OMNIRAY: MONOCULAR 3D RECONSTRUCTION WITH COMPOSITE RAY FIELDS

Anonymous authors
Paper under double-blind review

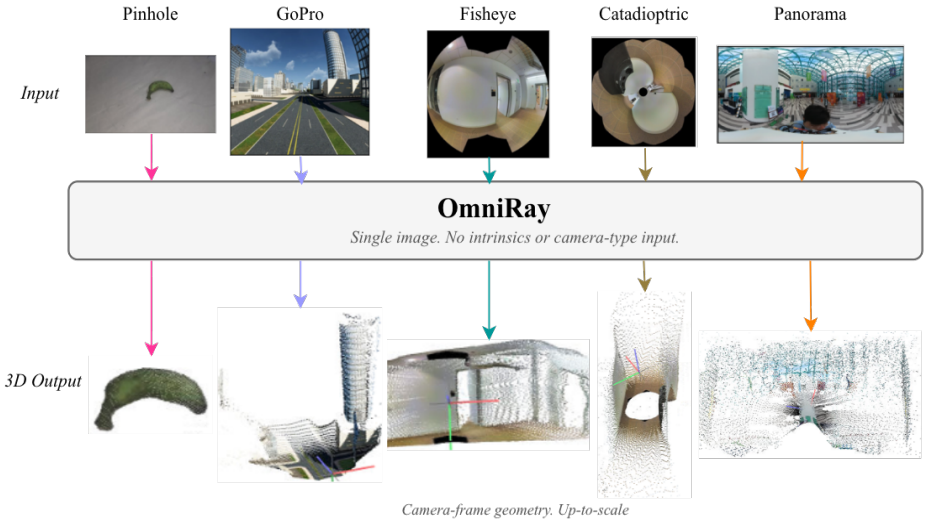


Figure 1: **OmniRay reconstructs 3D geometry from a single image across diverse camera geometries** without requiring camera intrinsics or camera-type input. Each column shows an input image and its predicted camera-frame pointcloud, recovered up to scale.

ABSTRACT

Monocular models recover 3D geometry from a single image in one pass, but most assume a pinhole camera and produce distorted geometry on the wide-angle imagery produced by fisheye and panoramic lenses. Camera-aware models remove this assumption by predicting a per-pixel ray field. However, they represent different camera geometries using a *single fixed basis* to capture both the overall camera projection and its finer variations. We argue that camera geometry has a hierarchy of complexity and that the ray representation should follow it. We present **OmniRay**, a feed-forward model whose ray module is composed of three components: a monotone radial base that captures radially symmetric structure, a spherical-harmonic residual for smooth deviations from it, and a free-form residual for remaining local variation. These components are composed as successive rotations on the sphere, allowing the representation to accommodate diverse camera geometries. Given a single image, OmniRay predicts a ray direction and an up-to-scale radial distance for every pixel, from which a 3D pointcloud is recovered without needing camera information at inference time. Extensive experiments across different camera configurations demonstrate that OmniRay establishes a new state-of-the-art in camera-agnostic monocular 3D reconstruction.

1 INTRODUCTION

Monocular 3D reconstruction recovers a 3D point for each pixel in an image. We consider central cameras, in which all viewing rays pass through a single optical center, so each 3D point splits into a ray direction and a distance along that ray. The ray direction depends only on the camera geometry, while the distance depends on the scene under observation. Recent feed-forward models perform well on monocular 3D estimation (Yang et al., 2024; Wang et al., 2025), but their architectures or training data focus on pinhole cameras. When applied to wide-angle images, these assumptions often

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

lead to distorted reconstructed geometry. Wide-angle cameras are common in robotics (Yogamani et al., 2019), augmented and virtual reality (Engel et al., 2023), and 360° capture (Liao et al., 2022), making this a practical limitation. One way to handle these images is to calibrate the camera and rectify them to a pinhole view before reconstruction, but a pinhole cannot represent rays at or beyond 90° from its optical axis. As such, undistortion typically leads to cropping away part of the field of view or strongly stretching the image near its periphery, leading to low-confidence predictions from conventional depth estimation or 3D foundation models (Guo et al., 2026).

Camera-aware models instead account for camera geometry directly (Bogdan et al., 2018; Tirado-Garín & Civera, 2025; Piccinelli et al., 2025). DeepFocal (Workman et al., 2015) and DeepCalib (Bogdan et al., 2018) first use CNNs to estimate focal length and distortion of fixed camera models. Follow-up works generalize camera parameter estimation to arbitrary camera models by estimating ray fields directly. In particular, UniK3D (Piccinelli et al., 2025) predicts a per-pixel ray field using a spherical-harmonic basis and a radial distance along each ray. It can therefore reconstruct pinhole, fisheye, and panoramic images without camera metadata. In this formulation, the same basis must describe both the overall camera projection and its finer variations. We argue that there exists a hierarchy of complexity in camera geometry estimation, which can lead to more efficient and performant design. Many camera projections have a simple structure: for approximately radially symmetric lenses, a smooth radial function is sufficient in describing how viewing directions change across the image. Spherical-harmonics can be used to complement radial functions.

We therefore propose **OmniRay**. It takes a single image and predicts a dense, up-to-scale 3D reconstruction without intrinsics or camera-type input (Fig. 1). A shared image encoder feeds two modules. The ray module predicts a monotone radial base and the two residuals, each of which refines the preceding ray field. The distance module predicts the radial distance of each pixel, and the corresponding 3D point is obtained by scaling its predicted ray. We train OmniRay on cameras ranging from pinhole to panorama and evaluate it on held-out scenes, datasets, and camera types. An oracle analysis examines which ray fields each representation can express, while benchmark experiments and ablations test the performance of the proposed components.

In summary, our contributions are:

1. A composite ray representation built from a monotone radial base, a spherical-harmonic residual, and a free-form residual.
2. **OmniRay**, a feed-forward model for single-image 3D reconstruction across central camera geometries, without needing intrinsics or camera-type input.
3. An oracle analysis of ray representations, and experiments across seen and held-out diverse camera types, scenes, and datasets, show that OmniRay improves over state-of-the-art by a large margin in ray field estimation and monocular 3D reconstruction.

2 Related Work

Camera calibration. Camera calibration determines the mapping from image pixels to rays in the camera frame. Classical methods parameterize this mapping with a low-dimensional model whose parameters are estimated from geometric observations (Zhang, 2000; Scaramuzza et al., 2006). Different camera geometries are handled by different parametric models: pinhole model with radial and tangential distortion for conventional lenses, Kannala & Brandt (2006) model for wide-angle optics, and unified camera models (Geyer & Daniilidis, 2000; Mei & Rives, 2007; Khomutenko et al., 2015; Usenko et al., 2018) for omnidirectional cameras. Each of these models fixes a particular functional form, so a lens whose behavior departs from that form introduces residual error. Generic models relax these assumptions with more degrees of freedom and capture complex lens behavior (Schops et al., 2020) at the cost of a higher-dimensional calibration problem.

