Learning Rank-1 Diffractive Optics for Single-shot High Dynamic Range Imaging

Qilin Sun[1] Ethan Tseng[2] Qiang Fu[1] Wolfgang Heidrich[1] Felix Heide[2]
[1]KAUST [2]Princeton University

Abstract

High-dynamic range (HDR) imaging is an essential imaging modality for a wide range of applications in uncontrolled environments, including autonomous driving, robotics, and mobile phone cameras. However, existing HDR techniques in commodity devices struggle with dynamic scenes due to multi-shot acquisition and post-processing time, e.g. mobile phone burst photography, making such approaches unsuitable for real-time applications. In this work, we propose a method for snapshot HDR imaging by learning an optical HDR encoding in a single image which maps saturated highlights into neighboring unsaturated areas using a diffractive optical element (DOE). We propose a novel rank-1 parameterization of the DOE which drastically reduces the optical search space while allowing us to efficiently encode high-frequency detail. We propose a reconstruction network tailored to this rank-1 parametrization for the recovery of clipped information from the encoded measurements. The proposed end-to-end framework is validated through simulation and real-world experiments and improves the PSNR by more than 7 dB over state-of-the-art end-to-end designs.

1. Introduction

High dynamic range (HDR) imaging has become a commodity imaging technique as evident by its applications across many domains, including mobile consumer photography, robotics, drones, surveillance, content capture for display, driver assistance systems, and autonomous driving. The pixels in conventional CMOS and CCD image sensors act as potential wells that saturate when the well capacity is reached. Unlike film, which provides a gradual compression of high intensities, digital image sensors thus suffer from a hard cutoff at some peak intensity, so that information about the saturated bright regions is irrevocably lost. By reducing the exposure time, brighter regions can be recovered, but at the cost of under-exposing, i.e., reducing signal photons in darker image regions. As a result, single captures of conventional sensors provide high fidelity only for low-contrast scenes, while struggling for high-contrast scenes at night with both low- and high-flux scene content.

Although existing HDR imaging methods in widely deployed consumer smartphone devices [20, 45, 23] successfully overcome this limitation by acquiring bursts of captures, the combined capture and processing time of multiple seconds [20] is prohibitive for many applications in robotics and autonomous driving that demand real-time feeds.

Faster multi-capture imaging methods [6, 56, 42], relying on only 2-3 low-dynamic range exposures and hardware exposure fusion, fail for higher dynamic scenes typical in automotive and robotics applications. As an alternative, emerging sensor designs multiplex exposures on the sensor [48, 64, 46], however, at the cost of spatial resolution required for spectral or spatial information. Optical splitting methods using multiple sensors [62] are often not practical in an application due to their cost and footprint. To tackle this issue, a line of recent work explores the hallucination of HDR images [9, 12] from single low-dynamic range (LDR) captures. These methods can only rely on semantic context but no measurement signal to recover the clipped HDR regions. In an alternative direction, Rouf et al. [55] proposed a hand-crafted star filter attachment to optically encode lost information by spreading out saturated highlights to nearby regions. Unfortunately, their approach only achieves low image quality far below that of recent hallucination approaches.

In this work, we revisit this idea, but by learning an optical HDR encoding in an end-to-end optimization. To this end, we jointly design the optical point spread function (PSF) together with the inverse reconstruction method, i.e., the post-capture processing that recovers the latent HDR scene from the input measurement, which we formulate as an (RAW-)image-to-image neural network. However, we found that applying existing end-to-end methods [60, 43] easily finds a local minimum of the vast design space, parameterized by an unconstrained diffractive phase plate optic. Instead, we parameterize the diffractive element in the proposed optical design with a rank-1 phase pattern. This constrained PSF design spaces makes allows us to tailor the

Figure 1: Due to pixel saturation, image detail in bright regions is lost in a single snapshot LDR image. Our camera, with the learned optic prototype (left), captures LDR images where high-intensity image content is encoded through a series of streaks (center left). This allows us to reconstruct the lost highlights (center, center right) with a specialized two-stage CNN.

architecture of the reconstruction network to the recovery from such streak-encoded measurement. We optimize the diffractive optic and reconstruction algorithm jointly in an end-to-end optimization which finds a local minimum that outperforms vanilla end-to-end designs with similar network capacity by more than 7 dB PSNR.

