This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

UniDAC: Universal Metric Depth Estimation for Any Camera

Girish Chandar Ganesan[1] Yuliang Guo[2] Liu Ren[2] Xiaoming Liu[1,3]
[1]Michigan State University [2]Bosch Research North America & Bosch Center for AI (BCAI)
[3]University of North Carolina at Chapel Hill
ganesang@msu.edu [yuliang.guo2,liu.ren]@us.bosch.com liuxm@cs.unc.edu
https://girish1511.github.io/UniDAC

Abstract

Monocular metric depth estimation (MMDE) is a core challenge in computer vision, playing a pivotal role in real-world applications that demand accurate spatial understanding. Although prior works have shown promising zero-shot performance in MMDE, they often struggle with generalization across diverse camera types, such as fisheye and 360° cameras. Recent advances have addressed this through unified camera representations or canonical representation spaces, but they require either including large-FoV camera data during training or separately trained models for different domains. We propose UniDAC, an MMDE framework that presents universal robustness in all domains and generalizes across diverse cameras using a single model. We achieve this by decoupling metric depth estimation into relative depth prediction and spatially varying scale estimation, enabling robust performance across different domains. We propose a lightweight Depth-Guided Scale Estimation module that upsamples a coarse scale map to high resolution using the relative depth map as guidance to account for local scale variations. Furthermore, we introduce RoPE-ϕ, a distortion-aware positional embedding that respects the spatial warping in Equi-Rectangular Projections (ERP) via latitude-aware weighting. UniDAC achieves state of the art (SoTA) in cross-camera generalization by consistently outperforming prior methods across all datasets.

1. Introduction

Depth estimation plays a pivotal role in bridging the gap between 2D and 3D vision, enabling a wide range of applications such as autonomous vehicles [37, 58], robotics [6, 49], and AR/VR [8, 50].

With the advent of deep learning, early works [3, 10, 29, 39, 66] achieve promising results. However, they fail to generalize across diverse scenes due to scale-depth ambiguity. Recent methods [26, 44, 61, 62, 64] address this limitation by predicting depth up to an unknown scale in addition to leveraging large-scale data [44, 61, 62] and diffusion priors [12, 26, 27]. However, despite their success, these up-to-scale methods encounter limitations in real-world applications where precise spatial measurements are essential, necessitating the inherently challenging and ill-constrained task of monocular metric depth estimation.

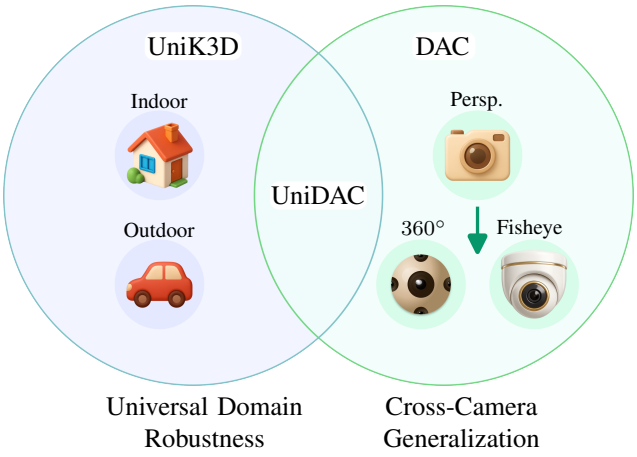
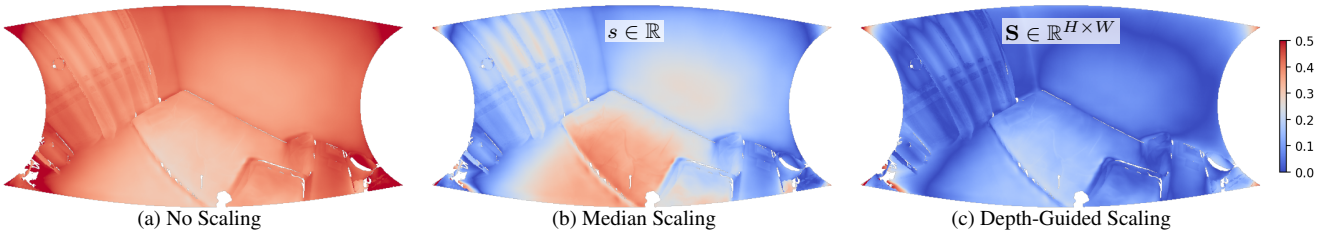


Figure 1. We propose **UniDAC**, a universal, domain-agnostic metric depth estimation framework that generalizes to any camera. Unlike prior methods that either rely on large-FoV data during training or require separate models for indoor and outdoor domains, UniDAC is trained solely on perspective images yet generalizes effectively to large-FoV inputs, leveraging a universal model to robustly handle both indoor and outdoor environments.

Recent works [18, 23, 40, 42, 57, 65] have demonstrated zero-shot performance in MMDE by conditioning upon camera parameters [40, 42], adopting a canonical space [23, 65] or explicitly estimating the scale [57], thereby overcoming the scale-depth ambiguity. However, despite their remarkable performance, these methods have been developed with images from perspective cameras and thereby struggle with large FoV images captured from fisheye or 360° cameras.



(a) No Scaling (b) Median Scaling (c) Depth-Guided Scaling

Figure 2. We show the Abs.Rel error between the predicted relative depth and the ground truth by performing (a) no scaling, (b) median scaling, and (c) depth guided scaling. Theoretically, the relative and metric depth are aligned with a single scale s. Practically, (b) we observe that the irregularities in the relative depth cannot be compensated with a single scalar s. Thus, to tackle this, (c) we propose a Depth-Guided Scale Estimation module that predicts a high-resolution scale map $\mathbf{S}$ respecting local variations.

Recently, [19, 41] demonstrate remarkable MMDE performance on diverse cameras. UniK3D [41] introduces a unified angular representation based on spherical harmonics to represent diverse camera models and subsequently condition the metric depth prediction. While UniK3D achieves strong cross-dataset generalization with a single model covering both indoor and outdoor domains, it relies on training with a diverse set of camera models, including large-FoV data that closely resemble the test distribution. In contrast, DAC [19] achieves strong cross-camera generalization from perspective-only training by adopting an Equi-Rectangular Projection (ERP) as a canonical representation space, along with FoV-aligned and multi-resolution training strategies. However, DAC requires separate models for indoor and outdoor domains, and its performance degrades when using a single unified model across both domains—partly due to its smaller-scale training compared to UniK3D.

In this work, we ask: *Can we design a monocular metric depth estimation (MMDE) framework that generalizes effectively across all camera types and scene domains, without substantially increasing model or dataset scale?* We propose UniDAC, a unified MMDE framework that generalizes across both diverse camera geometries and visual domains within a single model, as illustrated in Fig. 1. Our key insight is to achieve scene–camera unification by decomposing metric depth estimation into relative depth prediction and scale estimation - two components that inherently rely on different contextual scopes and benefit differently from pre-trained foundation model backbones. To this end, we partition encoder features into early and late-stage representations: the former captures local structural details well-suited for relative depth, while the latter encodes global, scene-level context crucial for estimating scale.

