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.

Depth Any Camera: Zero-Shot Metric Depth Estimation from Any Camera

Yuliang Guo[1*†], Sparsh Garg[2†], S. Mahdi H. Miangoleh[3], Xinyu Huang[1], Liu Ren[1]
[1]Bosch Research North America & Bosch Center for Artificial Intelligence (BCAI)
[2]Carnegie Mellon University [3]Simon Fraser University
[1][yuliang.guo2,xingyu.huang,liu.ren]@us.bosch.com
[2]sparshg@andrew.cmu.edu [3]smh31@sfu.ca
https://yuliangguo.github.io/depth-any-camera

Abstract

While recent depth foundation models exhibit strong zero-shot generalization, achieving accurate metric depth across diverse camera types—particularly those with large fields of view (FoV) such as fisheye and 360-degree cameras—remains a significant challenge. This paper presents Depth Any Camera (DAC), a powerful zero-shot metric depth estimation framework that extends a perspective-trained model to effectively handle cameras with varying FoVs. The framework is designed to ensure that all existing 3D data can be leveraged, regardless of the specific camera types used in new applications. Remarkably, DAC is trained exclusively on perspective images but generalizes seamlessly to fisheye and 360-degree cameras without the need for specialized training data. DAC employs Equi-Rectangular Projection (ERP) as a unified image representation, enabling consistent processing of images with diverse FoVs. Its core components include pitch-aware Image-to-ERP conversion with efficient online augmentation to simulate distorted ERP patches from undistorted inputs, FoV alignment operations to enable effective training across a wide range of FoVs, and multi-resolution data augmentation to further address resolution disparities between training and testing. DAC achieves state-of-the-art zero-shot metric depth estimation, improving δ_1 accuracy by up to 50% on multiple fisheye and 360-degree datasets compared to prior metric depth foundation models, demonstrating robust generalization across camera types.

1. Introduction

Depth estimation from monocular cameras is a foundational challenge for applications like autonomous driving, AR/VR, and robotics. While early deep learning methods relied on supervised training using single datasets and depth sensor [24] supervision [3, 25, 36], monocular depth estimation remains challenging due to scale-depth ambiguity. Expanding training datasets has been a key strategy to enhance robustness, with self-supervised approaches using sequential frames [13, 46, 48]. However, these methods often underperform due to self-supervision ambiguity, view inconsistencies and dynamic objects. Recent methods, such as MiDaS [37], leverage large-scale datasets with 3D supervision, normalizing scale differences across datasets to enable zero-shot testing. However, they primarily provide relative depth rather than metric depth.

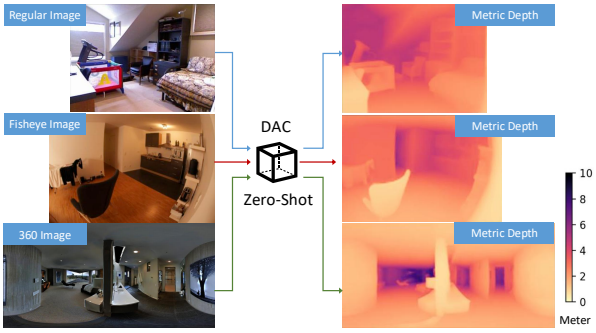

Figure 1. We introduce Depth Any Camera (DAC) framework, which leverages large-scale datasets containing perspective camera images to train a single depth estimation model capable of conducting zero-shot metric depth estimation on images captured various types of cameras, including those captured from large FoV fisheye and 360° cameras.

Recent methods tackle zero-shot metric depth estimation by addressing the challenges of inconsistent scaling factors in depth ground truth caused by varying camera intrinsic parameters. Several works demonstrate impressive generalization capabilities on novel images [4, 20, 33, 53, 57], establishing themselves as foundational depth models for downstream tasks. However, these approaches often struggle with large field-of-view (FoV) cameras like fisheye and

*Corresponding author.
†Equal technical contribution.

360° cameras, where performance significantly declines compared to standard perspective cameras.

As illustrated in Fig. 2, large FoV images can be represented in multiple formats, but generating the best-performing undistorted image for perspective-based depth models often leads to substantial FoV loss. Despite these limitations, large FoV inputs are crucial for efficiency-critical downstream applications such as large-scale detection [34, 52], segmentation [55, 58], SLAM [11, 44, 45, 62], interactive 3D scene generation [59], and robotic demonstration capturing [6, 14, 49].

Achieving zero-shot depth generalization across any FoV camera presents several challenges: (1) selecting a unified camera model to represent diverse FoVs, (2) effectively leveraging perspective training datasets to generalize to data spaces observable only from large FoV cameras, (3) managing drastically different training image sizes in the unified space caused by varying FoVs, and (4) handling resolution inconsistencies between training and testing phases.

In this paper, we present **Depth Any Camera (DAC)**, a novel zero-shot metric depth estimation framework that enables a depth model trained exclusively on perspective images to generalize across cameras with widely varying FoVs, including fisheye and 360° cameras (see Fig. 1). DAC employs Equi-Rectangular Projection (ERP) as a canonical representation to unify images from diverse FoVs into a shared space. A key innovation is the introduction of an efficient **pitch-aware Image-to-ERP conversion** based on grid sampling and Gnomonic Geometry [47], enabling seamless ERP-space data augmentations. Specifically, pitch-aware ERP conversion with pitch-angle augmentation projects perspective images into high-distortion regions of the ERP space, effectively simulating observations unique to large-FoV cameras. This enhances DAC's zero-shot generalization, allowing it to extrapolate beyond the perspective domain to a broader range of camera types. To facilitate learning from mixed datasets, we propose a **FoV alignment** process that normalizes diverse-FoV training samples to a predefined ERP patch size, preserving content while minimizing computational overhead. Additionally, multi-resolution augmentation is applied to address resolution mismatches, allowing the model to learn scale-equivariant features and adapt to a flexible range of testing resolutions. In summary, our contributions are as follows:

- We propose a novel zero-shot metric depth estimation framework capable of handling images from any camera type, including fisheye and 360° images, using a model trained exclusively on perspective data.
- We introduce an efficient pitch-aware perspective-to-ERP patch conversion that simulates the high-distortion characteristics of large-FoV cameras from perspective inputs, enhancing zero-shot generalization.
- We develop a FoV alignment process that enables effective training across cameras with diverse FoVs within a unified ERP space, along with a multi-resolution training strategy to address resolution mismatches between training ERP patches and testing images.
- Our method achieves State-of-The-Art (SoTA) zero-shot performance on all large FoV testing datasets, delivering up to a 50% improvement in δ_1 accuracy on indoor fisheye and 360° datasets, showcasing strong generalization across diverse camera types.

