OmniPoint

Universal Monocular Metric Pointcloud from Any Camera

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

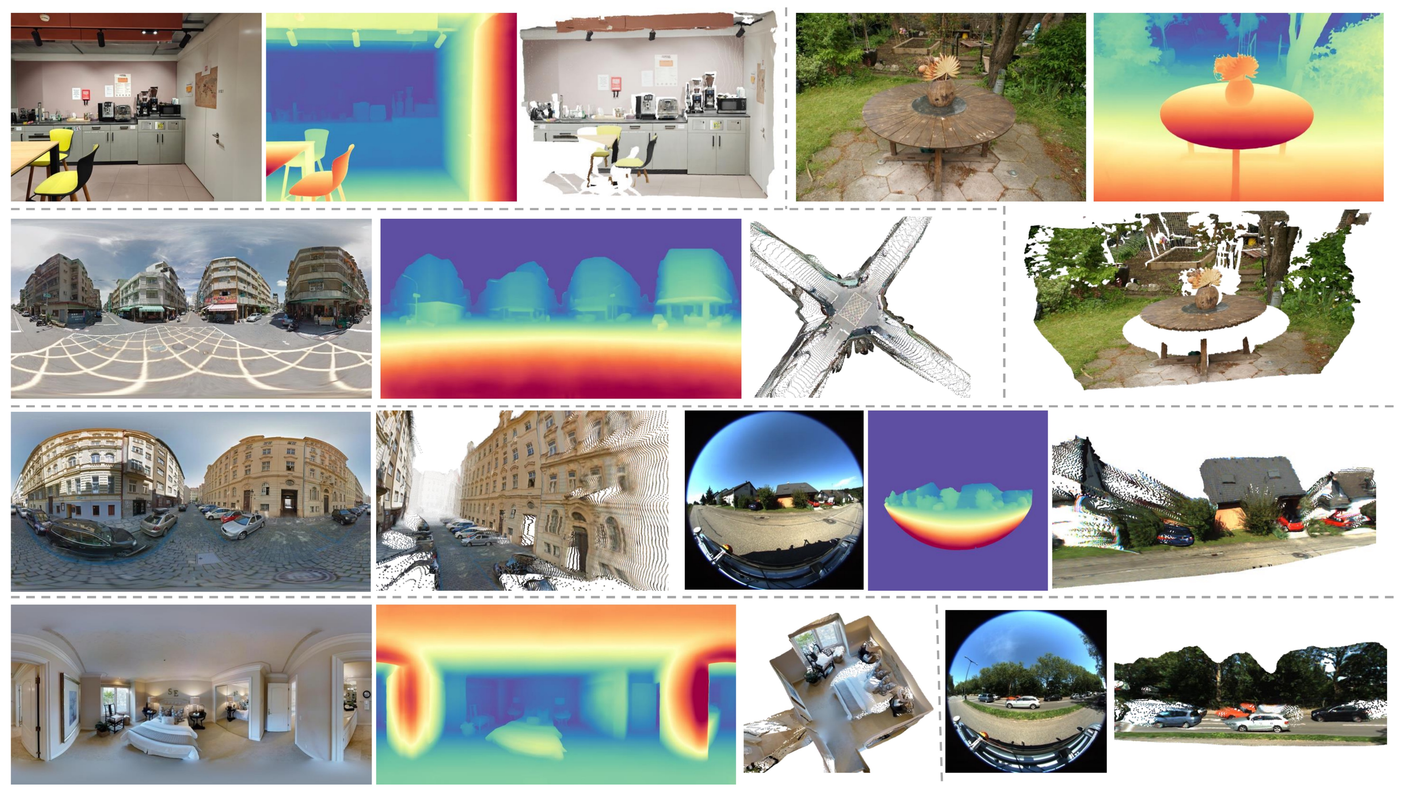
[1]Google DeepMind [2]ETH Zurich

Work done during an internship at Google DeepMind.

