Physically Grounded Monocular Depth via Nanophotonic Wavefront Encoding

Bingxuan Li[*,1] Jiahao Wu[*,2] Yuan Xu[*,2] Zezheng Zhu[2] Yunxiang Zhang[1]
Kenneth Chen[1] Yanqi Liang[2] Nanfang Yu[†,2] Qi Sun[†,1]

[1] New York University [2] Columbia University

Abstract. Depth foundation models (DFMs) offer strong learned priors for 3D perception from single RGB images but lack physical depth cues, leading to ambiguities in metric scale. We introduce metalenses, an emerging class of ultrathin planar optical elements, as a solution to physically encode missing metric depth cues via nanophotonics. In this paper, we bridge the gap between metalens and DFMs to achieve accurate metric monocular depth sensing. In a single monocular shot, our metalens embeds depth-dependent positional shifts into two polarized optical wavefronts. With an input adaptation strategty, we enable direct fine-tuning that aligns a pretrained DFM with the optical signals. To scale the training data, we further develop a comprehensive simulation pipeline that synthesizes metalens responses from RGB-D datasets, incorporating physical factors to minimize the sim-to-real gap. Experiments demonstrate that this approach outperforms both monocular metric depth estimation and depth-from-defocus baselines, showing an effective pathway for accurate monocular metric depth sensing.

Keywords: Depth foundation models · Metasurface

1 Introduction

Depth foundation models (DFMs) [34, 71] have recently achieved remarkable progress in monocular depth estimation by learning rich geometric priors from large-scale data, showing strong capabilities from *relative* to *metric* depth estimation [7,9,24,52,72]. However, the lack of physical depth cues from a monocular capture makes *metric* depth estimation inherently ill-posed, resulting in ambiguity and inaccuracy in applications requiring precise metric depth.

To enable physically grounded monocular depth estimation, providing DFMs with diverse modalities has emerged as a promising direction. Recent works leverage auxiliary sensors such as LiDAR [38, 40, 48] to provide accurate metric supervision. Yet, such systems depend on active, energy-consuming hardware, and the inclusion of additional sensors increases form factor and system complexity, deviating from a strict monocular setting. This raises a natural question:

* Equal contribution.
† Corresponding authors.

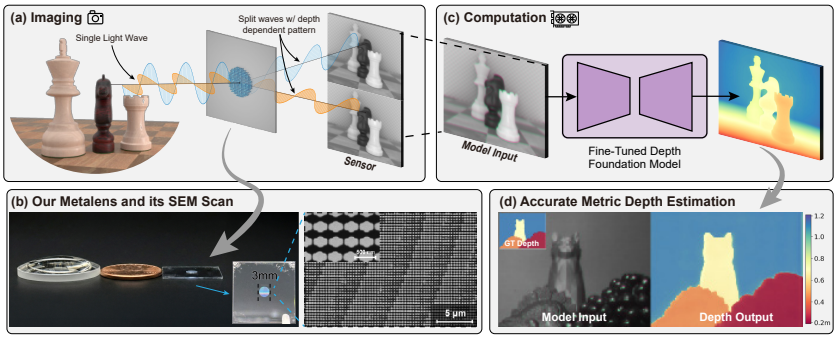


Fig. 1: *Overview of our system and method.* (a) Our birefringent metalens converts a 3D scene into two polarized images, encoding depth information in pixel-wise shifts between the images (see Fig. 3c). (b) The compact 3-mm-diameter metalens (right) consists of a two-dimensional array of 700-nm-tall TiO_2 nanopillars with anisotropic cross-sections, engineered to provide independent phase control for x- and y-polarized light. For scale, it is shown alongside a 1-inch plano-convex lens (left) and a U.S. 1-cent coin (middle). (c) These depth-dependent optical signals are converted into model inputs and processed by a fine-tuned depth foundation model. (d) Our method recovers metrically accurate depth by combining physical depth cues with learned image priors, enabling high-quality physically grounded monocular depth estimation.

can we ground DFMs in physics solely through passive optics within a compact monocular device, without relying on active sensing and additional sensors?

To answer this question, we introduce a new framework that physically grounds DFMs through passive light wave encoding in a single monocular capture. This is enabled by our custom-designed and fabricated birefringent metalens — an ultra-thin, planar element composed of nanophotonic structures for modulating optical wavefronts with subwavelength resolution (see Sec. 2.1 for background). As illustrated in Fig. 1, our metalens decomposes incoming light into two orthogonal polarization channels, each formed by a distinct depth-dependent point spread function (PSF). These two channels are formed along the same optical path and are projected onto the sensor in a single exposure, where the positional shift between the conjugate PSFs encodes metric depth. Both images originate from one viewpoint without multi-view parallax, making our approach fundamentally distinct from stereo.

Subsequently, through an input adaptation strategy, we transform the two polarization channels into a three-channel representation that embeds physical depth cues while retaining scene semantics. This enables a pretrained DFM to leverage its robust learned priors while simultaneously recovering metric scale from the optical signals without necessitating any architectural modifications. Specifically, we choose the Depth Anything V2 [71] as our model backbone. To solve the challenge in collecting large-scale training data, we develop a simulation pipeline that synthesizes the polarization channels from RGB-D datasets by physically modeling the birefringent metalens. While the simulation-to-real gap can degrade performance, we analyze its sources and introduce a novel

disocclusion-aware simulator that more accurately models the optical formation of asymmetric PSFs, supplemented by polarization-aware data augmentation.

We evaluate our approach in both simulated and physical experiments, demonstrating consistent improvement over state-of-the-art metric monocular depth estimators. Notably, we achieve performance comparable to PromptDA [40], which relies on LiDAR as an auxiliary sensor. Our method also outperform depth-from-defocus baselines [30, 59] in simulation, with ablation study verifying that both the optical frontend and the pretrained model backend drive the performance gains. These results underscore the potential of metalens in depth perception and its applicability to VR/AR, miniature robotics, medical endoscopy, and other embedded 3D vision systems. In summary, we make the following contributions:

- We introduce a new approach for physically grounded monocular depth estimation with birefringent metalens, featuring an input adaptation strategy that enables direct fine-tuning of a DFM.
- We present a disocclusion-aware optical forward model that accurately captures the image formation of asymmetric PSFs, paired with polarization-aware augmentation for improved simulation-to-real transfer.
- We demonstrate an integrated hardware-software depth sensing system, achieving highly accurate metric depth through the synergy of physical optical grounding and learned depth priors.

2 Background & Related Work

2.1 Metasurface and Metalens

A metasurface is a planar nanophotonic device composed of a 2D array of subwavelength dielectric pixels with different sizes and shapes chosen to locally control the optical phase delay, so that the array of pixels collectively mold the optical wavefront into a desired shape with subwavelength resolution [45, 73]. The pixels can also be designed to control the amplitude and polarization state of the scattered light wave so that the metasurface can impart designer polarization and amplitude profiles over the wavefront [5, 11, 29].

Metasurfaces have enabled ultra-compact optics for displays [22, 43, 74], optical computation [66], and color imaging [12, 64]. They also show promise for depth sensing, with prior work on active metasurfaces for structured-light projection [35, 37, 46], LiDAR beam steering [36, 49], and compact high-speed or high-accuracy systems [15, 33, 69]. These approaches rely on external illumination or electro-optic control. In contrast, passive metasurfaces can encode depth information in the optical response of metasurface-based lenses — known as **metalenses** — for example, in defocus [25] and chromatic aberration [62] of individual metalenses, and light fields of metalens arrays [14]. One promising route for depth sensing use a helical PSF to encode depth information [6, 16, 31, 32, 56].

