# UniK3D: Universal Camera Monocular 3D Estimation

Luigi Piccinelli[1] Christos Sakaridis[1] Mattia Segu[1]
Yung-Hsu Yang[1] Siyuan Li[1] Wim Abbeloos[2] Luc Van Gool[1,3]

[1]ETH Zürich [2]Toyota Motor Europe [3]INSAIT, Sofia University St. Kliment Ohridski

## Abstract

*Monocular 3D estimation is crucial for visual perception. However, current methods fall short by relying on oversimplified assumptions, such as pinhole camera models or rectified images. These limitations severely restrict their general applicability, causing poor performance in real-world scenarios with fisheye or panoramic images and resulting in substantial context loss. To address this, we present* UniK3D[1]*, the first generalizable method for monocular 3D estimation able to model any camera. Our method introduces a spherical 3D representation which allows for better disentanglement of camera and scene geometry and enables accurate metric 3D reconstruction for unconstrained camera models. Our camera component features a novel, model-independent representation of the pencil of rays, achieved through a learned superposition of spherical harmonics. We also introduce an angular loss, which, together with the camera module design, prevents the contraction of the 3D outputs for wide-view cameras. A comprehensive zero-shot evaluation on 13 diverse datasets demonstrates the state-of-the-art performance of UniK3D across 3D, depth, and camera metrics, with substantial gains in challenging large-field-of-view and panoramic settings, while maintaining top accuracy in conventional pinhole small-field-of-view domains. Code and models are available at github.com/lpiccinelli-eth/unik3d.*

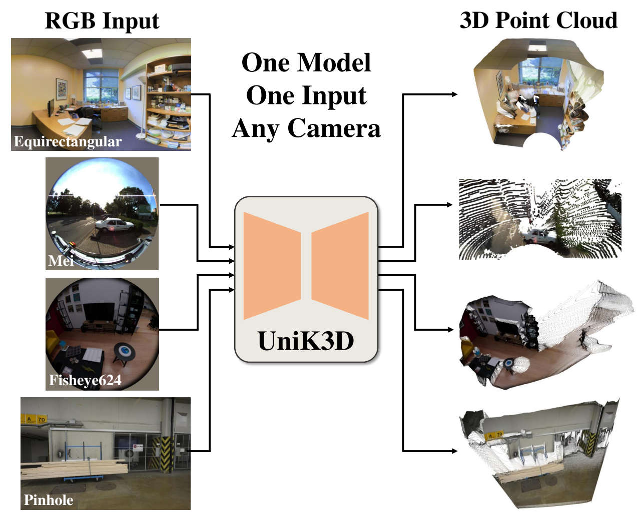


Figure 1. **UniK3D** introduces a novel and versatile approach that delivers precise metric 3D geometry estimation from a single image and for any camera type, ranging from pinhole to panoramic, without requiring any camera information. By leveraging (i) a flexible and general spherical formulation both for the radial dimension of 3D space and for the two camera-model-dependent orientation dimensions and (ii) advanced conditioning strategies. UniK3D outperforms traditional models without needing camera calibration or domain-specific tuning.

## 1. Introduction

Estimating 3D scene geometry is a fundamental task in computer vision since such 3D information serves as a crucial cue for action planning and execution [13, 80]. The scene's geometry 3D estimation task is vital for a wide range of applications, including autonomous navigation [49, 67] and 3D modeling [12], where accurate spatial understanding is essential. Recent advances in generalizable monocular depth estimation (MDE) [27, 56, 72] deliver impressive performance and visual quality across various domains, but these models are constrained to a relative output scale. Nonetheless, for practical applications, a consistent and reliable *metric-scaled* monocular depth estimate (MMDE) is crucial, as it enables accurate 3D reconstruction and geometric scene understanding necessary for embodied agents.

Existing methods have made considerable strides in the above direction of metric estimation. Earlier approaches assumed known camera intrinsics at test time [22, 76], while more recent works have relaxed this assumption [8, 53, 54]. However, these approaches still impose restrictive assumptions about input cameras, such as relying on a basic pinhole camera model [8, 53] or requiring access to ground-truth rectification parameters [76]. These simplifications substantially hinder the applicability and degrade the performance

[1]Pronounced "Unique-3D", with **K** denoting the intrinsics matrix.

of the above methods in real-world settings, where a wide range of camera projection models with strong non-linear deformations are common, such as fisheye or panoramic lenses. This limitation is more pronounced when estimating complete metric 3D geometry instead of only depth maps, as the former depends more heavily on the quality of camera estimation. Due to the restrictive assumptions in existing models, general camera estimation can not be effectively learned, even when models are exposed to images from varied camera types. Furthermore, the output space of previous state-of-the-art MMDE methods has inherent limitations, *e.g.* both disparity and log-depth prediction become mathematically ill-posed when the field of view (FoV) exceeds 180 degrees.

To address these challenges, we introduce *UniK3D*, the first framework for monocular metric 3D scene's geometry estimation that generalizes across a wide variety of camera models, from pinhole to fisheye and panoramic configurations, as shown in Fig. 1. Our method proposes a novel formulation for monocular 3D estimation which is spherical in two senses. First, UniK3D leverages a *fully spherical* output 3D space, modeling the range dimension through *radial distance* instead of perpendicular depth. This approach is especially beneficial at large angles from the optical axis, effectively resolving the ill-posed nature of traditional methods at extreme fields of view. Second, while building on the recently proposed decomposition [53] of camera prediction from depth estimation, UniK3D newly presents a general *spherical harmonics basis* as the *direct* output space of the camera module that represents the pencil of rays. Unlike previous works [8, 53] which predict explicit pinhole camera parameters and then encode [53] induced rays using a spherical basis, we remove the camera assumption and directly model the rays. As a result, UniK3D spans an unrestricted space of possible camera models, allowing for flexible and accurate depth prediction regardless of camera intrinsics. Our assumption-free spherical camera representation, with its flexibility, ensures that our model is well-suited for real-world deployment, where capturing scenes with non-standard cameras is common.

Our key contribution is the first camera-universal model for monocular 3D estimation that can accommodate any camera projective geometry. We achieve this through our unified spherical output representation that supports all inverse projection problems. By employing a fully spherical framework, our method ensures a complete disentanglement of projective *vs*. 3D scene geometry, as the dimension of an object projection on the image is a univocal function only w.r.t. radial distance and not w.r.t. depth. This disentanglement allows more consistent 3D reconstructions and enhances the stability of 3D outputs near the $xy$-plane, where depth approaches zero. Moreover, UniK3D models the camera rays as a decomposition across a finite spherical harmonics basis. This choice ensures representation generality and versatility, and at the same time maintains an accurate and compact representation for the resulting pencils of rays, also introducing inductive biases such as continuity and differentiability. In addition, we propose multiple novel strategies to ensure robust camera conditioning of our *radial module* such as an asymmetric angular loss based on quantile regression, static encoding, and curriculum learning.

We validate our approach through extensive zero-shot experiments on 13 widely used metric depth datasets, where UniK3D not only achieves state-of-the-art performance in monocular metric depth and 3D estimation, but also generalizes very well across various camera models, without either preprocessing or specific camera domains during training.

