OmniPoint: Universal Monocular Metric Pointcloud from Any Camera

Botao Ye[1,2⋆], Marc Pollefeys[2], Ming-Hsuan Yang[1], and Abhijit Kundu[1]

[1]Google DeepMind [2]ETH Zurich

Project Page: https://omnipoint.github.io/

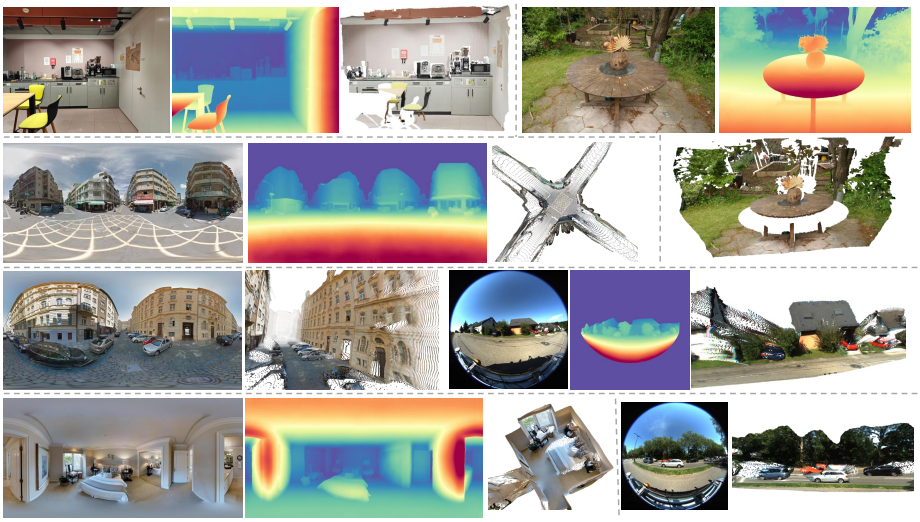

Fig. 1: Universal Monocular Point Cloud Estimation. OmniPoint reconstructs metric 3D point clouds from a single image using a unified framework that supports diverse camera models, including perspective, fisheye, and 360° panoramic images. It further accommodates flexible geometric conditioning by incorporating camera intrinsics and sparse depth information when available.

Abstract. Recovering metric 3D geometry from monocular images is a fundamental computer vision task, yet current methods remain heavily fragmented by fixed camera model assumptions and inflexible input schemes. We present OmniPoint, a unified framework designed to generalize metric reconstruction across diverse imaging sensors, including pinhole, fisheye, and equirectangular projections, while accommodating varying geometric priors. To overcome projection rigidity, OmniPoint abandons conventional planar depth regression. It instead adopts a decoupled ray and distance representation alongside a decoupled training objective, explicitly separating the camera projection model from the scene structure. To address the severe scarcity of training data for alternative cameras, we introduce a bidirectional augmentation strategy

⋆ Work done as an intern at Google DeepMind.

that explicitly bridges labeled perspective data and unlabeled omnidirectional domains in 3D space. Furthermore, to seamlessly integrate optional inputs like camera intrinsics or sparse depth without destabilizing the network through feature distribution shifts, we propose a robust information injection mechanism. This mechanism utilizes learnable input state embeddings to resolve architectural ambiguity and applies vectorized Gaussian smoothing to densify irregular measurements. Extensive experiments demonstrate that OmniPoint achieves state-of-the-art zero-shot performance across multiple benchmarks, establishing a robust new standard for unified monocular 3D reconstruction.

Keywords: Monocular Depth Estimation · Camera-Agnostic Model · Panoramic Depth Estimation

1 Introduction

Understanding the dense 3D structure of a scene from a single monocular image is a cornerstone challenge in computer vision. Recently, the field has witnessed a paradigm shift driven by large-scale datasets [41, 64] and visual foundation models [35, 42]. By learning powerful priors, modern approaches can now recover high-quality depth maps [22, 61] and pixel-wise 3D point clouds [38, 51] with remarkable zero-shot generalization across diverse, in-the-wild environments.

Despite this progress, current methodologies remain highly fragmented. Existing methods are typically *tied to a fixed camera model*, assuming a single projection model, most commonly pinhole [51, 61], or in specialized cases panoramic [27] or fisheye [60]. Furthermore, they are *inflexible with respect to input modalities* and cannot dynamically incorporate optional geometric cues such as camera intrinsics or sparse depth measurements (*e.g.*, from LiDAR or SfM). This fragmentation poses a significant barrier in practical applications such as robotics or autonomous driving, which are frequently equipped with heterogeneous sensor suites (*e.g.*, a forward-facing pinhole camera paired with surround-view fisheye lenses) and varying depth sensors. Deploying a robust 3D perception system in such scenarios currently requires maintaining multiple isolated models, which incurs prohibitive memory overhead and complicates sensor fusion.

To overcome this fragmentation, we introduce OmniPoint, a unified framework for monocular geometry estimation centered around a *camera-agnostic 3D representation*. Conventional approaches predict per-pixel depth values z under a pinhole assumption, making their output inherently tied to a specific camera model. For instance, affine-invariant disparity $(1/z)$ [61, 62] or log-depth $(\log z)$ [13, 51] require $z > 0$, precluding 360° environments and points behind the camera. More importantly, the projection formula $\mathbf{P} = z \cdot K^{-1}[u, v, 1]^\top$ assumes a linear mapping from pixel coordinates to camera rays, which fails for wide-FoV or non-pinhole cameras. While directly predicting per-pixel 3D coordinates (x, y, z) avoids this formula, it *entangles* camera projection with scene geometry: the network must implicitly learn the entire projection model for each camera type, making cross-camera generalization fundamentally difficult. Our

Table 1: Comparison of OmniPoint with state-of-the-art monocular geometry estimation methods. OmniPoint is the first approach to unify all camera types, output representations, and geometric input priors within a single framework.

