Split-Aperture 2-in-1 Computational Cameras - Supplemental Document -

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

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 - Supplemental Document - . *ACM Trans. Graph.* 43, 4, Article 141 (July 2024), 14 pages. https://doi.org/10.1145/3658225

In this Supplemental Document, we present additional results and method details in support of the findings from the main manuscript. Specifically, we first describe the fabrication of the diffractive optical element, introduce color-accurate Bayer imaging as an additional application of the proposed 2-in-1 camera, and additional experiments on exploring End-to-End optimization. Next, we describe pre-processing steps to obtain uncoded and coded captures. We then provide additional network details, PSF calibration process, fine-tuning description, ground truth acquisition, and additional experiments for each of the applications from the main document (snapshot HDR imaging, snapshot hyperspectral imaging, and absolute depth imaging).

Contents

*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 FABRICATION OF DIFFRACTIVE OPTICAL ELEMENT

We fabricate the DOEs with photo-lithography (PL) and reactive-ion etching (RIE) techniques with 5.4 micrometer pixel pitch. The fabrication procedures are as follows.

Sample Preparation. The substrate is a fused silica wafer with 4-inch diameter and 0.5 mm thickness. The wafer is first placed in Piranha bath at 115 °C for 10 min, followed by de-ionized water rinse and nitrogen drying to remove contaminants. A thin layer of Chromium (Cr) is deposited onto the wafer surface by sputtering as a hard mask layer for the following etching step.

Master Mask Fabrication. Master masks are fabricated by laser direct writing on 5-inch soda lime marks with Heidelberg μPG 501. In order to achieve 16-level structures, four master masks are required.

Patterning. The patterns on the master masks are transferred to the wafer by photo-lithography. We first prepare the wafer with HMDS (Hexamethyldisilazane) vapor priming at 150 °C for 20 min. Then, a thin film of photoresist AZ1505 is spin-coated on the wafer, with 0.6 μm thickness. The wafer is aligned with the master mask on a contact aligner (EVG6200 ∞) with a separation of 30 μm in between. We apply UV exposure (9 mJ/cm^2) to transfer the patterns from the mask to the photoresist. In the following development step, the exposed areas are removed by the developer AZ726MIF for 17 sec. Next, the open areas on the Cr layer are etched by Cr etchant ($HClO_4$ and $(NH_4)_2[Ce(NO_3)_6]$ solution) for 1 min to transfer the patterns on the hard mask on the wafer. The residual photoresist is removed by acetone afterwards.

Etching. The structures on the wafer are fabricated by dry etching (RIE) in a vaccum chamber. We use plasma of 15 sccm CHF_3 and 5 sccm O_2 at 10 °C and control the time of etching for the desired depth. The open areas on the wafer (no Cr covering) are selectively removed by the plasma. After the etching, the Cr layers are removed for the next step.

Finishing. We repeat the above patterning and etching steps for 4 times to create the 16-level structures. In each etching step, the target depths are 75 nm, 150 nm, 300 nm, and 600 nm for the design wavelength at 550 nm. In the finishing step, we deposit a Cr aperture around the clear region of the DOE to prevent unwanted light outside. The samples are diced with a dicing saw to the physical dimension of 10 mm × 10 mm.

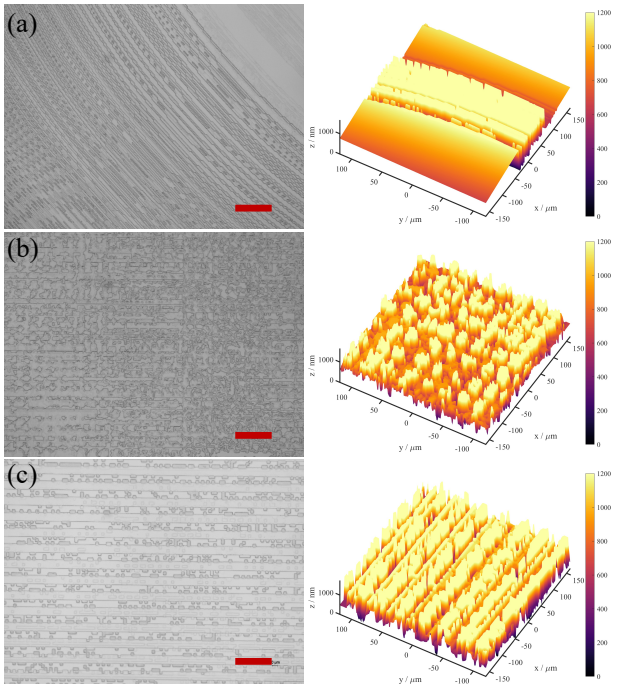

Fig. 1. Microscopic images and Zygo profile measurements for the fabricated DOEs. (a) Monocular Depth from Defocus. (b) Hyperspectral Imaging. (c) HDR imaging. Scale bar is 50 μm. Microscopic images are taken by Nikon Eclipse L200N, 20×. 3D height profiles are taken by Zygo NewView 7300, 20×.

2 COLOR-ACCURATE BAYER IMAGING

In this section, we describe an additional novel application. We present this application in simulation as it was conceived after the fabrication of the DOEs presented in the main manuscript.

Conventional optical systems aim to minimize chromatic aberrations and focus all light from a given direction to the same spot on the sensor. On the sensor, a color filter array, typically in a Bayer arrangement [Lukac 2018], rejects all light outside a specific range of wavelengths for a given pixel. Assuming a perfectly imaging system, one shortcoming of this approach is that only the light corresponding to one of the three color channels will be recorded for any given point in a scene, which may result in artificial colors being reconstructed.

Departing from this approach and inspired by Miyata et al. [2021], we investigate a DOE which deliberately maximize chromatic aberration and design the DOE to focus red, green, and blue light onto the corresponding Bayer filter locations. This allows us to improve the color accuracy in scenes where there are high-frequency details: in the grid-like regions of the scene that are correspondingly aligned with the Bayer sensor, unbiased samples can be collected for the R, G, and B channels instead of a single one of these channels. For other regions, the light is suppressed instead of providing incomplete R, G, and B samples. To this end, we aim to optically implement the target PSF

$$p'_{\text{Bayer}}(\lambda) = \begin{cases} \Delta(256, 256), & \lambda = \lambda_{\text{R}} \\ \Delta(257, 256), & \lambda = \lambda_{\text{G}} \\ \Delta(257, 257), & \lambda = \lambda_{\text{B}}, \end{cases} \tag{1}$$

where $\Delta(i, j)$ represents a dirac delta at pixel (i, j) and λ_{R}, λ_{G} and λ_{B} are corresponding RGB wavelength. We optimize the DOE phase profile to fit this PSF with

$$\mathcal{L}_{p,\text{Bayer}} = \mathcal{L}_1 \left(F\left(h_{\text{L}}, 0, \infty_{\text{O}}, \lambda_{\text{RGB}}\right), p'_{\text{Bayer}} \right), \tag{2}$$

where $\mathcal{L}_1$ distance between the simulated left PSF p_{L} and the target PSF with depth at optical infinity ∞_{O} for discrete RGB wavelength samples $\lambda_{\text{RGB}} = \{\lambda_{\text{R}}, \lambda_{\text{G}}, \lambda_{\text{B}}\}$.

