Split-Aperture 2-in-1 Computational Cameras

ZHENG SHI* and ILYA CHUGUNOV*, Princeton University, USA
MARIO BIJELIC, GEOFFROI CÔTÉ, and JIWOON YEOM, Princeton University, USA
QIANG FU, HADI AMATA, and WOLFGANG HEIDRICH, KAUST, Saudi Arabia
FELIX HEIDE, Princeton University, USA

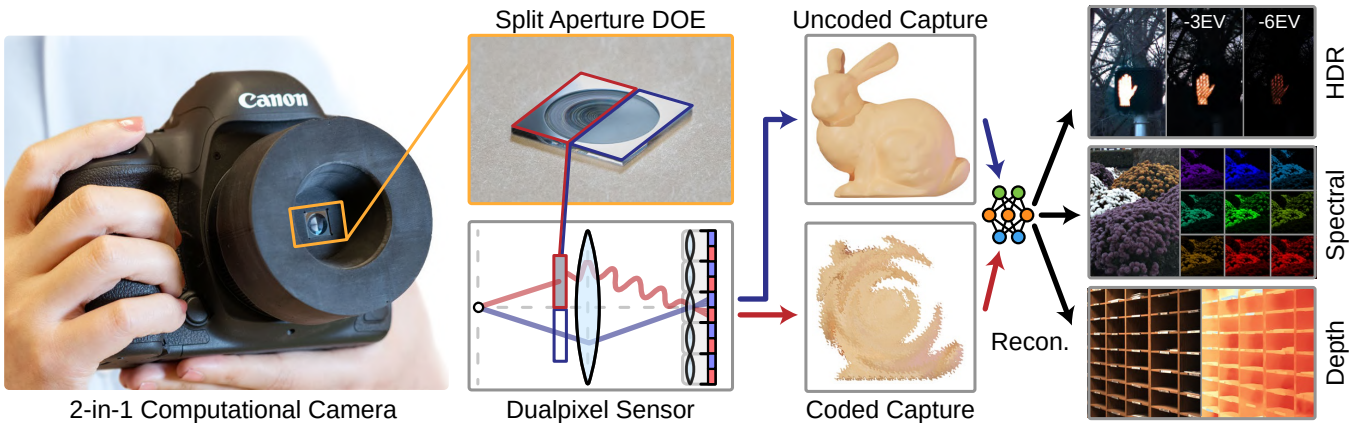

Fig. 1. Split-aperture 2-in-1 computational cameras enable simultaneous capture of both optically coded and conventional uncoded images without inverse image reconstruction or increasing the camera's physical footprint. To achieve this, we split the aperture into two halves: one half modulated by a diffractive optical element (DOE) to provide task-specific optical encoding, and the other remaining unmodulated. Then, with a commodity dual-pixel sensor found in many smartphone and DSLR cameras, we separate the two wavefronts into their coded and uncoded components. The uncoded image retains unperturbed high-frequency scene content, and enables conditional reconstruction of high-dynamic-range, hyperspectral, and depth measurements when paired with a task-specific coded capture; outperforming existing computational optics methods or RGB-to-X methods which operate with a single coded or uncoded image.

While conventional cameras offer versatility for applications ranging from amateur photography to autonomous driving, computational cameras allow for domain-specific adaption. Cameras with co-designed optics and image processing algorithms enable high-dynamic-range image recovery, depth estimation, and hyperspectral imaging through optically encoding scene information that is otherwise undetected by conventional cameras. However, this optical encoding creates a challenging inverse reconstruction problem for conventional image recovery, and often lowers the overall photographic quality. Thus computational cameras with domain-specific optics have only been adopted in a few specialized applications where the captured information cannot be acquired in other ways. In this work, we investigate a method that combines two optical systems into one to tackle this challenge. We split the aperture of a conventional camera into two halves: one which applies an application-specific modulation to the incident light via a diffractive optical element to produce a coded image capture, and one which applies no modulation to produce a conventional image capture. Co-designing the phase modulation of the split aperture with a dual-pixel sensor allows us to simultaneously capture these coded and uncoded images without increasing physical or computational footprint. With an uncoded conventional image alongside the optically coded image in hand, we investigate image reconstruction methods that are conditioned on the conventional image, making it possible to eliminate artifacts and compute costs that existing methods struggle with. We assess the proposed method with 2-in-1 cameras for optical high-dynamic-range reconstruction, monocular depth estimation, and hyperspectral imaging, comparing favorably to all tested methods in all applications.

CCS Concepts: • **Computing methodologies** → **Computational photography**.

Additional Key Words and Phrases: Computational Imaging, Co-Designed Optics, Dual-Pixel Sensor, HDR Imaging, Hyperspectral Imaging, Monocular Depth Estimation.

ACM Reference Format:
Zheng Shi, Ilya Chugunov, Mario Bijelic, Geoffroi Côté, Jiwoon Yeom, Qiang Fu, Hadi Amata, Wolfgang Heidrich, and Felix Heide. 2024. Split-Aperture 2-in-1 Computational Cameras . *ACM Trans. Graph.* 43, 4, Article 141 (July 2024), 19 pages. https://doi.org/10.1145/3658225

*Authors contributed equally to this work.

Authors' addresses: Zheng Shi, zhengshi@princeton.edu; Ilya Chugunov, chugunov@princeton.edu, Princeton University, USA; Mario Bijelic, mario.bijelic@princeton.edu; Geoffroi Côté, gcote@princeton.edu; Jiwoon Yeom, jy9976@princeton.edu, Princeton University, USA; Qiang Fu, qiang.fu@kaust.edu.sa; Hadi Amata, hadi.amata@kaust.edu.sa; Wolfgang Heidrich, wolfgang.heidrich@kaust.edu.sa, KAUST, Saudi Arabia; Felix Heide, Princeton University, USA, fheide@princeton.edu.

0730-0301/2024/7-ART141
https://doi.org/10.1145/3658225

1 INTRODUCTION

Digital cameras have enabled ubiquitous applications in personal photography, health, communication, remote sensing, and scientific imaging, as well as robotics and autonomous driving, where image data is consumed by downstream computer vision tasks. Despite the wide variety of applications, today's cameras are primarily designed to capture a perfect image with optics that still rely heavily on the linear model of optics derived by Gauss in the 19th century [Gauss 1843], independently of the domain—be that imaging for display or imaging for the input of a downstream object detector. Over the past twenty years, researchers have explored computational cameras [Nayar 2006] that integrate optics and computational models for specific tasks. These cameras, optimized for particular applications, could significantly enhance the functionality of traditional general-purpose camera systems. The co-design of optics and reconstruction algorithms has allowed for new capabilities such as single-shot high-dynamic-range [Metzler et al. 2020; Sun et al. 2020], large field-of-view [Peng et al. 2019], extended depth-of-field [Sitzmann et al. 2018], super-resolution [Sitzmann et al. 2018], and hyperspectral [Baek et al. 2021] imaging.

Despite their demonstrated potential, computational cameras remain limited to niche uses like microscopy [Pavani and Piestun 2009; Shechtman et al. 2014] or applications not involving co-designed optics and sensors, such as burst or multi-camera mobile phone imaging [Hasinoff et al. 2016]. A significant barrier is the ill-posed nature of the reconstruction problem, complicating signal extraction amidst sensor and photon noise and often leading to artifacts absent in traditional systems. Furthermore, the high computational demands of these methods, typically involving image-to-image neural networks, require substantial energy and bandwidth, a significant issue with the high resolutions of modern smartphones. Consequently, smartphones often employ conventional camera arrays, a costly yet more feasible alternative to computational optics.

In this work, we tackle these challenges by adding co-designed optics to a conventional optical system, simultaneously capturing an *optically coded and a conventional, uncoded image* without increasing the physical footprint of the camera. To merge two optical systems into one, we split both the aperture and sensor pixels of the proposed compound camera system. In the aperture plane, we modulate the phase of the captured light using a diffractive optical element (DOE). Specifically, we only modulate the wavefront in half of the aperture, leaving the other half unmodulated with no phase offset. To untangle the two wavefronts on the sensor, we make use of existing dual-pixel sensors, which collect separate images from each half of the camera aperture and have – primarily for autofocus – been broadly deployed in professional and smartphone cellphone cameras [Abuolaim et al. 2021; Garg et al. 2019]. The conventional capture serves as a condition for computational image reconstruction, as reconstructed measurements must be consistent with both the effects of the encoding and with the underlying scene content in the uncoded capture, reducing reconstruction artifacts otherwise caused by the loss of high-frequency details in the coded capture. Moreover, a conventional RGB image is immediately available for downstream applications – e.g., a viewfinder – without computationally expensive inverse image reconstruction.

We assess the proposed approach on diverse set of computational camera applications which rely on co-designed optics, both in simulation and with experimental prototypes. Specifically, we validate that the proposed method is capable of reconstructing high-dynamic-range measurements, monocular depth estimates, and hyperspectral images, outperforming existing methods in *all* cases. Specifically, we make the following contributions:

- We introduce a monocular single-shot imaging setup that captures an optically coded and a conventional uncoded image simultaneously.
- We propose a method for differentiable optics design that splits the aperture into two zones, each modulating half of the light entering the optical system. We disentangle the two optical paths by co-designing this split-aperture modulation with a dual-pixel sensor capture setup.
- We develop a conditional reconstruction framework designed to extract task-specific information from the encoded capture, conditioned on the uncoded conventional capture.
- We validate the proposed method in simulation and experimentally, confirming that the method is capable of combining two optical systems into one for optical high-dynamic-range reconstruction, monocular depth estimation, and hyperspectral imaging. We demonstrate significant improvements over tested RGB-to-X methods which rely on only an uncoded capture and computational optics approaches that only make use of a coded capture.

Optimized lens designs, network checkpoints, and code are available at: https://light.princeton.edu/2in1-camera.

Limitations. We rely on an academic nanofabrication facility to prototype the diffractive dual aperture device. As such, the quality of the fabricated phase plate is substantially lower than state-of-the-art nanofabrication processes. Although limiting the quality of the optical encoded aperture half, the proposed method allows for online compensation with the uncoded reference half, contrasting existing design methods. As a result of optical prototyping available to the authors, we design the proposed design as an add-on phase plate with a reduced aperture size. Future implementations may lift this prototype restriction by co-designing the compound optical system with the dual aperture encoding.

2 RELATED WORK

Differentiable Diffractive Optics. Conventional imaging systems employ compound refractive lens systems that are typically hand-engineered for image quality in isolation [Tseng et al. 2021], i.e., independently downstream camera tasks. Conventional refractive lens stacks are constrained by their smooth surface profile which can only provide smooth phase modulation, therefore limiting the design freedom to optically encode the desired task-specific scene information. To overcome these limitations, a large body of work in computational photography has explored the design of specialized lens system with diffractive optical elements (DOEs). With micron-scale surface profile, DOEs allow for fine-grained modulation of the phase of incident light via diffraction [Levin et al. 2007]. Researchers have shown that such optical systems can also be optimized via

back-propagation [Sitzmann et al. 2018; Wang et al. 2022], by modeling the imaging formation process with differentiable wave optics. Paired with a learnable reconstruction algorithm, such differentiable diffractive optics have not only allowed for high-quality color imaging [Peng et al. 2019], but have also enabled diverse applications in microscopy [Liu et al. 2022; Nehme et al. 2020], monocular depth imaging [Chang and Wetzstein 2019; Haim et al. 2018; Ikoma et al. 2021; Wu et al. 2019], high-dynamic range imaging [Metzler et al. 2020; Sun et al. 2020], hyperspectral imaging [Baek et al. 2021; Jeon et al. 2019; Li et al. 2022], and computer vision tasks [Shi et al. 2022; Tseng et al. 2021].

Snapshot Optics for Multimodal Acquisition. To enable additional imaging modalities beyond capturing RGB intensity, such as resolving the incident light in wavelength and time, existing capture systems often employ scanning-based or parallel acquisition. By design, scanning-based approaches [Brusco et al. 2006; Yoon et al. 2019] require a sequential acquisition that increases capture time, often prohibiting their use for dynamic scenes or real-time applications. Parallel imaging with multiple sensors does not increase the acquisition time but comes at the cost of a larger footprint [Gao and Wang 2016; Hagen et al. 2012] and a challenging alignment requirement when the same optical path is not shared. Another line of work [Baek et al. 2021; Jeon et al. 2019; Sun et al. 2020] explores snapshot optical systems that optically encode scene information with engineered PSFs that spatially multiplex over the sensor. While researchers have investigated recovering both RGB and additional multimodal image information from a single image, the image quality of the recovered images has been trailing that of conventional sensors due to the challenging reconstruction problem. In this work, we tackle this challenge and encode both a conventional capture and an encoded capture in the same optical path using a split aperture and a dual-pixel sensor.

Dual-pixel Sensors. Dual-pixel sensors rely on pixel technology where each pixel, without increasing pixel pitch, is split into two parts using two separate photodiode charge collection sites. As a result, for in-focus light, the µm-scale baseline between diodes in a dual-pixel sensor produces virtually zero “binocular-disparity” (pixel shift) between left and right views. Conversely, when the light is out-of-focus, it exhibits "defocus-disparity" – a depth-dependent change in the defocus blur between the left and right images. Early Canon sensors utilizes such defocus-disparity to perform phase difference autofocus (PDAF) [Kobayashi et al. 2016] – by comparing the signed average disparity value, autofocusing algorithms can determine the direction and extent of the defocus and move the lens accordingly to minimize the disparity. Today, dual-pixel sensors have become increasingly common among commercial cameras, including both DSLRs and smartphone imagers. While initially designed for autofocus, recent works have shown that the defocus-disparity from dual-pixel captures can be used for additional tasks such as depth estimation [Garg et al. 2019; Pan et al. 2021], reflection removal [Punnappurath and Brown 2019], and deblurring [Abuolaim and Brown 2020; Abuolaim et al. 2021]. In this work, in contrast to existing works that extract additional information from the natural defocus blur, we deliberately introduce additional phase modulation to half of the aperture using an optimized diffractive optical element. This allows us to extract task-specific information from the modulated side of the aperture while retrieving an undisturbed all-in-focus reference capture from the other side, without having to align the coded and uncoded captures.