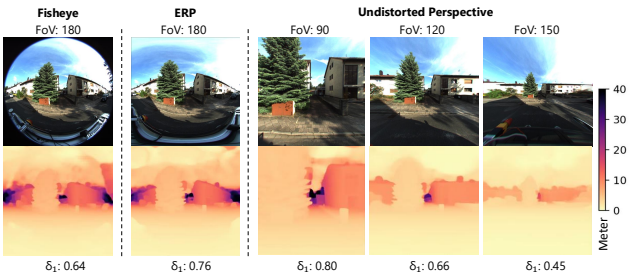

Figure 2. **Challenges on zero-shot test on large FoV camera images.** Metric depth estimation models trained on perspective images (e.g., Metric3Dv2 [20]) experience significant performance degradation when applied to fisheye images. Degradation is less pronounced when using an undistorted portion with a highly limited FoV or its ERP conversion.

2. Related Works

2.1. Zero-Shot Monocular Depth Estimation

Recent approaches to zero-shot metric depth estimation tackle the challenge of inconsistent scaling factors in depth ground truth due to varying camera intrinsic parameters [4, 16, 20, 33, 53, 54, 57]. Zoedepth [4] introduces an advanced network architecture, while DepthAnything [53, 54] employs a sophisticated unsupervised learning paradigm. However, their performance in metric depth estimation is limited without tackling inconsistency camera intrinsics. Metric3D [20, 57] and UniDepth [33] address scaling inconsistencies by converting images into a canonical camera space. Metric3D uses intrinsic parameters for this preprocessing, whereas UniDepth incorporates a network branch to estimate and convert intrinsics on the fly. Despite these advances, none of these methods achieve satisfactory zero-shot performance on large FoV images, presenting unique challenges in unifying diverse FoVs and supporting effective model learning.

2.2. Depth Estimation from Large FoV Cameras

Depth estimation for fisheye, 360° cameras has grown in popularity, as large FoVs capture richer contextual information that enhances depth estimation [1, 21, 26, 41, 60]. A key challenge for these cameras is managing position-dependent distortions, which vary by camera models. Approaches to address this include deformable CNNs [42, 50,

63], which adapt kernel shapes to compensate for distortions, as well as methods that segment ERP images to reduce distortion effects before merging [21, 38]. More recent methods leverage transformers to handle these distortions [10, 26, 41, 60]. While transformer-based networks have improved in-domain performance, they are approaching saturation, indicating that distortion is not the only challenge. Instead, the lack of large-scale FoV-specific training data is a key bottleneck for generalization. No current methods enable an unified depth estimation model trained on mixed large-scale perspective datasets to achieve zero-shot generalization on ERP or fisheye images.

3. Notations and Preliminaries

Depth Scaling Operations. Monocular depth estimation is inherently ill-posed, as different 3D object sizes and depths can produce the same 2D appearance. Deep learning models rely on learning an object's 3D dimensions from its 2D appearance [16, 17, 43] to infer depth, leading to the scaling operation illustrated in the right panel of Fig. 3. When using mixed camera data, apparent object size also depends on focal length, making accurate 2D-to-depth mapping dependent on appropriately scaling ground-truth depths when converting a perspective model to a canonical model, as shown in the left panel of Fig. 3. These scaling operations are central to the Metric3D [20, 57] pipeline and are integrated into our ERP-based approach.

EquiRectangular Projection (ERP). Equi-Rectangular Projection (ERP) is an image representation based on a spherical camera model, where each pixel corresponds to a specific latitude λ and longitude ϕ. A full ERP space spans $180°$ in latitude and $360°$ in longitude, making it ideal for handling diverse FoV cameras. The ERP image height is the only parameter needed to define the ERP space, allowing both training and testing images to be consistently converted into this space, regardless of the original FoV.

Transformations between standard images and ERP images use Gnomonic Projection transformation [47], which offers a closed-form mapping between tangent image coordinates (x_t, y_t) and spherical coordinates (λ, ϕ), assuming a tangent plane centered at (λ_c, ϕ_c) of a unit sphere. Specifically, as shown in Fig. 5, this mapping is given by:

$$x_t = \frac{\bar{x}}{\cos c} = \frac{\cos\phi \sin(\lambda - \lambda_c)}{\cos c} \tag{1}$$

$$y_t = \frac{\bar{y}}{\cos c} = \frac{\cos\phi_c \sin\phi - \sin\phi_c \cos\phi \cos(\lambda - \lambda_c)}{\cos c} \tag{2}$$

where $(\bar{x}, \bar{y}, \cos c)$ represents a point on the unit sphere, and c is the angular distance between (λ, ϕ) and (λ_c, ϕ_c), can be calculated as:

$$\cos c = \sin\phi_c \sin\phi + \cos\phi_c \cos\phi \cos(\lambda - \lambda_c) \tag{3}$$

We use these transformations to enable an efficient ERP conversion and data augmentation process, creating a streamlined pipeline that supports zero-shot generalization for depth estimation across various FoV cameras.

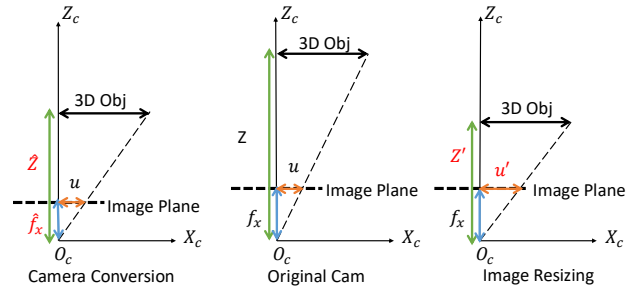

Figure 3. **Depth Scaling in Canonical Model Conversion and Image Resizing.** The apparent 2D size of an object u in an image depends on its 3D dimensions X, depth Z, and camera focal length f_x. *Left*: Converting a perspective camera model to a canonical model with a different focal length $\hat{f}_x$ requires scaling the depth values proportionally, so $\hat{Z} = \frac{\hat{f}_x Z}{f_x}$. *Center*: The original camera setup, showing the direct relationship between object size, depth, and focal length. *Right*: When the camera model is fixed but the image is resized to u', this simulates viewing the same 3D object at a different distance, necessitating depth scaling for accurate metric depth, with $Z' = \frac{uZ}{u'}$.

