GEOPOLAR: FEW-STEP RGB-TO-POLARIZATION DIFFUSION WITH GEOMETRY PRIORS

Anonymous authors
Paper under double-blind review

ABSTRACT

Polarization reveals surface orientation and reflection properties that are useful for depth and shape estimation, but capturing it requires dedicated sensors, so the cue is absent from ordinary RGB images. Generative models have therefore been proposed to synthesize polarization from RGB. However, the angle of polarization depends not only on surface orientation but also on whether reflection is diffuse or specular, and the two cases yield angles 90° apart that appearance alone cannot distinguish. We resolve this diffuse–specular ambiguity by leveraging the geometric prior of monocular foundation models. From the same RGB image, such a model recovers the scene geometry, which we convert into the angular coordinates of polarization rather than supplying it as raw normals. The network then only needs to select between the two candidate angles instead of regressing the angle from scratch. Because the foundation-model prior carries the geometric burden of the task, the generator can remain lightweight: a four-step diffusion model operating directly in pixel space. GEOPOLAR is an order of magnitude lighter than latent-diffusion baselines and remains lighter with its geometry model, yet it surpasses them on benchmark data at native resolution and on most metrics of a set captured with an unseen camera. Further analyses show that the gain follows the geometric content of the prior and persists under moderate errors in the normals.

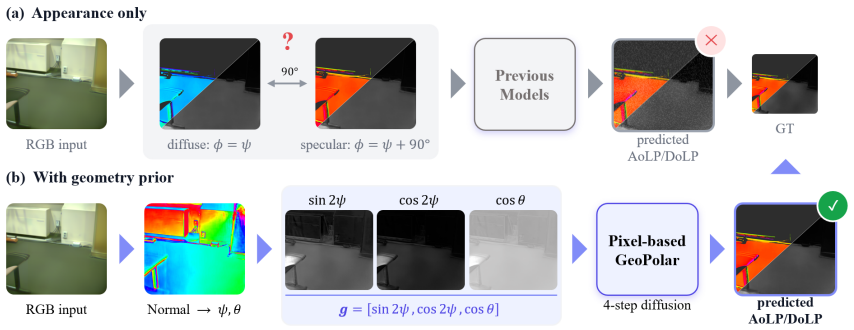


Figure 1: Why a geometry prior helps. (a) One RGB image is consistent with two AoLP fields 90° apart, one for diffuse and one for specular reflection, which appearance alone does not distinguish. (b) The normal of a monocular geometry model, encoded as $[\sin 2\psi, \cos 2\psi, \cos\theta]$, reduces the direction problem to the choice of a branch. Polarization panels show the AoLP above the diagonal and the DoLP below it.

1 INTRODUCTION

Polarization records how light interacted with the surface that reflected it. The angle and degree of linear polarization (AoLP and DoLP) reveal surface orientation and the reflection regime that intensity conceals, and they have served shape from polarization, depth and 3D reconstruction, reflection separation, and material and glass understanding (Miyazaki et al., 2003; Ba et al., 2020; Lei et al., 2022; Kadambi et al., 2015; Kalra et al., 2024; Cui et al., 2017; Zhu & Smith, 2019; Tian et al., 2023; Kong et al., 2012; Liang et al., 2022; Mei et al., 2022). The cue is rare, because capturing it requires a polarization sensor or several exposures, so polarimetric datasets remain small.

This scarcity motivates RGB-to-polarization estimation. A benchmark study established the task (Lin et al., 2025), and two latent-diffusion models now address it (Zhang et al., 2025; Luo

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.

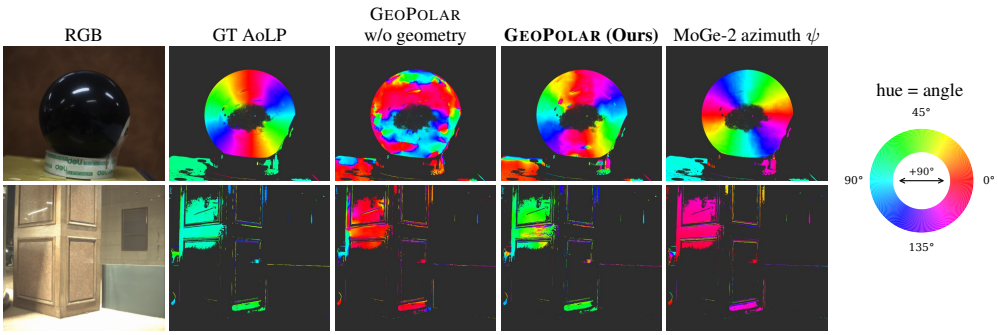


Figure 2: Appearance alone can select the wrong branch. Two specular scenes predicted by the same network trained without and with the geometry prior, next to the MoGe-2 azimuth ψ. Without the prior, the prediction breaks into patches or lies about 90° from the ground truth; with it, the prediction follows the ground truth (Section 5.2). Hue encodes the angle, and pixels with ground-truth DoLP below 0.1 are masked.

et al., 2026). All of them map appearance to polarization, although appearance alone does not determine polarization. For a dielectric, the angle of polarization follows the azimuth of the surface normal, and its degree follows the angle between that normal and the viewing ray (Atkinson & Hancock, 2006; Miyazaki et al., 2003). Diffuse and specular reflection place the angle 90° apart, so one surface orientation admits two directions, which we call its two branches, and a generator that sees only appearance must recover the geometry and choose a branch before it can commit to a direction.

Monocular foundation models now return geometry from a single image (Wang et al., 2025a;b; Yang et al., 2024), so it can be supplied to the generator instead of being recovered by it. The remaining design decision is its form. A raw normal map leaves the network to relate a Cartesian vector to a π-periodic image-plane angle, so we express the prior in the coordinates of polarization. From the normal and the viewing ray we compute the azimuth ψ and the zenith angle θ and pass $\mathbf{g} = [\sin 2\psi, \cos 2\psi, \cos\theta]$. In doubled angles the two branches differ only by a sign, so estimating the direction reduces to choosing that sign, and the third channel relates to the degree (Figure 1).

Because the prior carries the geometric part of the problem, the generator can be kept small. We use a residual-shifting diffusion model (Yue et al., 2025) in pixel space, whose chain starts from the RGB image so that four steps suffice, with a plain image transformer (Li & He, 2026) and a Stokes target supervised by an angular loss only where the angle is observable. The generator is an order of magnitude lighter than the latent baselines, yet it leads them on all three subsets of a multi-source benchmark and on most metrics of a set from a camera that none of them has seen. Against the same network without the prior, large angular errors on polarized pixels become much rarer, above all those near 90° that indicate the wrong branch (Section 5.2, Figure 2).

Contributions. **(i)** We identify the branch ambiguity as a principal failure of appearance-only RGB-to-polarization models and quantify it on real data (Section 5.2). **(ii)** We propose a geometry prior in polarization coordinates, $[\sin 2\psi, \cos 2\psi, \cos\theta]$ computed from monocular normals, which reduces direction estimation to a choice of branch (Section 3.2), and analyze its form and quality (Section 5). **(iii)** We build GeoPolar, a four-step pixel-space generator conditioned on the prior (Sections 3.3 and 3.4). **(iv)** We evaluate at native resolution on a multi-source benchmark and a new unseen-camera set (Section 4).

2 Related Work

Polarization imaging and RGB-to-polarization estimation. Polarization cues have served shape from polarization, depth estimation and 3D sensing, and reflection analysis (Miyazaki et al., 2003; Kadambi et al., 2015; Kalra et al., 2024; Ba et al., 2020; Lei et al., 2022; Kong et al., 2012), but their capture requires a polarization sensor or several exposures (Kurita et al., 2023; Hwang et al., 2025), so polarimetric datasets remain small (Jeon et al., 2024; Liang et al., 2022; Yao et al., 2025). Lin et al. (2025) established RGB-to-polarization estimation as a task. PolarAnything (Zhang et al., 2025) fine-tunes Stable Diffusion with a ControlNet-style encoder and predicts the doubled-angle AoLP with the DoLP, and GenPolar (Luo et al., 2026) predicts Stokes components, supervises the angle only where it is observable, and is distilled into a one-step generator. Both work in a pretrained

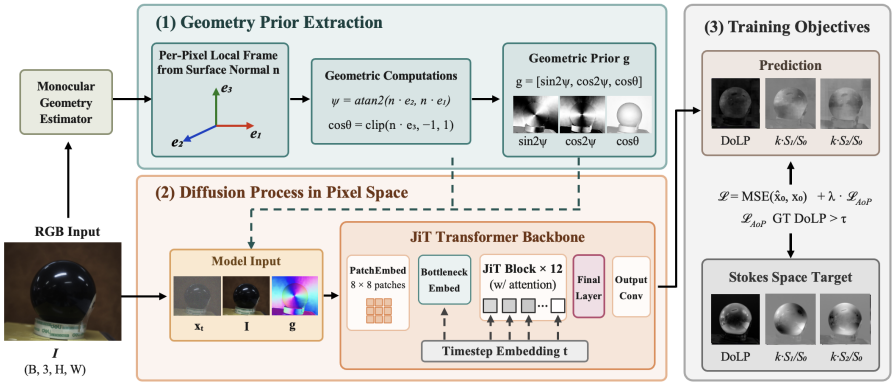


Figure 3: Overview of GeoPolar. (1) A frozen geometry model converts the RGB image into $\mathbf{g} = [\sin 2\psi, \cos 2\psi, \cos\theta]$. (2) $\mathbf{g}$ is concatenated with the image as the condition of a four-step residual-shifting chain in pixel space, with a plain image transformer as the denoiser. (3) The network predicts the Stokes target under a squared-error loss and an angular loss on polarized pixels.

latent space and condition on appearance alone. We keep the Stokes target and observability-aware supervision, generate in pixel space without distillation, and add a geometric condition.

Diffusion models for dense prediction. Conditional diffusion serves image-to-image translation (Saharia et al., 2022; Liu et al., 2023) and turns image priors into dense predictors (Ke et al., 2024; Lee et al., 2024), where deterministic or few-step sampling has proven effective (Xu et al., 2025; He et al., 2025). ResShift (Yue et al., 2025) shortens the chain by shifting the residual between a degraded and a clean image, and we run it from RGB to polarization. A plain pixel transformer (Li & He, 2026) serves as the denoiser, accepting any resolution and concatenated conditions.