However, lacking powerful computational backends with large-scale training, these prior methods are largely confined to single-object depth estimation or sparse feature matching. Consequently, they fail to reconstruct accurate, dense

depth maps for full complex scenes. We overcome this limitation by combining metalens-encoded physical depth cues with the rich depth priors embedded in DFMs. Our approach leverages these priors to enable high-resolution, metric depth estimation across the entire scene within a passive, single-sensor system.

2.2 Monocular Depth Estimation

Recovering 3D geometry from 2D images has long been a fundamental problem. Recent progress in monocular depth estimation has advanced 3D perception using. Models trained on large-scale datasets, including diffusion-based and vision transformer–based approaches, have evolved into DFMs [7, 8, 18, 70–72], demonstrating strong generalization across a wide range of scenes [9, 27, 28, 34, 51]. However, because single-view intensity lacks absolute depth cues, these models remain fundamentally scale-ambiguous. To resolve this, prompting DFMs with additional sensors such as LiDAR has been explored [40, 48]. Recent work also explored inference-time optimization strategy that uses defocus blur cues to resolve the scale ambiguity of Marigold [34,61]. However, LiDAR-based prompting requires a multi-sensor setup with active illumination [60], while inference-time optimization is prohibitively time-consuming (taking approximately five minutes on a modern GPU). In contrast, our approach employs a purely passive optical modulator, leveraging polarization-dependent PSF shifts to encode depth without the need for active sensors.

A parallel line of work lies in computational optics. Conventional depth-from-defocus (DfD) estimates depth from blur [2,26,63], but destroys high-frequency details and suffers from low sensitivity. To improve sensitivity, researchers design specialized masks [3, 30, 59, 67, 68, 75] to engineer distinct PSFs. Other approaches use dual-pixel sensors [19, 20, 47, 59] or conventional birefringent materials [4, 21, 42]. Despite these advancements, most systems train task-specific networks from scratch, without leveraging the depth priors of DFMs, which are crucial for reconstructing fine details and estimating depth in texture-less regions. Furthermore, engineered DfD introduces blur, while dual-pixel and polarization methods rely on specific sensors or extra components. Overcoming this, our metalens integrates PSF engineering, light focusing, and polarization multiplexing into one element which avoids severe blur and extra components.

3 Method

As illustrated in Fig. 1, our system integrates three components: a birefringent metalens that converts depth into and polarization channels (Sec. 3.1), a depth foundation model backbone and a encoding mechanism to inject physical cues (Sec. 3.2), and a disocclusion-aware optical forward model with alpha compositing and data augmentation that reduce simulation-to-real gaps (Sec. 3.3).

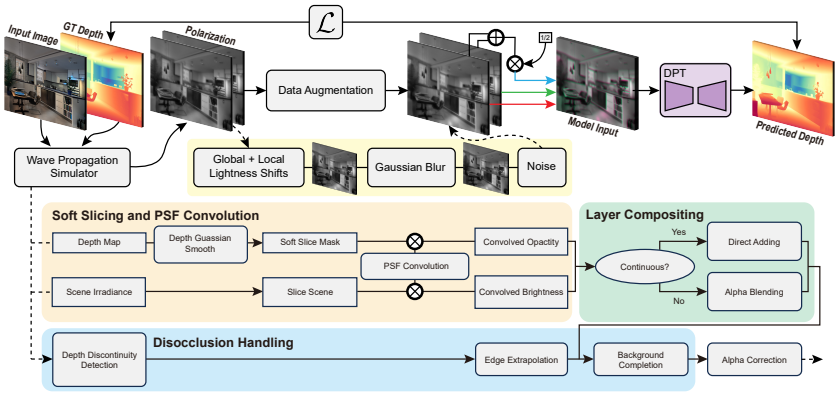


Fig. 2: *Illustration of our learning pipeline.* The top half illustrates our learning pipeline: synthetic RGB-D data are processed by our simulator to generate polarization image pairs, which are then augmented and transformed to model input. We adopt the DPT architecture of DepthAnything v2 [71] with pretrained weights for fine-tuning. The bottom half shows the workflow of our optical forward model, which integrates soft slicing, PSF convolution, disocclusion handling, and blending to eliminate simulation artifacts, narrowing the sim-to-real gap.

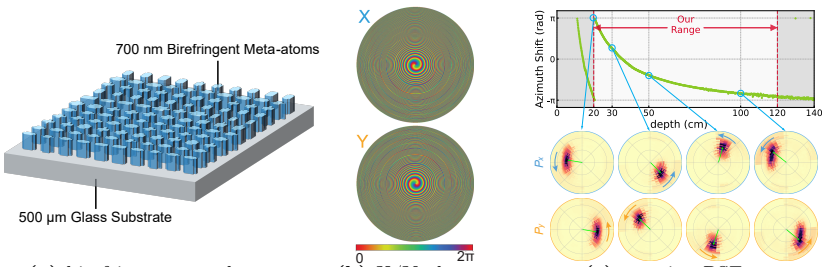
3.1 Birefringent Metalens for Polarization-Based Depth Encoding

To optimally leverage DFMs, we adopt polarization-multiplexed single-helix PSFs [53, 56]. Rotating PSFs provide noise-robust cues superior to standard defocus [23, 50], while their sharp profiles preserve high-spatial-frequency details essential for DFMs. Isolating single lobes via polarization eliminates ghosting to maximize the DFM's accuracy [21]. Additionally, replacing prior near-infrared designs [56] with our visible-light TiO_2 metasurface aligns input features with DFM priors while boosting depth sensitivity. Ablations (Tab. 4) verify that this configuration provides strong physical grounding for DFMs without joint training overhead.

Birefringent Metalens. We employ a birefringent metalens to *independently* modulate the phase ψ_k ($k \in \{x, y\}$) for x- and y-polarized light (Fig. 3a). For each polarization k, we decompose its phase profile as $\psi_k = \psi_{f,k} + \psi_{r,k}$. The $\psi_{f,k}$ term provides the focusing power. The $\psi_{r,k}$ component is engineered to create a *depth-dependent point spread function* (PSF, the blur on the sensor formed by a point light source), $\mathcal{P}_k(z)$. This PSF's shape varies with source depth z, an effect arising from the interplay between our engineered phase $\psi_{r,k}$ and the natural defocus that occurs as z deviates from the in-focus plane.

Depth Encoding with Rotating PSFs. Following [53,56], we design the phase $\psi_{r,k}$ to encode depth z as a PSF rotation. In the imaging plane's polar coordinates (r_i, ϕ_i), the engineered PSF for both polarizations, $\mathcal{P}_k$, rotates by the same depth-dependent angle $\Delta\phi_i(z)$:

$$\mathcal{P}_k(r_i, \phi_i; z) \approx \mathcal{P}_k(r_i, \phi_i - \Delta\phi_i(z); z_f), \tag{1}$$



(a) birefringent metalens **(b)** X/Y phases **(c)** rotating PSFs

Fig. 3: *Visualization of our metalens and PSFs at different depths.* (a) Schematic of the metalens. (b) Phase profiles for X- and Y-polarized light. (c) Monotonic relation between depth and PSF rotation angle in our designed depth range.

where z_f is the in-focus depth. We set the two polarized patterns 180° apart, so their relative disparity vector's angle directly tracks their co-rotation $\Delta\phi_i(z)$, enabling robust depth estimation [56].

To realize the PSF rotation, we partition the metalens at the pupil (radius R) into $N = 8$ concentric rings, each with a topological charge of n ($n = 1, \ldots, N$) [53]. In the pupil polar coordinates (r_m, ϕ_m), the x-polarized phase profile is:

$$\psi_{r,x}(r_m, \phi_m) = \left\{ n\,\phi_m \mid \sqrt{\frac{n-1}{N}} \leq \frac{r_m}{R} < \sqrt{\frac{n}{N}} \right\}. \tag{2}$$

The y-polarized phase profile $\psi_{r,y}$ is this pattern rotated by 180°: $\psi_{r,y}(r_m, \phi_m) = \psi_{r,x}(r_m, \phi_m - \pi)$. This design yields a PSF rotation angle $\Delta\phi_i(z)$ given by:

$$\Delta\phi_i(z) = \frac{\pi R^2}{N\lambda}\left(\frac{1}{z} - \frac{1}{z_f}\right), \tag{3}$$

where λ is the wavelength of light. The rotating PSF is illustrated in Fig. 3c. Further details are provided in the supplementary material.