4. Depth Any Camera

We propose **Depth Any Camera (DAC)**, a depth model training framework designed to achieve zero-shot generalization across diverse camera models, including perspective, fisheye, and 360° cameras. As illustrated in Fig. 4, images from different camera types and FoVs are transformed into a canonical ERP space during both the training and testing phases. For training, we leverage the extensive perspective image datasets by converting them into smaller ERP patches for efficient learning. During testing, large FoV images are similarly converted into the canonical ERP space, allowing the trained model to predict metric depth consistently, without getting confused by different camera intrinsic and distortion parameters.

Several key components are designed to address specific challenges in implementing the DAC framework. In Sec. 4.1, we present an efficient pitch-aware Image-to-ERP conversion method that simulate large-FoV images at patch level and supports online augmentation within the ERP space. Sec. 4.2 introduces a FoV alignment process, an effective data augmentation technique that maximizes content inclusion while minimizing computational waste on background areas, using a single predefined ERP patch size. In Sec. 4.3, we describe a multi-resolution data augmentation approach aimed at training a transformer-based network capable of handling a broad range of testing resolutions.

The proposed DAC framework is compatible with various depth estimation network architectures, which are not the primary focus of this paper. Without loss of generality,

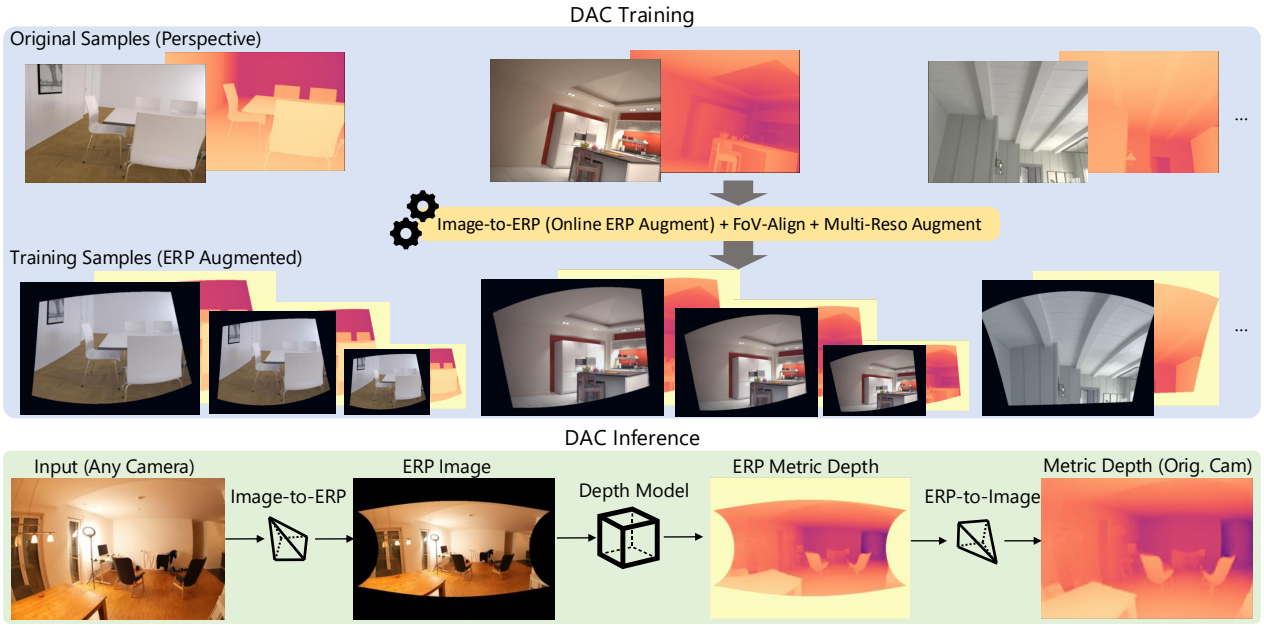

Figure 4. **Depth Any Camera Pipeline.** Our DAC framework converts data from any camera type into a canonical ERP space, enabling a model trained solely on perspective images to process large-FoV test data consistently for metric depth estimation. During training, we introduce an effective pitch-aware Image-to-ERP conversion with online data augmentation to simulate high-distortion regions unique to large-FoV images. The proposed FoV-Align process normalizes diverse-FoV data to a predefined ERP patch size, maximizing training efficiency. During inference, images from any camera type are converted into ERP space for depth estimation, with an optional step to map the ERP output back to the original image space for visualization.

we employ iDisc [32] for its simplicity and effectiveness, and for its incorporation of two prototypical attention modules, namely cross-attention and self-attention. We use the SIlog loss function [9] for training our models.

4.1. Pitch-Aware Image-to-ERP Conversion

An input image can be efficiently converted to its corresponding ERP patch through grid sampling combined with gnomonic projection. Assuming an ERP space with height H_E, width $W_E = 2H_E$, and the image center at latitude λ_c, longitude ϕ_c, with a target ERP patch size of $H_e \times W_e$, the ERP patch coordinates (u_e, v_e) can be mapped to spherical coordinates as $\phi = \frac{2\pi W_e}{W_E}(u_e - \frac{W_e}{2}) + \phi_c$, and $\lambda = \frac{\pi H_e}{H_E}(v_e - \frac{H_e}{2}) + \lambda_c$. Using Gnomonic Geometry presented in Eq. 1 and Eq. 2, we obtain the corresponding normalized image coordinate $(x_t, y_t, 1)$ in the tangent plane and $(\bar{x}, \bar{y}, \cos c)$ on the unit sphere. To map this coordinate to the actual image coordinate (u, v), we apply distortion and projection functions based on the given camera parameters:

$$(x_d, y_d) = f_d(\bar{x}, \bar{y}, \cos c, D_c) \tag{4}$$

$$(u, v) = f_p(x_d, y_d, K_c) \tag{5}$$

where f_d is the distortion function with distortion parameters D_c, and f_p is the projection function with intrinsic parameters K_c. If the input image has no distortion, projection function is directly applied to (x_t, y_t). Details on applying distortion models are included in Supplemental Sec. 8.

As shown in Fig. 5, with uniformly sampled grid points within the target ERP patch, each grid point can be mapped directly to a corresponding location in the input image. This mapping facilitates efficient transformation of the captured image into an ERP patch via grid sampling. Essentially, each grid point in the ERP patch maps to a specific floating-point position in the input image, and its value is obtained by interpolating from the neighboring pixel values.

