

Seeing Clearly and Deeply: An RGBD Imaging Approach with a Bio-inspired Monocentric Design

Zongxi Yu^{1,*}, Xiaolong Qian^{1,*}, Shaohua Gao^{1,2}, Qi Jiang¹, Yao Gao¹, Kailun Yang^{3,4,†}, and Kaiwei Wang^{1,†}

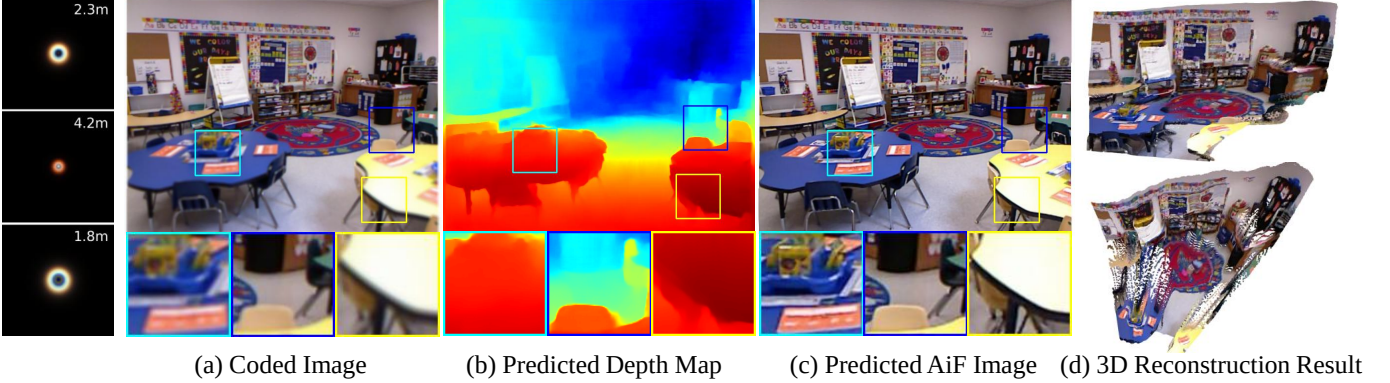


Fig. 1: Joint depth estimation and all-in-focus imaging with the proposed BMI Framework. (a) Coded image captured (simulated) by our bio-inspired monocentric lens. The optically encoded blur varies significantly with object distance, driven by the depth-dependent Point Spread Functions (PSFs). Example PSFs shown on the far left correspond to the average depths (2.3m, 4.2m, 1.8m) within the detail regions highlighted below. (b) Predicted depth map and (c) Predicted AiF image, jointly recovered from the single coded image (a) using our reconstruction network. (d) Resulting 3D point cloud reconstruction generated from the outputs (b) and (c). Detail regions, highlighted with colored boxes, showcase the relationship between the depth-encoded blur, the corresponding recovered depth values, and the restored image details. Please zoom in for the best view.

Abstract—Achieving high-fidelity, compact RGBD imaging presents a dual challenge: conventional compact optics struggle with RGB sharpness across the entire depth-of-field, while software-only Monocular Depth Estimation (MDE) is an ill-posed problem reliant on unreliable semantic priors. While deep optics with elements like DOEs can encode depth, they introduce trade-offs in fabrication complexity and chromatic aberrations, compromising simplicity. To address this, we first introduce a novel bio-inspired all-spherical monocentric lens, around which we build the Bionic Monocentric Imaging (BMI) framework, a holistic co-design. This optical design naturally encodes depth into its depth-varying Point Spread Functions (PSFs) without requiring complex diffractive or freeform elements. We establish a rigorous physically-based forward model to generate a synthetic dataset by precisely simulating the optical degradation process.

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LZ24F050003, in part by the Henan Province Key R&D Special Project (231111112700), in part by the National Natural Science Foundation of China (NSFC) under Grant No. 62473139, in part by the Hunan Provincial Research and Development Project (Grant No. 2025QK3019), in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2025B20), and in part by the State Key Laboratory of Autonomous Intelligent Unmanned Systems (the opening project number ZZKF2025-2-10).

¹Z. Yu, X. Qian, S. Gao, Q. Jiang, Y. Gao, and K. Wang are with the State Key Laboratory of Extreme Photonics and Instrumentation, Zhejiang University, Hangzhou 310027, China (e-mail: wangkaiwei@zju.edu.cn).

²S. Gao is also with the DJI Technology Co. Ltd., Shenzhen 518055, China.

³K. Yang is with the School of Artificial Intelligence and Robotics, Hunan University, Changsha 410012, China (e-mail: kailun.yang@hnu.edu.cn).

⁴K. Yang is also with the National Engineering Research Center of Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China.

*Equal contribution.

†Corresponding authors: Kaiwei Wang and Kailun Yang.

This simulation pipeline is co-designed with a dual-head, multi-scale reconstruction network that employs a shared encoder to jointly recover a high-fidelity All-in-Focus (AiF) image and a precise depth map from a single coded capture. Extensive experiments validate the state-of-the-art performance of the proposed framework. In depth estimation, the method attains an Abs Rel of 0.026 and an RMSE of 0.130, markedly outperforming leading software-only approaches and other deep optics systems. For image restoration, the system achieves an SSIM of 0.960 and a perceptual LPIPS score of 0.082, thereby confirming a superior balance between image fidelity and depth accuracy. This study illustrates that the integration of bio-inspired, fully spherical optics with a joint reconstruction algorithm constitutes an effective strategy for addressing the intrinsic challenges in high-performance compact RGBD imaging. The source code will be publicly available at <https://github.com/ZongxiYu-ZJU/BMI>.

Index Terms—Bio-inspired optics, image restoration, depth estimation, joint learning.

I. INTRODUCTION

High-fidelity three-dimensional (3D) environmental awareness, realized through dense RGBD imaging, is an essential prerequisite for advanced functions like obstacle avoidance and scene understanding [1], [2]. These capabilities are crucial for next-generation platforms, including robotics [3], Unmanned Aerial Vehicles (UAVs) [4], and Augmented/Virtual Reality (AR/VR) headsets [5]. However, this critical perceptual requirement stands in direct conflict with the relentless industry push towards miniaturized, lightweight, and low-power imaging systems. Specifically, traditional high-accuracy depth

acquisition methods, such as LiDAR [6], active structured light [7], or binocular stereo cameras [8]–[10], typically require bulky optical baselines or complex emitters, which directly increase system size and power consumption. Concurrently, under the constraints of pursuing compactness and a wide depth-of-field, conventional optical systems struggle to maintain high resolution and sharpness of the RGB image across the entire depth range, limiting the overall performance of the RGBD system. Consequently, achieving high-fidelity RGBD imaging using a single compact imaging module has become a primary objective in the field. While mainstream Monocular Depth Estimation (MDE) [11]–[14] relies on training neural networks on large-scale datasets to infer depth from semantic priors, recovering 3D geometry from a single 2D projection is an inherently ill-posed physical problem [15]. This reliance on semantics is fundamentally vulnerable, often resulting in failures in novel scenes or under ambiguous contextual cues [16]–[18].

A more fundamental methodology in computational imaging directly addresses the ill-posed nature of MDE. This approach involves engineering the system’s PSF to optically encode depth information, thereby rendering the recovery problem well-posed. Specifically, by designing an optical system whose PSF shape varies uniquely with object distance, the 3D geometry is no longer inferred from semantics, but is instead decoded from the physically captured, depth-dependent blur. However, a critical optical trade-off emerges in such monocular depth encoding strategies: maximizing the PSF’s sensitivity to depth often leads to excessive spatial blur, degrading the overall image quality. Existing work has explored this strategy of optical depth encoding by introducing specialized optical elements, such as phase masks [19], [20] and Diffractive Optical Elements (DOEs) [21] to generate complex PSFs dependent on depth. Yet, significant limitations persist in specialized optics-based PSF engineering. While DOEs excel at encoding information, they pose practical challenges. High-precision fabrication, often at sub-wavelength scales, is critical and inherent diffraction can cause energy loss to unwanted orders, reducing throughput and potentially creating artifacts [22]. Additionally, integrating DOEs also requires strict alignment tolerances, increasing assembly complexity and operational sensitivity [23]. This underscores a pivotal research gap: there is a need for a methodology that harnesses robust physical encoding, yet achieves this within a minimalist, all-spherical lens architecture, obviating the requirement for specialized supplementary elements.

To bridge this critical gap, we turn to nature for a more elegant and integrated solution. We first introduce a novel bio-inspired monocentric lens as the core of our computational imaging system. Inspired by the elegant simplicity of aquatic visual systems [24], our all-spherical, compact lens design naturally and intrinsically encodes scene depth into its PSFs, eliminating the need for complex, additive diffractive or freeform elements. Building upon this unique optical front-end, we establish the Bionic Monocentric Imaging (BMI) framework. This framework represents a holistic co-design integrating our compact optics with a dedicated dual-head reconstruction network trained on a physically-

realistic simulation model. Figure 1 visually demonstrates the framework’s core capability: jointly recovering a clear image and corresponding depth map from a single, optically encoded blurred input, highlighting its potential for 3D reconstruction. Our experimental results validate the effectiveness of this bio-inspired design philosophy. We demonstrate through comprehensive experiments that our method achieves a depth RMSE of 0.130, significantly surpassing leading software-only MDE methods and substantially outperforming other deep optics counterparts. Concurrently, our system strikes a state-of-the-art balance between depth precision and image fidelity, pairing this exceptional depth accuracy with top-tier image restoration quality, evidenced by an SSIM score of 0.960 and a perceptual LPIPS of 0.082.

In summary, this work delivers the following main contributions:

- We propose a novel bio-inspired monocentric lens, an all-spherical and lightweight design, to optically encode scene depth into depth-aware Point Spread Functions (PSFs). This establishes a compact optical front-end for computational RGB-D imaging that relies solely on conventional spherical surfaces.
- We establish a Bionic Monocentric Imaging (BMI) framework that integrates physical simulation with a reconstruction network to achieve high-quality RGBD imaging. This approach strikes an excellent balance between restored image fidelity and depth estimation accuracy.
- We verify through comprehensive simulation our system’s ability to obtain superior All-in-Focus (AiF) images and depth maps, and demonstrate its advantages for downstream visual perception tasks.