Polarization-Multiplexing Depth Encoding. The 2D image I_k is formed by integrating the depth-wise convolutions between scene slices $S(z)$ and their corresponding depth-dependent PSFs $\mathcal{P}_k(z)$. The rotating PSF induces slight, depth-dependent shifts in 2D images (Fig. 5). Since $\mathcal{P}_x$ and $\mathcal{P}_y$ are 180° apart, these shifts occur in opposite directions for polarized image pairs, creating a monotonic disparity vector that serves as a geometric depth cue. To capture both polarized images simultaneously, we engineer the focusing phase $\psi_{f,k}$ to introduce opposite vertical deflections, spatially separating them onto the sensor halves:

$$\psi_{f,k} = -\frac{2\pi}{\lambda}\begin{cases} \sqrt{x_m^2 + (y_m - \Delta y)^2 + f^2}, & k = x \\ \sqrt{x_m^2 + (y_m + \Delta y)^2 + f^2}, & k = y, \end{cases} \tag{4}$$

where (x_m, y_m) are the coordinates on the metalens.

3.2 Physically Grounded Monocular Depth

Monocular Backbone. Recent depth foundation models [70, 71] largely follow the architecture of Dense Prediction Transformer(DPT) [54]. Given an input RGB image $I \in \mathbb{R}^{C\times H\times W}$, a Vision Transformer (ViT) [17] encoder processes it into a hierarchy of token features T_i, where each stage S_i produces tokens $T_i \in \mathbb{R}^{C_i\times(\frac{H}{p}\times\frac{W}{p}+1)}$ with feature dimension C_i and patch stride p. The DPT decoder then reconstructs spatial feature maps $F_i \in \mathbb{R}^{C_i\times\frac{H}{p}\times\frac{W}{p}}$ from tokens and progressively fuses multi-level representations through a series of convolutional layers, culminating in a dense depth prediction $D \in \mathbb{R}^{H\times W}$. While diffusion-based monocular depth approaches [27, 34] have also emerged, their computational demands make them less suitable for real-time deployment. As such, we only adopt DPT-based architectures as our base model in this work.

Adapting Polarization Measurements for Monocular Models. Our camera produces two polarization observations, (I_x, I_y), whereas monocular depth models are pretrained to take a three-channel RGB image as input. Therefore, a central question is how to make these pretrained models compatible with our wavefront-encoded measurements, while preserving their learned visual priors.

To this end, we propose a simple input adaptation strategy that converts the two polarization channels into a three-channel input:

$$(I_x, I_y) \Rightarrow (I_x, I_y, (I_x + I_y)/2)\,.$$

This input matches the expected input format of monocular depth models without modifying any network layers or adding auxiliary branches. More importantly, it preserves scene structure in a form that remains compatible with the model's learned priors while injecting physically encoded metric depth cues. As illustrated in Fig. 4, a pretrained Depth Anything V2 model can still estimate high-quality depth from this input, suggesting that it retains sufficient natural image details for effective zero-shot transfer.

We also explored alternative designs. In particular, inspired from PromptDA [40], we attached a decoder-side fusion module following their design, while adapting the fusion input from one channel to two channels to accommodate (I_x, I_y). However, as shown in our ablation study (Tab. 5), this additional fusion module brings no meaningful advantage. In practice, the input adaptation is already sufficient to align pretrained monocular models with our polarization measurements while avoiding any extra parameters or architectural overhead.

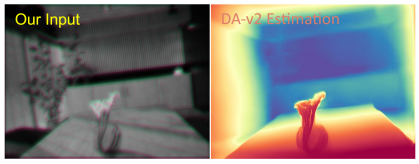


Fig. 4: *Qualitative validation of input compatibility.* DepthAnything V2 generates high-quality depth maps from our adapted input without fine-tuning. This demonstrates that our encoding preserves essential scene structure and remains aligned with the model's pretrained natural image priors.

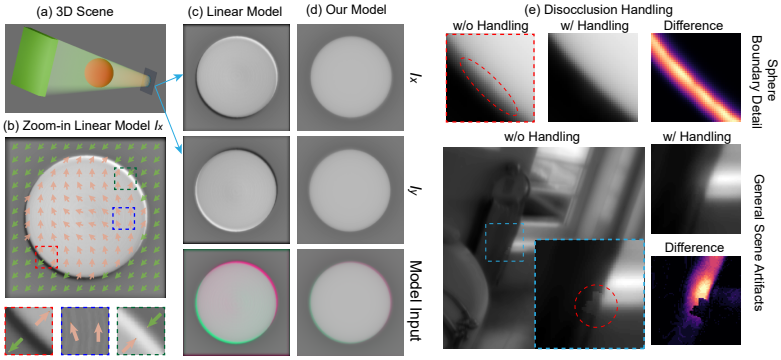


Fig. 5: *Reducing the simulation-to-real gap with a disocclusion-aware optical forward model.* (a) A 3D scene comprising a foreground sphere and a background. (b) Zoom-in of the linear model's I_x channel, with arrows indicating PSF position shifts. Insets highlight inherent artifacts: bright occlusion edges, dark disocclusion gaps, and aliasing fringes on steep surfaces. (c) Polarization pairs and model input generated by the standard linear convolution model. (d) Corresponding outputs from our disocclusion-aware model, which significantly mitigates these artifacts. (e) Qualitative ablation of the disocclusion handling module. Omitting this step introduces erroneous sphere boundary details (top) and severe artifacts in general scenes (bottom).

Simulator for Training Data. Training our model requires large-scale paired polarization and depth data. Since collecting this in the real world with pixel-wise accuracy is infeasible, we synthesize our training dataset by converting RGB-D images into depth-encoded polarization pairs using an optical forward model. We model the 3D scene capture as a discrete sum of layer-wise 2D convolutions between the per-plane scene irradiance and its corresponding PSF $\mathcal{P}_k(z)$:

$$I_k = \sum_{n=1}^{N} (S \odot M_n) * \mathcal{P}_k(z_n), \tag{5}$$

where I_k is the image intensity for polarization channel $k \in \{x, y\}$, S is the scene irradiance, and M_n is the binary mask isolating the n-th depth bin z_n. The PSFs $\mathcal{P}_k(z)$ are simulated using a Fast Fourier Transform (FFT) implementation of the Kirchhoff diffraction integral [10].

Although our simulated PSFs align closely with the measured ones (see supplementary), a pronounced simulation-to-real gap remains (Fig. 5). To address this and improve real-world generalization, we introduce a disocclusion-aware model with alpha-compositing and polarization-aware augmentation. These improvements are comprehensively validated through our ablation studies (Tab. 6).

3.3 Bridging Sim-to-Real Gap

Disocclusion-Aware Optical Forward Model Pronounced simulation-to-real discrepancies primarily arise in regions with rapid depth changes (Fig. 5b,c). The standard linear model (Eq. (5)) [13,21,68] fails here because it ignores occlusion

geometry. While prior alpha-compositing approaches [30] mitigate this, they assume symmetric blur and thus only handle occlusions. For general asymmetric optics (e.g., rotating PSFs), boundaries present a dual challenge: overlapping PSF shifts create bright occlusion edges, whereas diverging shifts leave dark *disocclusion* gaps. Additionally, steep depth gradients severely undersample the rapid PSF rotation, producing aliasing-like fringes.

