

Poppy: Polarization-based Plug-and-Play Guidance for Enhancing Monocular Normal Estimation

Irene Kim¹, Sai Tanmay Reddy Chakkerla¹, Alexandros Graikos¹, Dimitris Samaras¹, and Akshat Dave¹

Stony Brook University, Stony Brook NY 11794, USA
 {irnkim,schakkerla,agraikos,samaras,dave}@cs.stonybrook.edu

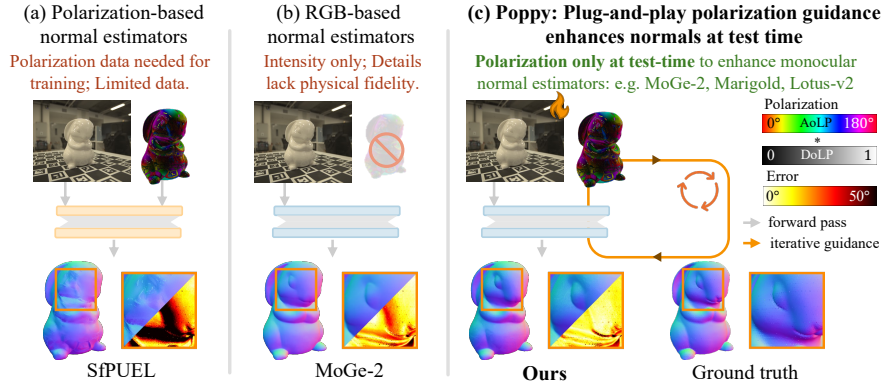


Fig. 1: Test-time polarization guidance to enhance normal estimation. (a) Polarization-based feed-forward models have limited generalizability due to the scarcity of polarization-normal training pairs. (b) RGB-only monocular normal estimators produce oversmoothed or hallucinated details on challenging surfaces – normals of the textureless bunny object appear flatter than ground truth. (c) Poppy introduces polarization guidance into pretrained RGB-only models at test time – improving normal accuracy without retraining.

Abstract. Monocular surface normal estimators trained on large-scale RGB-normal data often perform poorly in the edge cases of reflective, textureless, and dark surfaces. Polarization encodes surface orientation independently of texture and albedo, offering a physics-based complement for these cases. Existing polarization methods, however, require multi-view capture or specialized training data, limiting generalization. We introduce Poppy, a training-free framework that refines normals from any frozen RGB backbone using single-shot polarization measurements at test time. Keeping backbone weights frozen, Poppy optimizes per-pixel offsets to the input RGB and output normal along with a learned

reflectance decomposition. A differentiable rendering layer converts the refined normals into polarization predictions and penalizes mismatches with the observed signal. Across seven benchmarks and three backbone architectures (diffusion, flow, and feed-forward), Poppy reduces mean angular error by 23–26% on synthetic data and 6–16% on real data. These results show that guiding learned RGB-based normal estimators with polarization cues at test time refines normals on challenging surfaces without retraining. An interactive demo and code are available at <https://irnkim.github.io/poppy/>.

Keywords: surface normal estimation · polarization imaging · shape from polarization · test-time guidance · physics-based guidance

1 Introduction

Surface normal estimation is a core problem in computer vision, with applications in robotics, augmented reality, scene understanding, and 3D reconstruction. Monocular normal estimation aims to recover per-pixel 3D surface orientation from a single RGB image – an inherently ill-posed task because many surface normals can produce the same 2D appearance. Modern learning-based normal estimators reduce this ambiguity by learning geometric priors from large-scale paired RGB and normal datasets [6, 25, 32, 49, 50, 53].

However, state-of-the-art monocular estimators fail on three common surface types where RGB cues alone are unreliable: 1) highly reflective surfaces, where view-dependent highlights are misinterpreted as geometry; 2) textureless regions, which lack spatial variation and produce oversmoothed normals; and 3) dark objects, where low signal-to-noise ratio degrades predictions. RGB-normal datasets often under-represent these surfaces, compounding these errors (Fig. 1(b)). Thus, in this paper, in addition to RGB, we capture the polarization of light, only at test time, to refine normal estimation for these challenging scenarios.

Polarization provides a physically grounded complement to RGB. When unpolarized light reflects from a surface, it becomes partially polarized depending on surface orientation and material properties – making polarization informative for the failure cases above. Classical shape-from-polarization (SfP), however, is ill-posed: the azimuthal flip (π) ambiguity and the diffuse-specular ($\pi/2$) ambiguity together yield four candidate normals from a single measurement. Prior classical methods resolve these ambiguities by acquiring additional measurements – varying the illumination [1, 4, 17, 41], capturing multiple views [8, 15, 24, 35], or capturing additional modalities [27, 31] – increasing the capture burden beyond a single snapshot.

Learning-based SfP methods [5, 34, 38] aim to retain single-snapshot capture by leveraging data-driven priors learned from curated polarization-normal training pairs. These approaches, however, require paired polarimetric training data that is expensive to collect and narrow in scene diversity, which limits generalization to new materials and environments (Fig. 1(a)). Neither classical nor

learned SfP methods exploit the strong geometric priors already embedded in modern RGB normal estimators.

To bridge this gap, we introduce Poppy, a training-free framework that guides any differentiable RGB normal estimator with physics-based polarization constraints at test time (Fig. 1(c)). The weights of the normal estimator backbone are kept frozen. Poppy instead introduces learnable parameters before and after the network so that the predicted normals align with the observed polarization signal. Poppy is plug-and-play: it applies to diffusion-based [32], flow-based [25], and feed-forward [50] backbones without retraining or architectural changes. The pretrained backbone initializes normals using global scene structure from RGB as orientation constraints, resolving the azimuthal ambiguity; polarization guidance then corrects normal estimation errors that RGB cues alone miss.

Concretely, Poppy adds a learnable per-pixel image offset to the network input and a learnable per-pixel normal offset to the network output. The image offset acts as a global perturbation: input changes propagate through the backbone, steering the predicted geometry at a scene-wide scale. The normal offset refines high-frequency details that the backbone misses. A differentiable polarization rendering layer converts the refined normals and a learned specular radiance map into Stokes vectors, a representation of the polarization state, resolving the remaining diffuse-specular ambiguity. The loss between the predicted and measured Stokes vectors is minimized by optimizing the image offset, normal offset, and specular radiance maps.

We evaluate Poppy extensively across seven benchmarks – SfPUEL [38], NeRSP [24], NeISF [35], and DeepPol [16] (synthetic); SfPUEL, NeRSP, and PISR [8] (real) – covering nearly all publicly available polarization-normal datasets. Poppy enables consistent improvements in these benchmarks across three backbone architectures: Marigold [32], Lotus-v2 [25], and MoGe-2 [50] – with 23–26% reduction in mean angular error (MAE) for synthetic data and 6–16% reduction on real data. Poppy demonstrates improved normal quality on challenging materials, including reflective, textureless, and low albedo surfaces. The refined normals also improve downstream 3D mesh quality when used to aid an existing multi-view reconstruction method. The learned per-pixel specular radiance enables decomposition of the observed Stokes into diffuse and specular components.