Learning-based methods offer an alternative to this high-dimensional calibration process by learning camera geometry directly from images without input calibration. DeepCalib (Bogdan et al., 2018) regresses camera parameters, while Perspective Fields (Jin et al., 2023), GeoCalib (Veicht et al., 2024), and AnyCalib (Tirado-Garín & Civera, 2025) move toward dense geometric cues: a per-pixel up-vector and latitude field (Jin et al., 2023; Veicht et al., 2024), or a per-pixel field of view (Tirado-Garín & Civera, 2025), each of which describes the camera more generally than a parameter vector. Neural Ray Surfaces (Vasiljevic et al., 2020) take it further by learning a free-form ray surface from video observations and directly representing the pixel-to-ray mapping. These calibration approaches

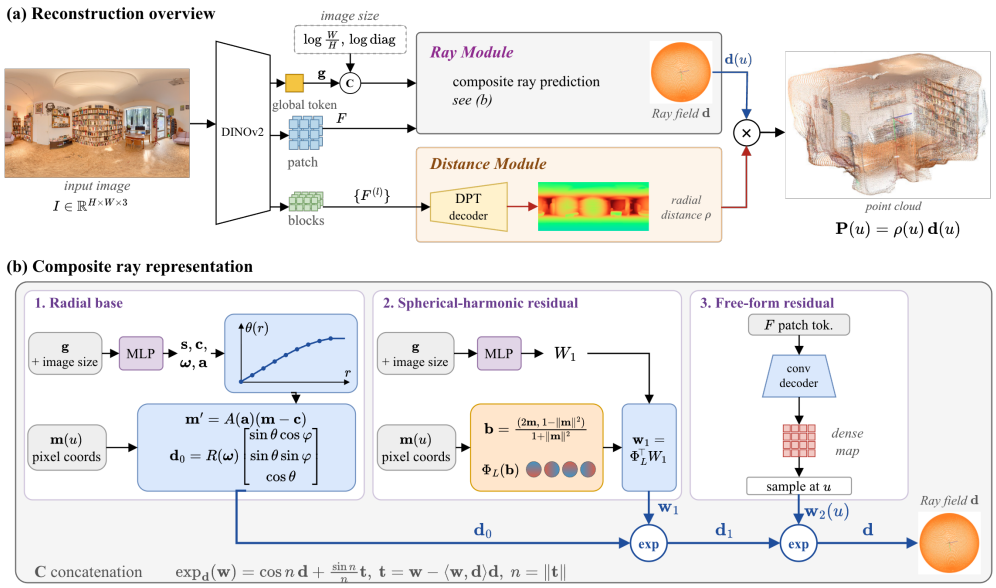


Figure 2: **OmniRay Overview.** Given a single image, a DINOv2 ViT-L/14 encoder produces global and dense features that drive two branches. The ***Ray Module*** progressively builds a camera-agnostic ray field from a monotone radial base, a spherical-harmonic residual, and a free-form residual. The ***Distance Module*** predicts a positive radial distance for each pixel. Combining the final ray $\mathbf{d}$ and distance ρ yields the local pointmap $\mathbf{P} = \rho\,\mathbf{d}$, recovering 3D geometry up to scale without camera intrinsics or a camera-type label.

shift from explicit parameter estimation toward learning the pixel-to-ray mapping, but leave open *how this mapping should be represented to remain both compact and expressive simultaneously*.

Monocular 3D. Recent foundation models have improved depth estimation through large-scale training and distillation (Yang et al., 2024), but do not explicitly model the camera geometry. Depth Anything 3 (Lin et al., 2025a) instead predicts depth with per-pixel rays by modeling camera geometry alongside scene structure. Camera-aware models specifically account for camera geometry in monocular depth estimation. UniK3D (Piccinelli et al., 2025) represents a per-pixel ray field with spherical-harmonics and predicts radial distance along each ray to handle diverse camera geometries without camera metadata. Depth Any Camera (Guo et al., 2025) and UniDAC (Ganesan et al., 2026) instead transform images into a common equirectangular representation to process different camera types before depth estimation, but require input camera intrinsics for the transformation. CAM3R (Guruprasad et al., 2026) and Wid3R (Jung et al., 2026) extend the spherical-harmonic ray representation to multi-view feed-forward 3D reconstruction. Together, these works highlight the value of explicitly accounting for the pixel-to-ray mapping for camera-agnostic 3D geometry, while leaving open *whether a single ray representation is necessary for all camera geometries*.

3 Method

Given a single image $I \in \mathbb{R}^{H\times W\times 3}$ from an arbitrary central camera, with no intrinsics, distortion parameters or camera-type label, OmniRay predicts for every pixel $u = (u_x, u_y)$ a unit ray $\mathbf{d}(u) \in S^2$ and a radial distance $\rho(u) > 0$.

A DINOv2 backbone (Oquab et al., 2023) encodes the image and its features drive two branches (Fig. 2): the **Ray Module** composes $\mathbf{d}$ from radial, spherical-harmonic, and free-form components, while the **Distance Module** predicts ρ. Their combination gives the 3D point in camera frame:

$$\mathbf{P}(u) = \rho(u)\,\mathbf{d}(u). \tag{1}$$

We predict radial distance ρ rather than perpendicular depth Z, related by $\rho = Z/d_z$, since $d_z \to 0$ as the ray approaches 90° from the optical axis and changes sign beyond it; ρ therefore remains well-defined across wide fields of view.

Notation. The DINOv2 backbone produces a dense feature map F and a pooled global token $\mathbf{g}$. We use the normalized image coordinate $\mathbf{m}(u) = (u - c_{\text{img}})/D$, where c_{img} is the image center and D is the image half-diagonal in pixels. This normalization places the image corners at $\|\mathbf{m}\| = 1$ for any resolution and aspect ratio. Image scale is provided separately to $\mathbf{g}$ through $\log(W/H)$ and the logarithm of the image diagonal. Ground-truth quantities are marked with a star, $\mathbf{d}^\star$ and $\rho^\star$, and V denotes the set of pixels at which they are defined.

3.1 Ray Module

The module comprises three stages, $\mathbf{d}_0 \to \mathbf{d}_1 \to \mathbf{d}$, in which radial base and spherical-harmonic residual are predicted from the global token $\mathbf{g}$, and free-form residual from the dense feature map F.

Radial base models the approximately radially symmetric structure of conventional central cameras with a monotone mapping from image radius to incidence angle.

Monotonicity restricts the base to injective radius-to-angle maps, so no two radii in the modeled range receive the same incidence angle, and the profile is evaluated directly rather than inverted numerically as a parametric projection model would require. A head maps $\mathbf{g}$ to $K + 8$ values (Fig. 2b): K increments $\mathbf{s} \in \mathbb{R}^K$ between fixed knot radii $0 = r_0 < \cdots < r_K$, with a principal-point offset $\mathbf{c}$, an axis-angle rotation $\boldsymbol{\omega}$, and three parameters forming an upper-triangular map $A(\mathbf{a})$ for anisotropic scaling and shear. These give the transformed coordinate $\mathbf{m}' = A(\mathbf{a})(\mathbf{m} - \mathbf{c})$, with radius r and azimuth φ. The increments are passed through a softplus to ensure positivity, accumulated into knot heights, and linearly interpolated,

$$\theta_k = \sum_{j \le k} \operatorname{softplus}(s_j), \quad \theta_0 = 0, \qquad \theta(r) = \operatorname{lerp}\big(\operatorname{clip}(r; r_0, r_K);\ \{r_k\},\ \{\theta_k\}\big), \tag{2}$$

so that $\theta(r)$ is strictly increasing within the knot span and constant outside it. We use $K = 16$ knots spanning $r \in [0, 1.3]$ in half-diagonal units. The corresponding ray is

$$\mathbf{d}_0 = R(\boldsymbol{\omega}) \begin{bmatrix} \sin\theta(r)\cos\varphi \\ \sin\theta(r)\sin\varphi \\ \cos\theta(r) \end{bmatrix}, \tag{3}$$

where $R(\boldsymbol{\omega})$ is rotation matrix of $\boldsymbol{\omega}$. The largest angle the base can produce is sum of its increments, so nothing caps θ at 90° and the base already represents lenses whose rays pass the equator.