To accurately render these complex dynamics, our pipeline (Fig. 2) introduces a dedicated disocclusion handling module. We first generate layer-wise opacity and brightness maps by convolving Gaussian-smoothed slice masks and scene irradiance with the PSF. While our baseline employs hybrid compositing (alpha blending for occlusions and direct summation for continuous regions), it inherently fails at disocclusion gaps. We tackle this by explicitly detecting depth discontinuities, applying edge extrapolation and background completion to reconstruct the missing background. Finally, an alpha correction step normalizes the output to ensure full opacity and suppress undersampling fringes. Our updated simulator significantly reduces simulation-to-real discrepancies (Fig. 5d); conversely, omitting this module produces inaccurate boundaries and severe artifacts (Fig. 5e), the consequence of which is illustrated in our quantitative ablation (Tab. 6).

Polarization-Aware Augmentation. Beyond simulator issues, several factors contribute to the sim-to-real gap, including (1) polarization imbalance from illumination and surface properties, (2) sensor and environmental noise, and (3) fabrication imperfections. To improve robustness, we introduce polarization-aware data augmentations: (i) global scaling for illumination changes, (ii) local brightness perturbations via a Gaussian mask for spatial polarization imbalance, (iii) Poisson and Gaussian noise for sensor and environmental effects, and (iv) Gaussian blur for fabrication-induced aberrations. As shown in Fig. 2, these augmentations regularize the model and improve tolerance to physical imperfections.

Few-Shot Real Adaptation. With the refined model and augmentation, most physics-induced gaps are mitigated. We address the remaining domain shift between simulated and real scenes by mixing a few real shots into the training set. Because dense depth is difficult to obtain, we manually segment objects and assign approximate planar depths (Fig. 7b).

3.4 Implementation Details

Metalens Fabrication. We design and fabricate a 3-mm-diameter metalens operating at λ=590 nm. The metalens consists of 700-nm-tall cross-shaped birefringent TiO_2 nanopillars patterned on a 500-μm-thick glass substrate; the nanopillars are arranged in a square lattice with a subwavelength pitch of 400 nm (Fig. 3a). The fabrication (detailed in supplementary material) involves three steps: (1) Electron-beam lithography patterning of a resist template, (2) atomic layer deposition of TiO_2 into the template, and (3) dry etching and plasma ashing to remove the resist and excess TiO_2, leaving the free-standing TiO_2 nanopillars.

Imaging Setup. We build a compact monocular depth imager (Fig. 7a), which consists of the metasurface mounted at a distance of 37.6 mm in front of a 20-MP, 1-inch monochrome CMOS sensor equipped with a 590-nm bandpass filter. The imager's in-focus depth is set to 35 cm, and the depth-sensing range is from 20 cm to 120 cm (Fig. 3). As a research prototype, the chosen hardware parameters aim to prove feasibility rather than maximize performance, which remains an important future optimization direction.

Training. We use DepthAnything v2 [71] as backbone and evaluate all three variants—ViT-Small, ViT-Base, and ViT-Large (denoted as Small, Base and Large). Starting from the metric-pretrained weights, we fine-tune the model on Hypersim [55] dataset. The depth range is linearly mapped to 0.2–1.2 m, followed by our data-preparation pipeline in Fig. 2. We use an L_1 and gradient loss L_{grad} [8] as $L = L_1 + 0.5\, L_{\text{grad}}$. We additionally mix in 5 manually annotated real samples with probability 0.05. The model is trained for 80k steps with a learning rate of 4×10^{-6} and batch sizes of 2 (Large) or 8 (Small/Base). Additional details are provided in the supplementary material.

4 Experiments

We evaluate our method through both simulation and real captures. We first compare against monocular metric depth estimation (MMDE) baselines in both simulated and physical experiments. We then compare with representative depth-from-defocus (DfD) baselines in simulation. Finally, we present ablations to isolate the contributions of our design choices.

4.1 Comparison with MMDE Baselines

Baselines and metrics. We compare against recent metric depth estimators, including Depth Anything v2/v3 [39, 71] (DepthAny. v2/v3), DepthPro [9], Lotus [27], Marigold [34], Metric3D v2 [28], MoGe v2 [65], UniDepth v2 [51], and ZoeDepth [7]. For each method, we use its largest available model variant. We report standard depth metrics, including MAE, RMSE, AbsRel, and $\delta_{0.5}$. For our method, we evaluate with all three DepthAny. v2 backbones. Since different models may operate over different metric depth ranges, we adopt a linear alignment $\{s, t\}$ to map predictions to the ground truth, following prior work [40, 59]. For a challenging and fair comparison, baseline results are optimized per image by the normalization that best aligns its predictions $\hat{D}$ with the ground-truth D: $(s^*, t^*) = \arg\min_{s,t} \| s\hat{D} + t - D \|_2^2$. This alignment removes global scale/shift ambiguity and can make baseline performance appear *higher* than their raw metric-depth accuracy. We further compare our *single-sensor* method with PromptDA [40], a recent *dual-sensor* RGB+LiDAR method. To emulate LiDAR on synthetic data, we downsample ground-truth depth by 10× to match the typical image-to-LiDAR ratio, adding uniform 1–2 cm noise to approximate iPhone LiDAR accuracy [1]. Finally, we fine-tune DepthAny. v2 without depth encoding on our dataset to isolate the benefit of the physical depth cues.

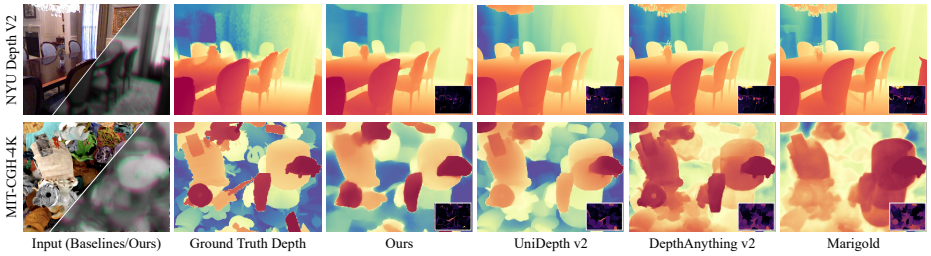


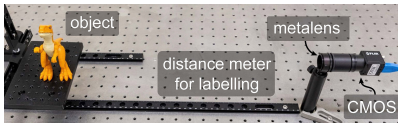
Fig. 6: *Qualitative comparison with monocular depth estimation baselines.* Bottom-right insets show the error map where dark colors indicate low error.

Simulation. We first evaluate our approach in simulation, with quantitative and qualitative results presented in Tab. 1 and Fig. 6, respectively. We experiment with two datasets: NYU Depth V2 [44], a standard indoor benchmark featuring dense LiDAR ground truth; MIT-CGH-4k [57,58], a synthetic dataset containing randomly placed 3D objects, serving as a zero-shot benchmark to assess generalization and our utilization of physical depth cues. For both benchmarks, our refined simulator is used to generate the necessary polarization images.

As shown in Tab. 1, our method achieves the best performance among all LiDAR-free baselines. On NYU Depth V2, our method demonstrate a clear advantage and is highly competitive with the LiDAR-assisted PromptDA, achieving lower RMSE, AbsRel, and $\delta_{0.5}$ errors alongside a comparable L1 error. On MIT-CGH-4k, where a lack of semantics severely degrades most baselines even after scale/shift alignment, our method retains strong accuracy. This confirms our model's ability to reliably decode metric depth from polarization wavefronts. Fine-tuning DepthAny. v2 does not improve performance on either benchmark,

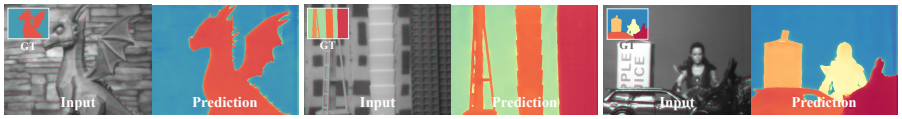
Table 1: *Quantitative comparisons on simulated experiment.* Train : fine-tuned on our dataset; Post. : post-aligned with GT using least-square fitting; w/ LiDAR : with additional simulated LiDAR input. Method with * is finetuned on our dataset. We highlight the top three results among LiDAR-free methods. Note that post-alignment removes global scale/shift ambiguity which can substantially improve the results..

