METRIC CALIBRATION FOR ARBITRARY MONOCULAR DEPTH ESTIMATION MODELS USING ULTRA-SPARSE MEASUREMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing monocular depth estimation methods recovering dense 3D structure from a single RGB image has achieved remarkable success while still suffers from inherent absolute metric scale ambiguity. Sparse metric-guided methods address this by incorporating metric depth measurements. In practice, however, available measurements are often far sparser than required, especially at long range. We refer to this setting as depth estimation with ultra-sparse metric measurements (USMM). To address this problem, we propose AnchorDepth, a training-free post-processing framework that calibrates monocular depth predictions using USMM. It anchors the predicted depth map to the real-world metric scale without any learnable parameters. AnchorDepth identifies approximately fronto-parallel planes from the output of any pretrained monocular depth model and fits local metric corrections within them. It then propagates the remaining calibration residuals using either bounded inverse-squared-distance interpolation (AnchorDepth-Interp) or edge-aware global optimization (AnchorDepth-Refine). Extensive evaluations on public benchmarks showed that AnchorDepth reduced AbsRel by up to 76% compared with the corresponding uncalibrated monocular depth models. Furthermore, considering that existing benchmarks mainly focus on short-range scenes and provide limited evaluation of long-range depth, we built two benchmarks of kilometer-scale: synthetic STREET, and real-world FAR. On both benchmarks, AnchorDepth reduced AbsRel by approximately 40% compared with state-of-the-art metric-guided methods. These results demonstrated that our method outperformed both sparse metric-guided and RGB-only baselines in the USMM setting.

1 INTRODUCTION

Monocular Depth Estimation (MDE) plays a fundamental role in embodied intelligence (Wofk et al., 2019) and autonomous driving (Xue et al., 2020). MDE recovers per-pixel depth from a single RGB image (Arampatzakis et al., 2023). However, the absolute metric scale cannot be determined from a single image: perspective projection maps 3D scenes to 2D observations in a many-to-one manner, making metric-scale recovery ill-posed (Hartley et al., 2003). Recent methods learn geometric priors from large-scale RGB-Depth data and recover detailed relative spatial structure (Ranftl et al., 2022; 2021; Yang et al., 2024a;b). Yet these methods require a large number of ground-truth points to align their predictions to the metric scale through a global affine fit. Metric-depth models address this scale ambiguity through metric supervision or geometric priors (Bhat et al., 2023; Hu et al., 2024; Piccinelli et al., 2024; Bochkovskiy et al., 2025), but their predictions may retain scale residuals that vary across scenes.

Sparse metric guidance incorporates measurements, such as low-resolution radar maps or LiDAR scans, to anchor depth predictions to the metric scale. Learning-based methods either take sparse measurements as direct network input (Uhrig et al., 2017; Ma & Karaman, 2018; Cheng et al., 2018; Park et al., 2020; Zhang et al., 2023; Tang et al., 2024; Yu et al., 2026b; Zuo et al., 2025) or use them to guide and adapt depth foundation models (Viola et al., 2025; Lin et al., 2025; Wang et al., 2026; Jeong et al., 2025; Yu et al., 2026a), while test-time rescaling offers a model-independent alternative (Marsal et al., 2025). These methods are generally designed for relatively dense measurements with structured sampling patterns. These assumptions become restrictive in practical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

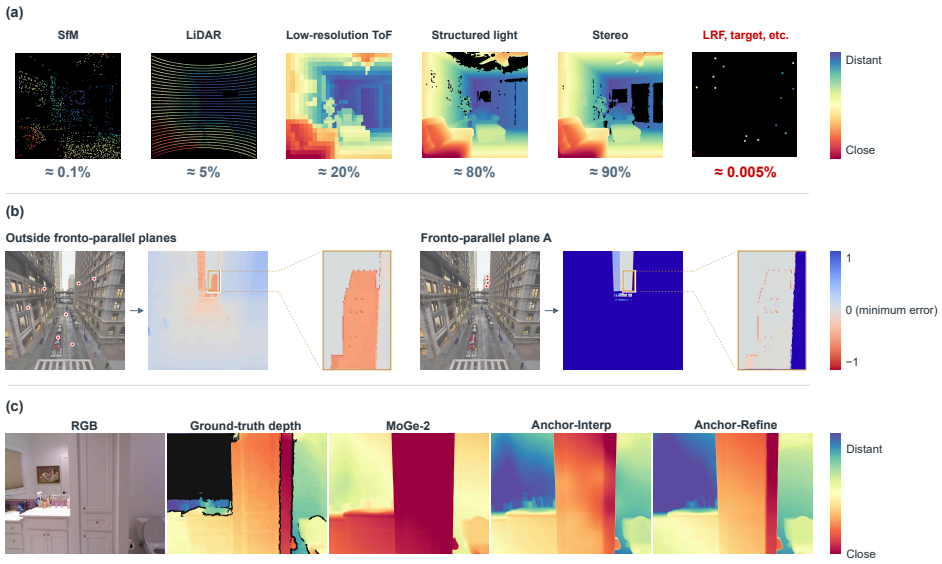

Figure 1: AnchorDepth anchors depth maps to the absolute scale using USMM. (a) Sparsity comparison across existing sensing modalities and the USMM setting. (b) Depth error analysis across different regions. Red dots in the RGB images mark the simulated measurement locations, made more visible by dimming the RGB images. We use all red dots to fit an affine transformation and align the predicted depth to the metric scale. In the error maps, blue and red indicate overestimated and underestimated depth, respectively, with gray marking the minimum absolute error. When the measurements are sampled outside fronto-parallel planes (left), the error on the distant fronto-parallel plane, shown enlarged on the right, remains large. Placing measurements on that plane (right) reduces its error to near zero, but the remaining regions become substantially worse. These observations motivate a plane-wise calibration strategy instead of a single global fit. (c) Comparison of depth maps before and after calibration.

long-range deployments. Common sensing modalities, such as ToF camera and LiDAR, can be expensive (Van Brummelen et al., 2018; Li & Ibanez-Guzman, 2020), and their signal-to-noise ratio degrades at long range (Royo & Ballesta-Garcia, 2019; Zheng et al., 2023).

Alternatives such as low-cost laser rangefinders, surveyed targets, or coarse visual estimates avoid some of these limitations by trading spatial sampling density for range and cost, but they typically yield only a few dozen irregularly distributed metric measurements (Li et al., 2023). For example, the HENYU SH003-A laser rangefinder costs only $56 on Amazon and has an advertised range of 1,100 to 1,400 meters, whereas the Seyond Falcon Prime LiDAR reaches only 500 meters despite costing around $5,300. We discuss possible measurement acquisition schemes in the Appendix A. We refer to this setting as depth estimation with ultra-sparse metric measurements (USMM). Such measurements may account for merely a few hundred-thousandths of the total pixel count, which is substantially lower than the pixel coverage of low-resolution radar maps (tens of percent) and LiDAR (a few percent). This sparsity falls outside the operating regime of existing sparse metric-guided methods (Hu et al., 2022; Wang et al., 2026). Figure 1a illustrates the difference in measurement density.

The large discrepancy between USMM and RGB observations makes direct joint processing challenging. Our key premise is that MDE models, despite their scale ambiguity, learn geometric priors from large-scale training data, which enable them to provide accurate *relative* depth estimates for local structures. These local structures can be effectively aligned to the metric scale through simple affine transformations. (Ranftl et al., 2022). We found that the affine transformations estimated for different structures were often inconsistent, as shown in Figure 1b. When the measurements are scattered across the scene, the calibration performs poorly on the distant fronto-parallel plane, which retains large errors. When we place all measurements on that plane, its error drops to nearly zero, but the error over the remaining regions grows substantially. We attributed this phenomenon

to the depth distribution bias in the training data. As shown in Appendix B, natural scenes are dominated by objects and surfaces with continuous depth variations, while perfectly fronto-parallel planar regions are relatively less frequent. Consequently, models may develop a bias toward predicting smoothly varying depth during training. These observations motivate separate treatment of approximately fronto-parallel planes and the remaining regions.

To address the challenges of USMM, we propose AnchorDepth, a model-agnostic, post-processing calibration framework for monocular depth predictions. AnchorDepth geometrically extracts approximately fronto-parallel planes from the output of a pretrained MDE model and fits local metric corrections individually on each plane. It then handles the remaining regions with two complementary operators: AnchorDepth-Interp propagates non-planar residuals using bounded inverse-squared-distance interpolation, and AnchorDepth-Refine treats the corrected planes as dense seeds and solves for an edge-aware smooth residual field. Both operators require no training and leave the underlying depth model unchanged, using only the RGB image and ultra-sparse metric measurements as input.

Existing outdoor benchmarks typically reach only about 200 meters, with the farthest extending to approximately 400 meters (Narasimhan et al., 2025), and therefore provide limited coverage of kilometer-scale scenes. We present STREET, a synthetic benchmark, and FAR, a real-world benchmark, to extend depth-estimation evaluation to long range. On both benchmarks, AnchorDepth reduced AbsRel by approximately 40% relative to state-of-the-art sparse metric-guided methods under the same measurement setting. Compared with RGB-only monocular depth models, AnchorDepth improved performance consistently across more than ten existing benchmarks, including a 76% AbsRel reduction relative to Depth Pro (Bochkovskiy et al., 2025). As visually shown in Figure 1c, our method outperformed existing MDE models.

2 Related Work

2.1 Monocular Depth Priors

Monocular depth estimation has progressed from supervised multi-scale CNNs (Eigen et al., 2014; Laina et al., 2016) to ordinal, adaptive-bin, and self-supervised formulations (Fu et al., 2018; Bhat et al., 2021; Godard et al., 2019). Large-scale mixed-dataset training with vision transformers produced robust zero-shot relative-depth priors in MiDaS, DPT, Depth Anything, and Depth Anything V2 (DA2) (Ranftl et al., 2022; 2021; Yang et al., 2024a;b). Metric predictors instead use metric supervision, canonical geometry, or camera conditioning: ZoeDepth combines relative and metric heads (Bhat et al., 2023), Metric3D v2 learns a canonicalized geometric representation (Hu et al., 2024), UniDepth and UniDepthV2 target camera- and domain-independent metric depth (Piccinelli et al., 2024; 2026), and Depth Pro emphasizes sharp metric predictions with efficient inference (Bochkovskiy et al., 2025). Current foundation models further extend zero-shot geometry through arbitrary-view reasoning (DA3), metric-scale geometry (MoGe-2), diffusion priors (Marigold and Pixel-Perfect Depth), and implicit neural fields (InfiniDepth) (Lin et al., 2026; Wang et al., 2025; Ke et al., 2024; Xu et al., 2025; Yu et al., 2026a). These predictors supply strong dense priors, but their remaining scale and shape errors can vary with camera, domain, and depth range, which motivates a post-processing calibration pipeline.

2.2 Sparse Metric Guidance

Depth completion aims to reconstruct dense scene depth by using sparse depth as a metric prior, combined with optional guidance information. Sparsity Invariant CNNs and Sparse-to-Dense established this formulation (Uhrig et al., 2017; Ma & Karaman, 2018); CSPN and NLSPN learn local or non-local affinities (Cheng et al., 2018; Park et al., 2020); and CompletionFormer, BP-Net, and LDCM extend completion with transformer, bilateral propagation, or large-model propagation (Zhang et al., 2023; Tang et al., 2024; Yu et al., 2026b). These methods learn a dense mapping from paired RGB and sparse-depth data, so their behavior is tied to the training domain, sensor characteristics, and sampling distribution. That dependency is a different assumption from calibrating an arbitrary frozen predictor after inference.

Foundation-model guidance reduces some dependence on conventional completion backbones but still changes how sparse measurements enter the model. Marigold-DC optimizes scale and shift

during guided diffusion (Viola et al., 2025), while OMNI-DC integrates sparse depth at multiple resolutions (Zuo et al., 2025). PromptDA injects low-resolution radar map as decoder prompts (Lin et al., 2025); PriorDA combines pixel-level alignment with distance-aware pre-filling (Wang et al., 2026); TestPromptDC adapts prompts at test time for zero-shot completion (Jeong et al., 2025); and InfiniDepth provides a dedicated DepthSensor configuration (Yu et al., 2026a). These systems are important sparse-guided baselines, but each incorporates the guidance into a model-specific inference or adaptation process. In contrast, AnchorDepth operates after monocular inference and applies the same calibration interface to frozen metric and relative-depth predictors.

Test-time scale or affine alignment offers a model-independent alternative by estimating a global transform from a few metric points (Marsal et al., 2025). Its single global mapping cannot express residuals that differ across surfaces. AnchorDepth addresses this representational gap by assigning metric evidence to geometric regions and propagating the remaining log residual according to image structure, without retraining or camera intrinsics.

3 METHOD

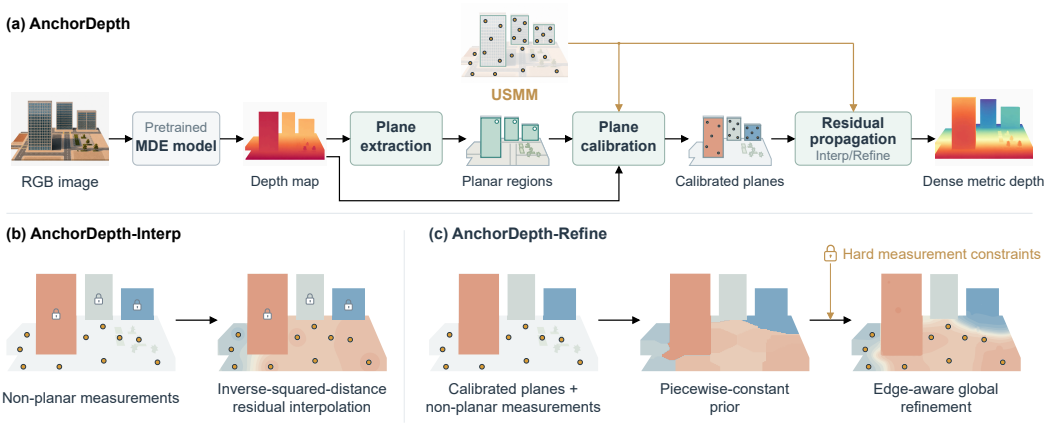

Figure 2: Given an RGB image and USMM, we extract approximately fronto-parallel planes from the outputs of a pre-trained depth model and calibrate each plane via region-wise affine fitting on local measurements. For the remaining regions, we employ two distinct strategies, AnchorDepth-Interp (b) and AnchorDepth-Refine (c), to propagate residuals and calibrate the depth map.