Although metric and relative depth are theoretically connected by a single global scalar (scene scale), in practice, relative depth often exhibits spatially non-uniform scaling across the scene, as illustrated in Fig. 2. To handle this, we introduce a Depth-Guided Scale Estimation (DGSE) module that first predicts a coarse, low-resolution scale map from global encoder features. We then upsample this map using the predicted relative depth as a non-parametric guidance signal, ensuring that spatially coherent regions share consistent scale values—while adding negligible computational overhead.

To better harness the potential of modern transformer architectures for large FoV images, we propose RoPE-ϕ, a distortion-aware rotary positional embedding specifically designed for the Equi-Rectangular Projection (ERP) domain. By incorporating latitude-aware weighting, RoPE-ϕ ensures that positional distances more faithfully correspond to geodesic proximity on the sphere. Comprehensive experiments demonstrate that UniDAC achieves robust generalization across both scenes and cameras, substantially outperforming prior SoTA approaches.

We summarize our contributions as follows:

- ✓ We propose UniDAC, a unified framework for monocular metric depth estimation that generalizes across diverse camera models and scene types using a single model.
- ✓ We design a Depth-Guided Scale Estimation module that generates a spatially adaptive scale map with minimal computational overhead.
- ✓ We introduce RoPE-ϕ, a distortion-aware RoPE, that respects the ERP geometry via latitude-based weighting.
- ✓ Extensive experiments show that UniDAC achieves state-of-the-art cross-camera generalization in MMDE.

2. Related Works

2.1. Perspective Monocular Depth Estimation

Monocular depth estimation can be broadly categorized into affine-invariant, up-to-scale depth estimation, and absolute, metric depth estimation.

Prior works [26, 44, 61, 62, 64] achieve notable generalization performance by leveraging large-scale labeled data [44, 64], extensive unlabeled data [61, 62], or strong generative priors [26] of pre-trained diffusion models [48]. However, despite their success, these methods encounter limitations in real-world applications where accurate metric depth is essential.

Recent works [18, 23, 40, 42, 65] exhibit robustness

in the zero-shot monocular metric depth estimation. [18, 23, 65] implicitly infer the scale of the scene by instilling the ground-truth camera parameters [18] or mapping the camera parameters to a canonical space [23, 65]. UniDepth [40, 42] implicitly predicts the scale in the form of camera parameters and applies them as a condition for the metric depth estimation, achieving incredible zero-shot metric depth performance without the need for intrinsic parameters during inference. Despite these advances, none of these methods achieves satisfactory zero-shot performance on large FoV images.

2.2. Large FoV Monocular Depth Estimation

Fisheye and 360° capture rich contextual information due to large FoV and can thereby benefit the task of depth estimation [1, 24, 30, 51, 67]. Due to the scarcity of large-FoV depth data, initial works [11, 24, 30, 46, 51, 55, 60, 67, 70] are constrained to in-domain setting. These works handle distortions by adjusting the convolutional kernels through deformable CNNs [55, 60, 70], partitioning ERP into segments [24, 46], and employing transformers to model the distortions [11, 30, 51, 67]. Although these methods achieve remarkable in-domain performance, they struggle to generalize across diverse cameras.

2.3. Cross-Camera Generalizable MMDE

A key challenge in building a foundational model for diverse cameras is the lack of large-scale, large-FoV datasets. UniK3D [41] combines perspective and large-FoV datasets using a spherical harmonics-based angular representation to share information across camera types. However, its performance is limited by training-time camera diversity and lacks cross-camera generalization. Recent methods [13, 19] achieve strong cross-camera generalization using only perspective images. [13] extends UniDepth [40] to fisheye images via steering tokens but is restricted to perspective and fisheye inputs. DAC [19] introduces a canonical Equi-Rectangular Projection (ERP) space with pitch-aware augmentation to simulate large-FoV images from perspective data, achieving SOTA cross-camera results. However, it requires separate models for indoor and outdoor domains. In contrast, UniDAC achieves cross-camera and cross-domain generalization with a single unified model.

3. Preliminaries

3.1. Metric Depth Decomposition

Given a metric depth map $\mathbf{D}^{\text{m}} \in \mathbb{R}^{H\times W}$, it can be decomposed into a relative depth map $\mathbf{D}^{\text{rel}} \in \mathbb{R}^{H\times W}$ and scale-shift scalars $\{s, t\} \in \mathbb{R}$ as follows:

$$\mathbf{D}^{\text{m}} = s\mathbf{D}^{\text{rel}} + t. \tag{1}$$

The relative depth map $\mathbf{D}^{\text{rel}}$ contains local information, such as object shapes and boundaries, whereas the scale-shift scalars are global factors that depend on the scene as a whole, such as indoor or outdoor. Therefore, the metric depth $\mathbf{D}^{\text{m}}$ can be split into local and global components and are represented using $\mathbf{D}^{\text{rel}}$ and $\{s, t\}$ respectively.

3.2. RoPE

Rotary Positional Embedding (RoPE) was first introduced by [54] in the language domain. RoPE embeds relative positional information among transformer tokens as position-dependent rotations. Specifically, given the n-th token $\mathbf{q}_n \in \mathbb{R}^{1\times d}$, RoPE rotates them using Euler's formula $e^{i\psi}$ to get $\mathbf{q}'_n = \mathbf{q}_n e^{in\psi}$, before applying the attention operations. RoPE practically implements the rotations by first converting the tokens $\mathbf{q}_n$ to the complex domain, $\bar{\mathbf{q}_n} \in \mathbb{C}^{1\times d/2}$, in addition to using multiple frequencies ψ_k across the channel dimensions. The rotations for N tokens are then compactly represented in a rotation matrix $\mathbf{R} \in \mathbb{C}^{N\times d}$ defined as:

$$\mathbf{R}(n, t) = e^{in\psi_k}. \tag{2}$$

[21] later proposes 2D-RoPE, designed to be compatible with vision transformers. Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, [21] replaces the 1D indexing n to 2D pixel indices $\mathbf{p}_n = \{u_n, v_n\} \in \mathbb{R}^{\{0,\dots,H\}\times\{0,\dots,W\}}$. Thus, the rotation matrix in Eq. (2) is modified to:

$$\mathbf{R}(n, 2k) = e^{iu_n\psi_k}; \mathbf{R}(n, 2k+1) = e^{iv_n\psi_k}, \tag{3}$$

where, the frequencies ψ_k are calculated as:

$$\psi_k = \{100^{-4k/d} : k \in \{0, \dots, d/4\}\}. \tag{4}$$

4. UniDAC

We propose UniDAC, a unified framework for monocular metric depth estimation with cross-camera generalization. We posit that one of the main issues with developing a unified model is the variation in depth ranges across different domains. For example, the maximum depth is typically around 10 meters for indoor scenes, whereas it's 80 meters for outdoor scenes. Thus, to prevent the model from getting confused trying to predict varied ranges, we decouple the metric depth prediction into domain-agnostic and domain-specific components, *i.e.*, relative depth map and scene scale, respectively. As mentioned in Sec. 3.1, the relative depth map depends only on local pixel variations in the image; thus, we utilize features rich in local information to estimate relative depth as described in Sec. 4.1. On the other hand, the scale and shift parameters depend on the scene as a whole, and we estimate it using the global information-rich features as detailed in Sec. 4.2. To account for the irregularities in the predicted relative depth, we estimate a high-resolution scale map as opposed to a 1-D scale

