POLARIS: POLARIMETRIC SPIN-2 EQUIVARIANT ADAPTERS FOR MONOCULAR METRIC DEPTH

Anonymous authors
Paper under double-blind review

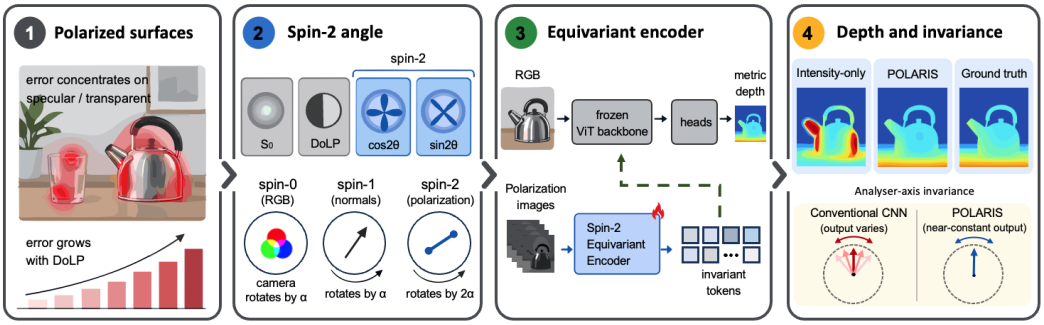


Figure 1: Overview of POLARIS. (1) The error of an intensity-only depth model grows with the degree of polarization (DoLP) of the returned light. (2) A camera roll turns the image and shifts the polarization angle, so polarization is a spin-2 field. (3) An equivariant encoder injects polarization into a frozen backbone, which it recovers exactly without polarization. (4) The branch refines the depth and is invariant to the analyser axis.

ABSTRACT

Monocular depth estimation from a single RGB image is scale-ambiguous: metric scale can be recovered only through learned priors. Existing methods therefore rely on pre-trained backbones mitigating scale ambiguity partially. Polarization provides such measurements, since the angle and degree of linear polarization depend on surface orientation and thus supply geometric cues absent from intensity alone. However, the angle of polarization is measured relative to the sensor's reference axis: a camera roll shifts it in exactly the same way as a change of that axis. An encoder that ignores this transformation therefore becomes dependent on the manufacturer's arbitrary choice of reference axis leading to spin-2 angle ambiguity. To resolve the angle ambiguity, we present **POLARIS**, which builds a rotation-equivariant polarimetric encoder and injects its orientation-independent features into a frozen geometry backbone through a prior-preserving interface. POLARIS lowers metric AbsRel by 16% over the strongest baseline on HAMMER and ranks first or second on all metrics on GM and RPS benchmarks. Our prediction result is rotation-invariant under an arbitrary analyser-axis rotation that changes the output of a conventional encoder. As a side benefit, polarization enables POLARIS to perceive a screen showing a scene as a flat plane, while RGB-only models, including POLARIS with polarization disabled, hallucinate the depth of the displayed content. On our own real-world captures, it also recovers fine structures missed by RGB-only baselines.

1 INTRODUCTION

Single-image metric depth has advanced quickly with geometric foundation models such as MoGe, Metric3D and Depth Anything (Wang et al., 2025b; Hu et al., 2024; Yang et al., 2024a). A common design predicts an affine-invariant point map together with a metric scale and deploys a large *frozen* pretrained backbone with light trainable heads. Despite strong average accuracy, a single intensity image leaves scale and shape ambiguous, and the residual error of such a model grows with the degree of polarization of the returned light (Section 3.2). Strong polarization indicates specular reflection or transmission, which violates the Lambertian assumptions on which these models implicitly rely.

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.

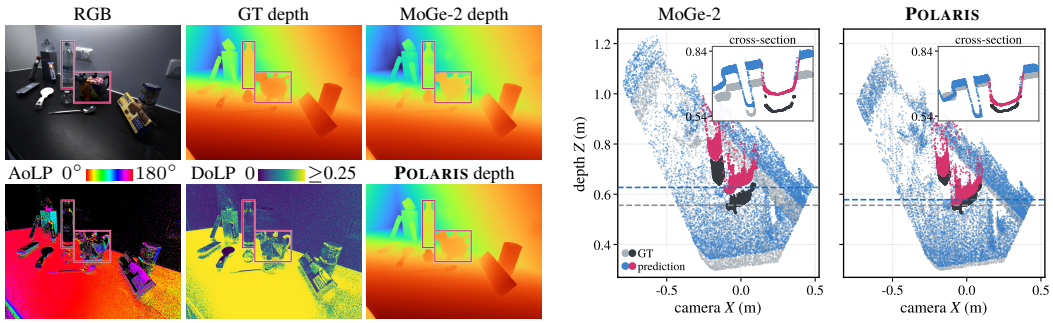


Figure 2: Metric depth and point clouds on an unseen HAMMER scene. The three depth maps share one colour scale, and the AoLP is black where DoLP < 0.1. The right block shows each prediction back-projected with the GT intrinsics and seen from above, without scale or shift alignment, with dashed lines at the median depths. Boxes and pink points mark a PET bottle and a ceramic teapot, and insets magnify a cross-section.

Polarization is also a physical cue bound to surface geometry through Fresnel reflection, and shape-from-polarization has long used the angle and degree of linear polarization (AoLP/DoLP) to recover normals (Atkinson & Hancock, 2006; Ba et al., 2020; Lei et al., 2022). Passive *metric* depth from polarization without a second depth source, on top of a frozen geometric foundation model, is nonetheless largely unexplored. Most polarimetric depth methods need sparse depth (Ikemura et al., 2024) or a stereo pair (Zhu & Smith, 2019), and SfP networks target normals.

The difficulty polarization presents to a standard encoder lies in its reference frame. AoLP is measured against the polarizer array fixed to the sensor, so a camera roll of α moves the pixels *and* shifts every measured angle by $-\alpha$. Turning the analyser axis by α shifts only the angles. Written as the complex field $z = \cos 2\theta + i \sin 2\theta$, the orientation acquires a factor $e^{-i2\alpha}$. Rotation order 2 is absent from RGB (order 0) and from surface normals (order 1). Sections 3.2 and 5.1 quantify this cost. A conventional CNN encoder is 35–87% inconsistent under $k \cdot 90°$ rotations, and its metric error varies by 2.73% when only the analyser axis turns, whereas an encoder built on the right symmetry is exact under these rotations and leaves the error unchanged.

Contributions. (i) POLARIS, a prior-preserving architecture that injects polarization into a frozen metric-depth backbone through a spin-2 equivariant encoder and position-aligned token injection. It reduces exactly to the RGB-only network when polarization is absent, so every polarization ablation is performed within one checkpoint (Fig. 1, Sec. 3). **(ii)** A measurement of the symmetry's contribution at the feature and at the task level, under camera rotation and under analyser-axis rotation. **(iii)** Metric depth on the HAMMER split against six recent monocular depth models and surface normals on a rendered and a real polarimetric set (Sec. 4).

2 Related Work

Monocular depth and geometric foundation models. Monocular depth has moved from regression networks (Eigen et al., 2014; Fu et al., 2018) through zero-shot models (Ranftl et al., 2022; 2021) to large pretrained priors. Depth Anything (Yang et al., 2024a;b) scales pseudo-labelled supervision on DINOv2 (Oquab et al., 2024), and Marigold (Ke et al., 2024) adapts diffusion priors. *Metric* depth additionally resolves the camera. ZoeDepth (Bhat et al., 2023) attaches metric heads, Metric3D (Yin et al., 2023; Hu et al., 2024) canonicalizes the camera, UniDepth (Piccinelli et al., 2024; 2026) predicts it jointly, Depth Pro (Bochkovskii et al., 2025) recovers a sharp metric map and the focal length in one pass, and MoGe (Wang et al., 2025b;c) regresses an affine-invariant point map and recovers focal length and shift at inference. Feed-forward geometry transformers such as VGGT (Wang et al., 2025a) predict camera, depth and point maps together. We adopt a frozen MoGe-2 backbone. The interface follows the parameter-efficient adaptation literature of adapters (Houlsby et al., 2019), low-rank updates (Hu et al., 2022) and zero-initialized control branches (Zhang et al., 2023), and adds an exact-identity condition on the injected modality.

