Deep Optics for Single-shot High-dynamic-range Imaging

Christopher A. Metzler Hayato Ikoma Yifan Peng Gordon Wetzstein
Stanford University
{cmetzler, hikoma, evanpeng, gordon.wetzstein}@stanford.edu

Abstract

High-dynamic-range (HDR) imaging is crucial for many applications. Yet, acquiring HDR images with a single shot remains a challenging problem. Whereas modern deep learning approaches are successful at hallucinating plausible HDR content from a single low-dynamic-range (LDR) image, saturated scene details often cannot be faithfully recovered. Inspired by recent deep optical imaging approaches, we interpret this problem as jointly training an optical encoder and electronic decoder where the encoder is parameterized by the point spread function (PSF) of the lens, the bottleneck is the sensor with a limited dynamic range, and the decoder is a convolutional neural network (CNN). The lens surface is then jointly optimized with the CNN in a training phase; we fabricate this optimized optical element and attach it as a hardware add-on to a conventional camera during inference. In extensive simulations and with a physical prototype, we demonstrate that this end-to-end deep optical imaging approach to single-shot HDR imaging outperforms both purely CNN-based approaches and other PSF engineering approaches.

1. Introduction

High dynamic range (HDR) imaging is one of the most widely used computational photography techniques with a plethora of applications, for example in image-based lighting [15], HDR display [59], and image processing [55, 5]. However, the dynamic range of a camera sensor is fundamentally limited by the full well capacity of its pixels. When the number of generated photoelectrons exceed the full well capacity, which is typically the case when imaging scenes with a high contrast, intensity information is irreversibly lost due to saturation. Ever shrinking pixel sizes, for example in mobile devices, exacerbate this problem because the full well capacity is proportional to the pixel size.

Several different strategies have been developed to overcome the limited dynamic range of available sensors. One class of techniques captures multiple low-dynamic-range (LDR) sensor images with fixed [26] or varying [40, 16, 46] exposure settings. Unfortunately, motion can be problematic when capturing dynamic scenes with this approach. Another class of techniques uses multiple optically aligned sensors [45, 66] to capture these exposures simultaneously, but calibration, cost, and device form factor can be challenging with such special-purpose cameras. Single-shot approaches are an attractive solution, but traditionally required custom exposure patterns to be multiplexed on the sensor [49, 24, 61]. More recently, single-shot HDR imaging approaches were proposed that hallucinate an HDR image from a single saturated LDR image (HDR-CNN, e.g. [18]). While successful in many cases, saturated scenes details often cannot be faithfully recovered via hallucination.

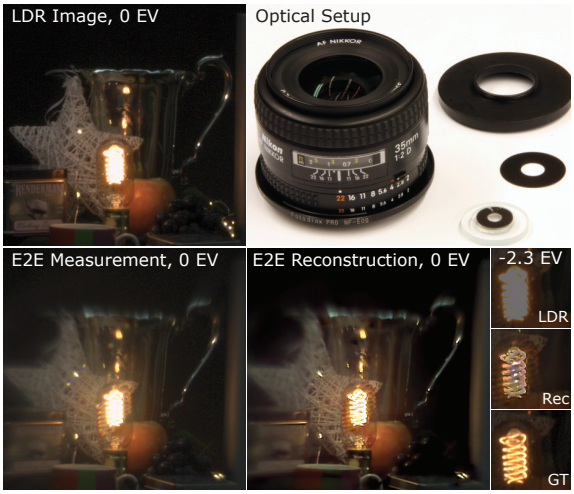

Figure 1: Conventional sensors are limited in their ability to capture high-dynamic-range (HDR) scenes. Details in brighter parts of the image, such as the light bulb, are saturated in a low-dynamic-range (LDR) photograph (top left). Our end-to-end (E2E) approach jointly optimizes a diffractive optical element (top right) and a neural network to enable single-shot HDR imaging. This deep optical imaging system records a single sensor image (bottom left) that contains optically encoded HDR information, which helps the network recover an HDR image (bottom right).

In this work, rather than hallucinating missing pixel values, we aim to preserve information about the saturated pixel values by encoding information about the brightest pixel values into nearby pixels via an optical filter with an optimized point spread function (PSF). Unlike previous attempts to encode HDR pixel information with an optical filter [58], we turn to machine learning to automatically design both the optical element and the reconstruction algorithm end-to-end, so as to maximize the information passed from the HDR scene to the low-dynamic-range (LDR) measurements. In essence, we construct an autoencoder where the encoding is performed optically and the decoding is performed computationally. Both encoder and decoder are trained in an end-to-end fashion, with the optimized optical element being fabricated and remaining fixed during inference.

In optimizing the encoder and decoder, our system must solve three challenging inverse problems at once. (1) Mapping a PSF to a manufacturable optical filter is implicitly a phase retrieval problem. (2) Using optically encoded information to fill in saturated regions is an inpainting problem. (3) Removing said optically encoded information from non-saturated regions is a deconvolution problem. Our work is the first to explore and successfully address this unique and challenging combination of inverse problems.

Using extensive simulations, we demonstrate that deep optics generally achieves better results than alternative single-shot HDR imaging approaches. This is intuitive, because compared with HDR-CNN approaches, our optimized PSF has more degrees of freedom to encode scene information in the sensor image, and compared with other optical encoding techniques, ours uses an optical element that is jointly optimized with the reconstruction algorithm, rather than heuristically chosen. We demonstrate the proposed camera system with a proof-of-concept prototype by fabricating a diffractive optical element that can simply be attached as a hardware add-on to a conventional camera lens.

Specifically, we make the following contributions

- We introduce an optical encoder and CNN-based decoder pipeline for single-shot HDR imaging.
- We present a new single-shot "multiplexing" approach to HDR imaging; the learned, grating-like diffractive optical element (DOE) creates shifted and scaled copies of the image which are used to reconstruct the brightest regions of the scene.
- We analyze the proposed system and demonstrate that it outperforms existing single-shot HDR methods.
- We fabricate the optimized diffractive optical element and validate the proposed system experimentally.

2. Related Work

HDR Imaging aims at overcoming the limited dynamic range of conventional image sensors using computational photography techniques. Many approaches rely on capturing several LDR images with different exposures and fusing them into a single HDR image [40, 16, 46, 26, 25]. Although motion between the LDR images can be a problem, many proposals have been introduced to deal with this problem [36, 20, 22, 30, 32], allowing even HDR video to be recorded from temporally varying exposures [34, 60, 33].