3.1 OVERVIEW

AnchorDepth uses ultra-sparse metric measurements to calibrate the prediction of a frozen MDE model. As shown in Figure 2a, AnchorDepth first detects approximately fronto-parallel planes from the output depth map of pretrained MDE model. Measurements within each detected plane then determine a local affine correction for that region. Once the planes have been corrected, the remaining regions are calibrated by propagating measurement residuals. AnchorDepth-Interp performs this propagation with efficient inverse-squared-distance interpolation, whereas AnchorDepth-Refine uses an edge-aware global solve to improve spatial consistency. Both variants therefore produce a metric depth map without updating the MDE model or introducing learnable parameters.

3.2 GENERAL CALIBRATION PROCEDURE

Let $\mathbf{I} \in \mathbb{R}^{3\times H\times W}$ be the input RGB image. The pretrained MDE model takes $\mathbf{I}$ as input and outputs a depth map $\mathbf{Z}_0 \in \mathbb{R}_{>0}^{H\times W}$. The method also receives M metric measurements, where M is the count of metric measurements. Measurement i contains its image coordinate $\mathbf{x}_i \in \mathbb{R}^2$ and depth $z_i \in \mathbb{R}_{>0}$. We use $\mathbf{x} \in \mathbb{R}^2$ for a pixel coordinate.

Together, these inputs support a two-stage calibration. The first stage detects a set of disjoint planes. For each plane, it fits an affine scale and bias by least squares from the measurements in that region

and calibrate to every pixel in the plane. This process produces the plane-corrected depths. Exact fitting rules are given in Appendix C.1.

The second stage therefore transfers metric information to the remaining regions in the log-residual domain. Let $\rho_i \in \mathbb{R}$ denote the residual of measurement i, $\mathbf{r} \in \mathbb{R}^{H\times W}$ a dense residual map, and $\mathbf{Z} \in \mathbb{R}^{H\times W}_{>0}$ the calibrated output. Their shared representation is

$$\rho_i = \log z_i - \log \mathbf{Z}_0(\mathbf{x}_i), \qquad \mathbf{Z}(\mathbf{x}) = \mathbf{Z}_0(\mathbf{x}) \exp\big(\mathbf{r}(\mathbf{x})\big). \tag{1}$$

The left equation in Eq. 1 turns the multiplicative error at each measurement into an additive residual that can be propagated across the image. After either variant estimates the dense residual field, the second equation decodes that field into positive metric depth.

3.3 Plane Extraction Method

This subsection produces the plane required by the first calibration stage. This module converts $\mathbf{Z}_0$ into inverse depth. Robust clipping and edge-preserving smoothing (Tomasi & Manduchi, 1998) prepare this representation for region extraction. Relative depth discontinuities then divide the image into connected candidates, and an affine inverse-depth surface is fitted to each candidate. The fit error, image-plane slope, and internal dispersion determine whether a candidate is approximately fronto-parallel. Candidates that satisfy all three tests become the planes used for plane-wise calibration. Because the entire procedure operates on the depth map, it requires no camera intrinsics. Appendix C.2 lists the exact tests and thresholds.

3.4 Residual Propagation

Plane-wise calibration provides reliable corrections inside the selected masks but leaves the rest of the image unresolved. Both variants address these remaining regions using the residual representation in Eq. 1. AnchorDepth-Interp directly interpolates the non-planar residuals and preserves the calibrated planes, prioritizing efficiency. AnchorDepth-Refine instead combines the plane and point corrections into a regional prior and refines it globally with edge-aware propagation, prioritizing spatial consistency. Numerical settings and the extension to denser measurements are provided in Appendices C.3 and D, respectively.

3.4.1 AnchorDepth-Interp: Inverse-Squared-Distance Residual Interpolation

AnchorDepth-Interp is designed for cases in which low calibration overhead is the main priority. Because the selected planes have already been corrected, it uses only the coordinates and residuals of the remaining measurements. Let $\mathcal{I}_{\mathrm{np}} \subseteq \{1, \dots, M\}$ denote their indices and $\epsilon_d \in \mathbb{R}_{>0}$ a distance stabilizer. For measurement i, let $w_i(\mathbf{x}) \in \mathbb{R}_{>0}$ be its influence at pixel $\mathbf{x}$. These quantities define the dense residual map $\widehat{\mathbf{r}}_{\mathrm{I}} \in \mathbb{R}^{H\times W}$ through bounded inverse-squared-distance interpolation (Shepard, 1968):

$$w_i(\mathbf{x}) = \max\{\|\mathbf{x} - \mathbf{x}_i\|_2^2, \epsilon_d\}^{-1}, \qquad \widehat{\mathbf{r}}_{\mathrm{I}}(\mathbf{x}) = \frac{\sum_{i\in\mathcal{I}_{\mathrm{np}}} w_i(\mathbf{x})\rho_i}{\sum_{i\in\mathcal{I}_{\mathrm{np}}} w_i(\mathbf{x})}. \tag{2}$$

The left equation in Eq. 2 gives nearby measurements greater influence, while ϵ_d prevents unstable weights at short distances. The second then averages the measured residuals using these weights. Finally, decoding $\widehat{\mathbf{r}}_{\mathrm{I}}$ with Eq. equation 1 calibrates the non-planar regions, while the affine-corrected plane depths remain unchanged.

3.4.2 AnchorDepth-Refine: Geodesic Prior and Edge-Aware Global Refinement

Distance alone does not indicate whether two pixels lie on the same surface. AnchorDepth-Refine therefore uses predicted depth boundaries to guide residual propagation. Its inputs are the affine-corrected plane depths, the remaining point residuals, and the log-depth guide $\boldsymbol{\ell} \in \mathbb{R}^{H\times W}$, defined by $\boldsymbol{\ell}(\mathbf{x}) = \log \mathbf{Z}_0(\mathbf{x})$. For each plane, the median pointwise log difference between its corrected depth and $\mathbf{Z}_0$ forms a plane marker. Each remaining measurement contributes its residual ρ_i as a point marker. Together, these markers initialize a marker-controlled watershed (Vincent & Soille, 1991), which spreads them into a piecewise-constant regional prior $\mathbf{r}_{\mathrm{p}} \in \mathbb{R}^{H\times W}$.

The regional prior provides broad coverage, but its global refinement must avoid transferring residuals across depth discontinuities. Let $\mathcal{E}$ denote neighboring pixel pairs $(\mathbf{x}, \mathbf{y})$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^2$. Let $\sigma \in \mathbb{R}_{>0}$ be a contrast scale and $w_{\min} \in \mathbb{R}_{>0}$ the minimum propagation weight. For each pair in $\mathcal{E}$, the propagation weight $w_{\mathbf{xy}} \in \mathbb{R}_{>0}$ is computed from the local log-depth contrast:

$$w_{\mathbf{xy}} = \max\left\{ \exp\left(-\frac{(\boldsymbol{\ell}(\mathbf{x}) - \boldsymbol{\ell}(\mathbf{y}))^2}{2\sigma^2} \right), w_{\min} \right\}. \tag{3}$$

Consequently, neighboring pixels with similar predicted depths receive a large weight and can exchange residuals strongly. At a depth cliff, the exponential term approaches zero and the weight falls to $w_{\min}$, thereby suppressing cross-boundary propagation.

These weights are then combined with the watershed prior in a global optimization. Let $\lambda \in \mathbb{R}_{\geq 0}$ be the prior-fidelity coefficient, let $\mathbf{r} \in \mathbb{R}^{H \times W}$ be the unknown residual field, and let $\widehat{\mathbf{r}}_{\mathrm{R}} \in \mathbb{R}^{H \times W}$ denote its output. The refinement solves

$$\begin{aligned} \widehat{\mathbf{r}}_{\mathrm{R}} = \arg\min_{\mathbf{r} \in \mathbb{R}^{H \times W}} & \left[\sum_{(\mathbf{x},\mathbf{y}) \in \mathcal{E}} w_{\mathbf{xy}} \big(\mathbf{r}(\mathbf{x}) - \mathbf{r}(\mathbf{y}) \big)^2 + \lambda \|\mathbf{r} - \mathbf{r}_{\mathrm{p}}\|_F^2 \right] \\ & \text{s.t.} \quad \mathbf{r}(\mathbf{x}_i) = \rho_i, \qquad i = 1, \ldots, M. \end{aligned} \tag{4}$$

The first term propagates residuals between neighboring pixels in proportion to their edge weights, whereas the second keeps the solution close to the regional prior. The hard constraints preserve every metric measurement. Thus, unlike AnchorDepth-Interp, AnchorDepth-Refine uses the plane corrections as a prior that can be adjusted during global refinement rather than freezing them. The optimized field $\widehat{\mathbf{r}}_{\mathrm{R}}$ is finally decoded through Eq. equation 1 to obtain the calibrated depth map. Appendix C.3 specifies the solver grid and numerical settings.

4 Experiments

4.1 Evaluation Benchmarks

Existing depth-estimation benchmarks primarily covered near-range scenes and therefore provided limited evidence about long-range deployment. We constructed two kilometer-scale benchmarks for long-range depth evaluation. STREET comprised 12,000 street-view images of virtual cities generated with a physics engine. It was divided by illumination into STREET-day and STREET-dark, each containing 6,000 images captured under simulated daytime or nighttime conditions. The maximum valid depth in STREET was 2,000 meters. FAR was captured with 3-D LiDAR scanners and comprised 500 images, with a maximum valid depth of approximately 730 meters. Further details of both datasets are provided in Appendix E.

For broader evaluation, we also assessed AnchorDepth on widely used depth-estimation benchmarks: Booster (Ramirez et al., 2024), Middlebury (Scharstein et al., 2014), NuScenes (Caesar et al., 2020), Sintel (Butler et al., 2012), ETH3D (Schops et al., 2017), SUN RGB-D (Song et al., 2015), iBims-1 (Koch et al., 2019), KITTI (Geiger et al., 2012), NYUv2 (Silberman et al., 2012), HAMMER (Jung et al., 2023), ScanNet (Dai et al., 2017), DIODE (Vasiljevic et al., 2019), and DDAD (Guizilini et al., 2020).

4.2 Comparison with Monocular Depth Estimation Methods

We evaluated AnchorDepth as a post-processing module for a diverse set of MDE models. The metric-depth estimators were DepthPro (Bochkovskiy et al., 2025), UniDepth V2 (Piccinelli et al., 2026), MoGe 2 (Wang et al., 2025), and DepthAnything 3 (Lin et al., 2026). The relative-depth estimators were Marigold (Ke et al., 2024), DepthAnything V2 (Yang et al., 2024b), Pixel-Perfect Depth (Xu et al., 2025), and InfiniDepth (Yu et al., 2026a). For each model, we used all or a subset of the datasets evaluated in its original paper and followed its official evaluation protocol. We used the raw outputs of metric-depth estimators and applied additional global scale alignment to the outputs of relative-depth estimators.

Figure 3 showed that AnchorDepth improved the evaluated monocular depth estimators. For metric-depth models, it calibrated predictions to the correct scale range and improved their performance.

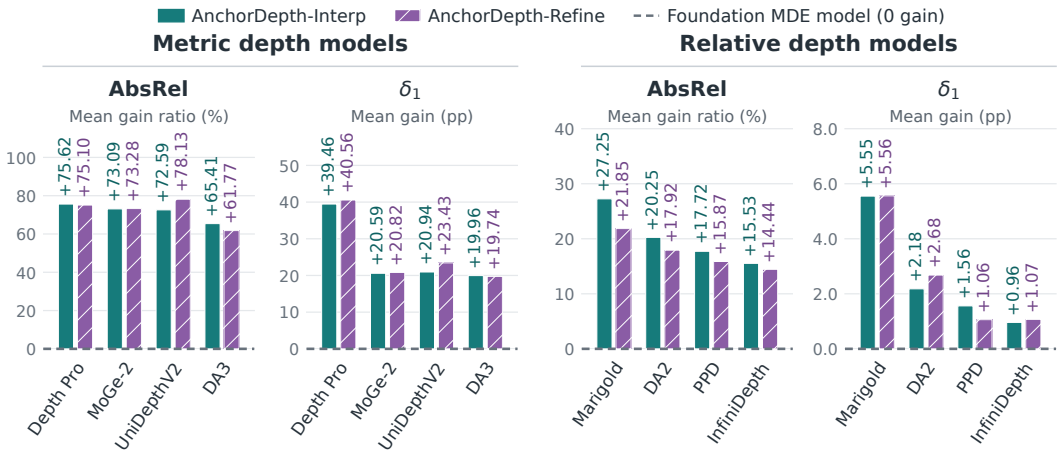

Figure 3: Average gains of the calibration methods under metric-depth and relative-depth evaluations. The two panels are vertically arranged with identical display widths.

For relative-depth models, it improved performance by correcting residual patterns independently across regions. The qualitative results in Figure A3 also indicated improvements in geometric shape accuracy. Detailed calibration results for each MDE model are provided in Appendix F.

4.3 Comparison with Sparse Metric Guidance Methods

We compared AnchorDepth with state-of-the-art sparse metric-guided approaches, including PromptDA (Lin et al., 2025), OMNI-DC (Zuo et al., 2025), Marigold-DC (Viola et al., 2025), RANSAC-Affine (Marsal et al., 2025), PriorDA (Wang et al., 2026), LDCM (Yu et al., 2026b), and InfiniDepth (Yu et al., 2026a). Following prior work, we evaluated these methods on five real-world datasets (Silberman et al., 2012; Geiger et al., 2012; Dai et al., 2017; Vasiljevic et al., 2019; Schops et al., 2017) and on STREET and FAR. Detailed dataset-specific preprocessing and sparse-input protocols are provided in Appendix G.1. Shared fairness controls are given in Appendix G.2. Table 1 showed that AnchorDepth performed favorably against existing trained sparse metric-guided approaches under USMM. For each competing method, we evaluated two sparse-input locations: the same structured measurements as AnchorDepth and the same number of points sampled at random locations. AnchorDepth consistently improved performance over the competing methods under both conditions. It reduced AbsRel by up to 40% (from 0.048 to 0.029) compared with the second-best method on our long-range benchmarks. We can see an improvement of 76% Table A5 also showed that AnchorDepth introduced only minimal additional runtime overhead.