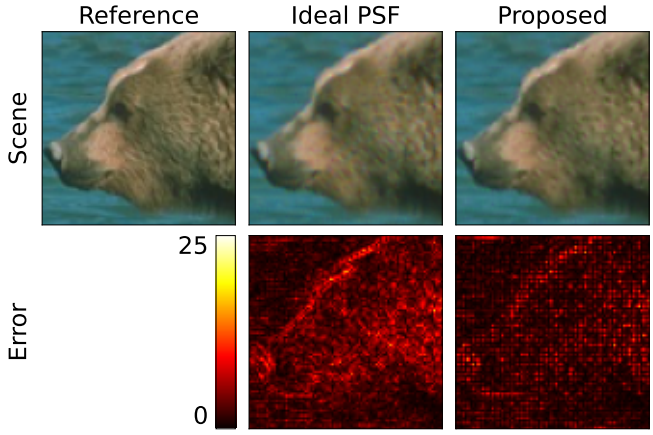

Fig. 2. Synthetically reconstructed scene after the simulation of a Bayer filter (top row), where the PSF is either an ideal Dirac delta or the color-shifted PSF of the proposed color-accurate Bayer imaging method. In both methods, nearest-neighbor interpolation is used to reconstruct all color channels from the Bayer image. In contrast to an ideal PSF, the proposed method leads to a more accurate reconstruction of the color channels, which we show as the mean absolute error over the U and V channels in LUV color space (bottom row).

Table 1. PSNR in RGB and LUV color spaces when synthetically reconstructing natural scenes after the simulation of a Bayer filter using our color-accurate Bayer imaging method. We compare to the PSF of an ideal image formation process (i.e., a centered Dirac delta PSF).

	RGB [dB]	L [dB]	U [dB]	V [dB]
Ideal PSF (debayering)	**28.9**	**35.5**	37.6	35.7
Ideal PSF (nearest neighbor)	**28.9**	**35.5**	37.7	35.7
Proposed	**28.9**	35.4	**39.5**	**37.8**

We assess the proposed method for color-accurate Bayer imaging in simulation using the BSDS500 dataset [Arbelaez et al. 2010], composed of 500 images of diverse real-world scenes. We consider the PSF from Eq. (1) for the proposed method, and an identity transformation for the baseline image. For all methods, we apply the PSF, simulate a naive RGGB Bayer filter (with a transmittance of either

1s or 0s for the corresponding color channels), and add Gaussian noise ($\sigma = 1$ pixel). From the simulated Bayer image, we reconstruct the original image using nearest-neighbor interpolation for the proposed method, and both nearest-neighbor interpolation and debayering for the baseline method. Then, we compare the PSNR in both RGB and LUV color spaces.

The findings in Tab. 1 and Fig. 2 validate that the proposed method achieves a better color accuracy than the application of a perfect PSF in the presence of a Bayer filter, albeit at the cost of slightly worse luminance reconstruction; incidentally, both methods yield similar reconstruction performance in RGB space. The proposed approach can find use in applications where color accuracy is important, e.g., in artistic or machine-vision applications.

3 END-TO-END OPTIMIZATION FOR HDR APPLICATION

In the main manuscript, we derive motivation for our split-aperture DOE designs from prior work that utilized a single camera setup. For instance, we co-optimize our HDR DOE design alongside the reconstruction network with a streak-like PSF initialization, as opposed to the more common random or focusing PSF initializations. In this section, we explore whether an end-to-end optimization conducted without a prior, can yield better results. We explore two methods for jointly optimizing the DOE and network designs, evaluating their performance through simulations as it was conceived after the fabrication of the DOEs presented in the main manuscript. We present qualitative and quantitative comparisons in Figure 3 and Table 2, respectively.

Table 2. **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 end-to-end learning with random PSF initialization, and recent differentiable proximal solver [2023].

	↓RMSE	↑PSNR	↓RMSE^H	↑PSNR^H
Δ-Prox [2023]	1.58	48.18	10.82	31.33
Rand Init	0.90	54.16	7.28	37.01
Proposed	**0.88**	**54.87**	**7.14**	**37.68**

First, we replace the streak-like PSF initialization with a random diffuser PSF, and jointly optimize both the DOE and the reconstruction network from scratch. The initialization influences the learned PSF to scatter energy similarly to a diffuser, creating a haze around highlights as observed in the 'Random Init' configuration. This scattering makes it challenging for the network to reconstruct fine details from the haze, resulting in a slight performance decrease compared to our proposed design.

Next, we explored Δ-Prox [Lai et al. 2023], a recent development that introduced a domain-specific language (DSL) and compiler for transforming optimization problems into differentiable proximal solvers. This method, which uses model-based proximal optimization, has demonstrated better local minima attainment compared to jointly optimizing the DOE design with a separate deep learning reconstruction network, as the proposed method. As shown in

Fig. 3. **End-to-End Snapshot HDR Methods in Simulation.** We compare the proposed method against end-to-end learning with random PSF initialization, and recent differentiable proximal solver [2023]. For each method, the leftmost column shows the learnt PSF, followed by simulated coded capture and the reconstructed scenes at 0EV, -3EV, and -6EV.

Figure 3 under the "Dprox" configuration, Δ-Prox converged to a markedly different PSF design, replicating highlights both locally and at greater distances, and outperformed all baselines discussed in our main manuscript, as shown in Table 2. However, unlike deep learning networks that can easily copy LDR content from the uncoded capture and focus primarily on highlight recovery, the unrolled ADMM solver must deconvolve the entire image using both coded and uncoded captures, resulting in decreased quality in LDR regions and more artifacts near highlights.

4 PRE-PROCESSING AND CROSS-TALK COMPENSATION

In our experiments, we use the Canon EOS 5D Mark IV dual-pixel sensor, which records raw captures in the Canon Raw 2nd edition (CR2) format. Our preprocessing begins with extracting two raw frames from each CR2 file, a process conducted using RawDigger software. Dual-pixel sensors are designed to output separate left and right views, but in practice, they produce a composite image (frame 1) consisting of the sum of these two views, alongside an individual view (frame 2). To obtain the second view, we subtract frame 2 from frame 1. For saturated pixels, we assign the same value to both views.

To compensate for crosstalk, we capture images of a white wall with half of the aperture blocked, either left or right, as shown in Fig. 4. We note that the cross-talk ratio differs among the Bayer color channels, likely a result of slight chromatic aberrations in the microlenses. Therefore, we compute the cross-talk ratios from the raw captures prior to debayering, addressing each color channel separately. To reduce the effect of sensor noise, we average the ratios for each column, creating a uniform weight for each. We further refine this by fitting a 1D smoothing spline across all columns,

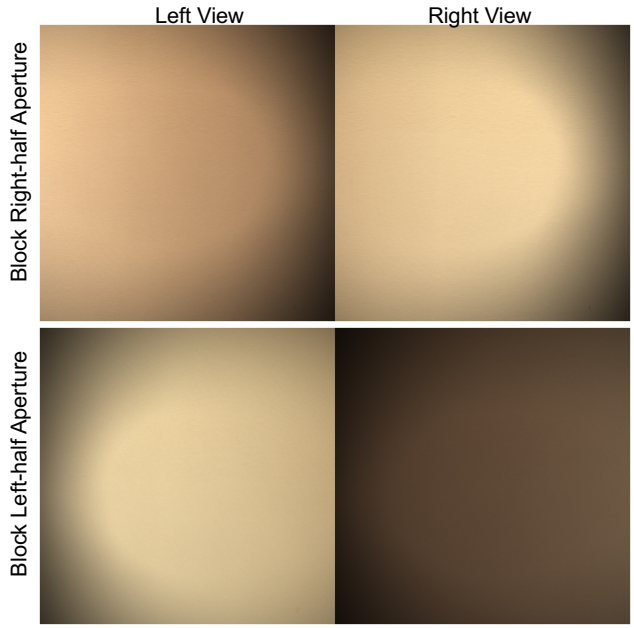

Fig. 4. Cross-talk Calibration Capture with Partial Aperture Blocked. We show here the central 3072 × 3072 pixels of the sensor during a calibration process where half of the aperture is blocked.

which smooths the transition of the cross-talk ratios, enhancing the calibration uniformity.