Shape from polarization. Classical SfP infers azimuth and zenith from AoLP/DoLP under the π-azimuth ambiguity and the diffuse/specular $\pi/2$ ambiguity (Atkinson & Hancock, 2006; Miyazaki et al., 2004), recovering height either by coupling polarization with shading (Smith et al., 2019) or by correcting a coarse depth map (Kadambi et al., 2015). Deep SfP replaces hand-built priors with

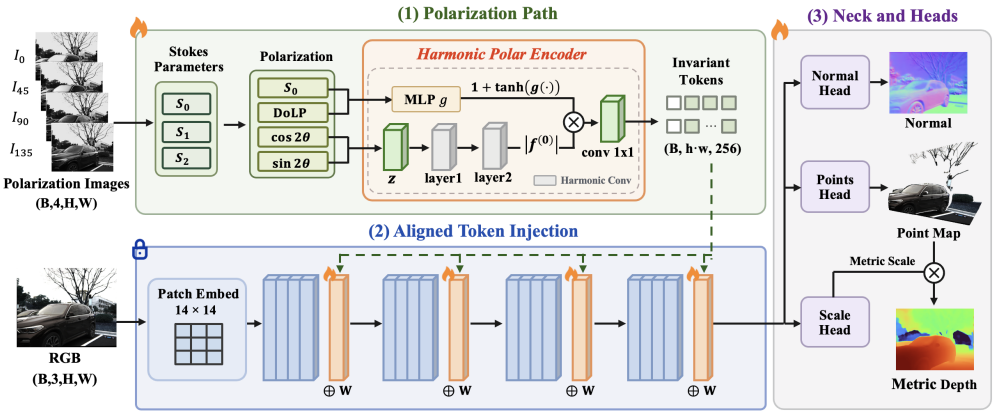


Figure 3: POLARIS architecture. (1) Four-angle polarization is reduced to $(S_0, \mathrm{DoLP}, \cos 2\theta, \sin 2\theta)$. The spin-2 equivariant harmonic encoder turns the axis field $\cos 2\theta + i \sin 2\theta$ into rotation-invariant tokens, and S_0 and DoLP scale these features. (2) After four of the 24 blocks of the *frozen* DINOv2-L encoder of MoGe-2, a linear map adds each token to the patch token at the same position. (3) The neck and heads are fine-tuned.

learned ones (Ba et al., 2020; Li et al., 2026) and extends to in-the-wild scenes (Lei et al., 2022), SVBRDF acquisition (Deschaintre et al., 2021) and transparent objects (Shao et al., 2023), but it targets normals rather than metric depth. Intensity-only normal estimators are a separate line, of which DSINE (Bae & Davison, 2024) is the backbone we adapt in Section 4.2.

Polarimetric depth, transparent and specular surfaces. These surfaces are the persistent failure case of intensity-only depth, motivating dedicated benchmarks (Sajjan et al., 2020; Zama Ramirez et al., 2024) and polarization cues for segmentation (Kalra et al., 2020). Polarimetric depth pipelines *fuse* polarization with a second depth source, a stereo pair (Zhu & Smith, 2019) or multi-view SLAM (Yang et al., 2018). PPFT (Ikemura et al., 2024), the closest to our interface, fuses polarization into a pre-trained depth-completion network through prompt blocks, but it needs a sensor depth map and fine-tunes the whole network. Monocular polarimetric networks learn depth with self- or cross-modal supervision (Blanchon et al., 2021; Verdié et al., 2022), and concurrent work adds a diffusion prior (Yu et al., 2026). HAMMER (Jung et al., 2023) and spectro-polarimetric captures (Jeon et al., 2024) provide real RGB+P data with dense ground truth.

Equivariant representation learning. Group-equivariant (Cohen & Welling, 2016) and steerable CNNs (Cohen & Welling, 2017; Weiler & Cesa, 2019) build symmetry into the architecture, and Harmonic Networks (Worrall et al., 2017) realize continuous rotation equivariance with circular-harmonic filters that map rotation order n through an order-m filter to order $n+m$. Polarization is one of the few modalities whose primary field has rotation order 2, a spin-2 field in the sense of spin-weighted spherical CNNs (Esteves et al., 2020), and we use order-$\{0, 2\}$ harmonic convolutions to carry it into rotation-*invariant* tokens consumable by a standard ViT (Dosovitskiy et al., 2021).

3 Method

3.1 Polarimetric representation

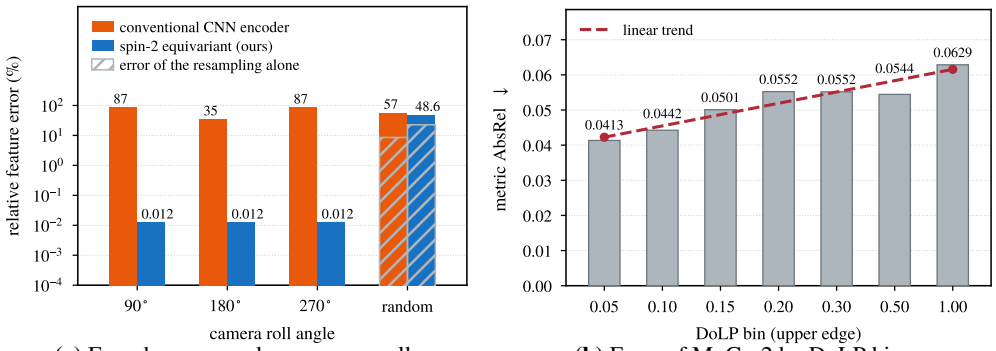
From four linear-polarizer measurements $I_0, I_{45}, I_{90}, I_{135}$ we form the first three Stokes parameters, the degree and angle of linear polarization, and the complex orientation field,

$$S_0 = \tfrac{1}{2}\textstyle\sum_i I_i, \qquad S_1 = I_0 - I_{90}, \qquad S_2 = I_{45} - I_{135},$$
$$\mathrm{DoLP} = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \qquad z = \cos 2\theta + i \sin 2\theta = \frac{S_1 + iS_2}{\sqrt{S_1^2 + S_2^2}}. \tag{1}$$

The network receives the per-pixel tuple $\mathbf{p} = (S_0, \mathrm{DoLP}, \cos 2\theta, \sin 2\theta)$. Unlike θ, which is defined only modulo π, the field z is continuous wherever polarization is defined.

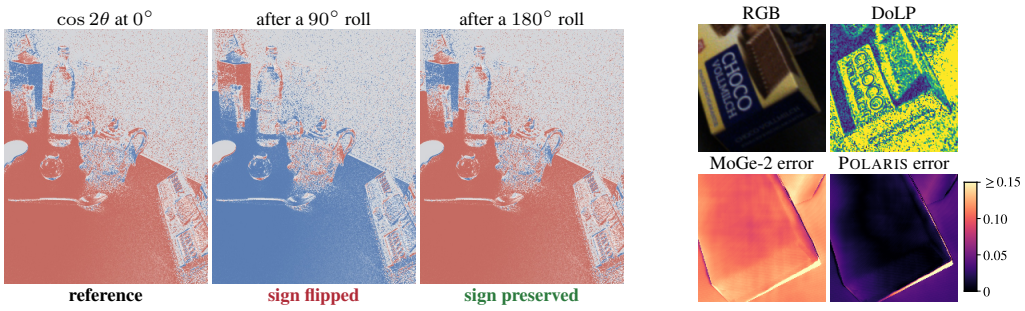
The transformation law distinguishes polarization from the other input modalities. The analyser axis is fixed to the sensor, so an image-plane rotation R_α turns the pixel lattice and shifts every measured angle at once. Writing $f \circ R_\alpha$ for the spatially rotated field,

$$S_0 \mapsto S_0 \circ R_\alpha, \qquad \mathrm{DoLP} \mapsto \mathrm{DoLP} \circ R_\alpha, \qquad z \mapsto e^{-i2\alpha}\,(z \circ R_\alpha), \tag{2}$$



(a) Encoder error under a camera roll. **(b)** Error of MoGe-2 by DoLP bin.

Figure 4: Encoder error under rotation and intensity-only error by degree of polarization. (a) The error compares the encoding of a rotated input, rotated back, with that of the original on real HAMMER polarization. Hatched bars give the error of the resampling alone, which no encoder avoids at angles other than multiples of 90°. (b) The error of MoGe-2, fine-tuned without polarization, rises with DoLP.



(a) The order-2 channel $\cos 2\theta$ under a camera roll. **(b)** A strongly polarizing region.

Figure 5: The rotation law and a strongly polarizing region on one real frame. (a) A 90° roll flips the sign of the order-2 channel, and a 180° roll keeps it. Pixels whose DoLP is below 0.06 are grey. (b) The error maps of MoGe-2 and POLARIS share one scale, and the colour bar gives the AbsRel.

so S_0 and DoLP, like image intensity, are scalars of rotation order 0, while z has order 2 and a surface normal order 1. At $\alpha = 90^\circ$ the phase is -1, the sign flip visible in Figure 5a, and at 180° it is $+1$. The Fresnel equations tie DoLP to the zenith angle and align the AoLP with the projected surface azimuth φ, so z is a noisy observation of $e^{i2\varphi}$. This doubling makes the field spin-2. The link is ambiguous, because the azimuth is determined only up to π and diffuse and specular reflection place the AoLP 90° apart. Rather than inverting the Fresnel model, which is ill-posed on real scenes, we build the order-2 rotation law into the architecture and let the network learn from the raw field.

3.2 COST OF IGNORING THE SYMMETRY

A convolutional encoder has no structure tying the channel transform in (2) to the spatial one, and Figure 4 compares the encoding of a rotated input, rotated back, with that of the original. At $k\cdot 90^\circ$ no pixel is resampled and the comparison is exact. A conventional encoder (Appendix B) is nonetheless 35–87% inconsistent there, while ours is exact up to floating-point error. The 180° column separates the two effects that 90° conflates. There the phase in (2) is $+1$, the order-2 channels move without the sign flip, and the conventional encoder's error drops from 87% to 35% accordingly. Section 5.1 measures depth under an analyser-axis rotation.