Our contributions are:

1. A training-free guidance framework that refines normals from RGB-based monocular estimators using polarization measurements at test time.
2. A plug-and-play guidance mechanism with learnable per-pixel image and normal offsets for global-then-local refinement of any differentiable backbone, together with a learned per-pixel specular radiance for handling diffuse-specular ambiguities.
3. Consistent improvements across three backbone architectures on seven benchmarks, demonstrating gains on reflective, textureless, and low-SNR surfaces that are typically challenging for normal estimation.

2 Related work

Monocular surface normal estimation methods predict per-pixel surface orientation from a single RGB image. Early discriminative approaches [18, 20, 42] train convolutional networks on large-scale supervision, with Omnidata [19] showing the benefit of multi-task pretraining. More recent discriminative methods incorporate stronger priors: DSINE [6] encodes per-pixel camera geometry, and MoGe [49, 50] regresses affine-invariant point maps via a transformer backbone. Diffusion-based methods [21, 22, 32, 37] repurpose pretrained generative models for geometry estimation, with Marigold [32], Lotus [26], StableNormal [53], and GenPercept [51] among the most prominent. Lotus-v2 [25] extends this to flow-matching with a detail-sharpening step. Despite strong average-case performance, all of these methods rely solely on RGB appearance cues and degrade on reflective, textureless, and low-albedo surfaces where photometric shading signals are ambiguous. Our work targets precisely these failure cases by introducing polarization-based physical constraints at test time.

Shape from Polarization methods recover surface normals from reflected light using diffuse and height-from-polarization models [2, 45, 47], specular and refractive cues [27, 31, 39, 40], or joint shading–polarization formulations [1, 4, 28, 41]. However, the polarization-to-normal mapping is inherently ambiguous due to π and $\pi/2$ symmetries in the Fresnel equations. Multi-view methods [3, 8, 13, 24, 35, 52] resolve this geometrically, while discriminative methods such as DeepSfP [5] and SfPUEL [38] learn to predict normals directly from polarization images. All of these approaches either require controlled capture, curated polarization training data, or multi-image acquisition. Most closely related to our work, PPFT [29] adapts a pretrained depth model to polarization inputs via prompt fusion tuning – requiring polarization training data and modifying model weights; in contrast, Poppy targets normals and operates entirely at test time without retraining.

Test-time guidance steers pretrained models at inference time rather than training specialized models on new data. For diffusion models, guidance methods back-propagate a task-specific likelihood during the reverse diffusion to steer generation towards measurement-consistent solutions [11], handling both linear and nonlinear inverse problems. More recent approaches have extended this to training-free energy guidance under arbitrary constraints [54], and faster, backpropagation-free sampling [23]. Guidance has been extensively used in depth-completion methods [30, 33, 46, 48], where sparse depth measurements or physics-based cues are used to guide pretrained monocular depth estimators at test time, as well as in medical image segmentation [10], where image-specific visual prompts are optimized at inference to bridge cross-domain distribution shifts without modifying model weights. Among the depth methods, Marigold-DC [48] is closest to our approach: it guides a diffusion-based depth model with sparse depth observations, whereas Poppy guides a normal estimation model with dense polarization measurements that encode shape through the Fresnel equations.

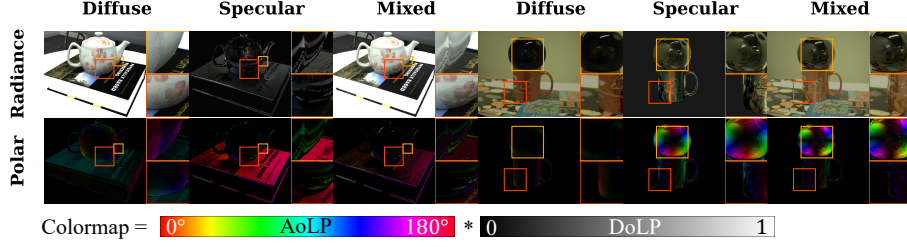


Fig. 2: Radiance decomposition. The mixed radiance S_0 is decomposed into diffuse radiance L_d and specular radiance L_s from our method. The learned L_s captures specular highlights and environment-dependent reflections (scaled $2\times$ for clarity), while L_d retains the object’s intrinsic diffuse shading and texture. From these radiance components and the predicted normals, the polarization maps (AoLP $\times$ DoLP) of the diffuse, specular, and combined components can be obtained.

3 Background

3.1 Representing polarization

Polarization characterizes the oscillation of light waves. While natural illumination is often unpolarized, reflection from surfaces induces partial polarization that depends on surface orientations and material properties. This makes polarization a physically grounded cue for geometry estimation.

Stokes vector. The polarization state of a light ray can be represented by the Stokes vector, $\mathbf{S} = [S_0, S_1, S_2, S_3]^\top$ [12]. S_0 represents total intensity. S_1 measures the difference between horizontally and vertically polarized components. S_2 measures the difference between components polarized at 45° and 135° . S_3 represents circular polarization.

A linear polarization camera captures intensity (I) through polarizers at multiple angles ($0/45/90/135^\circ$). The linear Stokes parameters are computed as

$$S_0 = I_0 + I_{90}, \quad S_1 = I_0 - I_{90}, \quad S_2 = I_{45} - I_{135}. \quad (1)$$

In most passive imaging scenarios, circular polarization is negligible compared to linear polarization. We therefore restrict our formulation to linear polarization and use the reduced representation, $\mathbf{S} = [S_0, S_1, S_2]^\top$.

Degree and Angle of Polarization. The Degree of Linear Polarization (DoLP), denoted ρ , measures the fraction of light that is linearly polarized – ranging from 0, for unpolarized light, to 1, for fully polarized light. DoLP can be obtained from the Stokes vector as $\rho = \frac{\sqrt{S_1^2 + S_2^2}}{S_0}$. The Angle of Linear Polarization (AoLP), denoted ϕ , describes the dominant polarization orientation. AoLP depends on the Stokes vector as $\phi = \frac{1}{2} \arctan(S_2, S_1)$. Together, (ρ, ϕ) provide a compact representation of the linear polarization state.

Diffuse and specular reflectance. When light reflects from a surface, it can be decomposed into diffuse and specular components. We provide an example decomposition learned from our method for visualization in Fig. 2. Diffuse reflection arises from subsurface scattering and typically produces weaker polarization, whereas specular reflection arises from surface reflection and follows Fresnel reflection laws [7]. Let L_d and L_s denote the diffuse and specular radiance components. The observed Stokes vector is modeled as the sum of diffuse and specular contributions: $\mathbf{S} = S_d + S_s$. Using the DoLP and AoLP of each component, the combined Stokes parameters can be written as [15, 36]

$$S_0 = L_d + L_s \quad (2a)$$

$$S_1 = L_d \rho_d \cos(2\phi_d) + L_s \rho_s \cos(2\phi_s) \quad (2b)$$

$$S_2 = L_d \rho_d \sin(2\phi_d) + L_s \rho_s \sin(2\phi_s), \quad (2c)$$

where $(\rho_{d/s}, \phi_{d/s})$ represent diffuse/specular DoLP, AoLP respectively.