Finally, we demosaic the calibrated raw captures using bilinear interpolation. The white balance for both captures is adjusted based on the gray world assumption applied to the uncoded capture.

5 OPTICALLY CODED SNAPSHOT HIGH DYNAMIC RANGE IMAGING

Reconstruction Network Architecture. We employ a DWDN-like network architecture to extract the optically encoded information, which initially conducts feature-based inverse filtering on the coded capture, followed by an encoder-decoder network for image reconstruction. See Tab. 3 and 4 for full network specification.

PSF Calibration Setup. A fiber tip (M37L01) coupled with a broadband Fiber-Coupled LED (MBB1F1) is used as a point light source. This light is collimated into plane wave illumination using two achromatic doublet lenses (AC254-150-A-ML).

Additional Details on the Training Process. To train our model, we gather 2039 HDR images with a mix of outdoor night scenes and indoor scenes from HDRi Haven, among which 1839 images are used for training purposes and 200 images are reserved for validation. To accommodate different image sizes, we take random 512 × 512 crops of the images for both training and validation. To ensure each crop contains saturated regions, we multiply the images by a scale factor such that 1% to 5% of pixels are saturated. After the scaling, we clip the pixel values to $[0, 2^8]$ and use the processed image as the target HDR data. During training, we also apply left-right flips as additional data augmentation.

Table 3. HDR Reconstruction Network architecture description (part 1). Specifically,"conv-k(a)-s(b)-LRelu" represents a convolution layer with an $a \times a$ kernel window, using the stride b, followed by a Leaky Relu ($\alpha = 0.02$) activation function, "Res-k(a)-Relu" represents a ResNet Block with an $a \times a$ kernel and Relu activation function, and "Interpolate-(a)" represents interpolate the input scale a. We use "convT" to denote transposed convolution, Wdeconv to denote Wiener-deconvolution using the coded PSF, and "concat" to denote concatenation.

Input	Layer Type	Output (# Channels)
Uncoded_capture	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	Uncoded_feature (6)
coded_capture	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	coded_feature (6)
coded_feature	Wdeconv	coded_deconv (6)
Interpolate-0.5(concat (Uncoded_feature, coded_feature, coded_deconv))	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_in (32)
scale1_in	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_encode1 (64)
scale1_encode1	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_encode2 (128)
scale1_encode2	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale1_decode2 (64)
scale1_decode2 + scale1_encode1	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale1_decode1 (32)
scale1_decode1 + scale1_in	Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_out (32)

Fine-tuning Process. Our reconstruction network implements a two-step fine-tuning process to ensure that the reconstructed images are aligned with both the measured attributes of the DOE and the characteristics of real-world captures. First, we fine-tune the network on synthetic data, replacing the theoretical forward PSF simulation in the model with the actual measured PSF values, acquainting the network with the true characteristics of the fabricated DOE. The next phase involves refining the network output to achieve cross-modal consistency. In this step, we reintroduce the reconstructed highlights into the image formation model. The goal here is to adjust the overall intensity so that the intensity of the simulated coded capture aligns with that of the real coded capture. This adjustment prevents the network from overfitting to the training distribution and ensures the output is consistent with real-world imaging scenarios.

Table 4. HDR Reconstruction Network architecture description (part 2). Specifically,"conv-k(a)-s(b)-LRelu" represents a convolution layer with an $a \times a$ kernel window, using the stride b, followed by a Leaky Relu ($\alpha = 0.02$) activation function, "Res-k(a)-Relu" represents a ResNet Block with an $a \times a$ kernel and Relu activation function, and "Interpolate-(a)" represents interpolate the input scale a. We use "convT" to denote transposed convolution, Wdeconv to denote Wiener-deconvolution using the coded PSF, and "concat" to denote concatenation.

Input	Layer Type	Output (# Channels)
concat(Uncoded_feature, coded_feature, coded_deconv, Interpolate-2(scale1_out))	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_in (32)
scale2_in	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_encode1 (64)
scale2_encode1	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_encode2 (128)
scale2_encode2	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale2_decode2 (64)
scale2_decode2 + scale2_encode1	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale2_decode1 (32)
scale2_decode1 + scale2_in	Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_out (32)
concat(uncoded_caputre , scale2_out)	conv-k5-s2-Relu	output (3)

Ground Truth Acquisition Procedure. For acquiring ground truth high dynamic range (HDR) data in outdoor test cases, we mount the camera on a tripod to capture the scene at various exposure settings (0EV, -3EV, and -6EV), without using the DOE. To account for the imperfect light efficiency of the DOE, we align the 95th percentile intensity value of the uncoded capture with that of the 0EV ground truth capture. Subsequently, we scale the intensities of all ground truth captures to correspond with this calibration. This method provides us with a reliable and consistent set of ground truth data, essential for evaluating the system performance in real-world HDR imaging scenarios.

Additional Simulation Results. In addition to the results presented in the main manuscript, we present additional qualitative simulation results in Fig.5. For each scene showcased, the leftmost column features the output from our method, followed by the reconstructed scenes at different exposure levels: 0EV, -3EV, and -6EV. Additionally, to emphasize the method's capability in detail resolution, we provide zoomed-in views of the saturated areas. These results further illustrate the effectiveness of our proposed 2-in-1 camera in capturing the rich details in both the bright and dark regions of the scene.

Table 5. Hyperspectral Reconstruction Network architecture description (part 1). Specifically,"conv-k(a)-s(b)-LRelu" represents a convolution layer with an $a \times a$ kernel window, using the stride b, followed by a Leaky Relu ($\alpha = 0.02$) activation function, "Res-k(a)-Relu" represents a ResNet Block with an $a \times a$ kernel and Relu activation function, and "Interpolate-(a)" represents interpolate the input scale a. We use "convT" to denote transposed convolution, Wdeconv to denote Wiener-deconvolution using the coded PSF, and "concat" to denote concatenation.

Input	Layer Type	Output (# Channels)
Uncoded_capture	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	Uncoded_feature (6)
coded_capture	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	coded_feature (31)
coded_feature	Wdeconv	coded_deconv (31)
Interpolate-0.5(concat (Uncoded_feature, coded_feature, coded_deconv))	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_in (64)
scale1_in	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_encode1 (128)
scale1_encode1	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_encode2 (256)
scale1_encode2	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale1_decode2 (128)
scale1_decode2 + scale1_encode1	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale1_decode1 (64)
scale1_decode1 + scale1_in	Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale1_out (64)

Additional Experimental Results. In Fig.6, we showcase further experimental results achieved with our proposed 2-in-1 camera in various outdoor settings. These additional results are compared with ground truth captures derived from bracketed exposures, as well as a learned LDR-to-HDR baseline method. These supplementary scenes serve as additional experimental validation of our method's effectiveness across diverse real-world environmental conditions.