This ERP conversion enables a powerful training pipeline when the latitude of the tangent plane center λ_c is defined by the camera's pitch angle in training. When camera orientation is available or can be estimated [22], perspective data can be projected to various latitudes in the ERP space, enabling the simulation of regions uniquely visible from large-FoV cameras, as shown in Fig. 5, and Supplemental Fig. 7. This **pitch-aware conversion** is crucial for improving the generalization of trained models to previously unobserved large-FoV data, as demonstrated in Sec. 5.3, because neural networks alone have limited capacity to generalize to extrapolated data spaces [51].

Another notable advantage is the seamless ability to perform **online augmentation** efficiently in the ERP space. For datasets with limited pitch variation [12, 15, 19, 61], a unique pitch augmentation can be efficiently applied by adding noise to λ_c, generating ERP patches with varying shapes, as shown in Fig. 4. In addition, common augmenta-

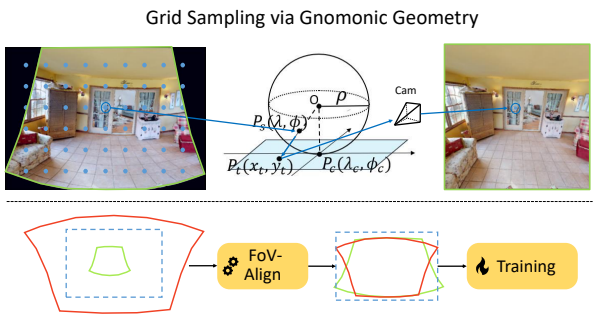

Figure 5. **Pitch-Aware ERP Conversion and FoV Alignment.** *Top*: Grid Sampling is applied for an efficient Image-to-ERP conversion. Each ERP grid sample's corresponding location in the input image is computed using gnomonic geometry and specific camera projection parameters. Given the patch center latitude λ determined by the camera's pitch angle, it makes the converted patch to represent high-distortion regions in the ERP space. *Bottom*: The FoV-Align process normalizes diverse-FoV ERP patches (shown in red and green) to match the height of a single predefined ERP patch (outlined in blue), ensuring efficient training.

tions, such as scaling, rotation, and translation—commonly applied to perspective images—can be directly applied to the normalized image coordinates (x_t, y_t) as follows:

$$\begin{bmatrix} x'_t \\ y'_t \end{bmatrix} = s_\sigma \begin{bmatrix} R_\sigma & T_\sigma \end{bmatrix} \begin{bmatrix} x_t \\ y_t \end{bmatrix} \tag{6}$$

where s_σ is a scale factor, R_σ is 2D rotation matrix, and T_σ is a 2D translation vector corresponding to the applied augmentations.

4.2. FoV Alignment

When training data include a wide range of camera FoVs within perspective images, such as in the HM3D [35] dataset produced by OmniData [8], the corresponding ERP regions can vary significantly in size, as shown in Fig. 5. There is no single crop size that can consistently capture most content information for certain images without wasting substantial computation on background padding for others. This creates a dilemma in prioritizing between training efficiency and richness of information, and it can also reduce training quality when samples exhibit drastically different content-to-background ratios.

To address this challenge, we introduce a simple yet effective FoV Alignment operation that adjusts the FoV of each input image to match the predetermined crop area FoV. Specifically, this process applies a specific scaling augmentation, as described in Sec. 4.1 and Eq. 6, specifically $\text{Fov}_e = \frac{H_e \pi}{H_E}$ and $s_\sigma = \frac{\text{Fov}_c}{\text{Fov}_e}$, where Fov_c is derived from actual camera parameters, and Fov_e is ERP patch's vertical FoV. As illustrated in Fig. 5, this approach allows a single predefined ERP patch size to maximize the inclusion of relevant content and minimize computational waste on background, making it ideal for an efficient training pipeline.

4.3. Multi-Resolution Training

Training ERP patches and testing images may prefer inconsistent resolutions for various reasons, e.g. drastically different aspect ratio, edge device limitation. When testing resolutions differ from the training patch size, model performance can degrade significantly, particularly with attention modules that aggregate different numbers of image tokens.

To address this issue, we adopt a multi-resolution training scheme. As illustrated in Fig. 4, each ERP patch is resized to two additional lower resolutions (typically 0.7 and 0.4 of the original) to incorporate varied image resolutions in training. The training feeds the model three batches of images at different resolutions and sums the losses.

5. Experiments

5.1. Experimental Setup

In-Domain Training Datasets. For indoor experiments, we use Habitat-Matterport 3D (HM3D) [35], Taskonomy [61], and Hypersim [39], totaling 670K perspective images with distinct characteristics. To streamline training, we use the first 50 scenes from HM3D and Taskonomy (tiny versions provided by OmniData [8]). For outdoor data, we use DDAD [15] and LYFT [19], totaling 130K images. Table 1 summarizes the datasets in use, showing varying distributions in FoV, pitch angles, sources, and image quality.

Zero-Shot Testing Datasets. We evaluate DAC on two 360° datasets—Matterport3D [5] and Pano3D-GV2 [2]—and two fisheye datasets—ScanNet++ [56] and KITTI360 [27]—all featuring larger FoVs than perspective images. Our primary experiments focus on these datasets to assess zero-shot generalization to large FoV cameras. Additional evaluations on NYUv2 [30] and KITTI [12] are provided in the supplementary material, demonstrating DAC's performance on perspective data relative to SoTA methods.

Evaluation Details. We assess DAC's generalization to large FoV cameras across both indoor and outdoor scenes, training separate models for each without data mixing to simplify training. For fair comparison, competing models are either re-trained with the same data splits or use their largest versions trained on extensive datasets. Evaluations are conducted using metric depth metrics: $\delta_1 \uparrow$, $\delta_2 \uparrow$, $\delta_3 \uparrow$, Abs Rel$\downarrow$, RMSE$\downarrow$, and log10$\downarrow$.

Baselines. We compare DAC with the following baselines:

- Metric3Dv2 [20]: A SoTA foundation model in zero-shot metric depth estimation, built upon perspective canonical camera model to standardize datasets.
- UniDepth [33]: A more recent SoTA foundation depth model leveraging network designs to handle diverse camera parameters. We test its ability to handle large FoV cameras not included in training.
- iDisc [32]: Selected as a network baseline due to its use of self-attention and cross-attention modules in a straight-

Table 1. **Overview of Datasets.** This table summarizes the training and testing datasets used in this work. The training datasets span a range of FoVs, pitch angles, and image quality, each potentially impacting performance on different test datasets in varying degrees.