Although many of these multi-shot approaches are successful in some scenarios, they can fail for fast motion and they can also be computationally expensive. To mitigate these limitations, multiple sensors can be optically combined to capture these exposures simultaneously [1, 45, 66]. However, this is costly and bulky and the system calibration can be challenging.

Motivated by these shortcomings, several approaches to single-shot HDR imaging have been proposed. Reverse tone mapping approaches aim at solving an ill-posed problem [6, 47, 56]. Using computational photography approaches, this problem can be made "less ill-posed", for example using spatially varying pixel exposures via neutral density filter arrays or spatially varying ISO settings [49, 70, 24, 61], using an optically coded point spread function (PSF) [58], or using a special modulo camera [73]. Most recently, convolutional neural networks (CNNs) have been employed to hallucinate realistic HDR images from a single LDR image [18, 19, 37].

Our work also uses a CNN to recover an HDR image from a single LDR image, but rather than hallucinating it, we use an optimized PSF to encode as much of the HDR image content as possible in the sensor image. While Rouf et al. [58] also used an optical filter to aim for the same goal, theirs was heuristically chosen and is limited in its ability to recover high-quality HDR images. We train a CNN end-to-end with an optimizable optical filter that achieves far superior image quality. We fabricate this optimized lens using grayscale lithography and demonstrate its ability to capture single-shot HDR images with a prototype camera system.

In concurrent work, Alghamdi et al. [2] explored a CNN-based image reconstruction approach from spatially coded LDR measurements, but they did not use an end-to-end approach to optimizing the optical coding strategy. Their algorithm requires a custom neutral density filter array on the sensors. Also in concurrent work, Martel et al. [43] recently proposed an end-to-end learning strategy for the spatially varying pixel exposures of a programmable "neural" sensor for

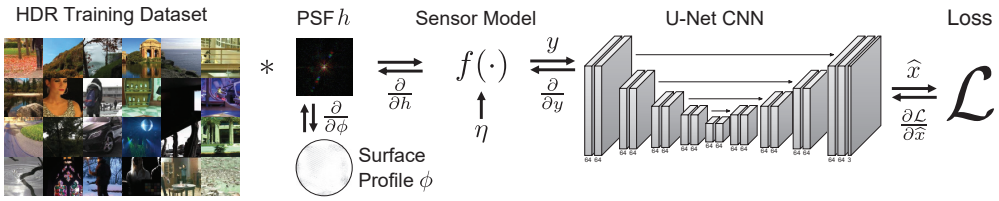

Figure 2: Illustration of the proposed end-to-end optimization framework. HDR images of a training set are convolved with the PSF created by a lens surface profile $\boldsymbol{\phi}$. These simulated measurements are clipped by a function $f(\cdot)$ to emulate sensor saturation and noise $\boldsymbol{\eta}$ is added. The resulting RGB image $\mathbf{y}$ is processed by a convolutional neural network (CNN) and its output compared with the ground truth HDR image using the loss function $\mathcal{L}$. In the learning stage, this loss is back-propagated into the CNN parameters and also into the height values $\boldsymbol{\phi}$ of the lens. During inference, a captured LDR image blurred by the optical PSF is fed directly into the pre-trained CNN to reconstruct an HDR image.

HDR imaging.

Computational Optics There is a long history of co-designing optics and image processing. In computational photography, research in this topic has focused on various applications such as extended-depth-of-field imaging [17, 14, 13], motion or defocus deblurring [54, 75, 74], depth estimation [38, 39], multispectral imaging [69, 11, 31], light field imaging [51, 68, 44], achromatic imaging [53], gigapixel imaging [12, 7], and lensless imaging [3, 4]. In computational microscopy, similar concepts are known as point spread function (PSF) engineering and have been used for optimizing the capabilities of single-molecule localization microscopy [52, 62]. In all of these examples, some optimality criterion is defined for the PSF, which is then optimized to work well for a particular choice of algorithm. This can be interpreted as co-design, whereas our approach builds on the emerging concept of end-to-end design, where optics and image processing are optimized jointly in an end-to-end fashion.

Deep Optics The idea of end-to-end optimization of optics and image processing has recently gained much attention. This concept has been demonstrated to provide significant benefits for applications in color imaging and demosaicing [8], extended depth of field and superresolution imaging [64], monocular depth imaging [27, 23, 71, 10], image classification [9], time-of-flight imaging [42, 65], computational microscopy [29, 28, 50, 35], and focusing light through scattering media [67]. To the best of our knowledge, this work is the first to explore deep optics for single-shot HDR imaging.

3. End-to-end HDR Imaging

A camera maps a scene $\mathbf{x}$ to a two-dimensional sensor image $\mathbf{y}$ as

$$\mathbf{y} = f\left(\mathbf{h} * \mathbf{x} + \boldsymbol{\eta}\right), \tag{1}$$

where $\mathbf{x} \in \mathbb{R}_{+}^{n_x \times n_y}$ is a discrete image with $n_x \times n_y$ pixels, each containing values that are proportional to the irradiance incident on the sensor. The irradiance is scaled such that the non-saturated values map to the range $\mathbf{x}_i \in [0, 1]$. Furthermore, $\boldsymbol{\eta}$ models signal-independent read noise, $\mathbf{h}$ is the optical point spread function (PSF) created by the camera lens, $*$ denotes the 2-D convolution operator, and $f(\cdot)$ is the camera's response function. This image formation model assumes that the PSF is shift-invariant, but the model could be generalized to describe PSFs that vary laterally or with depth.

We assume that the camera has a linear camera response function, which is typically the case when working with raw sensor data:

$$f\left(\mathbf{x}_i\right) = \begin{cases} 0, & \text{if } \mathbf{x}_i < 0, \\ \mathbf{x}_i, & \text{if } 0 \leq \mathbf{x}_i \leq 1, \\ 1, & \text{if } \mathbf{x}_i > 1. \end{cases} \tag{2}$$

Nonlinear camera response functions can be calibrated and inverted so as to mimic a linear response function [48]. We ignore the effects of quantization.

Our goal in this work is to jointly optimize the PSF $\mathbf{h}$ and a reconstruction algorithm $G : \mathbf{y} \mapsto \hat{\mathbf{x}}$ so as to recover $\mathbf{x}$ from $\mathbf{y}$ when $\|\mathbf{x}\|_\infty \gg 1$. To this end, we turn to differentiable optical systems and algorithms, which we describe in the following.