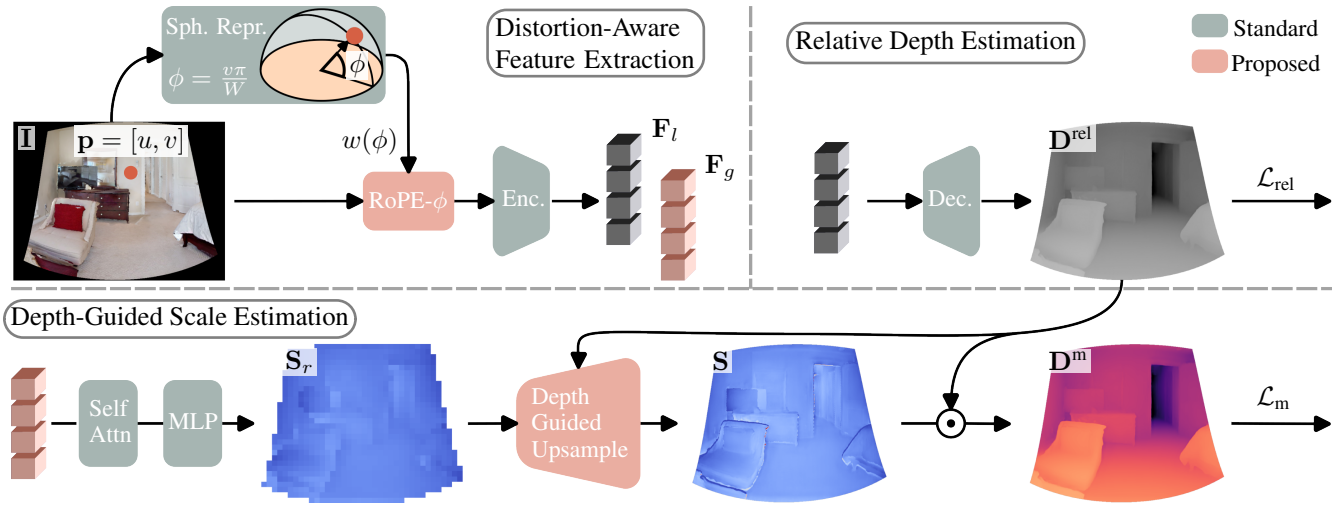


Figure 3. **Overview of proposed method.** UniDAC decouples metric depth estimation into relative depth and scale estimation. Relative depth relies on local scene information, while scene scale is domain-specific and depends on global scene information. Therefore, given an ERP image $\mathbf{I}$, we split the features from the encoder into local $\mathbf{F}_l$ and the global features $\mathbf{F}_g$. We predict the relative depth $\mathbf{D}^{\text{rel}}$ using the local features $\mathbf{F}_l$. We predict a scale map $\mathbf{S}$ from the global features $\mathbf{F}_g$ to account for the irregularities in $\mathbf{D}^{\text{rel}}$. We first predict a low-resolution scale map $\mathbf{S}_r$ and obtain the high-resolution $\mathbf{S}$ through our proposed Depth-Guided Scale (DGS) estimation module. The DGS upsamples $\mathbf{S}_r$ by using the $\mathbf{D}^{\text{rel}}$ as a guide to ensure the upsampling process respects object boundaries. The final metric depth $\mathbf{D}^{\text{m}}$ is calculated using $\mathbf{D}^{\text{rel}}$ and $\mathbf{S}$ as shown in Eq. (8). We introduce distortion-aware positional embedding, termed RoPE-ϕ, that applies a weight $w(\phi)$ to the RoPE rotations based on the latitude ϕ. We train using two losses $\mathcal{L}_{\text{rel}}$ and $\mathcal{L}_{\text{m}}$ applied on $\mathbf{D}^{\text{rel}}$ and $\mathbf{D}^{\text{m}}$, respectively.

through our proposed Depth-Guided Scale Estimation module (Sec. 4.2). Specifically, we achieve higher efficiency by first predicting a low-resolution scalar map (Sec. 4.2.1) and followed by a depth-guided upsampling (Sec. 4.2.2). An overview of our pipeline is shown in Fig. 3.

4.1. Relative Depth Estimation

Given an input image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, we derive set of features $\mathcal{F} = \{\mathbf{F}_l, \mathbf{F}_g\} \in \mathbb{R}^{h\times w\times C}$ from encoder $\mathcal{E}$, where $\mathbf{F}_l, \mathbf{F}_g$ are the local and global features respectively. $\mathbf{F}_l$ and $\mathbf{F}_g$ are split as the outputs from the early and final layers of $\mathcal{E}$, respectively. We then estimate the relative depth map $\mathbf{D}^{\text{rel}} \in \mathbb{R}^{H\times W\times 1}$ from local features $\mathbf{F}_l$ via a decoder $\mathcal{D}$.

The predicted relative depth map can have an arbitrary scale, which can affect the later scale estimation step. We normalize $\hat{\mathbf{D}}^{\text{rel}}$ by median scaling to get $\mathbf{D}^{\text{rel}}$:

$$\mathbf{D}^{\text{rel}} = \frac{\hat{\mathbf{D}}^{\text{rel}}}{\hat{s}}; \quad \hat{s} = \texttt{Median}(\hat{\mathbf{D}}^{\text{rel}}). \tag{5}$$

4.2. Depth-Guided Scale Estimation

The predicted relative depth map $\mathbf{D}^{\text{rel}}$ requires global scalars, *i.e.*, scale s and shift t, to be converted to metric depth. The global scalars are low-dimensional terms dependent on the scene as a whole and are not affected much by local variations in the scene. Therefore, we utilize the global features $\mathbf{F}_g$ from Sec. 4.1 to estimate s and t. Theoretically, the metric and relative depth are related to each other by a couple of 1-D scalars $\{s, t\}$ as shown in Eq. (1). However, in practice, the relative depth predicted by a network may be slightly stretched or compressed in different regions due to local errors or occlusions. Thus, to adjust for the irregularities in $\mathbf{D}^{\text{rel}}$, we predict a scalar map $\mathbf{S} \in \mathbb{R}^{H\times W}$. A naive solution is to predict $\mathbf{S}$ through a series of transpose convolutions, but this would incur an additional computational cost. Moreover, predicting a high-dimensional output $\mathbf{S}$ is challenging and may lead to poor estimates. Thus, we propose a low-resolution scale prediction followed by a lightweight non-parametric depth-guided upsampling to estimate $\mathbf{S}$. We predict t from the `CLS` token of the $\mathbf{F}_g$ by passing it through a shallow MLP.