Split/Multi Aperture Cameras. Researchers have long explored methods for capturing multiple images simultaneously using a single camera setup for diverse applications. In cinematography, split diopter or split-field diopter lenses are used to focus simultaneously at different depths, as noted by Malkiewicz [Malkiewicz and Mullen 2009]. These lenses cover only one half of a camera lens, rendering one half nearsighted and the other farsighted to create an illusion of deep focus. In the early 2000s, various split aperture camera designs were proposed for High Dynamic Range (HDR) imaging [Aggarwal and Ahuja 2004; Wang et al. 2005]. These designs typically employ mirrors and beamsplitters to direct incoming light to multiple CCD sensors, thus capturing the same scene at different brightness levels. Additionally, Green et al.[Green et al. 2007] proposed a design that splits the aperture into a central disc, enabling the capture of 2x2 images on the sensor with varying aperture sizes using relay optics and folding mirrors. Advancements in polarization-sensitive cameras have led to new methods, such as that by Ghanekar et al.[Ghanekar et al. 2022], which split the two lobes of the Double-Helix PSF and combines them with two polarizers in the pupil plane, capturing separate images across polarization channels to enhance depth reconstruction accuracy. Extending these advancements, our work integrates DOE with dual-pixel sensors, prevalent in commercial DSLR and smartphone cameras, to propose a versatile split-aperture 2-in-1 imaging system. This system is designed to produce both an application-specific modulated image and a conventional RGB capture without increasing the camera form factor, making it suitable for a wide range of coded imaging applications in consumer devices. Below, we review related work in the application domains explored with our proposed imaging method.

High-Dynamic-Range Imaging. While the ability to capture high-dynamic-range (HDR) scenes is crucial for many computer vision tasks such as nighttime self-driving, the dynamic range of a camera is fundamentally limited by the sensor well capacity, which results in standard commercial sensors having only a low dynamic range (LDR). As a result, standard sensors are not able to simultaneously capture a bright and a dark region outside of the sensor dynamic range in a single image, resulting in either saturated bright regions or low SNR in dark regions. Traditionally, multiple LDR images are captured with different exposure to be combined into an HDR image [Debevec and Malik 2008; Reinhard et al. 2010]. Recent burst techniques [Hasinoff et al. 2016] in smartphones are capable of acquiring HDR images for static scenes but fail for highly dynamic scenes, e.g., in outdoor automotive environments. Deep learning networks [Chen et al. 2023; Khan et al. 2019; Liu et al. 2020; Santos et al. 2020] have been used to generate plausible HDR content from a single LDR image based on imaging priors, but still fail to faithfully recover saturated details. To address this limitation, several approaches encode the HDR information into the captured LDR image using an additional DOE. Rouf et al. [2011] and Sun et al. [2020] spread the otherwise saturated information into unsaturated regions, aiming primarily to recover small saturated regions (3+ EV)

in night-time photography. While enabling high-fidelity reconstruction of the saturated regions, they leave noticeable artifacts in the unsaturated areas. The proposed 2-in-1 camera tackles this issue by acquiring a conventional LDR capture in addition to the optically encoded capture within a single shot.

Hyperspectral Imaging. Hyperspectral images, which capture much finer spectral band information than traditional 3-channel RGB, can facilitate applications in agricultural monitoring, material classification, and forensic science [Briottet et al. 2006; Näsi et al. 2015]. Scanning-based hyperspectral imaging methods [Brusco et al. 2006; Yoon et al. 2019] take multiple captures at each desired wavelength and isolate the spectral energy using additional bandpass filters. Existing snapshot hyperspectral imaging approaches [Baek et al. 2017] require multiple optical elements including dispersive optical elements (prisms), coded apertures and several lenses for relay and imaging purposes, making them bulky and impractical for most applications. To achieve a small-form-factor snapshot spectral imaging system, recent methods [Baek et al. 2021; Jeon et al. 2019; Li et al. 2022] rely on DOEs with spectrally varying point spread functions to encode the hyperspectral information. However, as a result of the ill-posed nature of the reconstruction problem, all of these snapshot imaging systems have in common that the estimated hyperspectral images suffer from artifacts and severe loss of high-frequency details. In this work, we design a 2-in-1 camera which leverages the unmodulated measurement as a guide image for single-shot hyperspectral reconstruction, and retains this high-frequency content.

Monocular Depth Estimation. Benefiting from large training sets, recent advances in deep neural networks have made it possible to accurately estimate the *relative* depth of objects based on monocular cues such as occlusions and relative object sizes [Bhat et al. 2023; Ranftl et al. 2021, 2022]. To estimate absolute depth from a single image, researchers have explored defocus blur as an additional monocular depth cue [Carvalho et al. 2018; Gur and Wolf 2019]. However, with a conventional camera, objects in front of the in-focus plane can provide the same defocus blur as objects behind the in-focus plane. To solve this ambiguity, a line of work [Chang and Wetzstein 2019; Haim et al. 2018; Ikoma et al. 2021; Wu et al. 2019] explores the design of depth-dependent PSFs with diffractive optical elements to unambiguously encode depth information. By design, these existing methods have in common that the spectral and spatial information of the scene are scrambled in the process, resulting in chromatic artifacts and blur in the recovered RGB-D images. Here, in addition to the depth-dependent encoded image, the proposed 2-in-1 camera also acquires an in-focus capture with the same optical path, which allows it to benefit from the monocular depth cues for depth estimation as well as recover an RGB image devoid of artifacts.

3 DUAL MODULATION IMAGE FORMATION

In this section, we describe the dual modulation image formation that enables the design of 2-in-1 computational cameras. We describe these proposed computational cameras in the following section.

We start by describing the operating principle of a dual-pixel sensor. Unlike traditional camera sensors, every single pixel on a dual-pixel sensor has two separate photodiodes located next to each other such that the light from two half-disks of the aperture is recorded independently, assuming a circular aperture, as shown in Fig. 2. Such an asymmetrical arrangement can be either a left-right or top-bottom pair; for simplicity, this paper considers a left-right arrangement. Where traditional dual-pixel cameras usually rely on a rotationally symmetric optical design to obtain two conventional captures, here we aim to generalize the design of dual-pixel cameras by introducing *split-aperture modulation*, where both captures are the result of independent light modulation that is asymmetric in the general case. Cameras designed under this principle incorporate two imaging modalities in the same optical light path, hence the concept of a *2-in-1* computational camera.

In the following, we derive the two PSFs of this image formation process via wave optics analysis. At any given point $(x, y, 0)$ within the aperture plane, a wavefront with wavelength λ originating from a singular on-axis scene point $(0, 0, -z)$—positioned at a distance z from the aperture plane—can be represented with a spherical phase profile

$$\phi_{\mathrm{s}}(x, y) = \frac{2\pi}{\lambda}\sqrt{x^2 + y^2 + z^2}. \tag{1}$$

The incident light passes through a DOE placed at the camera aperture plane, acquiring a phase modulation

$$\phi_{\mathrm{DOE}}(x, y) = \frac{2\pi(\mu(\lambda) - 1)}{\lambda} h(x, y). \tag{2}$$

Here, $h(x, y)$ is a 2D matrix representing the DOE height profile and $\mu(\lambda)$ is the wavelength-dependent refractive index of the DOE substrate. We allow both halves of the aperture height field h to take different forms

$$h(x, y) = \begin{cases} h_{\mathrm{L}}(x, y), & x < 0, \\ h_{\mathrm{R}}(x, y), & \text{otherwise.} \end{cases} \tag{3}$$

Immediately after the DOE, we assume a refractive lens focuses the incident light. Assuming both the DOE and refractive lens are co-located at the camera aperture plane, the lens can be modeled with a corresponding quadratic phase profile

$$\phi_{\mathrm{focus}}(x, y) = \frac{-2\pi}{\lambda} \frac{(x^2 + y^2)}{2f}, \tag{4}$$

where f is the focal length of the lens, representing the distance between the sensor and the camera aperture plane when the camera is focused at optical infinity.

After passing through the aperture plane, the incident light has accumulated phase

$$\phi_l = \phi_s + \phi_{\mathrm{DOE}} + \phi_{\mathrm{focus}}. \tag{5}$$

This modulated light then propagates to the sensor, where it is collected by separate photodiodes on each pixel. The PSF of this image formation process is

$$\begin{aligned} p &= |\mathcal{F}^{-1}\{\mathcal{F}\{u(x, y)\} \cdot \mathcal{H}\}|^2 \\ &= |\mathcal{F}^{-1}\{\mathcal{F}\{A(x, y)\exp(j\phi_l)\} \cdot \mathcal{H}\}|^2. \end{aligned} \tag{6}$$

Here, $u(x, y)$ is the complex-valued light field immediately after the refractive lens, $A(x, y)$ is the amplitude, and $\mathcal{H}$ is the light transport

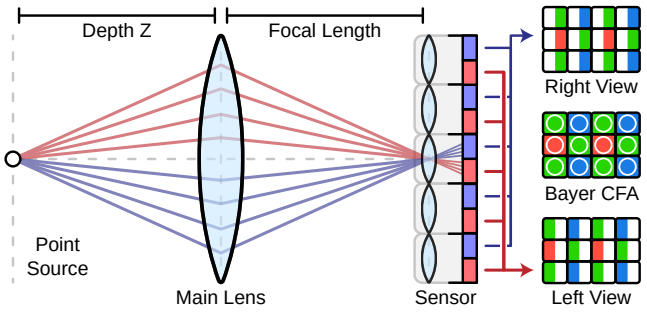

Fig. 2. **Dual-Pixel Sensing.** Increasingly prevalent in DSLRs and smartphones for enhanced autofocus, dual-pixel sensors feature two photodiodes per pixel, each capturing light from one half of the aperture independently. A microlens on each pixel ensures this division, creating a system akin to a miniaturized two-sample light field camera or a stereo system with an extremely small baseline. This dual-pixel split applies to every color channel, ensuring that both captures retain full-color information, effectively splitting the color filter array (CFA) into two separate CFAs.

term, which we model with the angular spectrum transfer function

$$\mathcal{H} = \begin{cases} \exp 2\pi j \frac{f}{\lambda}\sqrt{1-(\lambda f_x)^2-\lambda f_y)^2}, & \sqrt{f_x^2+f_y^2} < \frac{1}{\lambda} \\ 0, & \text{otherwise}, \end{cases} \tag{7}$$

where $k = \frac{2\pi}{\lambda}$ is the wave number and $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the Fourier transform and its inverse, respectively.

Due to the presence of the dual-pixel sensor combined with the micro-lens array, incoming light from the left and right sides of the aperture is collected separately by two interleaved pixel arrays. This is equivalent to $A(x, y)$ taking different and opposite forms for the left and right captures

$$A_{\mathrm{L}}(x,y) = \begin{cases} 1, & x < 0 \text{ and } x^2+y^2 < r_{\mathrm{a}}^2, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

$$A_{\mathrm{R}}(x,y) = \begin{cases} 1, & x > 0 \text{ and } x^2+y^2 < r_{\mathrm{a}}^2, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

where r_{a} is the aperture radius.

By successively substituting these expressions into Eq. (6), we derive the following PSF models

$$\begin{aligned} p_{\mathrm{L}} &= |\mathcal{F}^{-1}\{\mathcal{F}\{A_{\mathrm{R}}(x,y)\exp(j\phi_l)\}\cdot\mathcal{H}\}|^2, \\ p_{\mathrm{R}} &= |\mathcal{F}^{-1}\{\mathcal{F}\{A_{\mathrm{L}}(x,y)\exp(j\phi_l)\}\cdot\mathcal{H}\}|^2, \end{aligned} \tag{10}$$

where p_{L} describes the image formation process of the left capture while p_{R} is its counterpart for the right capture. Note, as illustrated in Fig. 2, light entering through one side of the aperture is focused onto the opposite side of the pixel.

Then, for a given scene I_{scene}, the dual-pixel sensor capture can be modeled as

$$\begin{aligned} I_{\mathrm{L}} &= \mathrm{clip}(p_{\mathrm{L}} * I_{\mathrm{scene}} + n_{\mathrm{L}}, 0, 1) \\ I_{\mathrm{R}} &= \mathrm{clip}(p_{\mathrm{R}} * I_{\mathrm{scene}} + n_{\mathrm{R}}, 0, 1), \end{aligned} \tag{11}$$

where n_{L} and n_{R} are the additional noise introduced by the sensor and the output intensity value is clipped to a low dynamic range of $[0, 1]$. We then simulate added read and signal noise with a Poissonian-Gaussian distribution.

4 2-IN-1 COMPUTATIONAL CAMERAS

In this section, we first introduce 2-in-1 computational cameras in general, irrespective of the task-specific optical encoding. We formalize the reconstruction problem of these cameras as a conditional inverse problem with access to both coded and uncoded image captures sharing the same optical path. Next, we propose several concrete 2-in-1 cameras that optically encode information that is challenging to measure directly, namely high-dynamic range, depth, and spectral information. An overview of the proposed 2-in-1 computational camera is illustrated in Fig. 3.

4.1 Task-Specific Split-Aperture Modulation

We first abstract F as the wave propagation forward model described in Sec. 3. Given a DOE height profile h in the same optical configuration as above, we express the corresponding PSFs (p_{L} and p_{R}) of a point light source of wavelength λ at depth z after passing through the left and right sides of the aperture as

$$p_{\mathrm{L}}, p_{\mathrm{R}} = F\left(h_{\mathrm{L}}, h_{\mathrm{R}}, z, \lambda\right), \tag{12}$$

where h_{L} and h_{R} represent the left and right halves of the DOE height profile. The forward modulation function F above is fully differentiable with respect to h, thereby allowing us to design the DOE via back-propagation. Specifically, for a given z and λ of interest, we pose the task-specific optical design problem as the following optimization problem

$$h^* = \arg\min_h \mathcal{L}_p\left(F\left(h_{\mathrm{L}}, h_{\mathrm{R}}, z, \lambda\right)\right), \tag{13}$$

where $\mathcal{L}_p$ is a penalty function that is defined using the PSFs (and potentially training data required for a reconstruction network).