Shape from polarization and monocular geometry. The link between polarization and surface normals has been studied for over three decades, almost always in the inverse direction, recovering a normal from a polarization measurement (Wolff & Boult, 1991; Rahmann & Canterakis, 2001; Miyazaki et al., 2003; Atkinson & Hancock, 2006). This inversion is ill-posed, because the azimuth is recovered only up to π, diffuse and specular reflection place the AoLP 90° apart, and the refractive index is unknown. Later methods resolve these ambiguities with added cues such as shading (Smith et al., 2019), a coarse depth map (Kadambi et al., 2015), or multiple views (Cui et al., 2017; Zhu & Smith, 2019). Deep SfP (Ba et al., 2020) provides a network with the ambiguous normal candidates of the physical model, Lei et al. (2022) show that the relation must be evaluated in the frame of the per-pixel viewing ray, and later work extends the formulation to reflectance and incident Stokes fields (Deschaintre et al., 2021; Li et al., 2024). We use the same relation in the forward direction. The ambiguities that make the inverse problem ill-posed also leave RGB-to-polarization underdetermined, so we likewise provide the network with the physically admissible candidates, computed in the frame of the viewing ray. This is now practical because geometric foundation models predict depth and surface normals from a single image (Wang et al., 2025a;b; Yang et al., 2024; Ke et al., 2024).

3 Method

Figure 3 summarizes GeoPolar. A frozen monocular geometry model estimates normals from the RGB image, which are converted into a prior in polarization coordinates (Section 3.2). The prior and the RGB image condition a four-step pixel-space diffusion model (Section 3.3), which predicts the Stokes vector with an angular loss applied only where the angle is observable (Section 3.4).

3.1 Polarization and Surface Geometry

From four analyzer measurements $I_0, I_{45}, I_{90}, I_{135}$, the linear Stokes parameters $S_0 = \frac{1}{2}\sum_i I_i$, $S_1 = I_0 - I_{90}$, and $S_2 = I_{45} - I_{135}$ give the degree and angle of linear polarization,

$$\rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}, \qquad \phi = \tfrac{1}{2}\operatorname{atan2}(S_2, S_1), \qquad \left(\frac{S_1}{S_0}, \frac{S_2}{S_0}\right) = \rho\,(\cos 2\phi,\ \sin 2\phi). \tag{1}$$

The polarization of reflected light depends on the surface geometry, the material, and the reflection mixture. For a smooth dielectric, Fresnel reflection ties both quantities to the surface normal $\mathbf{n}$, with azimuth ψ in the image plane and zenith angle θ from the viewing ray,

$$\phi \equiv \psi \pmod{\pi} \text{ (diffuse)}, \qquad \phi \equiv \psi + \tfrac{\pi}{2} \pmod{\pi} \text{ (specular)}, \qquad \rho = f(\theta), \tag{2}$$

where f differs between diffuse and specular reflection, vanishes at $\theta = 0$, and depends on the refractive index (Atkinson & Hancock, 2006; Miyazaki et al., 2003). A shift of $\pi/2$ in ϕ is a shift of π in 2ϕ, so in doubled angles the branches differ only by a sign,

$$(\cos 2\phi,\ \sin 2\phi) = \pm(\cos 2\psi,\ \sin 2\psi). \tag{3}$$

The normal is therefore an informative geometric cue but does not determine polarization alone. It fixes the direction up to a choice of branch, which depends on the reflection type, and the DoLP up to the material and the reflection mixture. A model that receives only the RGB image has to infer the normal implicitly, in addition to these factors. Supplying the normal as a prior leaves these ambiguities in place but relieves the generator of the geometric part of the inference, so that the network has to select the branch and set the magnitude rather than also recover the geometry.

3.2 Geometry prior in polarization coordinates

MoGe-2 (Wang et al., 2025b), frozen and run once per image, returns from the RGB image $\mathbf{I}$ a camera-space normal map $\mathbf{n}$, a validity mask, and the intrinsics $\mathbf{K}$.

From a normal to polarization coordinates. A raw normal map is a poor interface to (2), because the network would have to project the vector onto the image plane, convert it into a π-periodic angle, and relate the remaining component to the magnitude, all from a few thousand images. These data-independent steps are carried out analytically. We build an orthonormal frame on the viewing ray $\mathbf{d} \propto \mathbf{K}^{-1}[u, v, 1]^\top$ of each pixel (u, v), with $\mathbf{e}_3 = -\mathbf{d}$ toward the camera, $\mathbf{e}_1$ the image horizontal projected onto the plane orthogonal to $\mathbf{d}$, and $\mathbf{e}_2 = \mathbf{e}_3 \times \mathbf{e}_1$, so that ψ, like the AoLP of the data, runs from the image horizontal toward the image top, and compute

$$\psi = \operatorname{atan2}(\mathbf{n}\cdot\mathbf{e}_2,\ \mathbf{n}\cdot\mathbf{e}_1), \qquad \cos\theta = \mathbf{n}\cdot\mathbf{e}_3, \qquad \mathbf{g} = [\sin 2\psi,\ \cos 2\psi,\ \cos\theta] \in [-1, 1]^3. \tag{4}$$

Invalid pixels receive $\mathbf{g} = \mathbf{0}$, a "no information" value that no valid pixel attains. At $\theta = 0$, where ψ is undefined, we set $\psi = 0$; this has no effect because the DoLP vanishes there by (2).

What the encoding provides. By (3), the target's direction channels $(S_1/S_0, S_2/S_0)$ equal $\pm\rho\,(\cos 2\psi, \sin 2\psi)$, the first two channels of $\mathbf{g}$ in reverse order scaled by ρ, so the network selects only the branch and the magnitude. Because ψ and θ are computed against $\mathbf{d}$, they remain valid at any field of view, as the Fresnel relation requires (Lei et al., 2022). The third channel, the argument of f in (2), supplies the geometric factor of the DoLP, leaving the rest to the network.

3.3 Injecting the prior into a few-step pixel-space generator

Chain. Let $\mathbf{x}_0 \in \mathbb{R}^{H\times W\times 3}$ be the polarization target and $\mathbf{y} = \mathbf{I}$ the RGB image, both scaled to $[-1, 1]$. Following ResShift (Yue et al., 2025), the chain moves from $\mathbf{x}_0$ toward $\mathbf{y}$ instead of noise,

$$\mathbf{x}_t = \mathbf{x}_0 + \eta_t\,(\mathbf{y} - \mathbf{x}_0) + \kappa\sqrt{\eta_t}\,\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathrm{Id}), \qquad t = 1, \ldots, T, \tag{5}$$

with Id the identity and a monotone schedule from $\eta_1 \approx 0$ to $\eta_T \approx 1$. Because its far end is an informative image the chain can be short, and we use $T = 4$ steps and a noise scale of $\kappa = 2$. Inference starts from the noised RGB image, and the network predicts $\hat{\mathbf{x}}_0 = F_\Theta(\mathbf{x}_t, \mathbf{c}, t)$ at each step.

Conditioning path. The condition is $\mathbf{c} = [\mathbf{I}, \mathbf{g}]$, so the RGB image serves both as the far end $\mathbf{y}$ of (5) and as part of the condition. The network F_Θ, a plain pixel transformer (JiT) (Li & He, 2026) with 129.9M parameters, embeds 8×8 patches of the nine-channel input $[\mathbf{x}_t, \mathbf{c}]$, so the prior and the noisy state interact from the first attention layer, without an autoencoder, text encoder, or control branch.

3.4 Stokes target and observability-masked angular loss

The target is the scaled linear Stokes vector

$$\mathbf{x}_0 = \big[\rho,\ k\,S_1/S_0,\ k\,S_2/S_0\big] = \big[\rho,\ k\rho\cos 2\phi,\ k\rho\sin 2\phi\big], \qquad k = 1.5, \tag{6}$$

Table 1: Accuracy on the combined benchmark at native resolution. Subscript m denotes the polarization mask, PA is PolarAnything, and GEOPOLAR w/o geometry is the same network trained without the prior. **Bold** and underline mark the best and second best per column.

	RLP				DoFP				RSP			
Method	**AoLP$_m$** ↓	**DoLP$_m$** ↓	**SSIM** ↑	**PSNR** ↑	**AoLP$_m$** ↓	**DoLP$_m$** ↓	**SSIM** ↑	**PSNR** ↑	**AoLP$_m$** ↓	**DoLP$_m$** ↓	**SSIM** ↑	**PSNR** ↑
PA, T=20 ICCV25	40.1	0.245	0.867	26.28	33.2	0.135	0.837	29.03	19.8	0.118	0.976	40.67
PA, T=4	42.7	0.235	0.685	22.96	37.1	0.127	0.604	23.03	24.8	0.206	0.895	31.57
GenPolar ECCV26	33.4	0.132	0.981	36.84	27.2	0.113	0.970	34.84	21.9	0.101	0.982	43.44
GEOPOLAR w/o geometry	39.7	0.108	0.983	36.24	18.4	0.061	0.986	35.06	18.2	0.066	0.996	47.18
GEOPOLAR (Ours)	**28.2**	**0.105**	**0.985**	**38.01**	**11.7**	**0.055**	**0.989**	**37.73**	**13.4**	**0.065**	**0.996**	**49.04**