II. RELATED WORK

A. Paradigms in Bio-inspired Camera Design

Nature has served as a profound source of inspiration for the development of a myriad of bio-inspired optical systems. Through long-term interaction with the environment, creatures have evolved eyes whose diverse functions, such as a wide field of view, adjustable focus, and a deep field, provide valuable inspiration for the design of optical systems. The human eye, a typical chambered structure, includes a cornea, an iris, an adjustable lens, a gelatinous vitreous body, and a curved retina [25]. In particular, the curved retina provides a wide field of view by directly compensating for aberrations in the curved focal plane [26]. In the case of compound eyes, which are composed of thousands of individual photoreceptor units called ommatidia on a curved surface, key advantages include a wide field of view, a deep depth of field, and high sensitivity to motion [27], [28]. For aquatic eyes, fish possess symmetrical spherical lenses [29] because the cornea of fish cannot focus light in water, which gives them a wide field of view up to 160° [30]. To adjust focus, they compensate for the lens’s incompressibility by changing its position rather than its shape [31].

These remarkable biological models have spurred extensive research into novel imaging systems. However, a significant

portion of this work has either concentrated on mimicking the morphological aspects of these eyes, like [32]–[34], or has focused on replicating a singular, isolated function, such as [35]–[37]. Consequently, there has been limited exploration into the deep co-design of a bio-inspired lens’s unique optical properties with computational algorithms for multi-task recovery. Specifically, few studies leverage intrinsic optical features like depth-sensitive PSFs for complex tasks such as joint image restoration and depth estimation.

B. PSF-Aware Depth Estimation

Acquiring depth information is a fundamental task in computer vision. One major paradigm is binocular stereo vision [8]–[10], [38], which computes depth from the parallax between two cameras. Another major paradigm is monocular depth estimation, which infers depth from a single image. This field has progressed from the Convolutional Neural Network (CNN) architecture [39]–[41] to current state-of-the-art models [11]–[14], which achieve remarkable precision at the cost of significant computational resources.

Beyond these purely algorithmic methods, another line of research leverages the intrinsic optical properties of the camera to infer depth. A classic example is Depth from Defocus (DfD), because the amount of defocus blur of an object can be related to its depth, which estimates distance by measuring the sharpness of each pixel [42]–[45]. Modern computational imaging methods directly jointly design camera optics and networks, called deep optics. This is often accomplished by inserting specialized optical elements, like phase masks [19], [20] and freeform lens [46]. Similarly, the Diffractive Optical Element (DOE) is used to encode depth information in the PSF. Baek *et al.* [47] designed a learned DOE to create a PSF that varies with both depth and spectrum, allowing simultaneous single-shot hyperspectral and depth imaging. Ikoma *et al.* [48] proposed a rotationally symmetric DOE and jointly trained the optics with a network using an occlusion-aware image formation model for more accurate blur simulation at depth discontinuities. To address practical deployment challenges, Zhuge *et al.* [21] developed a calibration-free deep optics framework by combining ray tracing and diffraction to precisely simulate both on-axis and off-axis point spread functions, eliminating the need for physical system calibration. Furthermore, Wei *et al.* [49] explored the placement of the DOE, proposing an "off-aperture" encoding scheme to address off-axis aberrations in wide-FoV imaging by enabling local control of the wavefront, thereby achieving RGBD imaging.

Further research has explored various ways to engineer and model depth-aware PSFs for RGB-D tasks. For instance, Qian *et al.* [50] proposed a framework that utilizes depth-aware PSFs for aberration correction and depth estimation to achieve single-lens controllable depth-of-field imaging, whereas Luo *et al.* [51] established a comprehensive 4D-PSF model to guide a similar joint recovery process. Alternative encoding modalities have also been explored, such as in the work by Ghanekar *et al.* [52], which employed polarization to engineer a spiral PSF, separating its lobes to resolve depth ambiguities inherent in traditional rotating PSFs.

Although existing methods are effective, they come with significant overhead in terms of system size, computational cost, and hardware complexity. Our work, inspired by aquatic eyes, introduces a bio-inspired monocentric lens that naturally encodes depth into its PSF. This approach yields a compact and computationally efficient system for depth estimation, eliminating the need for specialized optical elements.

C. Joint Depth Estimation and Image Restoration

In certain applications, particularly in dynamic scene analysis or video processing where depth might be estimated independently or assumed, a known depth map is used to guide image restoration, referred to as depth-aware deblurring [53]–[55]. However, this paradigm is based on the availability of an accurate, pre-existing depth map. In specialized environments, such as underwater imaging [56]–[58], where scattering and blur are significant, or with computational cameras that optically encode depth, the captured image is inherently degraded, with depth information embedded directly within that degradation. This intrinsic link necessitates a joint solution for both depth estimation and image restoration, rather than a sequential approach.

Several methods have been proposed to address this coupled problem. Gur *et al.* [59] proposed a self-supervised method to jointly estimate a depth map and an all-in-focus image from a single defocused input by jointly training two networks for depth estimation and focus rendering, respectively. Anwar *et al.* [60] trained a cascade of two smaller networks to estimate a depth map, which is then used to compute kernels for restoring the AiF image by pixel-wise non-blind deconvolution. Architecturally, Nazir *et al.* [61] employed a shared encoder with two separate decoder heads for depth and deblurring. To further improve the coupling between tasks, Hou *et al.* [62] introduced a unified framework with specialized modules for task-aware fusion and spatial interaction within a shared encoder-dual decoder network. To fully exploit the depth-encoding capability of our bio-inspired fisheye system, we adapt the method proposed in [63] into a multi-task architecture. Specifically, we transform the original single-task deblurring network into a dual-head structure for joint image restoration and depth estimation. Our modified network employs a shared encoder to extract a unified feature representation, which is then simultaneously processed by two separate decoders for joint image restoration and depth estimation, achieving excellent performance on both tasks.

III. METHODOLOGY

This section details the proposed Bionic Monocentric Imaging (BMI) framework, which is centered on our novel, bio-inspired monocentric fisheye lens. This lens naturally encodes depth information into its PSFs, and the framework integrates this unique optical capability with a deep learning-based reconstruction pipeline, as illustrated in Figure 2. We begin in Sec. III-A by presenting the design of our bio-inspired monocentric fisheye lens, which naturally encodes depth information into its PSFs. In Sec. III-B, we describe the physically-based forward model used to generate a synthetic

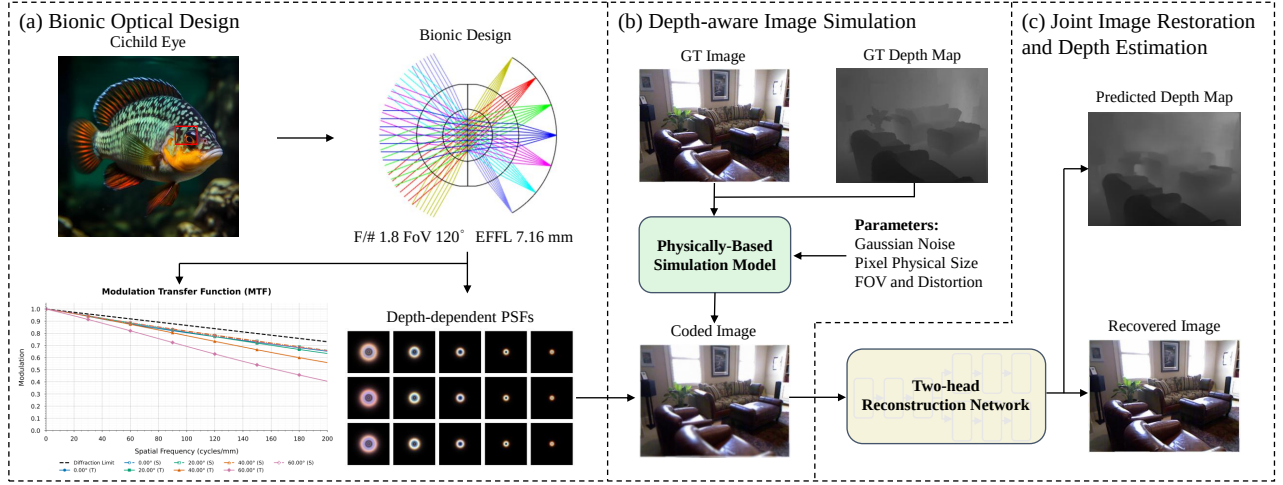


Fig. 2: Overview of the proposed Bionic Monocentric Imaging (BMI) framework. Our method consists of three main stages. (a) Bionic Optical Design: Inspired by the Cichlid Eye, we design a bio-inspired monocentric fisheye lens. The resulting Modulation Transfer Function (MTF) and depth-dependent Point Spread Functions (PSFs) are characterized. (b) Depth-aware Image Simulation: We build a physically-based forward model that uses the characterized PSFs to transform a ground truth (GT) image and its corresponding depth map into a coded image, simulating the degradation introduced by our lens. (c) Joint Image Restoration and Depth Estimation: A two-head reconstruction network takes the coded image as input and is trained to jointly recover a clear, restored image and its corresponding depth map.

training dataset, detailing how the characterized PSFs and an occlusion-aware model create realistically degraded images. Finally, in Sec. III-C, we introduce our designed dual-head reconstruction network, which is established to jointly recover a clear, all-in-focus image and a corresponding depth map from a single coded input.

A. Design of the Bio-inspired Fisheye Lens

Inspired by the acute depth sensitivity inherent in the visual systems of aquatic species [24], we propose a minimalist, compact, and integrated wide Field of View (FoV) optical system. This design aims to overcome a key limitation of conventional imaging systems: their insensitivity to depth cues, particularly in wide FoV scenarios. By integrating the functionalities of RGB texture acquisition and depth ranging into a single optical path, our system effectively resolves issues such as FoV mismatch and insufficient accuracy in depth computation that plague traditional multi-sensor approaches. As a result, our method demonstrates a significant advantage in the joint recovery of high-fidelity RGB images and accurate depth maps, proving its superiority in various downstream tasks.