In our approach, we opt not to introduce DOE phase modulation to one half of the aperture. Specifically, for the purposes of this paper, we keep the right half of the aperture unmodulated and have set $h_{\mathrm{R}} = 0$, and employ task-specific phase modulation to the left half of the aperture. As such, the left-side capture, I_{L}, functions like a conventional sensor capture. For clarity in subsequent discussions, we will refer to I_{L} as I_{uncoded} (uncoded capture). Conversely, the right-side capture, I_{R}, is designed to encode additional, task-specific scene information optically. In the later sections, this capture will be denoted as I_{coded} (coded capture).

$$\begin{aligned} p_{\mathrm{uncoded}}, p_{\mathrm{coded}} &= p_{\mathrm{L}}, p_{\mathrm{R}} = F\left(h_{\mathrm{L}}, 0, z, \lambda\right), \\ I_{\mathrm{uncoded}} &= I_{\mathrm{L}} = \mathrm{clip}(p_{\mathrm{uncoded}} * I_{\mathrm{scene}} + n_{\mathrm{L}}, 0, 1), \\ I_{\mathrm{coded}} &= I_{\mathrm{R}} = \mathrm{clip}(p_{\mathrm{coded}} * I_{\mathrm{scene}} + n_{\mathrm{R}}, 0, 1) \end{aligned} \tag{14}$$

This design choice offers two benefits. First, it allows for a readily available conventional capture without the need to solve a reconstruction problem and incur additional computational cost. Second, in contrast to computational cameras that only have access to the coded capture, the availability of an almost perfectly aligned conventional capture can be used to condition the reconstruction method and allow for test-time optimization. As such, the method can leverage additional monocular cues simultaneously captured in the conventional image, as we describe next.

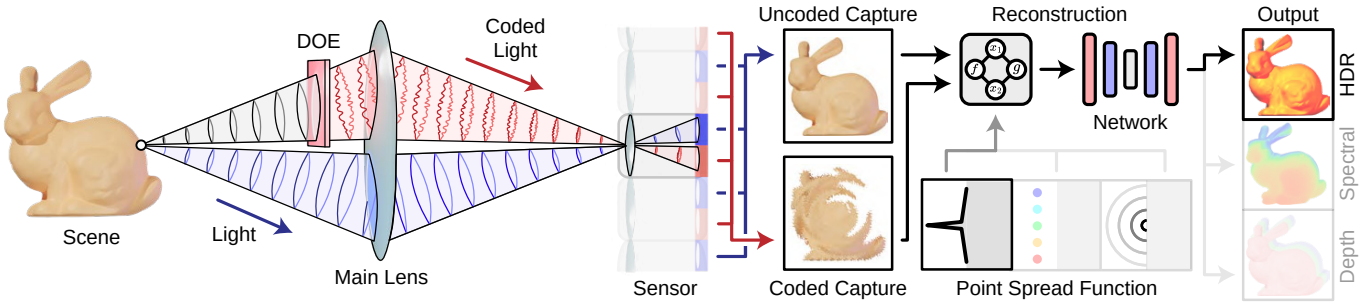

Fig. 3. **Split-Aperture 2-in-1 Computational Cameras.** To simultaneously capture an optically coded and a conventional uncoded image in a single shot, we propose task-specific phase modulation of one half of the aperture while keeping the other half unmodulated. The result is a hybrid imaging method: we acquire an uncoded image paired with an optically encoded capture tailored to specific scene information. The computational block of the system processes these dual images using a physics-based feature deconvolution block, relying on the task-specific optical point spread functions for a given application, and a conditional encoder-decoder backbone, enabling the extraction of task-specific image modalities supported and consistent with the uncoded capture. We demonstrate the effectiveness of this approach by demonstrating it for various computational photography applications, including high-dynamic-range, monocular depth, and multispectral imaging, each benefiting from customized optical encoding and tailored reconstruction networks.

4.2 Conditional Image Reconstruction

With the dual-pixel sensor captures described in Sec. 3 in hand, we design and optimize a task-specific reconstruction network G_{task} that has access to both the uncoded conventional capture I_{uncoded} and coded capture I_{coded}, as well as the task-specific PSF p_{coded}. Specifically,

$$\begin{aligned} I_{\text{recon}} &= G_{\text{task}}\left(I_{\text{coded}} | I_{\text{uncoded}}, p_{\text{coded}}\right) \\ W^{*}_{\text{task}} &= \arg\min_{W_{\text{task}}} \mathcal{L}_{\text{task}}\left(I_{\text{recon}}\right), \end{aligned} \tag{15}$$

where $\mathcal{L}_{\text{task}}$ is a task-specific reconstruction loss and W_{task} represents all the parameters of the reconstruction network G_{task}. In the following, we use this general design principle to devise several application-specific 2-in-1 computational cameras.

4.3 Coded Optics for High Dynamic Range Imaging

Existing work on snapshot HDR imaging has demonstrated that it is possible to optically encode localized high-dynamic-range information into nearby pixels using streak-like PSFs which optically spread out the saturated image content that is no longer saturated but superposed with the nearby image regions. Building on existing snapshot HDR imaging approaches [Rouf et al. 2011; Sun et al. 2020], which have successfully coded high-dynamic-range information into adjacent pixels using streak-like PSFs, we proceed to develop a 2-in-1 camera in line with this concept. We first optimize the free design phase half to align with the streak-like target PSF

$$\hat{p}'_{\text{HDR}}[x,y] = \begin{cases} 1, & x = \frac{R}{2} \\ 1, & y = \frac{R}{2} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

$$p'_{\text{HDR}}[x,y] = \frac{0.5}{\sum_{x,y} \hat{p}'_{\text{HDR}}[x,y]}, \tag{17}$$

where R denotes the DOE's radius in pixels, by minimizing the following penalty term

$$\begin{aligned} \mathcal{L}_{p,\text{HDR}} &= \mathcal{L}_1\left(p_{\text{coded}}, p'_{\text{HDR}}\right) \\ &= \mathcal{L}_1\left(F\left(h_{\text{L}}, 0, \infty_{\text{O}}, \lambda_{\text{RGB}}\right), p'_{\text{HDR}}\right), \end{aligned} \tag{18}$$

that is, the $\mathcal{L}_1$ distance between the simulated right-side PSF p_{coded} and the target streak-pattern PSF p'_{HDR} when depth is at optical infinity ∞_{O} for discrete sampled RGB wavelengths λ_{RGB}. Note that the PSFs are normalized to have a total energy of 0.5 since it only collects the light passing through half of the aperture.

Without loss of generality, we train a physics-based conditional reconstruction network G_{HDR} to reconstruct the original HDR scene. Since the unmodulated capture I_{uncoded} provides information in the unsaturated regions of the scene, G_{HDR} focuses on recovering the saturated highlight regions from the streak pattern that are produced by convolving the highlights with our designed PSF p_{coded}. Therefore, reconstructing the highlights becomes a deconvolution problem, and is tackled using a network with architecture similiar to Deep Wiener Deconvolution Network (DWDN) [Dong et al. 2020], which initially conducts feature-based inverse filtering on the coded capture, followed by an encoder-decoder network for image reconstruction. Specifically, the process is formalized as follows

$$\begin{aligned} I_{\text{recon}} &= G_{\text{HDR}}\left(I_{\text{coded}} | I_{\text{uncoded}}, p_{\text{coded}}\right) \\ &= \text{DWDN}\left(I_{\text{uncoded}}, D\left(I_{\text{coded}}, p_{\text{coded}}\right)\right), \end{aligned} \tag{19}$$

where D is a differentiable implementation of the Wiener filter. See Supplemental Material for additional details. We train this network joinly with the pre-trained DOE profile with the following loss

$$\begin{aligned} \mathcal{L}_{\text{HDR}} &= \mathcal{L}_1\left(I_{\text{recon}}, I_{\text{HDR}}\right) \\ &\quad + 0.1\mathcal{L}_1\left(I_{\text{recon}} M_{\text{highlight}}, I_{\text{HDR}} M_{\text{highlight}}\right), \end{aligned} \tag{20}$$

where I_{HDR} is the ground truth HDR scene, $\mathcal{L}_1$ is a per-pixel $\mathcal{L}_1$ loss and $M_{\text{highlight}}$ is a binary mask marking the highlight locations.

We train our model using a blend of outdoor night scenes and indoor scenes sourced from HDRi Haven. These scenes are scaled to have 1% to 5% of pixels saturated (values in the range $[1, 2^8]$).

4.4 Optically Coded Hyperspectral Imaging

Inspired by recent designs of small-form-factor snapshot spectral imaging systems [Baek et al. 2021; Jeon et al. 2019; Li et al. 2022],

we design a dual aperture camera that encodes 31-channel hyperspectral information into a 3-channel RGB capture by intentionally introducing chromatic aberration. We employ a grating-like phase profile for the coded aperture half to modulate the incident light, inducing a wavelength-dependent lateral shift in the focus point. This allows us to reconstruct the hyperspectral information based on the shift magnitude relative to the aberration-free uncoded capture. More specifically, our design process begins with a grating profile featuring slits that are 16 DOE pixels wide, as its first positive order diffraction shifts approximately 1 pixel to the left for every 10 nm change in wavelength. We then refine this profile to increase the total intensity in the first positive order while minimizing energy distribution in other areas, such as the zeroth or higher-order diffraction positions. To this end, we set the target PSF to

$$\begin{aligned} p'_{\mathrm{HS}}(\lambda) &= \Delta(\mathrm{GE}(\lambda)), \\ &= \Delta(\tan(\sinh(\frac{\lambda}{16\delta_{\mathrm{DOE}}}))\frac{f}{\delta_{\mathrm{camera}}}), \end{aligned} \tag{21}$$

where $\Delta(x, y)$ represents a Dirac delta at pixel (x, y), δ_{DOE} and δ_{camera} represent the pixel pitch size and $\mathrm{GE}(\lambda)$ computes the first positive order position for a grating profile with a 16 DOE pixel slit width, as per the grating equation.

To refine the DOE profile, we set

$$\begin{aligned} \mathcal{L}_{p,\mathrm{HS}} &= \mathcal{L}_1\left(p_{\mathrm{coded}}, p'_{\mathrm{HS}}\right) \\ &= \mathcal{L}_1\left(F\left(h_{\mathrm{L}}, 0, \infty_{\mathrm{O}}, \lambda_{\mathrm{HS}}\right), p'_{\mathrm{HS}}\right), \end{aligned} \tag{22}$$

as the $\mathcal{L}_1$ distance between the simulated right-side PSF p_{coded} and the target PSF when depth is at optical infinity ∞_{O} for discrete wavelength samples $\lambda_{\mathrm{HS}} \in [400, 700]$ nm.

Similar to G_{HDR} used for HDR scene reconstruction, we perform reconstruction conditioned on the uncoded capture I_{uncoded} with optically-coded capture I_{coded}, as well as the PSF of each sampled hyperspectral wavelength λ_{HS} to provide physics-based cues for the learned reconstruction backbone. We simulate the RGB captures and recover a latent image as

$$\begin{aligned} I_{\mathrm{uncoded}} &= \mathrm{clip}((p_{\mathrm{uncoded}} * I_{\mathrm{HS}})T_s + n_{\mathrm{L}}, 0, 1), \\ I_{\mathrm{coded}} &= \mathrm{clip}((p_{\mathrm{coded}} * I_{\mathrm{HS}})T_s + n_{\mathrm{R}}, 0, 1), \\ I_{\mathrm{recon}} &= G_{\mathrm{HS}}\left(I_{\mathrm{coded}} | I_{\mathrm{uncoded}}, p_{\mathrm{coded}}\right) \\ &= \mathrm{DWDN}\left(I_{\mathrm{uncoded}}, \{D\left(I_{\mathrm{coded}}, p_{\mathrm{coded}}(\lambda)\right)\}_{\lambda \in \lambda_{\mathrm{HS}}}\right), \end{aligned} \tag{23}$$

where T_s represents the RGB sensor response curve. The reconstruction network is supervised to minimize both the perceptual loss (LPIPS) in RGB space and Spectral Angle Mapper (SAM) [Kuching 2007] distance in hyperspectral space between the reconstructed scene and ground truth scene, which can be expressed as

$$\mathcal{L}_{\mathrm{HS}} = \mathcal{L}_{\mathrm{perc}}\left(I_{\mathrm{recon}}T_s, I_{\mathrm{HS}}T_s\right) + \mathrm{SAM}\left(I_{\mathrm{recon}}, I_{\mathrm{HS}}\right), \tag{24}$$

To improve the generalization of our computational camera, we train on images from two datasets: CZ_HSDB [Chakrabarti and Zickler 2011] and ICVL [Arad and Ben-Shahar 2016], and test on a separate dataset, CAVE [Yasuma et al. 2008]. Each image in these datasets is comprised of 31 spectral channels ranging from 400nm to 700nm at 10nm intervals.