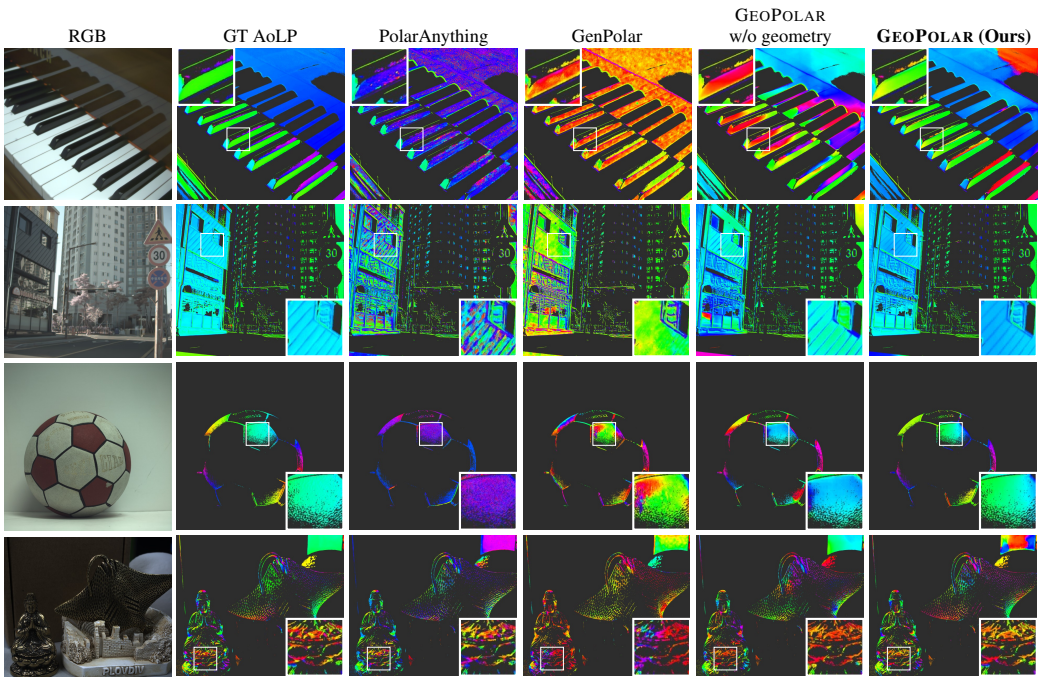


Figure 4: Qualitative AoLP results on DoFP (first two rows) and RLP (last two rows). On specular surfaces, the baselines and the model without geometry deviate over large regions, whereas GEOPOLAR follows the ground truth. Pixels with ground-truth DoLP below 0.1 are masked.

where k balances the variances of the direction channels and ρ. In training, ρ is mapped to $2\rho - 1$ and the direction channels are clipped to $[-1, 1]$. Unlike a unit-direction target (Zhang et al., 2025), which requires an angle even where $\rho \approx 0$, the Stokes target there is $(0, 0)$. The loss is

$$\mathcal{L} = \frac{1}{N}\left\|\hat{\mathbf{x}}_0 - \mathbf{x}_0\right\|_2^2 + \lambda \frac{1}{|\Omega|}\sum_{p\in\Omega}\left(1 - \cos 2\Delta\phi_p\right), \qquad \Omega = \{p : \rho_p > \tau\}, \tag{7}$$

where N is the number of entries of $\mathbf{x}_0$, $\cos 2\Delta\phi_p$ is the normalized inner product of the predicted and true direction channels, and $\lambda = \tau = 0.05$. The first term supervises magnitude and direction everywhere, and the bounded second term acts only on Ω, where the angle is defined.

3.5 TRAINING AND INFERENCE

Training uses random 512×512 crops at native resolution, without resizing or any other augmentation. Inference uses overlapping 512-pixel tiles at the training scale (Appendix A).

4 EXPERIMENTS

4.1 SETUP

Data. Training and evaluation use the combined benchmark, which follows the grouping of GenPolar (Luo et al., 2026) and whose ground truth is computed by (1). Its three subsets, with 50 test

Table 2: Transfer to the unseen-camera set. No model of Table 1 sees this set in training. **Bold** and underline mark the best and second best per column.

Method	$\text{AoLP}_\text{m}\downarrow$	$\text{DoLP}_\text{m}\downarrow$	SSIM↑	PSNR↑
PA, T=20	47.2	0.163	0.912	30.40
PA, T=4	47.1	0.155	0.715	24.54
GenPolar	37.6	0.173	0.964	**32.46**
GeoPolar w/o geometry	35.5	0.148	0.966	32.04
GeoPolar (Ours)	**30.5**	**0.140**	**0.967**	32.34

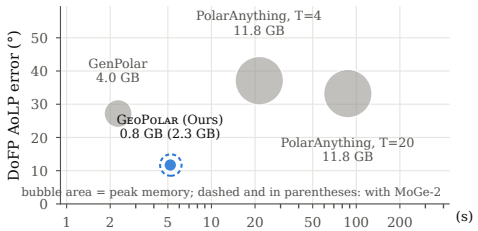


Figure 5: Accuracy against cost on DoFP. Time is measured per 1900×2100 image, and bubble area gives peak memory. GeoPolar is the most accurate and lightest model, with or without MoGe-2.

Table 3: Cost at native resolution. The times t_{512} and t_{full} are for a 512^2 input and a 1900×2100 image, and Mem is peak memory. **Bold** and underline mark the best and second best per column.

Method	Params (M)↓	Steps	t_{512} (ms)↓	t_{full} (s)↓	Mem (GB)↓	$\text{AoLP}_\text{m}\downarrow$
PA, T=20	1427.5	20	1893	87.47	11.82	33.2
PA, T=4	1427.5	4	490	21.40	11.82	37.1
GenPolar, T=4	1427.5	4	307	6.12	3.97	36.7
GenPolar, T=1	1427.5	1	**96**	**2.27**	3.99	27.2
GeoPolar (Ours)	**129.9**	4	170	5.20	**0.85**	**11.7**
GeoPolar (Ours) + MoGe-2	460.8	4	213	5.24	2.28	11.7

images each, are *RLP* (298 images), real indoor scenes captured through a rotating linear polarizer (Morimatsu et al., 2020; Rahman et al., 2025; Qiu et al., 2019; Wen et al., 2021; Zhou et al., 2025); *DoFP* (2,078 images), real division-of-focal-plane captures (Jeon et al., 2024; Yao et al., 2025); and *RSP* (2,000 images), a mixture of simulated and real measurements (Kurita et al., 2023), most of them barely polarized. We also evaluate on 198 four-angle captures of a new unseen-camera set, taken with a calibrated pair of LUCID polarization cameras indoors and outdoors, on which no model is trained (Appendix D).

Baselines. PolarAnything (Zhang et al., 2025) fine-tunes the Stable Diffusion 1.5 U-Net with a ControlNet-style encoder in its latent space and samples with $T = 20$ steps, and we also report $T = 4$. GenPolar (Luo et al., 2026) uses the same backbone, predicts Stokes components, and is distilled into a one-step generator. Because the exact training data of their released weights are not available to us, both are trained with their released code on our training split (Appendix E). The matched control, GeoPolar without geometry, is the same network trained with the same target, loss, data, and crops, but its condition is the RGB image alone, $\mathbf{c} = \mathbf{I}$, so that its patch embedding reads the six channels $[\mathbf{x}_t, \mathbf{I}]$ instead of the nine channels $[\mathbf{x}_t, \mathbf{I}, \mathbf{g}]$.

Protocol and metrics. Every test image is evaluated at native resolution. The AoLP is undefined for unpolarized light and noisy at low DoLP, so we measure the angle only on a mask of pixels whose brightness and ground-truth DoLP are both at least 0.1. AoLP_m is the mean angular error on this mask, modulo 180°, and DoLP_m the mean absolute DoLP error on it. PSNR and SSIM compare the four analyzer images reconstructed through Malus's law with the captured ones (Appendix B).

4.2 Main Results

Table 1 compares all models. GeoPolar is the most accurate model in every column of the three subsets, tied only in the SSIM of RSP, whereas the latent baselines are 5° to 25° behind in AoLP_m despite ten times the parameters of our generator and a pretrained image prior.

The network trained without the prior shares the architecture, target, and loss, so the comparison with it isolates the effect of the prior. The prior lowers AoLP_m on every subset, from 18.4° to 11.7° on DoFP, from 39.7° to 28.2° on RLP, and from 18.2° to 13.4° on RSP. Section 5.2 shows that it mainly removes errors close to 90°, that is, predictions on the wrong branch. The prior changes the DoLP only slightly (Appendix C), and PSNR follows the DoLP errors only loosely. In Figure 4, the model without geometry and the baselines deviate from the ground truth over large regions of the specular surfaces, in several cases by about 90°, whereas GeoPolar follows it closely.

Table 4: Form of the condition. Results are averaged over the three subsets, with the condition rewritten at test time. **Bold** and underline mark the best and second best above the rule.

Condition	$\text{AoLP}_\text{m}\downarrow$	$\text{DoLP}_\text{m}\downarrow$	PSNR↑
$\mathbf{g}$ (as trained)	**17.7**	**0.075**	**41.59**
azimuth only	23.9	0.083	36.85
zenith only	28.5	0.083	39.87
raw normal	34.3	0.110	38.02
$\mathbf{g} = \mathbf{0}$	33.3	0.095	37.19
GeoPolar w/o geometry	25.4	0.078	39.49

Table 5: Quality of the normals. Results are averaged over the three subsets, with a smooth random error of the given standard deviation added to the azimuth and zenith of the prior.

Error added	$\text{AoLP}_\text{m}\downarrow$	$\text{DoLP}_\text{m}\downarrow$	PSNR↑
0° (estimated)	17.7	0.075	41.59
5°	18.4	0.075	41.40
10°	19.8	0.078	40.93
20°	24.0	0.088	39.92
30°	28.0	0.098	39.38
45°	30.7	0.101	38.93
GeoPolar w/o geometry	25.4	0.078	39.49

Table 6: Branch of the predicted angle on the mask. ✓ is the share of pixels (%) within 22.5° of the ground truth, and ✗ the share beyond 67.5°, that is, on the wrong branch. **Bold** and underline mark the best and second best per column.

	RLP		DoFP		RSP	
Method	✓↑	✗↓	✓↑	✗↓	✓↑	✗↓
PA, T=20	21.9	**10.4**	45.4	14.3	68.0	6.9
PA, T=4	18.7	14.8	39.7	19.1	57.0	8.0
GenPolar	46.0	15.3	56.1	9.7	64.5	9.9
GeoPolar w/o geometry	38.9	24.9	74.6	8.5	75.9	10.6
GeoPolar (Ours)	**55.9**	14.0	**84.0**	**2.3**	**86.4**	**5.2**

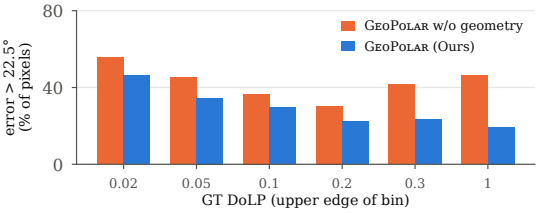


Figure 6: Angular errors above 22.5° by ground-truth DoLP for the network without and with the prior, with RLP and DoFP pooled. Bins are labeled by upper edge, and the prior lowers the share in every bin.