The design of our imaging system is derived from two core principles of piscine vision. First, emulating the optical properties of a fish’s spherical crystalline lens, the front end of the system employs a customized monocentric lens group. This configuration not only retains a wide 120° FoV, characteristic of fisheye lenses, but also effectively mitigates the peripheral distortion common in such designs through optimization with a curved sensor. This lays a robust foundation for capturing crisp, full-field RGB texture information. Second, by mimicking the core mechanism of depth perception in fish, our system eliminates the need for a separate depth sensor. It enables the simultaneous acquisition of both RGB and depth information

TABLE I: Bionic optical design for the monocentric lens.

Surface	Radius (mm)	Thickness (mm)	Material	Semi-diameter (mm)
1 (Sphere)	4.126	2.100	H-ZLAF3	4.070
2 (Sphere)	2.103	2.000	H-ZPK5	2.100
3 (Stop)	infinity	2.000	H-ZPK5	2.100
4 (Sphere)	-2.103	2.100	H-ZLAF3	2.100
5 (Sphere)	-4.126	3.040	-	4.070
Sensor	-7.199	-	-	0.755

in a single snapshot. The detailed design parameters are presented in Table I.

In terms of functional integration, the system achieves a significantly more compact form factor than conventional imaging systems. This is accomplished through an integrated design of optical components, featuring a minimalist lens structure co-packaged with a curved sensor module. The effective depth sensing range is extended to 10.0m, which not only addresses scenarios requiring fine-grained, close-range perception but meets the requirements of mid-range applications. This versatility makes the system highly adaptable for a variety of real-world scenarios, such as underwater exploration and visual navigation for compact robots.

B. Physically-Based Forward Model and Dataset Simulation

This section outlines the generation of a synthetic dataset for our proposed fisheye lens, a process that forms the critical link between the optical design and the reconstruction algorithm. We first generate a PSF map characterizing the lens’s behavior across various depths and spatial positions, which is then used in a physically-based simulation to encode depth-dependent blur into our training images.

PSF map construction. The PSF characterizes the spatially varying aberrations of the fisheye lens as a function of object depth and field position. PSFs at different depths and spatial positions are computed using ZEMAX [64] with internal lens

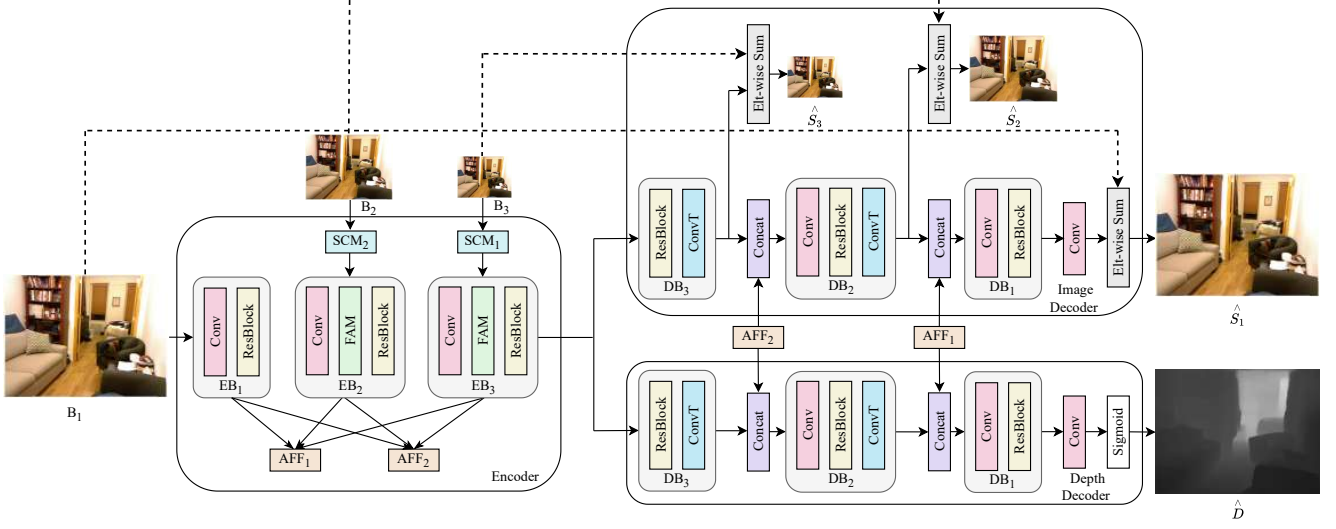


Fig. 3: The architecture of our reconstruction network for joint image restoration and depth estimation. The network utilizes a shared encoder to extract unified features from multi-scale input. These features are then fed into two separate decoder heads—one for multi-scale image restoration and the other for depth estimation—enabling the joint recovery of both tasks.

data. Specifically, we simulate PSFs across a depth range from $0.7m$ to $10.0m$ with $0.1m$ interval. We must clarify that while the optical design itself is fully capable of a 120° field of view, our experimental simulation is intentionally limited to a 6° half-field angle. This is not a limitation of the optics, but a necessary constraint for validation. This choice ensures compatibility with the target 640×480 sensor ($2.0\mu m$ pixel pitch) and mitigates the manufacturing challenges of highly curved sensors [26], [65], [66]. Crucially, it allows for a direct, “apples-to-apples” quantitative comparison against the standard, non-wide-angle NYU Depth V2 benchmark dataset. Evaluating the full 120° FoV would require a different sensor configuration and a dedicated wide-angle benchmark dataset, which is a key direction for future work. For three principal wavelengths, the PSFs were sampled on a 128×128 grid with a $0.4\mu m$ pixel pitch to capture over 99.9% of the energy, yielding a comprehensive PSF tensor, $PSF(c, \theta, d)$ that maps the lens’s response across depth, field, and wavelength.

For an axisymmetric system such as our fisheye lens, the full RGB PSF map can be accurately and efficiently generated from the initial PSF tensor, $PSF(c, \theta, d)$. Following the approach in [67], we employ interpolation, rotation, and resizing operations on the characterized PSFs to synthesize the complete, spatially-varying response for any scene point as:

$$PSF_{map}(c, h, w, d) = P_{resize} \circ P_{rot} \left(\sum_{\theta} W(\theta) \cdot PSF(c, \theta, d) \right), \quad (1)$$

where P_{resize} denotes the resizing operation to match the sampling pitch with the sensor’s pixel size, P_{rot} denotes the rotation operator, and $W(\theta)$ represents the normalized interpolation weights determined by an inverse square law.

Depth-aware Image Simulation with Occlusion. Conventional methods often simplify the computation of spatially varying blur by partitioning the image into patches and convolving each with a single, uniform PSF [50]. However, this patch-wise approximation introduces significant artifacts at depth discontinuities as it struggles to accurately model the abrupt kernel changes. For a more physically plausible simulation of the image formation process, we employ an image formation model with occlusion [48]. This model discretizes the scene into K distinct depth layers, where I_k and α_k represent the image content and a binary alpha mask for the k -th layer, respectively. By integrating this framework with our designed wavelength- and depth-dependent PSFs, the final image formation is expressed as:

$$P(c) = \sum_{k=0}^{K-1} \tilde{I}_k \prod_{k'=k+1}^{K-1} (1 - \tilde{\alpha}_{k'}) + \eta, \quad (2)$$

where $\tilde{I}_k = \frac{PSF_k(c) * I_k}{E_k(c)}$ and $\tilde{\alpha}_k = \frac{PSF_k(c) * \alpha_k}{E_k(c)}$. The term η represents additive Gaussian noise, and $*$ denotes the convolution operator. The normalization term $E_k(c) = PSF(c) * \sum_{k'=0}^k \alpha_{k'}$ preserves energy conservation across layers. For our implementation, we simulate at the principal wavelengths for the R, G, and B channels ($656.3nm$, $587.6nm$, and $486.1nm$). The scene’s depth, ranging from $0.7m$ to $10.0m$, is discretized into $K=94$ layers with $0.1m$ interval. To efficiently manage the spatially varying nature of the blur, all convolutions are accelerated via the Fast Fourier Transform (FFT) on 40×40 pixel sub-images.

C. Joint Image Restoration and Depth Estimation Network

Given the unique optical properties of our fisheye lens, the scene depth information and the resulting image degradation are intrinsically coupled within the PSF. This inherent

link allows for the simultaneous extraction of a depth map and restoration of a clear, all-in-focus image from a single captured frame. To accomplish this, we adapt a multi-scale image restoration network, MIMOUNet [63], into a dual-head architecture designed [62] for the joint task of depth estimation and image restoration.

As illustrated in Figure 3, our network adopts a multi-scale input strategy. The original coded image, B_1 , is downsampled by factors of 2 and 4 to generate B_2 and B_3 , respectively, with all three scales serving as the network input. The architecture features two branches for image restoration and depth estimation, which share a common encoder and feature fusion modules. This design ultimately produces multi-scale, all-in-focus image outputs and a final depth map prediction.

The training loss function L_{total} consists of image loss L_{cont} and L_{MSFR} as well as depth loss L_{Silog} as:

$$L_{total} = \gamma_{cont}L_{cont} + \gamma_{MSFR}L_{MSFR} + \gamma_{silog}L_{silog}. \quad (3)$$

For the image restoration branch, we employ a multi-scale content loss L_{cont} , defined as the L_1 distance between the restored images and the ground-truth images at each scale [68], as shown in Eq. (4). Furthermore, since a primary goal of deblurring is the recovery of high-frequency information, we introduce a multi-scale frequency-domain loss L_{MSFR} . This loss computes the L1 distance between the Fourier transforms of the restored and ground-truth images at each scale [69], as defined in Eq. (5).

$$L_{cont} = \sum_{k=1}^K \frac{1}{t_k} \left\| \hat{S}_k - S_k \right\|_1. \quad (4)$$

$$L_{MSFR} = \sum_{k=1}^K \frac{1}{t_k} \left\| \mathcal{F}(\hat{S}_k) - \mathcal{F}(S_k) \right\|_1. \quad (5)$$

For the depth estimation branch, we employ the scale invariant log error loss L_{Silog} [39]. This loss function is invariant to absolute global scale, focusing instead on penalizing errors in relative depth relationships. This property leads to significantly enhanced training stability. The L_{Silog} is defined as:

$$L_{Silog} = \frac{1}{N} \sum_{i=1}^N (\log d_i - \log \hat{d}_i) - \frac{1}{N^2} \sum_{i=1}^N (\log d_i - \log \hat{d}_i)^2. \quad (6)$$

IV. EXPERIMENTS

A. Optical Simulation and Datasets

Our experiments are based on a synthetic dataset generated using the optical properties of the proposed bio-inspired lens. As shown in Figure 4, the resulting PSFs exhibit a pronounced and systematic variation with object depth, transitioning from a large ring structure at close distances to a compact point for distant objects. This distinct depth-dependent property is the key to optically encoding scene information, forming the basis for our joint restoration and estimation tasks.