6 OPTICALLY CODED SNAPSHOT HYPERSPECTRAL IMAGING

PSF Calibration. Employing a similar setup as for the HDR application, we utilize a broadband fiber-coupled LED and a collimation lens as the light source. For PSF measurements across various wavelengths, we incorporate a linear variable VIS bandpass filter (Edmund 88-365) and a miniature spectrometer (Ocean Insight

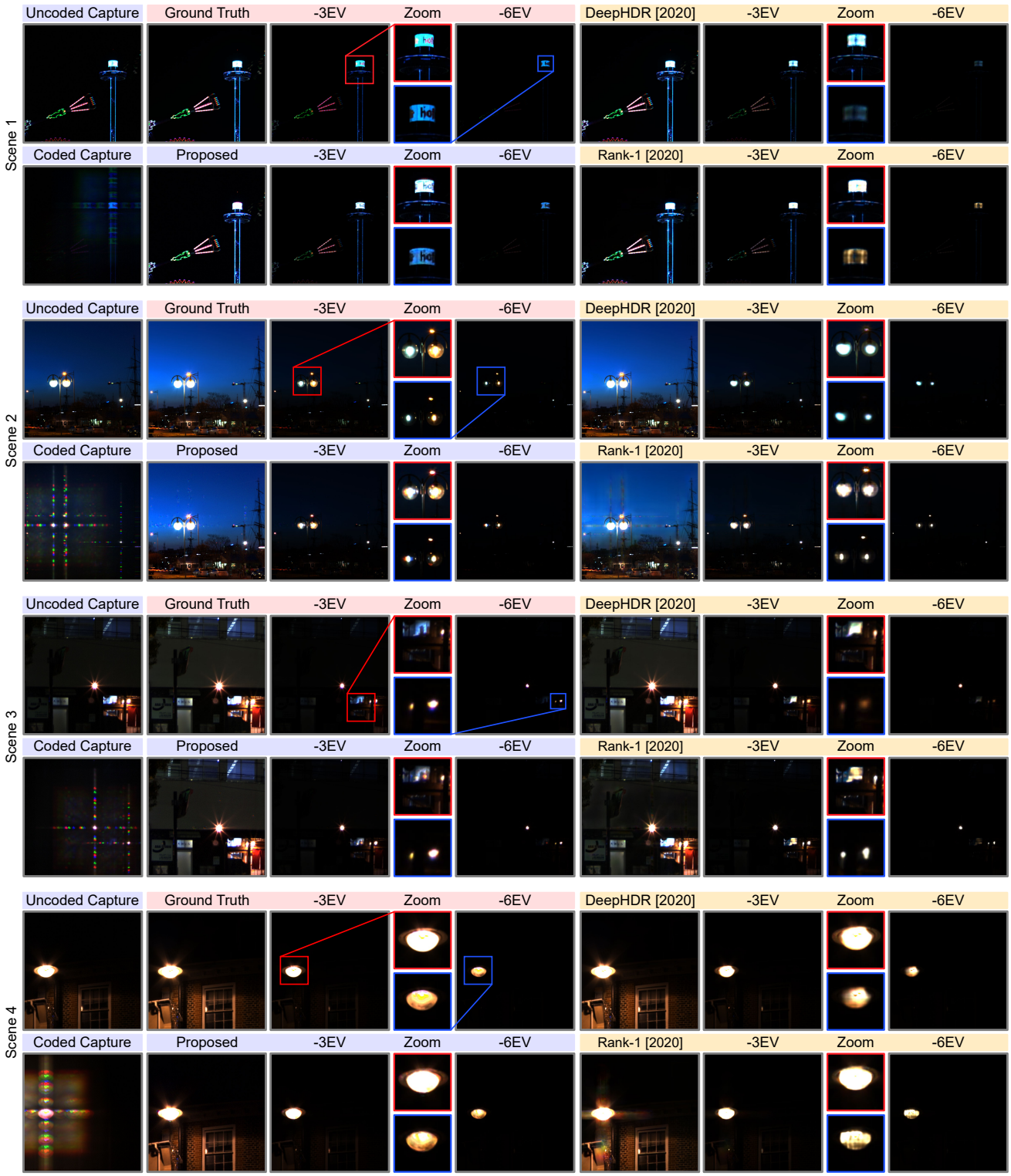

Fig. 5. **Additional Synthetic Results of Snapshot HDR Methods.** We assess the proposed method for snapshot HDR imaging in simulation by comparing the proposed method to the LDR-to-HDR method DeepHDR [Santos et al. 2020], and the DOE-based Rank-1 Optics approach [Sun et al. 2020]. DeepHDR, constrained 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.

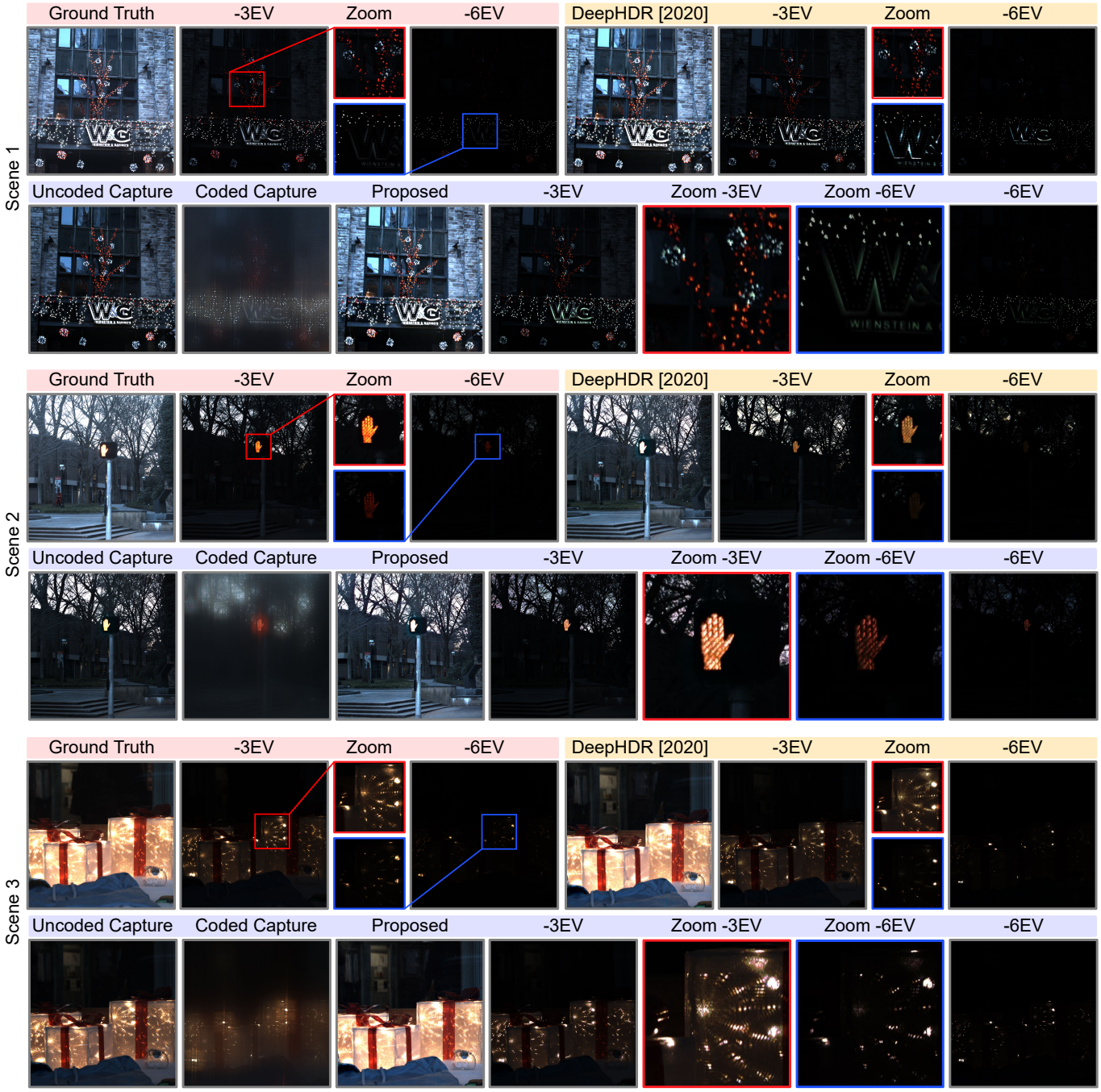

Fig. 6. **Additional Experimental Captures of Snapshot HDR Imaging.** We assess the proposed method experimentally for snapshot HDR imaging in outdoor settings, comparing our results with Ground Truth data obtained through bracketed exposures. The proposed method is able to recover fine detail of the 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.