3.1. Modeling the Optical Point Spread Function

As shown in Figure 1, our optical system is a conventional single lens reflex camera lens with a custom diffractive optical element (DOE) add-on. Similar to a photographic filter, the DOE is mounted directly on the lens. To model the light transport from a scene, through these optical elements, to the sensor, we build on a differentiable Fourier optics model [21], an approach closely related to recent work on end-to-end camera designs [64, 9, 71, 10]. Specifically, our aim is to find the microscopic surface profile $\boldsymbol{\phi}$ of the DOE that creates a PSF $\mathbf{h}$ which is optimally suited for the HDR image reconstruction algorithm.

Assuming that the scene is at optical infinity, the complex-valued wave field of a point located on the

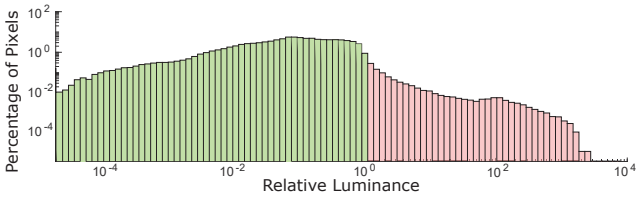

Figure 3: Our training dataset consisted of HDR images scaled such that between 1 and 2% of their pixels' values were saturated and clipped (red).

optical axis becomes a plane wave immediately before the DOE, i.e. $\mathbf{u}_{in} = \exp(ikz)$, where $k = \frac{2\pi}{\lambda}$ is the wave number and λ is the wavelength. The phase of this wave is affected by the DOE in a spatially varying manner by a complex-valued phase delay $\mathbf{t}_\phi$, which is directly calculated from the surface profile $\boldsymbol{\phi}$ as

$$\mathbf{t}_\phi(u,v,\lambda) = \mathbf{A}_\phi(u,v) \cdot \exp\left(ik\left(n(\lambda)-1\right)\boldsymbol{\phi}(u,v)\right). \quad (3)$$

Here, u, v are the lateral coordinates on the DOE surface, $n(\lambda)$ is the wavelength-dependent refractive index of the material that the DOE is made of and $A_{\boldsymbol{\phi}}(u,v)$ is a binary circular mask with diameter D_ϕ that models the aperture of the DOE as

$$\mathbf{A}_{\boldsymbol{\phi}}(u,v) = \begin{cases} 1, & \text{if } u^2+v^2 \leq \left(\frac{D_\phi}{2}\right)^2, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The wave field continues to propagate by some distance d_ϕ to the camera lens with focal length g. This lens induces the following phase delay:

$$\mathbf{t}_l(u',v') = \mathbf{A}_l(u',v') \cdot \exp\left(-i\frac{k}{2g}\left(u'^2+v'^2\right)\right). \quad (5)$$

Although the physical compound lens contains many optical elements that correct optical aberrations, a simplified thin lens model (Eq. 5) adequately describes the mathematical behavior of the lens.

Finally, the wave field propagates by a distance d_s to the sensor, where its intensity $\mathbf{h} = |\mathbf{u}_{sensor}|^2$ is recorded. Putting all of this together results in

$$\mathbf{h}_{\boldsymbol{\phi}}(x,y) = \left|P_{d_s}\left\{\mathbf{t}_l \cdot P_{d_\phi}\left\{\mathbf{t}_\phi \cdot \exp(ikz)\right\}\right\}\right|^2, \quad (6)$$

where $P_d\{u\}$ models free-space propagation of a wave field u by a distance d.

3.2. CNN-based Image Reconstruction

To recover $\mathbf{x}$ from $\mathbf{y}$ we use a convolutional neural network based on the well-known U-Net architecture [57]. Specifically, our U-Net uses skip connections and has 5 scales with 4 consecutive downsampling operations (maxpool) and 4 consecutive upsampling operations (transposed convolutions initialized with bilinear filter weights). At each scale of the U-Net, we include one additional convolutional layer; each convolutional layer is followed by a rectified linear unit (ReLU). BatchNorm layers are used after each upsampling layer and after the final convolutional layer. This architecture is inspired by Eilertsen's work [18] but slightly leaner, which resulted in faster convergence of the lens' surface profile. As illustrated in Figure 2, each network layer has 64 feature maps.

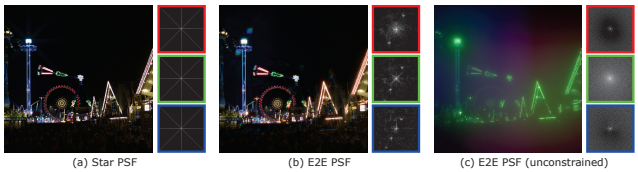

Figure 4: Simulated sensor images for an example from our evaluation set and point spread functions (PSFs) of several different optical coding approaches: (a) Rouf et al.'s star-shaped PSF [58], (b) our end-to-end optimized PSF with physically realizable constraints, and (c) our end-to-end optimized PSF without constraints. All three color channels of the PSFs are shown separately in the log domain. Whereas the star PSF continuously blurs the image along several radial streaks (a), the deep optics approach creates a PSF with several distinct peaks, which result in image content being copied at chromatically varying distances and at different intensity scales (b). The unconstrained PSF blurs the green color channel while focusing the red and blue channels.

3.3. End-to-end Training Details

We jointly optimize the PSF and the CNN via the end-to-end (E2E) framework illustrated in Figure 2. In particular, using thousands of HDR training images, we simulate (1) passing an HDR image through our optical system, (2) capturing a noisy and saturated LDR image with a sensor, and (3) reconstructing the HDR images with the CNN. We compute the corresponding loss and use Tensorflow's autodifferentiation capabilities to back-propagate the error and update the parameters θ of the CNN and the DOE height $\boldsymbol{\phi}$.

Loss Function Following Kalantari and Ramamoorthi [32], we originally experimented with minimizing the mean-squared-error (MSE) between the tonemapped reconstruction and ground-truth HDR images. While successful in training the CNN, this outlier-sensitive loss function caused the network to focus its efforts on the most overexposed and challenging images in the training data. Using this approach, a typical Dirac delta-type was found as a locally optimal solution for the PSF.