To generate the dataset, we apply our physically-based forward model, which incorporates occlusion handling, to the

NYU Depth V2 dataset [70]. This dataset is selected because it is a widely recognized and comprehensive benchmark for understanding indoor scenes, providing a standard protocol for training and evaluation that ensures a fair comparison with other state-of-the-art methods. Furthermore, its diversity is a key advantage, as it comprises 464 different scenes from a wide range of residential and commercial buildings. The high-quality RGB images, corresponding depth maps, and dense annotations make it particularly well-suited for the rigorous training and evaluation of our joint restoration and depth estimation framework. Following the official protocol, we utilize 754 images for training and validation, and the standard 654 images for testing. We benchmark our simulation against a conventional patch-wise method, where an image is divided into 16×16 pixel patches, each convolved with a single corresponding PSF.

To ensure the physical realism of our synthetic data, we employ an occlusion-aware image formation model rather than simpler approximations like the conventional patch-wise method. To quantitatively verify the superiority of the occlusion-aware model over the patch-wise approach in terms of physical realism, particularly regarding artifacts at depth discontinuities, we develop a specific metric, termed the Artifact Score (AS). The detailed definition, rationale, and computation of the AS, along with both visual and quantitative comparative results confirming the superiority of our chosen simulation method, are provided in Appendix B. This rigorous and validated simulation process yields a high-quality dataset crucial for training our reconstruction network effectively.

B. Implementation Details

All experiments are conducted on a single NVIDIA A800 GPU. The reconstruction network is trained using the Adam optimizer [71] with hyperparameters set to $\beta_1=0.9$ and $\beta_2=0.99$. We employ an initial learning rate of $5e-5$ and a batch size of 8. To improve model generalization, we apply data augmentation techniques, including random horizontal flips and rotations. During training, randomly cropped 256×256 patches are utilized as inputs. To balance the contributions of the image restoration and depth estimation tasks, the weights for the content loss L_{cont} , the frequency-domain loss L_{MSFR} , and the Silog loss L_{Silog} are set to 1.0, 0.1, 0.1, respectively. The model was trained for a total of 200,000 iterations.

C. Results and Comparative Analyses

We conduct our quantitative evaluation on the NYU Depth V2 dataset [70]. On this benchmark, our method is compared against state-of-the-art Monocular Depth Estimation (MDE) approaches, deep optics systems, and other optical designs. As detailed in Table II, performance is assessed using standard metrics for depth accuracy and image quality.

Evaluation metrics. We assess performance using a comprehensive set of standard metrics. For depth estimation accuracy, we report the following: Threshold: % of pixels s.t. $\max(\hat{y}_i/y_i, y_i/\hat{y}_i) < \text{thr}$, which measures the percentage of reliable pixel predictions; Absolute Relative Error (Abs Rel):

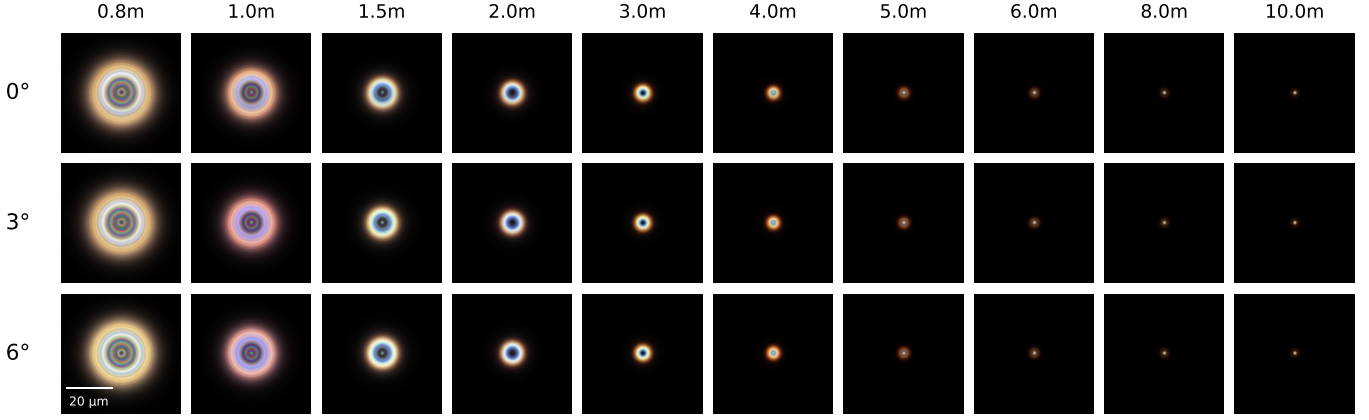


Fig. 4: Simulated PSFs of bio-inspired lens. The PSFs are shown for three different fields of view (rows: 0° , 3° , 6°) and ten object depths (columns: $0.8m$ to $10.0m$). Each PSF is visualized from a 128×128 data array. For better visualization, the intensity of each PSF has been normalized.

TABLE II: Quantitative comparison of joint image restoration and depth estimation. This table presents a detailed quantitative evaluation of our method against other state-of-the-art approaches, showcasing performance across various metrics for depth accuracy (δ , RMSE, Abs Rel) and image quality (PSNR, SSIM, LPIPS).

Method	Depth Accuracy $\uparrow$, $\delta <$			Depth Error $\downarrow$		Image Error			Others
	1.25	1.25^2	1.25^3	RMSE	Abs Rel	PSNR(dB) $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	
ZoeDepth [72]	0.955	0.995	0.999	0.270	0.075	-	-	-	-
VPD [73]	0.964	0.995	0.999	0.254	0.069	-	-	-	-
ECoDepth [74]	0.978	0.997	0.999	0.218	0.048	-	-	-	-
Depthanything [11]	0.984	0.998	1.000	0.206	0.056	-	-	-	-
Metric3Dv2 [12]	0.989	0.998	1.000	0.180	0.046	-	-	-	-
DeepOptics [46]	0.930	0.990	0.999	0.433	0.087	-	-	-	-
Phase3D [19]	0.932	0.989	0.997	0.382	0.093	-	-	-	0.01
Learnedoptics [48]	0.959	0.990	0.996	0.439	0.070	28.54	-	-	0.005
CF-DOE [21]	0.987	0.998	1.000	0.225	0.044	32.11	0.917	-	0.005
Doublegauss	0.977	0.996	0.999	0.183	0.042	25.74	0.832	0.272	0.005
Fresnel [75]	0.989	0.998	0.999	0.133	0.032	26.77	0.884	0.207	0.005
CF-DOE [21]+Ours(Network)	0.989	0.999	1.000	0.211	0.045	32.29	0.922	0.164	0.005
Ours	0.996	0.999	1.000	0.130	0.026	31.36	0.960	0.082	0.005

$\frac{1}{|\mathcal{T}|} \sum |\hat{y} - y|/y$, a scale-invariant metric for the mean error; and Root Mean Square Error (RMSE): $\sqrt{\frac{1}{|\mathcal{T}|} \sum (\hat{y} - y)^2}$, which is particularly sensitive to large outliers. For image restoration quality, we evaluate the Peak Signal-to-Noise Ratio (PSNR) to quantify pixel-wise fidelity; the Structural Similarity Index Measure (SSIM) [76] to assess structural similarity, and the Learned Perceptual Image Patch Similarity (LPIPS) [77] to measure perceptual distance in a deep feature space. In these metrics, $\hat{y}$ and y denote the predicted and ground-truth values, respectively, and $\mathcal{T}$ represents the set of valid pixels for evaluation.

Comparative performance analysis. As detailed in Table II, our comparative analysis is structured into three distinct categories to comprehensively evaluate our system’s performance. First, we benchmark our complete end-to-end pipeline (from optical simulation to final reconstruction) against state-of-the-art, software-only Monocular Depth Estimation (MDE) methods. Although these methods benefit from pristine, artifact-free input images, our system still outperforms them across

all depth error metrics. For instance, our approach achieves a significantly lower Absolute Relative Error (Abs Rel) of 0.026 and a Root Mean Square Error (RMSE) of 0.130, surpassing top-performing methods like Metric3Dv2 [12] (0.046 Abs Rel, 0.180 RMSE). This demonstrates that the physical depth cues encoded by our bio-inspired optics provide a tangible advantage over purely algorithmic inference.

Second, we compare our integrated system against other complete deep optics frameworks to evaluate the end-to-end performance. Our method strikes a superior balance between depth accuracy and image restoration quality. When compared to the entire system of CF-DOE [21], while their system achieves a slightly higher PSNR (32.11dB vs. our 31.36dB), our approach excels in depth estimation, significantly reducing the RMSE from their reported 0.225 to our 0.130. Furthermore, our method achieves the highest SSIM at 0.960, indicating best-in-class performance in recovering structural image quality. This highlights the effectiveness of our minimalist, bio-inspired design in achieving robust performance without complex hardware.