Train Dataset	# Imgs	Scene	xFoV (deg.)	# Cams	Pitch (deg.)	Source	Img Qual.	Test Dataset	Cam Type	xFoV (deg.)	Scene
HM3D-tiny [35]	310K	Indoor	$36^\circ - 124^\circ$	10K+	$3\sigma = 75^\circ$	Recon	Low	Matterport3D [5]	ERP	360°	Indoor
Taskonomy-tiny [61]	300K	Indoor	$45^\circ - 75^\circ$	10K+	$3\sigma = 24^\circ$	Real	High	Pano3D-GV2 [2]	ERP	360°	Indoor
Hypersim [39]	54K	Indoor	60°	1	$3\sigma = 60^\circ$	Sim	High+	ScanNet++ [56]	Fisheye	150°	Indoor
DDAD [15]	80K	Outdoor	$45^\circ - 60^\circ$	36+	$3\sigma \sim 10^\circ$	Real	High	KITTI360 [27]	Fisheye	180°	Outdoor
LYFT [19]	50K	Outdoor	$20^\circ - 60^\circ$	6+	$3\sigma \sim 10^\circ$	Real	High				

Table 2. **Zero-Shot Test on** 360° **and Fisheye Datasets.** DAC is compared with SoTA metric depth models across four large-FoV datasets.

Test Dataset	Methods	Train Dataset	Backbone	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$	Abs Rel↓	RMSE↓	log10↓
Matterport3D [5]	UniDepth [33]	Mix 3M	ViT-L [7]	0.2576	0.5114	0.7091	0.7648	1.3827	0.2208
	Metric3Dv2 [20]	Mix 16M	Dinov2 [31]	0.4381	0.7311	0.8735	0.2924	0.8842	0.1546
	Metric3Dv2 [20]	Indoor 670K	Dinov2 [31]	0.4287	0.7854	0.9333	0.2788	0.8961	0.1352
	iDisc [32]	Indoor 670K	Resnet101 [18]	0.5287	0.8260	0.9398	0.2757	0.7771	0.1147
	DAC (Ours)	Indoor 670K	Resnet101 [18]	**0.7727**	**0.9562**	**0.9822**	**0.156**	**0.6185**	**0.0707**
Pano3D-GV2 [2]	UniDepth [33]	Mix 3M	ViT-L [7]	0.2469	0.4977	0.7084	0.7892	1.2681	0.2231
	Metric3Dv2 [20]	Mix 16M	Dinov2 [31]	0.4040	0.6929	0.8499	0.3070	0.8549	0.1664
	Metric3Dv2 [20]	Indoor 670K	Dinov2 [31]	0.5060	0.8176	0.9360	0.2608	0.7248	0.1201
	iDisc [32]	Indoor 670K	Resnet101 [18]	0.5629	0.8222	0.9332	0.2657	0.6446	0.1122
	DAC (Ours)	Indoor 670K	Resnet101 [18]	**0.8115**	**0.9549**	**0.9860**	**0.1387**	**0.4780**	**0.0623**
ScanNet++ [56]	UniDepth [33]	Mix 3M	ViT-L [7]	0.3638	0.6461	0.8358	0.4971	1.1659	0.1648
	Metric3Dv2 [20]	Mix 16M	Dinov2 [31]	0.5360	0.8218	0.9350	0.2229	0.8950	0.1177
	Metric3Dv2 [20]	Indoor 670K	Dinov2 [31]	0.6489	0.8920	0.9558	0.1915	0.9779	0.0938
	iDisc [32]	Indoor 670K	Resnet101 [18]	0.6150	0.8780	0.9617	0.2712	0.4835	0.0972
	DAC (Ours)	Indoor 670K	Resnet101 [18]	**0.8517**	**0.9693**	**0.9922**	**0.1323**	**0.3086**	**0.0532**
KITTI360 [27]	UniDepth [33]	Mix 3M	ViT-L [7]	0.4810	0.8397	0.9406	0.2939	6.5642	0.1221
	Metric3Dv2 [20]	Mix 16M	Dinov2 [31]	0.7159	0.9323	0.9771	0.1997	4.5769	0.0811
	Metric3Dv2 [20]	Outdoor 130K	Dinov2 [31]	0.7675	0.9370	0.9756	0.1521	4.6610	0.0723
	iDisc [32]	Outdoor 130K	Resnet101 [18]	0.7833	0.9384	0.9753	0.1598	4.9122	0.0704
	DAC (Ours)	Outdoor 130K	Resnet101 [18]	**0.7858**	**0.9388**	**0.9775**	**0.1559**	**4.3614**	**0.0684**

forward yet effective network, and strong in-domain performance. As iDisc alone does not handle mixed camera parameters, we train it with Metric3Dv2 [20], and compare it to the same network trained with ours.

Implementation Details. In the DAC training pipeline, we set the full ERP height to $H_{\text{erp}} = 1400$ pixels, with an ERP patch size of 500×700 pixels for both indoor and outdoor models. We use 10° latitude augmentations for both and additionally 10° rotation augmentation for indoor. When training the iDisc [32] model using the Metric3D [57] pipeline, we use canonical focal lengths of $f_{\text{cano}} = 519$ (NYU dataset [30]) for indoor models and $f_{\text{cano}} = 721$ (KITTI dataset [12]) for outdoor models.

To test perspective models on ERP and fisheye images, specific adjustments are required. For 360° (ERP) images, which lack a defined focal length, we calculate a virtual focal length f_{virtual} based on pixels per latitude degree: $\frac{1}{f_{\text{virtual}}} = \tan\left(\frac{\pi}{H_{\text{erp}}}\right)$, scaling the predicted depth with $\frac{f_{\text{cano}}}{f_{\text{virtual}}}$ for ground-truth alignment. For fisheye images, aligning f_{cano} with the post-distortion focal length introduces significant errors, so we first convert fisheye images to ERP space and apply $\frac{f_{\text{cano}}}{f_{\text{virtual}}}$ for metric depth evaluation.

For testing resolution, if the original resolution is less than twice the training resolution, we use it directly; for larger resolutions, we maintain the aspect ratio and align it with the training resolution. Based on this rule, we evaluate Matterport3D [5] and Pano3D-GV2 [2] at 512×1024, ScanNet++ [56] at 500×750, and KITTI360 [27] at 700×700. For competing methods that are not adaptable to inconsistent resolutions compared to training, we report the higher score obtained from the two settings to ensure fairness.