		DAv2 [62]	MoGe [51]	MoGeV2 [52]	DepthPro [5]	UniK3D [36]	DAC [16]	DA^2 [27]	PromptDA [30]	PriorDA [55]	**OmniPoint**
Output	Affine-Inv Depth	✓	✓	✓	✓	✓	✓		✓	✓	✓
	Affine-Inv Points		✓	✓	✓	✓					✓
	Metric Points			✓	✓	✓					✓
Image Type	Pinhole	✓	✓	✓	✓	✓	✓		✓	✓	✓
	Fisheye					✓	✓				✓
	Panorama					✓	✓	✓			✓
Prior	Intrinsic					✓					✓
	Sparse Depth								✓	✓	✓

key insight is to explicitly *decouple* these two sources of variation. OmniPoint adopts a universal representation based on a ray direction $\mathbf{r}$ and a radial distance d for each pixel, defining a 3D point as $\mathbf{P} = d \cdot \mathbf{r}$. The ray direction $\mathbf{r}$ isolates the camera-dependent projection geometry, while the radial distance d encodes the camera-agnostic scene structure along that ray. This separation ensures that once the ray direction is known, the exact same distance prediction applies regardless of the camera type. Furthermore, we pair this representation with a *decoupled training objective* (Sec. 3.2) that prevents ray and distance errors from interfering during optimization, a subtle but critical design choice validated by our ablations (Tab. 6).

However, a universal representation alone is insufficient. A major bottleneck in achieving true cross-camera generalization is the severe scarcity of non-pinhole training data. To bridge this gap, we propose a systematic *bidirectional data augmentation* strategy. Rather than relying on simple 2D image transformations, we explicitly bridge labeled perspective data and unlabeled omnidirectional data in 3D. A *Perspective-to-Any* process synthetically projects large-scale pinhole datasets into fisheye and panoramic views to generate rich, paired supervision. Conversely, an *Any-to-Perspective* process samples virtual pinhole views from unlabeled in-the-wild panoramic images, generates pseudo-ground truth using our perspective model, and reprojects them to enforce cross-camera consistency.

Beyond camera generalization, another critical capability for a unified system is the flexible integration of geometric priors. A major obstacle in prior-aware models is that toggling optional inputs (*e.g.*, switching from RGB-only to RGB+sparse depth) causes severe feature distribution shifts, which destabilizes network training. Our insight is to explicitly signal the active configuration using learnable *input-state embeddings*, which dynamically guides the Vision Transformer (ViT) to resolve architectural ambiguity and correctly interpret varying input distributions. Furthermore, to effectively utilize sparse depth without suffering from its irregular and noisy nature, we introduce a robust information injection mechanism via fully vectorized Gaussian smoothing. Instead of feeding raw points directly, this mechanism normalizes and densifies the measurements into spatially consistent geometric guidance. This allows OmniPoint to seamlessly transition between acting as a zero-shot monocular estimator and a high-performance depth densification system.

Our key contributions are:

- OmniPoint, the first unified monocular geometry estimation framework that effectively handles heterogeneous camera models (pinhole, fisheye, 360°) and flexible priors (intrinsics, sparse depth) without architectural changes.
- A bidirectional augmentation strategy that effectively bridges the supervision gap between perspective and omnidirectional domains.
- A robust geometric information injection mechanism that prevents feature distribution shifts and leverages sparse inputs for highly accurate metric prediction.
- State-of-the-art zero-shot performance across a wide range of camera models and benchmarks, demonstrating unparalleled flexibility.

2 Related Work

Monocular Depth Estimation. Early approaches to monocular depth estimation relied on CNNs trained with supervised or self-supervised signals from stereo pairs or monocular video. A recent paradigm shift has been driven by models trained on large-scale, diverse datasets [40, 51, 61] and by pre-trained foundation models [22, 62]. These methods demonstrate strong zero-shot generalization to in-the-wild images. However, a fundamental limitation remains: they are designed for and evaluated primarily on perspective images under the pinhole camera assumption. Their planar-depth outputs are consequently ill-suited to other camera geometries.

Omnidirectional and Wide-FoV Depth. Another line of work targets specialized, non-pinhole camera systems. Methods have been proposed specifically for 360° panoramas [8, 27] and for fisheye or heavily distorted lenses [14, 26, 60], typically relying on camera-specific unprojection models. While effective within their respective domains, these models are highly specialized: a model trained for panoramas cannot process fisheye images, and neither can it operate on standard pinhole images, limiting their applicability in general scenarios. UniK3D [36] takes a step toward unifying multiple camera models within a single framework. However, due to the scarcity of wide-FoV and non-pinhole training data, its performance on panoramic and fisheye images remains suboptimal. In contrast, we show that our approach significantly outperforms UniK3D in these settings (see Fig. 4 and Tab. 1), owing to our bidirectional data augmentation strategy. Moreover, UniK3D is not designed to incorporate geometric prior information such as sparse depth or intrinsics.

Conditional Geometry Estimation. A third line of research incorporates available depth information into the estimation process. The most prominent examples are depth completion methods [45, 68], which densify sparse depth inputs (e.g., from LiDAR). More recent promptable approaches [30, 55] guide a general depth estimation model using sparse points or other auxiliary cues. However, these methods still rely on the pinhole camera assumption and cannot accommodate diverse camera models or leverage camera intrinsics as conditions.

In addition, they are tailored for conditioned estimation only and cannot be applied to non-conditioned RGB inputs.

OmniPoint is the first work to bridge these three previously disjoint research directions. It provides a single, unified model that functions simultaneously as (1) a general-purpose pinhole depth estimator, (2) a specialized wide-FoV estimator, and (3) a conditional, prior-aware model—without any architectural changes.

3 Method

OmniPoint is a feedforward framework that predicts dense 3D point clouds from a single monocular image captured by any camera type. Given an input image $\mathbf{I}$ and optional geometric priors such as camera intrinsics K and sparse depth $\mathbf{S}$, the model outputs a 3D point $\mathbf{P}_i$ for each pixel i, represented by its ray direction and radial distance from the camera center:

$$f_{\text{OmniPoint}}(\mathbf{I}, [K, \mathbf{S}]) = \{\hat{s}, \hat{\mathbf{R}}, \hat{\mathbf{D}}\}, \tag{1}$$

where $\tilde{\mathbf{R}}$ denotes the predicted ray directions, $\hat{\mathbf{D}}$ the corresponding radial distances, and s is the global metric scale factor. The final metric point map may be obtained via $\hat{s} \cdot \hat{\mathbf{R}} \cdot \hat{\mathbf{D}}$.

3.1 Camera-Agnostic 3D Representation

A core design choice in a camera-universal model is the output representation. We analyze common formulations and motivate our unified one.