Zero Shot	Train / Post. / w/ LiDAR	NYU Depth v2 MAE↓	RMSE↓	AbsRel↓	$\delta_{0.5}$↑	Zero Shot	MIT-CGH-4k MAE↓	RMSE↓	AbsRel↓	$\delta_{0.5}$↑
Yes	**Ours-Large**	0.023	0.040	0.039	0.951	Yes	**0.067**	**0.126**	0.105	0.764
	Ours-Base	**0.022**	**0.039**	**0.036**	**0.957**		0.068	0.129	**0.102**	**0.772**
	Ours-Small	0.025	0.043	0.043	0.936		0.076	0.137	0.125	0.724
	DepthAny. v2*	0.128	0.148	0.267	0.341		0.301	0.371	0.410	0.100
	DepthAny. v2*	0.054	0.079	0.098	0.731		0.180	0.220	0.371	0.241
	DepthAny. v2 [71]	0.043	0.067	0.079	0.805		0.151	0.190	0.308	0.300
	DepthAny. v3 [39]	0.037	0.062	0.069	0.846		0.134	0.172	0.276	0.349
	Depth Pro [9]	0.038	0.061	0.071	0.841		0.144	0.181	0.292	0.309
	Lotus [27]	0.069	0.093	0.127	0.575		0.162	0.201	0.330	0.268
	Marigold [34]	0.045	0.070	0.085	0.785		0.174	0.213	0.355	0.243
	Metric3D v2 [28]	0.056	0.082	0.110	0.766		0.212	0.252	0.442	0.183
	MoGe v2 [65]	0.034	0.058	0.063	0.867		0.133	0.169	0.272	0.346
	UniDepth v2 [51]	0.034	0.059	0.063	0.865		0.127	0.164	0.259	0.360
No	ZoeDepth [7]	0.041	0.062	0.076	0.805		0.205	0.247	0.428	0.204
	PromptDA [40]	0.021	0.042	0.036	0.955		0.058	0.113	0.099	0.802



(a) our physical setup

(b) model input, our depth prediction, and label



(c) our results on diverse backgrounds and objects; GT depth is shown on the top-left corner

Fig. 7: *Physical experiment setup and qualitative results.* We encourage readers to see our supplementary material for additional results.

indicating that the gains of our method do not arise from dataset-specific fine-tuning, but from the physically encoded depth information.

Physical experiments. To acquire physical measurements, we mounted our metalens-based depth camera prototype and target objects on an optical table with precise distance control (Fig. 7). We captured 42 scenes featuring 25 distinct objects, including single- and multi-object setups placed at various depths; 20 objects are unseen in our five-shot training set. Our model uses both polarization channels as input (Sec. 3.2), whereas baselines are provided with a grayscale image from one polarization channel. Obtaining dense LiDAR or stereo ground truth is challenging due to field-of-view mismatch and sparsity [59]. Following established practices [30,56], we assign reference depths to nearly planar objects using masks and known mounting distances, with averaged label uncertainty (< 1 cm) well below our performance margins.

As shown in Tab. 2, our method consistently outperforms all LiDAR-free baselines and achieves performance close to the LiDAR-assisted PromptDA. Even after applying optimal scale/shift alignment to baseline predictions, our approach maintains a clear margin, highlighting its effective use of physical depth cues. Qualitative results in Fig. 7 further show that our predictions are metrically accurate while preserving sharp object boundaries. Overall, these results demonstrate that our physically prompted model transfers robustly from simulation to real captures.

Table 2: *Quantitative comparisons on physical experiment.* Notations are consistent with the previous table.

Train / Post. / w/ LiDAR	MAE↓	RMSE↓	AbsRel↓	$\delta_{0.5}$ ↑
Ours-Large	0.036	**0.075**	0.060	0.888
Ours-Base	**0.032**	0.089	**0.055**	**0.895**
Ours-Small	0.048	0.098	0.081	0.818
DepthAny. v2*	0.061	0.100	0.128	0.678
DepthAny. v2	0.135	0.169	0.234	0.483
DepthAny. v3	0.111	0.146	0.198	0.561
Depth Pro	0.089	0.122	0.159	0.661
Lotus	0.140	0.181	0.261	0.479
Marigold	0.062	0.101	0.117	0.744
Metric3D v2	0.159	0.193	0.276	0.424
MoGe v2	0.063	0.095	0.119	0.709
UniDepth v2	0.107	0.145	0.196	0.592
ZoeDepth	0.109	0.146	0.175	0.561
PromptDA	0.030	0.086	0.054	0.951

4.2 Comparison with DfD

We consider two representative DfD baselines: DeepDfD [30] and Split-Aperture [59]. For fair evaluation, we follow their original metrics and dataset, FlyingThings3D [41]. To match their 1–5 m range, we retain the phase design in Eq. (2) and scale our metalens to a 5 mm diameter with a 50 mm focal length, comparable to baseline optics. We report results using our small backbone to match the baseline model capacity. As shown in Tab. 3, our method consistently outperforms the DfD baselines across all metrics. To attribute these gains, we next conduct ablations on both the optical design and the backbone model (Sec. 4.3).

Table 3: *Comparison on FlyingThings3D.*

Method	MAE↓	RMSE↓	$\log_{10}$↓	δ_1 ↑
Ours-Small	**0.022**	**0.132**	**0.005**	**0.995**
DeepDfD [30]	0.089	0.191	0.034	0.941
Split-Aperture [59]	0.086	0.147	0.011	0.993

4.3 Ablations and Analysis

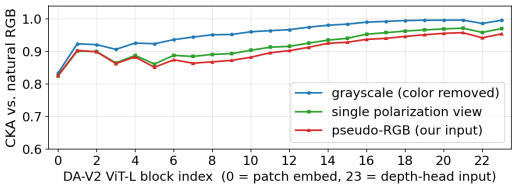
Deconstructing the performance gains over DfD baselines. To isolate the source of our improvements over DfD baselines, we conduct controlled ablations (Tab. 4) across three factors: metalens design (**Meta**), backbone architecture (**ViT**), and learned pretraining prior (**Prior**). For baselines, we use DeepDfD's optical design [30] which yields a compatible three-channel observation, and a U-Net of comparable capacity to the ViT.

Table 4: *Ablation study on performance gains over DFD baselines.*

Meta	**ViT**	**Prior**	MAE↓	RMSE↓	$\log_{10}$↓	δ_1 ↑
✓	✓	✓	**0.022**	**0.132**	**0.005**	**0.995**
×	✓	✓	0.035	0.133	0.009	0.994
✓	✓	×	0.061	0.268	0.014	0.984
✓	×	×	0.063	0.254	0.014	0.983
×	×	×	0.089	0.191	0.034	0.941

Our results reveal that the pretrained prior is the primary driver of performance; omitting it causes a steep drop in accuracy (row 1 vs. row 4), whereas fine-tuning a pretrained ViT with DeepDfD optics yields substantial gains (row 2 vs. row 5). Furthermore, our metalens design provides superior physical grounding for depth estimation, outperforming DeepDfD optics when using the same pretrained backbone (row 1 vs. row 2, further discussed in supplementary material). Finally, while the ViT marginally outperforms a similarly sized U-Net, the architectural difference alone is not significant (row 3 vs. row 4).