Spherical-harmonic residual. The radial base restricts the incidence angle to a monotone function of image radius, which becomes insufficient when the image covers a large portion of the viewing sphere (Fig. 3b). A spherical-harmonic residual supplies an angular correction to it (Fig. 2b). A head maps $\mathbf{g}$ to a single coefficient matrix $W_1 \in \mathbb{R}^{(L+1)^2 \times 3}$ for a real spherical-harmonic basis shared across all pixels. Thus, the correction field is represented by $3(L+1)^2$ predicted coefficients rather than an independent vector at each pixel. Each pixel evaluates the basis at an address on S^2 obtained from its normalized image coordinate through the inverse-stereographic lift,

$$\mathbf{b}(\mathbf{m}) = \frac{\big(2\mathbf{m},\ 1 - \|\mathbf{m}\|^2\big)}{1 + \|\mathbf{m}\|^2}, \qquad \mathbf{w}_1(u) = \Phi_L\big(\mathbf{b}(\mathbf{m}(u))\big)^\top W_1, \tag{4}$$

where Φ_L stacks the $(L+1)^2$ basis values. We use $L = 7$, giving 192 coefficients per image.

The correction is applied through the exponential map on S^2. For any ray $\mathbf{d}$, only the component of $\mathbf{w}$ orthogonal to $\mathbf{d}$ is a tangent direction, so the remaining component is discarded, leaving $\mathbf{t} = \mathbf{w} - \langle \mathbf{w}, \mathbf{d} \rangle \mathbf{d}$. Its norm $n = \|\mathbf{t}\|$ is the rotation angle in radians, and the ray is transported that far along the great circle through $\mathbf{d}$ in the direction of $\mathbf{t}$,

$$\exp_{\mathbf{d}}(\mathbf{w}) = \cos(n)\,\mathbf{d} + \frac{\sin n}{n}\,\mathbf{t}. \tag{5}$$

Free-form residual. The spherical-harmonic residual provides a compact global correction but remains limited by the finite-order basis. The free-form residual captures the remaining local variation using a lightweight convolutional decoder on the dense feature map F. The resulting correction map is interpolated at each pixel to obtain $\mathbf{w}_2(u)$, so its spatial detail is limited by the resolution of F. The two residuals are applied in turn through equation 5, giving the composite ray field

$$\mathbf{d}_1 = \exp_{\mathbf{d}_0}(\mathbf{w}_1), \qquad \mathbf{d} = \exp_{\mathbf{d}_1}(\mathbf{w}_2), \tag{6}$$

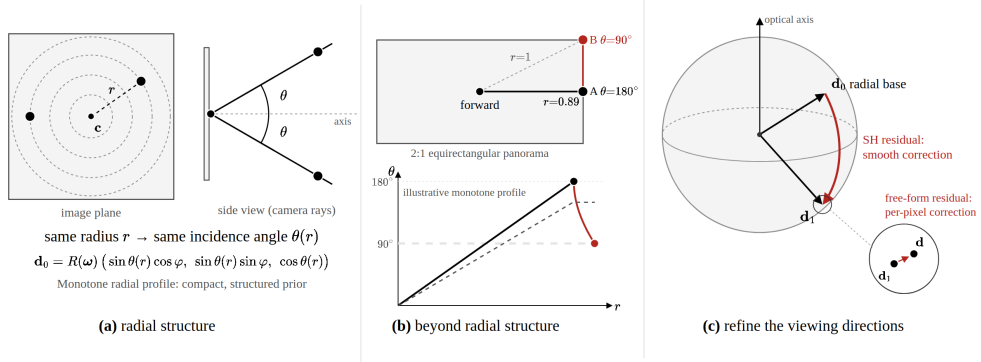


Figure 3: **Why combine the radial construction with the residual corrections?** (a) The radial representation gives the same incidence angle to pixels that lie at the same radius in the base coordinates. (b) In the centered 2:1 equirectangular projection, point B is closer to the center of the image than point A, yet has a lower incidence angle. A monotonic radial function cannot accommodate such ordering under fixed coordinate transformation and orientation. (c) The spherical-harmonic and free-form residuals improve upon the base direction via iterative correction on the unit sphere.

with $\mathbf{d}_0$ given by equation 3, and with $\mathbf{t}_1$ and $\mathbf{t}_2$ the tangential parts of $\mathbf{w}_1$ at $\mathbf{d}_0$ and of $\mathbf{w}_2$ at $\mathbf{d}_1$. Both residual heads are zero-initialized, so the composition starts from the radial base, and the norms $\|\mathbf{t}_1\|$ and $\|\mathbf{t}_2\|$ are the rotation angles each residual applies, measuring how far a camera departs from the representation beneath it and letting the representation adapt without model selection or routing.

3.2 Distance Module

The Ray Module determines the direction $\mathbf{d}$ along which each pixel observes the scene, but not how far along that ray the surface lies. The Distance Module predicts this remaining quantity: the up-to-scale radial distance $\rho(u)$ from the optical center. It reuses the same encoder pass, fusing four intermediate blocks with a DPT decoder (Ranftl et al., 2021) into a dense prediction sampled at each pixel, where a softplus gives a positive distance $\rho(u) = \mathrm{softplus}(f_\rho(u)) > 0$. Scaling each predicted ray by its distance as in equation 1 produces the camera-frame pointcloud $\{\mathbf{P}(u)\}$ in a single forward pass.

3.3 Training Objective

Angular ray loss supervises the final ray field and the two intermediate fields with the mean cosine residual $\ell(\mathbf{x}) = \frac{1}{|V|}\sum_{i\in V}\big(1 - \langle \mathbf{x}_i, \mathbf{d}_i^\star\rangle\big)$,

$$\mathcal{L}_{\text{ray}} = \ell(\mathbf{d}) + \gamma_0\,\ell(\mathbf{d}_0) + \gamma_1\,\ell(\mathbf{d}_1) + \mathcal{L}_{\text{region}}. \tag{7}$$

This deep supervision of $\mathbf{d}_0$ and $\mathbf{d}_1$ keeps each stage responsible for the structure it represents, so a correct final ray cannot be reached through an arbitrary base.

$\mathcal{L}_{\text{region}}$ weights the same per-pixel error more heavily where it is largest.

Writing $\bar{e}(S)$ for the mean error over a pixel set S, it is $\lambda_{\text{c}}\,\bar{e}(C) + \lambda_{\text{p}}\,\bar{e}(P) + \lambda_{\text{t}}\,\bar{e}(T)$ over an inner region C, an outer region P, and the tail T of 5% pixels with the largest error, with λ_{p} the largest of the three since error concentrates at the periphery for wide fields of view. The regions are thresholds on the image radius $\|\mathbf{m}\|$ relative to its maximum, so the split follows the imaged region of a circular fisheye as well as a rectangular frame and does not move with the predicted principal point.

Pointcloud loss supervises rays and distances jointly. Each cloud is normalized by its own mean point norm $z = \frac{1}{|V|}\sum_{i\in V}\|\mathbf{P}_i\|$, which removes the global scale that a single image leaves undetermined,

$$\mathcal{L}_{\text{pc}} = \frac{1}{|V|}\sum_{i\in V}\left\|\frac{\rho_i\mathbf{d}_i}{z_{\text{pred}}} - \frac{\rho_i^\star\mathbf{d}_i^\star}{z_{\text{gt}}}\right\|_2. \tag{8}$$

It is evaluated on $\mathbf{P} = \rho\,\mathbf{d}$ with the ray attached, so a ray error displaces the point even for an exact distance and the gradient reaches both modules.