3.2 Surface normals to Stokes

Polarization provides a cue for surface orientation because both DoLP and AoLP depend on the surface normal relative to the viewing direction.

Spherical coordinate conversion. Denoting the surface normal as $n = [n_x, n_y, n_z]^\top$, we convert it into spherical coordinates (θ, ψ) relative to the viewing direction v , where the elevation angle $\theta = \arccos(n \cdot v)$ and azimuth angle $\psi = \arctan(n_y, n_x)$.

DoLP from surface normals. Under Fresnel reflection theory, the degree of polarization depends on the elevation angle θ and the refractive index η (set to 1.5 per standard assumptions).

For diffuse reflection, $\rho_d(\theta) = \frac{(\eta-1/\eta)^2 \sin^2 \theta}{2+2\eta^2-(\eta+1/\eta)^2 \sin^2 \theta + 4 \cos \theta \sqrt{\eta^2 - \sin^2 \theta}}$, and for specular reflection, $\rho_s(\theta) = \frac{2 \sin^2 \theta \cos \theta \sqrt{\eta^2 - \sin^2 \theta}}{\eta^2 - \sin^2 \theta - \eta^2 \sin^2 \theta + 2 \sin^4 \theta}$.

AoLP from surface normals. The angle of polarization is determined by the azimuth angle ψ . For diffuse reflections, $\phi_d = \psi$, and for specular reflections, $\phi_s = \psi + \pi/2$. The orthogonality between these components reflects the phase difference induced by Fresnel reflection.

Stokes from surface normals. Normal-to-Stokes conversion is an inverse problem of SfP. Given a surface normal n and the specular radiance L_s , we compute the diffuse and specular DoLP/AoLP $(\rho_{d/s}, \phi_{d/s})$ and diffuse radiance $L_d = S_0 - L_s$. We can rewrite Eq. 2 in terms of n and L_s as $\hat{\mathbf{S}} = \mathcal{F}(n, L_s)$.

Ambiguities in Shape from Polarization. Shape from Polarization suffers from intrinsic ambiguities. AoLP is invariant under a 180° rotation, leading to a π ambiguity in azimuth. Additionally, diffuse and specular reflections produce orthogonal polarization orientations, resulting in a $\pi/2$ ambiguity when the dominant reflection component is unknown. These ambiguities motivate the integration of additional geometric priors using RGB cues, as we propose in our method.

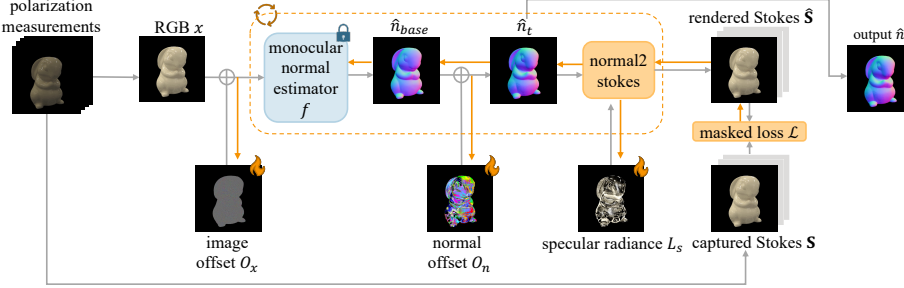


Fig. 3: Poppy pipeline. Given polarization measurements, we compute the observed Stokes vector $\mathbf{S}$, extract the RGB image x , and add learnable image offset O_x to x . A frozen backbone produces base normals $\hat{n}_{\text{base}}$; a learnable normal offset O_n yields the refined estimate $\hat{n}_t = \hat{n}_{\text{base}} + O_n$. Using Fresnel equations, the predicted Stokes vector $\hat{\mathbf{S}}$ is computed from $\hat{n}_t$ and a learnable specular radiance map L_s . We minimize the polarization consistency loss between $\hat{\mathbf{S}}$ and $\mathbf{S}$ to update the image offset O_x , normal offset O_n , and specular map L_s over T steps, while keeping backbone weights fixed.

3.3 Monocular normal estimators

A pretrained monocular estimator f predicts the surface normal as $n = f(x)$, by leveraging the rich visual priors learned from pretraining on large-scale datasets. Deep-learning models span diverse architectures: iterative diffusion-based models (*e.g.*, Marigold [32]), iterative flow-based models (*e.g.*, Lotus-v2 [25]), feed-forward transformer models (*e.g.*, MoGe-2 [50]).

4 Method

Given polarization measurements captured by a linear polarization camera, we recover surface normals that are consistent with physical light transport while preserving the strong geometric priors learned by modern RGB-based normal estimation networks. Instead of retraining a model, we formulate normal recovery as a test-time guidance problem. During the guidance, a differentiable rendering model enforces polarization constraints.

4.1 Rendering Stokes from monocular normals

Given the input polarization measurements $I_{\{0,45,90,135\}^\circ}$, we first compute the observed Stokes vector $\mathbf{S} = [S_0, S_1, S_2]^\top$ via Eq. 1, where the unpolarized intensity S_0 serves as the RGB input x to our pipeline.

A pretrained RGB normal estimator f takes x and produces a base surface normal estimate $\hat{n}_{\text{base}} = f(x)$. To synthesize the predicted Stokes vector $\hat{\mathbf{S}}$ from a given normal $\hat{n}_{\text{base}}$, we require the specular radiance L_s , following the procedure described in Sec. 3.2. In principle, L_s can be optimized iteratively via polarization-guided refinement until the residual $|\mathbf{S} - \hat{\mathbf{S}}|$ vanishes. However, when

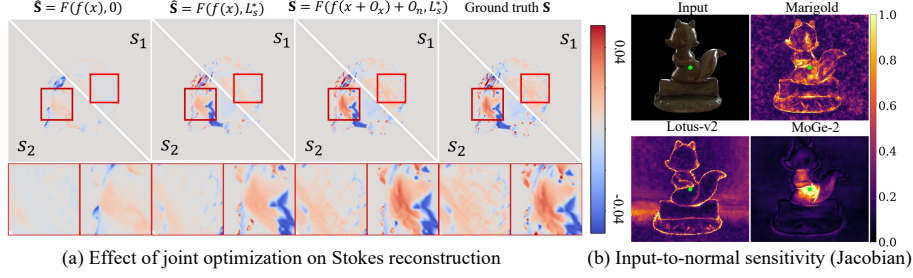


Fig. 4: (a) Stokes reconstruction of hedgehog scene (from NeRSP) when: $L_s = 0$ (diffuse only) in column 1; only L_s is learned from backbone predicted normals $\hat{n}_{\text{base}}$ in column 2; L_s , O_x , and O_n are jointly learned in column 3. (b) Jacobian magnitude maps for a selected input pixel (green dot), normalized by the 99th percentile for different backbones, showing how perturbations of a single input pixel influence the output normal map at a global level, across spatial locations.

f produces an erroneous normal estimate in geometrically challenging regions, prolonged optimization of L_s alone leads to either a persistent mismatch between the observed Stokes $\mathbf{S}$ and predicted Stokes $\hat{\mathbf{S}} = F(f(x), L_s)$, or overfitting of L_s to the incorrect normal (second column of Fig. 4(a)) – neither of which corrects the underlying geometric error.