## 2. Related Work

**Monocular Depth Estimation.** The introduction of end-to-end neural networks for MDE, first demonstrated by [15], revolutionized the field by enabling depth prediction through direct optimization, utilizing the Scale-Invariant log loss ($\mathrm{SI}_{\log}$). Since then, the field has evolved with increasingly sophisticated models, ranging from convolutional architectures [18, 30, 38, 51] to recent advancements using transformers [5, 52, 71, 77]. While these approaches have pushed the boundaries of MDE performance in controlled benchmarks, they often fail when faced with zero-shot scenarios, highlighting a persistent challenge: ensuring robust generalization across varying camera and scene domains and diverse geometric and visual conditions.
**Generalizable Monocular Depth Estimation.** To address the limitations of domain-specific models, recent research has focused on developing generalizable and zero-shot MDE techniques. These methods can be categorized into scale-agnostic approaches [27, 56, 65, 72, 73], which aim to mitigate scale ambiguity and emphasize perceptual depth quality, and metric depth models [6, 8, 22, 23, 53, 54, 76], which prioritize accurate geometric reconstruction. However, most existing MDE methods fall short of achieving truly zero-shot monocular metric 3D scene estimation. In particular, scale-agnostic approaches often require additional information to resolve scale ambiguities, while most of the metric-based models depend on a known camera or assume a simplistic pinhole camera configuration. Even the few models which are designed for zero-shot 3D scene estimation [8, 53, 76] remain constrained: they either explicitly assume a pinhole camera model [8, 53] or necessitate image rectification [76], effectively requiring test-time camera information and limiting their zero-shot generalizability to pinhole cameras.

On the contrary, UniK3D addresses these limitations by offering a unified solution that can handle any inverse projection problem. Our model can recover a coherent 3D point cloud from any single image, regardless of camera intrinsics, without any rectification or camera information at test time. This generality sets UniK3D apart, enabling robust and uni-

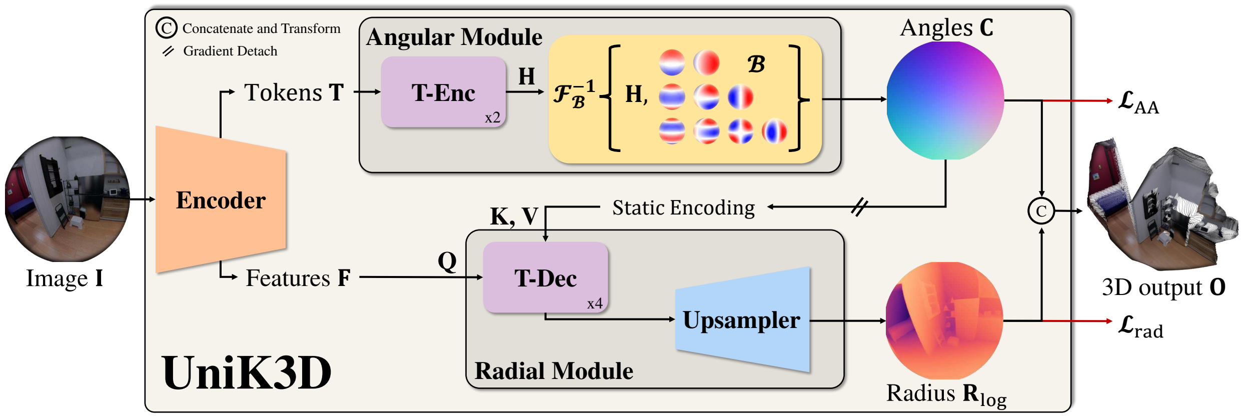


Figure 2. **Model architecture.** UniK3D utilizes solely the single input image to generate the 3D output point cloud ($\mathbf{O}$) for any camera. The projective geometry of the camera is predicted by the Angular Module. The camera representation corresponds to azimuth and polar angles ($\mathbf{C}$) of the backprojected pencil of rays on the unit sphere $\mathbb{S}^3$. The class tokens from the Encoder are processed by 2 Transformer Encoder (T-Enc) layers to obtain the 15 coefficients ($\mathbf{H}$) of the inverse Spherical transform $\mathcal{F}_{\mathcal{B}}^{-1}\{\mathbf{H}\}$ defined by a finite basis ($\mathcal{B}$) of spherical harmonics up to degree 3 with no constant component. Stop-gradient is applied to the angular information which conditions the Radial Module, simulating external information flow. The "static encoding" refers to sinusoidal encoding which matches the internal feature dimensionality. The Radial Module is composed of Transformer Decoder (T-Dec) blocks, one for each input resolution, which is utilized to condition the Encoder features on the bootstrapped camera representation. This conditioning injects prior knowledge on scene scale and projective geometry. The radial output ($\mathbf{R}_{\log}$) is obtained by processing the camera-aware features via a learnable upsampling module. The final output is the concatenation of the camera and radial tensors ($\mathbf{C}||\mathbf{R}_{\log}$). A closed-form coordinate transform is applied to obtain the Cartesian 3D output, but supervision is applied directly on angular coordinates, with our asymmetric angular loss $\mathcal{L}_{AA}$, and radial coordinates.

versal monocular metric 3D estimation that is required in diverse and challenging real-world applications.

**Camera Calibration.** Camera calibration is essential for estimating intrinsic parameters like focal length, principal point, and distortion coefficients to model the mapping from 3D world points to 2D image coordinates. Traditional parametric models, such as the pinhole model, Kannala-Brandt [25], Mei [42], Omnidirection [57], Unified Camera Model (UCM) [20], Enhanced UCM [28], and Double Sphere [61] models are effective for narrow- and wide-angle lenses but require controlled environments for accurate calibration. As models grow more complex, the risk of errors or divergence increases, especially under varying lighting or sensor noise. Additionally, each model has inherent limitations, *e.g*. UCM cannot represent tangential distortion, and Kannala-Brandt struggles beyond a 210° FoV.

By contrast, we take a different approach and model the camera backprojection as a linear combination of *spherical basis* functions, *i.e*. via an inverse spherical harmonics transformation, where the model simply infers the scalar expansion coefficients and the spherical domain boundaries.

## 3. UniK3D

Generalizable depth or 3D scene estimation models often face significant challenges when adapting to diverse camera configurations. Existing methods typically rely on rigid and camera-specific assumptions, such as the pinhole model or equirectangular models, or require extensive preprocessing steps like rectification. These constraints limit their applicability to real-world scenarios with non-standard camera projective geometries. By contrast, our model, UniK3D, introduces a novel framework that enables monocular 3D geometry estimation for any scene and any camera setup.

We begin by introducing the design of our 3D output space and the internal representation of the camera in Sec. 3.1. Our representation is intentionally formulated to be as general as possible, allowing to handle all inverse projection problems. Through our preliminary studies, we observed a consistent issue: the network predictions contracted to a reduced FoV, even when trained on a diverse set of camera types including large FoVs. Simple data re-balancing strategies proved insufficient to address this phenomenon. To overcome this, we have developed a series of architectural and design interventions, detailed in Sec. 3.2, aimed at preventing the backprojection contraction. In Sec. 3.3, we describe the architecture of our model, our optimization strategy, and the specific design and loss functions underpinning our approach. Fig. 2 displays an overview of our method.