4.3 Transfer to the unseen-camera set

Table 2 repeats the evaluation on the unseen-camera set. Every model degrades on this set, and the differences narrow. GeoPolar remains the most accurate model in angle, DoLP, and SSIM, with a margin of 5.0° in AoLP_m over the model without geometry, and GenPolar has the highest PSNR. Among the baselines, GenPolar transfers best, yet GeoPolar still leads it by 7.1° in AoLP_m and trails it in PSNR by only 0.12 dB, whereas PolarAnything falls to 47.2°, close to the chance level of 45°. Because no model is trained on this camera, the advantage of the prior does not depend on the sensors of the training data.

4.4 Efficiency

Table 3 lists the cost of every model on the same GPU with a batch size of one, as the median of 20 runs, and Figure 5 plots the DoFP angular error against the full-image time. The latent baselines comprise the U-Net, control branch, VAE, and text encoder of Stable Diffusion, 1.4B parameters in total, whereas GeoPolar is a single 130M-parameter network, or 461M parameters with MoGe-2, a third of theirs. At the same number of steps it is faster than both baselines on a 1900×2100 image, with a peak memory below 1 GB, or 2.3 GB including MoGe-2, a fifth of that of PolarAnything and below the 4.0 GB of GenPolar. Only the distilled one-step GenPolar is faster.

5 Discussion

This section examines where the prior acts, why it helps, and where it does not. Appendix G further tests whether the gain follows the content of the condition or merely its presence.

5.1 Error by polarization level

Figure 7 gives the angular error of every model per ground-truth DoLP bin (values in Table 7), and Figure 6 the share of errors above 22.5°, both over all pixels with a brightness of at least 0.1. The lowest bin, below 0.02, holds over two fifths of the real pixels, and there every model errs by more than 23°, on RLP near the chance level of 45°, so an unmasked error would be dominated by pixels whose angle is barely defined. GeoPolar is the most accurate model in every bin on DoFP and in all but one on RLP. On these two real subsets its margin over the model without geometry grows with the DoLP, from about 5° below 0.1 to about 15° above 0.2, where the angle is best defined.

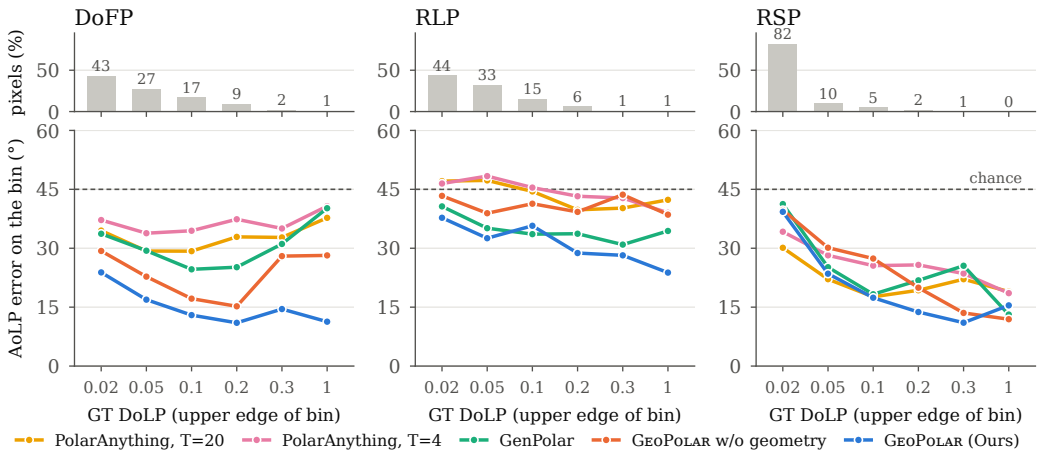


Figure 7: Angular error by DoLP. Bars give the share of pixels per ground-truth DoLP bin, and lines the AoLP error of each model. All models err strongly at low DoLP, and GeoPolar is the most accurate in every bin on DoFP and in all but one on RLP.

Table 7: Values of Figure 7. The table gives the share of pixels (%) and the AoLP error (°) in each ground-truth DoLP bin. **Bold** and underline mark the best and second best per column.

	DoFP						RLP						RSP					
Method	**<.02**	**.02–.05**	**.05–.10**	**.10–.20**	**.20–.30**	**≥.30**	**<.02**	**.02–.05**	**.05–.10**	**.10–.20**	**.20–.30**	**≥.30**	**<.02**	**.02–.05**	**.05–.10**	**.10–.20**	**.20–.30**	**≥.30**
share of pixels (%)	43.0	27.4	17.2	9.3	2.4	0.7	44.0	32.6	15.2	6.3	1.0	0.9	81.9	10.2	4.9	2.2	0.6	0.2
PolarAnything, T=20	34.5	29.3	29.3	32.9	32.8	37.7	47.1	47.3	44.5	39.8	40.2	42.3	**30.1**	**22.1**	17.6	19.3	22.1	19.0
PolarAnything, T=4	37.2	33.8	34.5	37.4	35.0	40.7	46.5	48.4	45.4	43.2	42.8	39.0	34.2	28.2	25.6	25.8	23.6	18.6
GenPolar	33.7	29.4	24.6	25.2	31.1	40.2	40.7	35.1	**33.6**	33.7	30.9	34.4	41.2	25.2	18.3	21.8	25.5	13.1
GeoPolar w/o geometry	29.3	22.8	17.2	15.2	28.0	28.2	43.3	38.9	41.3	39.3	43.6	38.6	39.7	30.1	27.4	20.0	13.5	**11.9**
GeoPolar (Ours)	**23.8**	**16.9**	**13.0**	**11.0**	**14.5**	**11.3**	**37.8**	**32.6**	35.7	**28.8**	**28.2**	**23.8**	39.3	23.5	**17.4**	**13.8**	**11.0**	15.5

The latent baselines show the opposite trend on DoFP, where their errors rise again in the highest bin, to 37.7° for PolarAnything and 40.2° for GenPolar, so they are least reliable where the angle is best defined. The model without geometry also rises there, from 15.2° to about 28°, whereas GeoPolar stays between 11° and 15°, so the prior largely avoids this rise.

5.2 Branch errors

A prediction on the wrong branch of (3) errs by about 90° however well-defined the angle is. Table 6 therefore divides the angular errors on the mask at 22.5° and 67.5°. Errors above 67.5° correspond to the wrong branch, which a uniformly random angle reaches with a probability of 25%. Without the prior, 8.5% of the DoFP and 24.9% of the RLP pixels lie there, and 2.3% and 14.0% with it, while the share within 22.5° rises from 74.6% to 84.0% and from 38.9% to 55.9%. GenPolar places 9.7% and 15.3% of the pixels on the wrong branch, whereas most errors of PolarAnything are intermediate, so the branch error is a principal but not the only failure of appearance-only models. Measured against the prior azimuth instead of the ground truth, the prior reduces the branch swaps by about half or more on every subset, and on DoFP GeoPolar is more accurate than the best branch choice for the prior azimuth, so it also refines the azimuth (Appendix H). Figure 6 shows that the prior lowers the share of large errors in every DoLP bin, and Figure 2 shows specular scenes, whose ground-truth AoLP is the azimuth of the prior rotated by 90° and therefore appears in the complementary hue.

5.3 Form of the geometry condition

Table 4 rewrites the condition of the trained model at test time as the azimuth alone or the zenith cosine alone (the other channels set to zero), the raw camera-space normal of the same MoGe-2 prediction, or $\mathbf{g} = \mathbf{0}$. Only the first row matches the training condition, so the table measures what each part of the prior contributes for this model, not what a model trained on another form would achieve. Every rewriting increases the error. The azimuth alone increases it least and remains more accurate than the model without geometry on average (23.9° against 25.4°), as expected, since the azimuth specifies the two admissible directions. The zenith cosine alone retains much less, and the

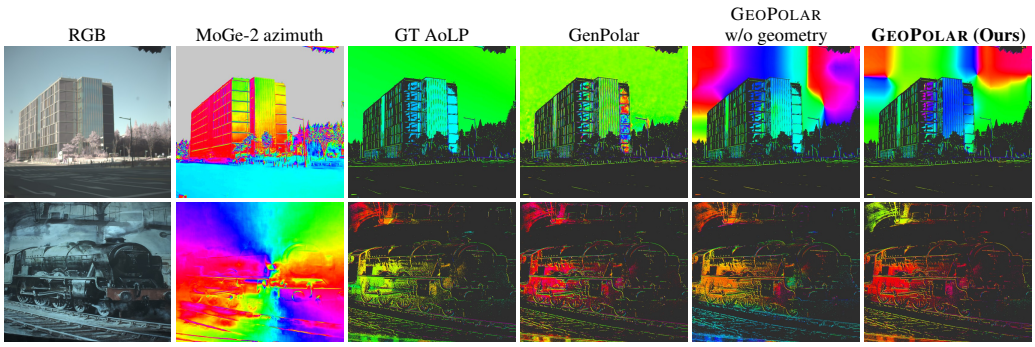


Figure 8: Failure cases of the prior. In the top row, the geometry model marks the sky as invalid (light gray), so GeoPolar receives no prior there, and GenPolar, which does not use geometry, is more accurate in the sky. In the bottom row, the geometry model returns the shape of a painted train instead of the flat wall, and the prior misleads GeoPolar (Section 6). Pixels with ground-truth DoLP below 0.1 are masked.

raw normal is worth no more than $\mathbf{g} = \mathbf{0}$, so the model relies on the polarization coordinates rather than on the normal as such. Conversely, zeroing the RGB image while keeping $\mathbf{g}$ brings the error close to chance (Appendix G), so the prior fixes the admissible directions and the appearance selects between them.

5.4 Quality of the normals

The prior is only an estimate. By (2) its azimuth should match the measured AoLP up to the branch. Table 5 tests the dependence on normal quality by adding a smooth random error to the azimuth and zenith of the prior (Appendix A). Errors up to about 10° change the result little. Near 20° the model is about as accurate as without the prior. Errors of 30° or more make GeoPolar less accurate than without geometry, indicating that the model relies on the prior rather than merely tolerating it.