Preservation of pretrained priors. We further analyze why the pretrained depth foundation model remains effective under our pseudo-RGB input adaptation. Although the chromatic channels are remapped to polarization views, the input preserves the spatial statistics most relevant to dense prediction, including edges, object boundaries, and texture gradients. To quantify the resulting feature shift, we feed the pretrained DA-V2 ViT-L the same Hypersim test scenes represented as grayscale, a single polarization view tiled to three channels, and our pseudo-RGB input, and then measure layer-wise CKA similarity to features extracted from the original RGB input.



As shown in Sec. 4.3, all variants maintain high similarity to RGB features, with CKA scores above 0.83 across layers and above 0.95 in the final three blocks. Moreover, pseudo-RGB remains within approximately 0.04 CKA of the grayscale baseline. This indicates that the chromatic-to-polarization remapping induces no larger representational shift than removing color alone, supporting the transferability of the pretrained priors.

Alternative designs. To validate our input adaptation strategy (**Adapt.**) in Sec. 3.2, we compare it against a decoder-side fusion approach (**Fusion**) in physical experiments. We adapt the fusion module in PromptDA [40] to our setting by modifying its input layers.

As shown in Tab. 5, our adaptation strategy (row 1) achieves the best overall performance. Adding the fusion module (row 2) yields only a marginal RMSE improvement while degrading other metrics, and relying solely on fusion (row 3) drastically reduces performance. This confirms that our input adaptation strategy is a simpler, more effective way to inject polarization wavefronts.

Table 5: *Ablation study comparing our input adaptation strategy against decoder-side fusion.*

Adapt.	**Fusion**	MAE↓	RMSE↓	AbsRel↓	$\delta_{0.5}$ ↑
✓	×	**0.032**	0.089	**0.055**	**0.895**
✓	✓	**0.032**	**0.087**	0.058	0.875
×	✓	0.063	0.113	0.109	0.727

Sim-to-real transfer. To decouple the benefits of physically accurate modeling from the inherent advantages of real-world fine-tuning, we ablate our pipeline (Tab. 6) under two training regimes: synthetic-only and with real data included. Removing disocclusion modeling (Tab. 6(a)) consistently degrades performance which confirms that real-world data cannot fully compensate for inaccurate optical modeling. Removing all augmentations (Tab. 6(c)) causes a drastic performance collapse in the synthetic-only regime, proving their necessity for zero-shot generalization. (Tab. 6(d–f)) reveals that modeling light imbalance is the most critical factor. Since our approach relies on polarization, we hypothesize it prevents the network from overfitting to absolute light intensities and forces it to focus on robust positional shifts.

Table 6: *Ablation study on sim-to-real transfer.* We disable simulator/augmentation components from full pipeline and evaluate with and without including real data.

Component	Synthetic only				Real data included			
	MAE↓	RMSE↓	AbsRel↓	$\delta_{0.5}$ ↑	MAE↓	RMSE↓	AbsRel↓	$\delta_{0.5}$ ↑
(a) Full	**0.107**	**0.149**	0.148	**0.546**	**0.032**	**0.088**	**0.055**	0.895
(b) w/o disocclusion	0.121	0.173	0.159	0.479	0.046	0.096	0.074	0.854
(c) w/o augmentation	0.198	0.260	0.226	0.240	0.038	0.097	0.066	0.865
(d) w/o light imbalance	0.116	0.159	0.157	0.451	0.034	0.089	0.058	0.868
(e) w/o blur	0.112	0.151	0.153	0.496	0.034	0.091	0.059	0.876
(f) w/o noise	0.111	0.162	**0.143**	0.534	0.033	0.092	0.056	**0.907**

5 Limitations and Future Work

Limitations. Our current system is intended as a proof of concept for physically grounding DFMs with nanophotonic wavefront cues, rather than a deployment-ready depth camera. The prototype is constrained by a system-level photon budget. Its 3-mm, $f/11.3$ aperture and 10-nm bandpass filter substantially reduce aperture–spectral throughput compared with mature $f/6$ RGB or DOE-based systems. As a result, the current setup is better suited to well-lit, near-range scenes with relatively longer exposures.

Polarization multiplexing does not directly discard the total collected signal, but it maps the two polarization views to separate sensor regions, reducing the effective field of view or sampling density. It can also increase sensitivity to background and read noise in non-shot-noise-limited regimes. In addition, the prototype currently has a limited field of view, increased tube length, and a narrower operating range than mature multi-lens or RGB+LiDAR systems.

Real-world polarization imbalance is another practical limitation. Our low-pass filtered mask and polarization-aware augmentation mitigate low-frequency lighting discrepancies and encourage the network to rely on positional shifts rather than intensity. However, high-frequency imbalance from specular reflections or actively polarized illumination remains challenging.

Future Work. Future work will explore end-to-end metalens–DFM co-design to jointly optimize optical encoding and depth inference. Larger-aperture metasurfaces and polarization-resolved sensors could improve photon throughput, compactness, field of view, and depth range, while reducing the need for narrow spectral filtering. We will also investigate training and calibration strategies that improve robustness under complex environmental polarization, especially for specular and partially polarized scenes.

6 Conclusion

We present a metalens-based depth imaging system that physically grounds depth foundation models using polarization-encoded nanophotonic wavefronts. Combining passive optical modulator with pretrained depth priors, we mitigate monocular scale ambiguity and enable accurate metric depth, substantially outperforming monocular depth estimation and prior depth-from-defocus methods that did not leverage depth priors. We hope this work will help bridge emerging foundation models and nanophotonics materials, enabling compact depth sensing for VR/AR, miniature robotics, medical endoscopy, and beyond.

References

1. Abdel-Majeed, H.M., Shaker, I.F., Abdel-Wahab, A., Awad, A.A.D.I.: Indoor mapping accuracy comparison between the apple devices' lidar sensor and terrestrial laser scanner. HBRC Journal **20**(1), 915–931 (2024)