### 3.1. Representations

**Output Space.** The output representation of UniK3D is designed to be universally compatible with any scene and camera configuration, providing a direct metric 3D scene estimate for each input image. Drawing from the disentanglement strategy presented in [53], our approach separates camera parameters from scene geometry. Specifically, we represent the camera using a dense tensor $\mathbf{C} = \theta||\phi$, where $\theta$

is the polar angle and $\phi$ is the azimuthal angle, consistent with standard spherical coordinates. However, we use the Euclidean *radius* (distance from the camera center) as the scene range component within a *fully spherical* framework, instead of relying on traditional perpendicular-depth-based representations. This design choice ensures that dimensions of projected objects in the image vary *univocally* with radius, a property that does not characterize the depth representation and renders the latter much harder to learn. Furthermore, the spherical framework enhances numerical stability when handling points near the $xy$-plane, a region where previous methods typically face challenges due to large gradients. We convert the spherical representation to Cartesian coordinates using a bijective transformation, accurately capturing the 3D geometry of the scene as the output 3D point cloud $\mathbf{O}$.

**Camera Internal Space.** In UniK3D, the dense pencil of rays which represents the viewing directions for the various pixels is expressed through a basis decomposition, providing a flexible and comprehensive angular representation. As shown in Fig. 2, our Angular Module predicts a tensor of coefficients $\mathbf{H}$, which is derived from the encoder's class tokens, denoted as $\mathbf{T}$. These coefficients correspond to a predefined basis: the Spherical Harmonics (SH) basis. We reconstruct the pencil of rays from $\mathbf{H}$ as follows:

$$\mathbf{C} = \mathcal{F}_{\mathcal{B}}^{-1}\{\mathbf{H}\} = \sum_{l=0}^{L} \sum_{m=-l}^{l} \mathbf{H}_{lm} \mathcal{B}_{lm}(\theta, \phi), \tag{1}$$

where $\mathbf{C}$ represents the reconstructed angular field and $\mathcal{F}_{\mathcal{B}}^{-1}$ denotes the inverse transform from the coefficient space to the angular space, using the SH basis $\mathcal{B}$. $\mathbf{B}_{lm}(\theta, \phi)$ are the SH basis functions, *i.e*. Legendre polynomials, and $\mathbf{H}_{lm}$ are the predicted coefficients. Here, $l$ and $m$ index the degree and order of the harmonics, respectively. This inverse transform is implemented as an inner product that maps from $\mathbb{R}^n \times \mathbb{S}^3$ to $\mathbb{S}^3$. The SH basis domain is defined by 4 parameters: the generalized "principal point" of the reference frame, *i.e*. the pole, and the horizontal and vertical FoVs. This formulation allows us to describe complex ray distributions *compactly* and implicitly, while ensuring important properties of the output, such as continuity and differentiability.

Additionally, the SH basis offers high sparsity, requiring only 15 harmonics for a 3$^{\text{rd}}$ degree basis without constant component and an equal number of coefficients to accurately represent intrinsics for most camera types. By leveraging this SH-based representation and defining the domain through the pole and FoV parameters, UniK3D achieves a robust and flexible framework that can handle virtually any camera geometry with only 19 parameters.

## 3.2. Preventing Distribution Contraction

**Asymmetric Angular Loss.** Neural networks tend to regress towards the most frequent modes in the training data, often neglecting the distribution tails. In our case, this bias would cause UniK3D to underrepresent wide-FoV angles in its outputs, since most visual datasets are heavily skewed towards small-FoV pinhole cameras. This leads to poor performance in scenarios requiring accurate wide-angle predictions. To overcome this issue, we introduce an asymmetric angular loss based on quantile regression, inspired by robust statistical estimators and decision theory principles, *i.e*. type-I and type-II errors [44]. Our loss function is defined as:

$$\mathcal{L}_{\text{AA}}^{\alpha}(\hat{\theta}, \theta^*) = \alpha \sum_{\hat{\theta} > \theta^*} \left| \hat{\theta} - \theta^* \right| + (1-\alpha) \sum_{\hat{\theta} \leq \theta^*} \left| \hat{\theta} - \theta^* \right|, \tag{2}$$

where $0 \leq \alpha \leq 1$ is the target quantile, $\hat{\theta}$ is the predicted angle, and $\theta^*$ is the ground-truth angle. This formulation adjusts the weighting of over- and underestimations of the polar angle $\theta$. When $\alpha = 0.5$, the loss degenerates to the standard Mean Absolute Error (MAE), but by tuning $\alpha$, we can emphasize underrepresented angles and balance the regression more effectively. Unlike naive dataset rebalancing–which would alter the underlying 3D scene diversity and introduce significant complexity, especially across multiple datasets–our loss addresses the angular imbalance directly and efficiently. By using quantile regression, we minimize the complexity to a simple search over the interval $[0, 1]$ for $\alpha$, making our method well-suited for large-scale and diverse training scenarios. This quantile-based strategy allows us to address the angular distribution bias without sacrificing simplicity and diversity, making it a robust and scalable solution.

**Enhancing Camera Conditioning.** In our initial experiments, we observed that our model struggled to effectively utilize camera conditioning following previous works [53], even when explicitly supplied with ground-truth camera rays during both training and testing. This issue was subtle for small-FoV pinhole cameras, but it became significant for large-FoV configurations. The root of the problem lies in weak conditioning: the model fails to disentangle camera parameters from geometric features, causing it to route local aberrations back to the encoder features' space, without integrating essential FoV information. As a result, even when prompted with accurate camera parameters at test time, the model might ignore, or be misled by, this information.

To address this, we hypothesize that camera data must be clear and explicitly structured from the beginning of training. To this end, we implement in UniK3D a static (non-learnable) encoding of camera rays and adopt a curriculum learning strategy, transitioning gradually from feeding GT camera parameters to predicted ones to the Radial Module. In particular, the GT camera is fed to the Radial Module with probability $1 - \tanh(\frac{s}{10^5})$, where $s$ is the current optimization step. To reinforce external conditioning, we detach gradients from the camera output that is fed to the Radial Module, hence preventing the model from relying on feedback mechanisms that could undermine the conditioning on