Planar Depth. The standard paradigm in monocular depth estimation [22,62] is to regress planar depth z orthogonal to the camera plane. This formulation inherently relies on the linear unprojection formula $\mathbf{P} = z \cdot K^{-1}[u, v, 1]^\top$. This mapping fundamentally breaks down for wide-FoV or non-pinhole cameras, where rays undergo complex radial distortions and do not project linearly. Moreover, planar depth cannot represent full 360° panoramic scenes. As the field of view approaches 180°, z diverges to infinity, and it becomes undefined for points located behind the camera ($z < 0$). Therefore, common prediction targets such as affine-invariant disparity ($1/z$) [61,62] and log-depth ($\log z$) [13,51] inherently require $z > 0$, making them unsuitable for modeling environments that extend beyond the front-facing half-space. Although some methods, such as UniDepth [38], additionally estimate camera rays, they still fall into this category as their primary prediction target remains planar depth.

Pixel-wise 3D Coordinates. An alternative approach [49,51,53] directly predicts (x, y, z) coordinates for each pixel. However, this is fundamentally *camera-dependent* and *ill-conditioned for generalization*, and merely shifts the burden entirely onto the network's latent capacity. The model is forced to memorize a highly non-linear mapping $\mathcal{F} : (u, v) \mapsto (x, y, z)$ that inherently mixes the scene's

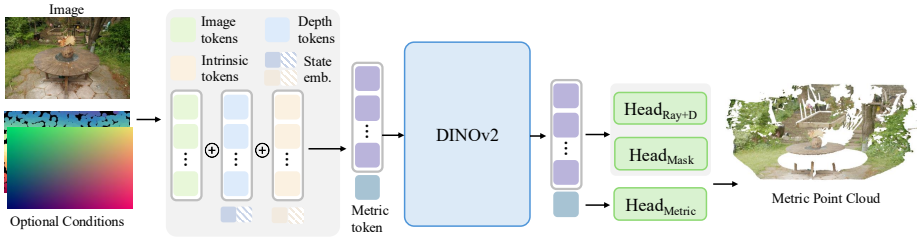

Fig. 2: Overview of OmniPoint. Our pipeline takes a single image from any camera and optional conditions (intrinsics, sparse depth). It uses a universal ray-distance representation to produce a metric point cloud.

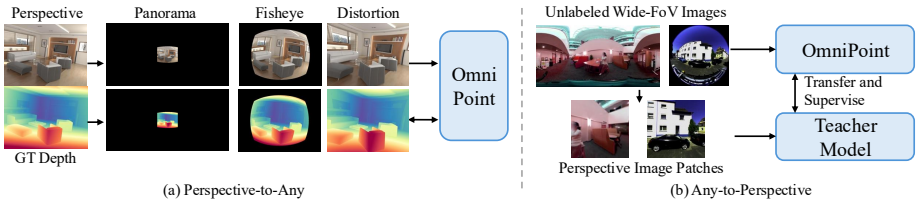

Fig. 3: Bidirectional Data Augmentation. (a) **Perspective-to-Any:** We synthesize novel fisheye and panoramic views from pinhole datasets. (b) **Any-to-Perspective:** We leverage unlabeled wide-FoV data by reprojecting virtual pinhole views for self-training.

structural distance with the specific camera's intrinsic projection pattern. Training a single neural architecture to dynamically alternate between the fundamentally different $\mathcal{F}$ mappings of pinhole, fisheye, and equirectangular images leads to optimization conflicts and poor zero-shot generalization across cameras.

Ray and Distance. To achieve true camera-agnostic estimation, we explicitly factorize the prediction into a unit ray direction $\mathbf{r} \in \mathbb{R}^3$ and a scalar radial distance $d \in \mathbb{R}^+$. The ray $\mathbf{r}$ perfectly absorbs any arbitrary lens distortion or projection geometry, leaving d to represent pure structural distance. By doing so, the distance prediction becomes completely invariant to the camera model. Furthermore, while recent camera-universal works like UniK3D [36] utilize Spherical Harmonics (SH) to parameterize the ray field, we find that SH introduces unwanted smoothing artifacts and struggles to capture abrupt geometric changes in highly distorted edge regions (See Fig. 4 and Fig. 5). Our simple, explicit ray-distance pairing performs comparably or better, providing a stable, high-fidelity target for network regression.

3.2 Decoupled Point Learning

Predicting both distance and ray direction introduces competing gradients if supervised naively. A standard L_1 point loss, $\mathcal{L}_{\text{naive}} = \sum_i \|\hat{d}_i \cdot \hat{\mathbf{r}}_i - d_i \cdot \mathbf{r}_i\|_1$, entangles the errors of the two predictions. For instance, if the network predicts

an incorrect ray direction $\hat{\mathbf{r}}_i$ but the correct distance $\hat{d}_i$, the resulting 3D point $\hat{\mathbf{P}}_i$ will still deviate from the ground truth. Consequently, the gradient update will inappropriately penalize the accurate distance prediction to compensate for the ray error. This cross-contamination destabilizes the learning process, particularly in highly distorted regions like the periphery of fisheye lenses.

To prevent this interference, we introduce a decoupled training objective that supervises the camera geometry and scene structure independently:

$$\mathcal{L}_{\text{ray}} = \sum_i \|\hat{\mathbf{r}}_i - \mathbf{r}_i\|_1, \quad \mathcal{L}_{\text{point}} = \sum_i \|s^* \cdot \hat{d}_i \cdot \mathbf{r}_i - d_i \cdot \mathbf{r}_i\|_1, \tag{2}$$

where s^* denotes the optimal scale obtained online via ROE alignment [51] between the predicted affine-invariant point map $(\hat{d}_i \cdot \mathbf{r}_i)$ and the ground-truth point map. In $\mathcal{L}_{\text{point}}$, we construct a proxy 3D point by projecting the *predicted* distance $\hat{d}_i$ along the *ground-truth* ray $\mathbf{r}_i$. This cleanly isolates the distance error, ensuring that the depth estimator is penalized strictly for structural inaccuracies, while $\mathcal{L}_{\text{ray}}$ independently guides the network to learn the correct unprojection mapping. As demonstrated in our ablations (Tab. 6), this decoupling is crucial for achieving high-fidelity point clouds across diverse camera geometries.

3.3 Bidirectional Data Augmentation

To overcome the severe scarcity of ground-truth 3D data for non-pinhole cameras, we introduce a systematic bidirectional data augmentation pipeline. This strategy explicitly bridges the gap between labeled perspective data and unlabeled omnidirectional domains operating in 3D space.