4.2.1. Patch-level Scale Estimation

The global features are low-resolution features corresponding to non-overlapping patches on the image $\mathbf{I}$. Since we want similar features to have similar scales, we employ a self-attention on $\mathbf{F}_g$ followed by a shallow-MLP to estimate a low-resolution scale map $\mathbf{S}_r \in \mathbb{R}^{h\times w}$. Thus, we get $\mathbf{S}_r = \texttt{MLP}(\texttt{SelfAttn}(\mathbf{F}_g))$.

$\mathbf{S}_r$ is a low-resolution spatial map that must be upsampled to obtain $\mathbf{S}$, while maintaining low computational overhead; hence, we adopt a non-parametric approach. Nearest-neighbor upsampling provides a simple solution, but it does not respect boundaries while upsampling. To ensure non-parametric upsampling while respecting inter-object boundaries, we utilize the predicted relative depth $\mathbf{D}^{\text{rel}}$.

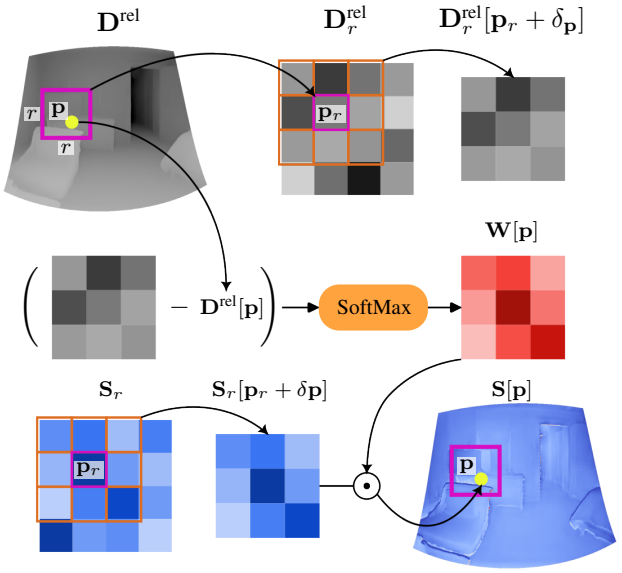


Figure 4. **Depth-Guided Upsampling.** We leverage the predicted relative depth $\mathbf{D}^{\text{rel}}$ as a guide to upsample the predicted low-resolution scale map $\mathbf{S}_r \in \mathbb{R}^{\frac{H}{r} \times \frac{W}{r}}$ to get $\mathbf{S} \in \mathbb{R}^{H \times W}$. We compare $\mathbf{D}^{\text{rel}}$ and its downsampled version $\mathbf{D}_r$ to get the local information in the form of weights $\mathbf{W} \in \mathbb{R}^{H \times W \times 9}$. We compare the spatial mapping between $\mathbf{S}$ and $\mathbf{S}_r$ and combine it with $\mathbf{W}$ to obtain $\mathbf{S}$. The Depth-Guided Upsampling is non-parametric and thus does not add computational overhead.

4.2.2. Depth-Guided Upsampling

The relative depth map $\mathbf{D}^{\text{rel}}$ already has local boundary information, which we leverage as a guide to upsample $\mathbf{S}_r$. We first obtain a low-resolution relative depth map $\mathbf{D}_r^{\text{rel}}$ by applying median pooling on $\mathbf{D}^{\text{rel}}$ with a kernel size and a stride of r. Thus, every pixel $\mathbf{p} = [u, v]$ in $\mathbf{D}^{\text{rel}}$ would be mapped to pixel $\mathbf{p}_r$ in $\mathbf{D}_r^{\text{rel}}$ such that $\mathbf{p}_r = [\lfloor \frac{u}{r} \rfloor, \lfloor \frac{v}{r} \rfloor]$. Then, with a neighborhood $\Omega = \{-1, 0, 1\}^2$ we obtain a distance matrix $\Delta \in \mathbb{R}^{H \times W \times 9}$ defined as follows:

$$\Delta[\mathbf{p}] = \{|\mathbf{D}^{\text{rel}}[\mathbf{p}] - \mathbf{D}_r^{\text{rel}}[\mathbf{p}_r + \delta_\mathbf{p}]| : \forall\ \delta_\mathbf{p} \in \Omega\}. \quad (6)$$

We apply softmax for every pixel over the negative distance values in Δ to get a per-pixel weight map such that $\mathbf{W}[\mathbf{p}] = \texttt{softmax}(-\Delta[\mathbf{p}]) \in \mathbb{R}^{H \times W \times 9}$. Δ and $\mathbf{W}$ contain the pseudo mapping from low-resolution data ($\mathbf{D}_r$) to high-resolution data ($\mathbf{D}$). We use this information as the routing signal to get the high-resolution scale data $\mathbf{S}$ from low-resolution scale data $\mathbf{S}_r$. Finally, the scalar map $\mathbf{S}$ is calculated as the weighted summation over the low-resolution scales as follows:

$$\mathbf{S}[\mathbf{p}] = \mathbf{W}[\mathbf{p}, :]^\intercal \mathcal{N}(\mathbf{S}_r, \mathbf{p}_r), \quad (7)$$

where, $\mathcal{N}(\mathbf{S}_r, p_r) = \{\mathbf{S}_r[\mathbf{p}_r + \delta_\mathbf{p}] : \forall\ \delta_\mathbf{p} \in \Omega\}$ is the neighborhood extractor.

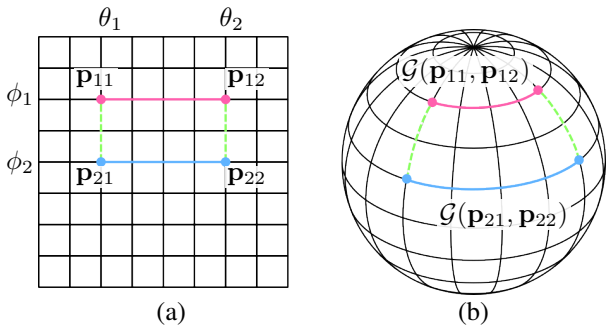


Figure 5. **Motivation for RoPE-ϕ.** We show the difference between (a) the pixel distance in ERP and (b) the corresponding geodesic distance on the curvature of the sphere. Although $|\mathbf{p}_{11} - \mathbf{p}_{12}| = |\mathbf{p}_{21} - \mathbf{p}_{22}|$ in the ERP, we see that $\mathcal{G}(\mathbf{p}_{11}, \mathbf{p}_{12}) < \mathcal{G}(\mathbf{p}_{21}, \mathbf{p}_{22})$ on the sphere. Geodesic distance respects the actual separation in the 3D space. Thus, we modify 2D-RoPE to reflect the geodesic distance to get RoPE-ϕ.