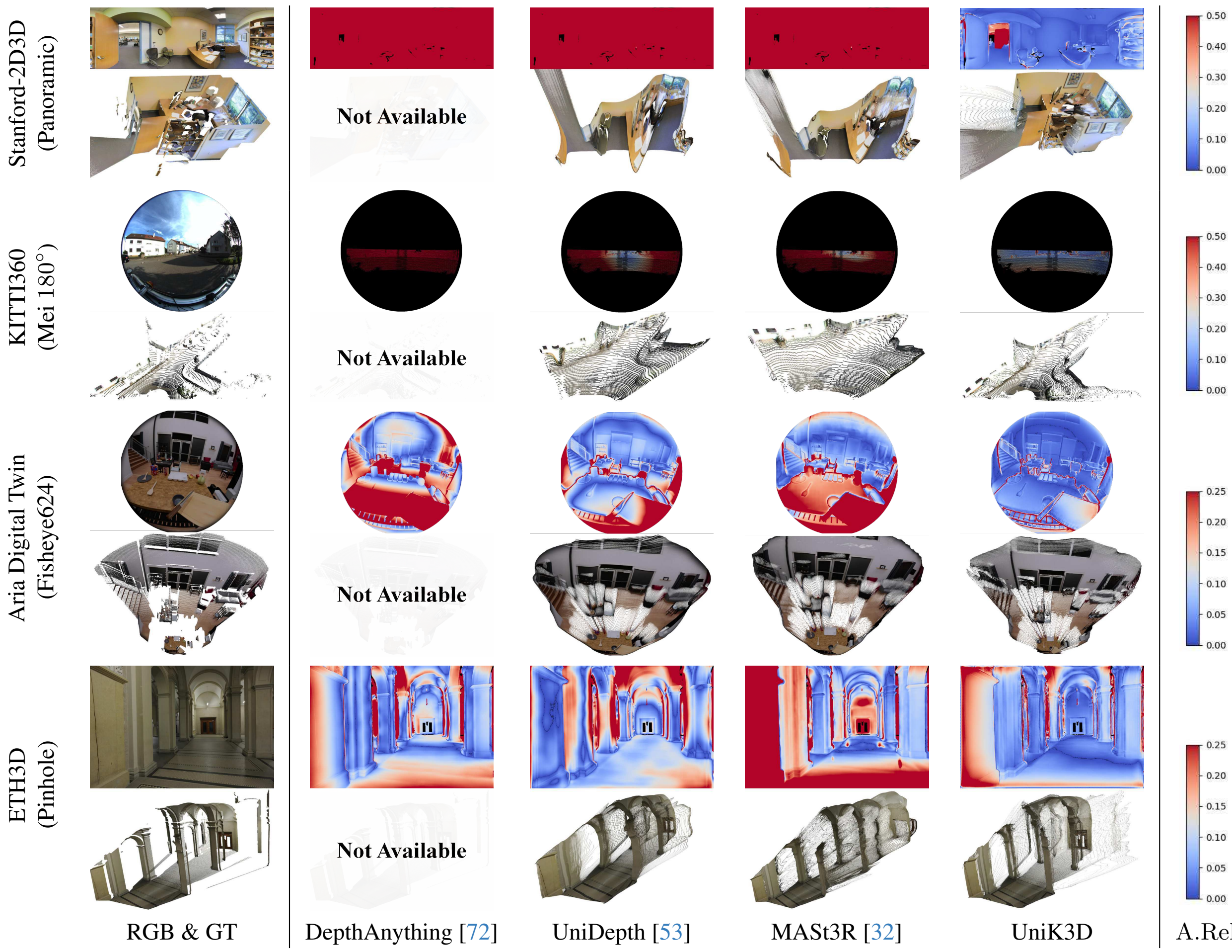


Figure 3. **Qualitative comparisons.** Each pair of consecutive rows represents one test sample. Each odd row displays the input RGB image and the 2D error map, color-coded with the *coolwarm* colormap based on absolute relative error (for panoramic images, the error is computed on distance rather than depth). To ensure a fair comparison, errors are calculated on GT-based shifted and scaled outputs for all models. Each even row shows the ground truth and predictions of the 3D point cloud. The last column displays the specific colormap ranges for absolute relative error. Key observations for each rows pair: (1) competing methods are limited to only positive depth and heavily distort the scenes for larger FoV; (2) in the case of representable but large FoV (180°), UniK3D output is the only one that does not suffer from pronounced FoV contraction; (3) for moderate-FoV images but with strong boundary distortion, *e.g.* fisheye, UniK3D can maintain planarity and overall scene structure; (4) our approach also delivers accurate 3D estimates for standard pinhole images.

the camera. Additionally, we disable learnable gains, such as LayerScale [60], in the cross-attention layers of the Radial Module's transformer decoder, to avoid shortcuts of the conditioning. These strategies ensure that the model effectively leverages camera information to adjust its encoder features, enhancing the robustness of 3D predictions.

### 3.3. Network Design

**Architecture.** Our network consists of an Encoder Backbone, an Angular Module, and a Radial Module, as illustrated in Fig. 2. Our encoder is ViT-based [14] and we extract dense features $\mathbf{F} \in \mathbb{R}^{h \times w \times C \times 4}$–where $(h, w) = (\frac{H}{14}, \frac{W}{14})$–along with class tokens $\mathbf{T}$. The Angular Module processes these class tokens, projecting them onto 512-channel representations that are split into 3 domain parameters and 15 spherical coefficient prototypes. These tokens pass through two layers of a Transformer Encoder (T-Enc) with 8 heads and are then projected onto scalar values. The values for the 3 domain parameters define the principal point (2) and the horizontal FoV (1), determining the intervals for the harmonics. We assume square pixels and thus do not learn an extra, fourth parameter for the vertical FoV, but rather compute this fourth parameter directly from the horizontal FoV. The 15 spherical coefficients undergo an inverse SH transformation according to (1), using a 3-degree SH basis. The gradient flowing from the Angular Module to the class tokens is multiplied by 0.1, as the magnitude of the camera-induced gradient for the encoder weights was empirically found to be ca. 10x higher than the radial-induced gradient.

Table 1. **Comparison on zero-shot evaluation for diverse camera domains.** Validation sets: *S.FoV* includes NYU, KITTI, IBims-1, ETH-3D, nuScenes, and Diode Indoor; *S.FoV$_{Dist}$* includes IBims-1, ETH-3D, and Diode Indoor with synthetic distortion; *L.FoV* includes ADT, ScanNet++ (DSLR), and KITTI360; *Pano* uses Stanford-2D3D. All models use a ViT-L backbone. Missing values (-) indicate the model's inability to produce the respective output. Metric3D and Metric3Dv2 cannot be evaluated on panoramic images as focal lengths are undefined. †: Requires ground-truth (GT) camera for 3D reconstruction. ‡: Requires GT camera for 2D depth map inference.

| **Method** | S.FoV | | | S.FoV$_{\text{Dist}}$ | | | L.FoV | | | Pano | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\delta_1^{\text{SSI}}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\delta_1^{\text{SSI}}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\delta_1^{\text{SSI}}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\delta_1^{\text{SSI}}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ |
| DepthAnything [72] | 92.2 | - | - | 94.3 | - | - | 47.5 | - | - | 10.4 | - | - |
| DepthAnythingv2 [73] | 92.4 | - | - | 88.9 | - | - | 48.7 | - | - | 11.3 | - | - |
| Metric3D†‡ [76] | 86.4 | 43.1 | - | 88.0 | 36.7 | - | 58.7 | 26.0 | - | - | - | - |
| Metric3Dv2†‡ [23] | 91.1 | 59.7 | - | 89.4 | 47.1 | - | 69.2 | 24.7 | - | - | - | - |
| ZoeDepth† [6] | 88.9 | 53.3 | - | 90.3 | 40.1 | - | 65.3 | 6.4 | - | 32.7 | 9.9 | - |
| UniDepth [53] | 94.9 | 59.0 | 85.0 | 94.0 | 43.0 | 70.5 | 68.6 | 16.9 | 19.8 | 33.0 | 2.0 | 1.7 |
| MASt3R [32] | 88.0 | 37.8 | 80.8 | 89.9 | 35.2 | <u>77.1</u> | 67.1 | 29.7 | 25.1 | 32.3 | 3.7 | 2.1 |
| DepthPro [8] | 87.4 | 56.0 | 79.6 | 80.6 | 29.4 | 71.7 | 64.5 | 26.1 | 32.1 | 31.8 | 1.9 | 1.9 |
| UniK3D-Small | 94.3 | 61.3 | 85.7 | 95.1 | 48.4 | 73.8 | 84.5 | 55.5 | 70.1 | 81.3 | 72.5 | <u>53.7</u> |
| UniK3D-Base | <u>95.5</u> | <u>64.9</u> | <u>86.1</u> | <u>96.5</u> | <u>50.2</u> | 75.1 | <u>87.4</u> | <u>67.7</u> | <u>79.9</u> | **83.6** | <u>73.7</u> | <u>53.7</u> |
| UniK3D-Large | **96.1** | **68.1** | **89.4** | **97.3** | **54.5** | **78.8** | **91.2** | **71.6** | **81.9** | <u>81.4</u> | **80.2** | **57.1** |