4.5 Monocular Depth from Coded Defocus

While typical monocular depth estimation methods rely on learned image-space depth cues such as relative object sizes to estimate the *relative* position of objects, a line of work [Chang and Wetzstein 2019; Haim et al. 2018; Ikoma et al. 2021; Wu et al. 2019] depth-from-defocus approaches with engineered PSFs that optically encode *absolute* depth using a DOE. We investigate a 2-in-1 computational camera that relies on depth-dependent concentric rings as a target PSF, following the design from Haim et al. [2018] but is conditioned on a simultaneously captured monocular RGB image. Specifically, for different depths z_k within 1–5 m, we set our target PSF as a semicircle with radius r growing 1 sensor pixel per 20 cm,

$$p'_{\mathrm{depth}}(z) = \mathrm{HCir}\left(-5z + 25\right), \tag{25}$$

where $\mathrm{HCir}(r)$ represents a semicircle with radius r pixels and $\mathrm{HCir}(0)$ is a Dirac peak. We propose to optimize the DOE phase pattern (h) to focus light effectively with a target half-ring pattern as

$$\mathcal{L}_{p,\mathrm{depth}} = \sum_k \sum_{x,y} \left(F\left(h_{\mathrm{L}}, 0, z_k, \lambda_{\mathrm{RGB}}\right) M\left(p'_{\mathrm{depth}}(z_k), \delta\right)\right)[x, y], \tag{26}$$

where $M\left(p'_{\mathrm{depth}}(z), \delta\right)$ is a mask for pixels located more than δ (set to 1 pixel) away from the target split ring $p'_{\mathrm{depth}}(z)$. To simulate captures, we partition the scene I_{Depth} into multiple depth planes from 1 m to 5 m in 0.25 m intervals, and convolve each depth layer with the corresponding PSF as

$$\begin{aligned} I_{\mathrm{uncoded}} &= \mathrm{clip}(\sum_{z \in \{1, 1.25, \ldots, 5\}} (p_{\mathrm{uncoded}}(z) * (I_{\mathrm{Depth}} M(z)) + n_{\mathrm{L}}, 0, 1), \\ I_{\mathrm{coded}} &= \mathrm{clip}(\sum_{z \in \{1, 1.25, \ldots, 5\}} (p_{\mathrm{coded}}(z) * (I_{\mathrm{Depth}} M(z)) + n_{\mathrm{R}}, 0, 1), \end{aligned} \tag{27}$$

where $M(z)$ is a mask of pixels with depth with in $z \pm 0.125$ m.

For conditional reconstruction, we employ ResNet18 [He et al. 2016] to extract features from both captures separately, before feeding them into a shared decoder. The network is optimized using the following loss

$$\mathcal{L}_{\mathrm{depth}} = \mathcal{L}_1(Z_{\mathrm{recon}}, Z_{\mathrm{GT}}) + \mathcal{L}_{\mathrm{grad}}(Z_{\mathrm{recon}}, Z_{\mathrm{GT}}), \tag{28}$$

with $\mathcal{L}1$ measuring per-pixel $\mathcal{L}1$ distance between the reconstructed depth (Zrecon) and ground truth (ZGT), and $\mathcal{L}_{\mathrm{grad}}$ comparing the gradient of reconstructed depth to the ground truth. To train and validate our approach, we use the FlyingThings3D dataset [Mayer et al. 2016], which contains pairs of all-in-focus RGB images and corresponding disparity maps.

5 ASSESSMENT

In this section, we evaluate the proposed method both in simulation and experimentally. We focus on three computational optics applications described in the previous section: recovery of high dynamic range (HDR) images from optically-encoded streak images, monocular depth imaging using coded depth-from-defocus, and hyperspectral image reconstruction via chromatic aberrations.

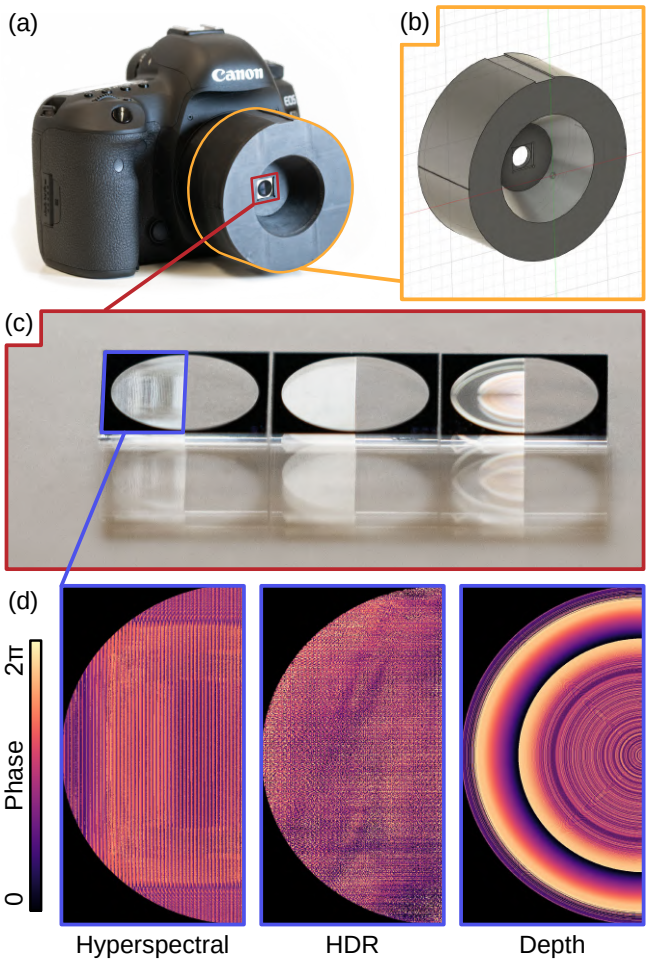

Fig. 4. **Experimental Prototype.** We built an experimental prototype, shown in (a), to evaluate the proposed method. The fabricated DOEs, shown in (c), are manufactured based on the phase profiles designed for each application, as shown in (d), utilizing a 16-level photolithography process. While this DOE is designed to be in the aperture plane of the target camera configuration, we opt to design and 3D print a DOE holder (b) to position the DOE adjacent to the lens cover glass, circumventing the need to dissect a commercial multi-element compound lens.

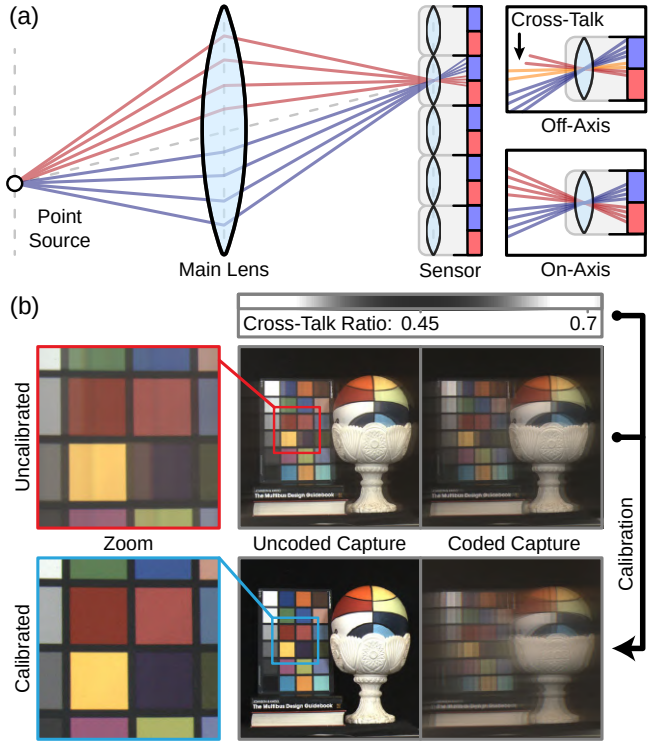

Fig. 5. **Cross-talk Dual-Pixel Calibration.** In theory, each photodiode in a dual-pixel sensor independently records light from only one half of the aperture. In practice, off-axis light causes crosstalk, with light intended for one photodiode being captured by the other. Here, (a) illustrates the angle-dependent crosstalk. We calibrate a fixed crosstalk ratio for the central 3072×3072 pixels, as indicated by the colorbar at the top of (b). This enables accurate retrieval of clean coded and uncoded captures across various applications, see (b).

5.1 Experimental Prototype

To assess the method experimentally, we fabricate a DOE tailored to each application and build a prototype 2-in-1 camera system, as illustrated in Fig. 4. The DOE fabrication involves a 16-level photolithography process on a fused silica wafer, see Supplemental Document for details. The DOE has a diameter of 8.64 mm and employs a chrome layer as an optical baffle. Our camera setup consists of a Canon EOS 5D Mark IV dual-pixel camera paired with a Canon EF 50mm f/1.8 STM lens. Although the DOE design is ideally suited for placement in the aperture plane of the target camera, we opt to design and 3D print a custom add-on DOE holder to avoid the complexities of modifying a commercial multi-element compound lens or constructing a 4F relay system. The holder positions the DOE directly next to the lens cover glass, minimizing the propagation distance between the DOE and aperture plane. Markers on the holder line up with the camera body to ensure that the DOE is center-aligned with the lens.

We proceed to calibrate the crosstalk between the left and right views in our camera system. Ideally, each pixel's two photodiodes should independently record light from the two half-disks of the aperture. However, in practice, this separation is imperfect, particularly for off-axis light. Consequently, some light intended for one side of the photodiodes is mistakenly recorded on the other, leading to crosstalk between the two captures, as illustrated in Fig. 5. This cross-talk can be formalized as

$$\begin{aligned} I_{\mathrm{L}}^{*} &= (1-\alpha_{l\to r})I_{\mathrm{L}} + \alpha_{r\to l}I_{\mathrm{R}} \\ I_{\mathrm{R}}^{*} &= (1-\alpha_{r\to l})I_{\mathrm{R}} + \alpha_{l\to r}I_{\mathrm{L}}. \end{aligned} \tag{29}$$

Here, I_{L} and I_{R} denote the theoretical left and right captures, while I_{L}^{*} and I_{R}^{*} represent the actual recorded views. The terms $\alpha_{l\to r}$ and $\alpha_{r\to l}$ represent the per-pixel cross-talk ratios from the left view to the right view and vice versa, respectively. Notably, this observed

Table 1. **Quantitative Evaluation of HDR Reconstruction Quality.** We measure the reconstruction quality in the overall image and the highlight regions, using the RMSE and PSNR, where PSNR is calculated with a maximum value of 2^8. We compare the proposed method against learned LDR-to-HDR methods, represented by DeepHDR [2020] and CEVR [2023], and recent DOE-based snapshot HDR imaging methods, represented by Rank-1 Optics [2020]. Additionally, we include a comparison where only the coded capture is used as the only input to the reconstruction network.

	↓RMSE	↑PSNR	↓RMSE^H	↑PSNR^H
DeepHDR [2020]	4.01	41.58	31.56	24.12
CEVR [2023]	3.17	42.55	25.32	25.29
Rank-1 Optics [2020]	2.96	43.64	20.93	26.81
Coded Capture Only	1.39	48.18	10.82	31.33
Proposed	**0.88**	**54.87**	**7.14**	**37.68**

cross-talk aligns with phenomena previously observed but we find it not being analyzed in dual-pixel sensor studies [Xin et al. 2021].

Given that the cross-talk ratio depends on the view angle rather than the scene itself, we look to calibrate the values of $\alpha_{l\to r}$ and $\alpha_{r\to l}$ to remove it. We begin by blocking the left half of the aperture (setting $I_{\mathrm{R}} = 0$) and capturing a white wall. Under these conditions, the captures can be simplified as

$$\begin{aligned} I_{\mathrm{L}}^* &= (1-\alpha_{l\to r})I_{\mathrm{L}} \\ I_{\mathrm{R}}^* &= \alpha_{l\to r} I_{\mathrm{L}}. \end{aligned} \tag{30}$$

From this, we calculate the cross-talk ratio from left to right view, that is

$$\alpha_{l\to r} = \frac{I_{\mathrm{R}}^*}{I_{\mathrm{L}}^* + I_{\mathrm{R}}^*}. \tag{31}$$

We repeat this process to determine the value of $\alpha_{r\to l}$. With these ratios, we accurately recover the clean encoded and uncoded captures

$$\begin{aligned} I_{\text{coded}} = I_{\mathrm{R}} &= \frac{I_{\mathrm{R}}^* - \alpha_{l\to r} I_{\mathrm{L}}^*}{1-\alpha_{r\to l}^2 - \alpha_{l\to r}\alpha_{r\to l} + \alpha_{l\to r}^2}, \\ I_{\text{uncoded}} = I_{\mathrm{L}} &= I_{\mathrm{L}}^* + I_{\mathrm{R}}^* - I_{\mathrm{R}}. \end{aligned} \tag{32}$$

As the crosstalk ratios are constant for our setup, this calibration is a one-time step, and we conduct this calibration process on the central 3072×3072 pixels of the sensor, using the same crosstalk ratios for every capture, whether it is for HDR imaging, depth estimation, or hyperspectral imaging.

5.2 Snapshot High Dynamic Range Imaging

We first assess the proposed method for coded high-dynamic range imaging introduced in Sec. 4.3 using both simulated captures and real-world captures obtained by our experimental prototype. Qualitative and quantitative comparisons using synthetic datasets are reported in Fig 6 and Tab 1, respectively, and experimental assessments are presented in Fig 7. Additional qualitative comparisons are available in the Supplemental Document.

5.2.1 Synthetic Assessment. We compare the proposed method to existing snapshot HDR imaging methods, and consider two types of baselines: (i) Learned LDR to HDR methods that synthesize content for the saturated regions, represented by the recent DeepHDR [Santos et al. 2020] and CEVR [Chen et al. 2023]; (ii) HDR imaging which optically encodes the saturated regions into the unsaturated regions to recover HDR, represented by Rank1 Optics [Sun et al. 2020]. Additionally, we include a comparison where only the coded capture is used as input to the reconstruction method. For these experiments, we generate the LDR input by either clamping the HDR ground truth with simulated noise to $[0, 1]$, or we simulate sensor measurements for the optical design of the respective baseline method and use the pretrained network weights provided by the authors. We assess reconstruction quality with RMSE and PSNR metrics, measuring both the overall image quality and specifically in the highlight regions. Qualitative and quantitative comparisons are reported in Fig 6 and Tab 1, respectively, with additional comparisons presented in the Supplemental Document.

DeepHDR [Santos et al. 2020] is a learning-based method for recovering overexposed pixels in a low dynamic range (LDR) image; pre-trained on a dataset of 2.5 million images for the task of inpainting, and fine-tuned on a set of 2,000 images for high dynamic range (HDR) generation. CEVR [Chen et al. 2023] extends the bit depth of LDR inputs and enables the generation of images across arbitrary, continuous exposure values (EVs) through a continuous exposure value representation. The method uses a cycle training strategy to supervise the model, allowing it to produce continuous EV LDR images without corresponding ground truths. However, due to the inherent lack of information, both DeepHDR and CEVR primarily synthesizes plausible image texture in overexposed regions that may not accurately represent the original scene.