Perspective-to-Any Synthesis. To generate rich, paired supervision for diverse camera models, we simulate wide-FoV cameras from existing large-scale pinhole datasets [7, 31, 41, 64]. Given a pinhole image I_p and its ground-truth depth D_p, we first unproject them into a dense 3D point cloud $\mathbf{P}$. We then define a virtual camera model f_{any} (e.g., a fisheye lens with specific radial distortion or an equirectangular panorama) and re-project $\mathbf{P}$ onto this new image plane. We apply a binary validity mask during re-projection to exclude void holes caused by occlusions, thereby creating a highly accurate synthetic training pair $(I_{\text{any}}, (\mathbf{R}_{\text{any}}, D_{\text{any}}))$.

Any-to-Perspective Self-Training. Conversely, to leverage the vast amount of unlabeled wide-FoV data available in-the-wild, we propose a self-training augmentation. Given an unlabeled fisheye or panoramic image I_{any}, we first sample one perspective patch I_p by re-projecting portions of the image using a virtual pinhole camera. We run our (already strong) pinhole-capable model to get pseudo-ground-truth depth D_p for the sampled patch. Notably, although we can sample multiple-perspective images from the wide-FoV images, such as panoramic images, we always only sample one path per sample during training, as it avoids inconsistent depth prediction between the predictions of different patches. These depths are then unprojected and stitched back into the original I_{img} coordinate frame, creating a new pseudo-label D_{any} for fine-tuning.

3.4 Optional Geometric Inputs

A critical capability for a unified geometry model is the flexible integration of auxiliary priors, such as camera intrinsics or sparse depth measurements. However, toggling these inputs causes severe feature distribution shifts that can destabilize the network. We address this through a robust information injection mechanism.

Input-State Embeddings. To resolve the architectural ambiguity caused by varying input configurations (*e.g.*, RGB-only vs. RGB + sparse depth), we explicitly signal the active configuration to the model. We introduce learnable *input-state embeddings* ($\mathbf{e}_{\text{intrinsic}}$ and $\mathbf{e}_{\text{depth}}$) that are added to the token sequence entering the Vision Transformer (ViT). Each embedding acts as a boolean indicator, adopting one learned vector when its respective prior is available and another when it is absent. This enables the backbone to dynamically correctly interpret the incoming feature distributions.

Robust Geometric Injection. When camera intrinsics are available, we precompute the ground-truth per-pixel ray map $\mathbf{R}_{\text{gt}}$, process it through a lightweight convolutional encoder, and fuse it with the image features.

When sparse depth $\mathbf{S}$ (*e.g.*, from LiDAR or SfM) is provided, feeding these raw, irregular points directly into the network often leads to noisy gradients. Instead of directly feeding raw sparse depth as input, we first normalize it by the mean of all valid depth values to ensure stable scaling and consistent depth ranges across different input conditions. Rather than introducing additional scale factors for metric depths as input, we perform a simple post-alignment between the sparse condition and the predicted depth when metric depth is provided. This works well because our model can sufficiently utilize normalized input, thus providing correspondence between output and input condition. Moreover, raw sparse depth $\mathbf{S}$ (from LiDAR, SfM, etc.) provides strong geometric cues but is often spatially irregular and noisy. We apply a fully vectorized Gaussian splatting procedure to obtain a smooth and spatially consistent depth map. Each valid sparse point $\mathbf{S}(u, v)$ is treated as the center of a Gaussian kernel whose influence is distributed within a fixed radius r. The spatial weight for an offset $(\Delta u, \Delta v)$ is defined as

$$w(\Delta u, \Delta v) = \exp\left(-\frac{\Delta u^2 + \Delta v^2}{2\sigma^2}\right), \tag{3}$$

where σ scales proportionally to the depth magnitude to maintain relative smoothness across varying depths. All pixel-wise contributions are accumulated in parallel using a scatter-add operation:

$$\mathbf{D}_{\text{smooth}}(x, y) = \frac{\sum_i w_i(x, y)\, S_i}{\sum_i w_i(x, y)}, \tag{4}$$

yielding a dense and spatially consistent smoothed depth map. We additionally construct a binary mask $\mathbf{M}$ indicating whether each pixel corresponds to an original valid sparse measurement. $\mathbf{D}_{\text{smooth}}$ and $\mathbf{M}$ are concatenated and projected into the feature space using a convolutional layer, then fused with image

embeddings to guide the ViT backbone. This mechanism allows OmniPoint to seamlessly transition into a highly accurate depth densification system.

3.5 Training

Loss Functions. In addition to $\mathcal{L}_{\text{point}}$ and $\mathcal{L}_{\text{ray}}$ (Sec. 3.2), we employ a metric alignment loss [52] to enforce global scale consistency between the predicted and ground-truth point clouds:

$$\mathcal{L}_{\text{metric}} = \| \log(\hat{s}) - \text{stopgrad}(\log(s^*)) \|_2^2. \tag{5}$$

Here, s^* is the optimal scalar between the predicted point cloud and the ground truth (see Sec. 3.2). The stop-gradient operator prevents gradients from flowing through s^*, stabilizing training while encouraging the network to recover the correct global scale. To further improve geometric fidelity, we incorporate normal and local consistency losses ($\mathcal{L}_{\text{normal}}$, $\mathcal{L}_{\text{local}}$) following [51], which enhance surface smoothness and preserve fine-grained structure. In addition, we apply a binary mask loss $\mathcal{L}_{\text{mask}}$ to mask out sky regions. The final training objective is given by

$$\begin{aligned} \mathcal{L} = \mathcal{L}_{\text{point}} + \lambda_{\text{ray}}\mathcal{L}_{\text{ray}} + \lambda_{\text{metric}}\mathcal{L}_{\text{metric}} \\ + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}} + \lambda_{\text{local}}\mathcal{L}_{\text{local}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}}, \end{aligned} \tag{6}$$

where $\lambda_\star$ controls the weight of each term.

Training Process. Training proceeds in three stages. (1) We pretrain on large-scale perspective datasets to establish a strong baseline geometry model. (2) We fine-tune on mixed synthetic and real data, including fisheye and panoramic images, using the bidirectional augmentation in Sec. 3.3. (3) Finally, we freeze the backbone and train only the mask head using $\mathcal{L}_{\text{mask}}$ with pseudo-labels from SegFormer [59], ensuring robust sky-region masking during inference.