Table 2. **Zero-shot comparison with equirectangular-specialized methods.** All methods are zero-shot tested on Stanford-2D3D [2]. Competing methods are all trained on equirectangular images. Our training set includes Matterport3D [10] with 2% sampling.

| **Method** | Train | $\delta_1\uparrow$ | A.Rel $\downarrow$ |
|---|---|---|---|
| BiFuse† [63] | Matterport3D | 86.2 | 12.0 |
| BiFuse++† [64] | Matterport3D | <u>91.4</u> | 10.7 |
| UniFuse† [24] | Matterport3D | 91.3 | <u>9.42</u> |
| UniK3D | Ours | **96.8** | **8.01** |

The Radial Module first processes the dense encoder features $\mathbf{F}$ through a Transformer Decoder (T-Dec) with 4 parallel layers, one for each resolution, and 1 head. These layers condition $\mathbf{F}$ on the sine-encoded angular representation $\mathbf{C}$ (cf. supplements). The conditioned features are then projected onto a 512-channel tensor, forming radial features $\mathbf{D} \in \mathbb{R}^{h\times w\times 512}$. These radial features are afterwards upsampled to the input resolution using residual convolutional blocks and learnable upsampling techniques, *i.e.* bilinear upsampling followed by a single $1\times 1$ convolution. The radial log-scale output $\mathbf{R}_{\text{log}} \in \mathbb{R}^{H\times W}$ is computed from the upsampled features and transformed to $\mathbf{R}$ via element-wise exponentiation. The final 3D spherical output $\mathbf{O} = \mathbf{C}||\mathbf{R}$ is converted to a Cartesian point cloud $\mathbf{O} \in \mathbb{R}^{H\times W\times 3}$ using a spherical-to-Cartesian coordinate transformation. Also, we predict a confidence map ($\mathbf{\Sigma}$) for the radial outputs by including a second projection head fed with upsampled $\mathbf{D}$ features, besides the first head of the Radial Module which computes $\mathbf{R}_{\text{log}}$.

**Optimization.** The optimization process is defined by three different losses. The angular loss $\mathcal{L}_{\text{AA}}$ is applied on $\theta$ and $\phi$ separately, with $\mathcal{L}_{\text{AA}}^{0.7}$ and $\mathcal{L}_{\text{AA}}^{0.5}$ for $\theta$ and $\phi$, respectively. The final angular loss can be expressed as

$$\mathcal{L}_{\text{A}}(\hat{\mathbf{C}}, \mathbf{C}^*) = \beta\mathcal{L}_{\text{AA}}^{0.7}(\hat{\theta}, \theta^*) + (1-\beta)\mathcal{L}_{\text{AA}}^{0.5}(\hat{\phi}, \phi^*), \quad (3)$$

with $\hat{(\cdot)}$ and $(\cdot)^*$ serving as prediction and GT identifiers, respectively, and $\beta = 0.75$. It is worth noting that $\mathcal{L}_{\text{AA}}^{0.5}$ corresponds to the standard, symmetric L1-loss, as the azimuthal dimension $\phi$ w.r.t. the principal point is *not* affected by prediction contraction. Our radial loss is the L1-loss between the predicted and GT log-radius obtained by the GT camera and depth: $\mathcal{L}_{\text{rad}} = \left\|\hat{\mathbf{R}}_{\text{log}} - \mathbf{R}^*_{\text{log}}\right\|_1$. The confidence loss is the L1-loss between the detached radial loss and the inverse predicted confidence, $\mathbf{\Sigma}$: $\mathcal{L}_{\text{conf}} = \left\||\hat{\mathbf{R}}_{\text{log}} - \mathbf{R}^*_{\text{log}}| - \mathbf{\Sigma}\right\|_1$. The loss is a linear combination of the three losses: $\mathcal{L}_{\text{A}} + \eta\mathcal{L}_{\text{rad}} + \gamma\mathcal{L}_{\text{conf}}$, with $\eta = 2$ and $\gamma = 0.1$.

# 4. Experiments

**Training Datasets.** The training dataset accounts for 26 different sources: A2D2 [21], aiMotive [41], Argoverse2 [68], ARKit-Scenes [4], ASE [16], BEDLAM [7], BlendedMVS [74], DL3DV [37], DrivingStereo [70], DynamicReplica [26], EDEN [31], FutureHouse [35], HOI4D [39], HM3D [55], Matterport3D [10], Mapillary-PSD [1], MatrixCity [33], MegaDepth [34], NianticMapFree [3], PointOdyssey [79], ScanNet [11], ScanNet++ (iPhone) [75], TartanAir [66], Taskonomy [78], Waymo [59], and WildRGBD [69]. More details are given in the supplement.

**Zero-shot Testing Datasets.** We evaluate the generalizability of models by testing them on 13 datasets not seen during training, grouped in 4 different domains which are defined based on their camera type: 1) small FoV (S.FoV), *i.e.* FoV $< 90°$, 2) small FoV with radial and tangential distortions (S.FoV$_{\text{Dist}}$), 3) large FoV (L.FoV), *i.e.* FoV $> 120°$, and 4) Panoramic (Pano) with $360°$ viewing angle. More specifically, the S.FoV group corresponds to the validation splits of NYU-Depth V2 [43], KITTI Eigen-split [19] and nuScenes [9], and the full IBims-1 [29], ETH-3D [58], and

Table 3. **Ablation on data.** *Data* indicates whether training images include strongly distorted cameras, either from real data or synthesized from pinhole cameras. Output representation: depth.

| | Model | Data | S.FoV | | $\text{S.FoV}_{\text{Dist}}$ | | L.FoV | | Pano | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ |
| 1 | Pinhole | ✗ | 55.1 | 79.2 | 31.7 | 60.0 | 41.2 | 35.1 | 8.4 | 4.2 |
| 2 | Pinhole | ✓ | 56.1 | 81.1 | 40.4 | 58.2 | 44.9 | 43.1 | 5.9 | 3.0 |
| 3 | SH | ✗ | 56.1 | 79.1 | 34.5 | 60.2 | 47.1 | 56.7 | 11.3 | 16.1 |
| 4 | SH | ✓ | 56.2 | 79.4 | 42.1 | 62.7 | 48.5 | 60.8 | 10.9 | 14.8 |

Diode Indoor [62]; the $\text{S.FoV}_{\text{Dist}}$ is composed by images artificially distorted from IBims-1, ETH-3D, and Diode Indoor (more details in the supplement); L.FoV is the mix of ADT [48], ScanNet++ (DSLR) [75], and KITTI360 [36]; and Panoramic (Pano) is to the full Stanford-2D3D [2] dataset.