4.4 Ablation Studies and Analysis

Table 2 shows that calibrating approximately fronto-parallel planes before propagating the remaining residuals improved both variants, with a larger gain for AnchorDepth-Refine because the calibrated planes provided dense prior seeds. Under global processing, AnchorDepth-Interp also outperformed RANSAC-Affine (Marsal et al., 2025), supporting the proposed residual representation. Additional ablations, and visualizations are provided in Appendix H.

AnchorDepth used the depth maps produced by the base depth-estimation models for calibration. Consequently, weaker base-model predictions could adversely affect its subsequent plane extraction and correction. To examine this dependency, we evaluated MoGe-2 models with different parameter counts (Wang et al., 2025) and measured how calibration performance changed with the quality of the input depth map. This setting also reflected compute-constrained edge deployments that used smaller models. The quantitative results are summarized in Table 3.

Table 1: Comparison with sparse metric guidance methods. *Matched* reuses the exact measurement locations selected for AnchorDepth, whereas *Random* uniformly samples the same number of measurements from valid locations in each image. The average number of measurements used per dataset is indicated in parentheses next to the dataset name. The best results are in **bold**, and the second-best are underlined.

Method	Locations	FAR(26.15)		STREET-day(18.78)		STREET-dark(18.48)		KITTI(18.08)		ETH3D(16.92)		NYUv2(16.93)		ScanNet(12.98)		DIODE(15.17)	
		AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑
PromptDA	Random	0.326	61.185	0.210	68.789	0.211	68.924	0.375	53.821	0.192	75.954	0.151	79.054	0.170	76.505	0.311	70.724
OMNI-DC	Random	0.226	75.203	0.087	90.431	0.090	90.028	0.158	73.324	0.082	90.262	0.132	82.593	0.156	77.683	0.166	74.687
Marigold-DC	Random	0.383	68.446	0.189	83.778	0.178	83.749	0.136	83.683	0.096	90.593	0.078	92.489	0.099	88.703	0.202	81.123
RANSAC-Affine	Random	0.179	78.290	0.036	98.102	0.045	96.566	0.070	94.753	0.044	96.978	0.041	97.480	0.051	96.250	0.091	90.829
PriorDA	Random	0.164	81.187	0.052	94.669	0.056	93.872	0.082	91.685	0.069	93.719	0.063	94.584	0.079	92.192	0.117	87.394
LDCM	Random	0.164	78.939	0.038	96.947	0.045	95.631	0.129	83.878	0.035	97.715	0.046	96.666	0.051	95.927	0.082	90.735
InfiniDepth	Random	0.561	83.462	0.030	97.597	0.031	97.369	0.052	96.579	0.029	98.664	0.044	96.786	0.066	94.438	0.145	85.965
PromptDA	Structured	0.881	44.075	0.401	59.676	0.471	62.465	0.316	55.190	0.153	79.522	0.131	83.086	0.129	83.894	0.140	85.657
OMNI-DC	Structured	0.252	68.668	0.071	93.730	0.092	90.758	0.109	82.966	0.052	94.713	0.057	93.085	0.048	96.922	0.053	95.678
Marigold-DC	Structured	1.175	47.972	0.482	69.639	0.350	72.070	0.144	82.537	0.067	94.254	0.065	94.332	0.068	93.622	0.067	93.148
RANSAC-Affine	Structured	0.344	57.891	0.075	93.263	0.101	87.784	0.089	91.429	0.045	97.149	0.040	97.404	0.044	97.097	0.074	91.552
PriorDA	Structured	0.245	72.546	0.109	91.827	0.222	89.517	0.083	91.574	0.044	96.652	0.054	95.640	0.055	95.753	0.053	94.802
LDCM	Structured	0.180	75.729	0.037	97.930	0.048	96.075	0.156	80.156	0.038	97.238	0.040	97.125	0.041	96.972	0.052	94.394
InfiniDepth	Structured	0.844	72.175	0.084	87.774	0.073	91.615	0.051	96.554	**0.022**	**99.384**	**0.032**	**98.297**	0.038	97.414	0.059	96.096
MoGe-2 Raw	–	0.642	10.911	0.115	95.647	0.116	96.393	0.093	93.107	0.116	89.790	0.083	94.753	0.101	89.796	0.191	61.268
AnchorDepth-Interp	Structured	0.164	82.539	0.034	98.147	0.031	98.048	**0.047**	**96.781**	0.025	98.913	**0.032**	98.192	**0.035**	**98.008**	0.051	95.083
AnchorDepth-Refine	Structured	**0.115**	**87.023**	**0.027**	**98.836**	**0.029**	**98.601**	0.057	96.087	0.026	98.897	0.034	97.948	**0.035**	97.866	**0.043**	**96.415**

Table 2: Ablation study on partition calibration, with MoGe-2-L as the foundation model.

Variant	DDAD		DIODE		ETH3D		HAMMER		iBims-1		KITTI		NYUv2	
	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑	AbsRel↓	δ_1(%)↑
Global RANSAC-Affine	0.225	69.608	0.081	90.828	0.057	94.964	0.036	99.048	0.036	98.012	0.090	91.865	0.048	96.797
Global AnchorDepth-Interp	0.110	89.626	0.045	96.765	0.034	98.427	0.029	99.479	0.030	98.702	0.044	97.903	0.035	98.083
AnchorDepth-Interp	0.102	90.187	0.034	97.759	0.027	98.974	0.027	99.517	0.027	98.692	0.040	97.975	0.032	98.501
Global AnchorDepth-Refine	0.155	74.046	0.166	68.324	0.099	88.959	0.259	65.425	0.145	80.673	0.174	65.433	0.068	96.785
AnchorDepth-Refine	0.099	91.609	0.029	98.294	0.026	98.935	0.036	99.179	0.023	98.838	0.043	97.920	0.031	98.447

4.5 On-device Evaluation

We evaluated AnchorDepth, MoGe-2, PriorDA, and InfiniDepth using a synchronized low-cost RGB–LiDAR rig. Sparse LiDAR measurements were used only for calibration, while object distances computed from the reconstructed point clouds provided the evaluation targets. AnchorDepth achieved relative errors of 0.012 outdoors and 0.030 indoors using only 38 and 33 measurements per frame, respectively (Table 4). The complete protocol is provided in Appendix H.3.

4.6 Evaluation on Out-of-distribution Scenes

We further evaluated AnchorDepth on UAV imagery, where general-purpose MDE models typically suffer from severe domain and scale shifts. As shown in Table 5, after calibrating the general-purpose MoGe 2 with AnchorDepth, it outperforms both general-purpose metric-depth models and the domain-specific model MoGe2-Aerial (Song et al., 2026), which extensively fine-tuned with large-scale data, across all evaluated UAV scene categories.

5 Conclusion & Limitation

We propose AnchorDepth, a training-free and model-agnostic post-processing calibration framework for monocular depth estimation with extremely sparse metric measurements. It maps the output of any depth estimator to the true metric scale using only a minimal number of sparse measurements, and partially corrects region-specific errors. We validate the effectiveness of AnchorDepth on proposed long-range dataset and existing benchmarks. Using merely dozens of sparse measurement points, AnchorDepth significantly outperforms any RGB-only depth estimation model, and also surpasses existing sparse-guided methods under the same level of sparsity.

Limitation and Future Work. AnchorDepth relies on the underlying depth predictions and cannot correct substantial geometric errors. Future work will further investigate sparse measurements to improve its ability to correct erroneous geometry.

Table 3: Ablation study of AnchorDepth on MoGe-2 models with different parameter scales.

Model	Params	Method	DDAD		DIODE		ETH3D		HAMMER		iBims-1		KITTI		NYUv2	
			AbsRel↓	δ_1(%) ↑	AbsRel↓	δ_1(%) ↑	AbsRel↓	δ_1(%) ↑	AbsRel↓	δ_1(%) ↑	AbsRel↓	δ_1(%) ↑	AbsRel↓	δ_1(%) ↑	AbsRel↓	δ_1(%) ↑
MoGe-2-S	35M	Raw	0.191	64.909	0.228	51.849	0.155	75.945	0.278	63.702	0.192	65.610	0.220	44.358	0.105	88.939
		AnchorDepth-Interp	0.131	85.571	0.054	95.648	0.050	97.064	0.038	98.755	0.044	97.132	0.050	96.457	0.044	97.686
		AnchorDepth-Refine	0.146	85.356	0.042	96.772	0.045	97.207	0.045	98.375	0.038	98.096	0.061	95.244	0.043	97.541
MoGe-2-B	104M	Raw	0.172	68.824	0.188	64.828	0.134	83.863	0.221	56.556	0.149	74.665	0.214	44.421	0.087	94.374
		AnchorDepth-Interp	0.116	88.294	0.043	96.820	0.036	98.270	0.033	99.360	0.035	98.118	0.044	97.413	0.040	98.145
		AnchorDepth-Refine	0.115	88.053	0.034	97.824	0.034	98.355	0.039	98.941	0.029	98.652	0.057	95.689	0.038	98.058
MoGe-2-L	331M	Raw	0.156	73.653	0.167	67.850	0.100	88.798	0.260	65.287	0.146	80.459	0.176	64.696	0.069	96.740
		AnchorDepth-Interp	0.102	90.187	0.034	97.759	0.027	98.974	0.027	99.517	0.027	98.692	0.040	97.975	0.032	98.501
		AnchorDepth-Refine	0.099	91.609	0.029	98.294	0.026	98.935	0.036	99.179	0.023	98.838	0.043	97.920	0.031	98.447

Table 4: Test results in real-world scenarios. AnchorDepth uses MoGe-2-S as the foundation model.

Scenario	Method	Relative error	Ground truth count per frame	Scan time per frame (ms)
Outdoor	MoGe-2	0.550	-	-
	PriorDA	0.157	27843	≈ 5000
	InfiniDepth-DepthSensor	0.091	29288	≈ 5000
	AnchorDepth	**0.012**	**38**	$\approx$**100**
InDoor	MoGe-2	0.313	-	-
	PriorDA	0.133	31737	≈ 5000
	InfiniDepth-DepthSensor	0.090	33549	≈ 5000
	AnchorDepth	**0.030**	**33**	$\approx$**100**

Table 5: Evaluation of metric depth estimation on UAV scenarios. All method evaluated without Ground-Truth intrinsics. AnchorDepth employed MoGe 2 as foundation model.

Method	AerialMetric-Oblique						AerialMetric-Decoupled							
	City		Natural		Rural		Building		Factory		Farm		Lawn	
	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1↑	AbsRel↓	δ_1
ZoeDepth (Bhat et al., 2023)	97.1	0.0	93.0	0.0	92.8	0.0	95.4	0.0	96.3	0.0	96.6	0.0	95.2	0.0
DepthPro (Bochkovskiy et al., 2025)	97.8	0.0	92.2	0.0	96.9	0.0	97.8	0.0	97.6	0.0	97.3	0.0	96.7	0.0
UniDepthV1 (Piccinelli et al., 2024)	83.9	0.0	80.3	0.0	86.0	0.0	87.4	0.0	81.1	0.0	89.5	0.0	84.6	0.0
UniDepthV2 (Piccinelli et al., 2026)	31.0	34.1	19.2	73.0	28.1	24.2	24.1	35.0	22.0	46.7	17.2	57.9	24.4	33.8
MoGe-2 (Wang et al., 2025)	48.4	5.1	45.9	23.0	56.7	0.3	29.0	26.6	26.8	34.8	71.0	0.0	32.1	19.5
MoGe2-Aerial (Song et al., 2026)	10.3	89.3	16.8	73.6	14.2	<u>81.2</u>	11.9	87.6	11.4	90.2	22.8	54.2	9.8	94.0
AnchorDepth-Interp	**3.2**	**98.8**	**3.3**	**98.3**	**2.5**	**99.8**	3.6	99.0	2.4	99.7	**1.0**	**100.0**	**1.3**	**100.0**
AnchorDepth-Refine	<u>3.5</u>	<u>98.7</u>	<u>3.9</u>	<u>98.2</u>	<u>2.9</u>	**99.8**	**2.8**	**99.4**	**1.8**	**99.8**	<u>1.4</u>	<u>99.9</u>	<u>1.7</u>	<u>100.0</u>

AI USE STATEMENT

In the course of this research and in preparing the manuscript, we utilized Large Language Models (LLMs) for two distinct purposes. First, during the manuscript preparation phase, an LLM was used to assist in refining the wording and improving the clarity of the English prose. Its role in this capacity was strictly limited to enhancing sentence structure, grammar, and the overall flow of the text. Beyond these specified roles, LLMs were not involved in the initial research design, data collection, or the generation of core scientific ideas. All substantive content, methodologies, and conclusions are entirely the original work of the authors.

REPRODUCIBILITY STATEMENT

To ensure the reproducibility of our work, we have detailed the reproduction specifics in the appendix. Appendix C describes the comprehensive hyperparameter settings for AnchorDepth, providing guidance for reproducing our evaluation pipeline. Appendix G outlines the specific evaluation protocols used in our experiments. The relevant code, dataset metadata, and detailed experimental configurations for AnchorDepth will be made publicly available upon the acceptance of this paper.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *IEEE transactions on pattern analysis and machine intelligence*, 34(11):2274–2282, 2012.

Vasileios Arampatzakis, George Pavlidis, Nikolaos Mitianoudis, and Nikos Papamarkos. Monocular depth estimation: A thorough review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(4):2396–2414, 2023.

Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Yuri Feigin, Peter Fu, Thomas Gebauer, Daniel Kurz, Tal Dimry, Brandon Joffe, Arik Schwartz, and Elad Shulman. ARKitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021.

Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Alexey Bochkovskiy, 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. In *The Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025.

Daniel J. Butler, Jonas Wulff, Garrett B. Stanley, and Michael J. Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision – ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pp. 611–625. Springer Berlin Heidelberg, 2012.

Yohann Cabon, Naila Murray, and Martin Humenberger. Virtual kitti 2. *arXiv preprint arXiv:2001.10773*, 2020.

Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11618–11628. IEEE, 2020.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. In *International conference on learning representations*, volume 2026, pp. 138846–138923, 2026.

Xinjing Cheng, Peng Wang, and Ruigang Yang. Depth estimation via affinity learned with convolutional spatial propagation network. In *Proceedings of the European Conference on Computer Vision*, 2018.

Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*, 2014.

Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361. IEEE, 2012.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

Jose L Gómez, Manuel Silva, Antonio Seoane, Agnes Borrás, Mario Noriega, Germán Ros, Jose A Iglesias-Guitian, and Antonio M López. All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. *Neurocomputing*, 637:130038, 2025.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3D packing for self-supervised monocular depth estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2482–2491. IEEE, 2020.