In experiments, ResNet101 [18]-backbone models are trained for 60k iterations with a batch size of 48, while Swin-L [28] and DINOv2 [31] models are trained for 120k iterations with a batch size of 48. Finally, to support all FoV types, depth is represented as *Euclidean Distance* from the camera center rather than *Z-buffer* format, as the latter is incompatible with spherical projections and would yield inaccurate low depth values for fisheye or ERP images.

5.2. Comparison with the SoTA

In this section, we compare DAC with primary baselines in zero-shot generalization tests on large FoV datasets, with quantitative results reported in Table 2, and qualitative results shown in Fig. 6. In indoor experiments, DAC significantly outperforms pre-trained models UniDepth [33] and Metric3Dv2 [20], even when using a lighter ResNet101 [18] backbone and a much smaller training dataset. DAC

achieves superior performance across both 360° datasets and the fisheye dataset ScanNet++ [56]. Compared to the iDisc [32] network trained with the Metric3Dv2 pipeline, DAC shows substantial improvements across all metrics on all datasets. Notably, DAC improves the next-best method by nearly 50% in the most differentiating metric, δ_1.

In outdoor tests, DAC significantly outperforms Metric3Dv2 [57] and UniDepth [33], even with much larger backbones. However, it achieves only marginal improvements over iDisc [32] under the same network configuration, with less pronounced gains compared to indoor settings. This is likely due to the limited camera pitch variance in the outdoor training data (Table 1), reducing the ability to simulate highly distorted regions. Moreover, KITTI360 [27] LiDAR points are concentrated in less distorted areas (Fig. 6), making the evaluation less distinctive.

Particularly, a notable observation is that while UniDepth [33] utilizes a network-based spherical conversion, it struggles with large FoV cameras, exposing the limitations of deep learning in extrapolated domains [51]. In contrast, DAC's success underscores the effectiveness of our geometry-based training pipeline.

Additional results of DAC models using SwinL [28] backbones are provided in Supplemental Table 5. These models outperform their ResNet101 [18] counterparts in most cases, except on the 360° datasets.

Table 3. **Impact of Key Components and Network.** We conduct the main ablation study on indoor datasets by training with HM3D [35] and performing zero-shot testing on Pano3D-GV2 [2] and ScanNet++ [56]. We compare the performance of the DAC framework with specific components removed, as well as different network architectures trained under the Metric3D [57] pipeline.

Test Dataset	Methods	$\delta_1 \uparrow$	$\delta_2 \uparrow$	A.Rel$\downarrow$
Pano3D-GV2 [2]	Metric3Dv2 [20]	0.5623	0.8341	0.2479
	iDisc-cnn [32]	0.3026	0.5565	0.3548
	iDisc [32]	0.4130	0.6844	0.3043
	DAC (Ours)	**0.7251**	**0.9254**	**0.1729**
	w\o Pitch-Aware ERP	0.4911	0.7904	0.2422
	w\o Pitch Aug 10°	0.6912	0.9311	0.188
	w\o FoV Align	0.4075	0.7585	0.2610
	w\o Multi-Reso	0.5128	0.7784	0.2437
ScanNet++ [56]	Metric3Dv2 [20]	0.4569	0.7463	0.2818
	iDisc-cnn [32]	0.4639	0.7653	0.3045
	iDisc [32]	0.5301	0.8048	0.3237
	DAC (Ours)	**0.6539**	**0.9083**	**0.1951**
	w\o Pitch-Aware ERP	0.4711	0.8068	0.2508
	w\o Pitch Aug 10°	0.6741	0.9066	0.1914
	w\o FoV Align	0.5428	0.8644	0.2200
	w\o Multi-Reso	0.5504	0.8464	0.2231

5.3. Ablation Study

Key Components and Network Architecture. We evaluate the effect of the FoV Align and Multi-Reso Training components by removing each individually, while keeping the rest of the DAC framework unchanged. This ablation is conducted on the challenging HM3D-tiny [35] indoor dataset, which includes varied camera FoVs, pitch angles, and lower-quality images from reconstructed scenes. We also test the impact of removing attention modules from iDisc [32] and compare to Metric3Dv2 [20] to isolate the influence of the iDisc architecture. Both iDisc-based methods and DAC use ResNet101 backbones, while Metric3Dv2 uses Dinov2. Table 3 provides a summary; full metrics and Matterport3D [5] results are in the Supplemental Table 6.

Table 4. **Impact of Train Dataset.** Models are trained separately on each training dataset and evaluated in zero-shot tests on 360° and fisheye datasets. Due to the unique characteristics of each training dataset, their contributions and importance to generalization across different testing datasets vary.

Test Datasets	Train Dataset	Methods	$\delta_1 \uparrow$	A.Rel$\downarrow$
Pano3D-GV2 [2]	HM3D-tiny [35]	Metric3Dv2 [20]	0.5623	0.2479
		iDisc [32]	0.4130	0.3043
		DAC (Ours)	**0.7251**	**0.1729**
	Taskonomy-tiny [61]	Metric3Dv2 [20]	0.3785	0.2959
		iDisc [32]	0.3888	0.4076
		DAC (Ours)	**0.6411**	**0.1972**
	Hypersim [39]	Metric3Dv2 [20]	0.3085	0.5583
		iDisc [32]	0.3372	0.3288
		DAC (Ours)	**0.5208**	**0.1792**
ScanNet++ [56]	HM3D-tiny [35]	Metric3Dv2 [20]	0.4569	0.2818
		iDisc [32]	0.5301	0.3237
		DAC (Ours)	**0.6539**	**0.1951**
	Taskonomy-tiny [61]	Metric3Dv2 [20]	0.6318	0.2148
		iDisc [32]	0.6743	0.1977
		DAC (Ours)	**0.7981**	**0.1447**
	Hypersim [39]	Metric3Dv2 [20]	0.5050	0.2269
		iDisc [32]	0.6656	0.2213
		DAC (Ours)	**0.7478**	**0.1762**

Table 3 highlights the pivotal role of pitch-aware ERP conversion in generalizing perspective-trained models to large FoV datasets by effectively simulating high-distortion regions uniquely observed in large FoV images (Fig. 7). This approach turns the wide pitch angle variance in datasets like HM3D [35] into an advantage. While additional pitch augmentation does not appear essential when the training dataset like HM3D intricately spans a large range of pitch angle. However, its effectiveness varies across datasets, as detailed in Supplemental Table 6.