**Evaluation Details.** All methods have been re-evaluated with a fair and consistent pipeline. In particular, we do not exploit any test-time augmentations and utilize the same set of weights for all zero-shot evaluations. We use the checkpoint corresponding to the zero-shot model for each method, *i.e*. not fine-tuned on KITTI or NYU. The metrics utilized in the main experiments are $\delta_1^{\text{SSI}}$, $\text{F}_\text{A}$, and $\rho_\text{A}$. Further metrics are reported in supplements. $\delta_1^{\text{SSI}}$ measures scale- and shift-invariant depth estimation performance. $\text{F}_\text{A}$ is the area under the curve (AUC) of F1-score [47] up to $1/20$ of the datasets' maximum depth and evaluates monocular 3D estimation. $\rho_\text{A}$ evaluates the camera performance and is the AUC of the average angular error of camera rays up to 15°, 20°, 30° for S.FoV, L.Fov, and Pano, respectively. We avoid parametric evaluations, such as those based on focal length or FoV, because they lack generality across diverse camera models. Instead, our chosen metrics ensure applicability to any camera type, preserving fairness and consistency in evaluation. Supplements show the fine-tuning ability of UniK3D by training the final checkpoint on KITTI and NYU-Depth V2 and evaluating in-domain, as per standard practice.

**Implementation Details.** UniK3D is implemented in PyTorch [50] and CUDA [45]. For training, we use the AdamW [40] optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with an initial learning rate of $5 \times 10^{-5}$. The learning rate is divided by a factor of 10 for the backbone weights for every experiment and weight decay is set to $0.1$. We exploit Cosine Annealing as learning rate scheduler to one-tenth starting from 30% of the whole training. We run 250k optimization iterations with a batch size of 128. The training time amounts to 6 days on 16 NVIDIA 4090. The dataset sampling procedure follows a weighted sampler, where the weight of each dataset is its number of scenes. Our augmentations are both geometric and photometric, *i.e*. random resizing and cropping for the former type, and brightness, gamma, saturation, and hue shift for the latter. We randomly sample the image ratio per batch between 2:1 and 9:16. Our ViT [14] backbone is initialized with weights from DINO-pre-trained [46] models. For the ablations, we run 100k training steps with a ViT-S backbone, with training pipeline as for the main experiments.

Table 4. **Ablation on camera model.** *Model* corresponds to the type of camera model for output rays and internal conditioning: pinhole, Zernike-polynomial coefficients, SH coefficients, or non-parametric, *i.e*. predicting one ray per pixel. All experiments are with full data, augmentation, model components, and radial output.

| | Model | S.FoV | | $\text{S.FoV}_{\text{Dist}}$ | | L.FoV | | Pano | |
|---|---|---|---|---|---|---|---|---|---|
| | | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ |
| 1 | Pinhole | 55.5 | 79.9 | 52.5 | 73.8 | 45.2 | 47.9 | 24.6 | 16.4 |
| 2 | Zernike | 56.6 | 80.9 | 39.9 | 51.3 | 49.9 | 54.6 | 31.8 | 17.9 |
| 3 | Non-Parametric | 56.4 | 81.0 | 45.2 | 62.8 | 42.0 | 42.8 | 51.7 | 20.1 |
| 4 | SH | 57.3 | 79.8 | 44.6 | 59.3 | 53.5 | 64.8 | 58.6 | 26.3 |

Table 5. **Ablation on output representation.** *Output* refers to the type of the 3rd dimension of the predicted output space: either Cartesian z-axis depth or spherical radius, *i.e*. distance. All experiments are with full data and augmentation.

| | Model | Output | S.FoV | | $\text{S.FoV}_{\text{Dist}}$ | | L.FoV | | Pano | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ | $\text{F}_\text{A}\uparrow$ | $\rho_\text{A}\uparrow$ |
| 1 | Pinhole | depth | 56.1 | 81.1 | 40.4 | 58.2 | 44.9 | 43.1 | 5.9 | 3.0 |
| 2 | Pinhole | radius | 56.0 | 81.1 | 39.5 | 57.6 | 44.4 | 48.9 | 10.1 | 4.9 |
| 3 | SH | depth | 56.2 | 79.4 | 42.1 | 62.7 | 48.5 | 60.8 | 10.9 | 14.8 |
| 4 | SH | radius | 56.8 | 76.7 | 35.0 | 43.7 | 51.8 | 61.1 | 53.8 | 22.0 |

## 4.1. Comparison with The State of The Art

Table 1 presents a comprehensive comparison of UniK3D against existing SotA methods across various FoV and image types. Our model consistently outperforms prior models, especially in challenging large-FoV and panoramic scenarios. For instance, in the L.FoV domain, UniK3D achieves a remarkable $\delta_{\text{SSI}}^1$ of $91.2\%$ and $\text{F}_\text{A}$ of $71.6\%$, outperforming the second-best method by more than $20\%$ and $40\%$, respectively. This substantial improvement underscores the robustness of our unified spherical framework in handling wide FoVs. In the Pano category, our model's $\delta_{\text{SSI}}^1$ and $\text{F}_\text{A}$ scores of $71.2\%$ and $66.1\%$ also set the new SotA, demonstrating its ability to effectively reconstruct 3D geometry even under extreme camera setups. These results validate that our design choices, including the SH-based camera model and radial output representation, are crucial for maintaining high performance in complex and diverse camera settings.

In addition, Fig. 3 clearly shows how UniK3D can estimate the 3D geometry of scenes from various and distorted cameras. This is in contrast to other methods that fail when facing unconventional or non-pinhole camera images, as depicted by the 2nd, 3rd, and 4th columns. It is important to highlight that Metric3D, Metric3Dv2, and ZoeDepth are evaluated using GT camera parameters for the $\text{F}_\text{A}$ score, while UniK3D, UniDepth, MASt3R, and DepthPro rely on their predicted cameras. Despite this added difficulty, UniK3D still demonstrates superior 3D reconstruction performance, showcasing its strength in handling real-world conditions where precise camera information is unavailable. Interestingly, our method does not sacrifice performance in more conventional, small-FoV scenarios. UniK3D keeps its top rank, with a $\delta_{\text{SSI}}^1$ of 94.3 in the S.FoV setting, outperforming previously leading methods. This balance highlights that our advancements in L.FoV representation do not undermine

Table 6. **Ablation on network components.** $\mathcal{L}_{AA}$ indicates if our asymmetric angular loss is used, L1-loss otherwise. *Cond* indicates if our design for enhanced camera conditioning from Sec. 3.2 is utilized. All experiments are with full data and augmentations, radial output representation, and an SH-based camera model.

| | $\mathcal{L}_{AA}$ | Cond | S.FoV | | S.FoV$_{\text{Dist}}$ | | L.FoV | | Pano | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\mathrm{F_A}\uparrow$ | $\rho_\mathrm{A}\uparrow$ | $\mathrm{F_A}\uparrow$ | $\rho_\mathrm{A}\uparrow$ | $\mathrm{F_A}\uparrow$ | $\rho_\mathrm{A}\uparrow$ | $\mathrm{F_A}\uparrow$ | $\rho_\mathrm{A}\uparrow$ |
| 1 | ✗ | ✗ | 56.8 | 76.7 | 35.0 | 43.7 | 51.8 | 61.1 | 53.8 | 22.0 |
| 2 | ✓ | ✗ | 57.7 | 80.9 | 39.5 | 52.1 | 52.9 | 64.2 | 56.1 | 24.4 |
| 3 | ✓ | ✓ | 57.3 | 79.8 | 44.6 | 59.3 | 53.5 | 64.8 | 58.6 | 26.3 |

the model's effectiveness for S.FoV tasks. $\mathrm{F_A}$ scores remain high in S.FoV and the $\rho_A$ metric shows that our model consistently provides accurate camera parameter estimation.