We demonstrate the proposed approach outperforms the state-of-the-art snapshot HDR methods in simulation. We prototype our design camera system by fabricating the DOE and demonstrate on a broad set of experimental in-the-wild captures, that this method generalizes to unseen scenarios, outperforming existing optical designs. Our method is most effective for recovering concentrated high-intensity light sources such as street lamps. In addition, we also show that the proposed network is effective in removing glare from in-the-wild automotive optics with windshield-induced streaks.

Specifically, we make the following contributions:

- We introduce a novel rank-1 parameterized optical design that learns to encode saturated information with a streak-like PSF.
- We co-design a tailored reconstruction network which first splits the unsaturated regions from the coded information and then recovers the saturated highlights from the encodings.
- We validate the proposed method in simulation and on real-world measurements acquired with a fabricated prototype system. The proposed method outperforms existing designs by over 7 dB in simulation.

2. Related Work

Multi-exposure HDR Imaging Traditionally, HDR imaging is performed by sequentially capturing LDR images for different exposures and then combining them through exposure bracketing [37, 7, 52, 15, 41, 19]. This approach is unsuitable for handling highly dynamic scenes and for fast captures necessary for real-time applications. More rapid HDR imaging can be realized with burst HDR acquisition [20, 45, 23]. However, these techniques still suffer from motion artifacts and require seconds for capture and processing for a single acquisition.

To alleviate motion artifacts, prior work has employed HDR stitching [31, 33], optical flow [36], patch matching [13, 14, 26, 30, 57], and deep learning [28, 29]. These techniques have even enabled HDR videography, but the post-processing cost makes them impractical for fast capture. Ultimately, these approaches attempt to find a trade-off between densely sampling different exposures and post-processing computation time.

HDR Snapshot Reconstruction A large body of work has explored reconstructing HDR content from a single LDR image, a process referred to as inverse tone-mapping. Early work in this domain utilized heuristic approaches [3, 8, 44, 53], but often does not provide satisfying HDR reconstructions [1, 40]. Building upon these works, deep learning has been used to hallucinate HDR content from LDR images [9, 12, 47, 10, 66, 34, 35, 63, 39, 50, 27]. These approaches generate plausible reconstructions of low-light regions but fail to recover saturated details accurately.

Several approaches encode information into the captured LDR image to allow for better estimation of highlights. This can be achieved by modifying the sensor architecture through spatially varying pixel exposures [48, 18], convolutional sparse coding [58], compressed sensing [16], or modulo cameras [68]. Drawbacks of these approaches include the need for expensive custom cameras and loss of detail in the low dynamic range. Furthermore, recovering highlights in scenes involving very large dynamic ranges is still a challenge for these approaches. Instead of modifying the sensor, other approaches place optical components in front of conventional cameras to affect the captured LDR image. Hirakawa et al. [24] utilized color filters to avoid saturation of any single color channel. Rouf et al. [55] proposed to use a known optical element to spread saturated information content into unsaturated regions. Although this allows for high fidelity estimation of highlights, these techniques leave noticeable artifacts in the unsaturated areas.

End-to-end Optics Design Joint optimization of optics and reconstruction has demonstrated superior performance over traditional heuristic approaches in color image restoration [4], microscopy[25, 32, 49, 59], monocular depth imaging [5, 17, 21, 65], super-resolution and extended depth of field [60], and time-of-flight imaging [38, 61].

We propose an end-to-end optimization framework for single-shot HDR imaging. Drawing inspiration from Rouf et al. [55], our optimized optic is a DOE that encodes clipped highlights into specific unsaturated regions. The ample design space of DOEs allows for rich optical encodings but has the unintended consequence of being challenging to optimize. As such, investigating a suitable parameterized model of the DOE becomes a critical design step. Recent work, in parallel to ours, explores end-to-end optimization of optics for HDR imaging by directly learning a heightmap for the DOE [43]. This approach causes the DOE to produce shifted scaled copies of saturated content that allow for HDR reconstruction but that are difficult to remove from the unsaturated regions. Another approach used by Sitzmann et al. [60] is to represent the DOE with Zernike polynomials, but this only allows the DOE to affect low frequencies and is inadequate for capturing high-frequency detail in HDR scenes.