To encourage the development of more powerful PSFs, we instead minimize the sum, over all the

batches, of per-batch, γ-corrected, ℓ_2-loss:

$$\mathcal{L}_{\text{Data}} = \sum_{\text{B}\subset\text{Batches}} \left\| (\mathbf{x}_\text{B} + \epsilon)^\gamma - (\hat{\mathbf{x}}_\text{B} + \epsilon)^\gamma \right\|_2, \quad (7)$$

where $\mathbf{x}_\text{B}$ denotes a vector consisting of all the images in batch B, $\hat{\mathbf{x}}_\text{B}$ denotes a vector consisting of all the reconstructions of the images in batch B, $\gamma = 1/2$, and ϵ is a small constant that avoids the non-differentiability around 0. That is, we effectively minimize the root mean-squared-error (RMSE) over batches, as opposed to the typical MSE loss.

In the context of regression, sums of ℓ_2-norms over groups encourage group-sparse solutions [63]. In our context, sums of ℓ_2-norms make the network more robust to outliers, i.e. it allows the network to fail for the most challenging reconstructions so long as the ℓ_2-norm of the error is small for most batches.

Incorporating Fabrication Constraints To ensure that the optimized height map can be manufactured, we clip its values to the maximum range during training and add an additional smoothness term on $\boldsymbol{\phi}$ to prevent the resulting surface profile to include many discontinuities. Specifically, we add the loss

$$\mathcal{L}_{\text{Reg}} = \nu \|\mathbf{D} * \boldsymbol{\phi}\|_2^2, \quad (8)$$

to the overall loss function, where $\mathbf{D}$ is a Laplacian filter and $\nu = 10^9$ is a weighting parameter.

Datasets Following Eilertsen et al. [18], we use training and validation datasets consisting of 2837 HDR images drawn from a combination of videos and images from a number of sources. We performed data augmentation by cropping, rescaling, and adjusting hue and saturation. The final training set contains just under 60,000 different HDR images with a resolution of 320×320 pixels. Our test set consists of 223 HDR images, of which 83 are still images and the rest frames drawn from every 10th frame of four separate video sequences (which were not used for training). To generate LDR/HDR image pairs, we simulated capturing LDR frames where we set the exposure such that between 1 and 2% of the pixels were saturated. A histogram of the pixel values in our training data is shown in Figure 3.

Miscellaneous Training Details We trained the end-to-end model using the Adam optimizer with a minibatch size of 8. We applied an exponential learning rate decay with an initial rate of 0.0001. We trained the network for 100 epochs, which took about 3 days on a Pascal Titan X graphics processing unit. Source code, trained models, and our captured data is available at `https://github.com/computational-imaging/DeepOpticsHDR`.

	HDR-VDP	PSNR-L	PSNR-γ
LDR	51.4	39.8	38.9
HDR-CNN [18]	58.6	42.1	42.9
U-Net	56.4	41.8	42.3
Star PSF+U-Net[58]	56.7	42.1	42.3
E2E PSF+U-Net	60.6	45.6	**44.3**
E2E PSF+U-Net (unconstrained)	**67.1**	**46.8**	40.7

Table 1: Quantitative evaluation for the test set. Several single-shot HDR imaging approaches are compared using a perceptual difference computed by HDR-VDP-2 and peak signal-to-noise ratio (PSNR) computed in the linear domain (L) and in the γ-corrected domain.

4. Analysis and Evaluation

In this section, we evaluate the proposed method in simulation and show comparisons to various other single-shot HDR imaging approaches.

Figure 4 shows simulated sensor images and PSFs for several options of optically coding the sensor image before reconstruction. The PSF that was optimized with the method proposed in the previous section is shown in (b) and also in Figure 6. This PSF contains several peaks, each creating a shifted and scaled copy of the sensor image superimposed on itself. In this way, the PSF serves to multiplex together different exposures of the image. The copies for individual color channels appear at slightly different location, which leads to visible chromatic aberrations in the sensor image; such chromatic aberrations are inevitable when using a single DOE design. We also optimize a PSF using the proposed optimization method but without parameterizing it by a physically realizable optical element (c). Therefore, the color channels of this PSF are completely independent of one another and individual pixel values can be arbitrary as long as the are in the range $[0, 1]$ and sum to 1 for each channel. We also show the star-shaped PSF proposed by Rouf [58].

We compare several reconstruction approaches in Figure 5. These include the conventional LDR image, Eilertsen et al.'s CNN applied to this LDR image (HDR-CNN), the proposed smaller U-Net applied to this LDR image, the U-Net applied to an image captured with the star PSF, and our end-to-end deep optics approach with physically realizable constraints. We used a U-net rather than the algorithm from [58] with the Star filter as CNNs dramatically outperform conventional algorithms at deconvolution [72]. For fair comparison, the U-Nets in each example are trained for the respective PSF. We show regions of interest in the insets along with visible differences predicted by HDR-VDP-2 [41]. In all cases, the proposed deep optics approach achieves the best results.