Moreover, UniK3D is competitive with specialized methods for equirectangular images, as demonstrated in Table 2. This shows how our model can incorporate different scene and camera domains at training time without compromising any domain-specific performance.

## 4.2. Ablation Studies

**Data.** Table 3 demonstrates the effect of training on datasets with and without large FoV and camera distortions. Incorporating images with strong camera distortions generally enhances performance across all domains, particularly in challenging cases such as S.FoV with distortion and L.FoV. This underscores the importance of diverse camera geometries in the training set to achieve better generalization. However, the improvement on Pano is limited due to the difficulty of representing panoramic images using a log-depth representation.
**Camera Model.** As shown in Table 4, employing SH as the basis for camera rays yields the best overall performance, particularly on L.FoV and Pano. This highlights the effectiveness of our basis function selection in capturing diverse camera models. By contrast, the non-parametric model underperforms in $\mathrm{F_A}$ and $\rho_\mathrm{A}$. Since the latter formulation is purely data-driven, we presume that it requires significantly more data to generalize well. It tends to underrepresent the tails of the data distribution, *i.e.* L.FoV and Pano, while performing adequately on more common domains, *i.e.* S.FoV with or without distortion. The Zernike-polynomial basis [17], typically used for modeling lens aberrations, struggles to represent spherical or equirectangular camera geometries due to its inherent planar structure.
**Output Space.** Table 5 compares different output representations for the third dimension of the predicted space: either the Cartesian z-axis (rows 1 and 3) or the spherical radius (rows 2 and 4). The results show that using the radius representation improves reconstruction metrics in Pano and L.FoV settings, as depth is less effective when dealing with FoVs near or exceeding 180 degrees. This improvement is realized only when the radial component is paired with a camera model capable of representing a wide range of geometries, *e.g.* our SH-based model (row 4 *vs.* row 2). However, the radius-based output space leads to poorer reconstruction for S.FoV with distortion (row 3 *vs.* row 4). This

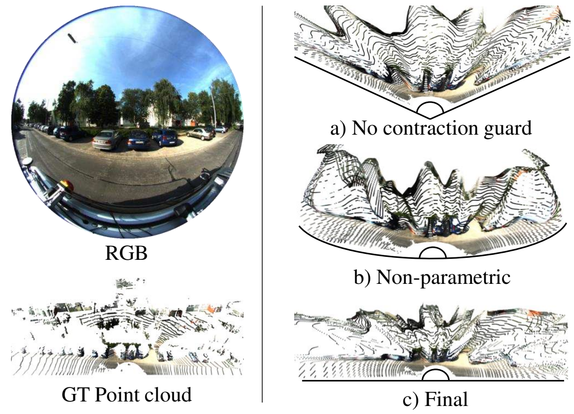


Figure 4. **FoV effects.** The image on the left showcases the challenge of representing the full 180° FoV, alongside the GT point cloud. The effect of FoV contraction occurs when no "guarding", *i.e.* asymmetric loss ($\mathcal{L}_{\text{AA}}$) and camera conditioning, is put in force, as shown in a). The total absence of any prior may lead to impossible and inconsistent backprojection, as shown in b). The final UniK3D is depicted in c), clearly showing the ability to recover large FoVs with a sensible camera backprojection model.

degradation occurs because the radius representation is more sensitive to minor angular variations, which disproportionately impacts accuracy in small but highly distorted views.
**Components.** Table 6 examines the impact of our asymmetric angular loss ($\mathcal{L}_{\text{AA}}$) and our strategies designed to enhance camera conditioning. Our full model, which leverages both the asymmetric loss and the improved conditioning (row 3), significantly outperforms those that do not, especially in distorted and L.FoV domains. This demonstrates the efficacy of our combined strategies in preventing contraction in backprojection and improving angular prediction accuracy. The overall gains are rather due to the synergy of combining these contributions. Moreover, these strategies aim at mitigating extreme cases, which may not be easily represented in aggregate quantitative results, but are clearly visible in qualitative samples as in Fig. 4.

## 5. Conclusion

We have presented UniK3D, the first universal framework for monocular 3D estimation that generalizes seamlessly across diverse camera models, from pinhole to fisheye and panoramic. Our approach introduces strategies to prevent FOV contraction and supports accurate metric 3D estimation through a flexible and robust design for backprojection with any generic camera model. While expanding the diversity and coverage of training data could even further enhance the robustness and applicability of UniK3D, the latter already achieves compelling generalization to unseen cameras and 3D scene domains far beyond the capabilities of the previous state of the art, with only a fair quantity of data.
**Acknowledgment.** This work is funded by Toyota Motor Europe via the research project TRACE-Zürich.

## References


[1] Manuel Lopez Antequera, Pau Gargallo, Markus Hofinger, Samuel Rota Bulò, Yubin Kuang, and Peter Kontschieder. Mapillary planet-scale depth dataset. In *The European Conference on Computer Vision (ECCV)*, pages 589–604. Springer International Publishing, 2020. 6

[2] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 6, 7

[3] Eduardo Arnold, Jamie Wynn, Sara Vicente, Guillermo Garcia-Hernando, Áron Monszpart, Victor Adrian Prisacariu, Daniyar Turmukhambetov, and Eric Brachmann. Map-free visual relocalization: Metric pose relative to a single image. In *European Conference on Computer Vision (ECCV)*, 2022. 6

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Advances in Neural Information Processing Systems (NIPS)*, 2021. 6

[5] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4008–4017, 2020. 2

[6] 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. 2, 6

[7] Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8726–8737, 2023. 6

[8] Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. *arXiv preprint arXiv:2410.02073*, 2024. 1, 2, 6

[9] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[10] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *Proceedings of the International Conference on 3D Vision (3DV)*, 2017. 6

[11] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[12] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 1

[13] Xingshuai Dong, Matthew A Garratt, Sreenatha G Anavatti, and Hussein A Abbass. Towards real-time monocular depth estimation for robotics: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(10):16940–16961, 2022. 1

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2021. 5, 7

[15] 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 (NeurIPS)*, pages 2366–2374. Neural information processing systems foundation, 2014. 2

[16] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 6

[17] Zernike Frits. Beugungstheorie des schneidenver-fahrens und seiner verbesserten form, der phasenkontrastmethode. *Physica: Nonlinear Phenomena*, 1:689–704, 1934. 8

[18] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2002–2011, 2018. 2

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 6

[20] Christopher Geyer and Kostas Daniilidis. A unifying theory for central panoramic systems and practical applications. In *The European Conference on Computer Vision (ECCV)*, 2000. 3

[21] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S. Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, Tiffany Fernandez, Martin Jänicke, Sudesh Mirashi, Chiragkumar Savani, Martin Sturm, Oleksandr Vorobiov, Martin Oelker, Sebastian Garreis, and Peter Schuberth. A2D2: Audi Autonomous Driving Dataset. *arXiv preprint arXiv:2004.06320*, 2020. 6