This failure mode stems from the fact that the error originates in $\hat{n}_{\text{base}}$. To address this, we seek a test-time guidance mechanism that corrects the normal estimate directly, without modifying frozen model weights, and ensures compatibility with any pretrained monocular RGB normal estimator in a plug-and-play fashion, independent of its architecture.

4.2 Polarization-guidance parameters

To correct the erroneous normal $\hat{n}_{\text{base}}$ while keeping the model weights frozen, we introduce learnable parameters applied before and after f . Formally, the refined normal at step t is expressed as $\hat{n}_t = g_a(f(g_b(x)))$, where g_b and g_a denote arbitrary functions applied to the input and output of f , respectively. We adopt the simplest instantiation, $g_b(x) = x + O_x$ and $g_a(\hat{n}_{\text{base}}) = \hat{n}_{\text{base}} + O_n$, yielding $\hat{n}_t = f(x + O_x) + O_n$, where $O_x \in \mathbb{R}^{H \times W \times 3}$ and $O_n \in \mathbb{R}^{H \times W \times 3}$ are per-pixel offsets applied to the input image x and the predicted normal $\hat{n}_{\text{base}}$, respectively.

Global Guidance. The image offset O_x is motivated by the observation that even a small perturbation in input space can produce a globally significant correction in the output normals. This behavior is explained by the Jacobian of different models f , visualized in Fig. 4(b), which shows that a change in a single input pixel has a broad influence over the output normal map (details in Supp. Sec. A). Consequently, the image offset O_x serves as an efficient mechanism for globally steering the predicted geometry toward a polarization-consistent solution.

Table 1: Aggregated normal estimation performance on real and synthetic benchmarks. Our polarization guidance (+ Poppy) consistently reduces MAE and RMSE across RGB backbones while it improves thresholded accuracy, confirming gains in both global surface orientations and fine-scale local geometric details. +Poppy denotes joint optimization of the image offset and normal offset. Best result per backbone within each data split is highlighted in bold.

Method	Real						Synthetic					
	Mean	Median	RMSE	Acc11.25	Acc22.5	Acc30	Mean	Median	RMSE	Acc11.25	Acc22.5	Acc30
SFPUEL [38]	20.68	17.08	26.15	0.37	0.70	0.80	17.03	12.91	23.33	0.46	0.79	0.87
DSINE [6]	23.06	19.07	28.83	0.27	0.62	0.76	24.13	19.94	30.00	0.26	0.60	0.75
Lotus [26]	16.69	13.96	21.70	0.42	0.78	0.88	19.99	17.04	25.06	0.30	0.69	0.84
StableNormal [53]	17.80	14.77	23.29	0.41	0.75	0.86	18.45	14.65	24.55	0.39	0.75	0.86
Marigold [32]	18.18	15.25	23.30	0.36	0.74	0.86	20.99	17.40	26.43	0.30	0.68	0.82
Marigold + Poppy	15.28	12.57	20.26	0.47	0.83	0.92	15.60	11.89	22.08	0.56	0.82	0.89
Lotus-v2 [25]	14.68	12.05	19.51	0.49	0.85	0.93	16.52	13.44	22.08	0.42	0.81	0.90
Lotus-v2 + Poppy	12.65	10.05	17.62	0.61	0.89	0.94	12.26	8.81	18.70	0.66	0.89	0.93
MoGe-2 [50]	13.10	10.51	18.11	0.57	0.87	0.94	14.13	11.33	20.01	0.51	0.87	0.94
MoGe-2 + Poppy	12.26	9.88	16.99	0.61	0.90	0.95	10.89	8.05	16.41	0.70	0.92	0.96

Local Guidance. The normal offset O_n , applied directly to the output, addresses the tendency of monocular normal estimators to produce smooth predictions. The normal offset O_n recovers high-frequency surface details that f fails to capture. Due to its direct coupling with the normal output, the normal offset O_n exhibits high sensitivity to the polarization loss. We therefore defer its optimization until L_s has converged sufficiently to yield a reliable estimate of Stokes $\hat{\mathbf{S}} = \mathcal{F}(f(x + O_x) + O_n, L_s)$ (third column in Fig. 4(a)). We assume orthographic projection and ignore viewing directions, an approximation that introduces negligible error unless the object is captured at close range.

4.3 Guidance objective and updates

During test-time guidance, we convert the refined normal $\hat{n}_t$ into Stokes $\hat{\mathbf{S}}$ for polarization consistency check. We optimize the learnable parameters $\Theta = \{L_s, O_x, O_n\}$ iteratively over T steps to minimize the discrepancy between the observed and predicted Stokes vectors. At each step t , the backbone f takes the offset input $x + O_x$ and produces the base normal $\hat{n}_{\text{base}} = f(x + O_x)$. The normal offset O_n is then applied to yield the refined normal $\hat{n}_t = \hat{n}_{\text{base}} + O_n$. Given $\hat{n}_t$ and the specular radiance L_s , the predicted Stokes vector $\hat{\mathbf{S}}(\hat{n}_t, L_s)$ is computed following Sec. 3.2. The polarization consistency loss is defined as $\mathcal{L} = \sum_p M(p) \sum_{i=0}^2 |S_i(p) - \hat{S}_i(p)|$, where p denotes a pixel and $M(p)$ is a validity mask that restricts the loss to physically meaningful measurements. Specifically, a pixel is included if it has a signal ($S_0 > 0.01$), is not saturated ($S_0 < 1$), and satisfies the physical Stokes constraint ($S_1^2 + S_2^2 \leq S_0^2$).

Global-then-local guidance. The three parameters are optimized in a staged schedule, each with a distinct learning rate. In the early phase ($t < 50$), we

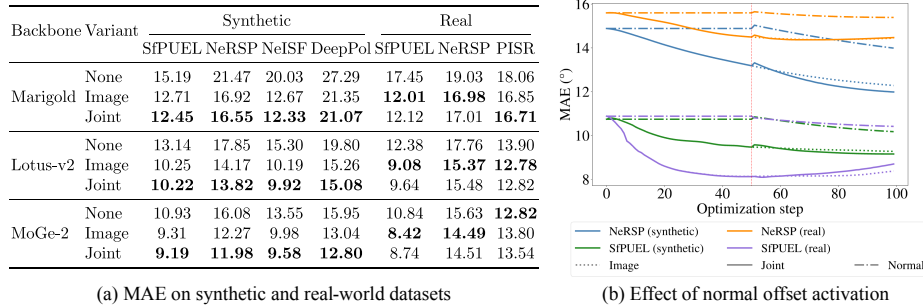


Fig. 5: Ablation of guidance variants on real and synthetic datasets. (a) MAE across three backbones with no guidance (None), image offset (Image), and joint image and normal offsets (Joint); guidance improves normals, with joint guidance performing slightly better on synthetic data. (b) MAE over optimization steps (MoGe-2 backbone); the red-dashed line at $t=50$ marks normal offset activation. On synthetic data, joint guidance accelerates error reduction. On real data, sensor noise slightly increases MAE though fine surface detail is still recovered.

only update L_s and O_x , allowing the specular radiance to converge and the predicted normals to reach global geometric plausibility. At $t = 50$, O_n is introduced to refine high-frequency surface details that the backbone fails to capture, and all three parameters are subsequently updated until $t = T$.