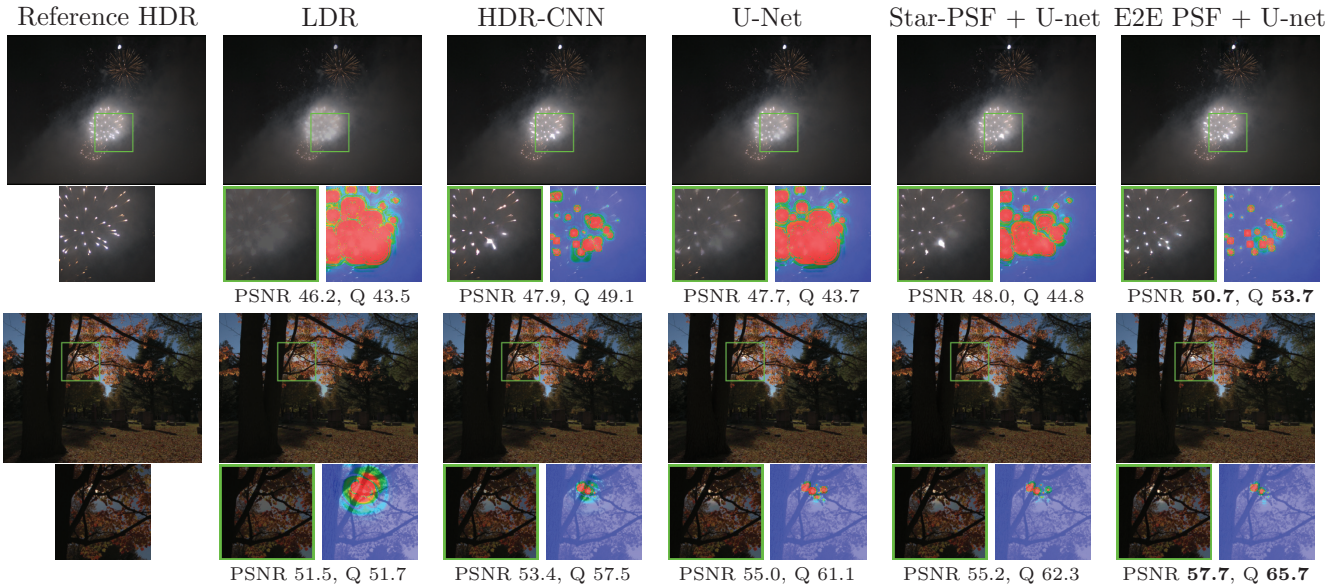

Figure 5: Comparison of single-shot HDR imaging approaches. In all examples, images are displayed at -1 stop and regions of interest at -3 stops. In the columns, we show ground truth HDR image, a corresponding LDR image, CNN-based reconstruction applied to the LDR image (HDR-CNN) [18], a slightly simpler U-Net applied to the LDR image, the star PSF [58] with a U-Net reconstruction, and our end-to-end approach (E2E). Color-coded insets show probabilities of perceiving the difference between reconstructions and ground truth HDR images, as computed by the HDR-VDP-2 visible differences predictor. In all cases, the E2E approach qualitatively and quantitatively (evaluated with peak signal-to-noise ratio (PSNR) and HDR-VDP Q value) outperforms other approaches.

Finally, we show a quantitative comparison of all these methods in Table 1. Here, we report the average perceptual difference as computed by HDR-VDP-2 as well as peak signal-to-noise ratio (PSNR) over the entire test set in the linear and γ-corrected domains. We observed that the end-to-end (E2E) approaches achieve the best image quality. The unconstrained E2E approach is usually better than the physically realizable version because it has more degrees of freedom. However, the PSNR-γ of the unconstrained PSF is slightly lower than that of the realizable approach. This is likely due to the fact that the network was trained using an outlier-robust loss, to which it over-fit. Recall, optimizing PSNR directly lead to sub-optimal convergence of the PSF (see Sec. 3.3).

5. Fabrication and Implementation

Lens Fabrication Once a phase profile is optimized, we fabricate the corresponding diffractive optical element (DOE) using polydimethyl-siloxane (PDMS) through replica molding. Figure 6 shows the optimized height profile (left) along with a 3D rendering of profilometer measurements of the fabricated DOE (center), which has a diameter of 5 mm. Qualitatively, the shapes are similar. We also show the simulated (top right) and the captured PSFs (bottom right). These PSFs match well and both create a Dirac peak in the center and lower-amplitude satellite peaks at slightly different locations for the three color channels. This PSF is created by the lens surface profile that resembles a grating-like structure. The captured PSF is slightly blurrier than the simulation and there is additional glare, both likely due to slight fabrication errors and interreflections between the lens elements.

System Integration We mount the fabricated DOE as an add-on to a conventional single lens reflex camera (Canon Rebel T5) equipped with a standard compound lens (Nikon Nikkor 35 mm). The DOE is fixed in a Thorlabs lens tube with rotating optic adjustment (SM1M05), which is coupled to the SLR lens via an optomechanical adapter (Thorlabs SM1A2). Figure 1 (right) shows the DOE and SLR lens. In this setup, the DOE is physically mounted at a slight distance to the compound camera lens. The exact distance between these optical elements is unknown because we model the primary camera lens as having a single refractive surface. While this is the easiest approach, a more detailed optical model of the compound lens may be desirable, although to model it appropriately, proprietary information from the lens manufacturer would have to be known. Moreover, the lack of anti-reflection coatings on the DOE may add interreflections and glare, and likely contributes to a slight mismatch between

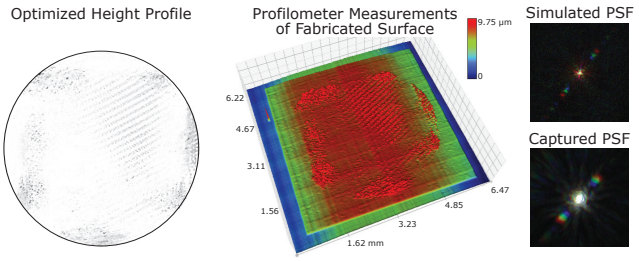

Figure 6: Optimized height profile of the diffractive optical element (DOE, left) along with profilometer measurements of the fabricated DOE. The DOE structure partially resembles that of a grating, which creates multiple peaks in the point spread function (PSF, right). Intuitively, this PSF creates three shifted and scaled copies of the input image. Although the measured PSF is slightly blurrier than the simulated PSF, likely due to imperfections in the fabrication process and approximations of our image formation model, their general shapes are comparable.

simulated and captured PSFs.

To calibrate our camera system, we capture several different exposures of a white light source behind a 75 µm-sized pinhole. These photographs are merged into a single HDR image [16]. This captured point spread function of our optical system is used to refine the pre-trained CNN by optimizing its parameters for the fixed captured PSF, as described in Section 3.3. Refining the CNN with a fixed PSF is significantly faster than training the end-to-end system from scratch and only takes a few hours.

6. Experimental Results

Using the prototype camera described in the previous section, we captured several HDR example scenes (see Fig. 7). These include two scenes recorded in a laboratory setting (top and center row) and one outdoor scene captured at night (bottom row). In Figure 7, captured measurements along with reconstructions computed by our CNN as well as reference LDR and HDR images and the result of the HDR-CNN [18] are shown. The ground truth HDR scene was computed by merging multiple images with different exposures [16]. In all of these examples, the captured LDR images include saturated areas that contain details, which are lost in the measurements, such as the filament of a light bulb (top row) or the structure of a light source on a wall (bottom row). We show these images as well as magnified closeups (right column) at varying exposure values (EVs) to best highlight these details. In should be noted that these are all examples where we expect the HDR-CNN approach to fail, because the network simply has no information about the detail in the saturated parts—the best it can do is to inpaint these regions. Inpainting results in smooth regions that exceed the dynamic range of the LDR sensor image but that do not resemble the actual content in these examples.