Overall objective. $\mathcal{L} = \lambda_{\text{ray}}\,\mathcal{L}_{\text{ray}} + \lambda_{\text{pc}}\,\mathcal{L}_{\text{pc}}$. Both modules and the shared encoder are trained end-to-end under this objective. Sec. A.1 gives the head architectures, the loss weights, and the full training schedule.

Table 1: **Oracle representation fitting.** Mean angular error in degrees after fitting each representation directly to ground-truth rays; lower is better. #Params denotes parameters per camera. Best results among compact representations are **bold**. The free field is shown separately as a high-capacity reference.

		Angular error (°) ↓		
Representation	#Params	Fisheye	Equirectangular	Catadioptric
Pinhole	6	5.731	48.080	50.434
Double sphere	8	0.021	22.932	41.876
Radial ($K = 16$)	24	**0.002**	22.802	0.104
Radial ($K = 32$)	40	**0.002**	22.801	**0.023**
SH ($L = 7$)	195	0.276	0.641	0.295
Radial$_{16}$ + SH$_7$ residual (ours)	216	0.075	**0.217**	0.174
Free field	76.8k	0.005	0.013	0.015

4 Experiments

4.1 Oracle test: representation ceilings

Before training anything, we measure what each candidate representation can express at all, by fitting its parameters directly to analytic ground-truth rays; Sec. A.2 gives the fitting protocol. Table 1 shows that compact parametric models succeed only on the geometries their assumptions match, and that increasing the number of knots in the monotone radial profile reduces the catadioptric error but leaves the equirectangular error at $22.8°$, the ceiling identified in Sec. 3.1. The spherical-harmonic field removes that ceiling but remains two orders of magnitude behind the radial profile on the fisheye. Only the composition is below $0.25°$ on all three geometries, and since it reduces to the radial base when the residual vanishes, its capacity is bounded below by that of the base.

4.2 Datasets and Camera-types

We train on two panoramic datasets, 2D3DS (Armeni et al., 2017), of indoor offices and corridors, and 360Loc (Huang et al., 2024), of larger indoor and outdoor environments, rendering perspective, fisheye, and panoramic views from each panorama; the 2D3DS perspective images are the real captures provided with the dataset. We evaluate on held-out scenes of both, and on two unseen datasets, PanoCity (Guo et al., 2026), of outdoor street-level captures, and CO3Dv2 (Reizenstein et al., 2021), of object-centric images. Two camera configurations are withheld from training entirely, a GoPro fisheye and a Nayar catadioptric camera (Nayar, 1997; Baker & Nayar, 1999), both satisfying our central-camera assumption. Sec. A.2 gives the splits and the camera specifications.

Random-lens Augmentation. During training we augment each panorama with randomly sampled synthetic central cameras, each defined by a viewing direction, a monotone radial profile drawn from positive spline increments, and an imaged sensor region. This single family covers filled and annular domains, from pinhole to wide-angle lenses. Draws whose profile falls close to a held-out geometry are rejected, so neither withheld camera nor its neighborhood appears in training. Sec. A.2 gives the sampling ranges and the rejection criteria, and Fig. 7 shows sampled examples.

4.3 Benchmark and Ablation

Baselines. UniK3D (Piccinelli et al., 2025) and Depth Anything 3 (DA3) (Lin et al., 2025a) jointly predict camera geometry and scene structure and are evaluated on all metrics; we fine-tune UniK3D on the 2D3DS and 360Loc panoramas so that it sees the same source imagery we train on. GeoCalib (Veicht et al., 2024) and AnyCalib (Tirado-Garín & Civera, 2025) calibrate the camera without predicting scene geometry. We convert their estimated cameras to dense ray fields for evaluation so all methods are scored identically.

Metrics. We evaluate the predicted ray fields, radial distance, and 3D pointclouds using the following metrics: Ray accuracy is the mean angular error $\arccos\langle \mathbf{d}, \mathbf{d}^\star \rangle$ in degrees, and radial distance is evaluated with AbsRel and $\delta < 1.25$. For pointclouds, accuracy and completeness

Table 2: **Main benchmark.** Results on seen cameras, held-out cameras, and held-out domains. Ang. denotes mean ray error in degrees; CD denotes median relative Chamfer distance, normalized by each scene's mean point norm. Metrics are macro-averaged over test images, and all methods use per-image median scale alignment. Lower is better; best results are **bold**. † denotes exclusion from our training data; a dash indicates that the method does not predict scene geometry. Full per-camera and per-dataset results and additional metrics are provided in Sec. A.3.

	Seen cameras		**Held-out cameras**				**Held-out domains**			
	Aggregate		Catadioptric†		GoPro†		PanoCity†		CO3Dv2†	
Method	Ang. (°) ↓	CD↓	Ang. (°) ↓	CD↓	Ang. (°) ↓	CD↓	Ang. (°) ↓	CD↓	Ang. (°) ↓	CD↓
GeoCalib (Veicht et al., 2024)	30.997	–	119.662	–	14.126	–	47.132	–	3.713	–
AnyCalib (Tirado-Garín & Civera, 2025)	28.655	–	107.129	–	10.031	–	42.335	–	2.711	–
DA3 (Lin et al., 2025a)	33.028	0.1999	121.566	1.0412	16.637	0.1097	50.533	0.3677	**2.041**	0.0288
UniK3D (Piccinelli et al., 2025)	6.336	0.0760	75.702	0.5948	6.342	0.0454	21.220	0.2120	2.676	**0.0277**
OmniRay w/o AUG	1.385	0.0592	30.867	0.3440	7.965	0.0641	6.352	0.0984	2.825	0.0330
OmniRay + RG	0.356	0.0205	**1.118**	**0.0113**	**5.180**	**0.0375**	1.592	0.0277	2.935	0.0411
OmniRay	**0.140**	**0.0179**	1.120	0.0131	5.533	0.0392	**1.350**	**0.0240**	2.967	0.0380

are the median nearest-neighbor distances normalized by the scene's mean point norm, and their mean is the Chamfer distance (CD). Predictions in different depth gauges are aligned per image by $s = \mathrm{median}(\rho^\star)/\,\mathrm{median}(\rho)$, and perpendicular depth is converted to radial distance by $\rho = Z/d_z$ using the predicted rays. All metrics are computed over pixels with valid ground truth.

Results. Table 2 summarizes performance across camera configurations and domains. On the seen camera types, OmniRay achieves a mean ray error of $0.14°$ and relative Chamfer distance of 0.0179, compared with $6.34°$ and 0.0760 for UniK3D and $33.03°$ and 0.1999 for Depth Anything 3. On the held-out configurations: for the Nayar catadioptric camera, all baselines exceed $75°$ while ours achieves $1.12°$ and 0.0131 CD; on GoPro, ours achieves $5.53°$ versus $6.34°$ for UniK3D. On PanoCity, ours achieves $1.35°$ ray error versus $21.22°$ and $50.53°$ for UniK3D and DA3, respectively. On CO3Dv2, our $2.97°$ ray error remains within a degree of both baselines.