The learnable parameters $\Theta = \{L_s, O_x, O_n\}$ are updated with an optimizer using gradient descent $\Theta^{(k+1)} = \Theta^{(k)} - \lambda_{\Theta} \nabla_{\Theta} \mathcal{L}$. After T iterations, the output surface normal is $\hat{n} = \frac{f(x+O_x)+O_n}{\|f(x+O_x)+O_n\|_2}$.

The learned specular radiance L_s and the refined output normal $\hat{n}$ together synthesize the predicted Stokes vector $\hat{\mathbf{S}}$ that closely matches the observed Stokes vector $\mathbf{S}$ as shown in the third column in Fig. 4(a). Consequently, minimizing this discrepancy drives the predicted normal $\hat{n}$ toward the true surface normal.

5 Results

5.1 Implementation details

Datasets. We evaluate on 14 real objects across 3 benchmarks – SfPUEL [38] (6; highly specular), NeRSP [24] (4; high reflectance, low illumination), PISR [8] (4; highly reflective, textureless) – and 18 synthetic objects across 4 benchmarks – SfPUEL (7; low SNR), NeRSP (4), NeISF [35] (2; moderate reflectance), DeepPol [16] (5; mixed diffuse and specular) – covering nearly all publicly available paired polarization–normal datasets. Real data require 3D scanning, which often fails on challenging materials, limiting large-scale real benchmarks.

Baseline models. We compare against the polarization-based model SfPUEL [38] and RGB-based methods DSINE [6], StableNormal [53], Lotus [26], Marigold [32], Lotus-v2 [25], and MoGe-2 [50], all under their default configurations. Poppy is

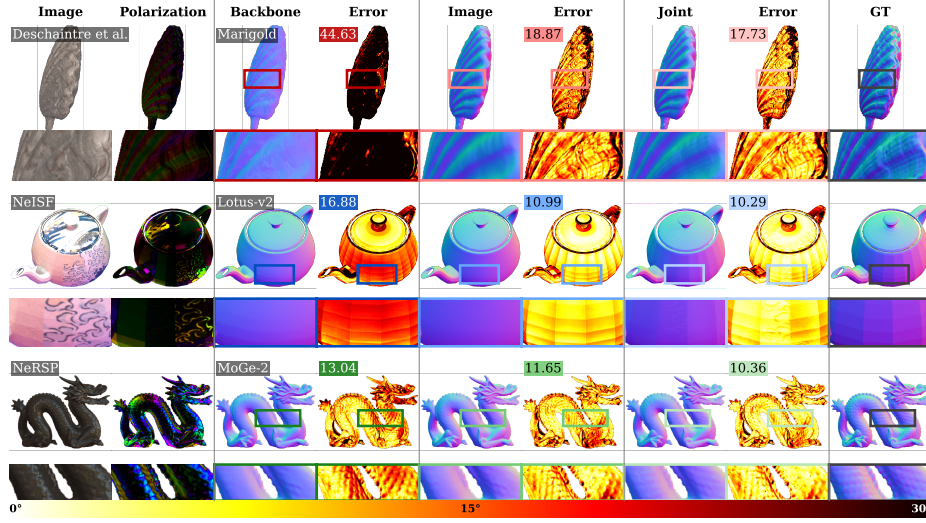


Fig. 6: Comparison of image (global) and joint (global-then-local) guidance on synthetic data. We compare three guidance conditions across three backbones: no guidance (Backbone), image offsets (Image), and joint image and normal offsets (Joint). Joint guidance recovers fine geometric structures – shell ridges, mesh facets, and dragon scales – that image-offset alone cannot resolve. Error maps (brighter is lower) and MAE shown; best viewed at high magnification.

applied to Marigold, Lotus-v2, and MoGe-2, representing diffusion-based, flow-based, and feed-forward backbones. We report mean angular error (MAE), median angular error (Median), angular RMSE, and accuracy under angular thresholds.

Optimization. We optimize with Adam for 100 steps. All three learnable parameters, O_x , O_n , and L_s , are initialized to 10^{-2} . Learning rates: $\lambda_{L_s} = 10^{-2}$ and $\lambda_{O_n} = 10^{-3}$ (activates at step 50 once L_s has converged) across all backbones; backbone-dependent λ_{O_x} : 10^{-3} (Marigold), 5×10^{-4} (Lotus-v2), 10^{-5} (MoGe-2).

Backbone inference. For Marigold, we use 4 denoising steps with 25 guidance steps each, without ensemble inference. For Lotus-v2, guidance precedes the detail-sharpening stage, followed by 10 sharpening iterations at default settings. For MoGe-2, gradients are backpropagated directly through the feed-forward pipeline. See Supp. Sec. C for per-backbone details.

Runtime and memory. At 768×768 resolution, backbone inference takes 3.00 s (MoGe-2), 4.02 s (Marigold), 3.33 s (Lotus-v2); Poppy adds 0.57 s, 0.95 s, and 1.73 s per step respectively, totaling 57 s, 95 s, and 173 s at 100 steps. Backbone-only inference requires 3.07, 7.47, and 38.14 GB at peak memory respectively; Poppy adds 12.26, 27.65, and 28.50 GB of overhead from retaining backbone