Table 6. Hyperspectral Reconstruction Network architecture description (part 2). Specifically,"conv-k(a)-s(b)-LRelu" represents a convolution layer with an $a \times a$ kernel window, using the stride b, followed by a Leaky Relu ($\alpha = 0.02$) activation function, "Res-k(a)-Relu" represents a ResNet Block with an $a \times a$ kernel and Relu activation function, and "Interpolate-(a)" represents interpolate the input scale a. We use "convT" to denote transposed convolution, Wdeconv to denote Wiener-deconvolution using the coded PSF, and "concat" to denote concatenation.

Input	Layer Type	Output (# Channels)
concat(Uncoded_feature, coded_feature, coded_deconv, Interpolate-2(scale1_out))	conv-k5-s1-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_in (64)
scale2_in	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_encode1 (128)
scale2_encode1	conv-k5-s2-LRelu Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_encode2 (256)
scale2_encode2	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale2_decode2 (128)
scale2_decode2 + scale2_encode1	Res-k5-Relu Res-k5-Relu Res-k5-Relu convT-k3-s2-Relu	scale2_decode1 (64)
scale2_decode1 + scale2_in	Res-k5-Relu Res-k5-Relu Res-k5-Relu	scale2_out (64)
scale2_out	conv-k5-s2-Relu	output (31)

USB4000-VIS-NIR-ES), enabling the creation of a narrowband light source around wavelength of interest. Limited by the LED spectral coverage, we calibrate the PSF from 450nm to 700nm at 10nm intervals. For wavelengths ranging from 400nm to 440nm, which are not directly measured, we interpolate the PSFs from the measurements for visualization and finetuning purposes.

Additional Details on the Training Process. We use 278 training images and 32 unseen testing images. To accommodate different image sizes, we take random 512×512 crops of the images for both training and validation purposes. During training, we also apply left-right flips and random channel shuffles for additional data argumentation.

Reconstruction Network Architecture. Similar to the snapshot HDR application, we employ a DWDN-like network architecture that performs feature-based inverse filtering on the coded capture, followed by an encoder-decoder network to reconstruct hyperspectral information from the dual captures. See Tab. 5 and 6 for a full network specification.

Fine-tuning Process. The fine-tuning procedure for our reconstruction network follows a two-stage approach similar to that used in HDR applications. In the initial stage, we fine-tune the network using synthetic data, substituting the theoretical forward PSF simulation in the model with the measured PSF values. However, due to the spectral coverage of our LED, we calibrate the PSF within the 450nm to 700nm range, at intervals of 10nm. For the 400nm to 440nm spectrum, which is not directly measured, we extrapolate the PSFs from the 450nm measurement. This extrapolation maintains the relative intensity among the 0th, 1st, and 2nd order diffractions, adjusting only the diffraction positions based on simulated values, Following this, we focus on enhancing the network output for cross-modal alignment. Here, the reconstructed hyperspectral scene is reintegrated into the image formation model. We then employ a pixel-wise $\mathcal{L}_1$ loss to ensure that both the simulated coded and uncoded captures correspond accurately with the actual captures.

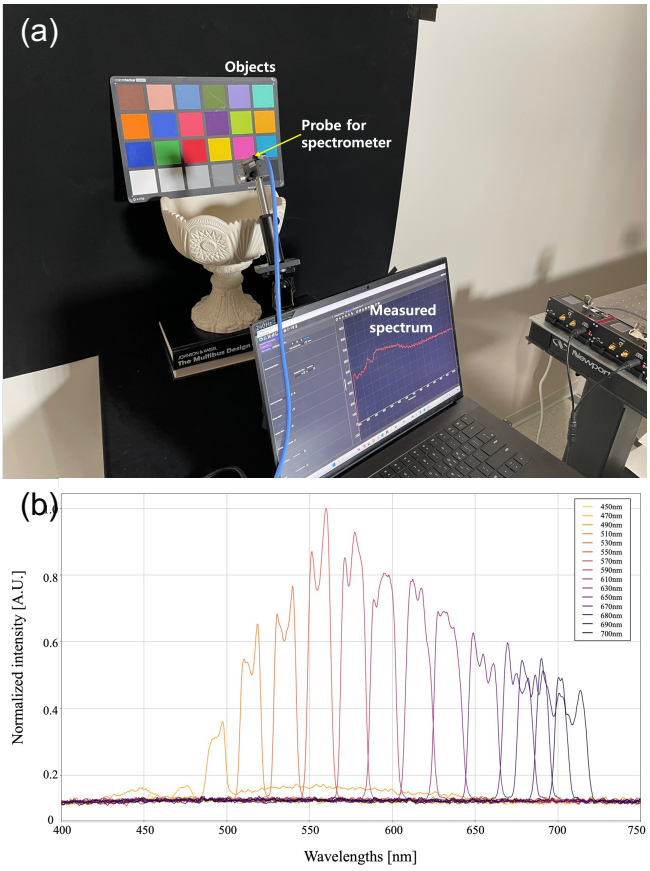

Fig. 7. (a) Ground Truth Acquisition Setup using a Spectrometer. (b) Spectral profile of the narrow-band light source used for PSF calibration.

Ground Truth Acquisition Procedure. To acquire accurate ground truth spectral intensity data, we employ a miniature spectrometer (Ocean Insight USB4000-VIS-NIR-ES), as shown in Fig.7(a). The spectral measurements are conducted under controlled lighting, which includes an overhead ceiling light and two-stage lights set to a color temperature of 5500K. We measure the spectral reflection at the center of each color and then normalize these measurements using

Table 7. Depth Reconstruction Network architecture description. Specifically,"conv-k(a)-s(b)-LRelu" represents a convolution layer with an $a \times a$ kernel window, using the stride b, followed by a Leaky Relu ($\alpha = 0.02$) activation function, "ResNet18" represents a ResNet18 backbone without the final decision layers. We use "convT" to denote transposed convolution, and "concat" to denote concatenation.

Input	Layer Type	Output (# Channels)
Uncoded_capture	ResNet18	Uncoded_feature (1024)
coded_capture	ResNet18	coded_feature (1024)
concat (Uncoded_feature, coded_feature))	convT-k2-s2-Relu conv-k3-s1-Relu conv-k3-s1	up5 (512)
up5	convT-k2-s2-Relu conv-k3-s1-Relu conv-k3-s1	up4 (256)
up4	convT-k2-s2-Relu conv-k3-s1-Relu conv-k3-s1	up3 (128)
up3	convT-k2-s2-Relu conv-k3-s1-Relu conv-k3-s1	up2 (64)
up2	convT-k2-s2-Relu conv-k3-s1-Relu conv-k3-s1	up1 (32)
up1	conv-k5-s1-Relu	output (1)

the data obtained from white and black references. For generating narrowband light sources targeting specific wavelengths for wavelength-dependent PSF calibration, as illustrated in Fig.7(b), we use a similar setup as for the HDR application complemented by a linear variable VIS bandpass filter (Edmund 88-365).

