 (2):721–728, 2020. 1, 2
- [15] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 5, 6, 8, 9, 10, 12
- [16] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *CVPR Deep Learning for Visual SLAM Workshop*, 2018. 2, 3, 5
- [17] Yaqing Ding, Václav Vávra, Snehal Bhayani, Qianliang Wu, Jian Yang, and Zuzana Kukelova. Fundamental matrix estimation using relative depths. In *ECCV*, 2024. 1, 2, 5, 6, 7, 9
- [18] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-net: A trainable cnn for joint detection and description of local features. *arXiv:1905.03561*, 2019. 2
- [19] Johan Edstedt, Ioannis Athanasiadis, Mårten Wadenbäck, and Michael Felsberg. Dkm: Dense kernelized feature matching for geometry estimation. In *CVPR*, 2023. 2
- [20] Johan Edstedt, Qiyu Sun, Georg Bökman, Mårten Wadenbäck, and Michael Felsberg. Roma: Robust dense feature matching. In *CVPR*, 2024. 2, 3, 5
- [21] Ainaz Eftekhari, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *CVPR*, 2021. 3
- [22] Ivan Eichhardt and Daniel Barath. Relative pose from deep learned depth and a single affine correspondence. In *ECCV*, 2020. 2
- [23] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *NeurIPS*, 2014. 2
- [24] Martin A. Fischler and Robert C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 1981. 5
- [25] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, 2018. 2
- [26] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *ECCV*, 2024. 2
- [27] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *CVPR*, 2019. 2
- [28] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2020. 1
- [29] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004. 2, 3, 5
- [30] Wenzel Jakob, Jason Rhinelander, and Dean Moldovan. pybind11 – seamless operability between c++11 and python, 2017. <https://github.com/pybind/pybind11>. 6
- [31] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *CVPR*, 2022. 1, 5, 9
- [32] Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *CVPR*, 2024. 1, 2, 3, 5, 7, 8, 9, 11
- [33] Viktor Larsson and contributors. PoseLib - Minimal Solvers for Camera Pose Estimation, 2020. 2, 5, 9, 11
- [34] Viktor Larsson, Kalle Astrom, and Magnus Oskarsson. Efficient solvers for minimal problems by syzygy-based reduction. In *CVPR*, 2017. 4, 10
- [35] Karel Lebeda, Jiri Matas, and Ondrej Chum. Fixing the Locally Optimized RANSAC. In *BMVC*, 2012. 3, 5

- [36] Karel Lebeda, Jiri Matas, and Ondrej Chum. Fixing the locally optimized ransac–full experimental evaluation. In *BMVC*, 2012. 2
- [37] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv:2406.09756*, 2024. 2, 5, 6, 7
- [38] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 2, 5, 6, 7, 8, 9
- [39] Zhenyu Li, Xuyang Wang, Xianming Liu, and Junjun Jiang. Binsformer: Revisiting adaptive bins for monocular depth estimation. *IEEE Transactions on Image Processing*, 2024. 2
- [40] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. LightGlue: Local Feature Matching at Light Speed. In *ICCV*, 2023. 2, 3, 5
- [41] Sheng Liu, Xiaohan Nie, and Raffay Hamid. Depth-guided sparse structure-from-motion for movies and tv shows. In *CVPR*, 2022. 1, 2, 5
- [42] Shing Yan Loo, Syamsiah Mashohor, Sai Hong Tang, and Hong Zhang. Deeprelativefusion: Dense monocular slam using single-image relative depth prediction. In *IROS*, 2021. 2
- [43] David G Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60:91–110, 2004. 2
- [44] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. *ACM TOG*, 39(4):71–1, 2020. 2
- [45] Quang-Tuan Luong and Olivier D. Faugeras. The fundamental matrix: Theory, algorithms, and stability analysis. *IJCV*, 1996. 5
- [46] David Nistér. An efficient solution to the five-point relative pose problem. In *CVPR*, 2003. 2, 3, 5
- [47] Suraj Patni, Aradhya Agarwal, and Chetan Arora. Ecodepth: Effective conditioning of diffusion models for monocular depth estimation. In *CVPR*, 2024. 2
- [48] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 1, 2
- [49] Matia Pizzoli, Christian Forster, and Davide Scaramuzza. Remode: Probabilistic, monocular dense reconstruction in real time. In *ICRA*, 2014. 2
- [50] René Ranftl and Vladlen Koltun. Deep fundamental matrix estimation. In *ECCV*, 2018. 2