ECCV 2026 | MALMÖ | SEPT 8–12

Project page: botaoye.github.io/omnipoint

1. Motivation



Metric point clouds from a single image — pinhole, fisheye, 360° — with intrinsics and sparse depth optional.

☹ **Fragmented today:** a separate model per projection; priors cannot be toggled.

😀 **OmniPoint:** one network, any camera, priors optional.

The only method covering every output, camera and prior

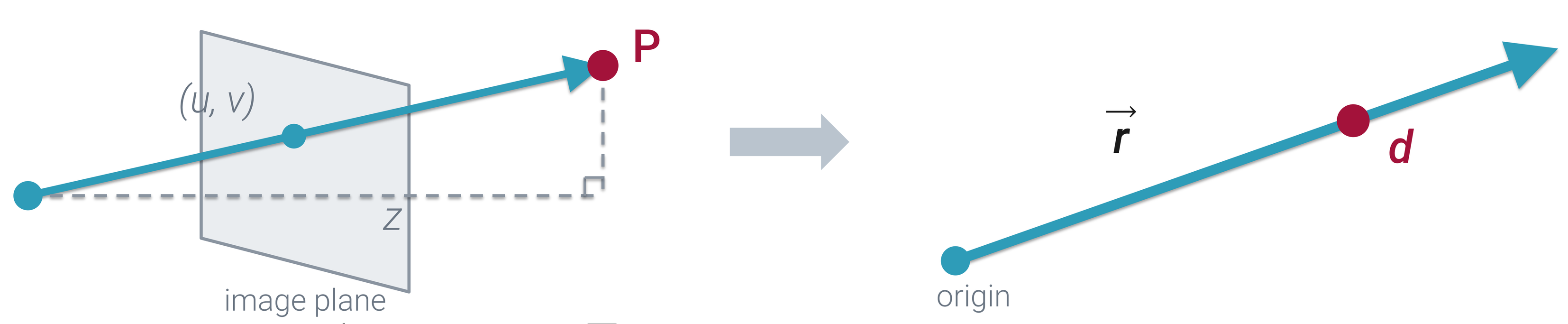
Capability	DA v2	MoGe V2	Depth Pro	UniK3D	DA²	PromptDA	OmniPoint
Affine-inv. depth	✓	✓	✓	✓	–	✓	✓
Affine-inv. points	–	✓	✓	✓	–	–	✓
Metric points	–	✓	✓	✓	–	–	✓
Pinhole	✓	✓	✓	✓	–	✓	✓
Fisheye	–	–	–	✓	–	–	✓
360° panorama	–	–	–	✓	✓	–	✓
Intrinsics prior	–	–	–	✓	–	–	✓
Sparse-depth prior	–	–	–	–	–	✓	✓

2. Camera-Agnostic Ray + Distance

☹ **Planar depth z:** needs a linear pixel→ray map; undefined past 180° FoV.

☹ **Raw (x, y, z):** entangles projection with scene structure.

😀 **Ray × distance:** r absorbs the camera, d carries the scene.



$$\mathbf{P} = z \cdot \mathbf{K}^{-1}[u,\, v,\, 1]^{\top} \qquad \mathbf{P} = d \cdot \mathbf{r}$$

Decoupled training objective

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

Distance is supervised along the **ground-truth ray** — ray errors never pollute it.

Representation	S.FoV	L.FoV	360°
Ray + D (ours)	4.04	6.37	5.79
XYZ point map	4.28	6.74	6.28