Additional Simulation Results. Complementing the main manuscript, Fig.8 contains further qualitative simulation results. For each scene, the leftmost column shows the sensor captures using our method, followed by reconstructions in both RGB and hyperspectral formats (alternate hyperspectral channels (410nm to 700nm at 20nm intervals). The RGB images are generated from hyperspectral reconstructions and sensor response curves. These additional results further validate the capability of our method in precisely reconstructing both spatial and spectral details with high fidelity.

Additional Experimental Results. Fig.9 reports additional experimental results obtained with our 2-in-1 camera for various indoor and outdoor environments. Given the limitations of our spectral measurement equipment in point-wise measurements, reference spectral curves were not acquired in these uncontrolled lighting conditions. Consequently, in these scenarios, we compare the results of our proposed method solely with those from a learned RGB-to-HS (hyperspectral) method. These experimental results further demonstrate the effectiveness of our approach in real-world settings under varying lighting conditions.

7 MONOCULAR DEPTH FROM CODED DEFOCUS

Additional Details on the Training Process. FlyingThings3D dataset contains 30K images, and is divided into 18K pairs for training, 4K pairs for validation and 8K pairs for testing. We take random 512×512 crops of the images for both training and validation purposes, and we also apply left-right flips as additional data augmentation during training. We set the target range in this dataset to the range from 1m to 5m.

Reconstruction Network Architecture. We utilize ResNet18 [He et al. 2016] as a feature extractor for both uncoded and coded captures, processing them independently before channeling them into a unified decoder. The ResNet18 components in our architecture begin with pre-trained weights, which are further refined with the rest of the network during the training phase. See Tab. 7 for a full network specification.

Fine-tuning Process. We first tine-tune the network using synthetic data, substituting the theoretical forward PSF simulation in the model with the measured PSF values. We notice model trained on FlyingThings3D dataset [Mayer et al. 2016] doesn't genalize well to the real captures so we additionally used Hypersim dataset [Roberts et al. 2021] during the fine-tuning process. Note Hypersim dataset was not used for the model used in synthetic evaluation. Additionally, to address the significant resolution disparity between the training scenes (512×512) and experimental captures (3072×3072), we apply a segmentation-based median filter to refine the depth output, where we apply Segment Anything [Kirillov et al. 2023] on the uncoded capture to obtain object boundaries, and perform masked medium filter to smooth out the depth output.

Ground Truth Acquisition Procedure. To obtain reference absolute depth information, we employ a solid-state LiDAR camera (Intel® RealSense™ LiDAR Camera L515). This camera is positioned adjacent to our proposed camera setup and we try to achieve a parallel alignment for consistency in data capture. The RealSense camera offers a depth output resolution of 1024×768. Given that the RealSense camera's field of view, at $70^\circ \times 55^\circ$, is substantially broader compared to the lens used in our setup, we manually crop the output to match the region corresponding to the scene captured by our proposed system.

Additional Simulation Results. In addition to the main manuscript, Fig. 10 presents further qualitative simulation results. For each scene, the leftmost two columns display the sensor captures using our method at double intensity. These are followed by depth reconstructions from the monocular depth estimation method MiDaS [Ranftl et al. 2022], and DOE-based depth from defocus method Deep DfD [Ikoma et al. 2021], alongside the ground truth data. The relative depth output from MiDaS is scaled to align with the known target depth range for consistent comparison. These additional results underscore our method's proficiency in accurately reconstructing absolute depth information in complex scenes.

Additional Experimental Results. Fig.11 reports additional experimental results obtained with our 2-in-1 camera for various indoor and outdoor environments. In the case of indoor scenes, we employ a solid-state LiDAR camera to gather absolute depth data from the scenes, serving as a reference, and in outdoor experiments, where the RealSense L515 sensor struggles to provide accurate depth, we

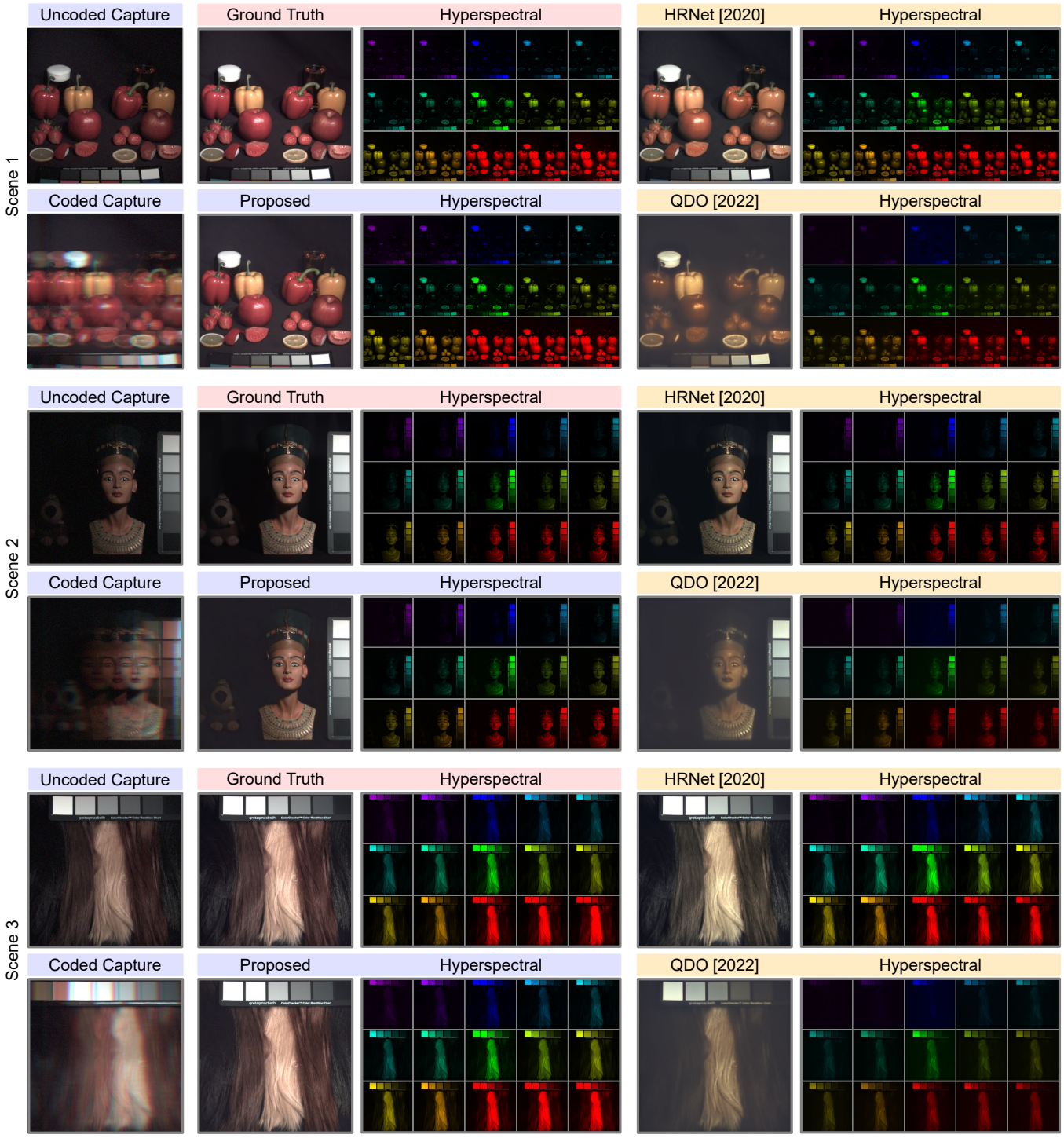

Fig. 8. **Additional Synthetic Results for Snapshot Hyperspectral Imaging.** We assess the proposed method for snapshot hyperspectral imaging with simulated ground truth spectral data (400nm to 700nm) and compare the RGB-to-Spectrum HRNet [Zhao et al. 2020], and DOE-based QDO systems [Li et al. 2022]. 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.