In this work, we found that by constraining the DOE height map model to a rank-1 phase pattern, our DOE learns to produce streak patterns that are easy to remove from the unsaturated regions but still allow for high fidelity reconstruction of saturated image content. We employ a structured multi-stage CNN, instead of a single-stage U-Net as in [43], to perform these tasks step by step.

3. Image Formation Model

Our image formation model is illustrated in Figure 2. In the following, we describe the individual parts of this forward model, directly parameterized in a way that can later be used for end-to-end optimization.

Point Light Source. Our optical model begins with a point light source placed 5 m in front of the DOE plane. Like most camera systems, the PSF for our optical model is depth dependent. We chose a 5 m focal point as a compromise for near and infinite scene depths. We also analyze the robustness with varying depth, please refer to the supplemental material for details.

The point source generates a spherical wave. Upon the arrival of the wavefront to the DOE plane the phase of the wavefront can be expressed as

$$\mathbf{u}_- = A_0 e^{jk\sqrt{x'^2+y'^2+z^2}}, \tag{1}$$

where A_0 is the amplitude, $k = 2\pi/\lambda$ is the wave number, and z is the distance from the point source and DOE center.

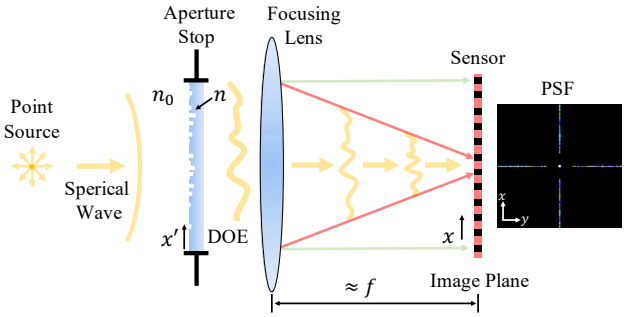

Figure 2: Our optical forward model consists of a point light source which generates a spherical wave that is modulated by the DOE and focusing lens before being captured by the sensor. The corresponding PSF is used to simulate images captured by our camera prototype.

DOE Layer and Rank-1 Factorization. We then use a DOE layer to modulate the incident wave and set the DOE plane as the aperture $\mathcal{A}(x', y')$ of the whole optical system. The modulated field can be expressed as

$$\mathbf{u}_+ = \mathcal{A}(x', y')\mathbf{u}_- e^{jk(n_\lambda - 1)\mathbf{h}(x', y')}, \tag{2}$$

where n_λ is the wavelength-dependent refractive index of the DOE and $\mathbf{h}(x', y')$ is the height map of the DOE.

Existing end-to-end frameworks have used an unconstrained height map model for the DOE, where each location in the $m \times m$ height map is a learnable parameter [60]. We found that this model has a tendency to produce local minima in the form of very local encodings such as shifted and scaled copies of highlights, as also shown in parallel work [43], but rarely produces non-local encodings such as streaks. Using these local encodings provides lower quality HDR reconstructions as they are difficult to separate from the unsaturated areas in the close neighborhood. We note that alternative parameterizations such as a truncated Zernike basis [60] are also not suitable for our application, because even though it can model non-local encodings, it is only suitable for low spatial frequencies and cannot encode high-frequency content.

To tackle this challenge, we propose a novel rank-1 decomposition of the 2D height map which not only can encode high frequencies but also reduces the number of parameters touched in training. The height map at location (x', y') is given by

$$\mathbf{h}(x', y') = h_{\text{max}}\, \sigma\left(\mathbf{v}\mathbf{q}^\top\right), \tag{3}$$

where $\mathbf{v} \in \mathbb{R}^{m\times 3}$ and $\mathbf{q} \in \mathbb{R}^{m\times 3}$ are trainable variable basis pairs whose outer product describes the DOE height map, σ is the sigmoid function, and $h_{\text{max}} = 1.125\ \mu$m is the maximum height that corresponds to 2π phase modulation at $\lambda = 550$ nm for fused silica. The sigmoid function $\sigma(x) = 1/(1 + e^{-x})$ is applied element-wise to $\mathbf{v}\mathbf{q}^\top$ to clamp the range to [0,1].