The final metric depth $\mathbf{D}^{\text{m}}$ is then estimated using the Hadamard product $\odot$ as follows:

$$\mathbf{D}^{\text{m}} = \mathbf{S} \odot \mathbf{D}^{\text{rel}} + t. \quad (8)$$

4.3. RoPE-ϕ

We take a step back and observe the encoder $\mathcal{E}$. $\mathcal{E}$ employs 2D-RoPE as the positional embedding scheme as defined in Sec. 3.2. The 2D-RoPE design ensures the relative positional embedding is dependent solely on the pixel locations $\mathbf{p} = [u, v]$. While this is desired for tasks involving perspective images, it is not suitable for ERP representation, which encodes spherical data. ERP is a 2D representation of the surface of the sphere. The pixel distance in ERP is the warped distance between points on the curvature of a sphere, referred to as the geodesic distance. We propose RoPE-ϕ, an enhanced 2D-RoPE that respects the geodesic distance while applying the rotations in 2D-RoPE.

Every pixel in an ERP denotes a direction in the 3D space. Specifically, given an ERP image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, a pixel $\mathbf{p}_n = \{u_n, v_n\}$ denotes direction $\{\theta_n = \frac{u_n}{H}, \phi_n = \frac{v_n}{W}\}$, where θ_n, ϕ_n are the longitude and latitude respectively. Given two pixels $\{\mathbf{p}_1, \mathbf{p}_2\}$ = $\{\{\theta_1, \phi_1\}, \{\theta_2, \phi_2\}\}$ in $\mathbf{I}$, and $\Delta\theta = |\theta_1 - \theta_2|; \Delta\phi = |\phi_1 - \phi_2|$, their geodesic distance on unit-sphere $\mathcal{G}(\mathbf{p}_1, \mathbf{p}_2)$ is defined as:

$$\mathcal{G}(\mathbf{p}_1, \mathbf{p}_2) = \arccos(\sin\phi_1 \sin\phi_2 + \cos\phi_1 \cos\phi_2 \Delta\theta). \quad (9)$$

We observe from Eq. (9) that when $\Delta\theta = 0$, *i.e.*, when the pixels are along the same longitude, $\mathcal{G} \propto \Delta\phi$, aligning with 2D-RoPE design. However, when $\Delta\phi = 0$, *i.e.*, when the pixels are along the same latitude, we can approximate that $\mathcal{G} \propto \cos\phi\Delta\theta$. In other words, given a pair of pixels on the same latitude ϕ and fixed longitudinal distance $\Delta\theta$,

the geodesic distance reduces as we move towards the pole, even though the pixel distance in ERP remains the same, as shown in Fig. 5. With this observation, we add a latitude-based cosine weighting to the 2D-RoPE and modify the rotation matrix $\mathbf{R}$ in Eq. (3) to get $\mathbf{R}_\phi$ as follows:

$$\mathbf{R}_\phi[n] = \mathbf{R}[n]^{w(\phi_n)} \; ; \; w(\phi) = \delta + (1-\delta)\cos\phi, \quad (10)$$

where δ controls the attenuation of weight towards the poles such that $w(\phi) \in [\delta, 1]$.

4.4. Optimization

We train UniDAC by optimizing relative depth loss $\mathcal{L}_{\text{rel}}$ and metric depth loss $\mathcal{L}_{\text{m}}$ on the predicted relative depth map $\mathbf{D}^{\text{rel}}$ and metric depth map $\mathbf{D}^{\text{m}}$, respectively. The losses $\mathcal{L}_{\text{rel}}, \mathcal{L}_{\text{m}}$ are both $\text{SI}_{\log}$ loss widely in MMDE. Given predicted depth map $\mathbf{D}$ and corresponding ground-truth depth map $\bar{\mathbf{D}}$, we estimate the logarithmic error $\epsilon_{\mathbf{p}} = \ln \bar{\mathbf{D}}[\mathbf{p}] - \ln \mathbf{D}[\mathbf{p}]$. The $\text{SI}_{\log}$ is then computed as follows:

$$\mathcal{L}_{SI_{\log}} = \sqrt{\frac{1}{n}\sum_{\mathbf{p}}(\epsilon_{\mathbf{p}})^2 - \frac{\lambda}{n^2}\left(\sum_{\mathbf{p}}(\epsilon_{\mathbf{p}})\right)^2}. \quad (11)$$

Using expectation $\mathbb{E}$ and variance $\mathbb{V}$, we rearrange Eq. (11) to get the following:

$$\mathcal{L}_{SI_{\log}} = \sqrt{\mathbb{V}[\epsilon_{\mathbf{p}}] + (1-\lambda)\mathbb{E}^2[\epsilon_{\mathbf{p}}]}. \quad (12)$$

We can observe from Eq. (12) that by setting $\lambda = 1$ we get a purely scale-invariant loss. The standard $\text{SI}_{\log}$ used in MMDE sets the $\lambda = 0.85$ allowing a mixture of scale-invariant and scale-variant components. We utilize this observation to define the losses $\mathcal{L}_{\text{rel}}$ and $\mathcal{L}_{\text{m}}$ as follows:

$$\mathcal{L}_{\text{rel}} = \mathcal{L}_{SI_{\log}}^{\lambda=1} \; ; \; \mathcal{L}_{\text{m}} = \mathcal{L}_{SI_{\log}}^{\lambda=0.85}. \quad (13)$$

5. Experiments

5.1. Setup

Training Datasets. Our training set comprises three indoor and four outdoor datasets captured from perspective cameras. The indoor datasets consist of HM3D [43], Hypersim [47], and Taskonomy [68], while the outdoor datasets consists of DDAD [17], LYFT [22], Argoverse2 [59], and A2D2 [16]. We use the tiny version of HM3D and Taskonomy provided by OmniData [9] to streamline training. We exclude images of rear-center camera in A2D2 and front-camera in Argoverse2 as explained in Sec. 7.1. Altogether, the training set comprises 1.1M perspective images; 670K from indoor datasets and 780K from outdoor datasets.

Testing Datasets. We evaluate the cross-camera generalization of UniDAC on two fisheye datasets, namely, ScanNet++ [63] and KITTI-360 [31], and two 360° datasets, namely, Pano3D-GV2 [2] and Matterport3D [5].

Baselines. We compare UniDAC to the following works:

- UniK3D [41]: A SoTA method in zero-shot MMDE that does not require any camera parameters in inference.
- DAC [19]: A SoTA method for cross-camera generalization that is trained only on perspective images. DAC_{I}, DAC_{O} denote models trained on indoor and outdoor datasets, respectively. DAC_{U} denotes a model trained from scratch on combined indoor and outdoor datasets. Since it was not released by [19], we train it using the same setup as our method for a fair comparison.

We additionally include UniDepth [40] and Metric3Dv2 [23] evaluated in DAC [19] for comparison.

Evaluation Metrics. We employ standard metrics in MMDE, *i.e.*, percentage of inliers ($\delta_i \uparrow$) with a threshold of 1.25^i, absolute mean relative error (A.Rel $\downarrow$), and root mean squared error (RMSE $\downarrow$).