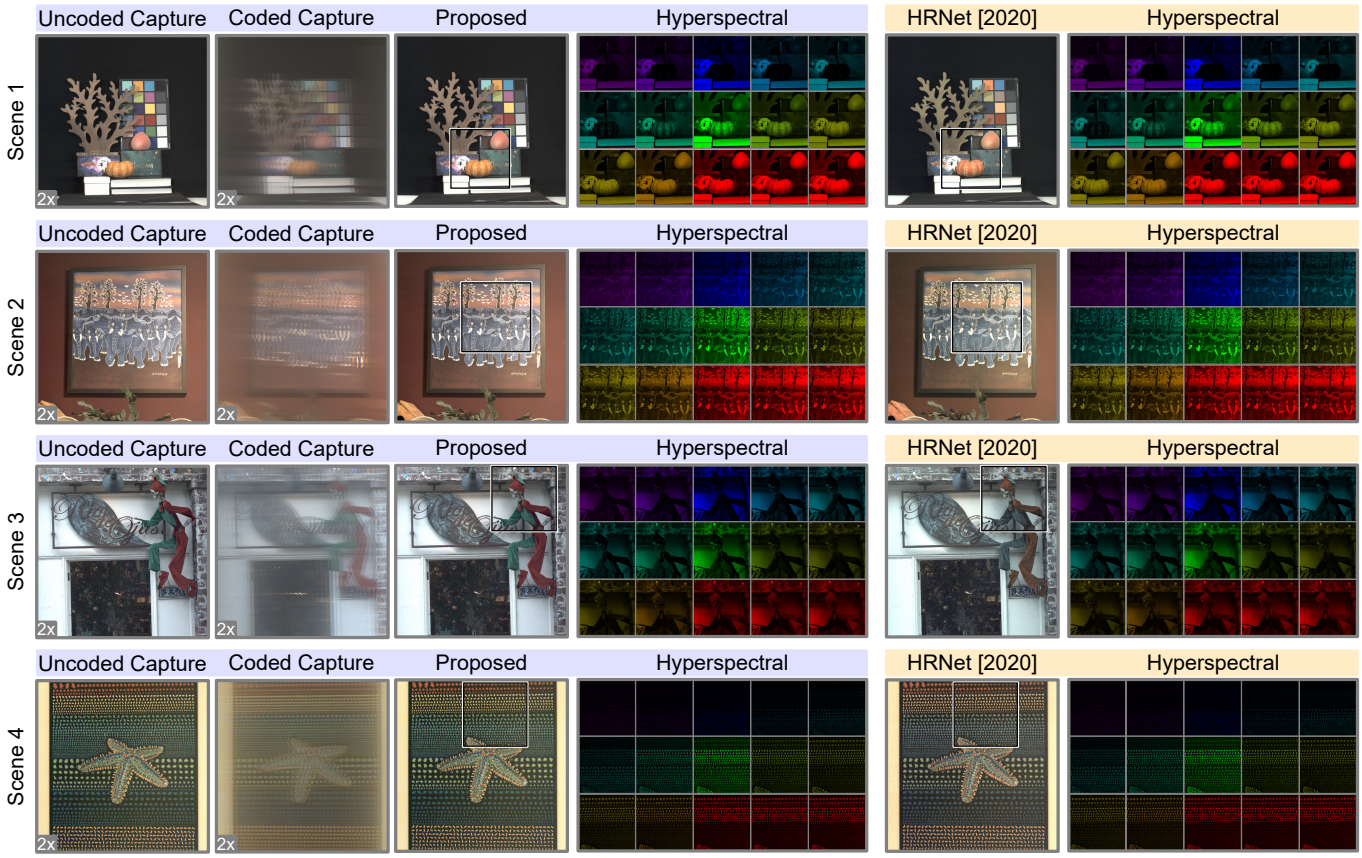

Fig. 9. **Additional Experimental Captures of Snapshot Hyperspectral Imaging.** We evaluate our method experimentally for snapshot hyperspectral imaging under varying lighting conditions. We compare our method to the learned RGB-to-HS technique, HRNet [2020], across different environments: outdoor (Scene 3), indoor with cool tone lighting (Scene 1), and warm lighting (Scenes 2 and 4). 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. While HRNet exhibits challenges in accurately reproducing colors at the spectrum boundaries, our proposed method demonstrates robust and consistent performance across all tested lighting scenarios.

limit our comparison to the MiDaS baseline. These experimental results further validate the proposed approach to reconstruct absolute scene depth in real-world settings.