6 Limitations

Where the geometry model marks a region invalid, the network receives no geometric information. In the scene of Figure 8, the error of GeoPolar rises from 16.0° on the building to 26.3° in the invalid sky, where GenPolar is more accurate (16.0°). A valid prior can also describe the wrong surface. On the painting in the second row of Figure 8, the geometry model returns the shape of the depicted train rather than the flat wall, and the error of GeoPolar rises to 27.8°, against 21.1° without geometry. An inaccurate normal, with errors of 30° or more, makes the result worse than without the prior (Section 5.4). Letting the network estimate where the prior can be trusted, through a learned per-pixel confidence that gates the prior and is trained with the prior randomly withheld, does not lower the error over the polarization mask (Appendix I), so we keep the simpler model as GeoPolar. The relation (2) assumes a dielectric, and metals and multilayer materials remain untested. The angle is tied to the analyzer axis of the training sensors, so a differently aligned camera requires rotating the direction channels. Combining the generator with a pretrained image prior remains open.

7 Conclusion

Polarization depends strongly on surface geometry, which appearance alone cannot determine. GeoPolar supplies this geometry explicitly, encoding the normals of a monocular foundation model as $[\sin 2\psi, \cos 2\psi, \cos\theta]$ so that the direction problem reduces to the choice of a branch. In a four-step pixel-space diffusion model, the prior lowers the angular error on polarized pixels well below that of the same model without geometry and of latent-diffusion baselines with ten times the parameters of its generator, and the gain follows the geometric content of the prior. Because the prior carries the geometric part of the task, the generator needs no pretrained image prior and, at the same number of steps, runs faster than both latent baselines. The gain persists under normal errors of about 10° and has nearly vanished at 20°, so more accurate monocular geometry can be expected to improve the generator further.

AI USE STATEMENT

In this work, we used generative AI tools, namely a coding agent based on a large language model, to implement methods, to clean and reformat datasets, and to translate drafts written by the authors into English. We have not used generative AI tools to generate synthetic data sets, to develop theoretical models or conceptual frameworks, to propose or refine hypotheses, to design research methodology or experiments or provide feedback on them, to interpret results, or to support qualitative and thematic data analysis, and the formulation of mathematical claims and the writing of proofs are not applicable to this work. Additionally, we used generative AI tools to create or edit software code and to search for related literature. We have reviewed all AI-assisted work. Every number in the paper was taken from the result files of the experiments and checked against them, the training and evaluation code was tested with short runs before the full runs, and every reference was checked against its published source. 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 synthesizes polarization from RGB images of indoor and outdoor scenes and objects. It involves no human subjects and no sensitive attributes. Our own captures were taken in and around office buildings, and the vehicle registration plates that are legible in them are masked in every panel of this paper. A few images of the public subsets contain distant pedestrians at a scale at which no individual is identifiable. The public datasets are used under their own licenses, and the remaining images were captured by the authors.

REPRODUCIBILITY STATEMENT

The encoding of the geometry prior is fully specified by (4), the diffusion process by (5), and the target and loss by (6) and (7). The training recipe, including the schedule η_t, is given in Section 3.5 and Appendix A, and the training of the baselines in Appendix E. The evaluation protocol, the mask thresholds, and the metric definitions are given in Section 4.1 and Appendix B. The construction of the combined benchmark from its three sources is described in Section 4.1, and that of the unseen-camera set in Appendix D. Code, configuration files, and the cached MoGe-2 geometry will be released.

REFERENCES

Gary A. Atkinson and Edwin R. Hancock. Recovery of surface orientation from diffuse polarization. *IEEE Transactions on Image Processing (TIP)*, 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.

Zhaopeng Cui, Jinwei Gu, Boxin Shi, Ping Tan, and Jan Kautz. Polarimetric multi-view stereo. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1558–1567, 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.

Jing He, Haodong Li, Wei Yin, Yixun Liang, Leheng Li, Kaiqiang Zhou, Hongbo Zhang, Bingbing Liu, and Ying-Cong Chen. Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

Inseung Hwang, Kiseok Choi, Hyunho Ha, and Min H. Kim. Benchmarking burst super-resolution for polarization images: Noise dataset and analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24899–24909, 2025.

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.

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, Alberto Dall'Olio, Raghav Khanna, Tomas Gerlich, Georgia Gi-annopoulou, Guy Stoppi, Daniel Baxter, Abhijit Ghosh, Rick Szeliski, and Kartik Venkataraman. A plentoptic 3D vision system. In *Proceedings of the SIGGRAPH Asia 2024 Conference Papers (SIGGRAPH Asia)*, pp. 1–12, 2024.

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.

Naejin Kong, Yu-Wing Tai, and Sung Yong Shin. A physically-based approach to reflection separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9–16, 2012.

Teppei Kurita, Yuhi Kondo, Legong Sun, and Yusuke Moriuchi. Simultaneous acquisition of high quality RGB image and polarization information using a sparse polarization sensor. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 178–188, 2023.

Hsin-Ying Lee, Hung-Yu Tseng, Hsin-Ying Lee, and Ming-Hsuan Yang. Exploiting diffusion prior for generalizable dense prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7861–7871, 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, Hajime Mihara, Alexander Gatto, Hajime Nagahara, and Yusuke Moriuchi. NeISF: Neural incident Stokes field for geometry and material estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21434–21445, 2024.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, 2026.

Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19800–19808, 2022.

Beibei Lin, Zifeng Yuan, and Tingting Chen. RGB-to-polarization estimation: A new task and benchmark study. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 37846–37876, 2025.

Guan-Horng Liu, Arash Vahdat, De-An Huang, Evangelos Theodorou, Weili Nie, and Anima Anandkumar. I^2SB: Image-to-image Schrödinger bridge. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *PMLR*, pp. 22042–22062, 2023.

Yidong Luo, Chenggong Li, Yuchao Feng, Boxin Shi, Junchao Zhang, and Xin Yuan. Stokes-informed diffusion for robust linear polarization estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 512–529, 2026.

Haiyang Mei, Bo Dong, Wen Dong, Jiaxi Yang, Seung-Hwan Baek, Felix Heide, Pieter Peers, Xiaopeng Wei, and Xin Yang. Glass segmentation using intensity and spectral polarization cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12622–12631, 2022.

Daisuke Miyazaki, Robby T. Tan, Kenji Hara, and Katsushi Ikeuchi. Polarization-based inverse rendering from a single view. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, volume 2, pp. 982–987, 2003.

Miki Morimatsu, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Monochrome and color polarization demosaicking using edge-aware residual interpolation. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2571–2575, 2020.

Simeng Qiu, Qiang Fu, Congli Wang, and Wolfgang Heidrich. Polarization demosaicking for monochrome and color polarization focal plane arrays. In *Vision, Modeling and Visualization (VMV)*, 2019.

Muhamad Daniel Ariff Bin Abdul Rahman, Yusuke Monno, Masayuki Tanaka, and Masatoshi Okutomi. Polarization denoising and demosaicking: Dataset and baseline method. In *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, pp. 2724–2729, 2025.

Stefan Rahmann and Nikos Canterakis. Reconstruction of specular surfaces using polarization imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pp. 149–155, 2001.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings (SIGGRAPH)*, pp. 1–10, 2022.

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 (TPAMI)*, 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.

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, 2025a.

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)*, volume 38, pp. 35928–35959, 2025b.

Sijia Wen, Yinqiang Zheng, and Feng Lu. A sparse representation based joint demosaicing method for single-chip polarized color sensor. *IEEE Transactions on Image Processing*, 30:4171–4182, 2021.

Lawrence B. Wolff and Terrance E. Boult. Constraining object features using a polarization reflectance model. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13 (7):635–657, 1991.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025.

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)*, volume 37, pp. 21875–21911, 2024.

Mingde Yao, Menglu Wang, King-Man Tam, Lingen Li, Tianfan Xue, and Jinwei Gu. PolarFree: Polarization-based reflection-free imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10890–10899, 2025.

Zongsheng Yue, Jianyi Wang, and Chen Change Loy. Efficient diffusion model for image restoration by residual shifting. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 47(1):116–130, 2025.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5728–5739, 2022.

Kailong Zhang, Youwei Lyu, Heng Guo, Si Li, Zhanyu Ma, and Boxin Shi. PolarAnything: Diffusion-based polarimetric image synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 26466–26476, 2025.

Shuangfan Zhou, Chu Zhou, Youwei Lyu, Heng Guo, Zhanyu Ma, Boxin Shi, and Imari Sato. PIDSR: Complementary polarized image demosaicing and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16081–16090, 2025.

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 Implementation details

Diffusion. We use $T = 4$ and $\kappa = 2$. The schedule η_t is exponential with exponent 0.3, a minimum noise level of 0.2, and $\eta_T = 0.98$, which gives $\eta = (0.010,\ 0.271,\ 0.580,\ 0.980)$ for $t = 1, \ldots, 4$. The network input is scaled by $1/(3\kappa\sqrt{\eta_t}+1)$, and the network predicts $\mathbf{x}_0$ directly. The reverse step forms the posterior mean of $\mathbf{x}_{t-1}$ from the prediction. Inference starts at $t = 4$ from $\mathbf{y} + \kappa\sqrt{\eta_4}\,\boldsymbol{\epsilon}$.

Backbone. The backbone uses a patch size of 8, a bottleneck width of 128, a hidden width of 768, 12 blocks, 12 heads, SwiGLU feed-forward layers with an expansion factor of 4, RMS normalization, adaptive layer normalization driven by a 256-dimensional sinusoidal timestep embedding passed through an MLP, two-dimensional sine and cosine positional embeddings computed for the input size, a final adaptively normalized linear layer to $8 \times 8 \times 3$ patches, and a 1×1 output convolution. It has 129.9M parameters in total.

Geometry. MoGe-2 (`moge-2-vitl-normal`) is run once per image at native resolution. Its normal map, validity mask, and intrinsics are encoded by (4).