Results in Table 3 also show that removing FoV Align or Multi-Reso Training significantly reduces DAC performance, particularly for zero-shot generalization on 360° images. Compared to the iDisc network trained with the Metric3D pipeline, DAC achieves notable improvements for large FoV cameras, with attention modules in iDisc proving effective for large FoV test data. Although Metric3Dv2 uses a heavier backbone, it shows limited zero-shot generalization on large FoV images without DAC. More comprehensive results can be found both in Supplemental Table 6.

Impact of Training Dataset. Each dataset has unique characteristics (Table 1). To evaluate their impact on generalization to large FoV data, we trained models separately on HM3D-tiny, Taskonomy-tiny [61], and Hypersim [39], then

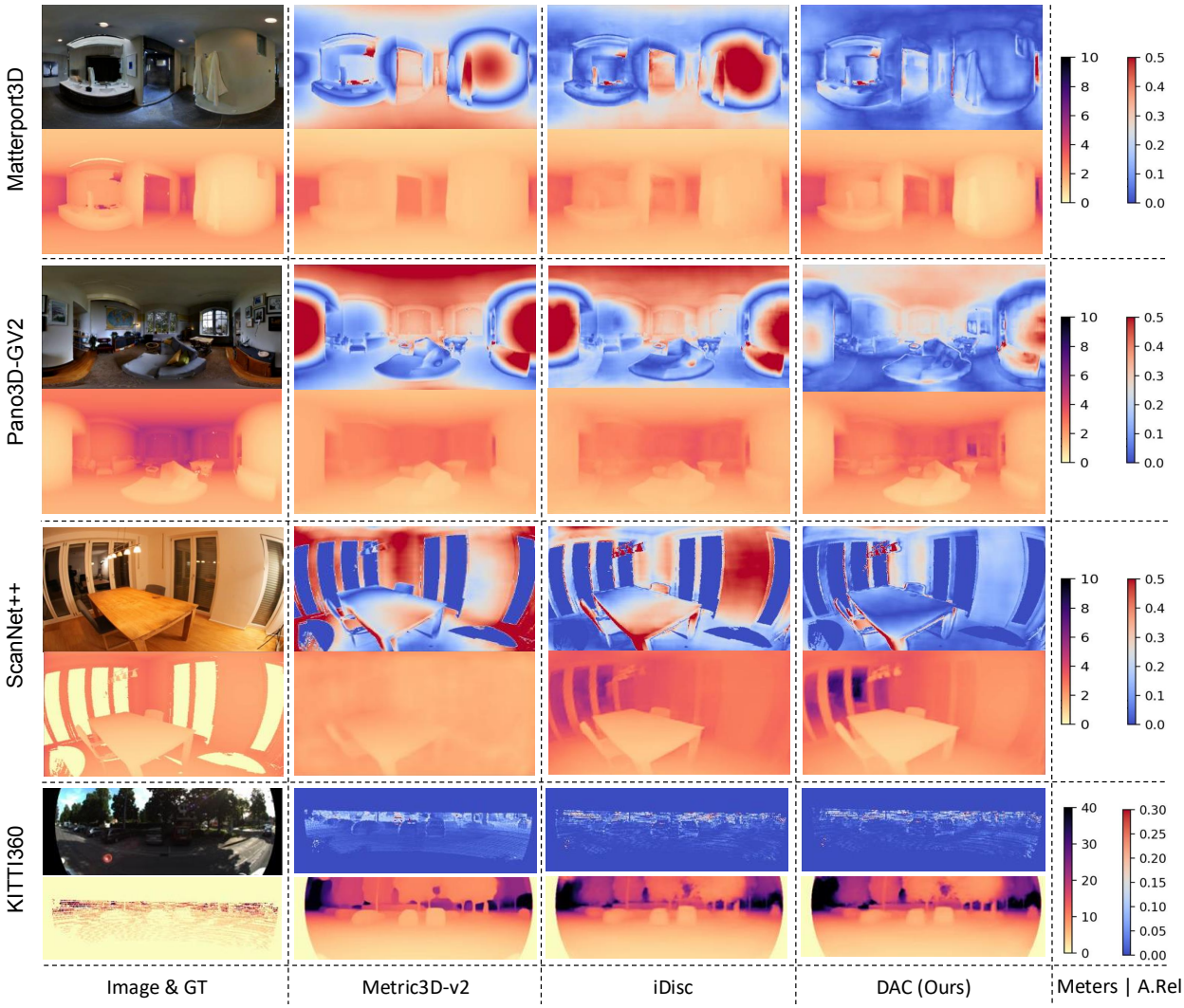

Figure 6. **Zero-Shot Qualitative Results.** For each dataset, an example is presented in two consecutive rows. The left column shows the original image and Ground-Truth depth map, followed by results from various methods. For each method, the top row displays the A.Rel map ↓ and the bottom row shows the predicted depth map. The color range for depth and A.Rel maps is indicated in the last column.

tested them in zero-shot mode on indoor large FoV datasets. Results are summarized in Table 4, with full results in Supplemental Table 7.

For Pano3D-GV2 [2], broader FoV and pitch angle coverage in HM3D training improve generalization across all methods, despite HM3D's lower quality due to rendering artifacts. For fisheye data in ScanNet++ [56], FoV diversity appears less crucial, as the single-camera Hypersim dataset, despite limited training data, outperforms HM3D, indicating that image quality plays a key role in ScanNet++ test.

Comparing individually trained models with results from mixed training in Table 2 shows that DAC effectively leverages the synergy of diverse datasets, significantly enhancing generalization to large FoV datasets.

6. Conclusion

We introduced the Depth Any Camera (DAC) framework for zero-shot metric depth estimation across diverse camera types, including perspective, fisheye, and 360° cameras. By leveraging a highly effective pitch-aware Image-to-ERP transformation, FoV alignment, and multi-resolution training, DAC addresses the challenges posed by varying FoVs and resolution inconsistencies and enables robust generalization on large FoV datasets. Our results demonstrate that DAC significantly outperforms state-of-the-art methods and adapts seamlessly to different backbone networks. In practice, DAC ensures that every piece of previously collected 3D data remains valuable, regardless of the camera type used in new applications.