A second measurement identifies the regions in which the cue is informative. Binning the test pixels by DoLP and scoring MoGe-2 fine-tuned on our data without polarization on each bin shows the metric error rising from 0.0413 where almost no light is polarized to 0.0629 where much of it is (Fig. 4b). An intensity-only model is therefore least accurate on the surfaces that polarize the light they return, and Figure 5b shows one such surface.

3.3 PRIOR-PRESERVING INJECTION

The DINOv2-L (Oquab et al., 2024) encoder of MoGe-2 (Wang et al., 2025c) is frozen (Fig. 3). A polarimetric encoder E maps $\mathbf{p}$ to one token $E(\mathbf{p})_j \in \mathbb{R}^{256}$ per backbone patch j. After four of the 24 transformer blocks, every patch token of the block output $\mathbf{x} \in \mathbb{R}^{N\times D}$ receives the polarimetric token at the same position,

$$\mathbf{x}_j \leftarrow \mathbf{x}_j + \mathbf{W}\Big(E(\mathbf{p})_j - \tfrac{1}{|J|}\textstyle\sum_{k\in J} E(\mathbf{p})_k\Big), \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{D\times 256}$ carries no bias and J is the set of patch positions. The class token is left unchanged. Two properties follow. *(a) Near-identity at initialization.* $\mathbf{W}$ starts from small random values, so training starts close to the pretrained solution. *(b) Exact identity without polarization.* If $\mathbf{p} \equiv \mathbf{0}$, then $E(\mathbf{0}) = \mathbf{0}$ and the update is zero for *any* $\mathbf{W}$. By property (b), every polarization ablation is performed *within* one checkpoint, since disabling the input yields exactly the RGB-only network. This RGB-only network uses the neck and heads fine-tuned together with the cue rather than those of the released MoGe-2. Neither a fully fine-tuned model nor a concatenation-based fusion admits such a comparison.

3.4 SPIN-2 EQUIVARIANT ENCODER

The encoder lifts only the orientation field z, of order 2, and processes it with circular-harmonic convolutions (Worrall et al., 2017). A filter of angular order m factorizes into a radial profile κ and a phase, $K_m(\Delta) = \kappa(\|\Delta\|)\, e^{im\angle\Delta}$, and cross-correlating an order-n map $f^{(n)}$ with it gives

$$\big(f^{(n)} \star K_m\big)[\mathbf{u}] = \sum_{\Delta} f^{(n)}[\mathbf{u}+\Delta]\; \kappa(\|\Delta\|)\; e^{im\angle\Delta}, \tag{4}$$

which is equivariant with a predictable phase. Rotating the input by α multiplies the output by $e^{-i(n+m)\alpha}$ and rotates it spatially, so order n maps to order $n{+}m$. We restrict feature orders to $\{0,2\}$ and filter orders to $m \in \{-2,0,+2\}$. The $m=-2$ filter takes the order-2 polarization field to order 0, to quantities that the rotation phase no longer affects. Nonlinearity acts on magnitudes only, $\sigma(z) = \mathrm{ReLU}(|z|+b)\, z/|z|$, which leaves the order unchanged, and the output takes order-0 magnitudes $|f^{(0)}|$, which are exactly rotation-invariant. The scalars S_0 and DoLP scale these magnitudes, through a real-valued gain $1+\tanh\big(g(S_0,\mathrm{DoLP})\big)$ computed by a pointwise MLP g. The gain has no phase and therefore preserves the invariance. A zero orientation field yields zero tokens for any gain. The tokens passed to the transformer are therefore real and invariant, and the transformer needs no equivariant machinery.

Proposition 1 (Exact C_4 equivariance)**.** *Suppose (i) the radial profile κ is* ring-tied*, that is, a single weight is shared by all filter taps at a given radius, and (ii) the centre tap vanishes for $m \neq 0$. Then the patch averaging and every harmonic-convolution layer commute with the group C_4 of 90° rotations of the pixel grid, and the order-0 magnitude output is C_4-invariant (Appendix A).*

With both conditions enforced, the error at $k\cdot 90^\circ$ is 0.012% (Fig. 4) against 35–87% for the conventional CNN. At other angles the pixel grid breaks the equivariance and the error rises, but from a much lower base and within a small factor of the error that the resampling alone causes.

3.5 TRAINING

The DINOv2-L encoder is frozen throughout, and the MoGe neck, its heads and the polarimetric branch are optimized with AdamW at 5×10^{-6} under a cosine schedule, batch size 8, for 16k steps, on the backbone's affine-invariant point and depth objective plus a small metric-scale term. The polarimetric input is zeroed with probability 0.3 during training, so by property (b) a single network operates both with and without the cue. All computation is in fp32. The polarimetric tokens are added at four evenly spaced ViT blocks, and the harmonic encoder has two layers of orders $\{0,2\}$, kernel 5 and hidden width 32 (cost in Appendix I, design variants in Appendix D).

4 EXPERIMENTS

Setup. Training uses HAMMER (Jung et al., 2023) scenes 2–11 (11,593 frames) together with two further real polarimetric sets, the outdoor driving captures of GM Polarimetric (Baltaxe et al., 2023)

Table 1: Metric depth on HAMMER (test split, 2,828 frames, calibrated camera). A dagger (†) marks released weights, and all other models are fine-tuned on our training data. Metric AbsRel, RMSE and δ_1 use no alignment, *s-* and *st-AbsRel* use the optimal per-frame scale and scale-and-shift, and *B-F1* and *e-AbsRel* measure depth boundaries (Section 4). The **best** and second-best results are marked per column.

	Metric depth			Aligned (shape)		Boundary	
Method	**AbsRel** ↓	**RMSE (mm)** ↓	δ_1 ↑	**s-AbsRel** ↓	**st-AbsRel** ↓	**B-F1** ↑	**e-AbsRel** ↓
Marigold † CVPR24	0.5626	394.6	0.0326	0.2198	0.0800	0.0530	0.4862
ZoeDepth arXiv23	0.1384	129.9	0.8556	0.0891	0.0717	0.0080	0.1444
VGGT CVPR25	0.1378	108.6	0.7921	0.0401	0.0335	0.0052	0.1405
Depth Pro † ICLR25	0.1168	83.6	0.8928	0.0512	0.0385	0.0987	0.1174
Metric3D v2 TPAMI24	0.0749	67.6	0.9806	0.0407	0.0255	0.0347	0.0870
MoGe-2 NeurIPS25	0.0471	38.1	**0.9981**	0.0260	0.0158	**0.1065**	0.0427
POLARIS (ours)	**0.0396**	**36.2**	0.9980	**0.0209**	**0.0148**	0.0966	**0.0417**

Table 2: Metric depth on GM and RPS (test splits, calibrated camera). A dagger (†) marks released weights, and the other models are fine-tuned on our training data. The RPS depth scale is fitted to MoGe-2 †, which favours MoGe-2-based models in AbsRel and RMSE, and the GM polarizer images are formed from its AoLP and DoLP maps. Marigold † predicts affine-invariant depth and is left out of the **best** and second-best marks.

	GM (1,361 **frames**)			**RPS** (50 **frames**)		
Method	**AbsRel** ↓	**RMSE (mm)** ↓	**s-AbsRel** ↓	**AbsRel** ↓	**RMSE (mm)** ↓	**s-AbsRel** ↓
Marigold † CVPR24	0.9810	21165	0.2160	0.8419	1910	0.1681
ZoeDepth arXiv23	0.1088	5940	0.1113	0.0960	302	0.0769
VGGT CVPR25	0.1104	5242	0.1022	0.2747	713	0.1964
Depth Pro † ICLR25	0.3799	10640	0.1298	0.1429	380	0.0799
Metric3D v2 TPAMI24	0.0920	4714	0.0607	0.0815	236	0.0458
MoGe-2 NeurIPS25	**0.0617**	4433	0.0637	0.0617	**200**	**0.0385**
POLARIS (ours)	0.0690	**4260**	**0.0591**	**0.0602**	205	0.0404

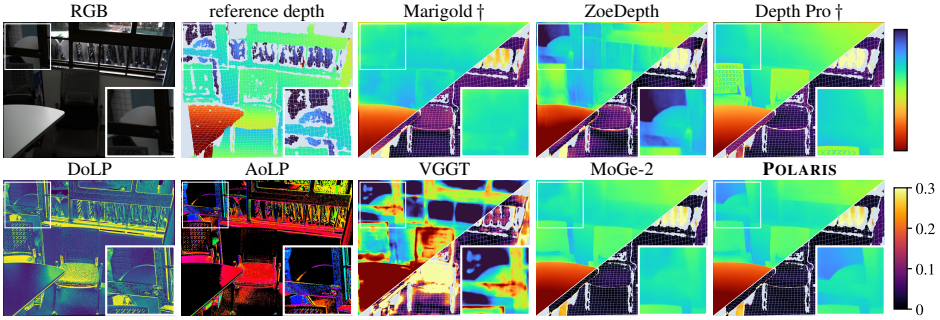


Figure 6: Metric depth on an RPS frame. The reference depth is sparse. In each prediction the upper-left triangle shows the depth without alignment and the lower-right triangle its metric error. The depth shares one colour scale, and Marigold †, which predicts affine-invariant depth, is aligned to the reference by scale and shift.

and the indoor and outdoor scenes of RPS (Tian et al., 2023). Evaluation is on the HAMMER test split, scenes 12–14 (2,828 frames), the official protocol also used by PPFT (Ikemura et al., 2024), and the test splits of GM (1,361 frames) and RPS (50 frames) are scored as well. Captures of our own stereo polarization rig (Appendix H, Table 8), whose reference depth is estimated from the stereo pair, are used for comparison outside HAMMER. The camera is calibrated throughout. MoGe-2 is fine-tuned on our training data with the schedule of Section 3.5, and ZoeDepth, VGGT and Metric3D v2 on the same data with a light recipe (Appendix B). Depth Pro, which releases no training code, and Marigold, which predicts affine-invariant depth, keep their released weights, marked by a dagger (†).