4 Experiments

4.1 Experimental Setup

Training Datasets. We employ 29 labeled datasets for training: ARKitScenes [2], ARKitScenes-HighRes [2], HOI4D [31], Matterport3D [10], ScanNet++ [64], SmartPortraits [24], Waymo [44], WildRGBD [58], 3D Ken Burns [34], BEDLAM [4], BlendedMVS [63], Dynamic Replica [21], EDEN [25], HyperSim [41], IRS [50], MVS-Synth [19], OmniObject3D [57], OmniWorld-Game [67], Spring [32], Synscapes [56], PointOdyssey [66], TartanAir [54], TartanAir v2 [54], Unreal4K [46], UrbanSyn [17], VirtualKITTI2 [7], Structure3D [65], and KITTI360 [28]. We additionally use the unlabeled panorama datasets Diverse360 [9] and 360+x [11] for our Any-to-Perspective training, where KITTI360 [28] is also included. Please refer to the supplementary material for detailed descriptions of these datasets.

Table 2: Relative geometric and depth comparison across small FoV (8 pinhole datasets: NYUv2 [33], KITTI [15], ETH3D [43], iBims-1 [23], GSO [12], Sintel [6], DIODE [47], and HAMMER [20]), large FoV (fisheye dataset: KITTI360 [28]), and 360° images (Stanford2D3D-S [1] and PanoSUNCG [48]). OmniPoint achieves SOTA or competitive performance across all camera types (**Best** , second-best , and third-best).

	S.FoV				L.FoV				360 Images			
	Depth		Point		Depth		Point		Depth		Point	
Method	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$
DA V1	6.43	95.3	–	–	–	–	–	–	–	–	–	–
DA V2	6.35	95.0	–	–	–	–	–	–	–	–	–	–
Metric3D V2	6.06	95.3	–	–	–	–	–	–	–	–	–	–
UniDepth V2	4.23	96.6	6.06	95.1	7.81	93.0	18.4	79.6	19.3	69.5	102.8	2.89
Depth Pro	5.28	95.6	7.49	93.3	26.6	54.6	36.2	35.5	–	–	–	–
MoGe V1	4.05	96.6	**5.25**	95.5	9.48	90.1	28.1	52.5	22.3	62.9	102.6	2.91
MoGe V2	**3.98**	**96.8**	5.59	95.5	12.0	85.0	23.4	60.9	25.1	56.1	102.8	2.85
DA3	4.78	95.5	6.33	93.7	12.8	83.4	29.9	47.5	25.5	55.7	101.5	2.75
UniK3D	4.14	96.7	5.61	95.5	8.77	92.7	11.5	90.6	10.4	90.0	11.4	89.4
DA^2	–	–	–	–	–	–	–	–	6.65	94.7	7.30	94.0
Ours	4.04	**96.8**	5.55	**95.6**	**6.37**	**94.0**	**6.66**	**93.8**	**5.79**	**95.1**	**5.90**	**95.1**

Implementation Details. We adopt DINOv2-ViT-Large [35] as the backbone and use DPT [39] heads to predict the 3D points and mask as in [52]. For the first training stage, we use 72 A100 GPUs with a batch size of 8 per GPU and train for 10k steps, which takes approximately 20 hours. The second stage uses the same hardware configuration but reduces the training schedule to 5k steps. Additional implementation details are provided in the supplementary material.

Baselines. We compare our approach with the following monocular geometry estimation methods: 1) *Affine-invariant depth estimator*: Depth Anything v1 [61], Depth Anything v2 [62], MoGe [51], and VGGT [49]; 2) *Metric depth estimator*: ZeoDepth [3], Metric3Dv2 [18], UniDepth [38], UniDepth v2 [37], Depth Pro [5], MoGe v2 [52], Depth Anything 3 [29], and UniK3D [36]; 3) *Panorama depth estimator*: DA^2 [27].

Evaluation Details. To ensure a strictly fair and rigorous comparison, we re-evaluate all baselines using identical input formats and a standardized metric computation pipeline. We report geometric performance on both depth and 3D point predictions using the standard relative error (Rel↓) and the percentage of robust inliers ($\delta_1\uparrow$) following [51].

4.2 Experimental Results and Analysis

Relative Geometry and Depth. We evaluate the relative depth and geometry performance of OmniPoint across different camera models, with comparisons shown in Tab. 2, Fig. 4, and Fig. 5. On standard perspective images (S.FoV),

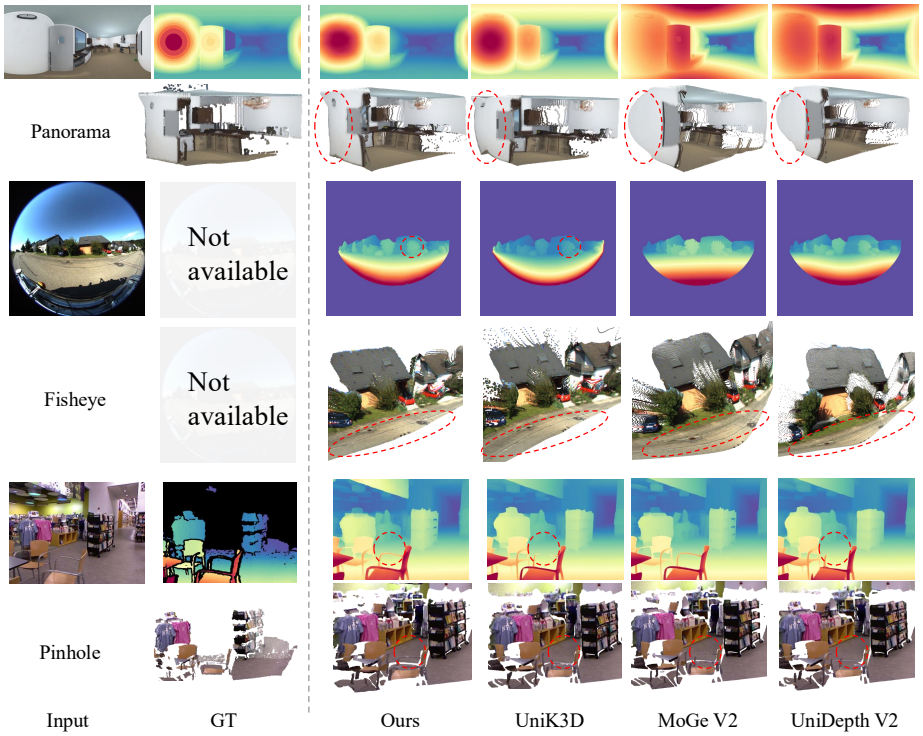

Fig. 4: Qualitative Comparison on Diverse Camera Models. We compare OmniPoint against baselines on panoramic (top), fisheye (middle), and pinhole (bottom) images. Our method produces geometrically consistent reconstructions, such as the vertical wall in the panorama and the straight road in the fisheye, where UniK3D [36] shows distortion. OmniPoint also recovers finer details in depth maps and point clouds.