Richard Hartley, Andrew Zisserman, et al. *Multiple view geometry in computer vision*, volume 2. Cambridge university press Cambridge, 2003.

Junjie Hu, Chenyu Bao, Mete Ozay, Chenyou Fan, Qing Gao, Honghai Liu, and Tin Lun Lam. Deep depth completion from extremely sparse data: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8244–8264, 2022.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Kaixuan Wang, Hao Chen, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024.

Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2821–2830. IEEE, 2018.

Chanhwi Jeong, Inhwan Bae, Jin-Hwi Park, and Hae-Gon Jeon. Test-time prompt tuning for zero-shot depth completion. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9443–9454. IEEE, 2025.

HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Ales Leonardis, Nassir Navab, and Benjamin Busam. On the importance of accurate geometry data for dense 3D vision tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 780–791. IEEE, 2023.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9492–9502, 2024.

Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco K"orner. Evaluation of CNN-based single-image depth estimation methods. In *Computer Vision – ECCV 2018 Workshops*, volume 11129 of *Lecture Notes in Computer Science*, pp. 331–348. Springer International Publishing, 2019.

Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth International Conference on 3D Vision (3DV)*, 2016.

Haoyu Li, Sihua Xiang, Lu Zhang, Jianzhong Zhu, Song Wang, and You Wang. A range ambiguity classification algorithm for automotive lidar based on fpga platform acceleration. *Frontiers in Physics*, Volume 11 - 2023, 2023.

You Li and Javier Ibanez-Guzman. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine*, 37(4): 50–61, 2020.

Haotong Lin, Sida Peng, Jingxiao Chen, Songyou Peng, Jiaming Sun, Minghuan Liu, Hujun Bao, Jiashi Feng, Xiaowei Zhou, and Bingyi Kang. Prompting depth anything for 4k resolution accurate metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17070–17080, 2025.

Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Yang Zhao, Sida Peng, Hengkai Guo, Xiaowei Zhou, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026.

Fangchang Ma and Sertac Karaman. Sparse-to-dense: Depth prediction from sparse depth samples and a single image. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018.

Rémi Marsal, Alexandre Chapoutot, Philippe Xu, and David Filliat. A simple yet effective test-time adaptation for zero-shot monocular metric depth estimation. In *2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 7858–7865, 2025.

Gautham Narayan Narasimhan, Heethesh Vhavle, Kumar Bhargav Vishvanatha, and James Reuther. Aevascenes: A dataset and benchmark for fmcw lidar perception, 2025. URL `https://scenes.aeva.com/`.

Jinsun Park, Kyungdon Joo, Zhe Hu, Chi-Kuei Liu, and In So Kweon. Non-local spatial propagation network for depth completion. In *Proceedings of the European Conference on Computer Vision*, 2020.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026.

Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: A benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1): 85–102, 2024.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

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 Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10892–10902. IEEE, 2021.

Santiago Royo and Maria Ballesta-Garcia. An overview of lidar imaging systems for autonomous vehicles. *Applied sciences*, 9(19):4093, 2019.

Daniel Scharstein, Heiko Hirschm"uller, York Kitajima, Greg Krathwohl, Nera Ne si'c, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *Pattern Recognition*, volume 8753 of *Lecture Notes in Computer Science*, pp. 31–42. Springer International Publishing, 2014.

Thomas Schops, Johannes L. Schonberger, 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 *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2538–2547. IEEE, 2017.

Donald Shepard. A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 ACM National Conference*, pp. 517–524, 1968.

Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from RGBD images. In *Computer Vision – ECCV 2012*, volume 7576 of *Lecture Notes in Computer Science*, pp. 746–760. Springer, 2012.

Shuran Song, Samuel P. Lichtenberg, and Jianxiong Xiao. SUN RGB-D: A RGB-D scene understanding benchmark suite. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 567–576. IEEE, 2015.

Zhongqiang Song, Guanying Chen, Yuqi Zhang, Yin Zou, Chuanyu Fu, Zhiyuan Yuan, Chuan Huang, Shuguang Cui, and Xiaochun Cao. Aerialmetric: Benchmarking and adapting uav monocular metric depth estimation in the real world. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2026.

Jie Tang, Fei-Peng Tian, Boshi An, Jian Li, and Ping Tan. Bilateral propagation network for depth completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Carlo Tomasi and Roberto Manduchi. Bilateral filtering for gray and color images. In *Proceedings of the Sixth International Conference on Computer Vision*, pp. 839–846, 1998.

Fabio Tosi, Yiyi Liao, Carolin Schmitt, and Andreas Geiger. Smd-nets: Stereo mixture density networks. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8938–8948. IEEE, 2021.

Jonas Uhrig, Nick Schneider, Lukas Schneider, Uwe Franke, Thomas Brox, and Andreas Geiger. Sparsity invariant cnns. In *2017 International Conference on 3D Vision (3DV)*, 2017.

Jessica Van Brummelen, Marie O'brien, Dominique Gruyer, and Homayoun Najjaran. Autonomous vehicle perception: The technology of today and tomorrow. *Transportation research part C: emerging technologies*, 89:384–406, 2018.

Igor Vasiljevic, Nick Kolkin, Shanyi Zhang, Ruotian Luo, Haochen Wang, Falcon Z. Dai, Andrea F. Daniele, Mohammadreza Mostajabi, Steven Basart, Matthew R. Walter, and Gregory Shakhnarovich. DIODE: A dense indoor and outdoor DEpth dataset. *arXiv preprint arXiv:1908.00463*, 2019.

Luc Vincent and Pierre Soille. Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(6): 583–598, 1991.

Massimiliano Viola, Kevin Qu, Nando Metzger, Bingxin Ke, Alexander Becker, Konrad Schindler, and Anton Obukhov. Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5359–5370, 2025.

Qiang Wang, Shizhen Zheng, Qingsong Yan, Fei Deng, Kaiyong Zhao, and Xiaowen Chu. Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2021.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems*, volume 38, pp. 40377–40408, 2025.

Wenshan Wang, Delong Zhu, Xiangwei Wang, Yaoyu Hu, Yuheng Qiu, Chen Wang, Yafei Hu, Ashish Kapoor, and Sebastian Scherer. Tartanair: A dataset to push the limits of visual slam. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4909–4916. IEEE, 2020.

Zehan Wang, Siyu Chen, Lihe Yang, Jialei Wang, Ziang Zhang, Hengshuang Zhao, and Zhou Zhao. Depth anything with any prior. In *The Fourteenth International Conference on Learning Representations (ICLR 2026)*, 2026.

Diana Wofk, Fangchang Ma, Tien-Ju Yang, Sertac Karaman, and Vivienne Sze. Fastdepth: Fast monocular depth estimation on embedded systems. In *2019 International Conference on Robotics and Automation (ICRA)*, pp. 6101–6108. IEEE, 2019.

Ke Xian, Jianming Zhang, Oliver Wang, Long Mai, Zhe Lin, and Zhiguo Cao. Structure-guided ranking loss for single image depth prediction. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 608–617. IEEE, 2020.

Gangwei Xu, Haotong Lin, Hongcheng Luo, Xianqi Wang, Jingfeng Yao, Lianghui Zhu, Yuechuan Pu, Cheng Chi, Haiyang Sun, Bing Wang, Guang Chen, Hangjun Ye, Sida Peng, and Xin Yang. Pixel-perfect depth with semantics-prompted diffusion transformers. In *Advances in Neural Information Processing Systems*, volume 38, pp. 193525–193549, 2025.

Feng Xue, Guirong Zhuo, Ziyuan Huang, Wufei Fu, Zhuoyue Wu, and Marcelo H. Ang. Toward hierarchical self-supervised monocular absolute depth estimation for autonomous driving applications. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 2330–2337, 2020.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024b.

Hao Yu, Haotong Lin, Jiawei Wang, Jiaxin Li, Yida Wang, Xueyang Zhang, Yue Wang, Xiaowei Zhou, Ruizhen Hu, and Sida Peng. Infinidepth: Arbitrary-resolution and fine-grained depth estimation with neural implicit fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26920–26930, 2026a.

Zhu Yu, Runmin Zhang, Lingteng Qiu, Kejie Qiu, Yisheng He, Siyu Zhu, Zilong Dong, Si-Yuan Cao, Hui-liang Shen, et al. Large depth completion model from sparse observations. In *International Conference on Learning Representations*, volume 2026, pp. 119433–119471, 2026b.

Youmin Zhang, Xianda Guo, Matteo Poggi, Zheng Zhu, Guan Huang, and Stefano Mattoccia. Completionformer: Depth completion with convolutions and vision transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *European Conference on Computer Vision*, pp. 519–535. Springer, 2020.

Ruxin Zheng, Shunqiao Sun, Hongshan Liu, and Teresa Wu. Deep-neural-network-enabled vehicle detection using high-resolution automotive radar imaging. *IEEE Transactions on Aerospace and Electronic Systems*, 59(5):4815–4830, 2023.

Yiming Zuo, Willow Yang, Zeyu Ma, and Jia Deng. Omni-dc: Highly robust depth completion with multiresolution depth integration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9287–9297, 2025.

APPENDIX

A METHODS FOR OBTAINING SPARSE GROUND TRUTH POINTS

AnchorDepth-Interp and AnchorDepth-Refine require a small set of metric depth values at image locations selected by their geometric front end. The acquisition problem is therefore to measure reliable depths in the target RGB camera frame, rather than to construct a dense ground-truth map. We represent the measurements as

$$\mathcal{A} = \{(u_i, v_i, z_i)\}_{i=1}^{N}, \qquad z_i > 0, \tag{5}$$

where (u_i, v_i) is a pixel on the processed RGB grid and z_i is the axial depth in meters. The same crop, resize, distortion model, and exposure time must be used for the RGB image and the depth measurements. Invalid, occluded, mixed-depth, dynamic, or internally interpolated values should be removed. If a sensor reports radial range r_i instead of camera-axis depth, the value should be converted with

$$\mathbf{q}_i = \mathbf{K}^{-1}[u_i, v_i, 1]^{\mathsf{T}}, \qquad z_i = \frac{r_i}{\|\mathbf{q}_i\|_2}. \tag{6}$$

A.1 DIRECT ACTIVE SENSING

LiDAR is a direct source of metric sparse points. A multi-beam scanner produces a point cloud that is transformed to the RGB camera using calibrated extrinsics and projected with the camera model. Returns near the requested pixels can then be retained after visibility, incidence-angle, and range-consistency checks. Timestamp synchronization or motion compensation is essential on moving platforms; otherwise a projected return may belong to a different scene configuration.

RGB-D cameras, time-of-flight (ToF) sensors, structured-light devices, and mobile-device LiDAR provide the same type of measurement at smaller ranges or higher image density. Their depth image should be registered to the target RGB camera and sampled only at valid sensor pixels. Per-pixel confidence, saturation flags, and the sensor's minimum and maximum reliable range are useful for rejecting unstable measurements.

When only a few specified locations are needed, a steered single-point ToF module, laser rangefinder, or pan–tilt LiDAR can measure each viewing direction directly. This option avoids unnecessary scanning, but it requires a calibrated pointing model, a rigid sensor mount, and latency compensation. The reported quantity must still be converted to the target camera's axial Z coordinate before being passed to the calibration interface.

A.2 PASSIVE MULTI-VIEW RECONSTRUCTION

A calibrated stereo or multi-camera rig obtains metric depth by triangulation. For a rectified pair, $z = fB/d$, where f is the focal length, B is the metric baseline, and d is the disparity. Only matches with adequate texture, positive depth, and low reprojection uncertainty should be exported as sparse points; occluded and ambiguous matches should be discarded. The stereo baseline must be known in metric units, otherwise the result has an unknown scale.

Structure-from-motion, visual-inertial odometry, and SLAM also produce sparse 3-D map points that can be projected into the target camera. Metric scale must be anchored by a known stereo baseline, inertial calibration, wheel odometry, GNSS, LiDAR, or a surveyed length. Points with large bundle- adjustment residuals, weak track support, or dynamic-scene inconsistency should not be used. These methods should export observed map points, not depths created by dense interpolation.

A.3 PHYSICAL TARGETS WITH KNOWN GEOMETRY

Planar coded targets such as checkerboards, dot grids, AprilTags, ArUco boards, and ChArUco boards provide accurate correspondences and a known metric scale. From the detected target corners and their surveyed dimensions, a calibrated PnP solver or a planar pose estimate gives the target plane in the camera frame. For a plane $\mathbf{n}^{\mathsf{T}}\mathbf{X} + d = 0$, the depth at a target pixel follows from

ray–plane intersection:

$$\mathbf{q} = \mathbf{K}^{-1}[u, v, 1]^{\mathsf{T}}, \qquad \mathbf{X} = \lambda \mathbf{q}, \qquad \lambda = -\frac{d}{\mathbf{n}^{\mathsf{T}}\mathbf{q}}, \qquad z = \lambda q_z. \tag{7}$$

This is a geometric measurement on the visible target plane. A target placed only for camera calibration does not provide the depth of an unrelated natural surface unless that surface and the target are connected by a rigid, measured transformation.

For non-planar layouts, surveyed 3-D landmarks, calibrated spheres, retroreflective markers, or a robot-mounted probe can provide known 3-D coordinates. Their pose, camera extrinsics, acquisition time, and survey uncertainty should be recorded. Reflective targets are especially useful when the ranging sensor must be directed to a small set of image locations or when ordinary image features are difficult to localize.

A.4 Minimal deployment protocol

In practice, the following procedure is sufficient for supplying valid sparse points:

1. select the requested locations in the original RGB coordinate system;
2. acquire RGB and range measurements synchronously, or compensate the platform motion;
3. calibrate the sensor-to-camera transform and project each measurement to the RGB grid;
4. retain the direct measurement with its confidence, timestamp, and source, without filling missing pixels by interpolation; and
5. apply visibility, range, motion, and reprojection-error checks before exporting (u_i, v_i, z_i) in meters.

B Analysis of the Proportion of Fronto-Parallel Planes

We compiled ten datasets widely used for training MDE models, Hypersim (Roberts et al., 2021), IRS (Wang et al., 2021), ARKitScenes (Baruch et al., 2021), Structured3D (Zheng et al., 2020), Virtual KITTI 2 (Cabon et al., 2020), UrbanSyn (Gómez et al., 2025), MVS-Synth (Huang et al., 2018), UnrealStereo4K (Tosi et al., 2021), TartanAir (Wang et al., 2020), HRWSI (Xian et al., 2020). From each dataset, we randomly sampled ten images and manually annotated the regions parallel to the camera plane. We then calculated the proportion of these regions relative to the total number of pixels, with the detailed results presented in the Table A1. It can be observed that these regions account for only a very small fraction of the overall pixels.