Evaluation metrics. Metric AbsRel and RMSE are computed without any per-image alignment, and RMSE is reported in millimetres. δ_1 is the fraction of pixels whose prediction and reference differ by less than a factor of 1.25. Two aligned variants separate shape from the global factor. The first, *s-AbsRel*, applies the least-squares optimal scale, a single free parameter, and the second, *st-AbsRel*, applies the optimal scale and shift, two free parameters, so st-AbsRel rewards neither a better metric factor nor a constant offset. Boundary F1 (*B-F1*) follows the scale-invariant occluding-contour criterion of Depth Pro (Bochkovskii et al., 2025) (Appendix B). Edge AbsRel (*e-AbsRel*) is the AbsRel restricted to a three-pixel band around the strongest fifth of the reference depth gradients, where thin structure lies.

Table 3: Surface normals on rendered and real polarimetric data. Errors are angles in degrees, and the **best** and second-best results are marked within each backbone group. Backbone rows are fine-tuned on SfPUEL without the branch, and no variant is trained on NeRSP.

	SfPUEL test (2,996 frames)			NeRSP (258 frames)		
Method	**Mean** ↓	**Median** ↓	11.25° ↑	**Mean** ↓	**Median** ↓	11.25° ↑
MoGe-2						
MoGe-2 NeurIPS25	11.026	7.371	67.36	14.907	11.683	48.42
+ conventional CNN encoder	9.936	6.456	71.62	14.524	11.259	49.97
+ POLARIS	**9.706**	**6.198**	**72.76**	**13.935**	**10.770**	**51.65**
DSINE						
DSINE CVPR24	16.846	12.016	46.95	26.136	22.016	20.66
+ conventional CNN encoder	11.617	7.550	66.13	17.051	**13.699**	38.15
+ POLARIS	**11.308**	**7.207**	**67.68**	**16.809**	13.768	**38.66**

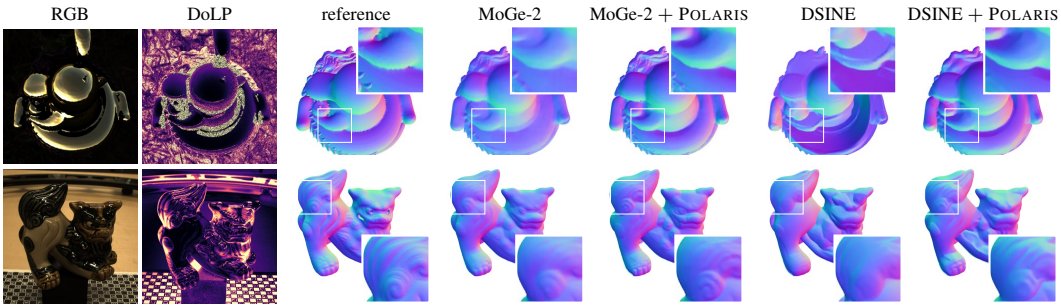


Figure 7: Qualitative surface normals. The top row shows an SfPUEL test frame and the bottom row a NeRSP view, unseen in training and cropped to the object. Colour encodes the normal with $+z$ toward the viewer, and RGB shows the unpolarized intensity S_0.

4.1 METRIC DEPTH

Table 1 compares POLARIS with six recent monocular depth models. First, against MoGe-2 fine-tuned on the same data without the branch, polarization lowers metric AbsRel from 0.0471 to 0.0396, a 15.9% reduction. Second, zeroing the polarimetric input of that checkpoint raises AbsRel to 0.0445, which is equivalent to deleting the branch, so the comparison is made within one set of weights. Third, the gain is not confined to the global scale, since st-AbsRel, measured after scale-and-shift alignment, also improves. Boundary F1, however, stays below that of MoGe-2. Figure 2 shows one unseen scene. On one colour scale with the ground truth, the baseline depth map is shifted towards larger depths, and its back-projected cloud, including the boxed bottle and teapot, lies behind the ground truth, the signature of a metric-scale error. POLARIS brings it close to the ground truth. Appendix E splits the gain by material, and Appendix F shows all models on another frame.

Table 2 scores the models on GM and RPS. On GM, POLARIS has the lowest RMSE and s-AbsRel and MoGe-2 the lowest AbsRel, so their difference is one of scale. On RPS, whose depth scale was fitted to MoGe-2 †, POLARIS has the lowest AbsRel and MoGe-2 the lowest RMSE and s-AbsRel.

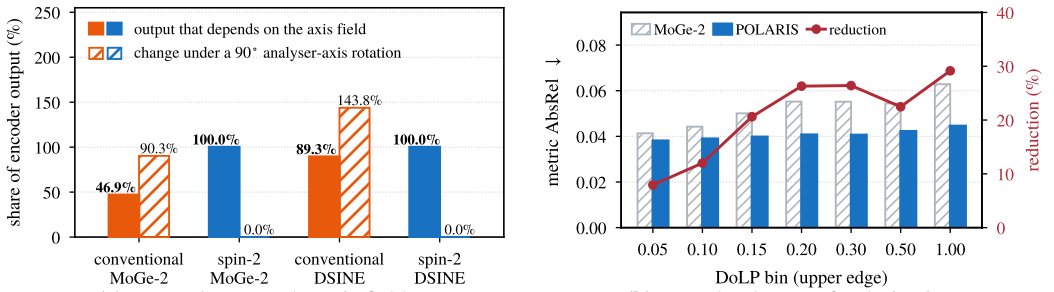
Figures 6 and 10 compare the baselines on two real frames outside HAMMER. The first comes from RPS, whose sparse reference is converted from disparity with a constant that is not published and that we fit to MoGe-2 †, so its metric error is indicative only. The second comes from our own stereo polarization rig, a sensor on which none of the models was trained, and its reference depth is estimated from the stereo pair by FoundationStereo (Wen et al., 2025), so it is itself an estimate. Over the shown part of our frame, the AbsRel of POLARIS is 0.135 against 0.182 for MoGe-2.

4.2 SURFACE NORMALS

Table 3 evaluates the encoder on normals against rendered and mesh-derived ground truth, since differentiating a sensor depth map amplifies exactly the specular and transparent errors under study. Each backbone row is fine-tuned on SfPUEL (Lyu et al., 2024) without the branch, so every row of a group sees the same data. On SfPUEL the adapter improves both backbones and the equivariant encoder is the better of the two, by $0.23°$ on MoGe-2 and $0.31°$ on DSINE. On NeRSP (Han et al.,

Table 4: Metric depth under a rotation of the analyser axis. Values are AbsRel ↓ on a subset of 120 HAMMER test frames, and *spread* is the range across rotations relative to each model's best. *Zeroed* removes the polarization, and *noise* randomizes the AoLP per pixel.

	Rotation of the analyser axis					Perturbed polarization	
Method	30°	60°	90°	**random°**	**spread ↓**	**zeroed**	**noise**
Conventional CNN encoder	0.04267	0.04384	0.04373	0.04311	2.73%	0.04763	0.04268
Δ encoder output	46.38%	79.48%	90.47%	83.32%	–	100.0%	47.3%
POLARIS (spin-2)	**0.03760**	**0.03760**	**0.03760**	**0.03760**	**0.00%**	0.04609	0.04594
Δ encoder output	0.00%	0.00%	0.00%	0.00%	–	100.0%	81.6%



(a) Dependence on the axis field. **(b)** Error by degree of polarization.

Figure 8: Analysis of the polarimetric branch. (a) On MoGe-2 and DSINE (rotation-invariant channels), solid bars give the relative change of each encoder's output when the AoLP is zeroed, and hatched bars its change when the analyser axis turns by 90°. (b) MoGe-2 (grey, the model of Fig. 4b) and POLARIS (blue) are scored on HAMMER test pixels binned by DoLP, and the line gives the reduction by POLARIS.

2024) the adapter improves both fine-tuned backbones as well, and either encoder lowers the DSINE error by more than 9°. Figure 7 shows one frame of each set.

4.3 ROLE OF THE POLARIMETRIC INPUT

An intensity image of a screen carries the depth cues of the scene it shows, so a model that reads intensity alone can take the screen for that scene. The polarimetric input separates the two. Figure 9 shows a monitor playing a landscape video in a dark room, captured by one camera of our rig, with the screen covering 30% of the frame. In the polarimetric input the screen appears as one surface. Its light is almost fully polarized, with a median DoLP of 0.91 against 0.03 for the room, and its AoLP takes one direction over the whole screen, 92.4° on average, whereas the intensity image shows the scenery. The crops shrink the margin around the screen from 80% to 0% and so remove the intensity cues to the plane, namely the corner of the monitor and the wall. MoGe-2, which reads intensity alone, predicts a nearly planar screen only while these cues are in view, and once the margin falls to 20% or less it predicts the depth of the scenery, with the railing near and the sky and the trees far. POLARIS predicts a plane at every margin, and with its polarimetric input zeroed the same network behaves like MoGe-2 (Fig. 9, Appendix G). Without the bezel, only the polarimetric input distinguishes the screen from the scene it shows.