The **w/o AUG** row isolates the random-lens sampling of Sec. 4.2. Without it the seen cameras remain strong at $1.39°$, but the held-out catadioptric and GoPro errors rise to $30.87°$ and $7.97°$; restoring it recovers $1.12°$ and $5.53°$ and improves the seen cameras to $0.14°$. Fig. 5a shows the same effect on a catadioptric example, where the unaugmented model exhibits larger ray-field and geometric errors. These results indicate that random-lens sampling is important for extending the composite representation beyond the camera configurations encountered during training. Adding ray guidance to each DPT scale following Lin et al. (2025b) (**+ RG**) leaves the metrics unchanged to within run-to-run variation; the injection is zero-initialized and could use the prompt if it helped, so we omit it.

Fig. 4 compares ray fields and camera-frame pointclouds on unseen PanoCity and CO3Dv2 images, with each ray field colored by its angular error. On the panorama, OmniRay keeps the road continuous while UniK3D breaks it across the image and DA3 fails to recover the ground structure. On the fisheye and GoPro rows, OmniRay keeps the road, pavement and lane markings straight where UniK3D bends them and DA3 also distorts the surrounding buildings. On the held-out catadioptric camera both baselines produce degraded reconstructions. On the CO3Dv2 pinhole example, OmniRay remains comparable to the baselines, demonstrating that the additional flexibility for wide-angle cameras does not come at the expense of pinhole imagery. On in-the-wild images, OmniRay also recovers coherent 3D structure, as shown in Fig. 6.

We also perform a controlled ray-only ablation as shown in Table 3 on a small held-out subset of 2D3DS to isolate the effect of each design choice. The radial base performs well on fisheye and pinhole cameras but fails on the panoramic case. Combining the radial base with a spherical *residual* improves the panorama while retaining the radial performance on forward-facing cameras. Adding the free-form residual improves the held-out performance.

5 Conclusion

We began with two questions: how the pixel-to-ray mapping can remain both compact and expressive, and whether a single representation is necessary for every camera geometry? **OmniRay** an-

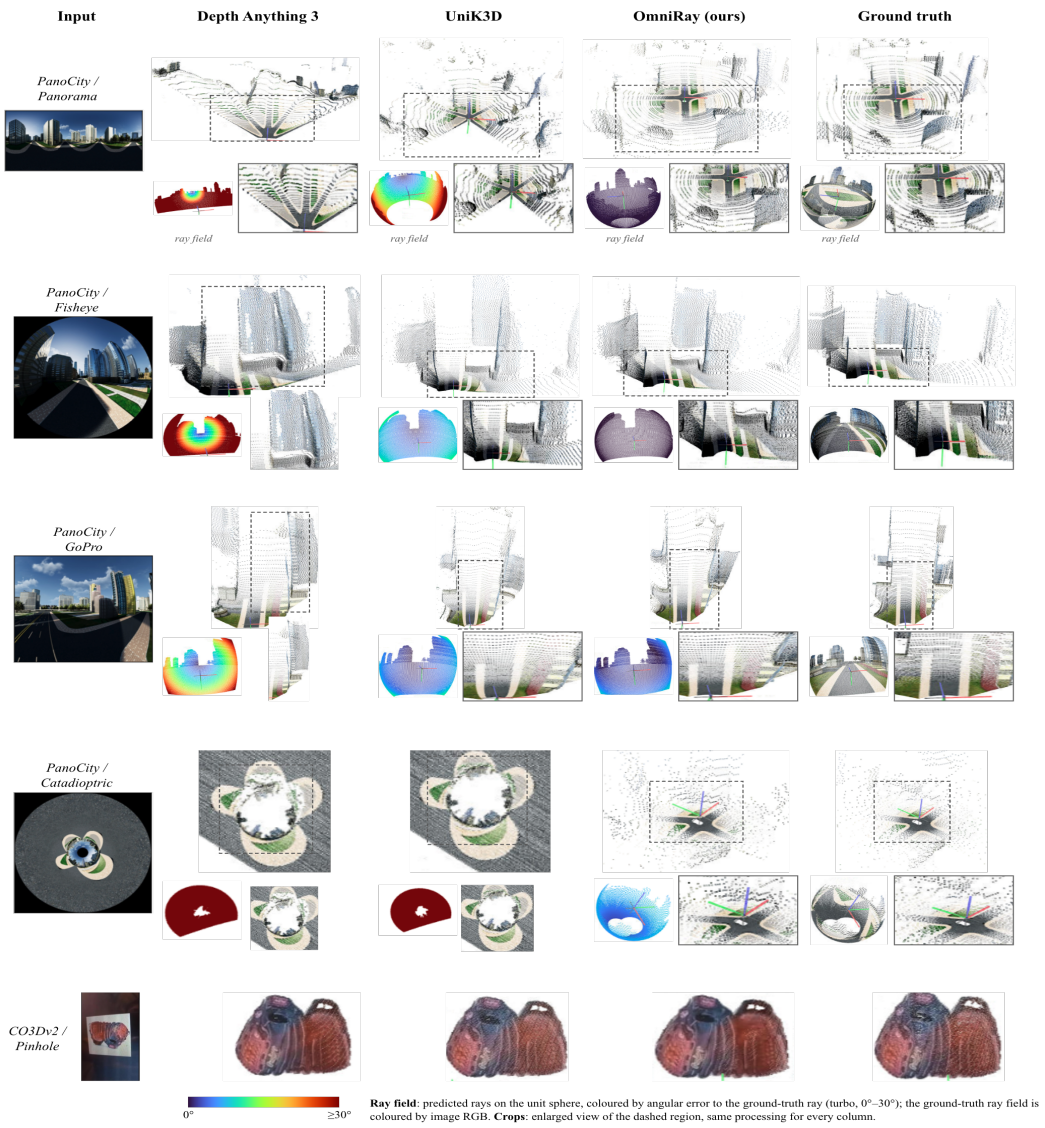


Figure 4: **Qualitative reconstructions on unseen datasets.** We compare OmniRay with Depth Anything 3 and UniK3D on PanoCity and CO3Dv2 across five camera configurations. Small visualizations show viewing-ray fields; larger renderings show the corresponding pointclouds, with ground truth in the rightmost column. OmniRay better preserves scene structure in the panoramic and catadioptric examples without camera intrinsics or camera-type input.

Table 3: **Ray representation ablation.** Mean angular ray error in degrees (↓) on a small held-out subset of 2D3DS. Seen camera families are evaluated on held-out instances; † denotes camera families excluded from training. Only ray prediction is evaluated. Best results are **bold**.

		Seen cameras			Held-out cameras†	
Ray representation	Training	Panorama	Fisheye	Pinhole	Catadioptric	GoPro
Spherical base	–	1.20	8.90	4.80	11.00	19.00
Radial base	–	22.80	0.30	0.20	1.50	**5.00**
Radial + SH residual	Joint	2.50	1.50	1.10	9.00	11.00
Radial + SH residual	Staged	1.10	0.30	0.30	2.00	7.00
Radial + SH + free-form (ours)	Staged	**0.12**	**0.16**	**0.16**	**1.20**	5.50