Table A1: The ratio of total pixels belonging to front-parallel planes.

Dataset	Hypersim	TartanAir	IRS	Virtual KITTI 2	UrbanSyn	UnrealStereo4K
Average	1.11%	0.007%	1.55%	0.20%	0.33%	1.30%
Dataset	ARKitScenes	Structured3D	MVS–Synth	HRWSI	**Total (100 frames)**	
Average	2.24%	3.82%	0.037%	0.023%	**1.06%**	

C Method Details

All fixed hyperparameters in this section are set before evaluation and shared across images and datasets; image-adaptive quantities are computed only by the rules stated below, and no test image is used for tuning. The following tables collect the implementation choices needed to reproduce the protocol. The main text gives the compact definitions of the two calibration variants, while this appendix provides their full numerical forms and safeguards.

C.1 ANCHOR ALLOCATION AND SAFEGUARDS

This subsection specifies the deterministic anchor budget and the fallbacks used when plane-wise calibration is ill-conditioned. Let K be the non-negative integer number of selected plane regions and let M be the positive integer number of metric anchors. The protocol considers the four largest candidate planes in area order. A retained plane must contain at least five valid metric locations; the same anchors are passed to AnchorDepth-Interp and AnchorDepth-Refine.

Table A2: Fixed anchor and numerical-safeguard policy.

Item	Fixed policy
Plane anchors	Five valid locations without replacement per selected plane.
Non-planar anchors	At most ten locations from the remaining valid domain.
Budget	Five anchors per plane plus at most ten non-planar anchors; $K \leq 4$ and $M \leq 30$.
Sampling	Per-image pseudorandom seed 0; non-planar sampling follows plane sampling.
Plane fit	Depth-domain ordinary least squares.

For plane P_k, let $\alpha_k, \beta_k \in \mathbb{R}$ be its fitted scale and bias, and let $\alpha, \beta \in \mathbb{R}$ be the optimization variables. Using the baseline $\mathbf{Z}_0$ and the measurements $(\mathbf{x}_i, z_i)$ defined in Section 3.2, the plane fit is

$$(\alpha_k, \beta_k) = \arg\min_{\alpha,\beta\in\mathbb{R}} \sum_{i:\,\mathbf{x}_i\in P_k} \big(\alpha \mathbf{Z}_0(\mathbf{x}_i) + \beta - z_i\big)^2. \tag{8}$$

The calibrated depth at $\mathbf{x} \in P_k$ is $\alpha_k \mathbf{Z}_0(\mathbf{x}) + \beta_k$. The safeguards below are applied after this fit.

If a plane fit is rank deficient or has a non-positive scale, we use the positive scale-only least-squares fallback. For the affine depth-calibration parameters $\alpha_k, \beta_k \in \mathbb{R}$,

$$\alpha_k = \frac{\sum_{i:\,\mathbf{x}_i\in P_k} \mathbf{Z}_0(\mathbf{x}_i) z_i}{\sum_{i:\,\mathbf{x}_i\in P_k} \mathbf{Z}_0(\mathbf{x}_i)^2}, \qquad \beta_k = 0. \tag{9}$$

Non-positive affine predictions retain $\mathbf{Z}_0$. With no valid non-planar anchor, AnchorDepth-Interp retains $\mathbf{Z}_0$ outside the selected planes; AnchorDepth-Refine returns the available plane-corrected estimate when its global solve is unconstrained.

C.2 PLANE EXTRACTION PARAMETERS

This subsection gives the fixed processing choices for identifying candidate fronto-parallel regions. Let $\mathbf{D}_{\text{seg}} \in \mathbb{R}^{H\times W}$ be the model-specific segmentation signal, let $\mathbf{Z}_0 \in \mathbb{R}^{H\times W}_{>0}$ be the metric baseline, and let $\mathbf{g} \in \mathbb{R}^{H\times W}$ be the inverse-depth representation used in Section 3.3. The signal is the native disparity-like output when available and $\mathbf{Z}_0$ otherwise. To obtain $\mathbf{g}$, we retain a disparity-like signal and invert a depth signal, as described in Section 3.3.

Let $p_q \in \mathbb{R}$ denote the qth percentile of valid entries of $\mathbf{g}$, let $\text{MAD} \in \mathbb{R}_{\geq 0}$ denote the median absolute deviation, let $S \in \mathbb{R}_{>0}$ be the robust image-level inverse-depth span, and let $L_{\text{ref}} \in \mathbb{R}_{>0}$ be the image-relative reference length. For candidate R, let $e_R \in \mathbb{R}_{\geq 0}$ be its affine-fit RMS residual, $\mathbf{g}_R \in \mathbb{R}^{|R|}$ its values, and let $a_{R,1}, a_{R,2} \in \mathbb{R}$ be the first two components of the affine fit vector $\mathbf{a}_R \in \mathbb{R}^3$. Let $\eta_{\text{fit}}(R), \eta_{\text{fronto}}(R), \eta_{\text{purity}}(R) \in \mathbb{R}_{\geq 0}$ denote the normalized fit, fronto-parallel, and purity scores, respectively:

$$\eta_{\text{fit}}(R) = \frac{e_R}{S}, \qquad \eta_{\text{fronto}}(R) = \frac{\sqrt{a_{R,1}^2 + a_{R,2}^2}\, L_{\text{ref}}}{S}, \qquad \eta_{\text{purity}}(R) = \frac{1.4826\ \text{MAD}(\mathbf{g}_R)}{S}. \tag{10}$$

The default protocol retains only approximately fronto-parallel candidates. The all-plane variant is used only in the supplementary analysis.

Table A3: Fixed parameters for model-specific plane extraction.

Stage	Value	Purpose
Value range	p_1–p_{99}; far rejection 0.02 for disparity and 0 for depth	Remove extreme values.
Bilateral filter	Diameter 7; range 0.05; spatial 7	Smooth $\mathbf{g}$ while preserving boundaries.
Depth cliff	Relative gradient 0.04; one-pixel dilation	Obtain candidate boundaries.
Depth crease	Gaussian $\sigma = 3$; denominator floor 0.10 median; p_{96} and $4\times$ median cutoff	Used only for depth-form outputs.
Candidate size	At least 600 pixels and 0.004 of valid area	Reject small regions.
Affine normalization	$S = \max(p_{99} - p_{01}, 3 \cdot 1.4826\ \mathrm{MAD}(\mathbf{g}))$	Make scores scale-normalized.
Region gates	$\eta_{\mathrm{fit}} \le 0.025$, $\eta_{\mathrm{fronto}} \le 0.05$, $\eta_{\mathrm{purity}} \le 0.030$; $L_{\mathrm{ref}} = 0.25$ times the shorter side	Select approximately fronto-parallel planes.
Mask post-processing	Five-pixel elliptical closing	Regularize retained masks.

C.3 Residual propagation parameters

This subsection gives the solver-grid implementation and numerical settings omitted from the main text. Equations equation 2 and equation 3 define the interpolation and edge weights, while Eq. equation 11 gives the grid-level refinement.

For each selected plane, we compute the pointwise log difference between its affine-corrected depth and $\mathbf{Z}_0$, and use the median difference as the plane marker. Each remaining measurement contributes its own log residual as a singleton marker. Let $\mathbf{r}_{\mathrm{p}} \in \mathbb{R}^{H\times W}$ denote the piecewise-constant prior obtained by spreading these markers with the watershed.

Let Ω_{s} be the solver grid and $\mathcal{E}_{\mathrm{s}} \subseteq \Omega_{\mathrm{s}} \times \Omega_{\mathrm{s}}$ its four-neighbor edges. For each $(\mathbf{x}, \mathbf{y}) \in \mathcal{E}_{\mathrm{s}}$, let $w_{\mathbf{xy}} \in \mathbb{R}_{>0}$ denote its propagation weight. Let $\mathbf{t}_{\mathrm{s}}, \boldsymbol{\ell}_{\mathrm{s}} \in \mathbb{R}^{|\Omega_{\mathrm{s}}|}$ be the prior and log-depth guide sampled on this grid, $\mathbf{q} \in \mathbb{R}^{|\Omega_{\mathrm{s}}|}$ the unknown residual, and $\lambda \in \mathbb{R}_{\ge 0}$ the prior-fidelity coefficient. Let π map image coordinates to the solver grid and let $\widehat{\mathbf{q}} \in \mathbb{R}^{|\Omega_{\mathrm{s}}|}$ denote the solved residual. The refinement objective is

$$\widehat{\mathbf{q}} = \arg\min_{\mathbf{q}\in\mathbb{R}^{|\Omega_{\mathrm{s}}|}} \left[\sum_{(\mathbf{x},\mathbf{y})\in\mathcal{E}_{\mathrm{s}}} w_{\mathbf{xy}} \big(\mathbf{q}(\mathbf{x}) - \mathbf{q}(\mathbf{y})\big)^2 + \lambda \sum_{\mathbf{x}\in\Omega_{\mathrm{s}}} \big(\mathbf{q}(\mathbf{x}) - \mathbf{t}_{\mathrm{s}}(\mathbf{x})\big)^2 \right] \quad \text{s.t.} \quad \mathbf{q}(\pi(\mathbf{x}_i)) = \rho_i, \qquad i = 1, \dots, M. \tag{11}$$

On the solver grid, Eq. equation 3 is evaluated using the sampled guide $\boldsymbol{\ell}_{\mathrm{s}}$.

Let Ω_v denote the valid image domain, $\boldsymbol{\ell} \in \mathbb{R}^{H\times W}$ the full-resolution log-depth guide, and $c(\mathbf{x}) \in \mathbb{R}_{>0}$ the watershed elevation at $\mathbf{x} \in \Omega_v$. The vector $\nabla_{\mathrm{Sobel}}\boldsymbol{\ell}(\mathbf{x}) \in \mathbb{R}^2$ is its Sobel gradient.

D Calibration with Denser Sparse Measurements

This section introduces AnchorDepth-Region, a supplementary variant that uses slightly more USMM to improve coverage while retaining the same metric baseline, plane calibration, and log-residual representation defined in Section 3.2. The plane candidates and plane-fit safeguards follow Appendix C.2 and Appendix C.1; AnchorDepth-Region changes only the allocation of non-planar measurements by replacing the ultra-sparse anchor budget with segment-level coverage sampling.

Table A4: Fixed parameters for residual propagation and global refinement.

Component	Fixed policy
AnchorDepth-Interp	IDW stabilizer $\epsilon_d = 1$ in squared-pixel units.
Watershed elevation	$c(\mathbf{x}) = 0.02 + 6\|\nabla_{\mathrm{Sobel}}\boldsymbol{\ell}(\mathbf{x})\|_2$.
Screening	$\lambda = 0.05$ in Eq. equation 11.
Edge weights	Floor $w_{\min} = 0.05$; σ is the 75th percentile of sorted-guide differences, floored at 10^{-3}.
Solver grid	Native valid grid up to 50,000 pixels; larger grids are resized to this cap by nearest-neighbor sampling.
Constraints and transfer	Anchor coordinates are mapped to the solver grid before Dirichlet constraints; the solved residual is bilinearly transferred back.
Non-finite solve	Retain the corresponding baseline depth $\mathbf{Z}_0$.

D.1 AnchorDepth-Region: Segment-Aware Coverage Sampling

AnchorDepth-Region distributes the additional non-planar measurements across image regions. Let $\mathcal{S} = \{S_j\}_{j=1}^{J}$ be the SLIC superpixel partition (Achanta et al., 2012) of the non-planar image area, with J segments, where J is a positive integer. The labels are computed once and kept fixed for all MDE models. Let $\mathcal{J}_{\mathrm{cand}} \subseteq \{1, \dots, J\}$ denote the segments containing at least one valid metric measurement. We rank these segments by their binary candidate-support count, retain at most fifty, and denote the retained set by $\mathcal{J}_{\mathrm{obs}} \subseteq \mathcal{J}_{\mathrm{cand}}$; one valid measurement is sampled from each retained segment.

For every segment, let $\boldsymbol{\mu}_j \in \mathbb{R}^2$ be its centroid and let $\rho_j \in \mathbb{R}$ denote its segment residual. For an observed segment $j \in \mathcal{J}_{\mathrm{obs}}$, let $(\mathbf{x}_j, z_j) \in \mathbb{R}^2 \times \mathbb{R}_{>0}$ denote its sampled measurement. Its residual is

$$\rho_j = \log z_j - \log \mathbf{Z}_0(\mathbf{x}_j), \qquad j \in \mathcal{J}_{\mathrm{obs}}. \tag{12}$$

An unobserved segment receives an inverse-distance interpolation of the observed segment residuals. Let $\widetilde{\omega}_{jk} \in \mathbb{R}_{>0}$ and $\omega_{jk} \in \mathbb{R}_{\geq 0}$ denote the unnormalized and normalized weights from observed segment $k \in \mathcal{J}_{\mathrm{obs}}$ to unobserved segment $j \notin \mathcal{J}_{\mathrm{obs}}$, respectively. Let $\ell \in \mathcal{J}_{\mathrm{obs}}$ be a summation index and let $\epsilon_d \in \mathbb{R}_{>0}$ be the same distance stabilizer as in Eq. equation 2. The interpolated residual and its weights are

$$\rho_j = \sum_{k \in \mathcal{J}_{\mathrm{obs}}} \omega_{jk}\rho_k, \qquad \omega_{jk} = \frac{\widetilde{\omega}_{jk}}{\sum_{\ell \in \mathcal{J}_{\mathrm{obs}}} \widetilde{\omega}_{j\ell}}, \quad \widetilde{\omega}_{jk} = \max\{\|\boldsymbol{\mu}_j - \boldsymbol{\mu}_k\|_2^2, \epsilon_d\}^{-1}. \tag{13}$$

Its fixed implementation value is given in Appendix C.3. The resulting piecewise-constant residual field $\widehat{\mathbf{r}}_{\mathrm{AnchorDepth\text{-}Region}} \in \mathbb{R}^{H \times W}$ is

$$\widehat{\mathbf{r}}_{\mathrm{AnchorDepth\text{-}Region}}(\mathbf{x}) = \rho_j, \qquad \mathbf{x} \in S_j. \tag{14}$$

Applying this field to the baseline with the log-residual decoding rule in Section 3.2 gives the calibrated non-planar depth. The plane estimates from the same section are retained on the selected regions.