5 DISCUSSION

5.1 ANALYSER-AXIS INVARIANCE

Table 4 measures the predicted depth when only the analyser axis rotates. This rotation is a camera roll with its pixel motion undone (Appendix C). It leaves the RGB input bit-identical and resamples nothing, so it isolates the polarimetric path exactly. Under this rotation the error of POLARIS does not change, while the error of the conventional encoder varies by 2.73%. The perturbed columns show whether each encoder uses the content of the polarization. Noise raises the error of POLARIS almost to its zeroed level, whereas the error of the conventional encoder rises only when the polarization is zeroed. This encoder therefore responds to the presence of polarization rather than to its content. The invariance discards only the absolute phase of the axis field. On both backbones the invariant output of the equivariant encoder depends entirely on the axis field, while the output of the conventional encoder changes by 90.3% on MoGe-2 and 143.8% on DSINE under a 90° axis rotation (Fig. 8a).

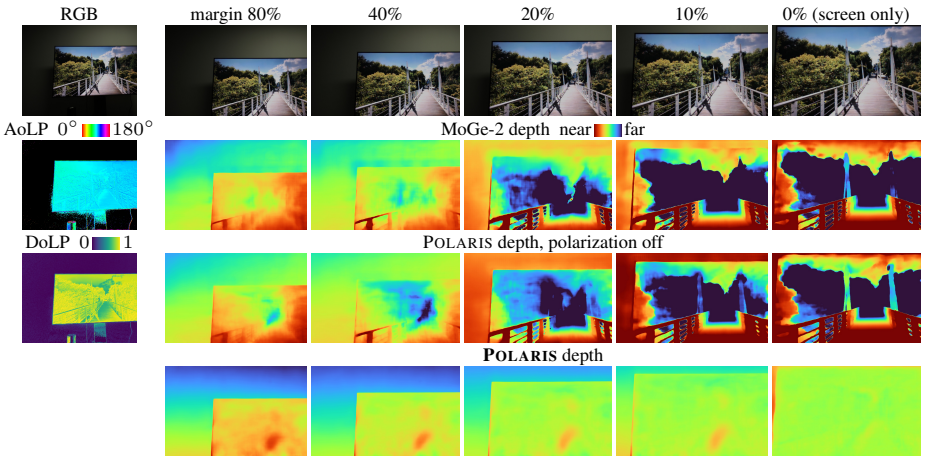


Figure 9: The polarimetric input on a screen that shows a scene. The first column shows the full frame, and the others crop it around the screen with a margin from 80% to 0%. Each depth map is coloured over its own median depth ±25%, so a fronto-parallel plane shows one colour. The AoLP is black where DoLP < 0.1.

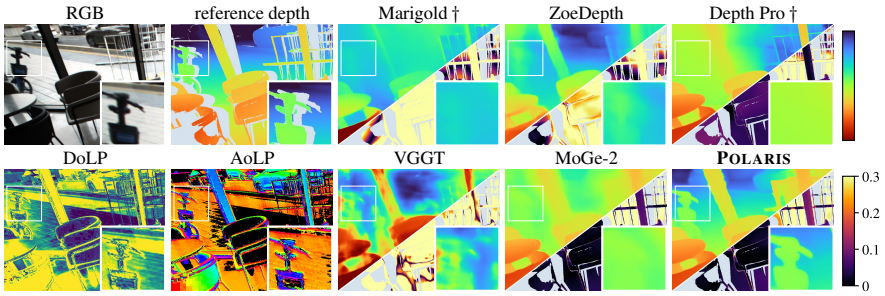


Figure 10: Metric depth on our own capture. The reference depth is estimated by stereo matching. In each prediction the upper-left triangle shows the depth without alignment and the lower-right triangle its metric error. The depth shares one colour scale, and Marigold † is aligned to the reference by scale and shift.

Against MoGe-2, the reduction that POLARIS brings rises from 8% in the least polarized bin to 26% at DoLP 0.15–0.3 and 29% in the most polarized bin (Fig. 8b).

6 LIMITATIONS

Our quantitative evidence is largely indoor and in-domain. Polarimetric data are scarce, so the model cannot be trained on large curated collections as RGB depth models are, and its outdoor depth loses sharpness. The disparity of RPS lacks a published metric constant, and the reference depth of our own captures is estimated by stereo matching rather than measured, so these two sets allow only a limited metric comparison. All numbers assume a calibrated camera, and the blind regime without calibration remains open. The magnitude read-out discards the global sign of the axis field, the sign that separates diffuse from specular reflection. Finally, POLARIS requires four-angle polarization at inference, and without it the model falls back to its RGB-only pass.

7 CONCLUSION

We presented POLARIS, which injects polarization into a frozen metric-depth backbone through a spin-2 rotation-equivariant encoder and a prior-preserving, position-aligned interface. The interface turns every polarization comparison into an ablation within one checkpoint, and the magnitude read-out makes the model insensitive to the manufacturer's choice of analyser axis. On HAMMER the polarimetric input lowers the metric error of an intensity-only model fine-tuned on the same data, most on strongly polarized surfaces, and it improves surface normals on two backbones. Noise in place of the angle removes the depth gain, so the encoder uses the content of the cue. Larger polarimetric datasets are the natural next step toward outdoor scenes.

AI USE STATEMENT

In this work, we used generative AI tools for proposing and refining hypotheses, developing conceptual frameworks, designing or providing feedback on the research methodology or experiments, implementing methods, cleaning and reformatting datasets, interpreting results, and translating text originally written in the authors' native language into English. We have not used generative AI tools for generating synthetic datasets, formulating mathematical claims, providing key ideas for proofs, writing proofs, or qualitative and thematic data analysis. Additionally, we used generative AI tools for drafting parts of the manuscript, editing it for readability, formatting references, searching for information, and identifying, summarizing, and analyzing related literature.

We have reviewed all AI-assisted work. The authors made all final decisions regarding the hypotheses, conceptual frameworks, research methodology, and experimental design of this work. Two authors verified that the AI-generated code matches the method described in the paper, and two authors manually inspected randomly sampled examples from the cleaned and reformatted datasets. The authors checked every AI-assisted interpretation against the raw experimental results and compared all translated text against the original. The authors also checked and revised all claims and numbers in AI-drafted text, and two authors confirmed that every cited paper actually exists. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work adds polarization to metric depth and surface-normal estimation, using public datasets and captures of our own stereo polarization rig. Frames of GM Polarimetric (Baltaxe et al., 2023) and RPS (Tian et al., 2023) may show bystanders and vehicle licence plates, and GM Polarimetric is distributed with GPS/INS readings. We use these data as released, make no attempt to identify any person, and do not use the GPS/INS readings. No figure in this paper shows an identifiable person or a readable licence plate. No human subjects were recruited for this work. Because our evidence is largely indoor, the model has not been validated for safety-critical uses such as driving.

REPRODUCIBILITY STATEMENT

The architecture is specified in Section 3, the training recipe in Section 3.5, and the evaluation protocol, with the official HAMMER scene split, the calibrated-camera setting and the error decomposition, in Section 4. Code and configurations will be released.

REFERENCES

Gary A. Atkinson and Edwin R. Hancock. Recovery of surface orientation from diffuse polarization. *IEEE Transactions on Image Processing*, 15(6):1653–1664, 2006.

Yunhao Ba, Alex Gilbert, Franklin Wang, Jinfa Yang, Rui Chen, Yiqin Wang, Lei Yan, Boxin Shi, and Achuta Kadambi. Deep shape from polarization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 554–571, 2020.

Gwangbin Bae and Andrew J. Davison. Rethinking inductive biases for surface normal estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9535–9545, 2024.

Michael Baltaxe, Tomer Pe'er, and Dan Levi. Polarimetric imaging for perception. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2023.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.

Marc Blanchon, Désiré Sidibé, Olivier Morel, Ralph Seulin, Daniel Braun, and Fabrice Mériaudeau. P2D: a self-supervised method for depth estimation from polarimetry. In *Proceedings of the International Conference on Pattern Recognition (ICPR)*, pp. 7357–7364, 2021.

Aleksei Bochkovskii, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan R. Richter, and Vladlen Koltun. Depth Pro: sharp monocular metric depth in less than a second. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 75602–75637, 2025.

Taco S. Cohen and Max Welling. Group equivariant convolutional networks. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2990–2999, 2016.

Taco S. Cohen and Max Welling. Steerable CNNs. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

Valentin Deschaintre, Yiming Lin, and Abhijeet Ghosh. Deep polarization imaging for 3D shape and SVBRDF acquisition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15567–15576, 2021.

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 *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.

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 (NIPS)*, pp. 2366–2374, 2014.

Carlos Esteves, Ameesh Makadia, and Kostas Daniilidis. Spin-weighted spherical CNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8614–8625, 2020.