our method achieves highly robust performance (96.8 δ_1), proving that our universal formulation preserves accuracy in traditional settings, matching strictly pinhole-specialized models like MoGe V2. The true superiority of our approach emerges in non-pinhole environments. When applied to large-FoV fisheye images, standard pinhole models fail at point map prediction due to severe projection distortion. In contrast, OmniPoint drastically surpasses all baselines, reducing the relative error of the universal baseline UniK3D from 11.5 down to 6.66 Rel. On 360° panoramic images, OmniPoint achieves the best depth accuracy (5.79 Rel), significantly outperforming the leading panorama-specific model DA^2 (6.65 Rel) while recovering sharper boundaries and undistorted structural lines, as visually evidenced in Fig. 5. The failure of UniK3D to maintain proper vertical structures on panoramas (Fig. 4) further validates the advantage of our bidirectional augmentation over purely latent unprojection representations. These results highlight our model's unique capability to serve as a universal geometry estimator.

Table 3: Metric Depth Estimation (δ_1) Performance. We evaluate zero-shot metric depth on six diverse benchmarks. Our model outperforms previous SOTA methods. Conditioning on intrinsics (+ K) or sparse depth (+ D) further improves results.

Method	NYUv2	KITTI	ETH3D	iBims-1	DIODE	HAMMER	Mean
ZoeDepth	91.9	85.4	33.7	67.2	29.3	3.23	51.8
UniDepth V2	92.8	**95.4**	69.5	**93.2**	51.8	46.8	74.4
Depth Pro	91.9	38.3	32.8	81.5	37.7	63.0	57.5
UniK3D	94.4	93.6	83.7	92.8	73.0	58.3	82.6
MoGe V2	**96.1**	62.9	90.8	83.0	66.4	65.6	77.5
Ours	86.2	88.0	**91.8**	87.7	**73.3**	**73.8**	**83.5**
Ours + K	88.3	90.7	93.0	88.3	75.2	74.5	85.0
Ours + D	98.5	98.4	94.1	99.0	98.4	99.3	98.0

Table 4: Comparison of different input configurations. Our optional geometric conditions consistently improve relative geometry estimation across metrics.

	Depth		**Point**	
Method	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$
No cond	4.04	96.8	5.55	95.6
Ours Intrin cond	3.98	96.8	5.08	96.2
Ours Sparse D cond	3.21	98.2	4.60	96.9
Ours D + K	3.15	98.3	3.59	97.9

Metric depth. Table 3 compares metric-scale depth estimation on various benchmarks. In the zero-shot, image-only setting, our model (83.5 Mean) achieves state-of-the-art performance, surpassing previous best metric estimators like UniK3D (82.6 Mean) and MoGe V2 (77.5 Mean). This confirms that our universal ray-distance representation and diverse training data lead to robust metric-scale understanding. Furthermore, we show the benefit of our flexible conditioning. By providing camera intrinsics, performance consistently improves, validating our intrinsic injection module. When provided with sparse depth priors, the performance sees a dramatic boost (98.0 Mean), demonstrating our model's capability to effectively utilize and densify sparse geometric information, akin to a high-performance depth completion system.

Geometric Conditioning. We provide an analysis of our geometric conditioning mechanism in Tab. 4 and Tab. 3. Starting from the unconditioned baseline (4.04 Depth Rel), injecting exact camera intrinsics immediately improves both depth (3.98 Rel) and point (5.08 Rel) accuracy by removing implicit unprojection ambiguities. When provided with sparse depth priors, the model seamlessly transitions into a high-performance densification system, yielding a massive boost across all metrics (*e.g.*, zero-shot metric δ_1 surges to an exceptional 97.5 Mean in Tab. 3). Finally, providing both priors simultaneously yields the highest precision (3.15 Depth Rel), confirming that our network efficiently fuses synergistic geometric information without suffering from feature distribution shifts.

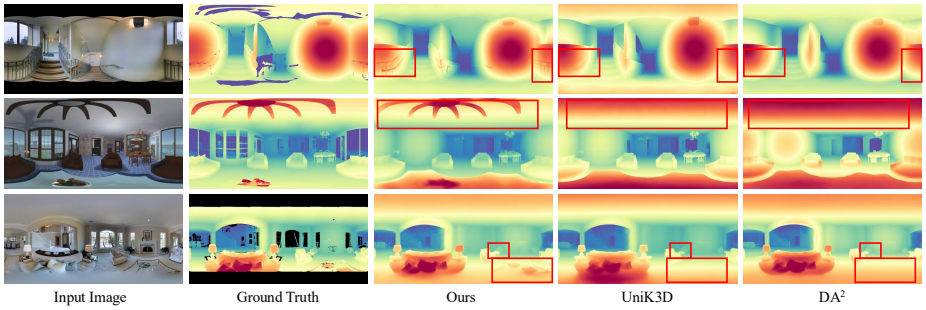

Fig. 5: Qualitative Comparison with Baselines on Panorama Image. Our method produces sharper panorama depth maps by adding our bidirectional data augmentation strategy.

Table 5: Output Representations. Our decoupled Ray+D representation consistently improves the performance across all camera types.

	S.FoV		L.FoV		360 Images	
Method	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$
Ray+D	**4.04**	**96.8**	**6.37**	**94.0**	**5.79**	**95.1**
XYZ	4.28	96.3	6.74	93.8	6.28	94.5

Table 6: Point Loss Mechanisms. Our decoupled loss (P + GT Ray) isolates distance errors for the highest accuracy.

	Depth		Point	
Method	Rel↓	$\delta\uparrow$	Rel↓	$\delta\uparrow$
P + GT Ray	**4.04**	**96.8**	**5.55**	**95.6**
P	4.05	96.8	5.65	95.2
D + Ray	5.01	96.4	7.88	93.5

4.3 Ablation Studies

Output Representations. Table 5 validates our ray-distance representation against a baseline that directly predicts global (x, y, z) coordinates. As shown, the XYZ formulation consistently lags behind our decoupled approach across all camera types, including standard perspective (4.04 vs 4.28 Rel), fisheye (6.37 vs 6.74 Rel), and panoramic images (5.79 vs 6.28 Rel). This explicitly confirms our core hypothesis: forcing a single neural network to implicitly memorize and alternate between diverse, complex unprojection mappings creates severe optimization conflicts. By explicitly decoupling projection geometry from scene structure, our Ray+D representation resolves these conflicts, enabling superior generalization across all camera models.