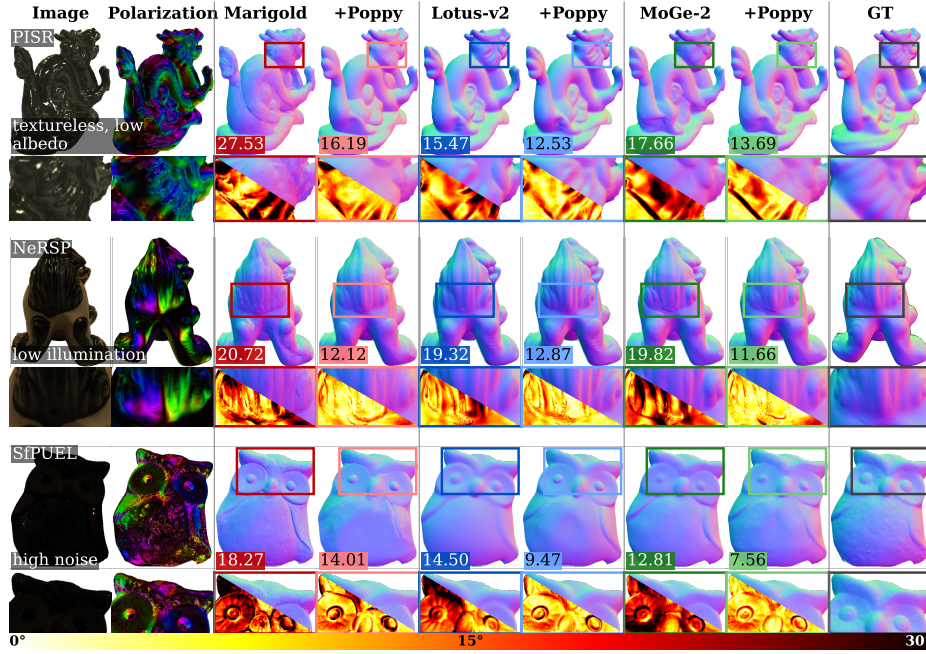


Fig. 7: Qualitative comparison on real-world scenes. Backbone predictions and Poppy results with pixel-wise error maps (brighter is lower) and MAE. Scenes span textureless, highly reflective, and low-SNR conditions (low illumination, low albedo, high noise); Poppy shows consistent improvement across all backbones.

activations for backpropagation and an fp16-to-fp32 precision switch required by the physics-based Stokes loss, while its own pixel-level offsets require only 21 MB. All run on NVIDIA RTX PRO 6000 Blackwell GPUs (95.6 GB VRAM).

5.2 Results and analysis

Quantitative comparisons with monocular normal baselines. Tab. 1 reports aggregated performance across all backbones and benchmarks. Adding polarization guidance (+Poppy) consistently improves MAE on both real and synthetic datasets, with MAE reductions of 6–26% and RMSE reductions of 6–18%, reflecting improved global surface orientation consistency. More notably, Acc11.25 (the fraction of pixels with angular error below 11.25°) improves by 7–31% on real data and 37–87% on synthetic data, demonstrating that Poppy not only corrects large-scale geometric errors but also recovers high-frequency surface details that RGB-based backbones systematically fail to capture.

Qualitative results on real-world scenes. To illustrate how these aggregate gains manifest in individual scenes, Fig. 7 presents qualitative comparisons on real-world scenes that are particularly challenging for RGB-based estimators. These

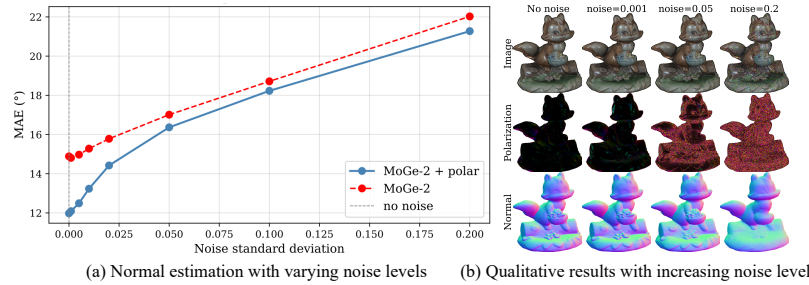


Fig. 8: Sensitivity to input noise. We add zero-mean Gaussian noise of increasing standard deviation σ to the NeRSP synthetic measurements. The plot reports mean angular error (MAE) for MoGe-2 with and without polarization guidance. Guidance improves accuracy at low noise levels, but the advantage diminishes with increasing σ .

scenes include textureless low-albedo surfaces, highly specular objects, and low-SNR conditions arising from low illumination, low albedo, or high measurement noise. Across all three backbone architectures and all challenging conditions, Poppy yields MAE reductions of 19–41% over the backbone predictions, with the largest improvements when RGB cues alone are most ambiguous.

Effect of global and local guidance. Fig. 5 and Fig. 6 analyze the individual contributions of the image offset O_x and the normal offset O_n . In Fig. 5(a), the image offset O_x alone accounts for the majority of the MAE reduction across all backbones and datasets, correcting large-scale geometric errors toward a globally polarization-consistent solution. Adding the normal offset O_n (Joint) yields a further reduction on synthetic data, though the additional gain is modest compared to the image offset alone; on real data, sensor noise causes it to slightly inflate MAE while still recovering fine surface detail qualitatively (Fig. 5(b)). Across all backbones and datasets, Poppy outperforms the backbone baseline, with the sole exception of MoGe-2 on PISR, which we attribute to a large perspective distortion that our orthographic assumption does not model. See Supp. Sec. D for perspective-aware guidance incorporating field-of-view.

The per-step MAE curve in Fig. 5(b) is consistent with this global-then-local behavior: MAE decreases primarily during the early optimization phase under the image offset; at $t=50$, when the normal offset is activated on synthetic data, a further sharp drop occurs as the remaining errors are refined. The normal offset recovers fine-scale geometric structures that the backbone over-smooths, such as shell ridges, mesh facets, and dragon scales (zoomed insets in Fig. 6).

Robustness to sensor noise. We examine how Poppy behaves when the polarization input itself is corrupted. Fig. 8 reports MAE after adding zero-mean Gaussian noise of varying standard deviation σ to the NeRSP synthetic polarization measurements. At low σ , Poppy reduces MAE relative to the no-guidance prediction on MoGe-2. As σ grows, the guided result converges toward the unguided backbone prediction – noisy cues are effectively ignored rather than amplified.

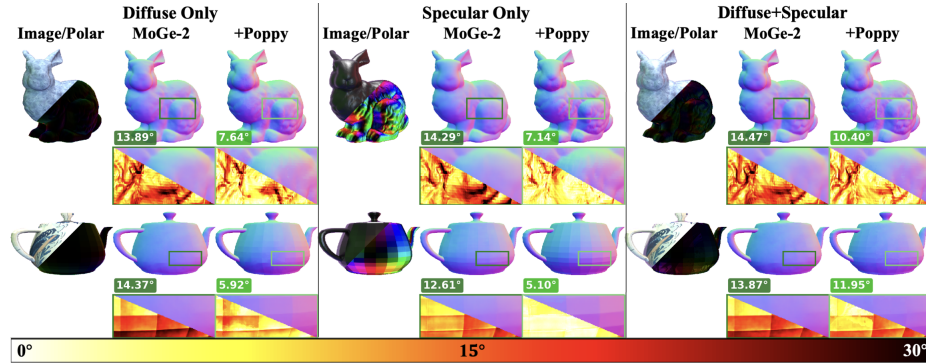


Fig. 9: Robustness to material reflectance properties. NeISF scenes with diffuse only, specular only, and mixed radiance. Specular only yields the largest improvement due to the strongest polarization signal, followed by diffuse; mixed gains are smaller due to the challenge of radiance decomposition. Error maps (brighter is lower) with MAE annotated.

Robustness to material properties. We isolate the effect of material reflectance using the NeISF synthetic dataset (Fig. 9), testing three conditions: diffuse only, specular only, and mixed reflectance. MoGe-2 without guidance achieves comparable MAE across all three conditions, but polarization guidance (+ Poppy) shows the most refinement on the specular case, where specular reflection produces stronger linear polarization. Diffuse surfaces yield the second-largest improvement, as even weaker polarization from diffuse radiance carries useful orientation information. In the mixed case, accurate diffuse-specular separation becomes a prerequisite; errors in this decomposition propagate into the polarization guidance, resulting in smaller but still meaningful gains.