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. 2

[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 *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, 2021. 5, 6, 7, 8, 2, 3, 4

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021. 1

[4] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *CoRR*, 2023. 1, 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. *International Conference on 3D Vision (3DV)*, 2017. 5, 6, 7, 2, 3, 4

[6] Cheng Chi, Zhenjia Xu, Chuer Pan, Eric Cousineau, Benjamin Burchfiel, Siyuan Feng, Russ Tedrake, and Shuran Song. Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots. *CoRR*, 2024. 2

[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 *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 6, 1, 2

[8] 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021. 5

[9] 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 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2014. 4

[10] 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

[11] Louis Gallagher, Ganesh Sistu, Jonathan Horgan, and John B. McDonald. A system for dense monocular mapping with a fisheye camera. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023 - Workshops, Vancouver, BC, Canada, June 17-24, 2023*, 2023. 2

[12] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The KITTI dataset. *Int. J. Robotics Res.*, 2013. 4, 5, 6, 1, 2

[13] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019. 1

[14] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2

[15] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, 2020. 4, 5, 6

[16] Vitor Guizilini, Igor Vasiljevic, Dian Chen, Rares Ambrus, and Adrien Gaidon. Towards zero-shot scale-aware monocular depth estimation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 2, 3

[17] Yuliang Guo, Abhinav Kumar, Cheng Zhao, Ruoyu Wang, Xinyu Huang, and Liu Ren. SUP-NeRF: A Streamlined Unification of Pose Estimation and NeRF for Monocular 3D Object Reconstruction. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024*, 2024. 3

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016. 6, 7, 1, 2

[19] 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 *4th Conference on Robot Learning, CoRL 2020, 16-18 November 2020, Virtual Event / Cambridge, MA, USA*, 2020. 4, 5, 6

[20] 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. *CoRR*, 2024. 1, 2, 3, 5, 6, 7, 4

[21] Hualie Jiang, Zhe Sheng, Siyu Zhu, Zilong Dong, and Rui Huang. Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics Autom. Lett.*, 2021. 2, 3

[22] Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023. 4, 1

[23] 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*, 2006. 2

[24] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel(r) realsense(tm) stereoscopic depth cameras. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21-26, 2017*, 2017. 1

[25] 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. *CoRR*, 2019. 1

[26] 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. 2, 3

[27] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2023. 5, 6, 7, 2

[28] 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 *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021. 6, 7, 1, 2

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

[30] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012. 5, 6, 1, 2

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. 6, 1, 2

[32] Luigi Piccinelli, Christos Sakaridis, and Fisher Yu. idisc: Internal discretization for monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023. 4, 5, 6, 7, 1, 2, 3

[33] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segù, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 2024. 1, 2, 5, 6, 7

[34] Elad Plaut, Erez Ben-Yaacov, and Bat El Shlomo. 3d object detection from a single fisheye image without a single fisheye training image. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2021, virtual, June 19-25, 2021*, 2021. 2

[35] 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 *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 5, 6, 7, 1, 3, 4

[36] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021. 1

[37] 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 Trans. Pattern Anal. Mach. Intell.*, 2022. 1

[38] 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

[39] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Ángel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021. 5, 6, 7, 1, 4

[40] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[41] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360$^{\circ }$ depth estimation. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part I*, 2022. 2, 3

[42] 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. 2

[43] Tom van Dijk and Guido de Croon. How do neural networks see depth in single images? In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, 2019. 3

[44] Yahui Wang, Shaojun Cai, Shi-Jie Li, Yun Liu, Yangyan Guo, Tao Li, and Ming-Ming Cheng. Cubemapslam: A piecewise-pinhole monocular fisheye SLAM system. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part VI*, 2018. 2

[45] Yahui Wang, Shaojun Cai, Shi-Jie Li, Yun Liu, Yangyan Guo, Tao Li, and Ming-Ming Cheng. Cubemapslam: A piecewise-pinhole monocular fisheye SLAM system. In *Computer Vision - ACCV 2018 - 14th Asian Conference on Computer Vision, Perth, Australia, December 2-6, 2018, Revised Selected Papers, Part VI*, 2018. 2

[46] Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel J. Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 2021. 1

[47] Eric W Weisstein. Gnomonic projection. 2, 3, 5

[48] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022. 1

[49] Ziniu Wu, Tianyu Wang, Zhaxizhuoma, Chuyue Guan, Zhongjie Jia, Shuai Liang, Haoming Song, Delin Qu, Dong Wang, Zhigang Wang, Nieqing Cao, Yan Ding, Bin Zhao, and Xuelong Li. Fast-umi: A scalable and hardware-independent universal manipulation interface. *CoRR*, 2024. 2

[50] 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. 2

[51] Keyulu Xu, Mozhi Zhang, Jingling Li, Simon Shaolei Du, Ken-ichi Kawarabayashi, and Stefanie Jegelka. How neural networks extrapolate: From feedforward to graph neural networks. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021. 4, 7

[52] Lu Yang, Liulei Li, Xueshi Xin, Yifan Sun, Qing Song, and Wenguan Wang. Large-scale person detection and localization using overhead fisheye cameras. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 2

[53] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 2024. 1, 2

[54] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *CoRR*, 2024. 2

[55] Yaozu Ye, Kailun Yang, Kaite Xiang, Juan Wang, and Kaiwei Wang. Universal semantic segmentation for fisheye urban driving images. In *2020 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2020, Toronto, ON, Canada, October 11-14, 2020*, 2020. 2

[56] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 5, 6, 7, 8, 2, 3, 4

[57] 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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, 2023. 1, 2, 3, 6, 7

[58] Senthil Kumar Yogamani, David Unger, Venkatraman Narayanan, and Varun Ravi Kumar. Fisheyebevseg: Surround view fisheye cameras based bird's-eye view segmentation for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*, 2024. 2

[59] Hong-Xing Yu, Haoyi Duan, Charles Herrmann, William T. Freeman, and Jiajun Wu. Wonderworld: Interactive 3d scene generation from a single image. *CoRR*, 2024. 2

[60] 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. 2, 3

[61] Amir Zamir, Alexander Sax, William B. Shen, Leonidas J. Guibas, Jitendra Malik, and Silvio Savarese. Taskonomy: Disentangling task transfer learning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019. 4, 5, 6, 7, 1

[62] Wei Zhang, Sen Wang, Xingliang Dong, Rongwei Guo, and Norbert Haala. BAMF-SLAM: bundle adjusted multi-fisheye visual-inertial SLAM using recurrent field transforms. In *IEEE International Conference on Robotics and Automation, ICRA 2023, London, UK, May 29 - June 2, 2023*, 2023. 2

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