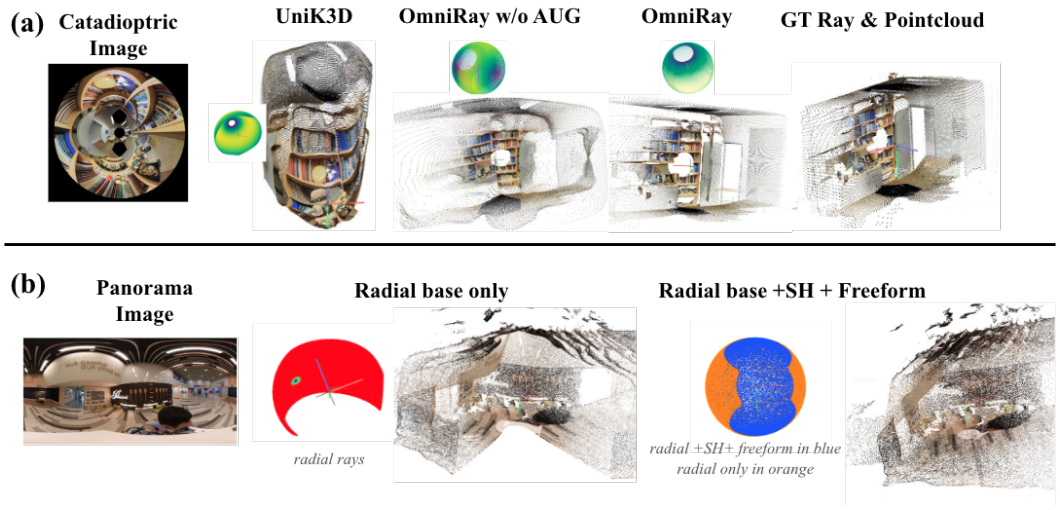


Figure 5: **Ablation.** **(a)** Held-out catadioptric camera: predicted ray field and pointcloud for UniK3D, OmniRay trained without random-lens augmentation, and the full model, against the ground truth. **(b)** Panoramic image: the radial base alone covers only part of the sphere (orange) and leaves the reconstruction torn, while the spherical-harmonic and free-form residuals complete the field (blue) and close the geometry.



Figure 6: **In-the-wild reconstructions.** Qualitative reconstructions on photographs collected from the internet, shown in order as a panorama, a fisheye image, and a pinhole image. Although these camera families are represented during training, the images come from different camera instances with unknown intrinsics and lens parameters. OmniRay requires no camera metadata at inference.

swers both by composing the ray field rather than mapping every camera into one basis. A monotone radial base carries the structure conventional lenses already possess, a spherical-harmonic residual supplies the smooth deviations a radial profile cannot express, and a free-form residual accounts for the remaining local variation, each applied as a rotation on S^2 so that the representation adapts to the camera without camera-type input or model selection. A shared distance module scales the predicted rays into a point cloud in a single forward pass, and sampling lenses continuously during training keeps the composition accurate on geometries it has never seen.

The oracle analysis establishes the ceilings that motivate this design, and the trained model follows them: a mean ray error of $0.14°$ on the camera families seen in training, and $1.12°$ on a catadioptric geometry withheld from it, where every baseline exceeds $75°$. The reconstruction is recovered only up to a single global scale, so all point clouds are compared after a per-image alignment. Since the representation is defined per image and per pixel, extending OmniRay to N-view reconstruction is a natural next step.

AI USE STATEMENT

In this work, we used generative AI tools for creating or modifying scientific figures or images, create or edit software code, edit a research paper to improve readability, propose a title or keywords for a research paper and writing assistance. We have not used generative AI tools for help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, propose or refine hypotheses, design or provide feedback on research methodology or experiments, implement methods, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. We reviewed LLM generated text to ensure it was consistent with our findings. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

No ethics concerns.

REPRODUCIBILITY STATEMENT

We have provided extensive architecture, training, and implementation details which can be used for reproducing our results. We also plan to open-source code upon publication.

REFERENCES

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

Simon Baker and Shree K Nayar. A theory of single-viewpoint catadioptric image formation. *International journal of computer vision*, 35(2):175–196, 1999.

Oleksandr Bogdan, Viktor Eckstein, Francois Rameau, and Jean-Charles Bazin. Deepcalib: A deep learning approach for automatic intrinsic calibration of wide field-of-view cameras. In *Proceedings of the 15th ACM SIGGRAPH European conference on visual media production*, pp. 1–10, 2018.

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

Girish Chandar Ganesan, Yuliang Guo, Liu Ren, and Xiaoming Liu. Unidac: universal metric depth estimation for any camera. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26953–26963, 2026.

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

Yijing Guo, Mengjun Chao, Luo Wang, Tianyang Zhao, Haizhao Dai, Yingliang Zhang, Jingyi Yu, and Yujiao Shi. Panovggt: Feed-forward 3d reconstruction from panoramic imagery. *arXiv preprint arXiv:2603.17571*, 2026.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26996–27006. IEEE, 2025.

Namitha Guruprasad, Abhay Yadav, Cheng Peng, and Rama Chellappa. Cam3r: Camera-agnostic model for 3d reconstruction. *arXiv preprint arXiv:2603.22631*, 2026.

Huajian Huang, Changkun Liu, Yipeng Zhu, Hui Cheng, Tristan Braud, and Sai-Kit Yeung. 360loc: A dataset and benchmark for omnidirectional visual localization with cross-device queries. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22314–22324, 2024.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17307–17316. IEEE, 2023.

Dongki Jung, Jaehoon Choi, Adil Qureshi, Somi Jeong, Dinesh Manocha, and Suyong Yeon. Wid3r: Wide field-of-view 3d reconstruction via camera model conditioning. *arXiv preprint arXiv:2602.05321*, 2026.

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.

Bogdan Khomutenko, Gaëtan Garcia, and Philippe Martinet. An enhanced unified camera model. *IEEE Robotics and Automation Letters*, 1(1):137–144, 2015.

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.

Haotong Lin, Sili Chen, Junhao Liew, Donny Y Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025a.

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

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

Shree K Nayar. Catadioptric omnidirectional camera. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pp. 482–488. IEEE, 1997.

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

Luigi Piccinelli, Christos Sakaridis, Mattia Segu, Yung-Hsu Yang, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unik3d: Universal camera monocular 3d estimation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1028–1039. IEEE, 2025.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12159–12168. IEEE, 2021.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10901–10911, 2021.

Davide Scaramuzza, Agostino Martinelli, and Roland Siegwart. A toolbox for easily calibrating omnidirectional cameras. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 5695–5701. IEEE, 2006.

Thomas Schops, Viktor Larsson, Marc Pollefeys, and Torsten Sattler. Why having 10,000 parameters in your camera model is better than twelve. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2535–2544, 2020.

Javier Tirado-Garín and Javier Civera. Anycalib: On-manifold learning for model-agnostic single-view camera calibration. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 8044–8055. IEEE, 2025.

Vladyslav Usenko, Nikolaus Demmel, and Daniel Cremers. The double sphere camera model. In *2018 International Conference on 3D Vision (3DV)*, pp. 552–560. IEEE, 2018.

Igor Vasiljevic, Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Wolfram Burgard, Greg Shakhnarovich, and Adrien Gaidon. Neural ray surfaces for self-supervised learning of depth and ego-motion. In *2020 International Conference on 3D Vision (3DV)*, pp. 1–11. IEEE, 2020.

Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. Geocalib: Learning single-image calibration with geometric optimization. In *European Conference on Computer Vision*, pp. 1–20. Springer, 2024.

Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271. IEEE, 2025.

Scott Workman, Connor Greenwell, Menghua Zhai, Ryan Baltenberger, and Nathan Jacobs. Deepfocal: A method for direct focal length estimation. In *2015 IEEE International Conference on Image Processing (ICIP)*, pp. 1369–1373. IEEE, 2015.

Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10371–10381. IEEE, 2024.

Senthil Yogamani, Ciarán Hughes, Jonathan Horgan, Ganesh Sistu, Padraig Varley, Derek O'Dea, Michal Uricár, Stefan Milz, Martin Simon, Karl Amende, et al. Woodscape: A multi-task, multi-camera fisheye dataset for autonomous driving. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9308–9318, 2019.

Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on pattern analysis and machine intelligence*, 22(11):1330–1334, 2000.

A Appendix

A.1 Implementation and training

Architecture. The encoder is DINOv2 ViT-L/14 (Oquab et al., 2023) with 1024-dimensional features. The radial head is a two-layer MLP with width 256 operating on the global token augmented with $(\log \text{aspect}, \log \text{diagonal})$, while the spherical head is a one-layer MLP of the same width on the same input. The radial head predicts $K + 8$ values with $K = 16$ knots spanning $r \in [0, 1.3]$ in half-diagonal units; the spherical head predicts $(L + 1)^2 \times 3$ coefficients with $L = 7$. These coefficients are shared across all pixels of an image. The free head is a convolutional decoder over the dense feature map. The Distance Module uses a width-256 DPT decoder (Ranftl et al., 2021) over encoder blocks $\{4, 11, 17, 23\}$ to predict the radial distance. Both residual heads are zero-initialized, while the radial head is initialized to a canonical camera, so training starts from a single radial lens.

Training. Images range from 896×1792 equirectangular panoramas to 512^2 lens crops. We train with a token cap of 8192 using a sampler for camera-type balancing. Within each image, we supervise 4096 pixels sampled on a stratified 32×32 grid, ensuring coverage across the image. Training proceeds in two phases. We first train the Ray Module alone, activating its three components in order, and then attach the Distance Module at the final stage and train the full model end to end. In this second phase we use AdamW with weight decay 0.01 and cosine decay after a warm-up, with learning rates of 5×10^{-6} for the encoder and ray heads and 5×10^{-5} for the distance decoder. The encoder runs in `bfloat16`, while the heads and losses remain in full precision. Training uses 8 NVIDIA A40 GPUs with distributed data parallelism.

A.2 Experimental Details.

Oracle fitting protocol. Each representation is fitted to the analytic ground-truth rays of a single camera by minimizing the cosine residual of equation 7, with a multi-start sweep over the axis sign,

roll and, where the representation uses one, the focal length. Every row is an independent fit under a fixed optimization budget, so each entry upper-bounds that representation's ceiling. The fitted error is the representation's approximation capacity for that camera and lower-bounds the error of any predictor built on it.

Splits and Camera sampling. For 2D3DS we train on areas 1-4 and evaluate on areas 5-6; for 360Loc we train on the Atrium and Piatrium scenes and evaluate on Concourse and Hall. Each synthetic camera samples a viewing orientation, an aspect ratio, a principal-point offset, an optional central blind spot, a rectangular or circular sensor boundary, and inner and outer incidence angles; panoramic cameras instead sample a vertical field of view, an azimuth and a small latitude stretch, with the horizon kept gravity-aligned as in the source captures. Draws near a held-out geometry are rejected: for the GoPro, any profile within $\delta = 15^\circ$ of its incidence-angle curve at every radius; for the catadioptric camera, any sample with an inner angle in $[25^\circ, 55^\circ]$ and a rim angle in $[140^\circ, 170^\circ]$. Note: The Nayar catadioptric camera (Nayar, 1997; Baker & Nayar, 1999) is a paraboloid mirror with orthographic projection, imaging a 40°–155° annulus about its axis; the mirror and camera share a single viewpoint. For PanoCity and CO3Dv2 we use the calibration provided with each dataset to obtain ground-truth rays. Ground-truth rays and distances follow analytically from each sampled camera. Fig. 7 shows images rendered under cameras drawn from this distribution.



Figure 7: **Random-lens augmentation.** Images rendered from a single panorama under cameras drawn from the sampling distribution of Sec. 4.2. A camera is a viewing direction, a monotone radial profile and an imaged sensor region, so one family spans narrow perspective crops, wide-angle and fisheye views, off-center and anisotropic sensors, and annular domains. The projection varies continuously from sample to sample, so no fixed map from image position to ray direction fits more than one of them. Draws falling close to either held-out geometry are rejected and never enter training.

A.3 Additional Results

Tables 4, 5 and 6 report every metric of Sec. 4.3 broken out per camera type, per dataset, and over the aggregate seen and held-out splits. Figures 8 and 9 show reconstructions on the 2D3DS and 360Loc test scenes, which share the camera families used in training but none of their scenes, complementing the unseen-dataset comparison of Fig. 4.



Figure 8: **Qualitative reconstructions on seen dataset: 2D3DS**

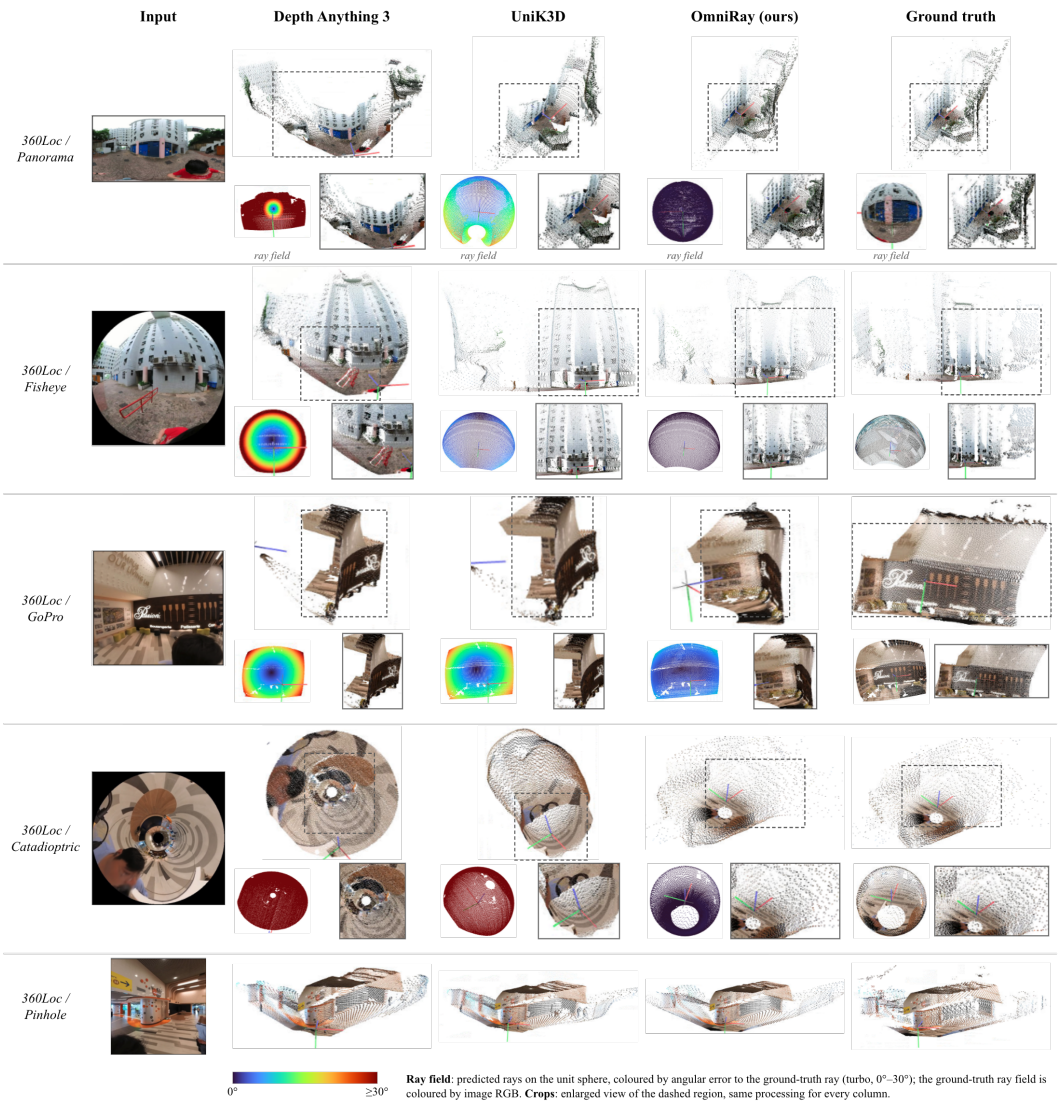


Figure 9: **Qualitative reconstructions on seen dataset: 360Loc**

Table 4: Per-camera results, all metrics. Macro-averaged over test images. † geometry withheld from training. Baselines are median-scale-aligned per image. Acc, Comp and CD are medians of the relative distance, divided by each scene's mean point norm. δ_1 is $\delta < 1.25$. A dash means the method predicts no scene geometry.