Training. We use random 512×512 crops at native resolution, without resizing or any other augmentation. Optimization uses AdamW with a learning rate of 5×10^{-5} decayed to 2×10^{-5} by a cosine schedule after $5{,}000$ warm-up iterations, a weight decay of 0, a batch of 32 (micro-batches of 8 with four accumulation steps), an exponential moving average (EMA) of the weights with a decay of 0.999, and mixed precision, in float16 with loss scaling in the early phase of training and in bfloat16 thereafter. The EMA weights are used at inference. We set $k = 1.5$, $\lambda = 0.05$, and $\tau = 0.05$, and in the angular term of (7) 10^{-2} is added to the squared norms of both direction pairs before normalization. The angular term is computed on the channel-averaged Stokes pair.

Inference. The latent baselines process the full image, padded to a multiple of eight, with the VAE encoder and decoder run on overlapping tiles as in the released inference code of PolarAnything, while their U-Net and control branch process the whole latent in one pass. Our models process images in tiles of 512 pixels with a stride of 384, the overlaps are averaged, and the four steps run in float32. All reported numbers are computed in this way. Because each tile is predicted from its own 512-pixel context, neighboring tiles disagree slightly and plain averaging leaves faint tile borders. In the qualitative figures we therefore merge the tiles with a linear ramp at a stride of half the tile, which removes the borders and changes the masked AoLP error of the scenes shown by up to about two degrees. The output is stored as $[\sin 2\hat{\phi}, \cos 2\hat{\phi}, \hat{\rho}]$ in 16-bit PNG, and the analyzer images for

Table 8: Degree of polarization on the mask, averaged over the three subsets. Every value is the mean of the three per-subset values of Table 9. Bias_m is the mean signed DoLP error, and σ ratio is the standard deviation of the predicted DoLP over that of the ground truth. **Bold** and underline mark the best and second best, with Bias_m and σ ratio ranked by their distance to 0 and 1. Our two models have the lowest error.

Method	$\text{DoLP}_\text{m}\downarrow$	Bias_m	σ ratio
PA, T=20	0.166	**+0.034**	1.75
PA, T=4	0.189	+0.128	1.88
GenPolar	0.115	-0.102	0.70
GeoPolar w/o geometry	0.078	-0.051	0.80
GeoPolar (Ours)	**0.075**	-0.049	**0.82**

PSNR are reconstructed per color channel as $I_\alpha = \mathbf{I}\,(1 + \hat{\rho}\cos(2\alpha - 2\hat{\phi}))$ from the RGB image $\mathbf{I}$, where α is the analyzer angle.

Test-time conditions. Figure 9 shows the conditions of Tables 4 and 5 for one test image. For the azimuth-only and zenith-only rows the other channels are set to zero, and the raw-normal row replaces $\mathbf{g}$ by the camera-space normal of the same MoGe-2 prediction. The error of Table 5 is Gaussian-filtered white noise with a correlation length of 4% of the image width, drawn independently for the azimuth and the zenith angle and scaled to the given standard deviation. The zenith angle is clipped to $[0°, 90°]$, and invalid pixels remain invalid.

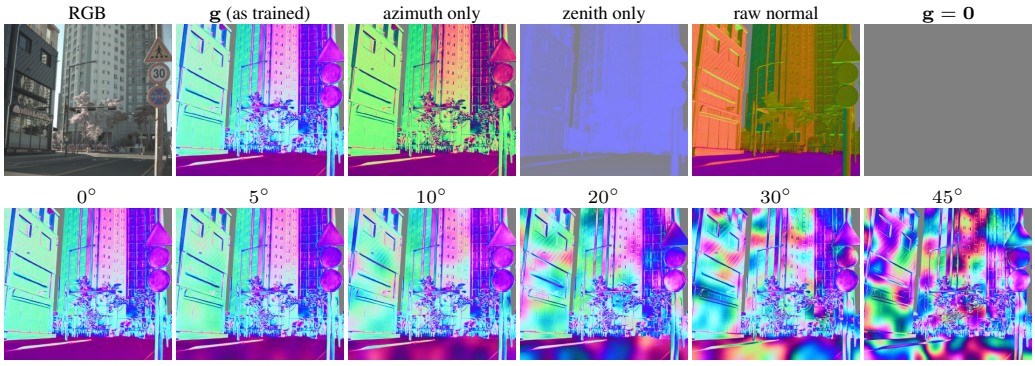


Figure 9: Conditions of Tables 4 and 5 for one test image. The top row shows the condition as trained and its test-time rewritings, and the bottom row the condition with a smooth random error of the given standard deviation added to its azimuth and zenith. The three channels of each condition are shown as red, green, and blue, with 0 as mid gray.

B Metric definitions

Let ϕ and $\hat{\phi}$ be the true and predicted AoLP, ρ and $\hat{\rho}$ the true and predicted DoLP, and b the brightness of a pixel, the luminance $0.299R+0.587G+0.114B$ of the RGB input in $[0, 1]$. The angular error is $\Delta\phi = \min(|\hat{\phi}-\phi|,\ \pi-|\hat{\phi}-\phi|)$. AoLP_m is the mean of $\Delta\phi$ over $M = \{p : b_p \geq 0.10,\ \rho_p \geq 0.10\}$, pooled over all pixels of a subset, and DoLP_m is the mean of $|\hat{\rho}-\rho|$ over M. PSNR is computed per image between the four reconstructed analyzer images and the captured ones in 8-bit and is averaged over images and angles. SSIM is computed on the same 8-bit images per color channel with a 7×7 window and averaged in the same way. The captured analyzer images and the polarization ground truth are not perfectly consistent, so PSNR is reported only as a secondary metric.

C Degree of polarization

The prior acts mainly on the angle. Table 8 reports the DoLP on the same mask, averaged over the three subsets. Both variants of our model have the lowest error and the spread of the DoLP closest to the true one, although they still underestimate it by about a fifth, whereas PolarAnything overestimates the spread by a factor of about 1.8, and GenPolar underestimates it more strongly,

Table 9: Degree of polarization per subset. DoLP_m and Bias_m are the mean absolute and signed DoLP errors on the polarization mask, σ ratio is the standard deviation of the predicted DoLP over that of the ground truth on the mask, ideally 1, and PSNR_ρ compares the whole DoLP map with the ground truth. **Bold** and underline mark the best and second best per column, with Bias_m and σ ratio ranked by their distance to 0 and 1.

	RLP				DoFP				RSP			
Method	**$\mathrm{DoLP}_\mathrm{m}\downarrow$**	**Bias_m**	**σ ratio**	**$\mathrm{PSNR}_\rho\uparrow$**	**$\mathrm{DoLP}_\mathrm{m}\downarrow$**	**Bias_m**	**σ ratio**	**$\mathrm{PSNR}_\rho\uparrow$**	**$\mathrm{DoLP}_\mathrm{m}\downarrow$**	**Bias_m**	**σ ratio**	**$\mathrm{PSNR}_\rho\uparrow$**
MAE	0.148	-0.136	0.55	26.50	0.132	-0.096	1.39	27.81	0.108	-0.037	1.37	22.31
Restormer	0.162	-0.159	0.26	25.87	0.156	-0.134	1.52	26.91	0.153	-0.141	0.71	16.69
PA, T=20	0.245	+0.179	1.72	14.54	0.135	-0.081	1.92	20.68	0.118	**+0.004**	1.61	22.67
PA, T=4	0.235	+0.193	**1.40**	10.07	0.127	**+0.010**	2.26	15.26	0.206	+0.180	1.98	13.64
GenPolar	0.132	-0.122	0.48	22.61	0.113	-0.101	0.86	26.70	0.101	-0.083	0.75	13.54
GeoPolar w/o geometry	0.108	-0.100	0.45	26.34	0.061	-0.026	0.86	27.88	0.066	-0.027	1.08	36.71
GeoPolar (Ours)	**0.105**	**-0.092**	0.55	**27.68**	**0.055**	-0.026	**0.97**	**30.00**	**0.065**	-0.028	**0.94**	**39.32**

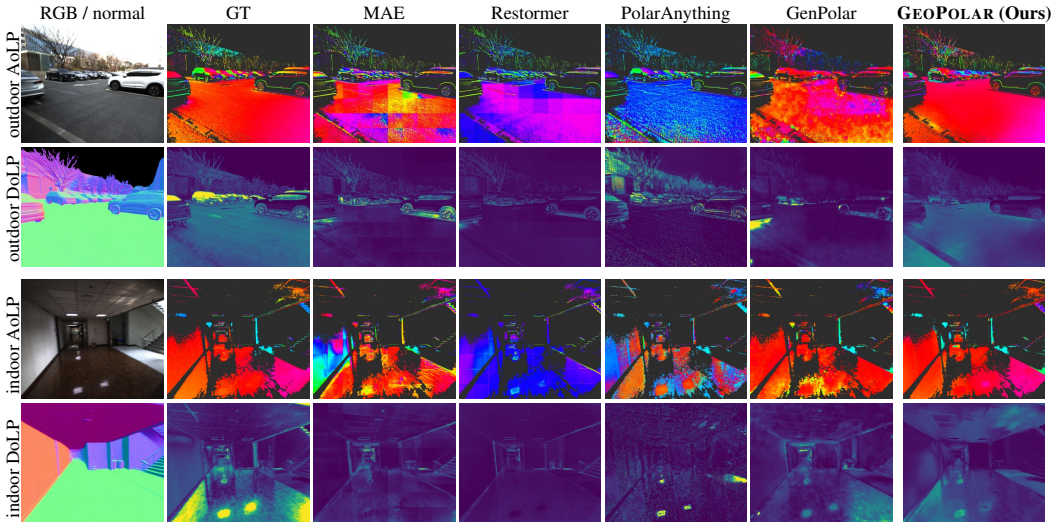


Figure 10: Comparisons on the unseen-camera set, on which no model is trained, with the layout of Figure 12.

with a larger error and a negative bias. Since the model without geometry shares this advantage, it does not come from the prior, and we attribute it mainly to the Stokes target in pixel space. The prior itself changes the DoLP only slightly, although its third channel is the argument of the Fresnel function, because the zenith angle determines the DoLP only up to the refractive index, the reflection mixture, and the roughness, none of which the geometry provides. Both pixel-space models also underestimate the DoLP on the polarized pixels, most on RLP (Table 9), consistent with a regression toward the many weakly polarized pixels. This remains the main limitation of the predicted DoLP.