5.3 Applications

Appearance decomposition. Beyond refining normals, Poppy yields a physics-based decomposition of captured appearance (S_0) into diffuse and specular components (L_d and L_s), as shown in Fig. 2 (top row). We provide image editing applications in Supp. Sec. J. From the decomposed appearance and refined normals, our forward model synthesizes the polarization properties of diffuse and specular components (Fig. 2 bottom row).

3D mesh reconstruction. In Fig. 10, we test our improved normals on downstream mesh reconstruction tasks. We use VCR-GauS [9] as our baseline to construct surface meshes, guided by multi-view RGB images and monocular surface normal estimates on those views. We observe an improvement in reconstructed mesh accuracy with our surface normals and an $\approx 6\%$ improvement in Chamfer distance as compared to RGB-only backbones. Note that we attempted using Gaussian Surfels [14]. However, this approach failed due to poor initialization points retrieved by COLMAP [43, 44] on glossy, textureless surfaces.

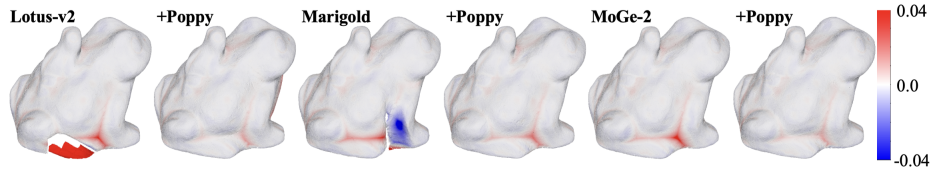


Fig. 10: Mesh reconstruction with refined surface normals. We reconstruct meshes using VCR-GauS [9] from multi-view images and corresponding normals estimated with and without polarization guidance. Polarization-refined normals yield better mesh quality across the three backbones. Signed distance to the ground-truth mesh is visualized for each reconstruction.

6 Conclusion

We introduced Poppy, a training-free framework that guides frozen RGB normal estimators with physics-based polarization constraints at test time, refining normals on reflective, textureless, and dark surfaces without polarimetric training data or backbone retraining. A learnable image offset propagates global corrections through the backbone; a normal offset recovers high-frequency local surface detail; and a learned specular radiance map resolves diffuse-specular ambiguity in single-view polarization. Because guidance parameters lie entirely outside the backbone, Poppy applies to diffusion-based [32], flow-based [25], and feed-forward [50] backbones in a plug-and-play manner. Across seven benchmarks, Poppy reduces MAE by 23–26% on synthetic data and 6–16% on real scenes. Metallic surfaces, while requiring a complex refractive index, remain applicable in practice: varying η yields negligible difference in MAE (Supp. Sec. E).

Although we focus on showing that polarization can guide any frozen monocular estimator at test time without retraining, our framework opens several directions for future work. The iterative optimization requires multiple forward passes through the backbone; however, we expect that better optimization strategies can reduce the computational overhead. The learnable image and normal offsets are lightweight yet effective for normal refinement; replacing them with neural network adapters could further increase the representation capacity at the expense of requiring minimal polarization data for fine-tuning. Extending robustness to higher sensor noise levels, handling perspective distortions for close-range captures, and modeling more complex material types such as transparent and translucent surfaces are all within reach of the modular framework. We see this work as a step toward steering frozen foundation models with physics-based cues, available solely at test time.

Acknowledgements

Alexandros Graikos and Dimitris Samaras are supported by National Science Foundation (NSF) grant IIS-2212046.

References

1. Atkinson, G.A.: Polarisation photometric stereo. *Computer Vision and Image Understanding* **160**, 158–167 (2017)
2. Atkinson, G.A., Hancock, E.R.: Recovery of surface orientation from diffuse polarization. *IEEE transactions on image processing* **15**(6), 1653–1664 (2006)
3. Atkinson, G.A., Hancock, E.R.: Shape estimation using polarization and shading from two views. *IEEE transactions on pattern analysis and machine intelligence* **29**(11), 2001–2017 (2007)
4. Atkinson, G.A., Hancock, E.R.: Surface reconstruction using polarization and photometric stereo. In: *International conference on computer analysis of images and patterns*. pp. 466–473. Springer (2007)
5. Ba, Y., Gilbert, A., Wang, F., Yang, J., Chen, R., Wang, Y., Yan, L., Shi, B., Kadambi, A.: Deep shape from polarization. In: *European Conference on Computer Vision*. pp. 554–571. Springer (2020)
6. Bae, G., Davison, A.J.: Rethinking inductive biases for surface normal estimation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 9535–9545 (2024)
7. Baek, S.H., Jeon, D.S., Tong, X., Kim, M.H.: Simultaneous acquisition of polarimetric svbrdf and normals. *ACM Trans. Graph.* **37**(6), 268 (2018)
8. Chen, G., He, Y., He, L., Zhang, H.: PIsr: Polarimetric neural implicit surface reconstruction for textureless and specular objects. In: *European Conference on Computer Vision*. pp. 205–222. Springer (2024)
9. Chen, H., Wei, F., Li, C., Huang, T., Wang, Y., Lee, G.H.: Vcr-gaus: View consistent depth-normal regularizer for gaussian surface reconstruction. *Advances in Neural Information Processing Systems* **37**, 139725–139750 (2024)
10. Chen, Z., Pan, Y., Ye, Y., Lu, M., Xia, Y.: Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 11184–11193 (2024)
11. Chung, H., Kim, J., McCann, M.T., Klasky, M.L., Ye, J.C.: Diffusion posterior sampling for general noisy inverse problems. In: *International Conference on Learning Representations* (2023)
12. Collett, E.: *Field guide to polarization*, vol. 15. SPIE press Bellingham, Washington (2005)
13. Cui, Z., Gu, J., Shi, B., Tan, P., Kautz, J.: Polarimetric multi-view stereo. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 1558–1567 (2017)
14. Dai, P., Xu, J., Xie, W., Liu, X., Wang, H., Xu, W.: High-quality surface reconstruction using gaussian surfels. In: *ACM SIGGRAPH 2024 conference papers*. pp. 1–11 (2024)
15. Dave, A., Zhao, Y., Veeraraghavan, A.: Pandora: Polarization-aided neural decomposition of radiance. In: *European conference on computer vision*. pp. 538–556. Springer (2022)
16. Deschaintre, V., Lin, Y., Ghosh, A.: Deep polarization imaging for 3d shape and svbrdf acquisition. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 15567–15576 (2021)
17. Ding, Y., Ji, Y., Zhou, M., Kang, S.B., Ye, J.: Polarimetric helmholtz stereopsis. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. pp. 5037–5046 (2021)