E Datasets Details

E.1 STREET

In the Unreal Engine 5 City Sample scene, we utilized the Movie Render Queue to offline render city depth data along a predefined camera trajectory. This scene shares the same city assets as The Matrix Awakens: An Unreal Engine 5 Experience. The urban layout, roads, and buildings were procedurally generated in Houdini; after the point clouds were imported into the engine, they were converted into instances by rule processors. The static meshes primarily leverage Nanite, while global illumination and reflections are computed by Lumen. Direct lighting is provided by directional lights, skylights, and emissive surfaces. The camera was captured along a continuous trajectory, maintaining a pitch of approximately -15 degree and a roll near zero (the negative sign indicates a downward tilt). For

Table A5: Method inference time. The image resolution is 560×672. The time costs of RANSAC-Affine, AnchorDepth-Interp , AnchorDepth-Refine and RANSAC-Affine consist of the inference time of the base depth estimation model and their own computational time.

Method	Inference time per frame(s)
PromptDA (Lin et al., 2025)	0.297
PriorDA (Wang et al., 2026)	0.462
OMNI-DC (Zuo et al., 2025)	1.256
Marigold-DC (Viola et al., 2025)	15.522
RANSAC-Affine (Marsal et al., 2025)	0.260+0.011
LDCM (Yu et al., 2026b)	0.790
InfiniDepth (Yu et al., 2026a)	0.325
MoGe-2 (Wang et al., 2025)	0.260
AnchorDepth-Interp	0.260+0.130
AnchorDepth-Refine	0.260+0.388

each frame, we saved an sRGB color image alongside a linear OpenEXR depth map aligned with the same viewpoint. We also recorded the horizontal field of view (FOV), focal length, and the camera's position and orientation in world coordinates. The three columns of the rotation matrix correspond to the camera's right, up, and backward vectors, respectively, with translation expressed in meters. Depth is defined as the axial distance along the camera's forward viewing axis, measured in centimeters. Pixels corresponding to the sky or exceeding the representable range are marked in an invalid mask and are not used as depth ground truth. For every camera vantage point, we simulated both daytime and nighttime conditions (denoted as STREET-day and STREET-dark, respectively). Given the rendering accuracy of the physics engine, we limited the depth evaluation range to 2,000 meters. Dataset examples are shown in Figure A1.

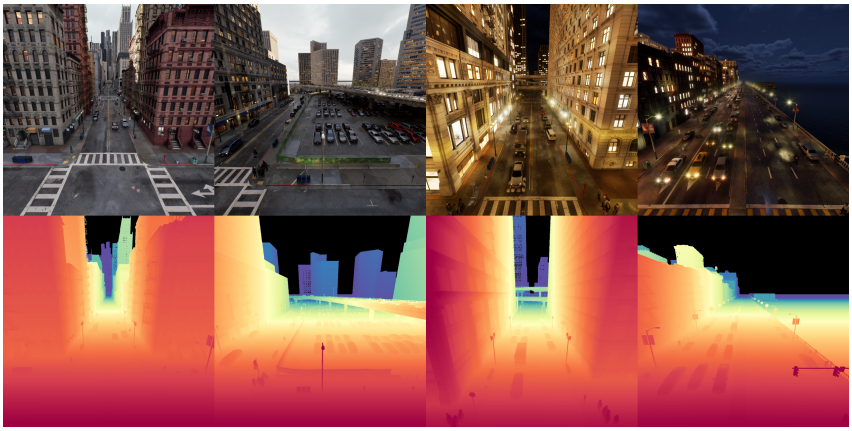

Figure A1: Examples of RGB and Depth Data from the STREET Dataset.

E.2 FAR

FAR is a real-world, long-range dataset. To capture data at the greatest possible distance while balancing equipment costs, we opted for a 3D laser scanner rather than LiDAR. During the data collection phase, we made every effort to use the longest-range 3D laser scanner available to us. We selected typical urban scenes, which generally feature regular structures and offer clear lines of sight over extended distances. Data collection was conducted during the day under favorable lighting conditions, resulting in a total of 500 RGB-D pairs. The maximum effective depth in the dataset is approximately 730 meters. Dataset examples are shown in Figure A2.

F Detailed monocular depth estimation calibration results

As shown in tables A6 to A13, we report the performance of Depth Pro, UniDepth V2, MoGe-2, DA3, DA2, Marigold, PPD, and InfiniDepth on their respective datasets, as well as the performance

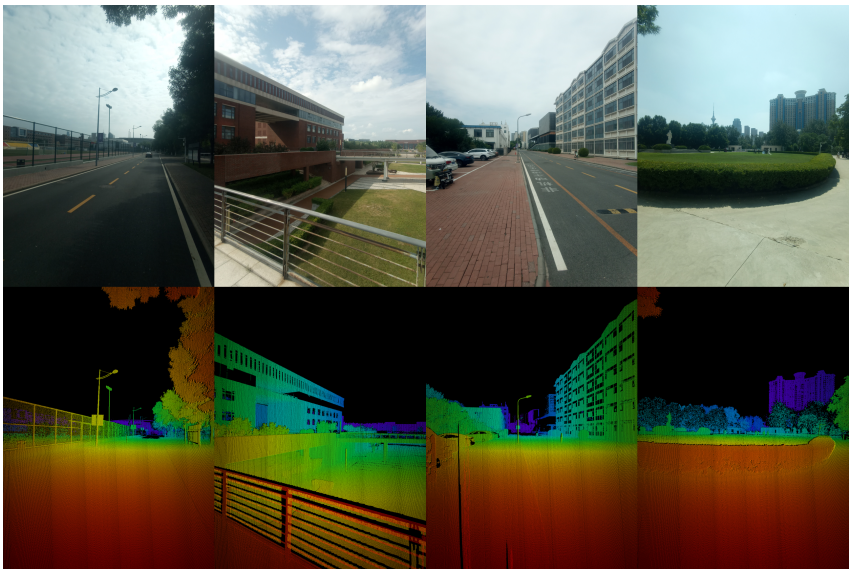

Figure A2: Examples of RGB and Depth Data from the FAR Dataset. We appropriately scaled up the valid measurement points in the depth maps to enhance visualization.

Table A6: Depth Pro results before and after three calibration schemes. All entries use official-known evidence and are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
Booster	Depth Pro	0.342	0.356	0.375	0.281	0.119	51.565	78.107	92.003	4.532
	AnchorDepth-Interp	0.022	0.025	0.039	0.034	0.010	99.696	**99.954**	99.970	3.184
	AnchorDepth-Refine	0.026	0.032	0.047	0.037	0.011	99.257	99.633	99.971	3.226
	AnchorDepth-Region	**0.014**	**0.016**	**0.033**	**0.027**	**0.006**	**99.740**	99.952	**99.980**	**2.653**
Middlebury	Depth Pro	0.161	0.442	0.504	0.174	0.070	75.131	97.260	99.982	7.511
	AnchorDepth-Interp	0.038	0.100	0.149	0.055	0.016	98.793	99.946	99.995	5.105
	AnchorDepth-Refine	0.036	0.095	0.145	0.053	0.016	98.794	**99.976**	99.995	4.848
	AnchorDepth-Region	**0.023**	**0.063**	**0.119**	**0.042**	**0.010**	**99.011**	99.959	**99.996**	**4.140**
NuScenes	Depth Pro	0.314	5.013	14.237	0.463	0.161	45.541	73.852	86.458	28.785
	AnchorDepth-Interp	0.217	3.057	**13.565**	**0.293**	0.073	79.302	90.146	**94.508**	**28.208**
	AnchorDepth-Refine	0.259	3.302	14.895	0.309	0.075	80.245	89.522	93.427	29.047
	AnchorDepth-Region	**0.200**	**3.012**	18.469	0.301	**0.061**	**83.832**	**90.898**	94.364	29.882
Sintel	Depth Pro	1.155	10.955	53.380	0.742	0.289	32.159	49.700	63.374	29.888
	AnchorDepth-Interp	0.238	3.682	37.156	0.238	0.062	82.985	92.185	95.667	22.606
	AnchorDepth-Refine	0.203	3.484	**35.916**	0.200	0.047	88.851	95.004	97.217	19.248
	AnchorDepth-Region	**0.125**	**1.940**	52.715	**0.187**	**0.038**	**90.521**	**95.678**	**97.623**	**18.585**
ETH3D	Depth Pro	0.378	3.249	3.648	0.562	0.240	33.374	55.133	67.917	8.265
	AnchorDepth-Interp	0.049	0.291	0.979	0.067	0.018	97.693	99.102	99.510	6.401
	AnchorDepth-Refine	0.054	0.299	1.017	0.072	0.019	97.722	98.965	99.324	6.581
	AnchorDepth-Region	**0.033**	**0.226**	**0.577**	**0.058**	**0.013**	**98.080**	**99.155**	**99.524**	**5.683**
SUN RGB-D	Depth Pro	0.144	0.321	0.396	0.159	0.060	81.111	98.019	99.553	8.063
	AnchorDepth-Interp	0.044	0.102	**0.199**	**0.078**	0.019	97.199	**99.214**	**99.714**	**7.487**
	AnchorDepth-Refine	0.043	0.100	**0.199**	**0.078**	0.018	**97.378**	99.120	99.623	7.517
	AnchorDepth-Region	**0.035**	**0.086**	0.553	**0.078**	**0.015**	97.230	99.066	99.639	7.708

of AnchorDepth when each method serves as the foundation model. The results demonstrate that AnchorDepth can significantly enhance depth estimation performance by simply employing USMM.

G Experimental Details

This section specifies the preprocessing, sparse-input construction, and metric aggregation used in the sparse-guidance and measurement-error experiments. Because the public benchmarks, STREET, and FAR differed in resolution and valid depth support, each group used a separate fixed protocol. All competing methods within a group received the same frozen sparse measurements.

Table A7: UniDepth v2 results before and after three calibration schemes. All values are rounded to three decimal places.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
SUN RGB-D	UniDepth v2	0.078	0.169	0.243	0.100	0.032	95.925	99.208	99.672	7.022
	AnchorDepth-Interp	0.036	0.081	**0.175**	**0.070**	0.015	97.762	**99.304**	**99.731**	**6.735**
	AnchorDepth-Refine	0.036	0.080	0.178	0.071	0.015	**97.840**	99.112	99.589	6.853
	AnchorDepth-Region	**0.030**	**0.070**	0.576	0.072	**0.012**	97.651	99.131	99.647	7.201
iBims-1	UniDepth v2	0.085	0.262	0.352	0.103	0.036	93.324	99.480	99.829	6.170
	AnchorDepth-Interp	0.032	0.112	0.237	0.061	0.013	98.133	99.391	99.815	5.926
	AnchorDepth-Refine	**0.026**	**0.095**	**0.223**	**0.057**	**0.011**	**98.506**	**99.508**	**99.832**	**5.596**
	AnchorDepth-Region	0.027	0.106	0.279	0.068	**0.011**	97.578	99.215	99.737	6.715
ETH3D	UniDepth v2	0.215	1.242	1.532	0.199	0.082	67.663	95.059	98.699	5.205
	AnchorDepth-Interp	0.026	0.162	0.526	0.046	0.011	**99.262**	**99.808**	**99.898**	4.502
	AnchorDepth-Refine	0.027	0.174	**0.509**	0.052	0.012	99.048	99.582	99.734	4.963
	AnchorDepth-Region	**0.020**	**0.152**	0.586	**0.044**	**0.008**	99.174	99.719	99.853	**4.406**
Sintel	UniDepth v2	0.596	18.342	27.037	0.650	0.249	34.570	54.668	68.345	32.163
	AnchorDepth-Interp	0.173	4.231	10.975	0.247	0.065	80.079	91.451	95.711	23.660
	AnchorDepth-Refine	0.124	3.189	**9.704**	**0.188**	0.044	89.789	**95.534**	**97.422**	**18.227**
	AnchorDepth-Region	**0.120**	**3.030**	31.599	0.204	**0.039**	**90.258**	95.208	97.149	20.225

Table A8: MoGe-2 results before and after three calibration schemes. All values are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
KITTI	MoGe-2	0.176	3.105	4.355	0.211	0.086	64.696	96.454	99.427	7.125
	AnchorDepth-Interp	0.040	0.911	**2.469**	**0.071**	0.017	**97.975**	**99.575**	**99.865**	**6.868**
	AnchorDepth-Refine	0.043	0.970	2.548	0.074	0.018	97.920	99.550	**99.865**	6.946
	AnchorDepth-Region	**0.036**	**0.881**	2.755	0.075	**0.015**	97.763	99.393	99.788	7.403
NYUv2	MoGe-2	0.069	0.170	0.231	0.090	0.030	96.740	99.394	99.808	6.109
	AnchorDepth-Interp	0.032	0.080	0.156	0.057	0.014	98.501	99.601	99.859	5.520
	AnchorDepth-Refine	0.031	0.079	**0.151**	0.056	0.013	98.447	**99.638**	**99.876**	**5.407**
	AnchorDepth-Region	**0.027**	**0.072**	0.190	**0.055**	**0.011**	**98.600**	99.584	99.843	5.491
ETH3D	MoGe-2	0.100	0.573	0.743	0.113	0.044	88.798	99.455	99.900	5.260
	AnchorDepth-Interp	0.027	0.193	0.385	0.046	0.012	98.974	99.807	99.910	4.405
	AnchorDepth-Refine	0.026	**0.176**	**0.365**	0.046	0.011	98.935	**99.828**	**99.912**	4.268
	AnchorDepth-Region	**0.021**	0.183	5.290	**0.041**	**0.008**	**99.194**	99.729	99.841	**4.060**
iBims-1	MoGe-2	0.146	0.465	0.549	0.156	0.062	80.459	98.481	98.841	5.655
	AnchorDepth-Interp	0.027	0.094	0.211	0.054	0.012	98.692	99.656	99.889	5.246
	AnchorDepth-Refine	0.023	**0.086**	**0.207**	**0.050**	0.010	**98.838**	**99.657**	**99.892**	**4.892**
	AnchorDepth-Region	**0.022**	0.088	0.255	0.055	**0.009**	98.535	99.529	99.790	5.454
DDAD	MoGe-2	0.156	4.662	9.081	0.216	0.070	73.653	94.890	98.623	**15.640**
	AnchorDepth-Interp	0.102	2.915	**7.483**	0.167	0.041	90.187	**96.947**	**98.644**	16.037
	AnchorDepth-Refine	0.099	**2.862**	7.649	**0.165**	0.038	**91.609**	96.926	98.497	15.762
	AnchorDepth-Region	**0.096**	3.091	27.767	0.180	**0.036**	91.139	96.403	98.166	17.888
DIODE	MoGe-2	0.167	2.040	2.898	0.208	0.081	67.850	94.276	98.373	8.465
	AnchorDepth-Interp	0.034	0.500	1.483	0.067	0.014	97.759	99.196	**99.596**	6.483
	AnchorDepth-Refine	0.029	0.470	**1.476**	0.066	0.013	98.294	99.122	99.371	6.260
	AnchorDepth-Region	**0.024**	**0.428**	2.617	**0.060**	**0.009**	**98.390**	**99.287**	99.581	**5.980**
HAMMER	MoGe-2	0.260	0.196	0.207	0.236	0.099	65.287	89.362	94.369	5.321
	AnchorDepth-Interp	0.027	0.023	0.038	0.043	0.012	99.517	**99.948**	**99.988**	4.102
	AnchorDepth-Refine	0.036	0.030	0.047	0.053	0.015	99.179	99.946	99.987	4.712
	AnchorDepth-Region	**0.017**	**0.015**	**0.031**	**0.034**	**0.007**	**99.559**	99.927	99.980	**3.381**