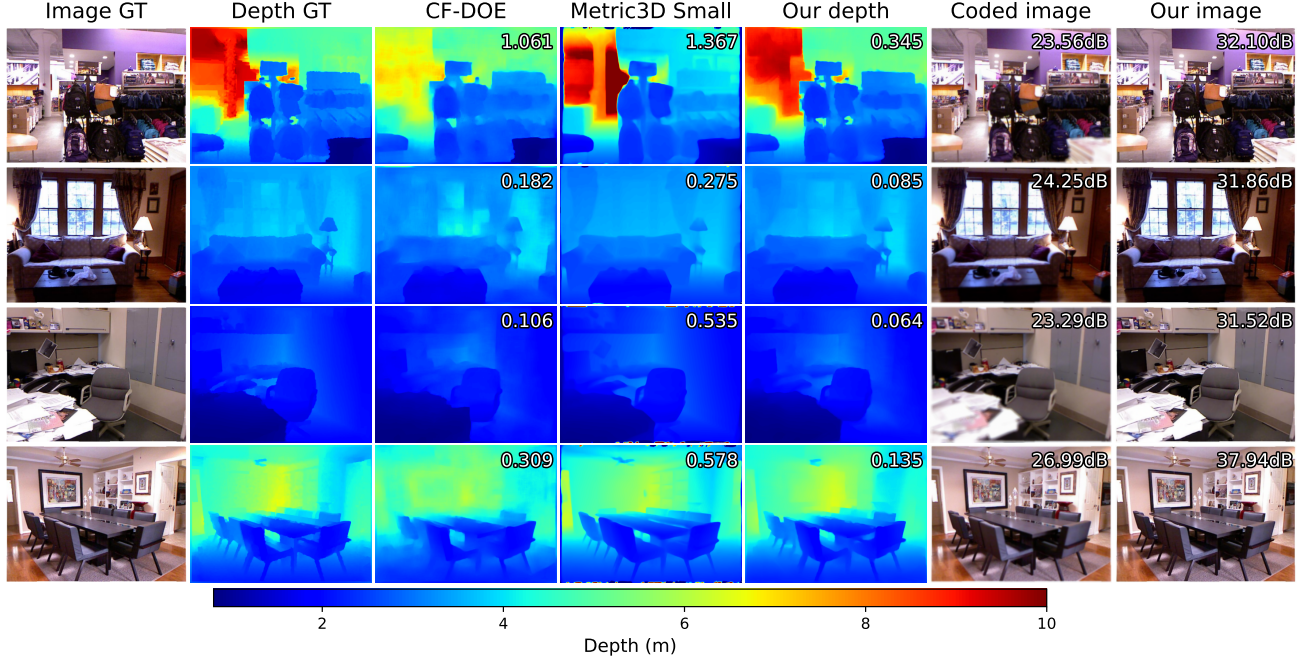


Fig. 5: Qualitative comparison of our method against other approaches on the NYU Depth V2 dataset, such as CF-DOE [21] and Metric3D Small [12]. The RMSEs of depth maps or the PSNRs of images compared with GTs are noted in the upper right corner. Our method produces depth maps with fewer artifacts and restored images with higher clarity and fidelity.

Finally, to specifically isolate and unequivocally demonstrate the superiority of our optical design itself, we conduct a rigorous ‘apples-to-apples’ comparison. In this controlled experiment, we kept the reconstruction network and simulation pipeline entirely fixed, only varying the front-end optics. To ensure a fair and meaningful comparison, we selected baselines that represent distinct categories of optical encoding. We compare against: (1) A conventional Doublegauss lens (see Appendix A), which, with its characteristic shallow depth-of-field, represents the classic Depth-from-Defocus (DfD) encoding approach. (2) Alternative computational imaging front-ends, specifically a Fresnel lens design [75] and the optical front-end from CF-DOE [21]. This methodology allows us to benchmark our novel encoding strategy not only against traditional defocus cues (Doublegauss) but also against other modern deep optics solutions (Fresnel, CF-DOE). As detailed at the bottom of Table II, our bio-inspired lens design comprehensively outperforms all other optics. This provides definitive proof that the core contribution to our system’s superior performance stems from the unique advantages of our bio-inspired optical design, which provides a higher-quality data foundation for the subsequent computational task.

Qualitative result analysis. Figure 5 presents the qualitative results on the NYU dataset. While our method incorporates a depth-aware model with occlusion, the resulting depth maps exhibit sharpness at certain object boundaries, a limitation partially attributable to the FFT-based patch implementation used for computational acceleration. Nevertheless, our approach demonstrates a clear advantage in quantitative evaluations over methods like Metric3D [12]. We also observe that the depth estimation accuracy degrades for distant objects. This is because the PSF’s variation with respect to depth becomes

less pronounced at greater distances, a phenomenon visible in the far-field regions of the living room scene in the first row. Figure 6 offers a detailed qualitative comparison of the image restoration performance. As highlighted by the magnified regions within the red boxes, our method demonstrates a superior capability in recovering fine image details. This is particularly evident in challenging textures, such as those in curtains, bookshelves, and ornaments, where our approach restores sharpness and clarity more effectively than the comparative methods.

D. Applications

To further validate the effectiveness and practical utility of our Bionic Monocentric Imaging framework beyond the benchmark comparisons on the NYU dataset, we demonstrate its application in several downstream tasks of significant practical relevance.

Application to 3D scene reconstruction. The ability of our system to jointly provide high-fidelity images and accurate depth maps makes it highly suitable for 3D scene reconstruction. Following the methodology for generating 3D point clouds from monocular depth data [78], we utilize our restored images and predicted depth maps to reconstruct the scenes. The results are visualized from two distinct perspectives in Figure 7. A frontal view (View A) effectively showcases the high quality of the restored texture from our image branch, whereas an oblique top-down view (View B) demonstrates the geometric accuracy of the reconstructed surfaces, thereby validating our depth estimation.

Application to RGBD semantic segmentation. To quantitatively evaluate our system’s utility in downstream tasks, we apply its outputs to RGBD semantic segmentation using

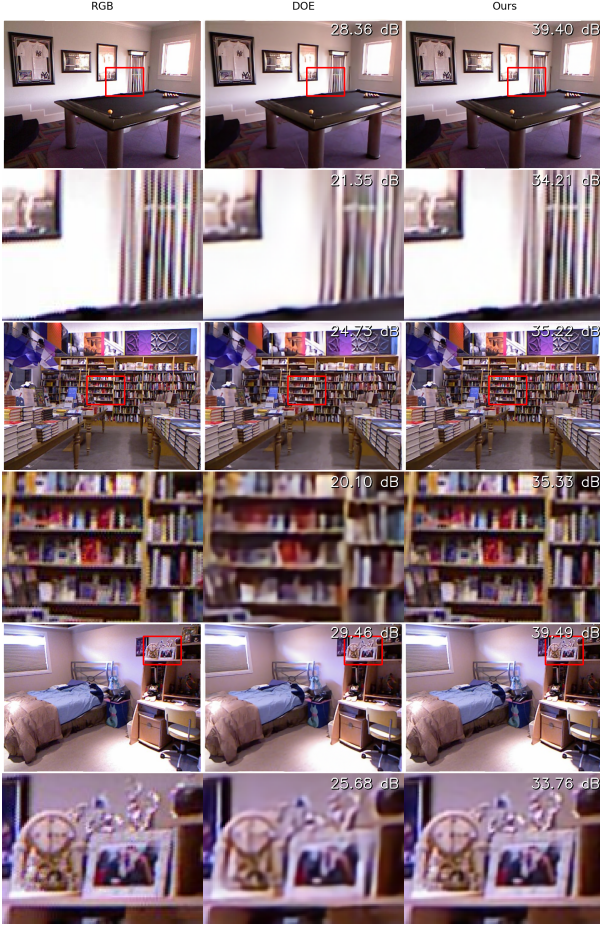


Fig. 6: Enlarged qualitative comparison for image restoration. The magnified regions, indicated by red boxes, compare our method with the DOE-based approach and the Ground Truth (GT).

ESANet [79], a lightweight framework known for its efficient fusion of multi-modal data. For a fair yet challenging assessment, our system is fed its own restored images and predicted depth maps. In contrast, baseline MDE methods are paired with original, pristine ground-truth RGB images. The qualitative results are presented in Figure 8. Although the quality of our restored images is slightly inferior to the original ground-truth, leading to minor performance degradation in complex scenes (e.g., the cluttered room in the first row) or on fine details (e.g., the bookshelf in the fourth row), our system results are highly competitive with software-only algorithms that benefit from pristine image inputs. This demonstrates the high quality of our jointly recovered image and depth data for practical applications. The quantitative evaluation of RGBD semantic segmentation is presented in Table III, using the ground-truth data from the NYU dataset as an upper-bound benchmark. Performance is measured by Overall Accuracy (OA), mean Accuracy (mAcc), mean Intersection over Union (mIoU), and model Parameters (#Params). When compared to other methods, our approach achieves highly competitive results (e.g., 41.17% mIoU for Ours vs. 42.19% for Metric3Dv2-giant) with a drastically smaller parameter count. More importantly, when isolating the front-end optics by fixing the segmentation network, our bis-inspired lens sig-

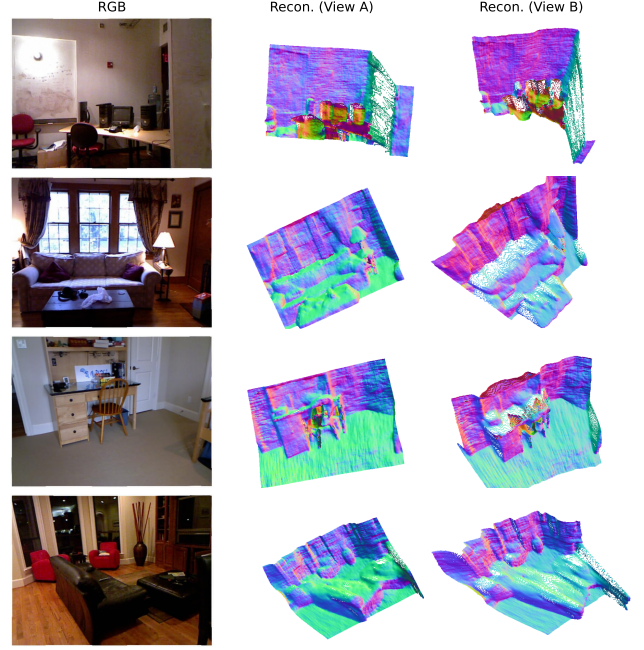


Fig. 7: Qualitative 3D reconstruction results. The figure shows the recovered scene geometry and texture from two different viewpoints (View A and View B).

TABLE III: Quantitative evaluation of RGBD semantic segmentation. Performance is compared on the NYU Depth V2 dataset using metrics including Overall Accuracy (OA), mean Accuracy (mAcc), mean Intersection over Union (mIoU), and model parameter count (#Params).