D The unseen-camera set

The unseen-camera set was captured with a calibrated pair of LUCID polarization cameras with a baseline of 220 mm, each recording four analyzer images of 1224×1024 pixels with a horizontal field of view of about $74°$. Its captures come from six scenes, three corridors and one room indoors, several of them under low light, and two outdoor scenes. We evaluate on 198 captures, of which 110 are outdoor and 88 indoor, rectified with the calibration of the stereo rig, and no model is trained on any part of the set. The polarization ground truth is computed from the four analyzer images with (1). Figure 10 compares the models on two captures.

Table 10: Released weights of the two baselines under the protocol of Table 1. The released GenPolar models may have been trained on images of our test split, so these rows are not a held-out evaluation.

	RLP				DoFP				RSP			
Method	**AoLP$_m$** ↓	**DoLP$_m$** ↓	**SSIM** ↑	**PSNR** ↑	**AoLP$_m$** ↓	**DoLP$_m$** ↓	**SSIM** ↑	**PSNR** ↑	**AoLP$_m$** ↓	**DoLP$_m$** ↓	**SSIM** ↑	**PSNR** ↑
PA, T=20	36.8	0.135	0.869	31.75	39.2	0.132	0.767	29.48	48.2	0.150	0.930	33.71
PA, T=4	42.2	0.131	0.721	26.30	47.9	0.107	0.553	22.37	46.5	0.168	0.857	29.63
GenPolar, T=1	**16.5**	**0.082**	**0.985**	**39.47**	**25.1**	**0.092**	**0.987**	**36.94**	**4.4**	**0.083**	**0.986**	**45.84**
GenPolar, T=4	28.4	0.125	0.977	38.20	35.6	0.145	0.982	36.13	6.5	0.094	0.983	44.74

E TRAINING OF THE BASELINES

GenPolar is trained on public datasets together with scenes captured by its authors, which are not available to us, and neither baseline releases the train/test split of its weights, so the data behind the released weights are not known exactly. We therefore train both baselines with their released code in our training setting, on our training split, and evaluate them with the protocol of Section 4.1.

PolarAnything fine-tunes the Stable Diffusion 1.5 U-Net with its ControlNet-style condition encoder for $1{,}000$ epochs with a batch of 16 and a learning rate of 4×10^{-5} in full precision. Following its released data pipeline, each training image is rescaled by a random factor between the smallest one that keeps a 512-pixel side and 1.5 before a 512×512 crop is taken. At test time it runs on the full image, padded to a multiple of eight, with 20 or 4 sampling steps.

GenPolar is trained in two stages with the parameterization stored in the configuration of its released checkpoint, noise prediction and Stokes components scaled by 5. The first stage is trained for 258 epochs with a learning rate of 4×10^{-5} and a batch of 1, and the second distills it into a one-step generator by distribution matching for 30 epochs with a learning rate of 10^{-5} in bfloat16. Both stages use native-resolution 512×512 crops without augmentation. For Table 3, the first stage is also sampled with 4 steps.

MAE and Restormer. The per-subset DoLP table and the qualitative comparisons of the appendix also include MAE (He et al., 2022) and Restormer (Zamir et al., 2022), two regression baselines of the RGB-to-polarization benchmark of Lin et al. (2025), trained on our training split and evaluated with the same protocol.

F RELEASED BASELINE WEIGHTS

The baselines in Table 1 are trained on our training split (Appendix E). Table 10 reports the weights released by the authors of the two baselines under the same protocol, for reference. PolarAnything remains far from the accuracy of our model with either set of weights. The released GenPolar models are far more accurate than their retrained counterparts on RLP and RSP. Their training data are not known exactly and may include images of our test split, because our test split is drawn from public datasets that GenPolar also uses. These rows are therefore not a held-out evaluation, and we do not compare against them. To test whether the gap is an artifact of our retraining, we also trained GenPolar with the parameterization stated in its paper, sample prediction and unscaled Stokes components, and with the parameterization of the released checkpoint from a first stage of 99 epochs. These reach a DoFP AoLP$_m$ of $33.0°$ and $32.5°$, less accurate than the row of Table 1 ($27.2°$).

No model has seen the unseen-camera set, so the released weights can be compared there without this caveat. Table 11 adds them below the rows of Table 2, evaluated with the same protocol on the rectified captures. The released one-step GenPolar, far more accurate than its retrained counterpart on the combined benchmark, reaches an AoLP$_m$ of $45.9°$ here, close to the chance level of $45°$ and $8.3°$ above the retrained model, and its four-step teacher reaches $52.7°$. Its DoLP error of 0.148 is lower than that of the retrained model and equal to that of GEOPOLAR without geometry. The gap in angle is not a difference of convention, since the best constant offset or reflection of the predicted angle lowers the error by about $5°$ and leaves it above $40°$, and whether a rotation by $90°$ helps changes from scene to scene, so the released model chooses the branch inconsistently on the new camera. This supports reading its accuracy on the combined benchmark as a consequence of a possible overlap between its training data and our test split. The released PolarAnything reaches

Table 11: Unseen-camera set with the released weights. Rows above the rule are those of Table 2, trained on our split; rows below are the weights released by the authors. Neither saw this set in training. **Bold** and underline mark the best and second best above the rule.

Method	$\text{AoLP}_\text{m}\downarrow$	$\text{DoLP}_\text{m}\downarrow$	SSIM↑	PSNR↑
PA, T=20	47.2	0.163	0.912	30.40
PA, T=4	47.1	0.155	0.715	24.54
GenPolar	37.6	0.173	0.964	**32.46**
GeoPolar w/o geometry	<u>35.5</u>	<u>0.148</u>	<u>0.966</u>	32.04
GeoPolar (Ours)	**30.5**	**0.140**	**0.967**	<u>32.34</u>
PA (released), T=20	38.6	0.157	0.880	30.43
PA (released), T=4	50.5	0.144	0.771	25.18
GenPolar (released), T=1	45.9	0.148	0.966	32.41
GenPolar (released), T=4	52.7	0.163	0.962	32.35

Table 12: Condition manipulation at test time. The table gives AoLP_m (°) on the combined benchmark when the condition of the trained model is replaced without retraining, and the last row is the model trained without geometry. **Bold** and underline mark the best and second best above the rule.

Condition	RLP↓	DoFP↓	RSP↓
as estimated	**28.2**	**11.7**	**13.4**
geometry zeroed	39.7	<u>29.9</u>	<u>30.4</u>
geometry shuffled	<u>38.5</u>	35.8	36.8
RGB zeroed	40.7	40.6	47.7
GeoPolar w/o geometry	39.7	18.4	18.2

$38.6°$ with 20 steps, more accurate than its retrained counterpart ($47.2°$) but still behind GeoPolar ($30.5°$) and GeoPolar without geometry ($35.5°$), and $50.5°$ with 4 steps.

G Condition manipulation at test time

Table 12 tests whether the gain depends on the content of the condition or merely on its presence, with the model of Table 1 on the same test set. The condition is manipulated at test time without retraining. Replacing the geometry with the "no information" value everywhere raises the error to that of the model without geometry on RLP and well above it on DoFP and RSP ($29.9°$ and $30.4°$, against $18.4°$ and $18.2°$), as expected for an input that the model never saw at that density. The geometry of another image is worse still on DoFP and RSP ($35.8°$ and $36.8°$), so the model follows the content of the prior rather than merely its presence. Zeroing the RGB image brings the error close to the chance level of $45°$ ($40.6°$ to $47.7°$), so the geometry alone does not determine the angle, consistent with the branch-selection mechanism. Geometry constrains the direction up to the sign, and appearance is needed to select the sign and to set the magnitude.

H Branch choice relative to the prior

Section 5.2 counts an error above $67.5°$ as a wrong branch, which does not separate a prediction on the other branch of a correct azimuth from a prediction whose azimuth is wrong by about $90°$. Table 13 therefore classifies the ground truth and every prediction relative to the azimuth ψ of the prior, as within $22.5°$ of ψ, within $22.5°$ of $\psi + 90°$, or neither, on the pixels of the mask where the prior is valid. On the pixels whose ground truth lies on one of the two branches, a prediction on the other one is a branch swap in the sense of (3). The table also reports three rules that use the prior alone, the diffuse branch everywhere, the specular branch everywhere, and the better of the two per pixel, an oracle whose error is that of the prior azimuth itself.

On the benchmark, the prior lowers the branch swaps from 29.3% to 15.4% on RLP, from 6.0% to 1.7% on DoFP, and from 13.7% to 7.2% on RSP, so the errors near $90°$ of the model without geometry are mostly genuine branch swaps. GeoPolar is more accurate than both constant rules on every subset of the benchmark, and on DoFP even than the oracle ($11.1°$ against $16.6°$), so it refines the azimuth of the prior as well as choosing its branch.

Table 13: Branch choice relative to the prior. The table gives $\mathrm{AoLP_m}$ (°) on the pixels of the mask with a valid prior, and the share of branch swaps (%) on the pixels whose ground truth lies on one of the two branches of the prior azimuth ψ. The first three rows use the prior alone. **Bold** and underline mark the best and second best among the models below the rule.

	RLP		DoFP		RSP	
Rule or model	**$\mathrm{AoLP_m}\downarrow$**	**Swap↓**	**$\mathrm{AoLP_m}\downarrow$**	**Swap↓**	**$\mathrm{AoLP_m}\downarrow$**	**Swap↓**
diffuse branch, $\hat{\phi}=\psi$	44.3	–	68.1	–	62.5	–
specular branch, $\hat{\phi}=\psi+90°$	45.7	–	21.9	–	27.5	–
better branch per pixel (oracle)	15.8	–	16.6	–	8.1	–
PA, T=20	40.1	**14.1**	32.7	13.5	<u>18.8</u>	<u>12.9</u>
GenPolar	<u>33.5</u>	18.0	28.9	12.8	19.4	13.4
GeoPolar w/o geometry	39.7	29.3	<u>16.0</u>	<u>6.0</u>	19.9	13.7
GeoPolar (Ours)	**28.2**	<u>15.4</u>	**11.1**	**1.7**	**13.2**	**7.2**

Table 14: Learned confidence in the prior. The first row is the model of Table 1. The columns give $\mathrm{AoLP_m}$ and $\mathrm{AoLP_{hi}}$ in degrees and $\mathrm{DoLP_m}$, where $\mathrm{AoLP_{hi}}$ is the AoLP error on pixels whose brightness and ground-truth DoLP are at least 0.3. The last two rows change the confidence model only at test time, and the better of the two trained models is in bold in each column. The predicted confidence separates correct from wrong geometry with an AUC of 0.80 (RLP), 0.88 (DoFP), and 0.88 (RSP).