Aiming to recover the saturated regions, Rank-1 Optics [Sun et al. 2020] proposes an optical system that employs a DOE with a rank-1 streak-like phase pattern. A co-optimized reconstruction network isolates the unsaturated regions from the encoded data and then restores the overexposed highlights. Unfortunately, this approach blends LDR image content with these streak PSFs, resulting in residual artifacts in the recovered HDR images.

The proposed 2-in-1 dual aperture camera is capable of acquiring coded HDR information and uncoded LDR information simultaneously. The uncoded capture provides a conventional image with unsaturated regions unaffected by an optical encoding, while the coded capture specializes in mapping HDR signals onto an LDR sensor. This encoding process, not required to preserve LDR signals, learns to create multiple scaled duplicates of overexposed regions across a broader exposure spectrum, effectively combining multiple exposures at several locations in a streak pattern. Unsurprisingly, the encoding is designed to work with an uncoded capture that contains the LDR information; without the uncoded capture the approach fails to reconstruct these regions of LDR content, resulting in low image quality.

5.2.2 Experimental Assessment. Next, we experimentally verify the proposed method for optically coded HDR. To this end, we first measure the PSF of the prototype system with the split aperture DOE installed. We employ exposure bracketing to measure the HDR PSFs, and the merged PSF measurements are presented at the top of Fig. 7. These measurements confirm that our fabricated DOE exhibits a PSF with the simulated streak shape. However, they also

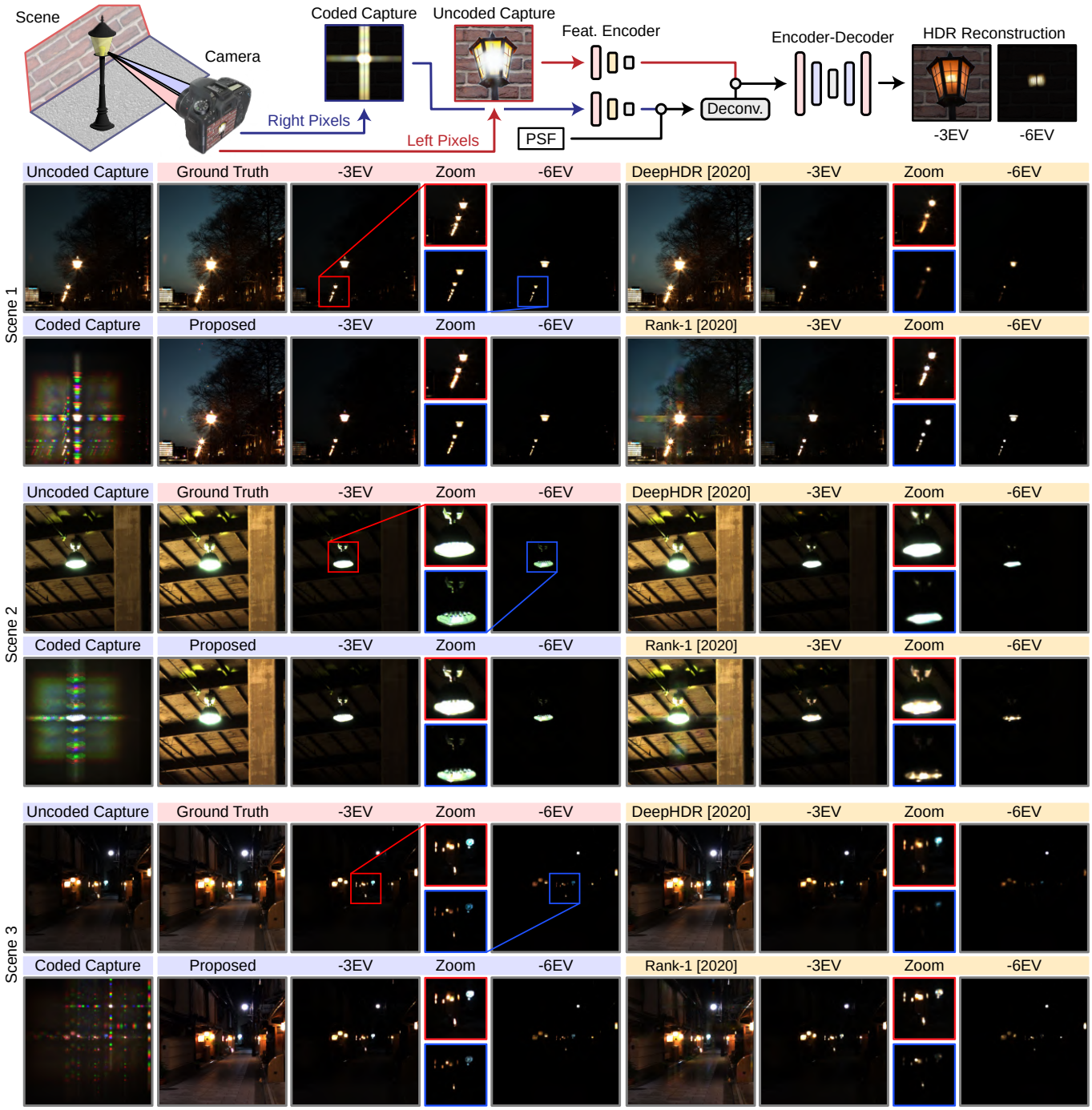

Fig. 6. **Snapshot HDR Methods in Simulation.** We assess the proposed method for snapshot HDR imaging in simulation by comparing the proposed method to the LDR-to-HDR method DeepHDR [Santos et al. 2020], and the DOE-based Rank-1 Optics approach [Sun et al. 2020]. For each scene, the leftmost column shows our method, followed by the reconstructed scenes at 0EV, -3EV, and -6EV. We also provide close-ups of the saturated regions to show the resolution of fine structures. DeepHDR, constrained by its LDR input, produces plausible HDR imagery but falls short in detailed recovery. Conversely, Rank-1 Optics occasionally struggles to differentiate HDR encoding from LDR content, resulting in visible streak artifacts. By simultaneously obtaining both LDR uncoded capture and coded capture, the proposed method is able to reconstruct highlight details without affecting the imaging quality of the LDR content.

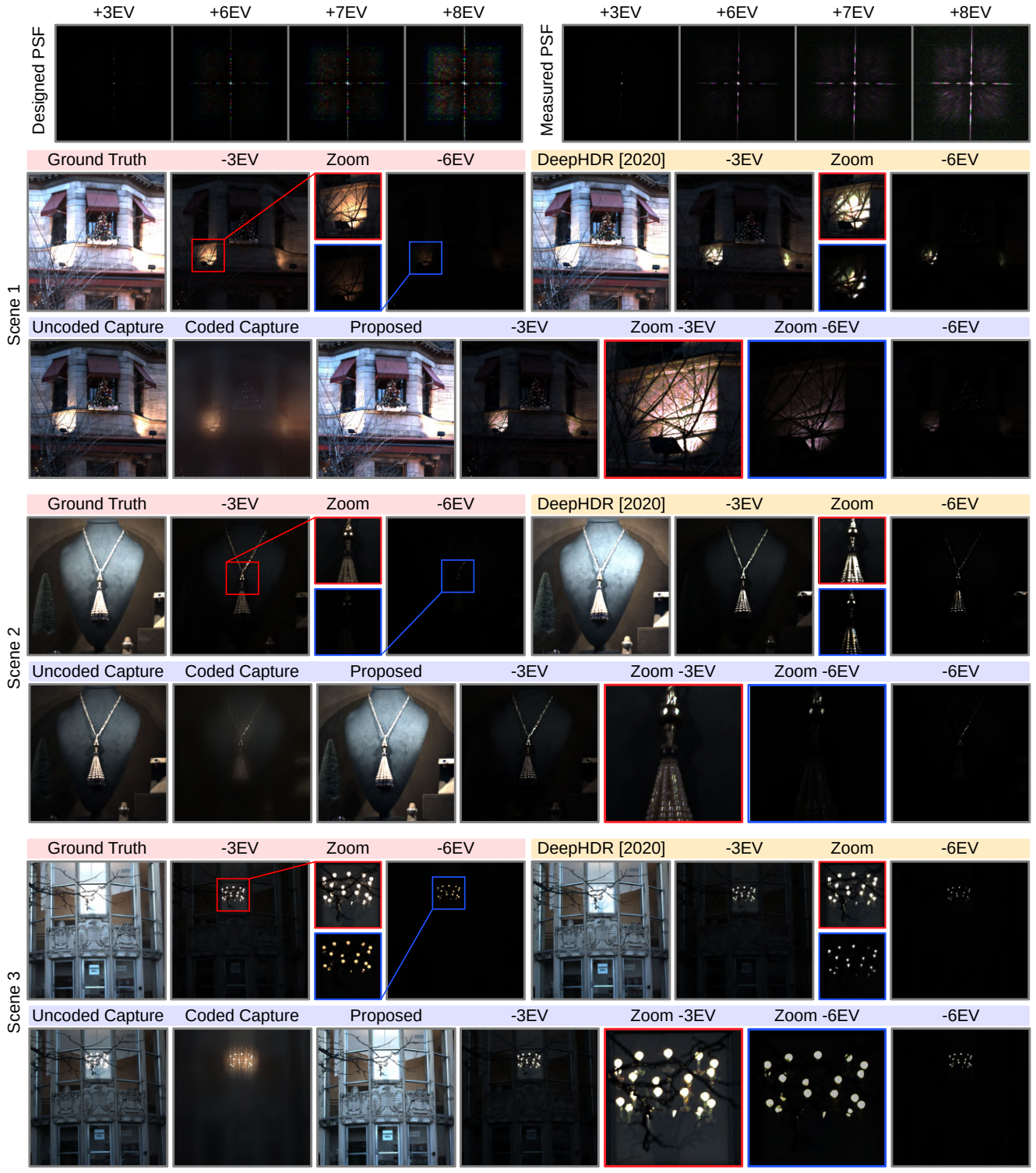

Fig. 7. **Experimental Assessment of Snapshot HDR Imaging.** We assess the proposed method experimentally for snapshot HDR imaging. The top row reports measured PSFs of the prototype with proposed split-aperture HDR DOE, which confirm that the DOE produces the intended streak PSFs. The next rows evaluate the method for outdoor scenes, comparing our results with Ground Truth data obtained through bracketed exposures. The proposed method is able to recover fine detail of the highlights, while the learned LDR-to-HDR method, DeepHDR, produces incorrect HDR estimates with image structure and intensity levels that significantly deviate from those in the ground truth captures. Please zoom into the electronic version of this document for details.

reveal pronounced chromatic aberrations in the PSF. We employ a two-step fine-tuning process for our reconstruction network to mitigate fabrication discrepancies in the learned DOE and distribution shifts between synthetic and real-world captures. First, we finetune the network using the PSF measured post-fabrication on synthetic data. This step provides the network with information on the characteristics of the fabricated DOE. Subsequently, we refine the output to ensure cross-modal consistency. Specifically, we employ a loss that ensures reintroducing the reconstructed highlights into the image forward model produces a simulated coded capture that closely matches the captured ones in intensity. This approach effectively aligns the reconstruction with both the measured DOE behavior and the real-world captures. Further details are provided in the Supplemental Document.

We proceed to test the proposed 2-in-1 computational camera in outdoor environments. For each scene, we first capture a reference image using the same camera without the DOE across varying exposures, to serve as reference for qualitative evaluation. Then, we attach the DOE to the lens and capture the scene again. The captured image is split into the uncoded and coded captures, which are then input into the reconstruction network to generate the reconstructed high dynamic range scene. Experimental results are reported in Fig.7 and the Supplemental Document. For every scene, we show the uncoded and coded captures from our prototype, the reconstructions from our method, and the reference captures obtained via exposure bracketing. Additionally, we compare these results with the learned LDR-to-HDR method DeepHDR [Santos et al. 2020], applied to the uncoded capture. We find that the proposed single-shot method is capable to recover fine details in saturated regions, such as tree branches, in contrast to, which DeepHDR often inaccurately hallucinates content, occasionally producing results that differ by orders of magnitude in intensity from the ground truth captures.

5.3 Snapshot Hyperspectral Imaging

Next, we assess the proposed method for aberration-guided hyperspectral imaging which is described in Sec. 4.4. For synthetic assessment, we use a different dataset CAVE [Yasuma et al. 2008] from the training datasets, and we also experimentally assess the proposed method on real-world captures. Qualitative and quantitative comparisons using synthetic datasets are reported in Fig 8 and Tab 2, respectively, and experimental assessments are presented in Fig 9. Additional qualitative comparisons are available in the Supplemental Document.

5.3.1 Synthetic Assessment. We compare the proposed methods to two types of snapshot hyperspectral imaging techniques: (i) learned RGB to hyperspectral methods, represented by the HRNet [Zhao et al. 2020], which reconstructs hyperspectral information from a single RGB capture; (ii) diffractive optics-based methods, which encode spectral information optically and recover it computationally, represented by QDO [Li et al. 2022]. Additionally, we include a comparison where only the coded capture is used as input in our proposed method. For these experiments, the model is fed either ground-truth RGB images with simulated noise or simulated sensor measurements based on the optics design specifications of the baseline methods. We use the same camera response curve as the proposed method for HRNet, as it is tailored for unknown, uncalibrated cameras [Arad et al. 2020]. For QDO, we use the pre-trained reconstruction models and camera settings provided by the authors for end-to-end optimization of the optics design and reconstruction. Qualitative and quantitative comparisons are reported in Fig 8 and Tab 2, respectively, while additional comparisons are presented in the Supplemental Document.

Table 2. **Quantitative Evaluation of Hyperspectral Reconstruction Quality.** We evaluate reconstruction quality using the SSIM, PSNR, and SAM metrics. We compare the proposed method against RGB-to-Spectrum methods, represented by HRNet [2020], and recent DOE-based snapshot spectral imaging systems, represented by QDO [2022]. Additionally, we include a comparison where only the coded capture is used as input to our reconstruction network.