Method	OA (%) $\uparrow$	mAcc (%) $\uparrow$	mIoU (%) $\uparrow$	#Params (M) $\downarrow$
Metric3Dv1-Large [80]	68.07	50.13	38.00	203.25
Metric3Dv2-small [12]	68.51	53.46	39.66	37.50
Metric3Dv2-giant [12]	70.02	56.49	42.19	1377.67
CF-DOE [21]	64.63	48.51	35.24	51.96
Fresnel [75]	65.75	48.86	35.85	-
Doublegauss	60.40	42.68	30.23	-
CF-DOE [21]+Ours(Network)	65.14	48.55	35.63	-
Ours	69.95	54.89	41.17	10.08
NYU-GT	77.02	64.68	51.59	-

nificantly outperforms systems based on conventional elements like Fresnel or Doublegauss lenses, demonstrating a clear advantage in data quality for this downstream task.

Application to underwater imaging. Motivated by the aquatic-life inspiration for our lens (the cichlid fish), we conduct an out-of-distribution robustness test to evaluate the framework’s performance in a challenging underwater environment. This experiment is conducted in a strict zero-shot manner; the framework is not retrained on any underwater data and relies solely on the model trained with our “in-air” physical model and the NYU dataset. We test this pre-trained model on a subset of the USOD10K dataset [81], specifically selecting images with a depth range of $0.8m \sim 10.0m$. The qualitative results, presented in Figure 9, demonstrate a remarkable ability to restore clarity by effectively removing the typical color cast and scattering artifacts present in the source images. The corresponding depth maps accurately capture the geometry of underwater objects. These visual findings are corroborated by the strong quantitative metrics in Table IV, which report a high

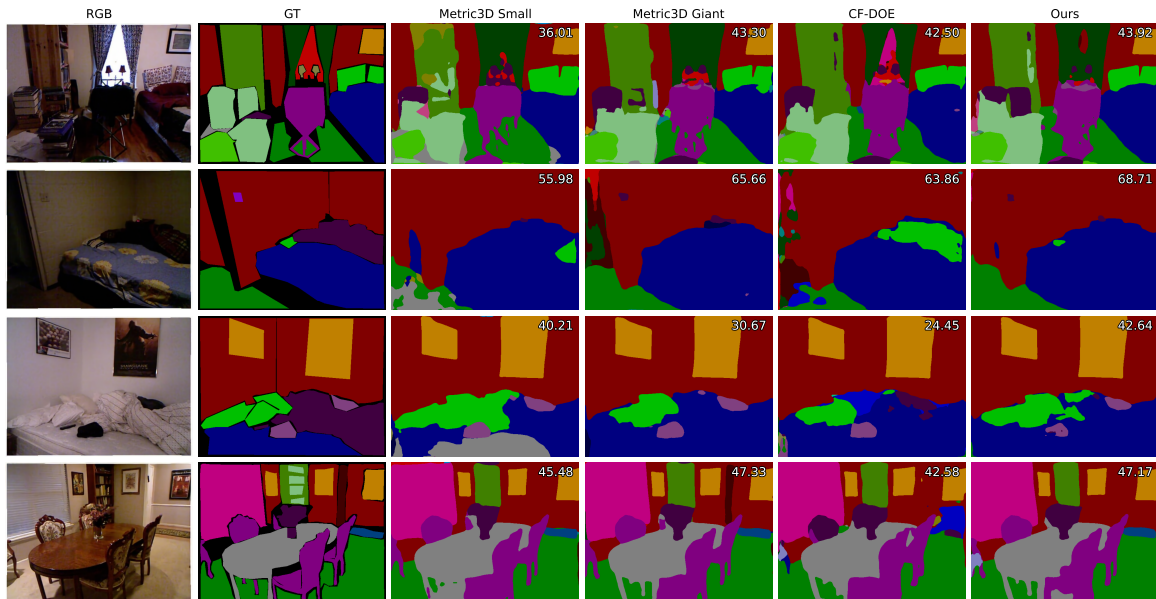


Fig. 8: Qualitative results for RGBD semantic segmentation. The maps highlight how our approach produces more accurate and coherent segmentation boundaries than others. The mIoU scores of segmentation compared with GTs are noted in the upper right corner.

TABLE IV: Quantitative results for underwater depth estimation and image restoration. The table details the quantitative results for our method in underwater scenes.

Depth Estimation	$\delta < 1.25 \uparrow$	RMSE $\downarrow$	Abs Rel $\downarrow$
Ours Depth	0.913	0.354	0.076
Image Restoration	PSNR(dB) $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$
Ours Image	34.21	0.917	0.110

PSNR of $34.21dB$ for image restoration and a depth accuracy ($\delta < 1.25$) of 0.913.

This strong performance on challenging, out-of-distribution underwater conditions serves as a significant robustness test for our framework. It is noteworthy that the features our network learned to deconvolve the specific, physically-based blur from our “in-air” model also demonstrated an ability to mitigate degradation (such as scattering and color cast) prevalent in underwater scenes. We must clarify that this result does not constitute a rigorous validation for aquatic physics, as the underlying physical model and training data were exclusively “in-air”. Nevertheless, this successful zero-shot generalization provides compelling evidence for the core advantage of our BMI framework. It indicates that the network learned to invert the physical degradation encoded by our bio-inspired lens, rather than merely overfitting to the training data. This inherent robustness, stemming from our holistic co-design of physics-aware optics and a dedicated reconstruction algorithm, underscores the framework’s effectiveness and its potential for real-world applicability.

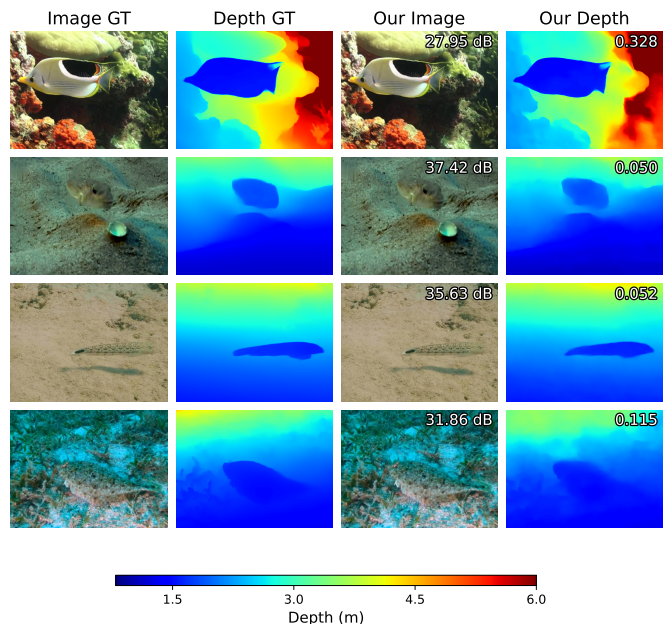


Fig. 9: Qualitative results of the bio-inspired monocentric lens for underwater imaging. This figure illustrates the image and depth outputs of our system in an underwater context.

V. CONCLUSION AND DISCUSSION

A. Conclusion

In this paper, we introduce the Bionic Monocentric Imaging (BMI) framework, a novel computational imaging approach for high-quality joint image restoration and depth estimation. Our framework uniquely leverages a bio-inspired, all-spherical monocentric lens that optically embeds depth cues into spatially-varying Point Spread Functions (PSFs). A deep reconstruction network, trained on a physically-realistic simu-

lation dataset, then encodes this optically-encoded information to simultaneously output a high-fidelity all-in-focus image and a precise depth map. Our comprehensive evaluation reveals the superiority of our optical encoding strategy. The physical depth cues captured by our lens empower our system to not only surpass leading software-only depth estimation techniques but also strike a state-of-the-art balance between image fidelity and depth accuracy when compared to other deep optics systems. Furthermore, the framework's practical viability and robustness are confirmed by its strong performance on downstream vision tasks, highlighted by its remarkable zero-shot generalization to challenging underwater environments. These results provide a compelling validation of our bio-inspired design philosophy.

B. Discussion and Future Work

This work demonstrates a key principle in computational imaging: a meticulously designed optical front-end, which enhances information quality at the physical source, can be a more efficient and promising technical route than relying solely on increasingly complex algorithms. However, this study also opens up several avenues for future research and exploration.

Deeper optical principles. The bio-inspired nature of our lens invites a more fundamental optical analysis. Although its effective depth encoding capabilities have been validated empirically, future work could delve into the precise theoretical relationship between depth encoding and specific aberrations such as spherical aberration and field curvature. Establishing this rigorous foundation would not only deepen our understanding but also guide the design of next-generation optical encoding systems.

Field-of-View extension. From a practical point of view, our current experimental validation is limited by constraints in available benchmark datasets and sensor dimensions, which, as discussed in Sec. III-B, necessitated limiting our simulation. A significant direction for future development is therefore to validate our framework at the larger field of view (FOV) that our monocentric design is fully capable of. Exploring the system's performance at wider angles would unlock its full potential for applications like robotics and autonomous navigation.

Physics-Informed reconstruction. On the algorithmic front, our reconstruction network could benefit from a deeper integration of optical priors. Future research could focus on designing architectures that explicitly incorporate physical information, such as the known characteristics of the PSFs, into their structure. This could lead to a more interpretable recovery framework and potentially yield even greater performance in both restoration and depth estimation tasks.