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

Yufei Han, Heng Guo, Koki Fukai, Hiroaki Santo, Boxin Shi, Fumio Okura, Zhanyu Ma, and Yunpeng Jia. NeRSP: neural 3D reconstruction for reflective objects with sparse polarized images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11821–11830, 2024.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 2790–2799, 2019.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2022.

Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3D v2: a versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46 (12):10579–10596, 2024.

Kei Ikemura, Yiming Huang, Felix Heide, Zhaoxiang Zhang, Qifeng Chen, and Chenyang Lei. Robust depth enhancement via polarization prompt fusion tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20710–20720, 2024.

Yujin Jeon, Eunsue Choi, Youngchan Kim, Yunseong Moon, Khalid Omer, Felix Heide, and Seung-Hwan Baek. Spectral and polarization vision: spectro-polarimetric real-world dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22098–22108, 2024.

HyunJun Jung, Patrick Ruhkamp, Guangyao Zhai, Nikolas Brasch, Yitong Li, Yannick Verdie, Jifei Song, Yiren Zhou, Anil Armagan, Slobodan Ilic, Aleš Leonardis, Nassir Navab, and Benjamin Busam. On the importance of accurate geometry data for dense 3D vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 780–791, 2023.

Achuta Kadambi, Vage Taamazyan, Boxin Shi, and Ramesh Raskar. Polarized 3D: high-quality depth sensing with polarization cues. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3370–3378, 2015.

Agastya Kalra, Vage Taamazyan, Supreeth Krishna Rao, Kartik Venkataraman, Ramesh Raskar, and Achuta Kadambi. Deep polarization cues for transparent object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8602–8611, 2020.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. Repurposing diffusion-based image generators for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9492–9502, 2024.

Chenyang Lei, Chenyang Qi, Jiaxin Xie, Na Fan, Vladlen Koltun, and Qifeng Chen. Shape from polarization for complex scenes in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12632–12641, 2022.

Chenhao Li, Taishi Ono, Takeshi Uemori, and Yusuke Moriuchi. Revisiting shape from polarization in the era of vision foundation models. *arXiv preprint arXiv:2603.04817*, 2026.

Youwei Lyu, Heng Guo, Kailong Zhang, Si Li, and Boxin Shi. SfPUEL: shape from polarization under unknown environment light. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 97184–97202, 2024.

Daisuke Miyazaki, Masataka Kagesawa, and Katsushi Ikeuchi. Transparent surface modeling from a pair of polarization images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(1):73–82, 2004.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. UniDepth: universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10106–10116, 2024.

Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. UniDepthV2: universal monocular metric depth estimation made simpler. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(3):2354–2367, 2026.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1623–1637, 2022.

Shreeyak S. Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, and Shuran Song. ClearGrasp: 3D shape estimation of transparent objects for manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, pp. 3634–3642, 2020.

Mingqi Shao, Chongkun Xia, Zhendong Yang, Junnan Huang, and Xueqian Wang. Transparent shape from a single view polarization image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9277–9286, 2023.

William A. P. Smith, Ravi Ramamoorthi, and Silvia Tozza. Height-from-polarisation with unknown lighting or albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(12):2875–2888, 2019.

Chaoran Tian, Weihong Pan, Zimo Wang, Mao Mao, Guofeng Zhang, Hujun Bao, Ping Tan, and Zhaopeng Cui. DPS-Net: deep polarimetric stereo depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3569–3579, 2023.

Yannick Verdié, Jifei Song, Barnabé Mas, Benjamin Busam, Aleš Leonardis, and Steven McDonagh. CroMo: cross-modal learning for monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3937–3947, 2022.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: visual geometry grounded transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306, 2025a.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5261–5271, 2025b.

Ruicheng Wang, Sicheng Xu, Yue Dong, Yu Deng, Jianfeng Xiang, Zelong Lv, Guangzhong Sun, Xin Tong, and Jiaolong Yang. MoGe-2: accurate monocular geometry with metric scale and sharp details. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 35928–35959, 2025c.

Maurice Weiler and Gabriele Cesa. General E(2)-equivariant steerable CNNs. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 14357–14368, 2019.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. FoundationStereo: zero-shot stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260, 2025.

Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Harmonic networks: deep translation and rotation equivariance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5028–5037, 2017.

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

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 21875–21911, 2024b.

Luwei Yang, Feitong Tan, Ao Li, Zhaopeng Cui, Yasutaka Furukawa, and Ping Tan. Polarimetric dense monocular SLAM. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3857–3866, 2018.

Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. Metric3D: towards zero-shot metric 3D prediction from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9043–9053, 2023.

Rongjia Yu, Tong Jia, Hao Wang, Xiaofang Li, Xiao Yang, Zinuo Zhang, and Cuiwei Liu. CDPR: cross-modal diffusion with polarization for reliable monocular depth estimation. *IEEE Transactions on Multimedia*, 2026.

Pierluigi Zama Ramirez, Alex Costanzino, Fabio Tosi, Matteo Poggi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Booster: a benchmark for depth from images of specular and transparent surfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(1):85–102, 2024.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3836–3847, 2023.

Dizhong Zhu and William A. P. Smith. Depth from a polarisation + RGB stereo pair. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7586–7595, 2019.

A Proof of Proposition 1

Let R_{90} denote the 90° rotation of the pixel grid and, following Eq. (2), let $Tf = e^{-in\pi/2}(f \circ R_{90})$ act on an order-n map f. The rotation maps the integer lattice onto itself and preserves distances, so it permutes the taps on every ring $\|\Delta\| = r$. Substituting $\Delta' = R_{90}\Delta$ in Eq. (4), with $\angle\Delta = \angle\Delta' - \pi/2$ and, by condition (i), $\kappa(\|\Delta\|) = \kappa(\|\Delta'\|)$, gives

$$(Tf) \star K_m = e^{-i(n+m)\pi/2}\big((f \star K_m) \circ R_{90}\big),$$

so each layer commutes with T and maps order n to order $n+m$. The centre tap has no angle, and condition (ii) removes it from every filter with $m \neq 0$, while the centre tap of K_0 carries no phase. The nonlinearity satisfies $\sigma(e^{i\gamma}z) = e^{i\gamma}\sigma(z)$ and acts pixelwise, the gain is a real pixelwise function of the order-0 scalars S_0 and DoLP, and averaging over 14×14 patches commutes with R_{90} when the image size is a multiple of 14. The order-0 output therefore rotates with the image without a phase, and so does its magnitude. A rotation of the analyser axis by β multiplies z by $e^{-i2\beta}$ and moves no pixel. The filters are complex-linear and σ commutes with a constant phase, so every feature acquires the same phase and $|f^{(0)}|$ is unchanged for every β. The residual of 0.012% at $k \cdot 90^\circ$ in Figure 4a is floating-point error.

B Implementation details

Fine-tuning. MoGe-2 is fine-tuned with the schedule of Section 3.5 without the polarimetric branch. VGGT, ZoeDepth and Metric3D v2 are fine-tuned on the same training data for up to 3k steps, updating the last two backbone blocks and the head with a log-depth loss.

Conventional encoder. The conventional CNN encoder applies ordinary convolutions and a 1×1 read-out to $\mathbf{p}$ on a downsampled grid.

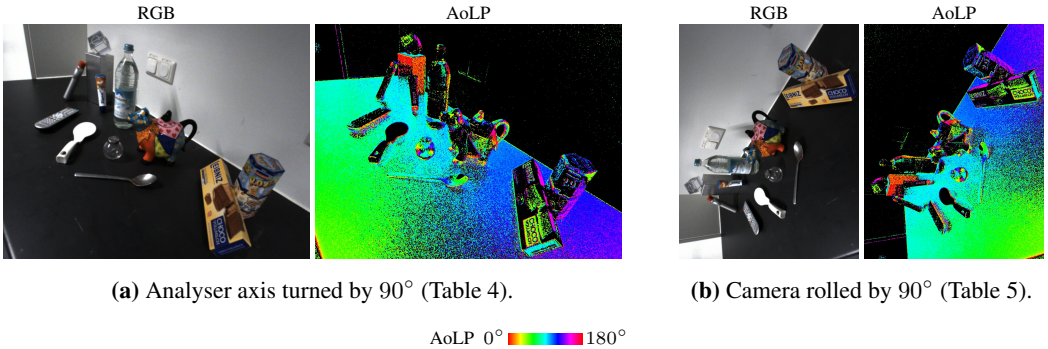
Surface normals. For MoGe-2 the branch is attached as for depth. For DSINE the encoder runs on a 4-pixel grid and also emits order-2 channels, and its output is added to four feature maps of the EfficientNet encoder through 1×1 convolutions that start at zero. Both backbones are fine-tuned on SfPUEL.

Boundary F1. Two neighbouring pixels form a contour when their depth ratio exceeds $1 + t/100$, and the F1 between the predicted and the reference contour sets is averaged over $t \in \{5, 10, 15, 20, 25\}$. The rule uses a ratio, so the score is invariant to the global scale.

GM polarizer images. GM Polarimetric releases AoLP and DoLP maps, from which we form the four polarizer images with the grey level of the RGB image as S_0.