- [51] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 44(3), 2022. 2, 3
- [52] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, 2016. 2
- [53] Paul D. Sampson. Fitting conic sections to "very scattered" data: An iterative refinement of the bookstein algorithm. *Comput. Graph. Image Process.*, 1982. 2, 5
- [54] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 2, 5
- [55] Torsten Sattler et al. RansacLib - A Template-based *SAC Implementation, 2019. 6
- [56] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. In *NeurIPS*, 2005. 2
- [57] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE TPAMI*, 31(5):824–840, 2008. 2
- [58] Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. The surprising effectiveness of diffusion models for optical flow and monocular depth estimation. In *NeurIPS*, 2024. 2
- [59] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *CVPR*, 2017. 5, 7, 9, 10, 11
- [60] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. In *Psychometrika*, 1966. 4
- [61] Mengyi Shan, Brian Curless, Ira Kemelmacher-Shlizerman, and Steve Seitz. Animating street view. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–12, 2023. 1
- [62] Jiuhan Song, Seonghoon Park, Honggyu An, Seokju Cho, Min-Seop Kwak, Sungjin Cho, and Seungryong Kim. Därf: boosting radiance fields from sparse inputs with monocular depth adaptation. In *NeurIPS*, 2023. 2
- [63] Henrik Stewénius, David Nistér, Fredrik Kahl, and Frederik Schaffalitzky. A minimal solution for relative pose with unknown focal length. *Image and Vision Computing*, 2008. 2, 3, 5
- [64] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 2
- [65] Keisuke Tateno, Federico Tombari, Iro Laina, and Nassir Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *CVPR*, pages 6243–6252, 2017. 2
- [66] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. *IEEE TPAMI*, 45(8):10247–10266, 2023. 2
- [67] Michał Tyszkiewicz, Pascal Fua, and Eduard Trulls. Disk: Learning local features with policy gradient. In *NeurIPS*, 2020. 2
- [68] Qing Wang, Jiaming Zhang, Kailun Yang, Kunyu Peng, and Rainer Stiefelhagen. Matchformer: Interleaving attention in transformers for feature matching. In *ACCV*, 2022. 2
- [69] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv:2410.19115*, 2024. 1, 2, 6, 7, 8, 9
- [70] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *CVPR*, 2024. 2, 5, 6
- [71] Tong Wei, Yash Patel, Alexander Shekhovtsov, Jiri Matas, and Daniel Barath. Generalized differentiable ransac. In *ICCV*, 2023. 2

- [72] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. [2](#)
- [73] Haofei Xu, Songyou Peng, Fangjinhua Wang, Hermann Blum, Daniel Barath, Andreas Geiger, and Marc Pollefeys. Depthsplat: Connecting gaussian splatting and depth. *arXiv:2410.13862*, 2024. [2](#)
- [74] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. [1](#), [2](#), [5](#), [6](#), [8](#), [9](#)
- [75] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv:2406.09414*, 2024. [1](#), [2](#), [3](#), [5](#), [6](#), [7](#), [8](#), [9](#)
- [76] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *ICCV*, 2023. [1](#), [2](#)
- [77] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In *NeurIPS*, 2022. [2](#)
- [78] Ziyao Zeng, Daniel Wang, Fengyu Yang, Hyoungseob Park, Stefano Soatto, Dong Lao, and Alex Wong. Wordepth: Variational language prior for monocular depth estimation. In *CVPR*, 2024. [2](#)
- [79] Xinyue Zhang, Jiaqi Yang, Xiangting Meng, Abdelrahman Mohamed, and Laurent Kneip. fcop: Focal length estimation from category-level object priors. *arXiv:2409.19641*, 2024. [2](#)
- [80] Wang Zhao, Shaohui Liu, Yezhi Shu, and Yong-Jin Liu. Towards better generalization: Joint depth-pose learning without posenet. In *CVPR*, 2020. [2](#)
- [81] Xiaoming Zhao, Xingming Wu, Weihai Chen, Peter CY Chen, Qingsong Xu, and Zhengguo Li. Aliked: A lighter keypoint and descriptor extraction network via deformable transformation. *IEEE Transactions on Instrumentation and Measurement*, 72:1–16, 2023. [2](#)
- [82] Qunjie Zhou, Torsten Sattler, Marc Pollefeys, and Laura Leal-Taixe. To learn or not to learn: Visual localization from essential matrices. In *ICRA*, 2020. [2](#)