REFERENCES

- [1] N. Robinson, B. Tidd, D. Campbell, D. Kulić, and P. Corke, "Robotic vision for human-robot interaction and collaboration: A survey and systematic review," *ACM Transactions on Human-Robot Interaction*, vol. 12, no. 1, pp. 1–66, 2023.
- [2] W. Zhou, Y. Yue, M. Fang, X. Qian, R. Yang, and L. Yu, "BCINet: Bilateral cross-modal interaction network for indoor scene understanding in RGB-D images," *Information Fusion*, vol. 94, pp. 32–42, 2023.
- [3] V. K. Bandari and O. G. Schmidt, "System-engineered miniaturized robots: From structure to intelligence," *Advanced Intelligent Systems*, vol. 3, no. 10, p. 2000284, 2021.
- [4] P. G. Fahlstrom, T. J. Gleason, and M. H. Sadraey, *Introduction to UAV systems*. John Wiley & Sons, 2022.
- [5] E.-L. Hsiang, Z. Yang, Q. Yang, P.-C. Lai, C.-L. Lin, and S.-T. Wu, "AR/VR light engines: perspectives and challenges," *Advances in Optics and Photonics*, vol. 14, no. 4, pp. 783–861, 2022.
- [6] N. Li *et al.*, "A progress review on solid-state LiDAR and nanophotonics-based LiDAR sensors," *Laser & Photonics Reviews*, vol. 16, no. 11, p. 2100511, 2022.
- [7] A. Forbes, M. De Oliveira, and M. R. Dennis, "Structured light," *Nature Photonics*, vol. 15, no. 4, pp. 253–262, 2021.
- [8] J.-R. Chang and Y.-S. Chen, "Pyramid stereo matching network," in *Proc. CVPR*, 2018, pp. 5410–5418.
- [9] V. Tankovich, C. Hane, Y. Zhang, A. Kowdle, S. Fanello, and S. Bouaziz, "HITNet: Hierarchical iterative tile refinement network for real-time stereo matching," in *Proc. CVPR*, 2021, pp. 14 362–14 372.
- [10] J. Cheng *et al.*, "MonSter: Marry monodepth to stereo unleashes power," in *Proc. CVPR*, 2025, pp. 6273–6282.
- [11] L. Yang, B. Kang, Z. Huang, X. Xu, J. Feng, and H. Zhao, "Depth anything: Unleashing the power of large-scale unlabeled data," in *Proc. CVPR*, 2024, pp. 10 371–10 381.
- [12] M. Hu *et al.*, "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*, vol. 46, no. 12, pp. 10 579–10 596, 2024.
- [13] A. Bochkovskii *et al.*, "Depth pro: Sharp monocular metric depth in less than a second," in *Proc. ICLR*, 2025.
- [14] L. Yang *et al.*, "Depth anything V2," in *Proc. NeurIPS*, vol. 37, 2024, pp. 21 875–21 911.
- [15] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.
- [16] Y. Zhao, S. Kong, D. Shin, and C. C. Fowlkes, "Domain decluttering: Simplifying images to mitigate synthetic-real domain shift and improve depth estimation," in *Proc. CVPR*, 2020, pp. 3327–3337.
- [17] M. Johnson-Roberson, C. Barto, R. Mehta, S. N. Sridhar, K. Rosaen, and R. Vasudevan, "Driving in the matrix: Can virtual worlds replace human-generated annotations for real world tasks?" in *Proc. ICRA*, 2017, pp. 746–753.
- [18] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari, "Robust monocular depth estimation under challenging conditions," in *Proc. ICCV*, 2023, pp. 8143–8152.
- [19] Y. Wu, V. Boominathan, H. Chen, A. Sankaranarayanan, and A. Veeraraghavan, "PhaseCam3D — Learning phase masks for passive single view depth estimation," in *Proc. ICCP*, 2019, pp. 1–12.
- [20] M. Mel, M. Siddiqui, and P. Zanuttigh, "End-to-end learning for joint depth and image reconstruction from diffracted rotation," *The Visual Computer*, vol. 40, no. 9, pp. 5961–5977, 2024.
- [21] Z. Zhuge *et al.*, "Calibration-free deep optics for depth estimation with precise simulation," *Optics and Lasers in Engineering*, vol. 180, p. 108313, 2024.
- [22] U. Levy, D. Mendlovic, and E. Marom, "Efficiency analysis of diffractive lenses," *Journal of the Optical Society of America A*, vol. 18, no. 1, pp. 86–93, 2001.
- [23] S. N. Khonina, N. L. Kazanskiy, R. V. Skidanov, and M. A. Butt, "Advancements and applications of diffractive optical elements in contemporary optics: A comprehensive overview," *Advanced Materials Technologies*, vol. 10, no. 4, p. 2401028, 2025.
- [24] M. Kim *et al.*, "An aquatic-vision-inspired camera based on a monocentric lens and a silicon nanorod photodiode array," *Nature Electronics*, vol. 3, no. 9, pp. 546–553, 2020.
- [25] D. Atchison, *Optics of the human eye*. CRC Press, 2023.
- [26] W. Gao, Z. Xu, X. Han, and C. Pan, "Recent advances in curved image sensor arrays for bioinspired vision system," *Nano Today*, vol. 42, p. 101366, 2022.
- [27] D. Floreano, J.-C. Zufferey, M. V. Srinivasan, and C. Ellington, *Flying insects and robots*. Springer, 2009.
- [28] J. Herault, *Biologically inspired computer vision: fundamentals and applications*. John Wiley & Sons, 2015.
- [29] Z.-P. He, X. Han, W.-Q. Wu, Z.-S. Xu, and C.-F. Pan, "Recent advances in bioinspired vision systems with curved imaging structures," *Rare Metals*, vol. 43, no. 4, pp. 1407–1434, 2024.
- [30] W. S. Jagger and P. Sands, "A wide-angle gradient index optical model of the crystalline lens and eye of the octopus," *Vision Research*, vol. 39, no. 17, pp. 2841–2852, 1999.

- [31] M. F. Land, "The optics of animal eyes," *Contemporary Physics*, vol. 29, no. 5, pp. 435–455, 1988.
- [32] Y. M. Song *et al.*, "Digital cameras with designs inspired by the arthropod eye," *Nature*, vol. 497, no. 7447, pp. 95–99, 2013.
- [33] K. Zhang *et al.*, "Origami silicon optoelectronics for hemispherical electronic eye systems," *Nature Communications*, vol. 8, no. 1, p. 1782, 2017.
- [34] E. K. Lee *et al.*, "Fractal web design of a hemispherical photodetector array with organic-dye-sensitized graphene hybrid composites," *Advanced Materials*, vol. 32, no. 46, p. 2004456, 2020.
- [35] M. Ott, "Visual accommodation in vertebrates: mechanisms, physiological response and stimuli," *Journal of Comparative Physiology A*, vol. 192, no. 2, pp. 97–111, 2006.
- [36] L. Li, Q.-H. Wang, and W. Jiang, "Liquid lens with double tunable surfaces for large power tunability and improved optical performance," *Journal of Optics*, vol. 13, no. 11, p. 115503, 2011.
- [37] Z. Rao *et al.*, "Curvy, shape-adaptive imagers based on printed optoelectronic pixels with a kirigami design," *Nature Electronics*, vol. 4, no. 7, pp. 513–521, 2021.
- [38] L. Ou, Y. Liu, X. Bai, and Y. Peng, "Learning RGBD imaging via asymmetrically focused stereo cameras," *The Visual Computer*, pp. 1–13, 2025.
- [39] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *Proc. NeurIPS*, vol. 27, 2014, pp. 2366–2374.
- [40] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 3DV*, 2016, pp. 239–248.
- [41] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, "Deep ordinal regression network for monocular depth estimation," in *Proc. CVPR*, 2018, pp. 2002–2011.
- [42] P. Favaro, A. Mennucci, and S. Soatto, "Observing shape from defocused images," *International Journal of Computer Vision*, vol. 52, no. 1, pp. 25–43, 2003.
- [43] A. Chakrabarti and T. Zickler, "Depth and deblurring from a spectrally-varying depth-of-field," in *Proc. ECCV*, vol. 7576, 2012, pp. 648–661.
- [44] M. Carvalho, B. L. Saux, P. Trounev-Peloux, A. Almansa, and F. Champagnat, "Deep depth from defocus: How can defocus blur improve 3D estimation using dense neural networks?" in *Proc. ECCVW*, vol. 11129, 2018, pp. 307–323.
- [45] Z. Wu, Y. Monno, and M. Okutomi, "Self-supervised spatially variant PSF estimation for aberration-aware depth-from-defocus," in *Proc. ICASSP*, 2024, pp. 2560–2564.
- [46] J. Chang and G. Wetzstein, "Deep optics for monocular depth estimation and 3D object detection," in *Proc. ICCV*, 2019, pp. 10 192–10 201.
- [47] S.-H. Baek *et al.*, "Single-shot hyperspectral-depth imaging with learned diffractive optics," in *Proc. ICCV*, 2021, pp. 2631–2640.
- [48] H. Ikoma, C. M. Nguyen, C. A. Metzler, Y. Peng, and G. Wetzstein, "Depth from defocus with learned optics for imaging and occlusion-aware depth estimation," in *Proc. ICCP*, 2021, pp. 1–12.
- [49] H. Wei *et al.*, "Learned off-aperture encoding for wide field-of-view RGBD imaging," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- [50] X. Qian *et al.*, "Towards single-lens controllable depth-of-field imaging via depth-aware point spread functions," *IEEE Transactions on Computational Imaging*, vol. 11, pp. 305–320, 2025.
- [51] J. Luo, Y. Nie, W. Ren, X. Cao, and M. Yang, "Correcting optical aberration via depth-aware point spread functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 46, no. 8, pp. 5541–5555, 2024.
- [52] B. Ghanekar, V. Saragadam, D. Mehra, A.-K. Gustavsson, A. C. Sankaranarayanan, and A. Veeraraghavan, "PS² F Polarized spiral point spread function for single-shot 3D sensing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 47, no. 8, pp. 6134–6145, 2025.
- [53] L. Li, J. Pan, W.-S. Lai, C. Gao, N. Sang, and M.-H. Yang, "Dynamic scene deblurring by depth guided model," *IEEE Transactions on Image Processing*, vol. 29, pp. 5273–5288, 2020.
- [54] C. Zhu *et al.*, "Deep recurrent neural network with multi-scale bi-directional propagation for video deblurring," in *Proc. AAAI*, vol. 36, no. 3, 2022, pp. 3598–3607.
- [55] G. F. Torres, J. Kalliolia, S. Tripathy, E. Acar, and J.-K. Kämäräinen, "DAVIDE: Depth-aware video deblurring," in *Proc. ECCVW*, vol. 15631, 2024, pp. 161–179.
- [56] J. Zhou, T. Yang, W. Ren, D. Zhang, and W. Zhang, "Underwater image restoration via depth map and illumination estimation based on a single image," *Optics Express*, vol. 29, no. 19, pp. 29 864–29 886, 2021.
- [57] P. Hambarde, S. Murala, and A. Dhall, "UW-GAN: Single-image depth estimation and image enhancement for underwater images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–12, 2021.
- [58] B. Xiao, X. Gao, and H. Huang, "Optimizing underwater image restoration and depth estimation with light field images," *Journal of Marine Science and Engineering*, vol. 12, no. 6, p. 935, 2024.
- [59] S. Gur and L. Wolf, "Single image depth estimation trained via depth from defocus cues," in *Proc. CVPR*, 2019, pp. 7683–7692.
- [60] S. Anwar, Z. Hayder, and F. Porikli, "Deblur and deep depth from single defocus image," *Machine Vision and Applications*, vol. 32, no. 1, p. 34, 2021.
- [61] S. Nazir, L. Vaquero, M. Mucientes, V. M. Brea, and D. Coltuc, "Depth estimation and image restoration by deep learning from defocused images," *IEEE Transactions on Computational Imaging*, vol. 9, pp. 607–619, 2023.
- [62] S. Hou, M. Fu, and W. Song, "Joint learning of image deblurring and depth estimation through adversarial multi-task network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7327–7341, 2023.
- [63] S.-J. Cho, S.-W. Ji, J.-P. Hong, S.-W. Jung, and S.-J. Ko, "Rethinking coarse-to-fine approach in single image deblurring," in *Proc. ICCV*, 2021, pp. 4621–4630.
- [64] Zemax OpticStudio® - Design and Analysis Software for Optical Systems, Zemax, Inc., Fremont, CA, USA, 2023.
- [65] K. Joaquina *et al.*, "Curved CMOS imaging sensor: development and reliability test results," in *Proc. ICSO*, vol. 12777, 2023, pp. 2863–2872.
- [66] C. s.a.s., "Premier capteur courbe commercial à application scientifique : une révolution dans le domaine de l'imagerie issue de la recherche en astronomie; first commercial curved and freeform sensor delivered to neuroscience: a revolution for imagery, fruit of research in astronomical instrumentation," CURVE s.a.s. (Spin-off du CNRS et Aix Marseille Université), Tech. Rep., 2020.
- [67] Y. Gao, Q. Jiang, S. Gao, L. Sun, K. Yang, and K. Wang, "Exploring quasi-global solutions to compound lens based computational imaging systems," *IEEE Transactions on Computational Imaging*, vol. 11, pp. 333–348, 2025.
- [68] S. Nah, T. Hyun Kim, and K. Mu Lee, "Deep multi-scale convolutional neural network for dynamic scene deblurring," in *Proc. CVPR*, 2017, pp. 257–265.
- [69] L. Jiang, B. Dai, W. Wu, and C. C. Loy, "Focal frequency loss for image reconstruction and synthesis," in *Proc. ICCV*, 2021, pp. 13 899–13 909.
- [70] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. ECCV*, vol. 7576, 2012, pp. 746–760.
- [71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. ICLR*, 2015.
- [72] S. F. Bhat, R. Birkel, D. Wofk, P. Wonka, and M. Müller, "ZoeDepth: Zero-shot transfer by combining relative and metric depth," *arXiv preprint arXiv:2302.12288*, 2023.
- [73] W. Zhao, Y. Rao, Z. Liu, B. Liu, J. Zhou, and J. Lu, "Unleashing text-to-image diffusion models for visual perception," in *Proc. ICCV*, 2023, pp. 5706–5716.
- [74] S. Patni, A. Agarwal, and C. Arora, "ECoDepth: Effective conditioning of diffusion models for monocular depth estimation," in *Proc. CVPR*, 2024, pp. 28 285–28 295.
- [75] Y. Peng, Q. Sun, X. Dun, G. Wetzstein, W. Heidrich, and F. Heide, "Learned large field-of-view imaging with thin-plate optics," *ACM Transactions on Graphics*, vol. 38, no. 6, pp. 1–14, 2019.
- [76] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.
- [77] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. CVPR*, 2018, pp. 586–595.
- [78] rnlee1998, "3D-reconstruction-for-monocular-depth-estimation," 2022, <https://github.com/rnlee1998/3D-Reconstruction-for-Monocular-Depth-Estimation> (Accessed: Sep. 1, 2025).
- [79] D. Seichter, M. Köhler, B. Lewandowski, T. Wengelfeld, and H.-M. Gross, "Efficient RGB-D semantic segmentation for indoor scene analysis," in *Proc. ICRA*, 2021, pp. 13 525–13 531.
- [80] W. Yin *et al.*, "Metric3D: Towards zero-shot metric 3D prediction from a single image," in *Proc. ICCV*, 2023, pp. 9009–9019.
- [81] L. Hong, X. Wang, G. Zhang, and M. Zhao, "USOD10K: A new benchmark dataset for underwater salient object detection," *IEEE Transactions on Image Processing*, vol. 34, pp. 1602–1615, 2025.