Point loss	Rel↓
P + GT ray (ours)	5.55
naive P	5.65
disjoint D + Ray	7.88

3. OmniPoint Pipeline

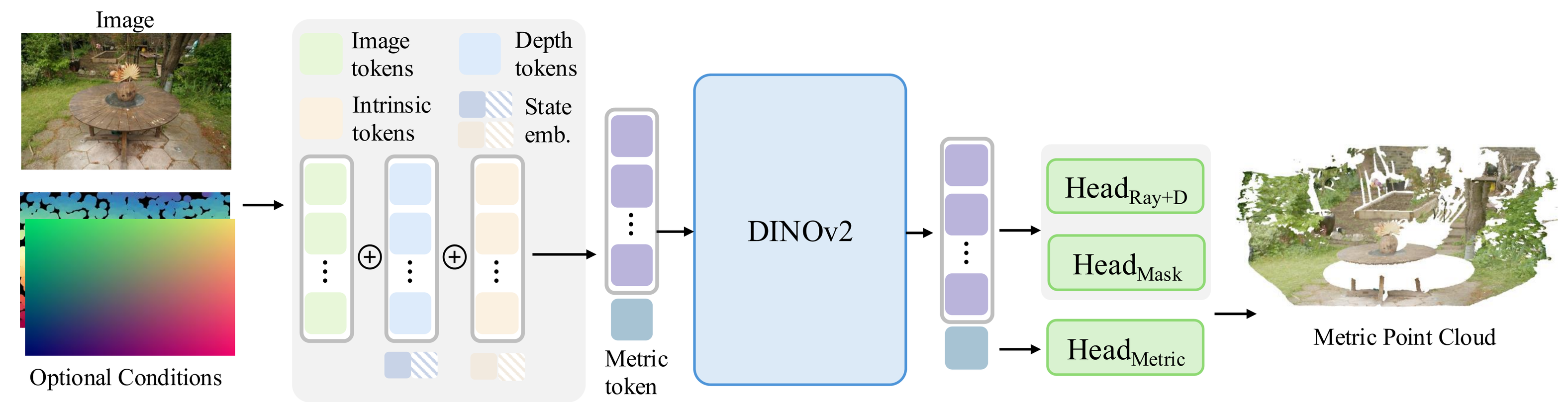


Image tokens + optional intrinsic / sparse-depth tokens → one DINOv2 ViT-L → DPT heads for ray + distance, sky mask, metric scale.

Optional priors — without destabilising the backbone

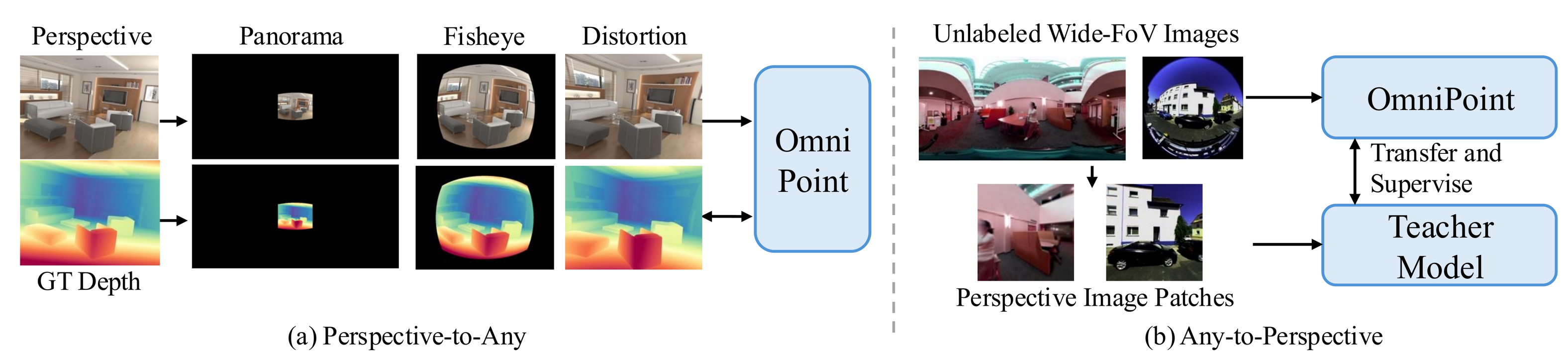
① **Input-state embeddings:** tokens tell the ViT which priors are active.