C Depth under a camera roll

Table 5 reports metric depth under camera rolls by multiples of $90°$ on the frames of Table 4. A roll turns the RGB and the polarization images and shifts every measured angle (Fig. 11). At these angles no pixel is resampled, and the prediction is turned back before scoring. Each baseline runs through its own inference path with the weights of Table 1. The $0°$ column is the unrotated reference of Table 4. Turning a rolled input back upright restores the RGB image and keeps the angle shift, so the last column repeats the $90°$ column of Table 4. The roll raises the error of most models, and that of VGGT changes least. Polaris has the lowest error at $0°$, $90°$ and $270°$, whereas at $180°$ Metric3D v2, VGGT and the conventional encoder are more accurate.



(a) Analyser axis turned by $90°$ (Table 4). **(b)** Camera rolled by $90°$ (Table 5).

Figure 11: Inputs of the two rotation experiments on one HAMMER test frame. (a) Turning the analyser axis keeps the RGB image and shifts every AoLP by $90°$. (b) A camera roll turns the images of (a) by $90°$. The AoLP is black where DoLP < 0.1.

D Design variants

Table 6 lists design variants of Polaris on HAMMER. The conventional CNN rows use an ordinary convolutional encoder of about the same size, fed the full tuple $\mathbf{p}$ or the axis field alone. The harmonic rows feed normalized Stokes parameters in place of $\mathbf{p}$, or S_0 and DoLP as additional inputs in place of the gain. The rows marked $+$ add to Polaris a correction path in the decoder or an azimuth resolved with the predicted normals, and the indented row also synthesizes the polarization of objects from their ground-truth geometry during training.

Table 5: Metric depth under a camera roll (the 120 HAMMER test frames of Table 4, calibrated camera). Values are AbsRel ↓. The input is rolled by the given angle and the prediction is turned back before scoring. *Turned back* rolls the input by 90° and turns it back upright before the prediction. Marigold † predicts affine-invariant depth and is left out of the **best** and second-best marks.

	Camera roll				90°,
Method	0°	90°	180°	270°	**turned back**
Marigold † CVPR24	0.58287	0.57506	0.52236	0.61203	0.58287
ZoeDepth arXiv23	0.12337	0.30238	0.59879	0.31269	0.12337
VGGT CVPR25	0.14911	0.15612	0.14376	0.18165	0.14911
Depth Pro † ICLR25	0.14257	0.35469	0.69857	0.49585	0.14257
Metric3D v2 TPAMI24	0.06599	0.14709	**0.11018**	0.15318	0.06599
MoGe-2 NeurIPS25	0.04664	0.14116	0.22932	0.13178	0.04664
Conventional CNN encoder	0.04263	0.13419	0.20587	0.12859	0.04373
POLARIS (ours)	**0.03760**	**0.10298**	0.21557	**0.09525**	**0.03760**

Table 6: Design variants on HAMMER (test split, 2,828 frames, calibrated camera). Rows change the input, the encoder or the training of POLARIS, and the metrics follow Table 1. The table lists variants that do not lower the AbsRel of POLARIS, and the best result per column is in bold.

	Metric depth			Aligned (shape)		Boundary	
Variant	**AbsRel ↓**	**RMSE (mm) ↓**	δ_1 ↑	**s-AbsRel ↓**	**st-AbsRel ↓**	**B-F1 ↑**	**e-AbsRel ↓**
MoGe-2, no branch	0.0471	38.1	**0.9981**	0.0260	0.0158	0.1065	0.0427
Conventional CNN encoder	0.0442	37.7	**0.9981**	0.0227	0.0149	0.1034	0.0421
Conventional CNN, axis field only	0.0451	38.3	**0.9981**	0.0228	0.0150	**0.1066**	0.0430
Harmonic, normalized Stokes input	0.0462	39.5	**0.9981**	0.0213	0.0150	0.1029	0.0456
Harmonic, S_0 and DoLP as inputs	0.0429	37.6	0.9974	0.0251	0.0175	0.1033	0.0442
POLARIS + decoder correction	0.0427	37.9	0.9979	0.0213	0.0150	0.0934	0.0442
POLARIS + azimuth from normals	0.0411	36.8	0.9979	0.0214	0.0151	0.0982	0.0427
+ object polarization from geometry	0.0439	38.3	0.9980	0.0215	**0.0148**	0.1022	0.0450
POLARIS	**0.0396**	**36.2**	0.9980	**0.0209**	**0.0148**	0.0966	**0.0417**

E METRIC DEPTH BY MATERIAL

Figure 12 shows how the instance labels of HAMMER are grouped, and Table 7 splits the metric error of Table 1 by group.

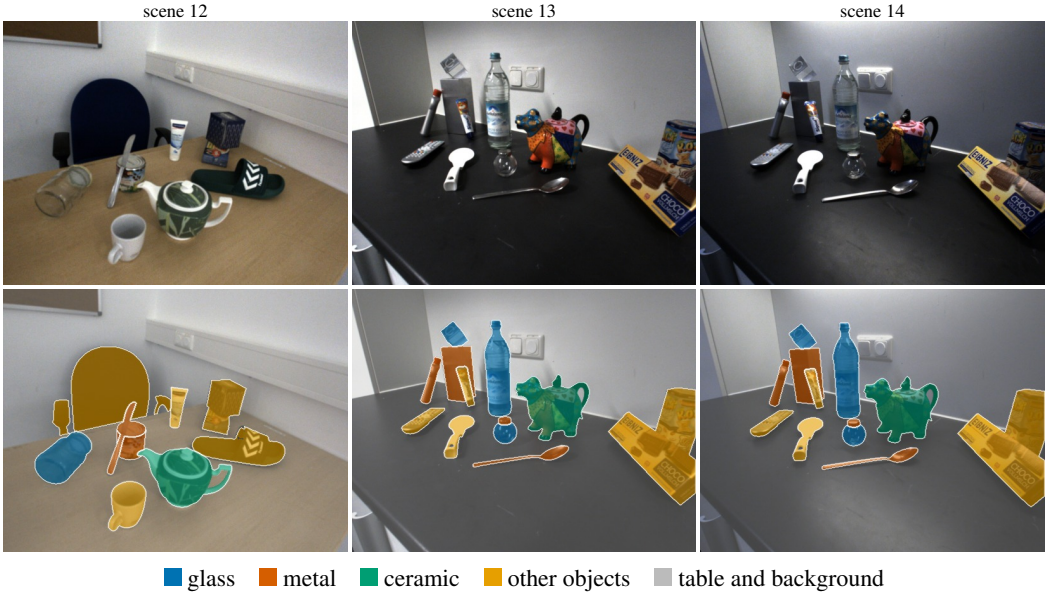


Figure 12: Material groups on HAMMER. The instance labels of HAMMER are grouped into five materials, shown on one test frame of each scene.

Table 7: Metric depth by material on HAMMER (test split, 2,828 frames, calibrated camera). Values are AbsRel ↓ without alignment, averaged over the frames in which a material covers at least 200 pixels. The polarization-off column zeroes the polarimetric input of the same POLARIS weights.

Material	Frames	Pixels	MoGe-2	Conv. CNN	POLARIS, pol. off	POLARIS
Glass	1,414	1.6%	0.0567	0.0570	0.0569	**0.0513**
Metal	1,414	1.6%	0.0358	**0.0329**	0.0394	0.0342
Ceramic	1,414	1.5%	0.0528	0.0497	0.0533	**0.0409**
Other objects	1,882	7.1%	0.0511	0.0507	0.0468	**0.0417**
Table and background	2,828	88.2%	0.0460	0.0439	0.0436	**0.0387**

F QUALITATIVE COMPARISON ON HAMMER

Figure 13 compares the depth and the metric error of the models of Table 1 on one HAMMER test frame.