		Ray	Radial Dist.		Pointcloud Quality			
Camera	**Method**	**ang°** ↓	**AbsRel**↓	δ_1 ↑	**CD**↓	**Acc**↓	**Comp**↓	**F@10cm**↑
pinhole	GeoCalib	2.196	–	–	–	–	–	–
	AnyCalib	1.043	–	–	–	–	–	–
	DA3	6.604	0.1508	0.862	0.0701	0.0862	0.0540	0.300
	UniK3D	3.619	0.1371	0.871	0.1058	0.1802	0.0315	0.356
	ours	**0.155**	**0.0769**	**0.925**	**0.0242**	**0.0355**	**0.0130**	**0.533**
fisheye	GeoCalib	30.001	–	–	–	–	–	–
	AnyCalib	22.134	–	–	–	–	–	–
	DA3	28.727	0.3225	0.562	0.1888	0.1111	0.2665	0.066
	UniK3D	7.055	0.2155	0.768	0.0543	0.0569	0.0517	0.198
	ours	**0.154**	**0.0906**	**0.924**	**0.0159**	**0.0197**	**0.0121**	**0.531**
panoramic	GeoCalib	60.795	–	–	–	–	–	–
	AnyCalib	62.788	–	–	–	–	–	–
	DA3	63.752	0.3352	0.547	0.3408	0.1956	0.4859	0.009
	UniK3D	8.336	0.2154	0.774	0.0678	0.0723	0.0633	0.174
	ours	**0.110**	**0.0593**	**0.949**	**0.0137**	**0.0163**	**0.0111**	**0.585**
catadioptric†	GeoCalib	119.662	–	–	–	–	–	–
	AnyCalib	107.129	–	–	–	–	–	–
	DA3	121.566	0.1617	0.771	1.0412	0.7232	1.3592	0.000
	UniK3D	75.702	0.2523	0.656	0.5948	0.5047	0.6849	0.004
	ours	**1.120**	**0.0396**	**0.969**	**0.0131**	**0.0130**	**0.0131**	**0.633**
GoPro†	GeoCalib	14.126	–	–	–	–	–	–
	AnyCalib	10.031	–	–	–	–	–	–
	DA3	16.637	0.1657	0.840	0.1097	0.0811	0.1383	0.079
	UniK3D	6.342	0.1665	0.857	0.0454	0.0449	0.0458	**0.235**
	ours	**5.533**	**0.0842**	**0.932**	**0.0392**	**0.0354**	**0.0429**	0.215

Table 5: Per-dataset results, all metrics. Macro-averaged over test images. † unseen in training. Conventions as in Table 4. CO3Dv2 contains perspective and fisheye imagery only, so its zero-shot axis is the domain rather than the camera, and its object-masked COLMAP ground truth covers roughly 12% of pixels.

		Ray	Radial Dist.		Pointcloud Quality			
Scenes	**Method**	**ang°** ↓	**AbsRel**↓	δ_1 ↑	**CD**↓	**Acc**↓	**Comp**↓	**F@10cm**↑
2D3DS	GeoCalib	40.713	–	–	–	–	–	–
	AnyCalib	37.520	–	–	–	–	–	–
	DA3	43.775	0.2124	0.747	0.3614	0.1744	0.5483	0.224
	UniK3D	19.432	0.1272	0.864	0.1487	0.0981	0.1994	0.411
	ours	**1.619**	**0.0568**	**0.958**	**0.0236**	**0.0226**	**0.0246**	**0.767**
360Loc	GeoCalib	48.223	–	–	–	–	–	–
	AnyCalib	42.021	–	–	–	–	–	–
	DA3	48.064	0.2872	0.629	0.3212	0.2121	0.4303	0.048
	UniK3D	19.979	0.2709	0.676	0.1601	0.1455	0.1747	0.154
	ours	**1.274**	**0.0713**	**0.947**	**0.0160**	**0.0160**	**0.0160**	**0.590**
PanoCity†	GeoCalib	47.132	–	–	–	–	–	–
	AnyCalib	42.335	–	–	–	–	–	–
	DA3	50.533	0.1819	0.774	0.3677	0.3318	0.4037	0.001
	UniK3D	21.220	0.1940	0.815	0.2120	0.2719	0.1522	0.015
	ours	**1.350**	**0.0823**	**0.914**	**0.0240**	**0.0333**	**0.0147**	**0.141**
CO3Dv2†	GeoCalib	3.713	–	–	–	–	–	–
	AnyCalib	2.711	–	–	–	–	–	–
	DA3	**2.041**	0.0573	0.950	0.0288	**0.0199**	0.0377	0.282
	UniK3D	2.676	**0.0478**	**0.963**	**0.0277**	0.0233	**0.0322**	**0.293**
	ours	2.967	0.0889	0.914	0.0380	0.0291	0.0469	0.173

Table 6: Aggregate results over the 4500-sample benchmark. Conventions as in Table 4. Seen cameras are perspective, fisheye and panoramic; the held-out geometries are the catadioptric mirror and the GoPro fisheye.

		Ray	Radial Dist.		Pointcloud Quality			
Split	**Method**	**ang°** ↓	**AbsRel**↓	δ_1 ↑	**CD**↓	**Acc**↓	**Comp**↓	**F@10cm**↑
all	GeoCalib	45.356	–	–	–	–	–	–
	AnyCalib	40.625	–	–	–	–	–	–
	DA3	47.457	0.2272	0.716	0.3501	0.2394	0.4608	0.091
	UniK3D	20.211	0.1974	0.785	0.1736	0.1718	0.1754	0.193
	ours	**1.414**	**0.0701**	**0.940**	**0.0212**	**0.0240**	**0.0184**	**0.499**
seen	GeoCalib	30.997	–	–	–	–	–	–
	AnyCalib	28.655	–	–	–	–	–	–
	DA3	33.028	0.2695	0.657	0.1999	0.1310	0.2688	0.125
	UniK3D	6.336	0.1893	0.804	0.0760	0.1031	0.0488	0.242
	ours	**0.140**	**0.0756**	**0.933**	**0.0179**	**0.0238**	**0.0120**	**0.550**
held-out†	GeoCalib	66.894	–	–	–	–	–	–
	AnyCalib	58.580	–	–	–	–	–	–
	DA3	69.102	0.1637	0.806	0.5754	0.4021	0.7488	0.040
	UniK3D	41.022	0.2094	0.756	0.3201	0.2748	0.3654	0.120
	ours	**3.326**	**0.0619**	**0.951**	**0.0261**	**0.0242**	**0.0280**	**0.424**