Implementation. We implement UniDAC in PyTorch [38] and CUDA [36]. We initialize ViT-L [7] backbone with DINO [53] pre-trained weights as our encoder and use DPT [45] as our decoder. We utilize the AdamW optimizer initialized with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.999$, and an initial learning rate of 0.0001. Cosine Annealing, proposed by [33], is employed as the learning rate scheduler with the final learning rate set to one-tenth of the initial learning rate. We train the model for 120k iterations with a batch size of 128. We follow the training setup of [19] and apply FoV-alignment, multi-resolution sampling, and ERP augmentations. All ablations are performed by training on HM3D and KITTI-360 datasets using the ViT-B backbone.

5.2. Comparison with SoTA

Tab. 1 compares UniDAC with [19, 23, 40, 41] in universal domain robustness on large FoV datasets. We observe that UniDAC outperforms all the prior methods trained with perspective images on both indoor and outdoor datasets and sets the SoTA in cross-camera generalization. UniK3D's [41] training set contains images from large FoV cameras, and thus does not constitute a fair comparison to the other methods trained only on perspective images. However, UniDAC still outperforms UniK3D [41], demonstrating the strong generalization prowess of UniDAC. We omit the evaluation on Matterport3D [5] from Tab. 1 since it is present in the training set of [41] and would be an unfair comparison to other methods.

We observe from Tab. 1 that UniDAC achieves $\sim$26% improvement in δ_1 metric in the ScanNet++ dataset compared to both UniK3D and DAC_{U}. While we improve the Pano3D-GV2 by $\sim$8.4% over DAC_{U}, we achieve similar performance as that of UniK3D. As mentioned previously, UniK3D contains Matterport3D, a 360° dataset similar to Pano3D-GV2, and we still come close to UniK3D on the Pano3D-GV2 dataset, demonstrating the robustness of our

Table 1. **Zero-shot evaluation of universal domain robustness**. We evaluate all unified methods on a zero-shot setting across both indoor and outdoor domains. All models are trained on a mix of indoor and outdoor datasets. UniK3D is trained on large FoV images, while the rest of the methods are only trained on perspective images. [Key: **Best**, Second Best, S: Swin-L [32], V: ViT-L [7]]

Methods	Dataset Size	ScanNet++			Pano3D-GV2			KITTI-360		
		$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
UniK3D [41] - V	7.94M	0.651	0.253	0.285	**0.785**	0.170	0.400	0.817	0.244	**2.400**
Metric3Dv2 [23] - V	16.20M	0.536	0.223	0.895	0.404	0.307	0.855	0.716	0.200	4.580
UniDepth [40] - V	3.83M	0.364	0.497	1.166	0.247	0.789	0.127	0.481	0.294	6.564
DAC_U [19] - S	0.79M	0.658	0.233	0.464	0.684	0.203	0.507	0.708	0.186	5.079
UniDAC- V	1.45M	**0.918**	**0.097**	**0.277**	0.768	**0.161**	**0.394**	**0.836**	**0.141**	3.977

Table 2. **Zero-shot evaluation of cross-camera generalization**. We evaluate all methods on a zero-shot setting across diverse cameras. $*_I$, $*_O$ indicate models trained only on the indoor and outdoor datasets, respectively. [Key: **Best**, Second Best, R: ResNet101 [20], S: Swin-L [32], V: ViT-L [7]]

Methods	ScanNet++			Pano3D-GV2			Matterport3D			KITTI-360		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
DAC_I [19] - R	0.852	0.132	0.309	0.812	0.139	0.478	0.773	0.156	0.619	0.082	0.464	10.046
DAC_I [19] - S	0.854	0.128	0.287	0.729	0.184	0.483	0.723	0.179	0.591	0.245	0.365	9.508
DAC_O [19] - R	0.256	0.901	2.312	0.340	0.616	1.713	0.390	0.548	1.630	0.786	0.156	4.361
DAC_O [19] - S	0.109	1.412	3.539	0.323	0.870	2.056	0.330	0.924	2.164	0.822	0.149	3.751
DAC_U [19] - S	0.658	0.233	0.464	0.684	0.203	0.507	0.662	0.215	0.662	0.708	0.186	5.079
UniDAC- V	**0.918**	**0.097**	**0.277**	**0.768**	**0.161**	**0.394**	**0.745**	**0.175**	**0.442**	**0.836**	**0.141**	**3.977**

method. We outperform prior methods on KITTI-360, including UniK3D by $\sim$2%, even though [41] contains aiMotive [34], an outdoor fisheye dataset in its training set.

5.2.1. Comparison with DAC

Tab. 2 compares UniDAC against unified and domain specific DAC models across all testing datasets. As expected, DAC_I [19] and DAC_O [19] performance drops catastrophically when evaluated using outdoor and indoor datasets, respectively. Although DAC_U [19] is trained on both indoor and outdoor data, it fails to generalize well since it tries to learn a single global scale for both domains. UniDAC outperforms DAC_U, consistently and significantly across indoor and outdoor domains, underscoring the effect of decoupling metric depth into relative and scale estimation.

In addition to outperforming DAC_U, Tab. 2 shows that UniDAC outperforms DAC_I on ScanNet++. Although UniDAC beats DAC_I with the Swin backbone on Pano3D-GV2 and Matterport3D, we fall short of beating DAC_I with the ResNet backbone. We believe this is because transformers struggle with scale equivariance as noted in [19]. While comparing between transformer backbones, *i.e.*, Swin and ViT, we outperform DAC_I on Pano3D-GV2 and Matterport3D with $\sim$4% and $\sim$2.2% respectively.

5.3. Ablation Studies

Impact of Scale Map. We analyze the benefits of scale map estimation by comparing with single-scale estimation and direct metric-depth estimation without decoupling. For direct metric-depth estimation, we remove the scale estimation module and use the decoder's output as the metric depth. We keep the relative depth estimation architecture the same, but replace the Depth-Guided Scale estimation module with a single scalar estimation. We follow [57] and utilize the `[CLS]` token from the global features $\mathbf{F}_g$ of the encoder. We estimate the scale as $s = \texttt{MLP}(\mathbf{F}_g\texttt{[CLS]}) \in \mathbb{R}$, where `MLP` is a shallow fully-connected network.

Table 3. **Ablation on Scale Estimation.** We study the impact of depth-guided scale estimation by comparing it against single-scale estimation. The model performances are compared in a zero-shot setting. '-' indicates we directly estimate metric depth without decoupling relative depth and metric scale. [Key: **Best**]

Scale	ScanNet++			KITTI-360		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
–	0.782	0.152	0.401	0.563	0.286	5.461
$s \in \mathbb{R}$	0.773	0.166	0.426	0.601	0.245	5.257
$\mathbf{S} \in \mathbb{R}^{H \times W}$	**0.792**	**0.140**	**0.396**	**0.622**	**0.239**	**5.057**

Tab. 3 shows that directly predicting metric depth performs well on ScanNet++ but fails on KITTI-360. This is due to strong indoor bias in the ablation data: HM3D (310K) vs. DDAD (80K). Predicting a scale map $\mathbf{S}$ gives 2% performance gain over single-scale and thus validates our approach.