G.1 Dataset-Specific Evaluation Protocols

Public benchmarks. The sparse-guidance comparison used the official evaluation samples from five datasets: 652 KITTI Eigen images, 454 ETH3D images, 654 NYUv2 test images, 800 fixed ScanNet validation images, and 325 DIODE indoor validation images. We followed the established valid-depth ranges and masks: $(10^{-3}, 80)$ m with the benchmark and Eigen masks for KITTI, $(0.1, 200)$ m for ETH3D, $(10^{-3}, 10)$ m with the Eigen mask and raw sensor depth for NYUv2, $(10^{-3}, 10)$ m for ScanNet, and $(10^{-3}, 300)$ m intersected with the native depth mask for DIODE. No additional crop was applied outside these masks.

All five datasets were mapped to a common $H \times W = 560 \times 672$ grid by aspect-ratio-preserving resizing and centered letterboxing. RGB downsampling used area interpolation. To prevent invalid depth values from contaminating the resized ground truth, valid-weighted depth and the validity weights were resized separately, while the binary evaluation mask used nearest-neighbor sampling.

Table A9: Depth Anything 3 results before and after three calibration schemes. All values are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
KITTI	DA3 metric	0.097	1.950	3.627	0.136	0.045	91.481	98.487	99.606	10.347
	AnchorDepth-Interp	0.061	1.283	**2.945**	**0.101**	0.026	95.251	**99.078**	**99.731**	**9.680**
	AnchorDepth-Refine	0.078	1.569	3.484	0.116	0.032	93.786	98.263	99.395	10.612
	AnchorDepth-Region	**0.055**	**1.229**	8.271	0.103	**0.023**	**95.395**	98.755	99.571	10.203
NYUv2	DA3 metric	0.073	0.201	0.288	0.100	0.033	96.498	99.068	99.498	7.470
	AnchorDepth-Interp	0.040	0.105	**0.193**	0.068	0.017	97.807	99.453	**99.802**	6.496
	AnchorDepth-Refine	0.045	0.113	0.208	0.073	0.018	97.502	98.941	99.414	6.886
	AnchorDepth-Region	**0.031**	**0.089**	0.529	**0.063**	**0.013**	**98.114**	**99.457**	99.792	**6.291**
ETH3D	DA3 metric	0.114	0.798	1.021	0.136	0.053	87.645	96.956	99.809	6.067
	AnchorDepth-Interp	0.032	0.224	**0.440**	0.054	0.014	98.888	**99.744**	**99.887**	5.165
	AnchorDepth-Refine	0.036	0.240	0.459	0.058	0.015	98.611	99.689	99.851	5.468
	AnchorDepth-Region	**0.025**	**0.190**	0.443	**0.049**	**0.011**	**98.889**	99.717	99.872	**4.880**
DIODE	DA3 metric	0.325	1.444	1.631	0.427	0.178	12.502	67.527	91.572	11.801
	AnchorDepth-Interp	0.060	0.203	0.425	0.112	0.025	94.325	98.261	**99.290**	10.733
	AnchorDepth-Refine	0.058	0.201	**0.419**	0.110	0.023	**95.301**	98.160	99.227	**10.497**
	AnchorDepth-Region	**0.045**	**0.160**	0.529	**0.109**	**0.018**	95.271	**98.293**	99.232	10.825
SUN RGB-D	DA3 metric	0.079	0.186	0.275	0.108	0.035	95.571	98.997	99.579	8.099
	AnchorDepth-Interp	0.045	0.102	**0.199**	0.079	0.019	97.221	**99.195**	**99.698**	**7.528**
	AnchorDepth-Refine	0.046	0.105	0.208	0.081	0.019	97.207	98.956	99.501	7.775
	AnchorDepth-Region	**0.035**	**0.083**	0.319	**0.077**	**0.014**	**97.274**	99.079	99.634	7.694

Table A10: Depth Anything V2 results before and after three calibration schemes. All values are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
KITTI	DA2	0.096	2.045	4.215	0.135	0.041	91.459	98.121	99.324	12.499
	AnchorDepth-Interp	0.076	1.611	**3.708**	0.118	0.032	93.739	98.485	99.469	11.213
	AnchorDepth-Refine	0.091	1.884	4.033	0.131	0.038	92.410	98.118	99.251	11.867
	AnchorDepth-Region	**0.055**	**1.242**	6.641	**0.102**	**0.023**	**95.552**	**98.892**	**99.627**	**10.121**
NYUv2	DA2	0.050	0.139	0.233	0.075	0.021	97.114	99.129	99.705	7.222
	AnchorDepth-Interp	0.041	0.114	**0.215**	0.068	0.017	97.493	99.361	**99.819**	6.550
	AnchorDepth-Refine	0.044	0.121	0.231	0.072	0.018	97.570	99.104	99.558	6.835
	AnchorDepth-Region	**0.031**	**0.093**	1.274	**0.062**	**0.012**	**98.253**	**99.495**	99.790	**6.119**
Sintel	DA2	0.284	2.717	5.914	0.332	0.098	70.951	86.898	93.059	31.077
	AnchorDepth-Interp	0.226	2.180	**5.292**	0.296	0.080	77.317	88.840	93.894	28.218
	AnchorDepth-Refine	0.219	2.139	5.370	0.285	0.073	80.715	90.556	94.758	26.940
	AnchorDepth-Region	**0.115**	**1.181**	5.831	**0.203**	**0.044**	**88.719**	**94.764**	**97.097**	**20.022**
ETH3D	DA2	0.057	0.485	0.948	0.082	0.025	96.180	98.674	99.316	7.384
	AnchorDepth-Interp	0.043	0.351	0.831	0.070	0.018	97.201	99.070	99.614	6.645
	AnchorDepth-Refine	0.047	0.386	0.867	0.074	0.020	96.923	99.150	99.618	6.833
	AnchorDepth-Region	**0.028**	**0.259**	**0.577**	**0.053**	**0.012**	**98.476**	**99.616**	**99.835**	**5.193**
DIODE	DA2	0.071	0.267	0.524	0.123	0.031	93.270	97.977	99.148	11.609
	AnchorDepth-Interp	0.059	0.225	**0.505**	0.118	0.025	94.107	97.981	99.152	11.302
	AnchorDepth-Refine	0.057	0.224	0.508	0.115	0.024	94.742	98.054	99.146	11.151
	AnchorDepth-Region	**0.046**	**0.187**	1.399	**0.109**	**0.018**	**95.148**	**98.229**	**99.210**	**10.797**

RGB padding replicated boundary values, whereas depth and mask padding remained invalid. Metrics were computed on this common grid.

For the fixed condition, we ran MoGe-2 on the common-grid RGB image and formed the sparse measurements from at most four extracted planes, using five valid points per plane and at most ten additional non-planar points. This produced 10–30 measurements per image. The random condition used the same per-image measurement count, sampled uniformly without replacement from valid pixels after excluding the fixed coordinates. The frozen fixed and random measurement sets were shared by all competing methods. The average measurement counts were 18.08, 16.92, 16.93, 12.98, and 15.17 for KITTI, ETH3D, NYUv2, ScanNet, and DIODE, respectively.

STREET. Each STREET split contained 6,000 paired RGB and depth images at $H \times W = 1000 \times 1000$ resolution. Metric depth was obtained from the first EXR channel after division by 100, and valid pixels satisfied $10^{-3} < d < 2000$ m. RGB images were resized to an external $H \times W = 512 \times 512$ grid with area interpolation, without cropping or padding. Ground-truth depth and its native validity mask were transferred to this grid with nearest-neighbor sampling, after which the validity range was reapplied. A new MoGe-2-based measurement set was generated on

Table A11: Marigold results before and after three calibration schemes. All values are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
KITTI	Marigold	0.177	2.632	3.981	0.337	0.091	73.056	91.877	96.802	29.513
	AnchorDepth-Interp	0.129	2.098	3.793	0.280	0.066	83.436	95.168	98.052	26.380
	AnchorDepth-Refine	0.136	2.400	4.230	0.281	0.068	81.284	94.734	97.980	25.844
	AnchorDepth-Region	**0.065**	**1.377**	**3.259**	**0.116**	**0.028**	**94.187**	**98.556**	**99.552**	**11.480**
NYUv2	Marigold	0.071	0.167	0.248	0.100	0.030	94.042	98.534	99.494	9.423
	AnchorDepth-Interp	0.052	0.128	0.225	0.086	0.022	96.205	99.007	99.628	8.270
	AnchorDepth-Refine	0.051	0.127	0.221	0.084	0.022	96.414	98.927	99.587	8.078
	AnchorDepth-Region	**0.035**	**0.098**	**0.213**	**0.071**	**0.015**	**97.629**	**99.312**	**99.728**	**7.025**
ETH3D	Marigold	0.117	0.698	0.991	0.156	0.049	87.453	94.839	97.740	13.589
	AnchorDepth-Interp	**0.068**	**0.425**	**0.760**	0.118	0.030	93.890	98.084	99.216	11.364
	AnchorDepth-Refine	0.074	0.492	0.829	0.126	0.034	93.006	97.603	98.984	11.564
	AnchorDepth-Region	0.070	0.508	13.749	**0.084**	**0.019**	**96.736**	**98.858**	**99.391**	**8.205**
DIODE	Marigold	0.386	2.875	**4.346**	0.375	0.115	70.233	83.346	89.403	34.072
	AnchorDepth-Interp	0.298	2.522	4.572	0.322	0.087	76.664	86.793	91.616	30.764
	AnchorDepth-Refine	0.331	2.540	4.692	0.317	0.082	79.414	88.470	92.419	30.020
	AnchorDepth-Region	**0.239**	**2.214**	5.093	**0.304**	**0.071**	**81.428**	**89.611**	**93.382**	**29.960**
ScanNet	Marigold	0.082	0.130	0.188	0.114	0.034	92.785	98.139	99.436	10.656
	AnchorDepth-Interp	0.059	0.097	0.171	0.098	0.025	95.117	98.672	99.518	9.385
	AnchorDepth-Refine	0.059	0.099	0.169	0.097	0.025	95.269	98.642	99.494	9.304
	AnchorDepth-Region	**0.039**	**0.069**	**0.154**	**0.078**	**0.016**	**97.030**	**99.045**	**99.649**	**7.713**

Table A12: Pixel-Perfect Depth results before and after three calibration schemes. All values are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
KITTI	PPD	0.100	1.944	3.903	0.145	0.043	90.965	97.446	98.978	13.383
	AnchorDepth-Interp	0.078	1.622	**3.640**	0.131	0.034	92.753	97.892	99.181	12.414
	AnchorDepth-Refine	0.091	1.873	3.996	0.141	0.039	91.520	97.533	98.941	12.969
	AnchorDepth-Region	**0.055**	**1.214**	7.204	**0.104**	**0.023**	**95.774**	**98.806**	**99.547**	**10.330**
NYUv2	PPD	0.056	0.158	0.276	0.096	0.025	95.687	98.494	99.310	8.935
	AnchorDepth-Interp	0.050	0.138	0.261	0.084	0.021	96.364	98.898	99.560	7.998
	AnchorDepth-Refine	0.049	0.138	**0.259**	0.087	0.022	96.263	98.712	99.447	8.171
	AnchorDepth-Region	**0.031**	**0.091**	0.514	**0.066**	**0.013**	**97.923**	**99.398**	**99.764**	**6.535**
ETH3D	PPD	0.064	0.472	0.884	0.118	0.032	94.482	97.998	98.971	10.782
	AnchorDepth-Interp	**0.052**	**0.396**	**0.845**	0.101	0.024	95.651	98.489	99.284	9.650
	AnchorDepth-Refine	0.056	0.433	0.886	0.109	0.027	94.812	98.146	99.117	10.072
	AnchorDepth-Region	0.066	0.586	8.531	**0.062**	**0.013**	**98.124**	**99.448**	**99.694**	**6.112**
DIODE	PPD	0.102	1.342	2.364	0.190	0.056	90.703	94.996	96.364	14.339
	AnchorDepth-Interp	0.082	0.999	2.759	0.124	0.027	94.629	97.701	98.662	11.889
	AnchorDepth-Refine	0.070	0.851	**2.176**	0.135	0.029	94.415	97.512	98.480	12.575
	AnchorDepth-Region	**0.039**	**0.487**	8.201	**0.084**	**0.013**	**97.680**	**98.922**	**99.385**	**8.275**
ScanNet	PPD	0.056	0.101	0.179	0.095	0.025	95.900	98.753	99.463	8.973
	AnchorDepth-Interp	0.049	0.091	**0.171**	0.088	0.021	96.116	98.947	99.616	8.371
	AnchorDepth-Refine	0.052	0.095	0.177	0.090	0.022	96.024	98.764	99.488	8.547
	AnchorDepth-Region	**0.031**	**0.056**	0.176	**0.073**	**0.013**	**97.657**	**99.160**	**99.643**	**7.208**

the 512×512 grid using the same maximum of four planes, five points per plane, and ten non-planar points. The count-matched random condition sampled valid grid locations uniformly without replacement without explicitly excluding fixed coordinates. Both measurement sets were frozen and shared across methods. The average counts were 18.78 for STREET-day and 18.48 for STREET-dark.