② **Gaussian smoothing:** sparse points become a dense, consistent guide.

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

Conditioning design (sparse depth)	Rel↓	δ_1↑
Ours (full)	3.21	98.2
w/o Gaussian smoothing	3.80	97.5
w/o input-state embedding	3.38	97.7

4. Bidirectional Data Augmentation

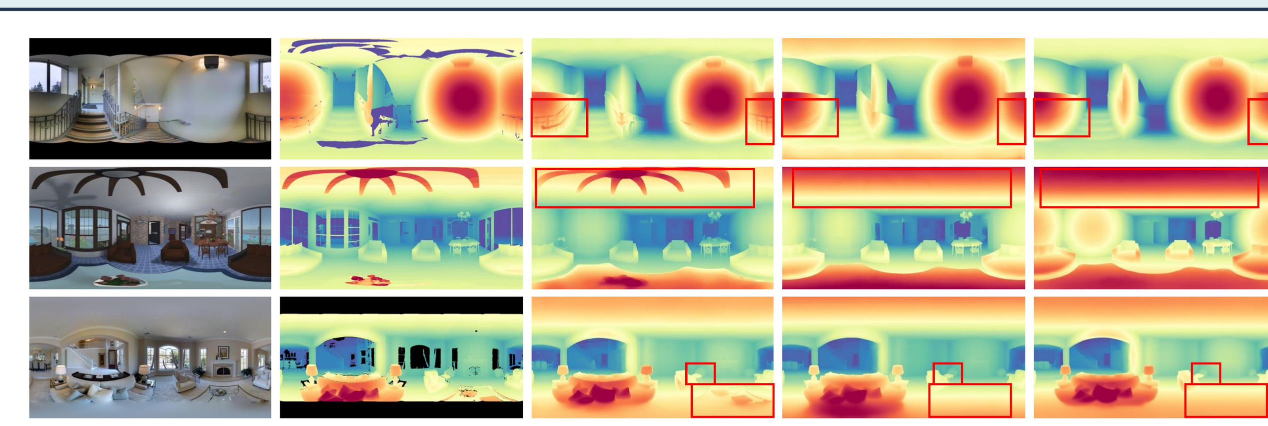


Non-pinhole ground truth is scarce — so we bridge labelled perspective and unlabelled 360° data in 3D.

→ **Perspective-to-Any:** re-project labelled pinhole data into virtual fisheye and 360° cameras.