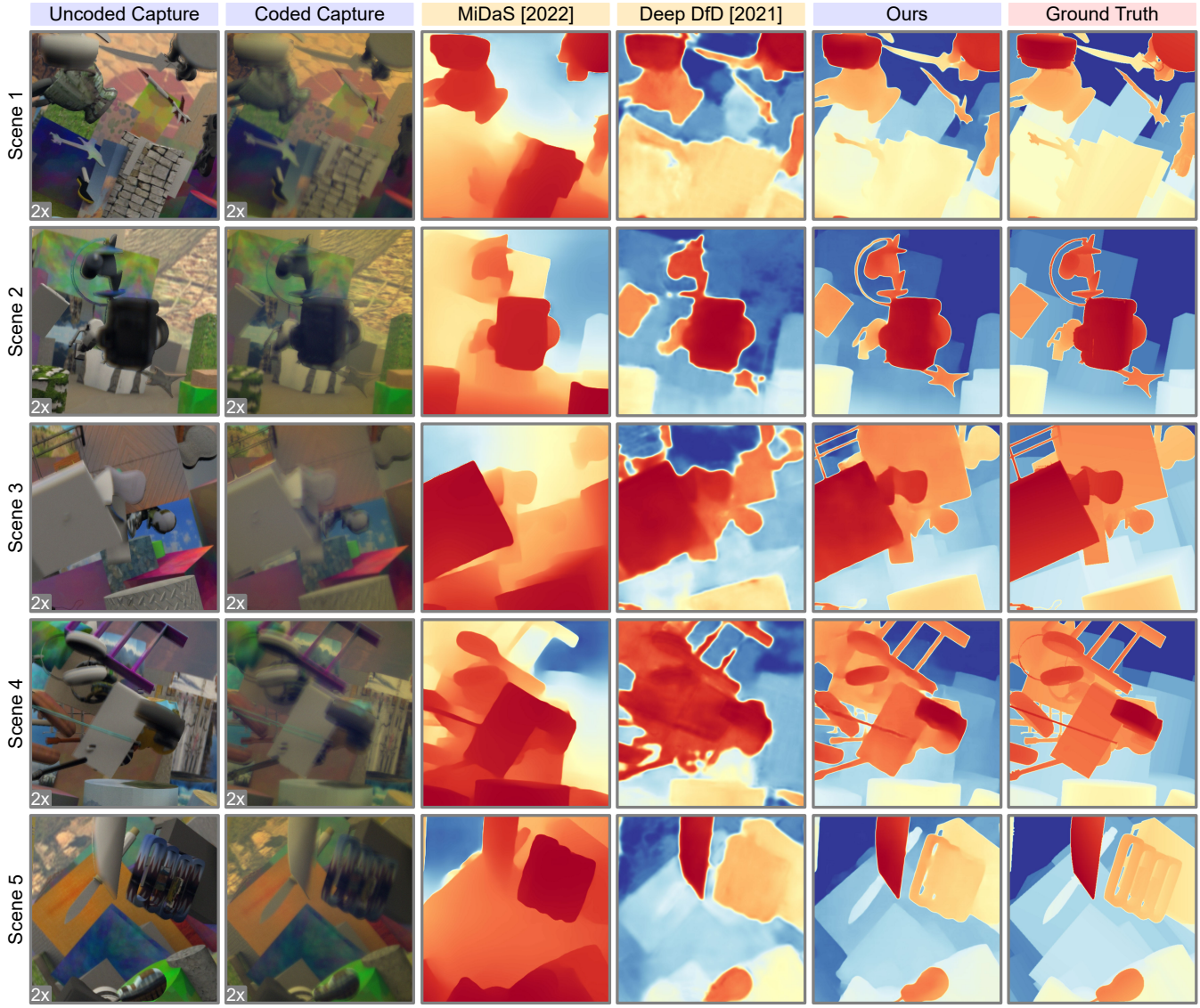

Fig. 10. **Additional Synthetic results of Monocular Depth Imaging.** We assess our approach for monocular depth estimation in simulation by comparing our method to the monocular depth estimation method MiDaS [Ranftl et al. 2022], and DOE-based depth from defocus method Deep DfD [Ikoma et al. 2021]. 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.

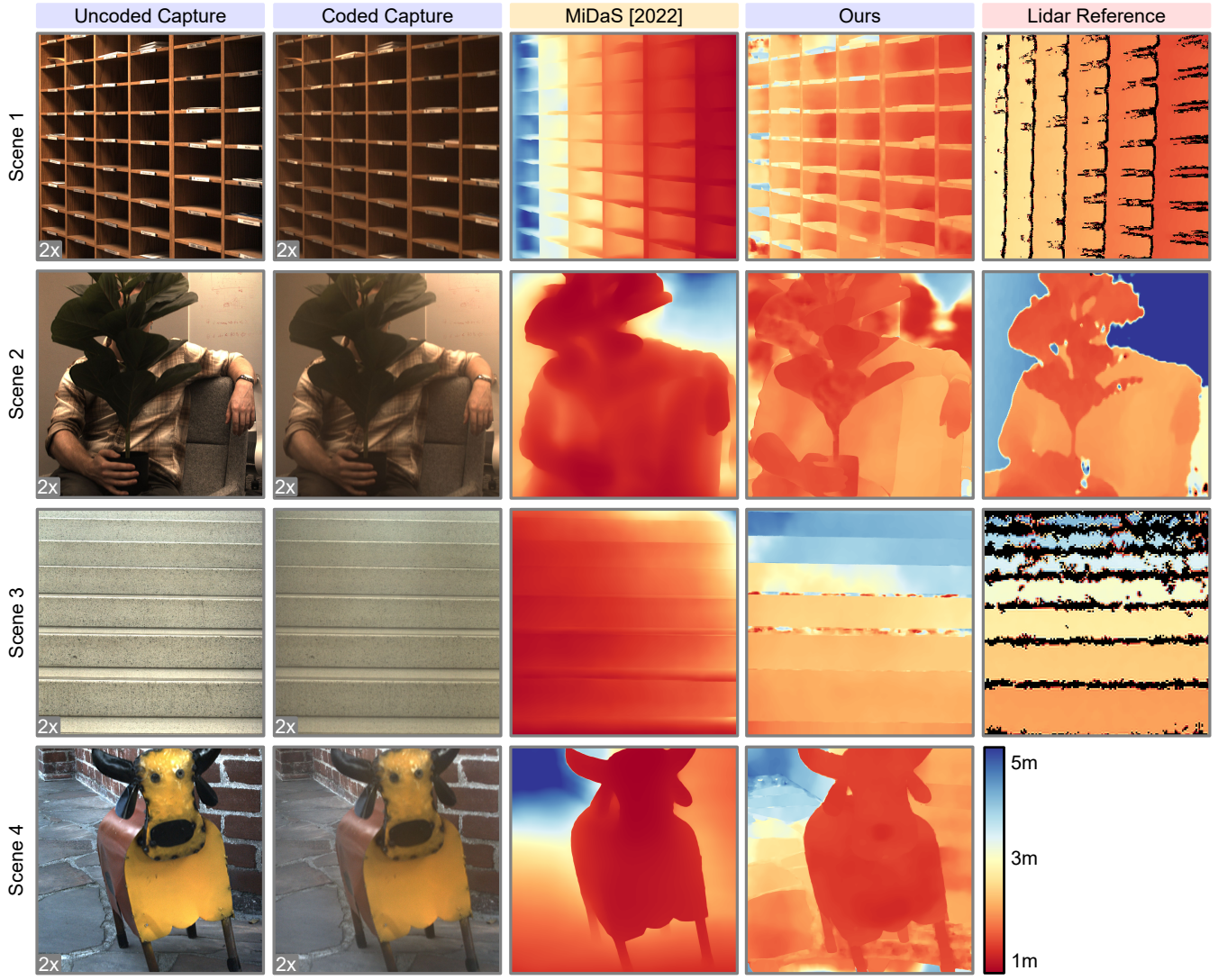

Fig. 11. **Additional Experimental Captures of Optically Coded Depth Imaging.** We evaluate the proposed method in both indoor (Scenes 1 to 3) and outdoor (Scene 4) environments, and compared it against the monocular depth method MiDaS [Ranftl et al. 2022] applied to the uncoded capture and rescaled to the target depth range. Areas where the RealSense camera was unable to provide measurements are indicated with a black mask. The depth reconstructions produced by our proposed method demonstrate a close alignment with the RealSense reference data. In contrast, MiDaS is limited to provide a plausible relative depth map and often inaccurately merges unconnected objects into a singular, continuous depth profile.

REFERENCES

Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE transactions on pattern analysis and machine intelligence* 33, 5 (2010), 898–916.

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.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment Anything. *arXiv:2304.02643* (2023).

Zeqiang Lai, Kaixuan Wei, Ying Fu, Philipp Härtel, and Felix Heide. 2023. ∇-Prox: Differentiable Proximal Algorithm Modeling for Large-Scale Optimization. *ACM Transactions on Graphics (TOG)* 42, 4, Article 105 (2023), 19 pages. https://doi.org/10.1145/3592144

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.

Rastislav Lukac. 2018. *Single-sensor imaging: methods and applications for digital cameras*. CRC Press.

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.

Masashi Miyata, Naru Nemoto, Kota Shikama, Fumihide Kobayashi, and Toshikazu Hashimoto. 2021. Full-color-sorting metalenses for high-sensitivity image sensors. *Optica* 8, 12 (2021), 1596–1604.

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).

Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. 2021. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In *ICCV*. https://arxiv.org/pdf/2011.02523.pdf

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).

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)*.

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.