2. Alexander, E., Guo, Q., Koppal, S., Gortler, S., Zickler, T.: Focal flow: Measuring distance and velocity with defocus and differential motion. In: European conference on computer vision. pp. 667–682. Springer (2016)
3. Antipa, N., Kuo, G., Heckel, R., Mildenhall, B., Bostan, E., Ng, R., Waller, L.: Diffusercam: lensless single-exposure 3d imaging. Optica **5**(1), 1–9 (Jan 2018). https://doi.org/10.1364/OPTICA.5.000001, https://opg.optica.org/optica/abstract.cfm?URI=optica-5-1-1
4. Baek, S.H., Gutierrez, D., Kim, M.H.: Birefractive stereo imaging for single-shot depth acquisition. ACM Transactions on Graphics (TOG) **35**(6), 1–11 (2016)
5. Balthasar Mueller, J.P., Rubin, N.A., Devlin, R.C., Groever, B., Capasso, F.: Metasurface polarization optics: independent phase control of arbitrary orthogonal states of polarization. Phys. Rev. Lett. **118**(11), 113901 (Mar 2017), http://dx.doi.org/10.1103/PhysRevLett.118.113901
6. Berlich, R., Bräuer, A., Stallinga, S.: Single shot three-dimensional imaging using an engineered point spread function. Optics Express **24**(6), 5946–5960 (2016). https://doi.org/10.1364/OE.24.005946, https://opg.optica.org/oe/abstract.cfm?URI=oe-24-6-5946
7. Bhat, S.F., Birkl, R., Wofk, D., Wonka, P., Müller, M.: Zoedepth: Zero-shot transfer by combining relative and metric depth (2023). https://doi.org/10.48550/ARXIV.2302.12288, https://arxiv.org/abs/2302.12288
8. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024)
9. Bochkovskii, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S.R., Koltun, V.: Depth pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073 (2024)
10. Born, M., Wolf, E.: Principles of optics: electromagnetic theory of propagation, interference and diffraction of light. Elsevier (2013)
11. Cao, Z., Li, N., Zhu, L., Wu, J., Dai, Q., Qiao, H.: Aberration-robust monocular passive depth sensing using a meta-imaging camera. Light: Science & Applications **13**(1), 236 (2024)
12. Chakravarthula, P., Sun, J., Li, X., Lei, C., Chou, G., Bijelic, M., Froesch, J., Majumdar, A., Heide, F.: Thin on-sensor nanophotonic array cameras. ACM Transactions on Graphics (TOG) **42**(6), 1–18 (2023)
13. Chang, J., Wetzstein, G.: Deep optics for monocular depth estimation and 3d object detection. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 10193–10202 (2019)
14. Chen, M.K., Liu, X., Wu, Y., Zhang, J., Yuan, J., Zhang, Z., Tsai, D.P.: A meta-device for intelligent depth perception. Advanced Materials **35**(34), 2107465 (2023). https://doi.org/https://doi.org/10.1002/adma.202107465, https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.202107465
15. Chen, R., Shao, Y., Zhou, Y., Dang, Y., Dong, H., Zhang, S., Wang, Y., Chen, J., Ju, B.F., Ma, Y.: A semisolid micromechanical beam steering system based on micrometa-lens arrays. Nano Letters **22**(4), 1595–1603 (2022). https://doi.org/10.1021/acs.nanolett.1c04493, https://doi.org/10.1021/acs.nanolett.1c04493, doi: 10.1021/acs.nanolett.1c04493
16. Colburn, S., Majumdar, A.: Metasurface generation of paired accelerating and rotating optical beams for passive ranging and scene reconstruction. ACS Photonics **7**(6), 1529–1536 (2020). https://doi.org/10.1021/acsphotonics.0c00354, https://doi.org/10.1021/acsphotonics.0c00354, doi: 10.1021/acsphotonics.0c00354

17. Dosovitskiy, A.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
18. 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: ECCV (2024)
19. Garg, R., Wadhwa, N., Ansari, S., Barron, J.T.: Learning single camera depth estimation using dual-pixels. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019)
20. Ghanekar, B., Khan, S.S., Sharma, P., Singh, S., Boominathan, V., Mitra, K., Veeraraghavan, A.: Passive snapshot coded aperture dual-pixel rgb-d imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25348–25357 (2024)
21. Ghanekar, B., Saragadam, V., Mehra, D., Gustavsson, A.K., Sankaranarayanan, A.C., Veeraraghavan, A.: Ps^2f: Polarized spiral point spread function for single-shot 3d sensing. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
22. Gopakumar, M., Lee, G.Y., Choi, S., Chao, B., Peng, Y., Kim, J., Wetzstein, G.: Full-colour 3d holographic augmented-reality displays with metasurface waveguides. Nature pp. 1–7 (2024)
23. Greengard, A., Schechner, Y.Y., Piestun, R.: Depth from diffracted rotation. Optics Letters **31**(2), 181–183 (2006). https://doi.org/10.1364/OL.31.000181, https://opg.optica.org/ol/abstract.cfm?URI=ol-31-2-181
24. Guizilini, V., Vasiljevic, I., Chen, D., Ambruș, R., Gaidon, A.: Towards zero-shot scale-aware monocular depth estimation. In: 2023 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 9199–9209 (2023). https://doi.org/10.1109/ICCV51070.2023.00847
25. Guo, Q., Shi, Z., Huang, Y.W., Alexander, E., Qiu, C.W., Capasso, F., Zickler, T.: Compact single-shot metalens depth sensors inspired by eyes of jumping spiders. Proceedings of the National Academy of Sciences **116**(46), 22959–22965 (2019). https://doi.org/doi:10.1073/pnas.1912154116, https://www.pnas.org/doi/abs/10.1073/pnas.1912154116
26. Gur, S., Wolf, L.: Single image depth estimation trained via depth from defocus cues. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7683–7692 (2019)
27. 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. arXiv preprint arXiv:2409.18124 (2024)
28. Hu, M., Yin, W., Zhang, C., Cai, Z., Long, X., Chen, H., Wang, K., Yu, G., Shen, C., Shen, S.: Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
29. Huang, H., Overvig, A.C., Xu, Y., Malek, S.C., Tsai, C.C., Alù, A., Yu, N.: Leaky-wave metasurfaces for integrated photonics. Nat. Nanotechnol. **18**(6), 580–588 (2023), https://doi.org/10.1038/s41565-023-01360-z
30. Ikoma, H., Nguyen, C.M., Metzler, C.A., Peng, Y., Wetzstein, G.: Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In: 2021 IEEE International Conference on Computational Photography (ICCP). pp. 1–12. IEEE (2021)
31. Jin, C., Afsharnia, M., Berlich, R., Fasold, S., Zou, C., Arslan, D., Staude, I., Pertsch, T., Setzpfandt, F.: Dielectric metasurfaces for distance measurements and

three-dimensional imaging. Advanced Photonics **1**(3), 036001 (2019), https://doi.org/10.1117/1.AP.1.3.036001
32. Jin, C., Zhang, J., Guo, C.: Metasurface integrated with double-helix point spread function and metalens for three-dimensional imaging. Nanophotonics **8**(3), 451–458 (2019). https://doi.org/doi:10.1515/nanoph-2018-0216, https://doi.org/10.1515/nanoph-2018-0216
33. Juliano Martins, R., Marinov, E., Youssef, M.A.B., Kyrou, C., Joubert, M., Colmagro, C., Gâté, V., Turbil, C., Coulon, P.M., Turover, D., Khadir, S., Giudici, M., Klitis, C., Sorel, M., Genevet, P.: Metasurface-enhanced light detection and ranging technology. Nature Communications **13**(1), 5724 (2022). https://doi.org/10.1038/s41467-022-33450-2, https://doi.org/10.1038/s41467-022-33450-2
34. 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 (CVPR) (2024)
35. Kim, G., Kim, Y., Yun, J., Moon, S.W., Kim, S., Kim, J., Park, J., Badloe, T., Kim, I., Rho, J.: Metasurface-driven full-space structured light for three-dimensional imaging. Nature Communications **13**(1), 5920 (2022). https://doi.org/10.1038/s41467-022-32117-2, https://doi.org/10.1038/s41467-022-32117-2
36. Kim, I., Martins, R.J., Jang, J., Badloe, T., Khadir, S., Jung, H.Y., Kim, H., Kim, J., Genevet, P., Rho, J.: Nanophotonics for light detection and ranging technology. Nature Nanotechnology **16**(5), 508–524 (2021). https://doi.org/10.1038/s41565-021-00895-3, https://doi.org/10.1038/s41565-021-00895-3
37. Li, Z., Dai, Q., Mehmood, M.Q., Hu, G., yanchuk, B.L., Tao, J., Hao, C., Kim, I., Jeong, H., Zheng, G., Yu, S., Alù, A., Rho, J., Qiu, C.W.: Full-space cloud of random points with a scrambling metasurface. Light: Science & Applications **7**(1), 63 (2018). https://doi.org/10.1038/s41377-018-0064-3, https://doi.org/10.1038/s41377-018-0064-3
38. Liang, Y., Hu, Y., Shao, W., Fu, Y.: Distilling monocular foundation model for fine-grained depth completion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22254–22265 (2025)
39. Lin, H., Chen, S., Liew, J.H., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
40. Lin, H., Peng, S., Chen, J., Peng, S., Sun, J., Liu, M., Bao, H., Feng, J., Zhou, X., Kang, B.: Prompting depth anything for 4k resolution accurate metric depth estimation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 17070–17080 (2025)
41. Mayer, N., Ilg, E., Häusser, P., Fischer, P., Cremers, D., Dosovitskiy, A., Brox, T.: A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2016), http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16, arXiv:1512.02134
42. Meuleman, A., Baek, S.H., Heide, F., Kim, M.H.: Single-shot monocular rgb-d imaging using uneven double refraction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2465–2474 (2020)
43. Nam, S.W., Kim, Y., Kim, D., Jeong, Y.: Depolarized holography with polarization-multiplexing metasurface. ACM Transactions on Graphics (TOG) **42**(6), 1–16 (2023)
44. Nathan Silberman, Derek Hoiem, P.K., Fergus, R.: Indoor segmentation and support inference from rgbd images. In: ECCV (2012)