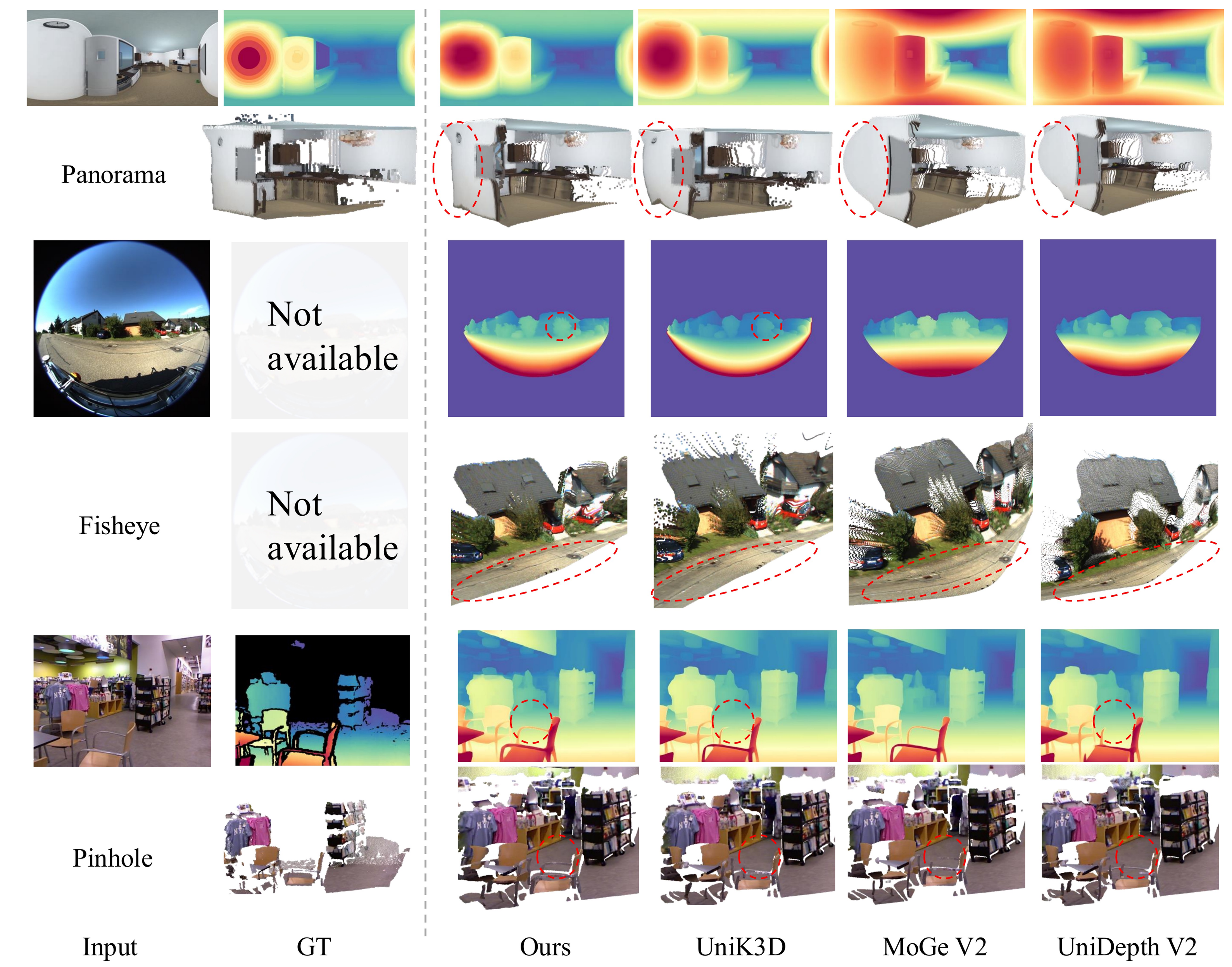
← **Any-to-Perspective:** pseudo-label virtual pinhole crops of unlabelled wide-FoV images.

P2A	A2P	360° Rel↓	360° δ_1↑
–	–	8.45	89.8
✓	–	6.21	94.5
–	✓	8.21	90.9
✓	✓	5.79	95.1



Sharper panorama depth than the specialist DA².

5. Qualitative Results



Vertical walls, straight roads, sharp boundaries

6. Quantitative Results

6.66 fisheye point Rel↓ UniK3D 11.5

5.79 360° depth Rel↓ DA² 6.65

83.5 metric δ_1 mean best image-only

98.0 metric δ_1 mean with sparse depth

Zero-shot relative geometry — Rel↓ (%)

Method	S.FoV		L.FoV		360° Images	
	Depth	Point	Depth	Point	Depth	Point
UniDepth V2	4.23	6.06	7.81	18.4	19.3	102.8
Depth Pro	5.28	7.49	26.6	36.2	–	–
MoGe V1	4.05	**5.25**	9.48	28.1	22.3	102.6
MoGe V2	**3.98**	5.59	12.0	23.4	25.1	102.8
DA3	4.78	6.33	12.8	29.9	25.5	101.5
UniK3D	4.14	5.61	8.77	11.5	10.4	11.4
DA²	–	–	–	–	6.65	7.30
OmniPoint (ours)	4.04	5.55	6.37	6.66	5.79	5.90

Zero-shot metric depth — δ_1↑ (%)

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