Decoupled Loss. We validate our decoupled point loss (Sec. 3.2) in Tab. 6. Independent, disjoint supervision of distance and ray (D + Ray) completely fails to enforce spatial geometric consistency, resulting in poor point cloud accuracy (7.88 Rel). Conversely, a naive entangled point loss (P) allows gradient cross-contamination, where unprojection errors inappropriately penalize structural predictions, visibly degrading point quality compared to our method (5.65 vs 5.55 Rel). By projecting the predicted distance strictly along the ground-truth

Table 7: Bidirectional Augmentation. Perspective-to-Any (P2A) synthesis and Any-to-Perspective (A2P) self-training both independently improve panoramic predictions. Used synergistically, they effectively overcome the extreme scarcity of non-pinhole 3D data.

P2A	A2P	Rel↓	$\delta\uparrow$
–	–	8.45	89.8
✓	–	6.21	94.5
–	✓	8.21	90.9
✓	✓	**5.79**	**95.1**

Table 8: Conditioning Mechanisms. Removing Gaussian smoothing severely degrades sparse-guided performance by injecting irregular noise, while removing the state embedding causes feature distribution shifts. Both components are vital for robust prior injection.

Method	Rel↓	$\delta\uparrow$
Ours	**3.21**	**98.2**
w/o Smooth	3.80	97.5
w/o State Emb	3.38	97.7

ray (P + GT Ray), our loss safely isolates structural errors, yielding the most stable optimization target.

Bidirectional Augmentation. We ablate our data augmentation strategy for 360° images in Tab. 7. Without any augmentation, the model heavily overfits to the limited available data, resulting in poor panoramic geometry (8.45 Rel). Our Perspective-to-Any synthesis generates crucial synthetic supervision, providing the single largest performance jump (to 6.21 Rel). Any-to-Perspective self-training provides highly complementary pseudo-supervision directly from real-world omnidirectional scenes. Combining both techniques synergistically yields our final state-of-the-art result of 5.79 Rel.

Condition Design. Table 8 validates the specific components of our geometric conditioning module (Sec. 3.4). Removing the Gaussian splatting procedure severely degrades sparse-guided performance (from 3.21 to 3.80 Rel), proving that raw sparse points are too irregular to provide stable gradients and must be normalized into dense, spatially consistent guidance. Furthermore, removing the learnable state embeddings causes a distinct performance drop (3.38 Rel). This confirms that explicitly signaling the active priors is strictly necessary to resolve architectural ambiguity and prevent harmful feature distribution shifts within the ViT backbone.

5 Conclusion

We introduce OmniPoint, a unified framework for monocular geometry estimation that jointly handles arbitrary camera models and diverse input priors. The core of our method is a universal ray-distance representation that fundamentally decouples projection geometry from structural geometry. To overcome data scarcity, we proposed a systematic bidirectional data augmentation pipeline bridging perspective and omnidirectional imagery. Finally, our robust prior-injection mechanism seamlessly integrates sparse depth and camera intrinsics

when available. Extensive experiments demonstrate that OmniPoint achieves state-of-the-art zero-shot performance across pinhole, fisheye, and panoramic domains, establishing a robust new standard for 3D vision.

References

1. Armeni, I., Sax, S., Zamir, A.R., Savarese, S.: Joint 2d-3d-semantic data for indoor scene understanding. arXiv preprint arXiv:1702.01105 (2017) 10
2. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021) 9
3. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth. arXiv preprint arXiv:2302.12288 (2023) 10
4. Black, M.J., Patel, P., Tesch, J., Yang, J.: Bedlam: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In: CVPR (2023) 9
5. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. In: ICLR (2024) 3, 10
6. Butler, D.J., Wulff, J., Stanley, G.B., Black, M.J.: A naturalistic open source movie for optical flow evaluation. In: ECCV (2012) 10
7. Cabon, Y., Murray, N., Humenberger, M.: Virtual kitti 2. arXiv preprint arXiv:2001.10773 (2020) 7, 9
8. Cao, Z., Zhu, J., Zhang, W., Ai, H., Bai, H., Zhao, H., Wang, L.: Panda: Towards panoramic depth anything with unlabeled panoramas and mobius spatial augmentation. In: CVPR (2025) 4
9. Cao, Z., Zhu, J., Zhang, W., Wang, L.: Any360d: Towards 360 depth anything with unlabeled 360 data and möbius spatial augmentation. arXiv e-prints pp. arXiv–2406 (2024) 9
10. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. arXiv preprint arXiv:1709.06158 (2017) 9
11. Chen, H., Hou, Y., Qu, C., Testini, I., Hong, X., Jiao, J.: 360+ x: A panoptic multi-modal scene understanding dataset. In: CVPR (2024) 9
12. Downs, L., Francis, A., Koenig, N., Kinman, B., Hickman, R., Reymann, K., McHugh, T.B., Vanhoucke, V.: Google scanned objects: A high-quality dataset of 3d scanned household items. In: ICRA (2022) 10
13. Eigen, D., Puhrsch, C., Fergus, R.: Depth map prediction from a single image using a multi-scale deep network. Advances in neural information processing systems **27** (2014) 2, 5
14. Feng, H., Wang, W., Deng, J., Zhou, W., Li, L., Li, H.: Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In: ICCV (2023) 4
15. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: The kitti dataset. The international journal of robotics research **32**(11), 1231–1237 (2013) 10
16. Guo, Y., Garg, S., Miangoleh, S.M.H., Huang, X., Ren, L.: Depth any camera: Zero-shot metric depth estimation from any camera. In: CVPR (2025) 3