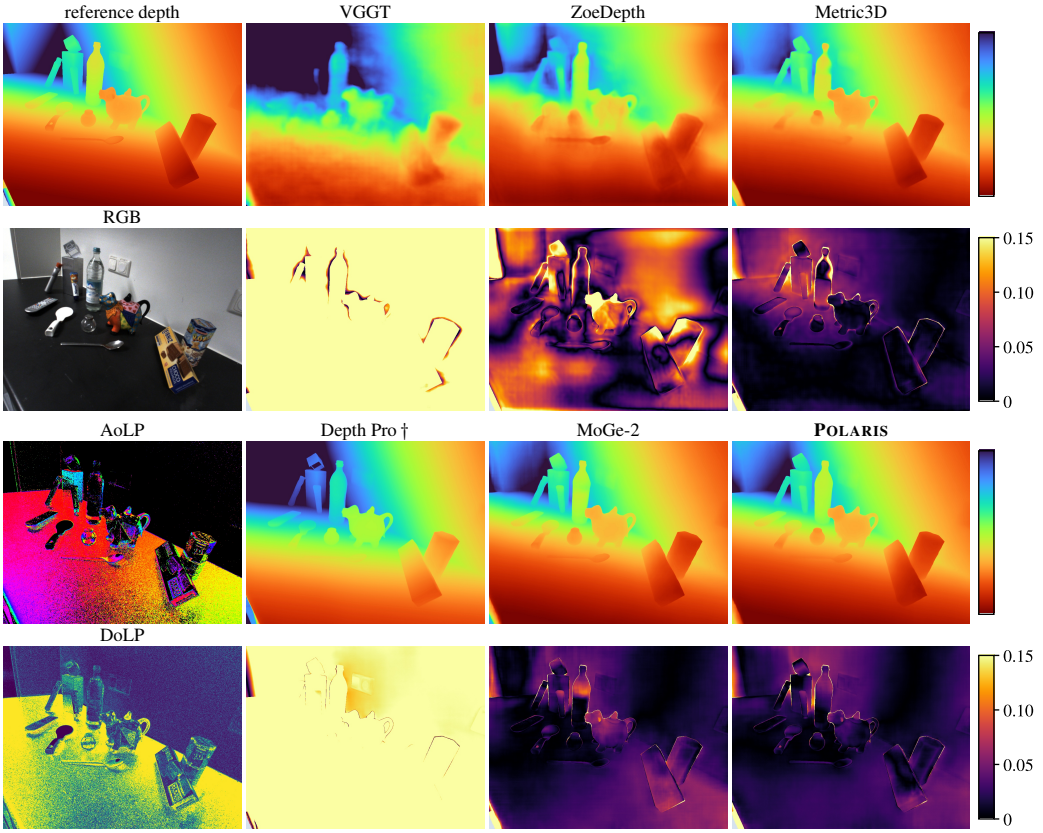


Figure 13: Metric depth on a HAMMER test frame. The first column shows the input and the reference depth. For each model, the upper panel is its depth, on one colour scale shared with the reference, and the lower panel its metric error $|d - d^*|/d^*$ without alignment.

G FURTHER SCREENS

Figure 14 repeats Figure 9 on four further captures of the monitor.

H OUR STEREO POLARIZATION RIG

Hardware. Two LUCID colour polarization cameras with Sony IMX250 polarization sensors stand side by side with a baseline of 220.15 mm. Each sensor places linear polarizers at 0°, 45°, 90° and

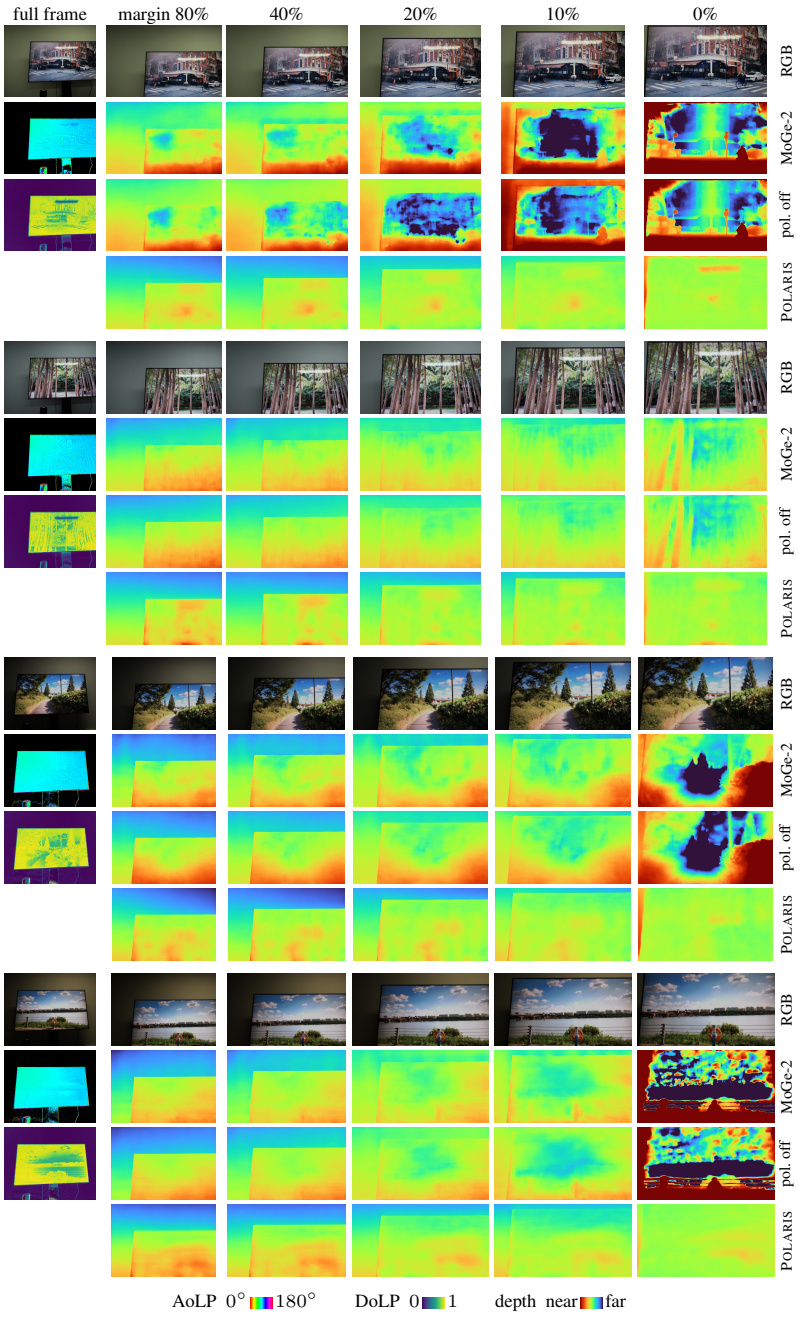


Figure 14: The polarimetric input on four further screens. In each block the first column shows the full frame, its AoLP and its DoLP, and the rows show the crops and the depth of MoGe-2, of POLARIS with its polarimetric input zeroed (pol. off) and of POLARIS. Margins and colours follow Figure 9.

$135°$ over every 2×2 block of pixels, so each frame gives four 1224×1024 angle images of the 2448×2048 sensor, stored as 8-bit JPEG. The lenses have a focal length of 4 mm.

Rectification and input. Each camera is rectified with maps refitted to the raw and rectified image pairs of the capture software under the 12-parameter distortion model of OpenCV, with a reprojection error of 0.4 px. After rectification both cameras share a focal length of 815.4 px, a principal point of $(774.7, 526.5)$ and a horizontal field of view of $73.8°$. A remaining vertical offset of about 1.2 px is removed in every frame by the median vertical offset of SIFT matches. The RGB input is the mean of the four angle images, and the polarimetric input follows Eq. (1), as for HAMMER.

Reference depth. FoundationStereo (Wen et al., 2025) with its released ViT-L weights and 32 iterations predicts the disparity at full resolution, and the depth is the focal length times the baseline divided by the disparity. The depth of the right view is obtained by mirroring both images. A pixel is kept when the disparities of the two views agree within 1 px, its match falls inside the other image, it lies outside the rectification border and its depth lies between 0.3 and 40 m, which keeps about 85% of the pixels. The reference is itself an estimate. Through a window it gives the depth of the scene behind the glass, and the sky and the distant background fall outside its range.

Captures. We use 74 frames with both views. Six scenes, four indoor and two outdoor, give 66 frames chosen for sharpness and exposure without duplicates, and two further scenes, indoor and outdoor, give 8 frames at two times. The median depth of a scene lies between 1.8 and 9.7 m. Table 8 scores the models of Table 1 on these frames. The RMSE of Depth Pro † is dominated by the 0.07% of pixels that it places at its 10 km output limit.

Table 8: Metric depth on our own captures (74 frames, stereo reference depth, calibrated camera). Metrics follow Table 1. Marigold † predicts affine-invariant depth and is left out of the **best** and second-best marks.

	Metric depth			Aligned (shape)		Boundary	
Method	**AbsRel** ↓	**RMSE (mm)** ↓	δ_1 ↑	**s-AbsRel** ↓	**st-AbsRel** ↓	**B-F1** ↑	**e-AbsRel** ↓
Marigold †	0.9411	6873.9	0.0000	0.2980	0.2284	0.0394	0.9400
ZoeDepth	0.3224	3307.5	0.5023	0.2383	0.2026	0.0150	0.4382
VGGT	0.3548	3376.3	0.3036	0.3327	0.3001	0.0222	0.3794
Depth Pro †	0.1457	26660.2	0.8791	0.2076	0.3337	**0.1682**	0.2681
Metric3D v2	**0.0903**	2308.4	**0.9112**	0.1991	0.1696	0.0559	**0.1593**
MoGe-2	0.1495	2620.4	0.8015	0.1826	0.1346	0.0479	0.2223
POLARIS (ours)	0.1431	**2277.4**	0.8266	**0.1459**	**0.1035**	0.0362	0.2112

I COMPUTATIONAL COST

Table 9: Cost of the polarimetric branch. Latency (median of 30 runs) and peak memory are measured on one 1088×832 frame in fp32 on an RTX PRO 6000 Blackwell. *Added* counts parameters beyond MoGe-2.

Model	**Params (M)**	**Added (M)**	**Latency (ms)**	**Peak memory (GB)**
MoGe-2 † NeurIPS25	330.9	–	126	2.56
MoGe-2	330.9	0.00	126	2.56
POLARIS, conventional CNN	332.0	1.07	126	2.60
POLARIS	332.0	1.07	127	2.60

Table 9 gives the cost of the branch. It adds 1.07M parameters (21.5K in the harmonic encoder and 1.05M in the linear injection layers), about 1 ms and 0.04 GB. With the backbone frozen, fine-tuning updates 23.4M parameters, 7.0% of the model.