18. Do, T., Vuong, K., Roumeliotis, S.I., Park, H.S.: Surface normal estimation of tilted images via spatial rectifier. In: European Conference on Computer Vision. pp. 265–280. Springer (2020)
19. Eftekhari, A., Sax, A., Malik, J., Zamir, A.: Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10786–10796 (2021)
20. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE international conference on computer vision. pp. 2650–2658 (2015)
21. Fu, X., Yin, W., Hu, M., Wang, K., Ma, Y., Tan, P., Shen, S., Lin, D., Long, X.: Geowizard: Unleashing the diffusion priors for 3d geometry estimation from a single image. In: European Conference on Computer Vision. pp. 241–258. Springer (2024)
22. Garcia, G.M., Zeid, K.A., Schmidt, C., De Geus, D., Hermans, A., Leibe, B.: Fine-tuning image-conditional diffusion models is easier than you think. In: Proceedings of the Winter Conference on Applications of Computer Vision. pp. 753–762 (2025)
23. Graikos, A., Jojic, N., Samaras, D.: Fast constrained sampling in pre-trained diffusion models. In: Advances in Neural Information Processing Systems (2025)
24. Han, Y., Guo, H., Fukai, K., Santo, H., Shi, B., Okura, F., Ma, Z., Jia, Y.: Nersp: Neural 3d reconstruction for reflective objects with sparse polarized images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11821–11830 (2024)
25. He, J., Li, H., Sheng, M., Chen, Y.C.: Lotus-2: Advancing geometric dense prediction with powerful image generative model. arXiv preprint arXiv:2512.01030 (2025)
26. He, J., Li, H., Yin, W., Liang, Y., Li, L., Zhou, K., Zhang, H., Liu, B., Chen, Y.C.: Lotus: Diffusion-based visual foundation model for high-quality dense prediction. In: International Conference on Learning Representations (2025)
27. Huynh, C.P., Robles-Kelly, A., Hancock, E.R.: Shape and refractive index from single-view spectro-polarimetric images. *International journal of computer vision* **101**(1), 64–94 (2013)
28. Ichikawa, T., Purri, M., Kawahara, R., Nobuhara, S., Dana, K., Nishino, K.: Shape from sky: Polarimetric normal recovery under the sky. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14832–14841 (2021)
29. Ikemura, K., Huang, Y., Heide, F., Zhang, Z., Chen, Q., Lei, C.: Robust depth enhancement via polarization prompt fusion tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20710–20720 (2024)
30. Jeong, C., Bae, I., Park, J.H., Jeon, H.G.: Test-time prompt tuning for zero-shot depth completion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9443–9454 (2025)
31. Kadambi, A., Taamazyan, V., Shi, B., Raskar, R.: Polarized 3d: High-quality depth sensing with polarization cues. In: Proceedings of the IEEE international conference on computer vision. pp. 3370–3378 (2015)
32. Ke, B., Obukhov, A., Huang, S., Metzger, N., Daudt, R.C., Schindler, K.: Repurposing diffusion-based image generators for monocular depth estimation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9492–9502 (2024)

33. Lee, H., Kim, K.S., Kwon, B.K., Oh, T.H.: Zero-shot depth completion via test-time alignment with affine-invariant depth prior. In: The 39th Annual AAAI Conference on Artificial Intelligence. pp. 3877–3885 (2025)
34. Lei, C., Qi, C., Xie, J., Fan, N., Koltun, V., Chen, Q.: Shape from polarization for complex scenes in the wild. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12632–12641 (2022)
35. Li, C., Ono, T., Uemori, T., Mihara, H., Gatto, A., Nagahara, H., Moriuchi, Y.: Ne-Isf: Neural incident stokes field for geometry and material estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21434–21445 (2024)
36. Li, Z., Zhong, Z., Nobuhara, S., Nishino, K., Zheng, Y.: Fooling polarization-based vision using locally controllable polarizing projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24706–24715 (2024)
37. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9970–9980 (2024)
38. Lyu, Y., Guo, H., Zhang, K., Li, S., Shi, B.: Sfpuel: Shape from polarization under unknown environment light. *Advances in Neural Information Processing Systems* **37**, 97184–97202 (2024)
39. Miyazaki, D., Tan, R.T., Hara, K., Ikeuchi, K.: Polarization-based inverse rendering from a single view. In: Proceedings Ninth IEEE International Conference on Computer Vision. pp. 982–987. IEEE (2003)
40. Morel, O., Meriaudeau, F., Stolz, C., Gorria, P.: Polarization imaging applied to 3d reconstruction of specular metallic surfaces. In: *Machine Vision Applications in Industrial Inspection XIII*. vol. 5679, pp. 178–186. SPIE (2005)
41. Ngo Thanh, T., Nagahara, H., Taniguchi, R.: Shape and light directions from shading and polarization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2310–2318 (2015)
42. Qi, X., Liu, Z., Liao, R., Torr, P.H., Urtasun, R., Jia, J.: Geonet++: Iterative geometric neural network with edge-aware refinement for joint depth and surface normal estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **44**(2), 969–984 (2022)
43. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4104–4113 (2016)
44. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: *European conference on computer vision*. pp. 501–518. Springer (2016)
45. Smith, W.A.P., Ramamoorthi, R., Tozza, S.: Height-from-polarisation with unknown lighting or albedo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **41**(12), 2875–2888 (2019)
46. Talegaonkar, C., Suresh, N.G., Novack, Z., Belhe, Y., Nagasamudra, P., Antipa, N.: Repurposing marigold for zero-shot metric depth estimation via defocus blur cues. In: *Advances in Neural Information Processing Systems* (2025)
47. Tozza, S., Smith, W.A., Zhu, D., Ramamoorthi, R., Hancock, E.R.: Linear differential constraints for photo-polarimetric height estimation. In: *Proceedings of the IEEE international conference on computer vision*. pp. 2279–2287 (2017)

48. Viola, M., Qu, K., Metzger, N., Ke, B., Becker, A., Schindler, K., Obukhov, A.: Marigold-dc: Zero-shot monocular depth completion with guided diffusion. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2025)
49. Wang, R., Xu, S., Dai, C., Xiang, J., Deng, Y., Tong, X., Yang, J.: Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 5261–5271 (2025)
50. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. arXiv preprint arXiv:2507.02546 (2025)
51. Xu, G., Ge, Y., Liu, M., Fan, C., Xie, K., Zhao, Z., Chen, H., Shen, C.: What matters when repurposing diffusion models for general dense perception tasks? In: International Conference on Learning Representations (2025)
52. Yang, L., Tan, F., Li, A., Cui, Z., Furukawa, Y., Tan, P.: Polarimetric dense monocular slam. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3857–3866 (2018)
53. Ye, C., Qiu, L., Gu, X., Zuo, Q., Wu, Y., Dong, Z., Bo, L., Xiu, Y., Han, X.: Stablenormal: Reducing diffusion variance for stable and sharp normal. *ACM Transactions on Graphics (ToG)* **43**(6), 1–18 (2024)
54. Yu, J., Wang, Y., Zhao, C., Ghanem, B., Zhang, J.: Freedom: Training-free energy-guided conditional diffusion model. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23174–23184 (2023)