45. Ni, X., Emani, N.K., Kildishev, A.V., Boltasseva, A., Shalaev, V.M.: Broadband light bending with plasmonic nanoantennas. Science **335**(6067), 427–427 (Jan 2012), http://dx.doi.org/10.1126/science.1214686
46. Ni, Y., Chen, S., Wang, Y., Tan, Q., Xiao, S., Yang, Y.: Metasurface for structured light projection over 120° field of view. Nano Letters **20**(9), 6719–6724 (2020). https://doi.org/10.1021/acs.nanolett.0c02586, https://doi.org/10.1021/acs.nanolett.0c02586, doi: 10.1021/acs.nanolett.0c02586
47. Pan, L., Chowdhury, S., Hartley, R., Liu, M., Zhang, H., Li, H.: Dual pixel exploration: Simultaneous depth estimation and image restoration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4340–4349 (June 2021)
48. Park, J.H., Jeong, C., Lee, J., Jeon, H.G.: Depth prompting for sensor-agnostic depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9859–9869 (2024)
49. Park, J., Jeong, B.G., Kim, S.I., Lee, D., Kim, J., Shin, C., Lee, C.B., Otsuka, T., Kyoung, J., Kim, S., Yang, K.Y., Park, Y.Y., Lee, J., Hwang, I., Jang, J., Song, S.H., Brongersma, M.L., Ha, K., Hwang, S.W., Choo, H., Choi, B.L.: All-solid-state spatial light modulator with independent phase and amplitude control for three-dimensional lidar applications. Nature Nanotechnology **16**(1), 69–76 (2021). https://doi.org/10.1038/s41565-020-00787-y, https://doi.org/10.1038/s41565-020-00787-y
50. Pavani, S.R.P., Piestun, R.: Three dimensional tracking of fluorescent microparticles using a photon-limited double-helix response system. Optics express **16**(26), 22048–22057 (2008)
51. Piccinelli, L., Sakaridis, C., Yang, Y.H., Segu, M., Li, S., Abbeloos, W., Van Gool, L.: Unidepthv2: Universal monocular metric depth estimation made simpler. arXiv preprint arXiv:2502.20110 (2025)
52. Piccinelli, L., Yang, Y.H., Sakaridis, C., Segu, M., Li, S., Van Gool, L., Yu, F.: UniDepth: Universal monocular metric depth estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024)
53. Prasad, S.: Rotating point spread function via pupil-phase engineering. Optics Letters **38**(4), 585–587 (2013). https://doi.org/10.1364/OL.38.000585, https://opg.optica.org/ol/abstract.cfm?URI=ol-38-4-585
54. Ranftl, R., Bochkovskiy, A., Koltun, V.: Vision transformers for dense prediction. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12179–12188 (2021)
55. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: International Conference on Computer Vision (ICCV) 2021 (2021)
56. Shen, Z., Zhao, F., Jin, C., Wang, S., Cao, L., Yang, Y.: Monocular metasurface camera for passive single-shot 4d imaging. Nature Communications **14**(1), 1035 (2023)
57. Shi, L., Li, B., Kim, C., Kellnhofer, P., Matusik, W.: Towards real-time photorealistic 3d holography with deep neural networks. Nature **591**(7849), 234–239 (2021)
58. Shi, L., Li, B., Matusik, W.: End-to-end learning of 3d phase-only holograms for holographic display. Light: Science & Applications **11**(1), 247 (2022)
59. Shi, Z., Chugunov, I., Bijelic, M., Côté, G., Yeom, J., Fu, Q., Amata, H., Heidrich, W., Heide, F.: Split-aperture 2-in-1 computational cameras. ACM Trans.

Graph. **43**(4) (jul 2024). https://doi.org/10.1145/3658225, https://doi.org/10.1145/3658225

60. Sun, J.Q., Weng, H., Xing, X., Yeum, C.M., Crowley, M.: View invariant learning for vision-language navigation in continuous environments. IEEE Robotics and Automation Letters **11**(5), 5861–5868 (2026). https://doi.org/10.1109/LRA.2026.3669785
61. 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 (2025), https://arxiv.org/abs/2505.17358
62. Tan, S., Yang, F., Boominathan, V., Veeraraghavan, A., Naik, G.V.: 3d imaging using extreme dispersion in optical metasurfaces. ACS Photonics **8**(5), 1421–1429 (2021). https://doi.org/10.1021/acsphotonics.1c00110, https://doi.org/10.1021/acsphotonics.1c00110, doi: 10.1021/acsphotonics.1c00110
63. Tang, H., Cohen, S., Price, B., Schiller, S., Kutulakos, K.N.: Depth from defocus in the wild. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2740–2748 (2017)
64. Tseng, E., Colburn, S., Whitehead, J., Huang, L., Baek, S.H., Majumdar, A., Heide, F.: Neural nano-optics for high-quality thin lens imaging. Nature communications **12**(1), 6493 (2021)
65. 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 (2025), https://arxiv.org/abs/2507.02546
66. Wei, K., Li, X., Froech, J., Chakravarthula, P., Whitehead, J., Tseng, E., Majumdar, A., Heide, F.: Spatially varying nanophotonic neural networks. Science Advances **10**(45), eadp0391 (2024)
67. Wijayasingha, L., Alemzadeh, H., Stankovic, J.A.: Camera-independent single image depth estimation from defocus blur. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3749–3758 (January 2024)
68. Wu, Y., Boominathan, V., Chen, H., Sankaranarayanan, A., Veeraraghavan, A.: Phasecam3d—learning phase masks for passive single view depth estimation. In: 2019 IEEE International Conference on Computational Photography (ICCP). p. 1–12. IEEE (2019)
69. Yan, T., Zhou, T., Guo, Y., Zhao, Y., Shao, G., Wu, J., Huang, R., Dai, Q., Fang, L.: Nanowatt all-optical 3d perception for mobile robotics. Science Advances **10**(27), eadn2031 (2024). https://doi.org/doi:10.1126/sciadv.adn2031, https://www.science.org/doi/abs/10.1126/sciadv.adn2031
70. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10371–10381 (2024)
71. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
72. Yin, W., Zhang, C., Chen, H., Cai, Z., Yu, G., Wang, K., Chen, X., Shen, C.: Metric3d: Towards zero-shot metric 3d prediction from a single image. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9043–9053 (2023)
73. Yu, N., Genevet, P., Kats, M.A., Aieta, F., Tetienne, J.P., Capasso, F., Gaburro, Z.: Light propagation with phase discontinuities: generalized laws of reflection and refraction. science **334**(6054), 333–337 (2011)

74. Zheng, C., Zhao, G., So, P.: Close the design-to-manufacturing gap in computational optics with a'real2sim'learned two-photon neural lithography simulator. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–9 (2023)
75. Zheng, Y., Salman Asif, M.: Joint image and depth estimation with mask-based lensless cameras. IEEE Transactions on Computational Imaging **6**, 1167–1178 (2020). https://doi.org/10.1109/TCI.2020.3010360