7. Discussion

In summary, we propose an end-to-end approach to jointly train an optical encoder, i.e. the point spread function created by a custom optical element, and electronic decoder, i.e. a convolutional neural network, for the application of single-shot high-dynamic-range imaging. As opposed to CNN-based methods that operate directly on conventional LDR images, our deep optics approach has the ability to optically encode details from bright parts of the scene into the LDR measurements. In particular, our method uses a unique multiplexing solution to HDR imaging, which it developed automatically, wherein the PSF superimposes multiple shifted exposures on top of one another. The proposed framework builds on the emerging idea of end-to-end optimization of optics and image processing, but to our knowledge it is the first to explore this general methodology for single-shot HDR imaging.

Limitations Conceptually, HDR-CNN approaches that operate directly on conventional LDR images solve an inpainting problem. Accordingly, and unlike our deep optics approach, the worst case solution of HDR-CNNs is a conventional LDR image, which is directly recorded and which may not always need additional post-processing to begin with. Our method changes the optical image formation, so post-processing becomes a necessary part of the imaging pipeline. Slight reconstruction artifacts in our experimental results (Figs. 7), which are primarily due to imperfections in the PSF calibration, can be found in both non-saturated and saturated parts of the image due to the need for processing the entire LDR image, rather than just its saturated parts.

In essence, the inverse problem in our method is more closely related to deconvolution problems than to inpainting problems, although for extremely bright scenes where even the lower intensity copies of the sensor image saturate, inpainting is unavoidable and the additional scene copies in our measurements may actually be harmful. Therefore, careful curation of the training set is crucial, because it needs to include HDR images with values adequately representing those observed during inference. A network is likely going to fail to produce high-quality results for conditions that it has not been trained for.

Unlike multi-shot approaches, our single-shot HDR imaging technique does not suffer from ghosting artifacts. However, in dark imaging conditions long exposures, which potentially cause motion blur, may be

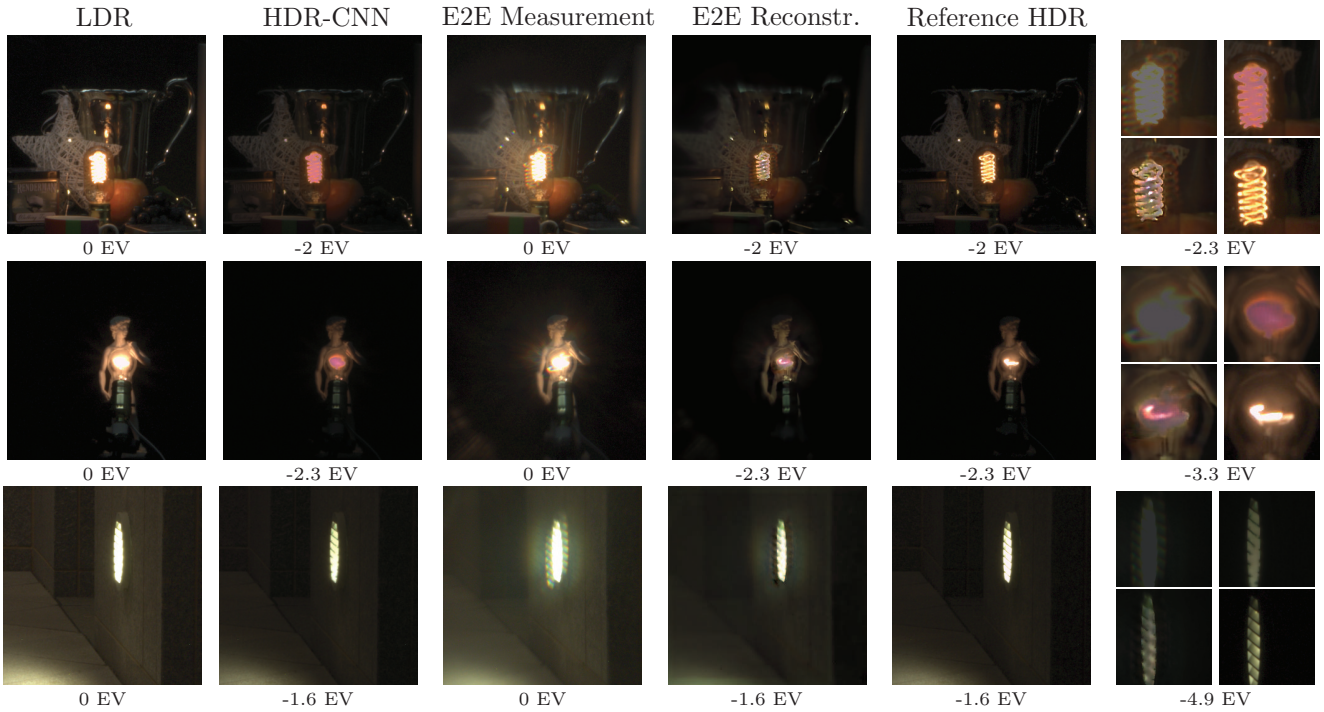

Figure 7: Experimental results of two indoor scenes (top rows) and one outdoor scene at night (bottom row). The limited dynamic range of the sensor loses details in the brighter parts of the captured LDR image (first column) as compared with the reference HDR image (fifth column). A CNN operating directly on the LDR images hallucinates brighter content in these saturated parts, but it is missing the detail (second column). The measurements captured with our prototype camera optically encode this detail in the image by superimposing several scaled and shifted copies of the image on itself (third column). This information is used by our CNN to recover the missing parts of the scene while digitally removing the image copies (fourth column). The closeups, showing E2E measurements, HDR-CNN, E2E Reconstruction, and ground truth HDR, demonstrate that deep optics is more successful in recovering bright detail of HDR scenes than other single-shot HDR imaging approaches (right column).

necessary.

The fabricated diffractive optical element creates a PSF that closely resembles the simulated PSF. Yet, blur and glare, likely due to interreflections between optical elements, are problematic. Optical blur makes the deconvolution problem harder and glare causes the PSF to be shift variant (see the Supplement), which limits the effective field of view of our captured data (depth-of-field is unaffected). Thus, although the add-on approach of mounting the DOE in front of an SLR camera lens is convenient and flexible, integrating the DOE into the aperture plane of the primary lens may produce better results, as it more closely resembles our paraxial image formation model. Alternatively, the image formation model could be generalized to a non-paraxial model. Finally, anti-reflection coatings may help mitigate glare in the optics.

Future Work End-to-end methods enable designing optics tailored for a particular task, rather than just capturing the sharpest image. Evaluating the benefits of end-to-end optimization of optics and image processing for other applications, including multispectral, light field, and lensless imaging or computational microscopy, is an interesting avenue of future work.