Predictions were bilinearly resized from the external grid to the native 1000×1000 grid. Metrics were computed over the native valid mask and averaged over images within each split.

FAR. FAR RGB and metric-depth images had a native resolution of $H \times W = 4160 \times 3120$ and the same 4:3 aspect ratio as the common model grid. All methods used an external resolution of $H \times W = 1792 \times 1344$, obtained by full-image resizing without cropping or padding. The dense ground truth remained at its native resolution for evaluation. Each common-grid coordinate (y, x) was mapped to the nearest native pixel center as follows. Here, H_{native} and W_{native} denote the native height and width, and H_{common} and W_{common} denote the corresponding common-grid

Table A13: InfiniDepth RGB-only results before and after three calibration schemes. All values are rounded to three decimal places. The δ_1, δ_2, and δ_3 values are reported as percentages.

Dataset	Method	AbsRel↓	MAE↓	RMSE↓	RMSE-log↓	Log10↓	$\delta_1(\%)\uparrow$	$\delta_2(\%)\uparrow$	$\delta_3(\%)\uparrow$	SILog↓
KITTI	InfiniDepth	0.060	1.563	4.445	0.097	0.024	95.778	98.709	99.484	9.082
	AnchorDepth-Interp	0.050	1.342	4.134	0.090	0.021	96.699	**99.153**	**99.672**	8.667
	AnchorDepth-Refine	0.053	1.441	4.363	0.092	0.022	96.502	99.008	99.650	8.761
	AnchorDepth-Region	**0.040**	**1.092**	**4.124**	**0.087**	**0.017**	**97.144**	99.099	99.668	**8.610**
NYUv2	InfiniDepth	0.081	0.228	0.564	0.097	0.027	95.431	98.358	99.016	9.194
	AnchorDepth-Interp	0.069	0.199	**0.545**	0.087	0.022	96.377	98.845	99.381	8.327
	AnchorDepth-Refine	0.070	0.200	0.549	0.087	0.021	96.556	98.794	99.335	8.316
	AnchorDepth-Region	**0.034**	**0.114**	2.920	**0.069**	**0.013**	**97.829**	**99.291**	**99.743**	**6.829**
ETH3D	InfiniDepth	0.069	0.472	1.596	0.081	0.021	97.116	99.047	99.430	7.559
	AnchorDepth-Interp	0.059	0.412	**1.594**	0.074	0.017	97.850	99.400	99.661	7.076
	AnchorDepth-Refine	0.061	0.481	1.758	0.075	0.017	97.830	99.396	99.649	7.048
	AnchorDepth-Region	**0.041**	**0.333**	7.576	**0.061**	**0.010**	**99.009**	**99.610**	**99.741**	**5.990**
DIODE	InfiniDepth	0.072	0.315	0.972	0.085	0.021	95.929	98.426	99.458	7.839
	AnchorDepth-Interp	0.061	0.279	0.945	0.074	0.016	97.088	98.987	99.588	7.078
	AnchorDepth-Refine	0.058	0.276	0.945	0.071	0.015	97.683	99.007	99.544	6.643
	AnchorDepth-Region	**0.018**	**0.080**	**0.328**	**0.049**	**0.008**	**98.733**	**99.508**	**99.803**	**4.815**
ScanNet	InfiniDepth	0.085	0.159	0.430	0.099	0.026	95.241	98.292	99.185	9.220
	AnchorDepth-Interp	0.071	0.136	0.409	0.089	0.022	96.301	98.839	99.488	8.555
	AnchorDepth-Refine	0.072	0.141	0.422	0.090	0.021	96.279	98.692	99.407	8.570
	AnchorDepth-Region	**0.032**	**0.059**	**0.194**	**0.075**	**0.013**	**97.570**	**99.071**	**99.591**	**7.414**

dimensions:

$$y_{\text{native}} = \lfloor (y + 0.5) H_{\text{native}} / H_{\text{common}} \rfloor ,$$
$$x_{\text{native}} = \lfloor (x + 0.5) W_{\text{native}} / W_{\text{common}} \rfloor . \tag{15}$$

The sparse metric value was read directly from the corresponding native pixel, without interpolating the ground truth.

The fixed condition used the MoGe-2-based measurements generated on the common grid. The random condition sampled the same number of common-grid cells with valid native depths, uniformly and without replacement, without explicitly excluding fixed coordinates. All methods reused the same frozen coordinates and metric values. The average measurement count was 26.15. Predictions were bilinearly resized to the native grid. Metrics were computed over all finite, positive native ground-truth pixels and averaged over images.

G.2 Metrics and Fairness Controls

For every dataset, metrics were first computed independently over each image's protocol-valid mask and then averaged with equal weight across images. We did not pool all valid pixels across the dataset. Sparse anchors remained inside the evaluation mask. All methods reused the frozen per-image measurement sets, and no measurements were reselected during inference. Every sparse-depth input was constructed solely from the selected measurements; unselected dense ground-truth values were never exposed to the models. When predictions were restored to the evaluation resolution, only the predictions were bilinearly interpolated. The protocol-specific native or common-grid masks were preserved.

The reported metrics included AbsRel, MAE, RMSE, RMSE-log, log10, SqRel, δ_1, δ_2, δ_3, and SILog where applicable. In the measurement-error experiments, AbsRel and δ_1 were computed over the complete valid region, including the measurements.

H Additional Results

H.1 Robustness analysis

We evaluated AnchorDepth under localization bias and different measurement- noise levels. For localization bias, we replaced the depth reading at each selected ground-truth point with the reading at a randomly selected neighboring point. All perturbations used the same frozen, MoGe-derived sparse ground-truth points selected for each image. AnchorDepth-Interp and AnchorDepth-Refine

used identical measurement coordinates and perturbed depth values in each condition. The perturbations affected only the sparse depth readings. The original MoGe-2-L predictions, dense ground truth, and evaluation masks remained unchanged. The seven MoGe benchmarks used their official target grids, homographies, and valid masks. Table A14 summarizes the quantitative results.

Measurement-Error Protocol The robustness analysis started from the frozen MoGe-derived measurements for each image. Each image contained at most four planes, five points per plane, and ten non-planar points, for at most 30 measurements. AnchorDepth-Interp and AnchorDepth-Refine shared the same coordinates and perturbed values in every condition. Perturbations affected only the sparse readings. The MoGe-2-L prediction, dense ground truth, and evaluation mask remained unchanged.

For each selected sparse point (y_i, x_i), we considered candidate neighbors (y, x) satisfying

$$0 < (y - y_i)^2 + (x - x_i)^2 \le r^2,$$

where $r \in \{3, 5, 10\}$ pixels. Candidates had to lie inside the image and pass the protocol-valid mask. The center point was explicitly excluded. We uniformly sampled one candidate for the current seed and replaced the sparse depth reading with its ground-truth depth, while keeping the input coordinate at (y_i, x_i). Sampling was discrete, with no bilinear interpolation. If no valid candidate existed, we retained the original reading. The shift fallback/anomaly column reports the fraction of processed sparse points using this fallback.

Different radii used independent deterministic random streams, with five independent seeds per radius. The two AnchorDepth variants shared the sampled neighbor and its ground-truth depth in each condition.

Point locations remained unchanged in the LiDAR-like noise experiments. For each clean depth d_i in meters, each seed, image, and point received fixed random primitives: a Gaussian core error $\epsilon_{\text{core},i} \sim \mathcal{N}(0, \sigma_i^2)$, where $\sigma_i = 0.01\,\text{m} + 0.001 d_i$; a clipped Student-$t$ outlier

$$\epsilon_{\text{out},i} = \text{clip}(0.05 d_i T_i, -0.5 d_i, 0.5 d_i), \qquad T_i \sim t_{\nu=3};$$

and a uniform random variable $u_i \sim \mathcal{U}(0, 1)$. For an outlier probability $p \in \{5\%, 10\%, 25\%, 50\%\}$, the perturbed measurement was

$$d_i' = d_i + \epsilon_{\text{core},i} + \mathbf{1}[u_i < p]\epsilon_{\text{out},i}.$$

The four noise levels shared all three random primitives, yielding nested outlier sets. Non-positive perturbed depths were resampled. The table reports the realized outlier fraction rather than the noise amplitude. We averaged images within each dataset and reported the mean and population standard deviation across the five random seeds.

H.2 Additional visual results

We provide additional visual results to demonstrate the superiority of AnchorDepth. It can be observed that AnchorDepth significantly corrects the ambiguous scale in the original predictions of the MDE model. And more importantly, AnchorDepth is capable of correcting geometric structure errors to a certain extent, as shown in the last row of Figure A3.

H.3 On-device evaluation test details

Because the low-cost LiDAR produced an insufficiently dense reference point cloud, we did not use standard depth-estimation metrics. Instead, we used SAM3 (Carion et al., 2026) to segment the target objects from the RGB images, and then projected the depth maps produced by each method onto the point cloud using accurately calibrated camera intrinsics to compute the relevant distances within the masked regions. The evaluation covered three tasks: measuring object-to-camera distances, inter-object distances, and object dimensions. We used relative error, defined as the error divided by the ground-truth value, as the evaluation metric. We conducted evaluations across 32 common real-world scenarios (15 indoor and 17 outdoor), collecting a total of 154 RGB images along with their corresponding distance information, and separately recorded the number of ground-truth points, runtime, and overall performance of each method under indoor and outdoor conditions.

Table A14: Robustness of AnchorDepth-Interp and AnchorDepth-Refine on STREET-day and STREET-dark. Shift and LiDAR-like noise conditions use five random seeds; entries are mean (standard deviation).

Perturbation	AnchorDepth-Interp		AnchorDepth-Refine	
	AbsRel ↓	δ_1 ↑	AbsRel ↓	δ_1 ↑
STREET-day				
Clean	0.03365	0.98147	0.02672	0.98841
Shift–3 px	0.03877 (0.00083)	0.97868 (0.00028)	0.02903 (0.00059)	0.98833 (0.00007)
Shift–5 px	0.04149 (0.00044)	0.97561 (0.00015)	0.03647 (0.00085)	0.98766 (0.00012)
Shift–10 px	0.04834 (0.00060)	0.97168 (0.00029)	0.03888 (0.00166)	0.98591 (0.00025)
LiDAR noise–5%	0.03490 (0.00001)	0.98119 (0.00001)	0.02869 (0.00005)	0.98718 (0.00025)
LiDAR noise–10%	0.03604 (0.00004)	0.98095 (0.00006)	0.03062 (0.00021)	0.98602 (0.00032)
LiDAR noise–25%	0.03932 (0.00003)	0.98021 (0.00005)	0.03600 (0.00020)	0.98328 (0.00054)
LiDAR noise–50%	0.04456 (0.00008)	0.97891 (0.00006)	0.04521 (0.00032)	0.97813 (0.00090)
STREET-dark				
Clean	0.03121	0.98048	0.02897	0.98667
Shift–3 px	0.03401 (0.00025)	0.97829 (0.00032)	0.02967 (0.00004)	0.98659 (0.00005)
Shift–5 px	0.03700 (0.00033)	0.97658 (0.00034)	0.03095 (0.00008)	0.98619 (0.00007)
Shift–10 px	0.04322 (0.00031)	0.97378 (0.00019)	0.03534 (0.00018)	0.98533 (0.00010)
LiDAR noise–5%	0.03235 (0.00003)	0.98024 (0.00005)	0.03081 (0.00006)	0.98575 (0.00014)
LiDAR noise–10%	0.03346 (0.00006)	0.97998 (0.00010)	0.03254 (0.00009)	0.98478 (0.00019)
LiDAR noise–25%	0.03664 (0.00010)	0.97927 (0.00013)	0.03777 (0.00011)	0.98200 (0.00036)
LiDAR noise–50%	0.04161 (0.00019)	0.97794 (0.00019)	0.04668 (0.00025)	0.97685 (0.00050)

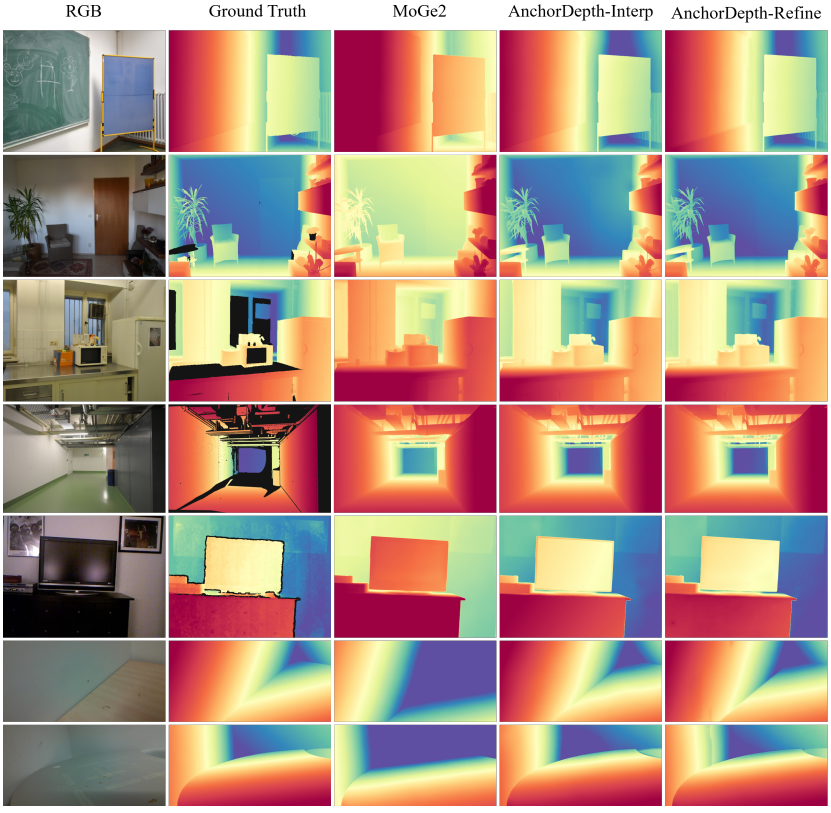

Figure A3: Comparison of results before and after AnchorDepth calibration.

Each comparison method is evaluated using its official evaluation protocol. For PriorDA and InfiniDepth, we input the 5000ms point cloud collected by low-cost LiDAR into the models for sparse prior guidance.