	RLP			DoFP			RSP		
Model	**$\mathrm{AoLP_m}\downarrow$**	**$\mathrm{AoLP_{hi}}\downarrow$**	**$\mathrm{DoLP_m}\downarrow$**	**$\mathrm{AoLP_m}\downarrow$**	**$\mathrm{AoLP_{hi}}\downarrow$**	**$\mathrm{DoLP_m}\downarrow$**	**$\mathrm{AoLP_m}\downarrow$**	**$\mathrm{AoLP_{hi}}\downarrow$**	**$\mathrm{DoLP_m}\downarrow$**
w/o confidence	**28.2**	18.7	0.105	**11.7**	**9.0**	0.055	**13.4**	15.5	**0.065**
with confidence	28.3	**14.6**	**0.102**	13.1	10.6	**0.054**	13.9	**11.6**	0.069
test time: gate off (r=1)	28.0	14.3	0.105	13.3	9.5	0.054	14.1	11.4	0.069
test time: no geometry	32.8	36.1	0.091	21.6	26.4	0.058	20.2	13.2	0.068

I A LEARNED CONFIDENCE IN THE PRIOR

The prior is wrong where the geometry model describes a reflected or transmitted scene, or where the surface faces the camera and its azimuth is undefined. We test whether the generator can learn where to trust it. A small convolutional head with 53k parameters predicts from the RGB image and $\mathbf{g}$ a confidence $r \in [0,1]$ per pixel. The confidence scales the two azimuth channels of the condition and is appended to it as a seventh channel. The head is supervised only during training, by whether the measured AoLP agrees with the azimuth up to the branch, with agreement within $15°$ as positive and a distance beyond $30°$ as negative, through a binary cross-entropy term of weight 1. During training the prior is also replaced by the "no information" value for 10% of the images and, with probability 0.5, in the regions where it disagrees with the measurement, so that the network learns to generate without it. At test time the head needs no ground truth.

The model without confidence in Table 14 is the model of Table 1. The model with confidence is initialized from an intermediate checkpoint of it and trained for the rest of its schedule with the head, the gate, and the random removal of the prior.

Over the polarization mask the confidence does not lower the error, since $\mathrm{AoLP_m}$ is unchanged on RLP and rises by $0.5°$ on RSP and by $1.4°$ on DoFP. Its effect appears on the strongly polarized pixels, those with a brightness and a ground-truth DoLP of at least 0.3, whose error $\mathrm{AoLP_{hi}}$ it lowers from $18.7°$ to $14.6°$ on RLP and from $15.5°$ to $11.6°$ on RSP, while it raises it on DoFP from $9.0°$ to $10.6°$. Setting $r = 1$ at test time changes the errors by about one degree or less, so the gain comes from training with the head and with the prior withheld rather than from the gate at inference. Removing the prior at test time raises $\mathrm{AoLP_{hi}}$ to $36.1°$ on RLP and $26.4°$ on DoFP, so the model still relies on the prior where it is available. Figure 11 shows scenes where the confidence is low on glass and on surfaces facing the camera and the error decreases, and a scene in its last row where the model with confidence is less accurate although it trusts the prior. Because the gain is limited to the strongly polarized pixels of two subsets, where the normal of the geometry model is least reliable, and the mask error does not decrease, the main results use the model without confidence.

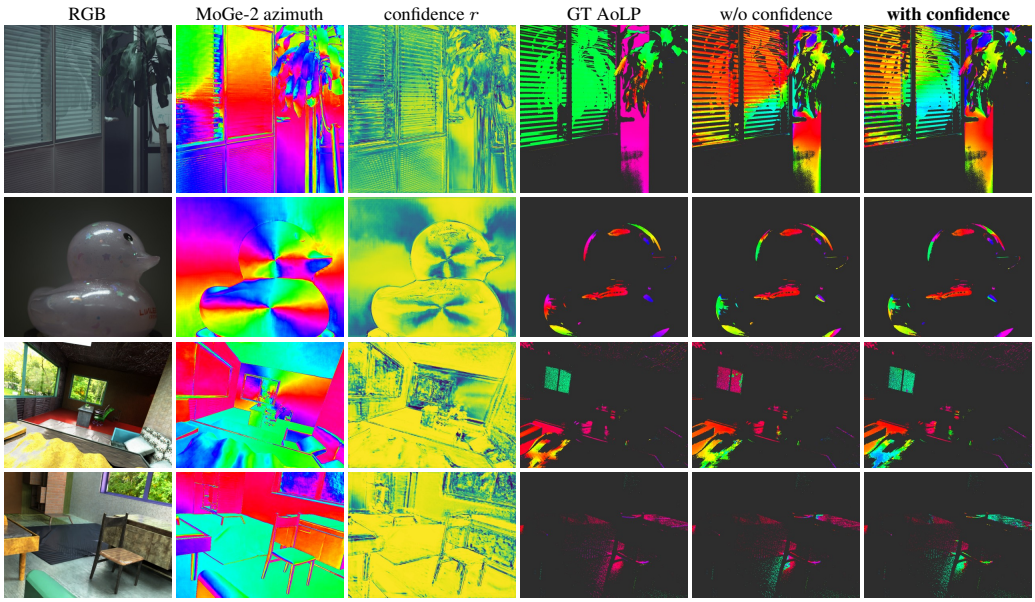


Figure 11: Learned confidence in the prior. The columns show the confidence r, from dark for distrust to yellow for trust and gray for an invalid prior, and the AoLP of the same network without and with the confidence gate. The AoLP is masked where the ground-truth DoLP or the brightness is below 0.1.

J ADDITIONAL QUALITATIVE RESULTS

Figure 12 compares all methods on two scenes of each captured subset, chosen as scenes on which GEOPOLAR is more accurate than every baseline on both the masked AoLP and the masked DoLP error.

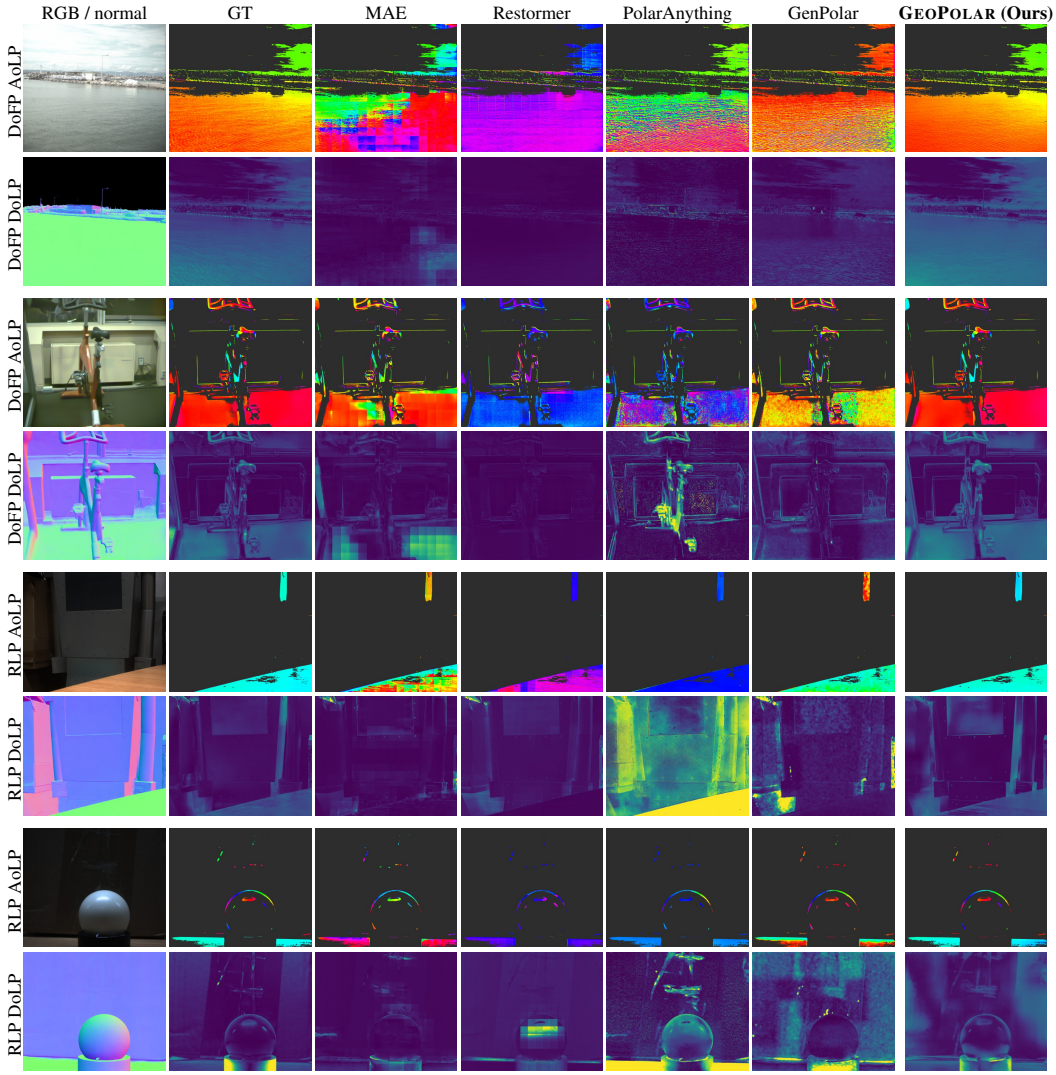


Figure 12: Additional comparisons on the combined benchmark for two scenes of each captured subset. For each scene, the upper row shows the AoLP as hue, masked where the brightness or the ground-truth DoLP is below 0.1, and the lower row the DoLP from 0 (dark blue) to 0.5 (yellow). The first column shows the RGB input and the MoGe-2 normal from which the prior is computed.