Acknowledgments

C.M. was supported the by an ORISE Intelligence Community Postdoctoral Fellowship. G.W. was supported by an NSF CAREER Award (IIS 1553333), a Sloan Fellowship, by the KAUST Office of Sponsored Research through the Visual Computing Center CCF grant, and a PECASE by the ARL. Part of this work was performed at the Stanford Nano Shared Facilities (SNSF)/Stanford Nanofabrication Facility (SNF), supported by the National Science Foundation under award ECCS-1542152.

References

[1] Manoj Aggarwal and Narendra Ahuja. Split aperture imaging for high dynamic range. *International Journal*

of Computer Vision, 58(1):7–17, 2004.

[2] M. Alghamdi, Q. Fu, A. Thabet, and W. Heidrich. Reconfigurable snapshot hdr imaging using coded masks and inception network. In *International Symposium on Vision, Modeling and Visualization*, 2019.

[3] N. Antipa, S. Necula, R. Ng, and L. Waller. Single-shot diffuser-encoded light field imaging. In *Proc. IEEE ICCP*, pages 1–11, 2016.

[4] M. S. Asif, A. Ayremlou, A. Sankaranarayanan, A. Veeraraghavan, and R. G. Baraniuk. Flatcam: Thin, lensless cameras using coded aperture and computation. *IEEE Trans. Computational Imaging*, 3(3):384–397, 2017.

[5] F. Banterle, A. Artusi, K. Debattista, and A. Chalmers. *Advanced High Dynamic Range Imaging: Theory and Practice*. AK Peters (CRC Press), 2011.

[6] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356, 2006.

[7] D. J. Brady, M. E. Gehm, R. A. Stack, D. L. Marks, D. S. Kittle, D. R. Golish, E. M. Vera, and S. D. Feller. Multiscale gigapixel photography. *Nature*, 486:386–389, 2012.

[8] Ayan Chakrabarti. Learning sensor multiplexing design through back-propagation. In *Advances in Neural Information Processing Systems*, pages 3081–3089, 2016.

[9] Julie Chang, Vincent Sitzmann, Xiong Dun, Wolfgang Heidrich, and Gordon Wetzstein. Hybrid optical-electronic convolutional neural networks with optimized diffractive optics for image classification. *Scientific reports*, 8(1):12324, 2018.

[10] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. In *Proc. ICCV*, 2019.

[11] Inchang Choi, Daniel S. Jeon, Giljoo Nam, Diego Gutierrez, and Min H. Kim. High-quality hyperspectral reconstruction using a spectral prior. *ACM Trans. Graph. (SIGGRAPH Asia)*, 36(6):218:1–218:13, 2017.

[12] O. Cossairt, D. Miau, and S.K. Nayar. Gigapixel computational imaging. In *Proc. ICCP*, 2011.

[13] Oliver Cossairt and Shree Nayar. Spectral focal sweep: Extended depth of field from chromatic aberrations. In *Proc. ICCP*, pages 1–8, 2010.

[14] Oliver Cossairt, Changyin Zhou, and Shree Nayar. Diffusion coded photography for extended depth of field. *ACM Trans. Graph. (SIGGRAPH)*, 29(4):31:1–31:10, 2010.

[15] P. Debevec. Image-based lighting. *IEEE Computer Graphics and Applications*, 22(2):26–34, 2002.

[16] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proc. SIGGRAPH*, pages 369–378, 1997.

[17] Edward R. Dowski and W. Thomas Cathey. Extended depth of field through wave-front coding. *OSA Appl. Opt.*, 34(11):1859–1866, 1995.

[18] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM Transactions on Graphics (TOG)*, 36(6):178, 2017.

[19] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph.*, 36(6):177:1–177:10, 2017.

[20] Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. In *Proc. ICCP*, pages 1–7, 2009.

[21] Joseph W Goodman. *Introduction to Fourier optics*. Roberts and Company Publishers, 2005.

[22] Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. Automatic noise modeling for ghost-free hdr reconstruction. *ACM Trans. Graph.*, 32(6):201:1–201:10, 2013.

[23] H. Haim, S. Elmalem, R. Giryes, A. M. Bronstein, and E. Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Trans. Computational Imaging*, 4(3):298–310, 2018.

[24] Saghi Hajisharif, Joel Kronander, and Jonas Unger. Adaptive dualiso hdr reconstruction. *EURASIP Journal on Image and Video Processing*, 2015, 2015.

[25] S. W. Hasinoff, F. Durand, and W. T. Freeman. Noise-optimal capture for high dynamic range photography. In *Proc. CVPR*, pages 553–560, 2010.

[26] Samuel W Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics (TOG)*, 35(6):192, 2016.

[27] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27(9):4676–4689, 2018.

[28] Eran Hershko, Lucien E. Weiss, Tomer Michaeli, and Yoav Shechtman. Multicolor localization microscopy and point-spread-function engineering by deep learning. *OSA Opt. Express*, 27(5):6158–6183, 2019.

[29] Roarke Horstmeyer, Richard Y Chen, Barbara Kappes, and Benjamin Judkewitz. Convolutional neural networks that teach microscopes how to image. *arXiv preprint arXiv:1709.07223*, 2017.

[30] J. Hu, O. Gallo, K. Pulli, and X. Sun. Hdr deghosting: How to deal with saturation? In *Proc. CVPR*, 2013.

[31] Daniel S Jeon, Seung-Hwan Baek, Shinyoung Yi, Qiang Fu, Xiong Dun, Wolfgang Heidrich, and Min H Kim. Compact snapshot hyperspectral imaging with diffracted rotation. *ACM Transactions on Graphics (TOG)*, 38(4):117, 2019.

[32] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph. (SIGGRAPH)*, 36(4), 2017.

[33] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based High Dynamic Range Video. *ACM Trans. Graph. (SIGGRAPH Asia)*, 32(6), 2013.

[34] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Trans. Graph. (SIGGRAPH)*, 22(3):319–325, 2003.

[35] Michael Kellman, Emrah Bostan, Michael Chen, and Laura Waller. Data-driven design for fourier ptychographic microscopy. In *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–8. IEEE, 2019.

[36] E. A. Khan, A. O. Akyuz, and E. Reinhard. Ghost removal in high dynamic range images. In *Proc. ICIP*, pages 2005–2008, 2006.

[37] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proc. ECCV*, pages 596–611, 2018.