Our rank-1 parameterized model encourages global optical encodings, such as streaks, while still permitting local encodings, such as peaks. Furthermore, this parameterization produces a grating-like height map which is more suitable for DOE fabrication. In addition to the rank-1 parameterization, we also use an additional constraint to assist the framework in finding proper optical encodings. To ensure that highlights are encoded without severely affecting low-light regions, we adopt a regularization loss to constrain 94% of the energy into the center of the PSF and to spread the remaining 6% into the surroundings. We found that if we take our converged height map and continue optimizing without our rank-1 parameterization, then the optimized height profile is still maintained, which suggests that we do indeed find a good local optimum. Please refer to the Supplemental Material for details.

Focusing-Lens Layer. We place a well-corrected lens (approximated as a thin lens) immediately behind the DOE. This lens is responsible for focusing the image, and allows the DOE to be purely optimized for the HDR encoding without also requiring the focusing operation for broadband illumination. The wave field $\mathbf{u}_l$ can be expressed as

$$\mathbf{u}_l = \mathbf{u}_+ e^{jk(f-\sqrt{x'^2+y'^2+f^2})}. \tag{4}$$

Fresnel Propagation Layer. We use the Fresnel approximation here to describe the field propagation from the focusing-lens to sensor. Specifically, the field $\mathbf{u}_s$ at the sensor plane can be expressed as

$$\mathbf{u}_s = \mathcal{F}^{-1}\{\mathcal{F}\{\mathbf{u}_l\}\mathcal{H}\}, \tag{5}$$

where $\mathcal{H}(f_x, f_y) = e^{jkL}e^{-j\pi\lambda L(f_x^2+f_y^2)}$, with $f_x = 1/2\Delta x'$ and $f_y = 1/2\Delta y'$, is the Fresnel propagation kernel and L is the distance between the normal lens and sensor plane. Finally, the PSF corresponding to the entire image formation model is given by $\mathbf{p} \propto |\mathbf{u}_s|^2$.

Sensor Model. The image captured by the sensor $\mathbf{I}_s$ is given by

$$\mathbf{I}_s = \mathbf{s}(\mathbf{I} * \mathbf{p} + \eta), \tag{6}$$

where $\mathbf{I}$ is the high dynamic range ground truth image, $\mathbf{p}$ is the point-spread function of the optical system, η is sensor noise, and $\mathbf{s}(\cdot)$ is the camera response function that clips to [0, 1]. Note that $\mathbf{I}_s$ and $\mathbf{I}$ are both continuous variables.

4. End-to-end Design and Reconstruction

The proposed end-to-end imaging system consists of three main parts: a differentiable optical model, a robust network for recovering and separating the unsaturated image $\mathbf{I}_\mathbb{U}$ (i.e., pixel values in $[0, 1]$) from the residual information $\mathbf{I}_r$ encoded by the saturated image $\mathbf{I}_\mathbb{S}$ (i.e., pixel values in $[1, 2^8]$), and a reconstruction network for inferring $\mathbf{I}_\mathbb{S}$ from $\mathbf{I}_r$. In a final step, the recovered unsaturated component $\mathbf{I}_\mathbb{U}$ and the recovered highlight component $\mathbf{I}_\mathbb{S}$ are combined using a fusion network to predict the latent HDR image $\mathbf{I}$.

Differentiable Optical Model We implement the optical model as described in Section 3.

Residual Splitting Network. We first discuss the network for reconstructing $\mathbf{I}_\mathbb{U}$ and separating this unsaturated part from $\mathbf{I}_r$. Our residual splitting network $f_\mathbb{U}$ takes in the coded LDR sensor capture $\mathbf{I}_s$ and outputs a prediction $\hat{\mathbf{I}}_\mathbb{U}$ for the unsaturated image and a prediction $\hat{\mathbf{I}}_r$ for the residual information:

$$\hat{\mathbf{I}}_\mathbb{U}, \hat{\mathbf{I}}_r = f_\mathbb{U}\left(\mathbf{I}_s\right). \tag{7}$$

Inspired by recent work on separation of reflection from transmission in single-shot images [67], the network first uses a pre-trained VGG model to extract feature maps. Specifically, we used the pre-trained VGG-19 network to extract "conv1_2", "conv2_2", "conv3_2", "conv4_2" and "conv5_2" feature maps and bilinearly upsampled them to the input image size. These feature maps, along with the input image, are then compressed to 64 channels by using a 1×1 convolution layer before being fed through seven 3×3 dilated convolution layers with dilation rates from 1 to 64 (Dilated Full-Resolution Reconstruction Block in Fig 3). Each dilated convolution layer has 64 channels. Finally, a 1×1 convolution layer is used to reduce to six channels, three of which correspond to $\hat{\mathbf{I}}_\mathbb{U}$ and the other three correspond to $\hat{\mathbf{I}}_r$. Each dilated convolution layer is followed by a Leaky ReLu activation (slope = 0.2) and an instance normalization layer. The loss on the unsaturated pixels $\mathcal{L}_\mathbb{U}$ as shown in Figure 3 forces this network to effectively split streaks from the unsaturated image $\hat{\mathbf{I}}_\mathbb{U}$.