	↑PSNR [dB]	↑SSIM	↓SAM
HRNet [2020]	26.76	0.88	0.40
QDO [2022]	23.60	0.67	0.45
Coded Capture Only	24.46	0.72	0.45
Proposed	**32.96**	**0.90**	**0.15**

HRNet [Zhao et al. 2020], the top performer in the NTIRE 2020 Spectral Reconstruction from RGB Image Challenge [Arad et al. 2020] Real Image Track, utilizes a 4-level hierarchical regression network to convert RGB images into hyperspectral ones, extrapolating missing spectral information through priors learned during training without requiring specific setups or camera details. While it delivers high-quality outputs and comparable SSIM scores to our method, its low SAM score highlights its limitations in spectral accuracy, particularly in individual spectral channels.

QDO [Li et al. 2022] proposes quantization-aware deep optics for diffractive snapshot hyperspectral imaging, using a joint optimization of the quantized DOE and a Res-UNet reconstruction network, with consideration of fabrication constraints. However, the quantization in the design phase, while simplifying fabrication, severely limits design freedom and final image quality.

The proposed 2-in-1 camera system merges two optical functionalities into a single optical system, enabling the capture of spectral information through optical encoding while circumventing the spatial resolution loss typically associated with DOE spectral coding, achieving a margin larger than 6 dB in PSNR and a 0.25 dB margin in SAM over the existing method. The reconstruction quality stems from the simultaneously captured uncoded capture. When removing the uncoded capture, the proposed method exhibits limitations similar to other diffractive optics-based methods, validating the proposed 2-in-1 camera design.

5.3.2 Experimental Assessment. In the following, we experimentally verify the proposed optically-coded hyperspectral imager. Fig. 9 shows these PSF measurements visualized in the RGB domain, compared to the designed PSFs. These measurements confirm that the fabricated DOE reproduces the wavelength-dependent shift in the focal point from simulation, resulting in a rainbow-like PSF. However, the measurements also reveal slight blur and varying diffractive efficiencies across different wavelengths. To account for this manufacturing inaccuracy, we employ a similar two-step fine-tuning process as described above for our reconstruction network to compensate

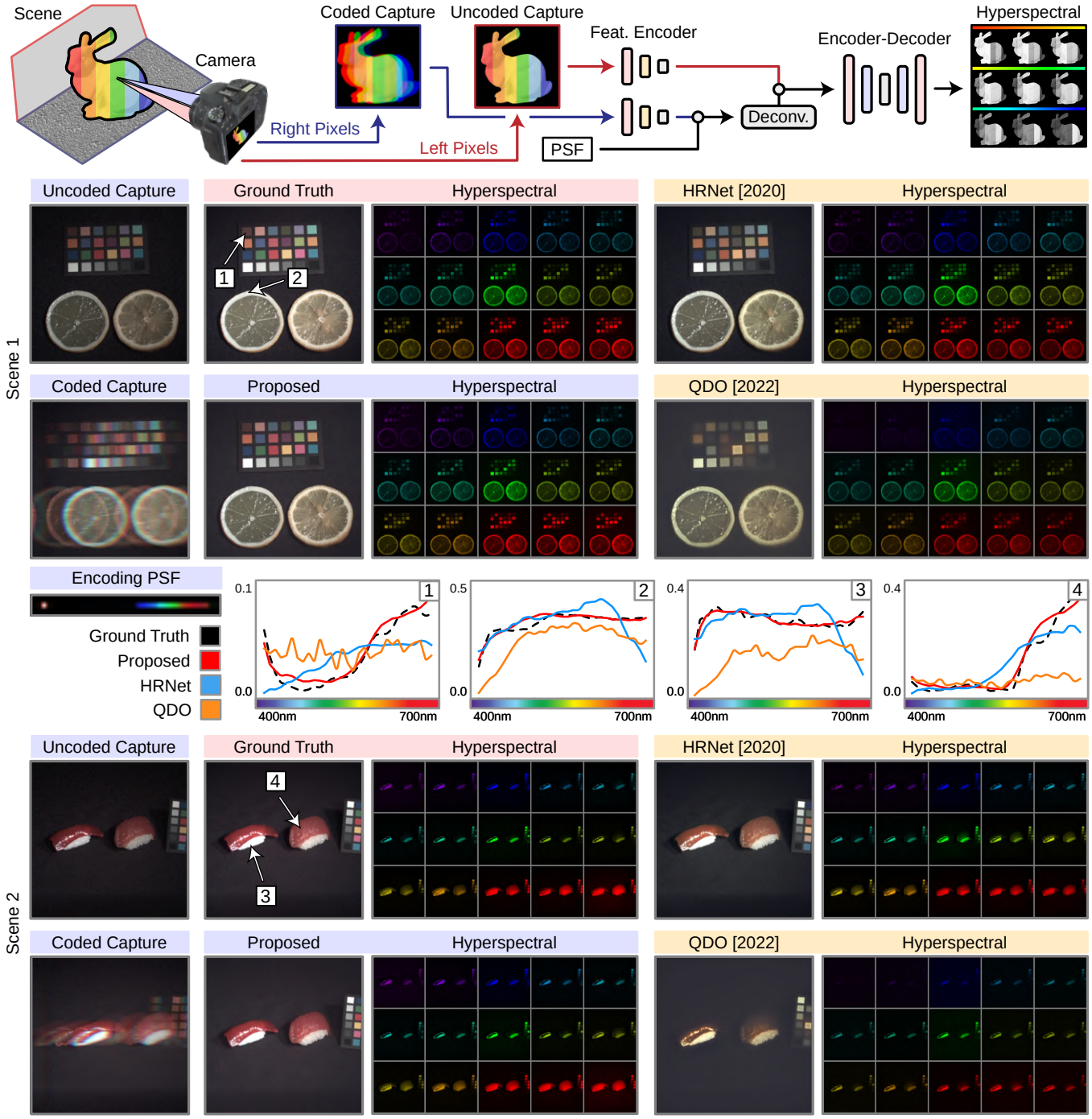

Fig. 8. **Snapshot Hyperspectral Imaging in Simulation.** We assess the proposed method for snapshot hyperspectral imaging with simulated ground truth spectral data (400nm to 700nm) and compare the RGB-to-Spectrum HRNet [Zhao et al. 2020], and DOE-based QDO systems [Li et al. 2022]. For each scene, the leftmost column shows the sensor captures using our method, followed by reconstructions in both RGB and hyperspectral formats. The RGB images are generated from hyperspectral reconstructions and sensor response curves. Due to space constraints, we display alternate hyperspectral channels (410nm to 700nm at 20nm intervals). We also present spectral validation plots of all approaches for four specific points, marked on the Ground Truth RGB image. QCO, limited by its heavily quantized design and spatial resolution loss from optical encoding, faces challenges in high-quality reconstruction. HRNet, while generating plausible results, tends to overfit to its training dataset, particularly at both ends of the spectrum. Our method, capturing both uncoded and coded images, achieves high fidelity in recovering spatial and spectral details.

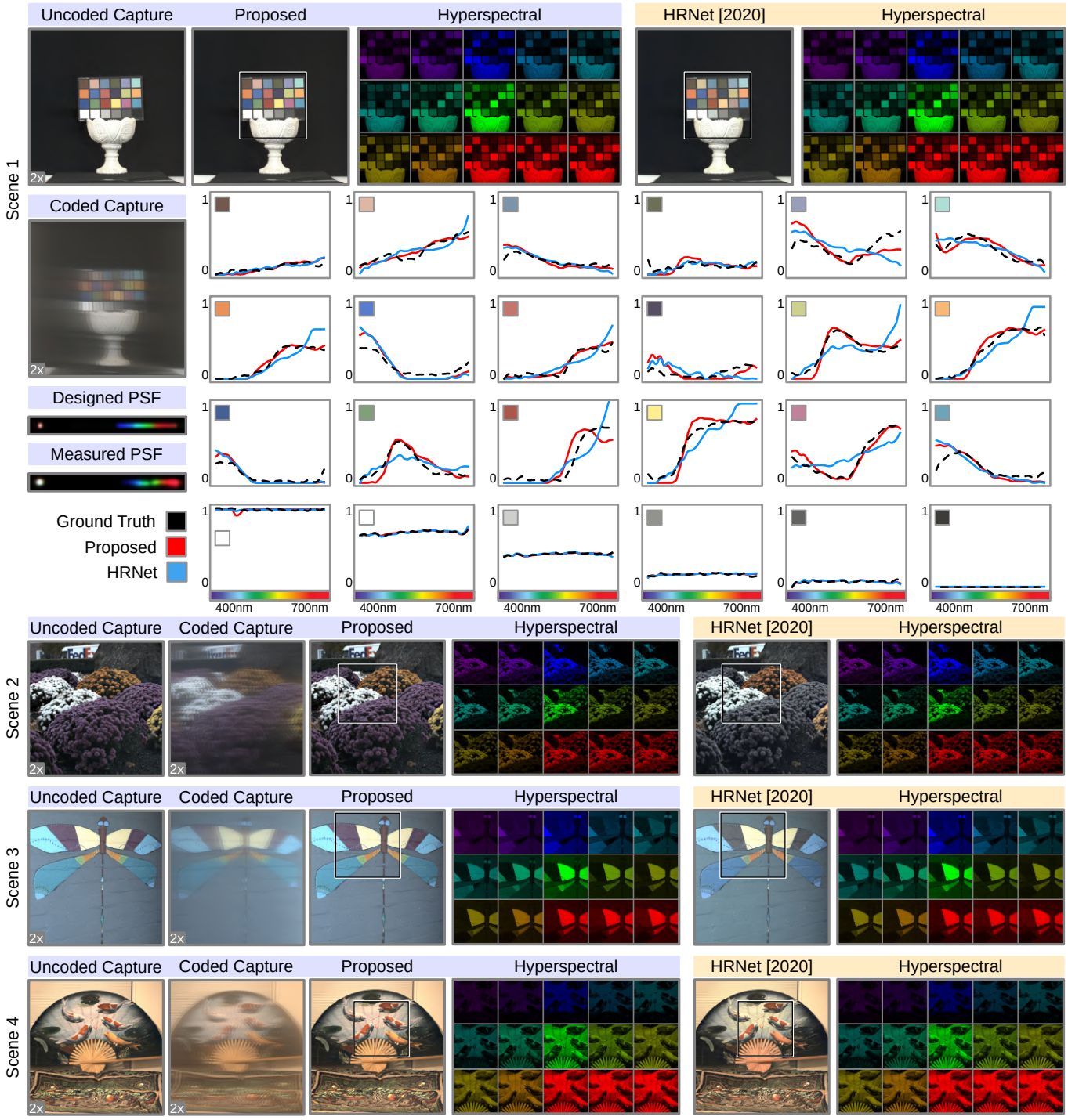

Fig. 9. **Experimental Assessment of Snapshot Hyperspectral Imaging.** We evaluate our method experimentally for snapshot hyperspectral imaging under varying lighting conditions, and compare it against the learned RGB-to-HS approach, HRNet [2020]. On the left, we display the measured PSFs, which verify that the DOE generates the intended rainbow-like PSF. Scene 1 on the top illustrates results from a lab environment, and we include spectral validation plots for all 24 color blocks, with ground truth spectra obtained via a miniature spectrometer. The subsequent rows validate the method in outdoor (Scenes 2 and 3) and indoor (Scene 4) settings. The spectral reconstructions from our method align closely with the measured spectral intensities, in contrast to HRNet, which exhibits notable inaccuracies, especially at the spectrum boundaries. This discrepancy is also evident in out-of-lab experiments, where HRNet struggles with color accuracy, see red and purple images. In the absence of Ground Truth RGB captures, we present the uncoded and coded captures at double intensity, where the uncoded capture serves as a pseudo-ground truth in the RGB domain.

for fabrication inconsistencies in the learned DOE and address differences between synthetic and real-world captures.

We proceed to evaluate the spectral accuracy of the proposed 2-in-1 computational camera in a lab setting before presenting outdoor experimental results. We first perform an evaluation with a MacBeth ColorChecker target under controlled lighting and use a miniature spectrometer to measure the hyperspectral intensity at the center of each color patch, creating a reference spectrum. Next, we capture the same scene with our prototype under identical lighting conditions. The reconstructed spectral curves of each color are then compared against these reference measurements. Additionally, we compare our method to the learned RGB-to-Spectrum method HRNet [Zhao et al. 2020] on the uncoded capture. Reconstructions with both RGB and hyperspectral visualization, along with spectral validation plots, are presented in Fig. 9 and the Supplemental Document. In these comparisons, the spectral curves reconstructed by the proposed method exhibit close alignment with the measured spectral intensities. Conversely, HRNet, while producing plausible results, tends to be less accurate and particularly struggles at the spectrum ends. We also test the proposed method in real-world scenarios, both indoors and outdoors, to assess its adaptability to various lighting conditions. Due to limitations in our point-wise spectral measurement instrumentation, we did not obtain reference spectral curves in these uncontrolled lighting environments, and compare the results of our proposed method solely with those from a learned RGB-to-HS (hyperspectral) method. Additional details on hyperspectral intensity measurements and additional comparisons can be found in the Supplemental Document.

5.4 Monocular Depth from Coded Defocus

Next, we assess the proposed method for depth from coded defocus as an application introduced in Sec. 4.5 using both simulated captures and real-world captures obtained by our experimental prototype. Qualitative and quantitative comparisons using synthetic data are reported in Fig 10 and Tab 3, respectively, and experimental results are reported in Fig 11. Additional qualitative comparisons are available in the Supplemental Document.

Synthetic Assessment. We compare the proposed method to two types of existing monocular depth methods: (i) learned monocular (relative) depth methods that estimate the relative depth from a single capture, represented by MiDaS [Ranftl et al. 2022] and Zoedepth [Bhat et al. 2023] (ii) (absolute) depth from defocusing methods that estimate absolute depth from optically encoded depth defocusing cues, represented by Deep DfD [Ikoma et al. 2021]. Additionally, we include a comparison where only the coded capture is used as input to the reconstruction method. For MiDas, we use the MiDaS v3-large weight and provide the simulated uncoded capture as an input. For Zoedepth, we tested all 3 provided model checkpoints on our test scenes and reported the highest score (achieved by ZoeD-M12-N). And for Deep DfD, we use the pretrained network weights, optics design, camera settings and image formation model provided by the authors. We evaluate reconstruction quality using several metrics: MAE, RMSE, average logarithmic error (average $\log_{10}$ error), and the percentage of pixels where the ratio of predicted to ground truth depth is within a factor of 1.25, denoted as δ<1.25. Qualitative and quantitative findings are reported in Fig 10 and Tab 3, respectively, while additional comparisons are presented in the Supplemental Document.