APPENDIX A LENS DATA

Our experiments include comparisons against several optical front-ends to rigorously evaluate the contribution of our novel bio-inspired design within the BMI framework. Alongside the proposed monocentric lens, we incorporate both a standard Double Gauss lens design (detailed in Table A.1), and a Fresnel lens design (detailed in Table A.2) as baseline references, representing conventional optical systems. The detailed parameters for the Fresnel lens used in this work are as follows: Surface 2 (Fresnel) has 4th, 6th, 8th, and 10th order terms of $1.492E-5$, $1.139E-7$, $-5.886E-10$, and 0, respectively. Surface 3 (Fresnel) has corresponding order terms of $1.696E-5$, $2.631E-8$, $1.240E-10$, and $-2.977E-13$.

By simulating the coded images generated by these conventional lenses using the same physically-based forward model (Sec. III-B) and processing them with the identical reconstruction network (Sec. III-C), we can effectively isolate the performance impact solely attributable to the front-end optical design. As demonstrated quantitatively in Table II, this controlled comparison unequivocally validates the unique advantages of our bio-inspired monocentric lens’s intrinsic depth-encoding capabilities over these established optical configurations for the task of joint image restoration and depth estimation.

TABLE A.1: Lens data for the Doubleguass used in this paper.

Surface	Radius(mm)	Thickness(mm)	Material	Semi-diameter(mm)
1 (Sphere)	15.977	6.293	LAF2	9.964
2 (Sphere)	Infinite	0.0039		9.964
3 (Sphere)	7.666	4.222	PSK3	6.612
4 (Sphere)	-73.042	1.134	SF1	6.612
5 (Sphere)	4.435	1.951		3.644
6 (Stop)	Infinite	1.249		3.599
7 (Sphere)	-7.388	1.215	SF1	3.644
8 (Sphere)	8.109	5.212	LAF2	4.611
9 (Sphere)	-10.497	2.012		4.611
10 (Sphere)	7.464	4.695	LAF2	4.183
11 (Sphere)	64.825	3.656		4.183
Sensor				0.801

TABLE A.2: Lens data for the Fresnel used in this paper.

Surface	Radius(mm)	Thickness(mm)	Material	Semi-diameter(mm)
1 (Stop)	infinity	13.248		4.500
2 (Fresnel)	226.656	10.000	PMMA	4.963
3 (Fresnel)	-23.164	41.874		5.095
Sensor				1.615

APPENDIX B OPTICAL SIMULATION COMPARISON

To quantitatively evaluate the simulation’s physical realism, we propose a global Artifact Score (AS). The rationale for this metric is rooted in human visual perception: simulation artifacts, such as ringing or blockiness from patch-wise processing, are most prominent and disruptive in smooth regions of an image (*e.g.*, walls), whereas they can be visually masked by the high-frequency content of natural object boundaries. Our score is therefore designed to specifically quantify these perceptually jarring imperfections. The metric isolates these

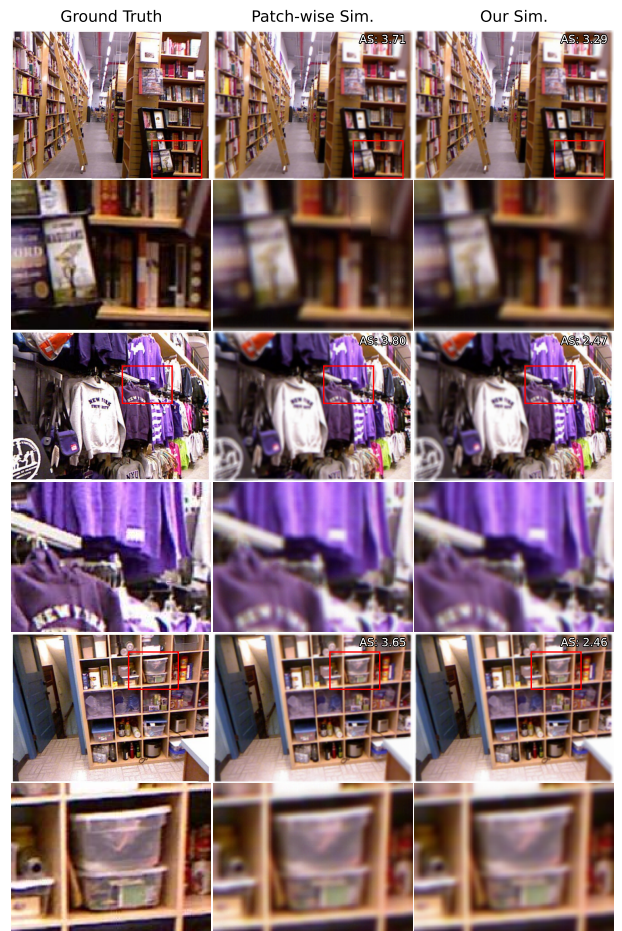


Fig. A.1: Comparison of simulation pipelines. The figure compares our occlusion-aware model against the conventional patch-wise method across three scenes. Magnified regions highlight performance at object boundaries, with the global AS for each simulation noted in the upper-right corner.

non-physical artifacts by calculating the average response of a Laplacian operator (L)—which is highly sensitive to high-frequency noise—within the smooth areas of the simulated image. We identify these areas by creating a binary mask (M_s) where smooth regions are marked as 1 and edge regions as 0. This mask is generated by applying a Canny edge detector to the ground-truth image, followed by a dilation operation to robustly exclude edge-adjacent areas. The score is formally defined as:

$$AS = \frac{\sum_{x,y} L(x,y) \cdot M_s(x,y)}{\sum_{x,y} M_s(x,y)}. \quad (\text{B.1})$$

A lower AS signifies a more physically plausible simulation with fewer visual artifacts. As shown in Figure A.1, our occlusion-aware simulation consistently yields a lower AS across various scenes compared to the patch-wise approach. This demonstrates that our method produces more accurate and physically realistic results, particularly at depth discontinuities, while maintaining computational efficiency.