[38] Anat Levin, Rob Fergus, Frédo Durand, and William T. Freeman. Image and depth from a conventional camera with a coded aperture. *ACM Trans. Graph. (SIGGRAPH)*, 26(3), 2007.

[39] Anat Levin, Samuel W. Hasinoff, Paul Green, Frédo Durand, and William T. Freeman. 4d frequency analysis of computational cameras for depth of field extension. *ACM Trans. Graph. (SIGGRAPH)*, 28(3):97:1–97:14, 2009.

[40] Mann, Picard, S. Mann, and R. W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combining differently exposed pictures. In *Proceedings of IS&T*, pages 442–448, 1995.

[41] Rafat Mantiuk, Kil Joong Kim, Allan G. Rempel, and Wolfgang Heidrich. Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Trans. Graph. (SIGGRAPH)*, 30(4):40:1–40:14, 2011.

[42] Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrian Jarabo, Min H. Kim, Xin Tong, and Diego Gutierrez. Deeptof: Off-the-shelf real-time correction of multipath interference in time-of-flight imaging. *ACM Trans. Graph. (SIGGRAPH Asia)*, 36(6):219:1–219:12, 2017.

[43] J.N.P. Martel, L.K. Müller, S.J. Carey, P. Dudek, and G. Wetzstein. Neural Sensors: Learning Pixel Exposures for HDR Imaging and Video Compressive Sensing with Programmable Sensors. *Proc. IEEE ICCP*, 2020.

[44] Kshitij Marwah, Gordon Wetzstein, Yosuke Bando, and Ramesh Raskar. Compressive light field photography using overcomplete dictionaries and optimized projections. *ACM Trans. Graph. (SIGGRAPH)*, 32(4):46:1–46:12, 2013.

[45] M. McGuire, W. Matusik, H. Pfister, B. Chen, J. F. Hughes, and S. K. Nayar. Optical splitting trees for high-precision monocular imaging. *IEEE Computer Graphics and Applications*, 27(2):32–42, 2007.

[46] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum (Eurographics)*, volume 28, pages 161–171. Wiley Online Library, 2009.

[47] Laurence Meylan, Scott Daly, and Sabine Süsstrunk. The reproduction of specular highlights on high dynamic range displays. *IS&T/SID 14th Color Imaging Conference (CIC)*, 2006.

[48] T. Mitsunaga and S. K. Nayar. Radiometric self calibration. In *Proc. IEEE CVPR*, volume 1, pages 374–380, 1999.

[49] S. K. Nayar and T. Mitsunaga. High dynamic range imaging: spatially varying pixel exposures. In *Proc. CVPR*, volume 1, pages 472–479 vol.1, 2000.

[50] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Tomer Michaeli, and Yoav Shechtman. Dense three dimensional localization microscopy by deep learning. *arXiv preprint arXiv:1906.09957*, 2019.

[51] Ren Ng, Marc Levoy, Mathieu Bredif, Gene Duval, Mark Horowitz, and Pat Hanrahan. Light field photography with a hand-held plenoptic camera. Tech Report CSTR 2005-02, 2005.

[52] Sri Rama Prasanna Pavani, Michael A. Thompson, Julie S. Biteen, Samuel J. Lord, Na Liu, Robert J. Twieg, Rafael Piestun, and W. E. Moerner. Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function. 106(9):2995–2999, 2009.

[53] Yifan Peng, Qiang Fu, Felix Heide, and Wolfgang Heidrich. The diffractive achromat full spectrum computational imaging with diffractive optics. *ACM Trans. Graph. (SIGGRAPH)*, 35(4):31:1–31:11, 2016.

[54] Ramesh Raskar, Amit Agrawal, and Jack Tumblin. Coded exposure photography: Motion deblurring using fluttered shutter. *ACM Trans. Graph. (SIGGRAPH)*, 25(3):795–804, 2006.

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

[56] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: On-the-fly reverse tone mapping of legacy video and photographs. *ACM Trans. Graph. (SIGGRAPH)*, 26(3), 2007.

[57] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

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

[59] Helge Seetzen, Wolfgang Heidrich, Wolfgang Stuerzlinger, Greg Ward, Lorne Whitehead, Matthew Trentacoste, Abhijeet Ghosh, and Andrejs Vorozcovs. High dynamic range display systems. *ACM Trans. Graph. (SIGGRAPH)*, 23(3):760–768, 2004.

[60] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B. Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph. (SIGGRAPH Asia)*, 31(6):203:1–203:11, 2012.

[61] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum (Eurographics)*, pages 153–163, 2016.

[62] Yoav Shechtman, Steffen J. Sahl, Adam S. Backer, and W. E. Moerner. Optimal point spread function design for 3d imaging. *Phys. Rev. Lett.*, 113:133902, 2014.

[63] Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

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

[65] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. In *Proc. CVPR*, 2018.

[66] Michael D. Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. *ACM Trans. Graph. (SIGGRAPH)*, 30(4):41:1–41:10, 2011.

[67] Alex Turpin, Ivan Vishniakou, and Johannes D Seelig. Light scattering control with neural networks in transmission and reflection. *Arxiv: 180505602 [Cs]*, 2018.

[68] Ashok Veeraraghavan, Ramesh Raskar, Amit Agrawal, Ankit Mohan, and Jack Tumblin. Dappled photography: Mask enhanced cameras for heterodyned light fields and coded aperture refocusing. *ACM Trans. Graph. (SIGGRAPH)*, 26(3), 2007.

[69] Ashwin Wagadarikar, Renu John, Rebecca Willett, and David Brady. Single disperser design for coded aperture snapshot spectral imaging. *OSA Appl. Opt.*, 47(10):B44–B51, 2008.

[70] G. Wetzstein, I. Ihrke, and W. Heidrich. Sensor saturation in fourier multiplexed imaging. In *Proc. CVPR*, pages 545–552, 2010.

[71] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d – learning phase masks for passive single view depth estimation. In *Proc. ICCP*, 2019.

[72] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *Advances in neural information processing systems*, pages 1790–1798, 2014.

[73] H. Zhao, B. Shi, C. Fernandez-Cull, S. Yeung, and R. Raskar. Unbounded high dynamic range photography using a modulo camera. In *Proc. ICCP*, pages 1–10, 2015.

[74] C. Zhou, S. Lin, and S. K. Nayar. Coded aperture pairs for depth from defocus. In *Proc. ICCV*, 2009.

[75] C. Zhou and S. Nayar. What are good apertures for defocus deblurring? In *Proc. IEEE ICCP*, pages 1–8, 2009.