[22] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rareș Ambruș, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9233–9243, 2023. 1, 2

[23] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *arXiv preprint arXiv:2404.15506*, 2024. 2, 6

[24] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360 panorama

depth estimation. *IEEE Robotics and Automation Letters (RA-L)*, 6(2):1519–1526, 2021. 6

[25] Juho Kannala and Sami Sebastian Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI*, 28:1335–1340, 2006. 3

[26] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Dynamicstereo: Consistent dynamic depth from stereo videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[27] 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 (CVPR)*, pages 9492–9502, 2024. 1, 2

[28] Bogdan Khomutenko, Gaëtan Garcia, and Philippe Martinet. An enhanced unified camera model. *IEEE Robotics and Automation Letters (RA-L)*, 1:137–144, 2016. 3

[29] Tobias Koch, Lukas Liebel, Marco Körner, and Friedrich Fraundorfer. Comparison of monocular depth estimation methods using geometrically relevant metrics on the IBims-1 dataset. *Computer Vision and Image Understanding (CVIU)*, 191:102877, 2020. 6

[30] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. *Proceedings of the International Conference on 3D Vision (3DV)*, pages 239–248, 2016. 2

[31] Hoang-An Le, Thomas Mensink, Partha Das, Sezer Karaoglu, and Theo Gevers. Eden: Multimodal synthetic dataset of enclosed garden scenes. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1579–1589, 2021. 6

[32] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. *arXiv preprint arXiv:2406.09756*, 2024. 5, 6

[33] Yixuan Li, Lihan Jiang, Linning Xu, Yuanbo Xiangli, Zhenzhi Wang, Dahua Lin, and Bo Dai. Matrixcity: A large-scale city dataset for city-scale neural rendering and beyond. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3205–3215, 2023. 6

[34] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2041–2050, 2018. 6

[35] Zhen Li, Lingli Wang, Xiang Huang, Cihui Pan, and Jiaqi. Yang. Phyir: Physics-based inverse rendering for panoramic indoor images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 6

[36] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 2022. 7

[37] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. DL3DV-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22160–22169, 2024. 6

[38] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 38:2024–2039, 2015. 2

[39] Yunze Liu, Yun Liu, Che Jiang, Kangbo Lyu, Weikang Wan, Hao Shen, Boqiang Liang, Zhoujie Fu, He Wang, and Li Yi. Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21013–21022, 2022. 6

[40] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *7th International Conference on Learning Representations, ICLR 2019*, 2017. 7

[41] Tamas Matuszka, Ivan Barton, Ádám Butykai, Péter Hajas, Dávid Kiss, Domonkos Kovács, Sándor Kunsági-Máté, Péter Lengyel, Gábor Németh, Levente Pető, Dezső Ribli, Dávid Szeghy, Szabolcs Vajna, and Balint Viktor Varga. aimotive dataset: A multimodal dataset for robust autonomous driving with long-range perception. In *International Conference on Learning Representations (ICLR) Workshop on Scene Representations for Autonomous Driving*, 2023. 6

[42] Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pages 3945–3950, 2007. 3

[43] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *The European Conference on Computer Vision (ECCV)*, 2012. 6

[44] Jerzy Neyman and Egon Sharpe Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706):289–337, 1933. 4

[45] John Nickolls, Ian Buck, Michael Garland, and Kevin Skadron. Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for? *Queue*, 6(2):40–53, 2008. 7

[46] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 7

[47] Evin Pınar Örnek, Shristi Mudgal, Johanna Wald, Yida Wang, Nassir Navab, and Federico Tombari. From 2d to 3d: Re-thinking benchmarking of monocular depth prediction. *arXiv preprint arXiv:2203.08122*, 2022. 7

[48] Xiaqing Pan, Nicholas Charron, Yongqian Yang, Scott Peters, Thomas Whelan, Chen Kong, Omkar Parkhi, Richard Newcombe, and Yuheng Carl Ren. Aria digital twin: A new benchmark dataset for egocentric 3d machine perception. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20133–20143, 2023. 7

[49] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 8024–8035. Curran Associates, Inc., 2019. 7

[51] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3Depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[52] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. iDisc: Internal discretization for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[53] 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 (CVPR)*, pages 10106–10116, 2024. 1, 2, 3, 4, 5, 6

[54] 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. *arXiv:2502.20110*, 2025. 1, 2

[55] Santhosh Kumar Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alexander Clegg, John M Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (HM3d): 1000 large-scale 3d environments for embodied AI. In *Advances in Neural Information Processing Systems (NIPS)*, 2021. 6

[56] 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 (T-PAMI)*, 44(3):1623–1637, 2020. 1, 2

[57] Davide Scaramuzza. Omnidirectional camera. In *Computer Vision, A Reference Guide*, 2014. 3

[58] Thomas Schöps, Johannes L. Schönberger, Silvano Galliani, Torsten Sattler, Konrad Schindler, Marc Pollefeys, and Andreas Geiger. A multi-view stereo benchmark with high-resolution images and multi-camera videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[59] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2446–2454, 2020. 6

[60] Hugo Touvron, Matthieu Cord, Alexandre Sablayrolles, Gabriel Synnaeve, and Hervé Jégou. Going deeper with image transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 32–42, 2021. 5

[61] Vladyslav C. Usenko, Nikolaus Demmel, and Daniel Cremers. The double sphere camera model. *International Conference on 3D Vision (3DV)*, pages 552–560, 2018. 3

[62] Igor Vasiljevic, Nicholas I. 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. *CoRR*, abs/1908.00463, 2019. 7

[63] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 6

[64] Fu-En Wang, Yu-Hsuan Yeh, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. Bifuse++: Self-supervised and efficient bi-projection fusion for 360 depth estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (T-PAMI)*, 45 (5):5448–5460, 2022. 6

[65] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. *arXiv preprint arXiv:2410.19115*, 2024. 2

[66] 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)*, pages 4909–4916. IEEE, 2020. 6

[67] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 1

[68] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next generation datasets for self-driving perception and forecasting. In *Advances in Neural Information Processing Systems*, 2021. 6

[69] Hongchi Xia, Yang Fu, Sifei Liu, and Xiaolong Wang. Rgbd objects in the wild: Scaling real-world 3d object learning from rgb-d videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22378–22389, 2024. 6

[70] Guorun Yang, Xiao Song, Chaoqin Huang, Zhidong Deng, Jianping Shi, and Bolei Zhou. Drivingstereo: A large-scale dataset for stereo matching in autonomous driving scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[71] Guanglei Yang, Hao Tang, Mingli Ding, Nicu Sebe, and Elisa Ricci. Transformer-based attention networks for continuous

pixel-wise prediction. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16249–16259, 2021. 2

[72] 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 (CVPR)*, pages 10371–10381, 2024. 1, 2, 5, 6

[73] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 2, 6

[74] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1790–1799, 2020. 6

[75] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 6, 7

[76] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3d: Towards zero-shot metric 3d prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9043–9053, 2023. 1, 2, 6

[77] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3906–3915. IEEE, 2022. 2

[78] Amir R Zamir, Alexander Sax, William B Shen, Leonidas Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018. 6

[79] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetzstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 19855–19865, 2023. 6

[80] Brady Zhou, Philipp Krähenbühl, and Vladlen Koltun. Does computer vision matter for action? *Science Robotics*, 4, 2019. 1