Impact of RoPE-ϕ. We evaluate the impact of our proposed distortion-aware positional embedding RoPE-ϕ by

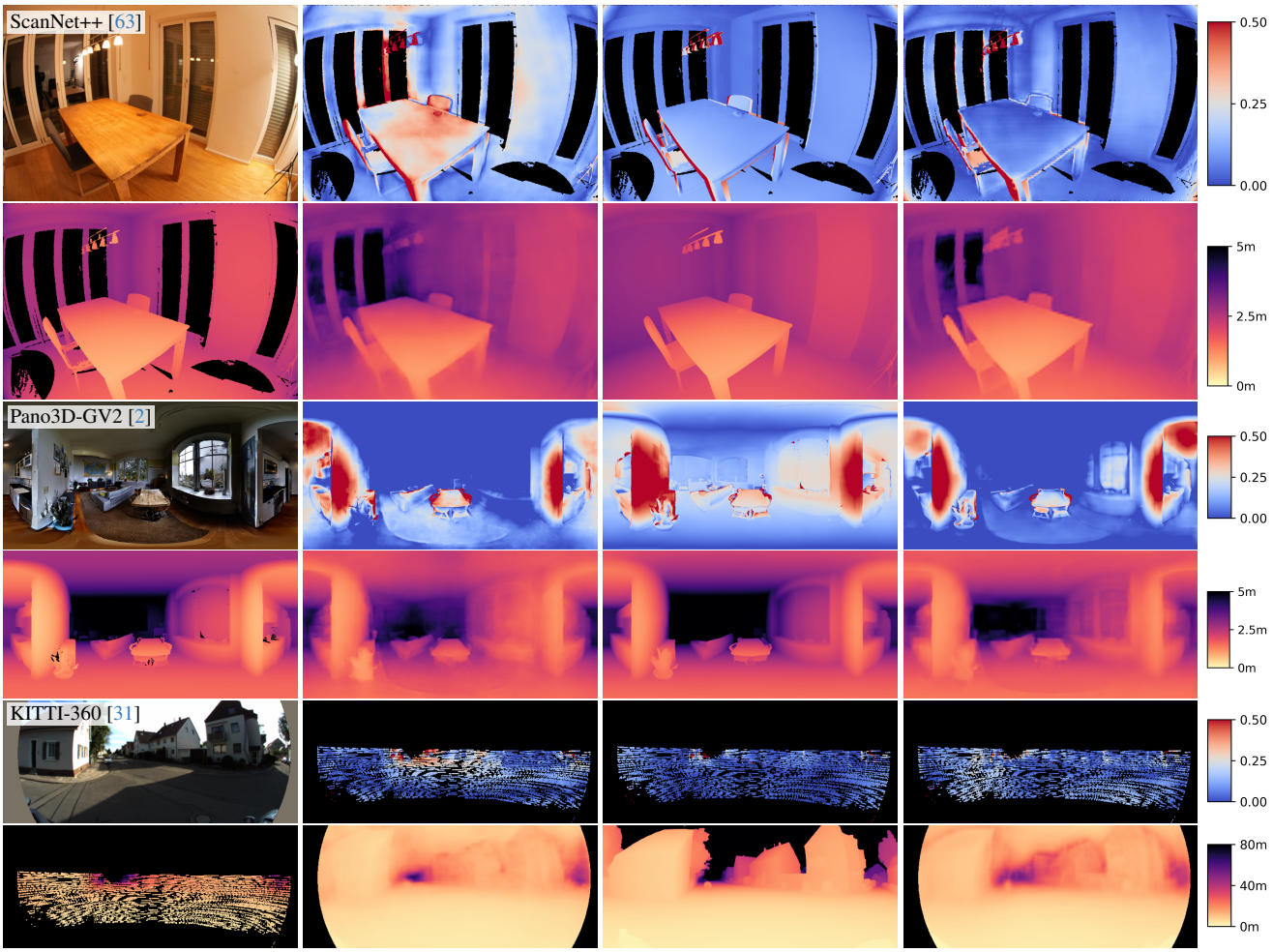


Figure 6. **Qualitative Results.** Every pair of consecutive rows corresponds to a single sample. Odd rows display the input RGB image, and A.Rel error between predicted and GT depth maps. Even rows display the GT depth map and predicted depth maps. UniDAC reduces error in the distorted region as seen around the table in ScanNet++ compared to DAC$_U$ and on the walls in Pano3D-GV2 compared to UniK3D.

comparing it with 2D RoPE defined in Sec. 3.2. We can observe from Eq. (10) that simply setting $\delta = 1$ would give us 2D RoPE, thus indicating that 2D RoPE is a special case of RoPE-ϕ. We observe from Tab. 4 that ScanNet++ benefits from RoPE-ϕ more than KITTI-360. We believe that this is because KITTI-360 contains valid depth along a narrow band around close to the equator as opposed to ScanNet++, which provides a dense-depth map spanning a large FoV.

Table 4. **Ablation on RoPE.** We evaluate the benefit of RoPE-ϕ, comparing it against a model trained with 2D RoPE. The model performances are compared in a zero-shot setting. [Key: **Best**]

Pos. Emb	ScanNet++			KITTI-360		
	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$	$\delta_1 \uparrow$	A.Rel $\downarrow$	RMSE $\downarrow$
RoPE	0.750	0.177	0.452	0.592	0.254	5.313
RoPE-ϕ	**0.792**	**0.140**	**0.396**	**0.622**	**0.239**	**5.057**

6. Conclusion

We presented UniDAC, a unified framework for monocular metric depth estimation that universally generalizes across diverse camera models and scene domains using a single model. By decoupling metric depth into relative depth and scale map, we enable robust estimation across indoor and outdoor domains with a single model. Our proposed Depth-Guided Scale Estimation effectively fuses global and local information to produce scale-aware predictions with minimal overhead. Furthermore, we introduced RoPE-ϕ, a distortion-aware positional encoding tailored for ERP images, enhancing the model's performance on large FoV images. Extensive experiments demonstrate that UniDAC achieves state-of-the-art performance in cross-camera and cross-domain generalization, establishing a strong foundation for scalable and universal depth perception systems.