Table 3. **Quantitative Evaluation of Monocular Depth Imaging Accuracy.** We evaluate reconstruction quality using several metrics: MAE, RMSE, average logarithmic error ($\log_{10}$), and the percentage of pixels where the ratio of predicted to ground truth depth is within a factor of 1.25 (δ<1.25). We compare the proposed method against monocular depth methods, represented by MiDaS [2022] and Zoedepth [2023], and recent DOE-based depth from defocus method, represented by Deep DfD [2021]. MiDaS generates only relative depth information, which we adjust to align with the known target depth range. Zoedepth is designed to provide monocular depth estimation with metric scale. Therefore, we report its performance using both the original output and the output scaled to the known target depth range. Additionally, we include a comparison where only the coded capture is used as input to the proposed method.

	↓MAE [m]	↓RMSE [m]	↓$\log_{10}$	↑δ<1.25
ZoeDepth [2023]	1.566	1.702	0.242	0.123
ZoeDepth [2023] (re-scaled)	1.042	1.203	0.158	0.384
MiDaS [2022] (re-scaled)	0.736	0.918	0.107	0.609
Deep DfD [2021]	0.356	0.485	0.051	0.894
Coded-Only	0.145	0.223	0.018	0.985
Proposed	**0.086**	**0.147**	**0.011**	**0.993**

MiDaS [Ranftl et al. 2022] computes relative inverse depth (disparity) from a single image and achieves high zero-shot cross-dataset performance by effectively leveraging mixed-dataset training. Nonetheless, like all monocular depth methods, MiDaS is constrained by its singular input modality, resulting in scale ambiguity: without a reference frame, the method faces challenges in accurately determining the true size of objects or their precise distance from the camera. Zoedepth [Bhat et al. 2023] is a recent work that combines relative and metric depth estimation for monocular depth with metric scale. It employs a two-stage process: first, pre-training an encoder-decoder on relative depth, followed by fine-tuning with domain-specific heads for metric depth. Although it predicts reasonable depth scales (1 to 10 meters for test scenes of 1 to 5 meters), the absolute depth measurements are often geometrically inaccurate.

Deep DfD [Ikoma et al. 2021] aims to recover absolute depth from a single defocused shot, using a jointly optimized DOE and a depth reconstruction network. While it effectively recovers absolute depth, the lack of a sharp in-focus capture limits its ability to discern fine details in cluttered areas.

Our proposed 2-in-1 computational camera lifts the limitations of both approaches. It merges two optical systems, obtaining a high-resolution, sharp capture of the scene along with a coded capture that encodes absolute depth information from defocus cues. These capabilities enable our method to effectively recover absolute depth with fine details. As for the other applications, we again confirm the effectiveness of access to the uncoded capture in the reported results, validating the proposed approach.

Experimental Assessment. Finally, we experimentally validate the proposed method for monocular depth imaging. We first measure the depth-dependent PSFs of our prototype system equipped with the fabricated DOE at various distances from the source, ranging from 1m to 5m at 0.5m intervals. The PSF measurements, shown at

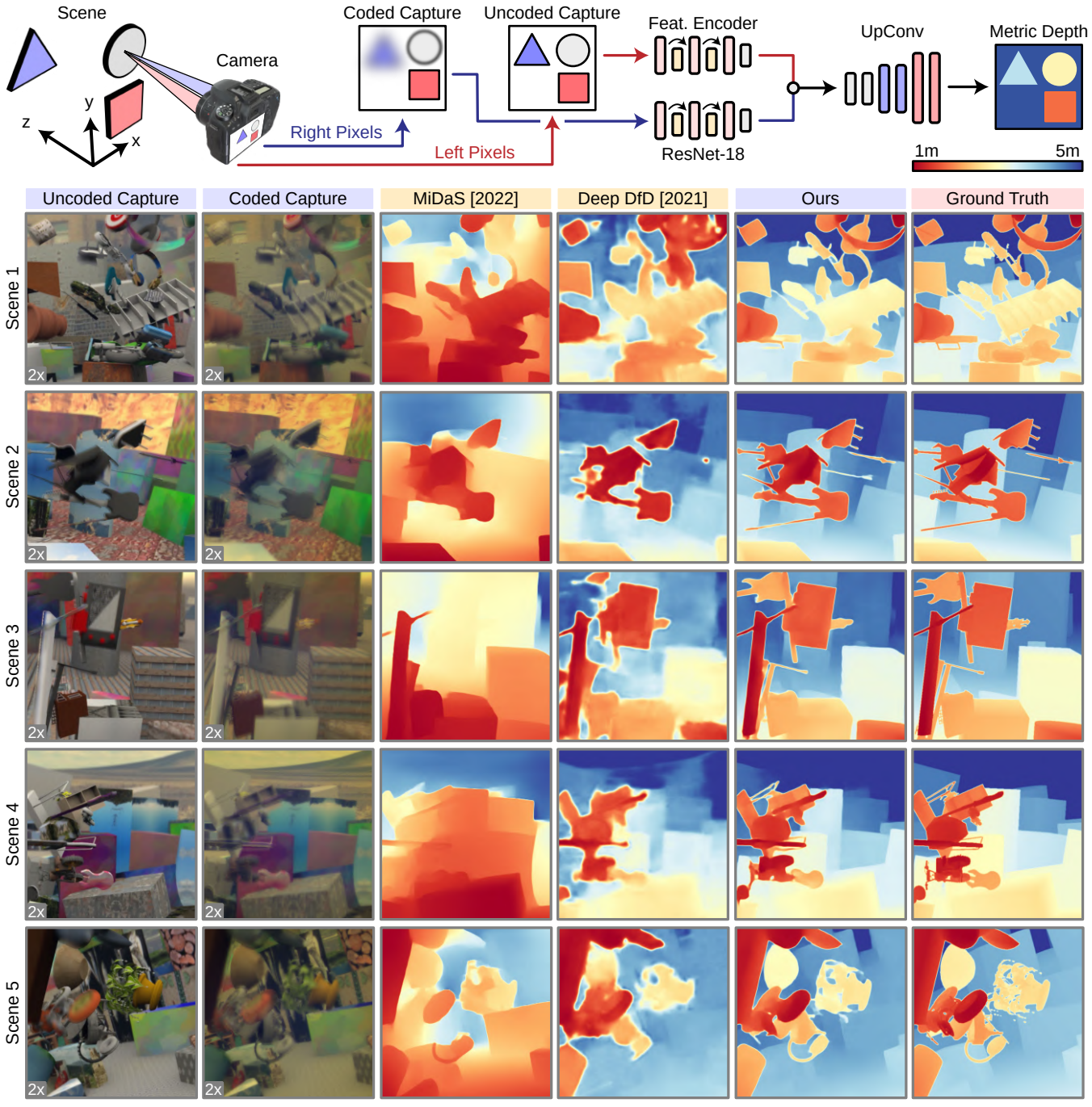

Fig. 10. **Monocular Depth Imaging in Simulation.** We assess our approach for monocular depth estimation in simulation by comparing our method to the monocular depth estimation method MiDaS [Ranftl et al. 2022], and DOE-based depth from defocus method Deep DfD [Ikoma et al. 2021]. For each scene, the leftmost two columns display the sensor captures using our method at double intensity, followed by depth reconstructions from different methods. We scale MiDas relative depth output to match the known target depth range. While MiDaS estimates a qualitatively plausible depth map, their *estimation remains relative and misrepresent the spatial relationship of non-adjacent objects.* Deep DfD, capable of recovering depth scale, faces challenges in resolving fine details. Our method, leveraging both the sharp details from the in-focus uncoded capture and the depth cues from the coded captures, is able to accurately capture both the scale and details in the scene.

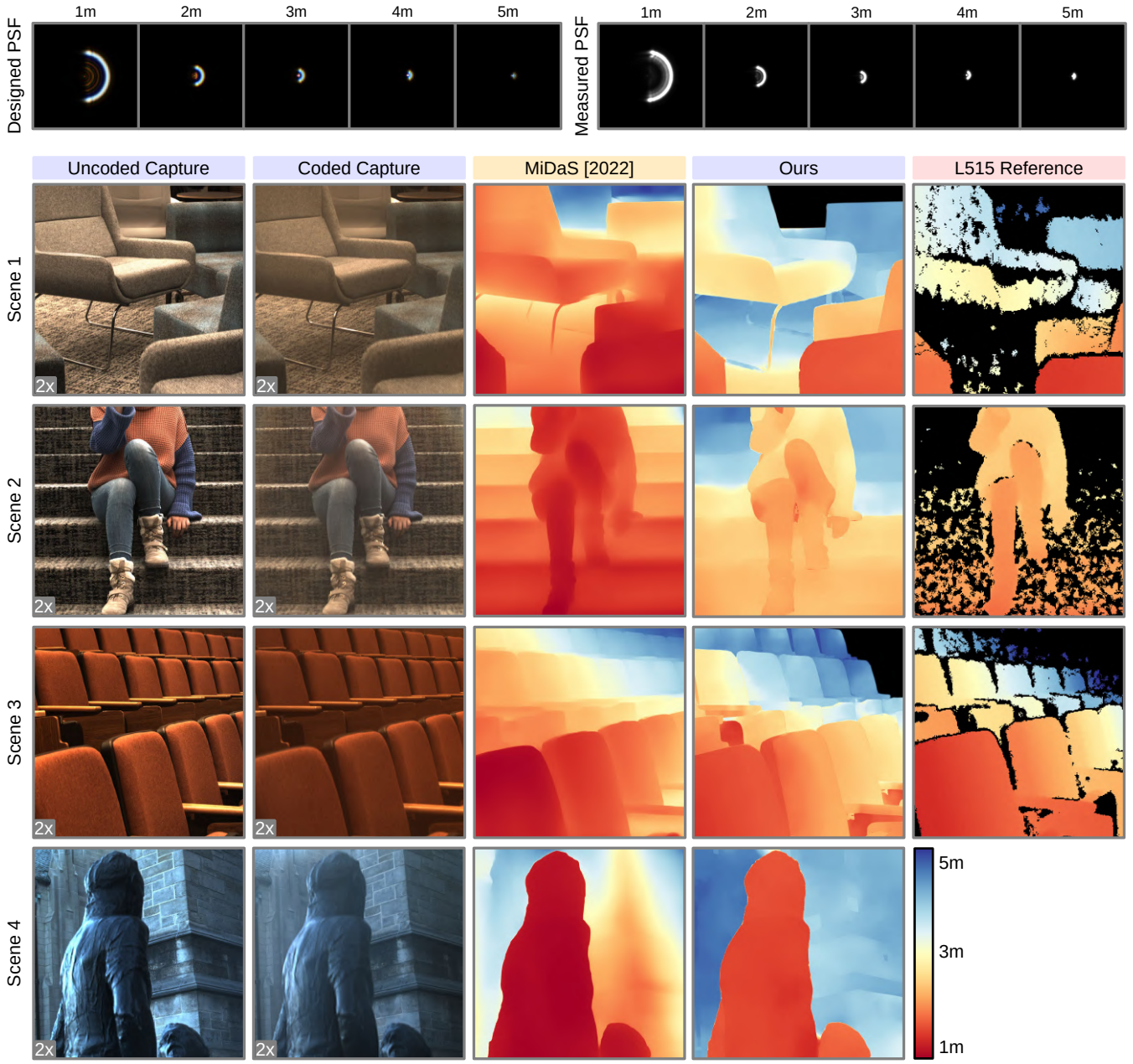

Fig. 11. **Experimental Assessment of Optically Coded Depth Imaging.** We evaluate the proposed method in both indoor (Scenes 1 to 3) and outdoor (Scene 4) environments and compare it to the monocular depth method MiDaS [Ranftl et al. 2021] run on the uncoded capture and rescaled to the target depth range. The top row of our results shows the measured depth-dependent encoding PSFs from our prototype, covering depths from 1m to 5m, which contain the intended half-ring PSF as per our design. For each scene, the leftmost two columns display the sensor captures using our method at double intensity, followed by depth reconstructions from different methods. In the case of indoor scenes, we employ a solid-state LiDAR camera (RealSense L515) to gather absolute depth data from the scenes, serving as a reference. Areas where the RealSense L515 camera was unable to provide measurements are indicated with a black mask. The depth reconstructions produced by our proposed method are in close alignment with the RealSense L515 reference depth. In contrast, MiDaS is limited to provide a plausible relative depth map and often inaccurately merges unconnected objects into a singular, continuous depth profile.

the top of Fig. 11, are then compared to the designed PSFs. These measurements confirm that the DOE accurately reproduces the depth-dependent half-ring PSFs as predicted in our simulations. For network fine-tuning, we use the post-fabrication measured PSFs with synthetic data to help compensate for manufacturing discrepancies. We evaluate depth reconstruction in both indoor and outdoor settings, employing a solid-state LiDAR camera (Intel® RealSense™ LiDAR Camera L515) to obtain reference absolute depth data for each indoor scene. However, we note that reference captures from this camera can be unreliable under high ambient illumination or for surfaces that are either highly reflective or highly light-absorbing. We report the experimental results in Fig.11 and the Supplemental Document. For each scene, we present the uncoded and coded captures from our prototype at double intensity, the depth reconstructions of the proposed method, and the RealSense L515 measurement for reference. We also compared against the monocular depth estimation method MiDaS [Ranftl et al. 2021] run on the uncoded capture and adjusted to the appropriate depth range. Benefiting from the defocus encoding, our single-shot method successfully recovers absolute depth information that aligns closely with the RealSense L515 reference. In outdoor experiments, where the RealSense L515 sensor fails to estimate accurate depth under strong ambient light, we limit our comparison to the MiDaS baseline method.