17. Gómez, J.L., Silva, M., Seoane, A., Borràs, A., Noriega, M., Ros, G., Iglesias-Guitian, J.A., López, A.M.: All for one, and one for all: Urbansyn dataset, the third musketeer of synthetic driving scenes. Neurocomputing **637**, 130038 (2025) 9
18. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE TPAMI (2024) 10
19. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR (2018) 9
20. Jung, H., Ruhkamp, P., Zhai, G., Brasch, N., Li, Y., Verdie, Y., Song, J., Zhou, Y., Armagan, A., Ilic, S., et al.: On the importance of accurate geometry data for dense 3d vision tasks. In: CVPR (2023) 10
21. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Dynamicstereo: Consistent dynamic depth from stereo videos. In: CVPR (2023) 9
22. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: CVPR (2024) 2, 4, 5
23. Koch, T., Liebel, L., Fraundorfer, F., Korner, M.: Evaluation of cnn-based single-image depth estimation methods. In: ECCV Workshops (2018) 10
24. Kornilova, A., Faizullin, M., Pakulev, K., Sadkov, A., Kukushkin, D., Akhmetyanov, A., Akhtyamov, T., Taherinejad, H., Ferrer, G.: Smartportraits: Depth powered handheld smartphone dataset of human portraits for state estimation, reconstruction and synthesis. In: CVPR (2022) 9
25. Le, H.A., Mensink, T., Das, P., Karaoglu, S., Gevers, T.: Eden: Multimodal synthetic dataset of enclosed garden scenes. In: WACV (2021) 9
26. Lee, J., Cho, G., Park, J., Kim, K., Lee, S., Kim, J.H., Jeong, S.G., Joo, K.: Slabins: Fisheye depth estimation using slanted bins on road environments. In: ICCV (2023) 4
27. Li, H., Zheng, W., He, J., Liu, Y., Lin, X., Yang, X., Chen, Y.C., Guo, C.: Da2: Depth anything in any direction. arXiv preprint arXiv:2509.26618 (2025) 2, 3, 4, 10
28. Liao, Y., Xie, J., Geiger, A.: Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. IEEE TPAMI **45**(3), 3292–3310 (2022) 9, 10
29. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025) 10
30. Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B.: Prompting depth anything for 4k resolution accurate metric depth estimation. In: CVPR (2025) 3, 4
31. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: CVPR (2022) 7, 9
32. Mehl, L., Schmalfuss, J., Jahedi, A., Nalivayko, Y., Bruhn, A.: Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In: CVPR (2023) 9
33. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012) 10
34. Niklaus, S., Mai, L., Yang, J., Liu, F.: 3d ken burns effect from a single image. ACM TOG **38**(6), 1–15 (2019) 9

35. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023) 2, 10
36. Piccinelli, L., Sakaridis, C., Segu, M., Yang, Y.H., Li, S., Abbeloos, W., Van Gool, L.: Unik3d: Universal camera monocular 3d estimation. In: CVPR (2025) 3, 4, 6, 10, 11
37. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. arXiv preprint arXiv:2502.20110 (2025) 10
38. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: Unidepth: Universal monocular metric depth estimation. In: CVPR (2024) 2, 5, 10
39. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: ICCV (2021) 10
40. Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., Koltun, V.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE TPAMI **44**(3), 1623–1637 (2020) 4
41. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: ICCV (2021) 2, 7, 9
42. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2
43. Schops, T., Sattler, T., Pollefeys, M.: Bad slam: Bundle adjusted direct rgb-d slam. In: CVPR (2019) 10
44. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: CVPR (2020) 9
45. Tang, J., Tian, F.P., An, B., Li, J., Tan, P.: Bilateral propagation network for depth completion. In: CVPR (2024) 4
46. Tosi, F., Liao, Y., Schmitt, C., Geiger, A.: Smd-nets: Stereo mixture density networks. In: CVPR (2021) 9
47. Vasiljevic, I., Kolkin, N., Zhang, S., Luo, R., Wang, H., Dai, F.Z., Daniele, A.F., Mostajabi, M., Basart, S., Walter, M.R., et al.: Diode: A dense indoor and outdoor depth dataset. arXiv preprint arXiv:1908.00463 (2019) 10
48. Wang, F.E., Hu, H.N., Cheng, H.T., Lin, J.T., Yang, S.T., Shih, M.L., Chu, H.K., Sun, M.: Self-supervised learning of depth and camera motion from 360 videos. In: ACCV (2018) 10
49. Wang, J., Chen, M., Karaev, N., Vedaldi, A., Rupprecht, C., Novotny, D.: Vggt: Visual geometry grounded transformer. In: CVPR (2025) 5, 10
50. Wang, Q., Zheng, S., Yan, Q., Deng, F., Zhao, K., Chu, X.: Irs: A large naturalistic indoor robotics stereo dataset to train deep models for disparity and surface normal estimation. arXiv preprint arXiv:1912.09678 (2019) 9
51. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: CVPR (2025) 2, 3, 4, 5, 7, 9, 10
52. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. In: NeurIPS (2025) 3, 9, 10
53. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: CVPR (2024) 5

54. Wang, W., Zhu, D., Wang, X., Hu, Y., Qiu, Y., Wang, C., Hu, Y., Kapoor, A., Scherer, S.: Tartanair: A dataset to push the limits of visual slam. In: IROS (2020) 9
55. Wang, Z., Chen, S., Yang, L., Wang, J., Zhang, Z., Zhao, H., Zhao, Z.: Depth anything with any prior. arXiv preprint arXiv:2505.10565 (2025) 3, 4
56. Wrenninge, M., Unger, J.: Synscapes: A photorealistic synthetic dataset for street scene parsing. arXiv preprint arXiv:1810.08705 (2018) 9
57. Wu, T., Zhang, J., Fu, X., Wang, Y., Ren, J., Pan, L., Wu, W., Yang, L., Wang, J., Qian, C., et al.: Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In: CVPR (2023) 9
58. Xia, H., Fu, Y., Liu, S., Wang, X.: Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In: CVPR (2024) 9
59. Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P.: Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in neural information processing systems **34**, 12077–12090 (2021) 9
60. Xie, S., Wang, D., Liu, Y.H.: Omnividar: Omnidirectional depth estimation from multi-fisheye images. In: CVPR (2023) 2, 4
61. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024) 2, 4, 5, 10
62. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. In: NeurIPS (2024) 2, 3, 4, 5, 10
63. Yao, Y., Luo, Z., Li, S., Zhang, J., Ren, Y., Zhou, L., Fang, T., Quan, L.: Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In: CVPR (2020) 9
64. Yeshwanth, C., Liu, Y.C., Nießner, M., Dai, A.: Scannet++: A high-fidelity dataset of 3d indoor scenes. In: ICCV (2023) 2, 7, 9
65. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: ECCV (2020) 9
66. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In: ICCV (2023) 9
67. Zhou, Y., Wang, Y., Zhou, J., Chang, W., Guo, H., Li, Z., Ma, K., Li, X., Wang, Y., Zhu, H., et al.: Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. arXiv preprint arXiv:2509.12201 (2025) 9
68. Zuo, Y., Deng, J.: Ogni-dc: Robust depth completion with optimization-guided neural iterations. In: ECCV (2024) 4