References

[1] Hao Ai, Zidong Cao, Yan-Pei Cao, Ying Shan, and Lin Wang. Hrdfuse: Monocular 360° depth estimation by collaboratively learning holistic-with-regional depth distributions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023. 3

[2] Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Pano3d: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3727–3737, 2021. 6, 8, 1, 3, 5

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4009–4018, 2021. 1

[4] 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*, pages 11621–11631, 2020. 2

[5] 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. *arXiv preprint arXiv:1709.06158*, 2017. 6, 1

[6] 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

[7] 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. 6, 7

[8] Ruofei Du, Eric Turner, Maksym Dzitsiuk, Luca Prasso, Ivo Duarte, Jason Dourgarian, Joao Afonso, Jose Pascoal, Josh Gladstone, Nuno Cruces, et al. Depthlab: Real-time 3d interaction with depth maps for mobile augmented reality. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, pages 829–843, 2020. 1

[9] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 6

[10] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 1

[11] Hao Feng, Wendi Wang, Jiajun Deng, Wengang Zhou, Li Li, and Houqiang Li. Simfir: A simple framework for fisheye image rectification with self-supervised representation learning. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 3

[12] Xiao Fu, Wei Yin, Mu Hu, Kaixuan Wang, Yuexin Ma, Ping Tan, Shaojie Shen, Dahua Lin, and Xiaoxiao Long. Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024. 1

[13] Suchisrit Gangopadhyay, Jung-Hee Kim, Xien Chen, Patrick Rim, Hyoungseob Park, and Alex Wong. Extending foundational monocular depth estimators to fisheye cameras with calibration tokens. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5198–5209, 2025. 3

[14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[15] Christopher Geyer and Kostas Daniilidis. A unifying theory for central panoramic systems and practical implications. In *European conference on computer vision*, pages 445–461. Springer, 2000. 1, 2

[16] Jakob Geyer, Yohannes Kassahun, Mentar Mahmudi, Xavier Ricou, Rupesh Durgesh, Andrew S Chung, Lorenz Hauswald, Viet Hoang Pham, Maximilian Mühlegg, Sebastian Dorn, et al. A2d2: Audi autonomous driving dataset. *arXiv preprint arXiv:2004.06320*, 2020. 6, 1

[17] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 6, 1

[18] 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*, pages 9233–9243, 2023. 1, 2, 3

[19] Yuliang Guo, Sparsh Garg, S Mahdi H Miangoleh, Xinyu Huang, and Liu Ren. Depth any camera: Zero-shot metric depth estimation from any camera. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26996–27006, 2025. 2, 3, 6, 7, 8, 1, 4, 5

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[21] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024. 3, 2

[22] John Houston, Guido Zuidhof, Luca Bergamini, Yawei Ye, Long Chen, Ashesh Jain, Sammy Omari, Vladimir Iglovikov, and Peter Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *Conference on Robot Learning*, pages 409–418. PMLR, 2021. 6, 1

[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. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10579–10596, 2024. 1, 2, 3, 6, 7

[24] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics Autom. Lett.*, 2021. 3

[25] Juho Kannala and Sami S Brandt. A generic camera model and calibration method for conventional, wide-angle, and fish-eye lenses. *IEEE transactions on pattern analysis and machine intelligence*, 28(8):1335–1340, 2006. 1

[26] 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*, 2024. 1, 2

[27] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv preprint arXiv:2505.09358*, 2025. 1

[28] Tobias Koch, Lukas Liebel, Friedrich Fraundorfer, and Marco Korner. Evaluation of cnn-based single-image depth estimation methods. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2

[29] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 1

[30] Yuyan Li, Yuliang Guo, Zhixin Yan, Xinyu Huang, Ye Duan, and Liu Ren. Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022. 3

[31] 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*, 45(3):3292–3310, 2022. 6, 8, 1, 2, 3

[32] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 7

[33] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*, 2016. 6

[34] 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 2023 Workshop on Scene Representations for Autonomous Driving*, 2023. 7

[35] Christopher Mei and Patrick Rives. Single view point omnidirectional camera calibration from planar grids. In *Proceedings 2007 IEEE International Conference on Robotics and Automation*, pages 3945–3950. IEEE, 2007. 1, 2

[36] 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. 6

[37] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3142–3152, 2021. 1

[38] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[39] 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*, pages 21477–21487, 2023. 1, 2

[40] 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*, pages 10106–10116, 2024. 1, 2, 3, 6, 7

[41] Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1028–1039, 2025. 2, 3, 6, 7, 8, 1, 4, 5

[42] 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, 2025. 1, 2, 3

[43] Santhosh K Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X Chang, et al. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai. *arXiv preprint arXiv:2109.08238*, 2021. 6, 1

[44] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 1, 2

[45] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12179–12188, 2021. 6

[46] Manuel Rey, Mingze Yuan Area, and Christian Richardt. 360monodepth: High-resolution 360 monocular depth estimation. in 2022 ieee. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[47] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb,

and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *International Conference on Computer Vision (ICCV) 2021*, 2021. 6, 1

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[49] Anupa Sabnis and Leena Vachhani. Single image based depth estimation for robotic applications. In *2011 IEEE Recent Advances in Intelligent Computational Systems*, pages 102–106. IEEE, 2011. 1

[50] Irawati Nurmala Sari, Weiwei Du, et al. Depth map estimation of single-view image using smartphone camera for a 3-dimension image generation in augmented reality. In *2023 Sixth International Symposium on Computer, Consumer and Control (IS3C)*, pages 167–170. IEEE, 2023. 1

[51] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360$ˆ{\circ }$ depth estimation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, 2022. 3

[52] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 2

[53] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 6

[54] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 3

[55] Yu-Chuan Su and Kristen Grauman. Learning spherical convolution for fast features from 360° imagery. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 2017. 3

[56] Javier Tirado-Garín and Javier Civera. Anycalib: On-manifold learning for model-agnostic single-view camera calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8044–8055, 2025. 1

[57] Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. Moge-2: Accurate monocular geometry with metric scale and sharp details. *arXiv preprint arXiv:2507.02546*, 2025. 1, 7

[58] 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

[59] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, et al. Argoverse 2: Next generation datasets for self-driving perception and forecasting. *arXiv preprint arXiv:2301.00493*, 2023. 6, 1

[60] Yuwen Xiong, Zhiqi Li, Yuntao Chen, Feng Wang, Xizhou Zhu, Jiapeng Luo, Wenhai Wang, Tong Lu, Hongsheng Li, Yu Qiao, Lewei Lu, Jie Zhou, and Jifeng Dai. Efficient deformable convnets: Rethinking dynamic and sparse operator for vision applications. 2024. 3

[61] 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*, pages 10371–10381, 2024. 1, 2

[62] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 1, 2

[63] 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*, pages 12–22, 2023. 6, 8, 1, 3, 4

[64] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021. 1, 2

[65] 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*, pages 9043–9053, 2023. 1, 2, 3

[66] 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*, pages 3916–3925, 2022. 1

[67] Ilwi Yun, Chanyong Shin, Hyunku Lee, Hyuk-Jae Lee, and Chae-Eun Rhee. Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 3

[68] Amir R Zamir, Alexander Sax, William Shen, Leonidas J Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3712–3722, 2018. 6, 1

[69] Shengjie Zhu, Girish Chandar Ganesan, Abhinav Kumar, and Xiaoming Liu. Replay: Remove projective lidar depthmap artifacts via exploiting epipolar geometry. In *European Conference on Computer Vision*, pages 393–411. Springer, 2024. 2

[70] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets V2: more deformable, better results. 2019. 3