6 CONCLUSION

We investigate a computational camera design capable of capturing both a conventional and optically encoded image in a single shot. Over the last two decades, a substantial body of work has proposed computational optics that encode scene information into image measurements. However, they inherently sacrifice conventional image quality by design. This makes computational recovery fundamentally challenging, often necessitating a separate camera for uncoded capture in practice. To address this, we adapt dual-pixels sensors for split-aperture split-wavefront capture. Dividing the aperture into two modulated and unmodulated halves, we acquire domain-specific computational and conventional images in a single shot and single camera system at no additional computational cost. We then demonstrate the utility of this aligned, uncoded capture for a range of computational imaging tasks, outperforming existing single-image methods.

While this marks a step towards practical computational optics, we implemented the method with a reduced aperture due to nanofabrication constraints. However, considering the typically small-aperture configuration of smartphone cameras, this prototype restriction could be well addressed in the future via co-design with smartphone optics. Additionally, while we simulate and utilize the defocus-disparity for out-of-focus light in our depth reconstruction application, we assume that the scene is all-in-focus for other applications. For potential future applications to perform coded reconstruction of out-of-focus scene content (e.g., image deblurring), it could be beneficial to jointly simulate and solve for scene depth. The proposed design principle also paves the way for a wide range of future task-specific computational camera setups, including active illumination dual aperture cameras, optical signatures for authenticated communication, and optical neural network computation.

ACKNOWLEDGMENTS

Ilya Chugunov was supported by an NSF GRFP (2039656). Felix Heide was supported by an Amazon Science Research Award, Packard Foundation Fellowship, Sloan Research Fellowship, Sony Young Faculty Award, the Project X Fund, and NSF CAREER (2047359). This work was in part supported by KAUST baseline funding. The DOE design was fabricated in the KAUST Nanofabrication Core Lab. Authors also appreciates Thomas Eboli for fruitful discussions.

REFERENCES

Abdullah Abuolaim and Michael S Brown. 2020. Defocus deblurring using dual-pixel data. In *European Conference on Computer Vision*. Springer, 111–126.

Abdullah Abuolaim, Radu Timofte, and Michael S Brown. 2021. NTIRE 2021 challenge for defocus deblurring using dual-pixel images: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 578–587.

Manoj Aggarwal and Narendra Ahuja. 2004. Split aperture imaging for high dynamic range. *International Journal of Computer Vision* 58 (2004), 7–17.

Boaz Arad and Ohad Ben-Shahar. 2016. Sparse Recovery of Hyperspectral Signal from Natural RGB Images. In *European Conference on Computer Vision*. Springer, 19–34.

Boaz Arad, Radu Timofte, Ohad Ben-Shahar, Yi-Tun Lin, and Graham D Finlayson. 2020. Ntire 2020 challenge on spectral reconstruction from an rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 446–447.

Seung-Hwan Baek, Hayato Ikoma, Daniel S Jeon, Yuqi Li, Wolfgang Heidrich, Gordon Wetzstein, and Min H Kim. 2021. Single-shot Hyperspectral-Depth Imaging with Learned Diffractive Optics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2651–2660.

Seung-Hwan Baek, Incheol Kim, Diego Gutierrez, and Min H Kim. 2017. Compact single-shot hyperspectral imaging using a prism. *ACM Transactions on Graphics (TOG)* 36, 6 (2017), 1–12.

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

X Briottet, Y Boucher, A Dimmeler, A Malaplate, A Cini, Marco Diani, HHPT Bekman, P Schwering, T Skauli, I Kasen, et al. 2006. Military applications of hyperspectral imagery. In *Targets and backgrounds XII: Characterization and representation*, Vol. 6239. SPIE, 82–89.

Nicola Brusco, S Capeleto, M Fedel, Anna Paviotti, Luca Poletto, Guido Maria Cortelazzo, and G Tondello. 2006. A system for 3D modeling frescoed historical buildings with multispectral texture information. *Machine Vision and Applications* 17, 6 (2006), 373–393.

Marcela Carvalho, Bertrand Le Saux, Pauline Trouvé-Peloux, Andrés Almansa, and Frédéric Champagnat. 2018. Deep Depth from Defocus: how can defocus blur improve 3D estimation using dense neural networks?. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

A. Chakrabarti and T. Zickler. 2011. Statistics of Real-World Hyperspectral Images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 193–200.

Julie Chang and Gordon Wetzstein. 2019. Deep optics for monocular depth estimation and 3d object detection. (2019), 10193–10202.

Su-Kai Chen, Hung-Lin Yen, Yu-Lun Liu, Min-Hung Chen, Hou-Ning Hu, Wen-Hsiao Peng, and Yen-Yu Lin. 2023. Learning continuous exposure value representations for single-image hdr reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12990–13000.

Paul E Debevec and Jitendra Malik. 2008. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*. 1–10.

Jiangxin Dong, Stefan Roth, and Bernt Schiele. 2020. Deep wiener deconvolution: Wiener meets deep learning for image deblurring. *Advances in Neural Information Processing Systems* 33 (2020), 1048–1059.

Liang Gao and Lihong V Wang. 2016. A review of snapshot multidimensional optical imaging: measuring photon tags in parallel. *Physics reports* 616 (2016), 1–37.

Rahul Garg, Neal Wadhwa, Sameer Ansari, and Jonathan T Barron. 2019. Learning single camera depth estimation using dual-pixels. In *Proceedings of the IEEE/CVF international conference on computer vision*. 7628–7637.

Carl Friedrich Gauss. 1843. *Dioptric examinations by CF Gauss.* in the Dieterich bookstore.

Bhargav Ghanekar, Vishwanath Saragadam, Dushyant Mehran, Anna-Karin Gustavsson, Aswin C. Sankaranarayanan, and Ashok Veeraraghavan. 2022. PS2F: Polarized Spiral Point Spread Function for Single-Shot 3D Sensing. *IEEE Trans. Pattern Analysis and Machine Intelligence (TPAMI) / Special Issue of ICCP* (August 2022).

Paul Green, Wenyang Sun, Wojciech Matusik, and Fredo Durand. 2007. Multi-aperture photography. In *Acm Siggraph 2007 Papers*. 68–es.

Shir Gur and Lior Wolf. 2019. Single Image Depth Estimation Trained via Depth From Defocus Cues. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nathan A Hagen, Liang S Gao, Tomasz S Tkaczyk, and Robert T Kester. 2012. Snapshot advantage: a review of the light collection improvement for parallel high-dimensional measurement systems. *Optical Engineering* 51, 11 (2012), 111702.

Harel Haim, Shay Elmalem, Raja Giryes, Alex Bronstein, and Emanuel Marom. 2018. Depth Estimation From a Single Image Using Deep Learned Phase Coded Mask. *IEEE Transactions on Computational Imaging* 4 (2018), 298–310.

Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. 2016. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (ToG)* 35, 6 (2016), 1–12.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.

Hayato Ikoma, Cindy M Nguyen, Christopher A Metzler, Yifan Peng, and Gordon Wetzstein. 2021. Depth from defocus with learned optics for imaging and occlusion-aware depth estimation. In *2021 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–12.

Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. 2019. Compact snapshot hyperspectral imaging with diffracted rotation. (2019).

Zeeshan Khan, Mukul Khanna, and Shanmuganathan Raman. 2019. Fhdr: Hdr image reconstruction from a single ldr image using feedback network. In *2019 IEEE Global Conference on Signal and Information Processing (GlobalSIP)*. IEEE, 1–5.

Masahiro Kobayashi, Michiko Johnson, Yoichi Wada, Hiromasa Tsuboi, Hideaki Takada, Kenji Togo, Takafumi Kishi, Hidekazu Takahashi, Takeshi Ichikawa, and Shunsuke Inoue. 2016. A low noise and high sensitivity image sensor with imaging and phase-difference detection AF in all pixels. *ITE Transactions on Media Technology and Applications* 4, 2 (2016), 123–128.

Sarawak Kuching. 2007. The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis. *Journal of Computer Science* 3, 6 (2007), 419–423.

Anat Levin, Rob Fergus, Frédo Durand, and William T Freeman. 2007. Image and depth from a conventional camera with a coded aperture. *ACM transactions on graphics (TOG)* 26, 3 (2007), 70–es.

Lingen Li, Lizhi Wang, Weitao Song, Lei Zhang, Zhiwei Xiong, and Hua Huang. 2022. Quantization-Aware Deep Optics for Diffractive Snapshot Hyperspectral Imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19780–19789.

Guoxuan Liu, Ning Xu, Huaidong Yang, Qiaofeng Tan, and Guofan Jin. 2022. Miniaturized structured illumination microscopy with diffractive optics. *Photonics Research* 10, 5 (2022), 1317–1324.

Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. 2020. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1651–1660.

Kris Malkiewicz and M David Mullen. 2009. *Cinematography*. Simon and Schuster.

N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox. 2016. 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)*. http://lmb.informatik.uni-freiburg.de/Publications/2016/MIFDB16 arXiv:1512.02134.

Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. 2020. Deep optics for single-shot high-dynamic-range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1375–1385.

Roope Näsi, Eija Honkavaara, Päivi Lyytikäinen-Saarenmaa, Minna Blomqvist, Paula Litkey, Teemu Hakala, Niko Viljanen, Tuula Kantola, Topi Tanhuanpää, and Markus Holopainen. 2015. Using UAV-based photogrammetry and hyperspectral imaging for mapping bark beetle damage at tree-level. *Remote Sensing* 7, 11 (2015), 15467–15493.

Shree K Nayar. 2006. Computational cameras: Redefining the image. *Computer* 39, 8 (2006), 30–38.

Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Lucien E Weiss, Onit Alalouf, Tal Naor, Reut Orange, Tomer Michaeli, and Yoav Shechtman. 2020. DeepSTORM3D: dense 3D localization microscopy and PSF design by deep learning. *Nature methods* 17, 7 (2020), 734–740.

Liyuan Pan, Shah Chowdhury, Richard Hartley, Miaomiao Liu, Hongguang Zhang, and Hongdong Li. 2021. Dual pixel exploration: Simultaneous depth estimation and image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4340–4349.

Sri Rama Prasanna Pavani and Rafael Piestun. 2009. 3D microscopy with a double-helix point spread function. In *Three-Dimensional and Multidimensional Microscopy: Image Acquisition and Processing XVI*, Vol. 7184. SPIE, 65–71.

Yifan Peng, Qilin Sun, Xiong Dun, Gordon Wetzstein, Wolfgang Heidrich, and Felix Heide. 2019. Learned large field-of-view imaging with thin-plate optics. *ACM Trans. Graph.* 38, 6 (2019), 219–1.

Abhijith Punnappurath and Michael S. Brown. 2019. Reflection Removal Using a Dual-Pixel Sensor. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. 2022. Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 3 (2022).

Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. 2010. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann.

Mushfiqur Rouf, Rafał Mantiuk, Wolfgang Heidrich, Matthew Trentacoste, and Cheryl Lau. 2011. Glare encoding of high dynamic range images. In *CVPR 2011*. IEEE, 289–296.

Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. 2020. Single image HDR reconstruction using a CNN with masked features and perceptual loss. *arXiv preprint arXiv:2005.07335* (2020).

Yoav Shechtman, Steffen J Sahl, Adam S Backer, and William E Moerner. 2014. Optimal point spread function design for 3D imaging. *Physical review letters* 113, 13 (2014), 133902.

Zheng Shi, Yuval Bahat, Seung-Hwan Baek, Qiang Fu, Hadi Amata, Xiao Li, Praneeth Chakravarthula, Wolfgang Heidrich, and Felix Heide. 2022. Seeing through obstructions with diffractive cloaking. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–15.

Vincent Sitzmann, Steven Diamond, Yifan Peng, Xiong Dun, Stephen Boyd, Wolfgang Heidrich, Felix Heide, and Gordon Wetzstein. 2018. End-to-end optimization of optics and image processing for achromatic extended depth of field and super-resolution imaging. *ACM Trans. Graph. (TOG)* 37, 4 (2018), 114.

Qilin Sun, Ethan Tseng, Qiang Fu, Wolfgang Heidrich, and Felix Heide. 2020. Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ethan Tseng, Ali Mosleh, Fahim Mannan, Karl St-Arnaud, Avinash Sharma, Yifan Peng, Alexander Braun, Derek Nowrouzezahrai, Jean-Francois Lalonde, and Felix Heide. 2021. Differentiable Compound Optics and Processing Pipeline Optimization for End-to-end Camera Design. *ACM Transactions on Graphics (TOG)* 40, 2, Article 18 (2021).

Congli Wang, Ni Chen, and Wolfgang Heidrich. 2022. dO: A differentiable engine for deep lens design of computational imaging systems. *IEEE Transactions on Computational Imaging* 8 (2022), 905–916.

Hongcheng Wang, Ramesh Raskar, and Narendra Ahuja. 2005. High dynamic range video using split aperture camera. In *IEEE 6th Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras, Washington, DC, USA*. Citeseer.

Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. 2019. PhaseCam3D — Learning Phase Masks for Passive Single View Depth Estimation. *2019 IEEE International Conference on Computational Photography (ICCP)* (2019), 1–12.

Shumian Xin, Neal Wadhwa, Tianfan Xue, Jonathan T Barron, Pratul P Srinivasan, Jiawen Chen, Ioannis Gkioulekas, and Rahul Garg. 2021. Defocus map estimation and deblurring from a single dual-pixel image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2228–2238.

F. Yasuma, T. Mitsunaga, D. Iso, and S.K. Nayar. 2008. *Generalized Assorted Pixel Camera: Post-Capture Control of Resolution, Dynamic Range and Spectrum*. Technical Report.

Jonghee Yoon, James Joseph, Dale J Waterhouse, A Siri Luthman, George SD Gordon, Massimiliano Di Pietro, Wladyslaw Januszewicz, Rebecca C Fitzgerald, and Sarah E Bohndiek. 2019. A clinically translatable hyperspectral endoscopy (HySE) system for imaging the gastrointestinal tract. *Nature communications* 10, 1 (2019), 1–13.

Yuzhi Zhao, Lai-Man Po, Qiong Yan, Wei Liu, and Tingyu Lin. 2020. Hierarchical regression network for spectral reconstruction from RGB images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 422–423.