Highlight Reconstruction Network. After splitting the unsaturated image from the residual encoding we then use the residual to reconstruct highlights. Since the residual encoding was produced by convolving the highlights with our designed PSF, reconstructing the highlights becomes a deconvolution problem. Our network $f_\mathbb{S}$ thus takes in the residual prediction $\hat{\mathbf{I}}_r$ and outputs a prediction $\hat{\mathbf{I}}_\mathbb{S}$ of the highlights:

$$\hat{\mathbf{I}}_\mathbb{S} = f_\mathbb{S}\left(\hat{\mathbf{I}}_r\right). \tag{8}$$

We rely on a variation of the U-Net architecture [54] to deal with this deconvolution task. Specifically, our U-Net has five scales with four consecutive downsamplings (max-pools) and four consecutive upsamplings (nearest neighbor upsampling following by a 3×3 convolution layer). Each layer uses a 3×3 kernel window except for the first layer with 7×7 and the last layer with 1×1. Since the coded information is in the range $[0, 1]$ while the values to reconstruct are in the range $[1, 2^8]$, we avoid using normalization in the last two convolution layers to allow the network to learn a large range of output values. Similar to the splitting network, the loss $\mathcal{L}_\mathbb{S}$ encourages high-fidelity highlight reconstructions.

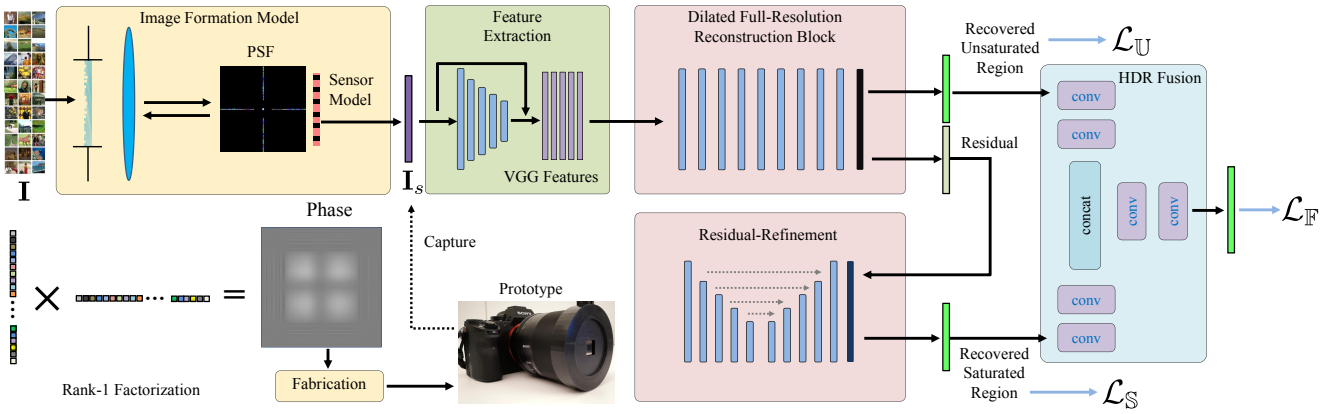

Figure 3: Our end-to-end pipeline consists of the image formation model and CNN reconstruction. Our CNN is divided into several stages that focus on separating the encoding from the captured LDR image, recovering the highlights, and fusing the recovered unsaturated and saturated regions to produce the final HDR prediction. After fabrication our image formation model is replaced by real-world captures.

Fusion Network. In order to avoid boundary artifacts caused by combining $\hat{\mathbf{I}}_{\mathbb{U}}$ and $\hat{\mathbf{I}}_{\mathbb{S}}$ into a single image, we adopt a light-weight fusion network $f_{\mathbb{F}}$ to combine them and create the final predicted HDR image $\hat{\mathbf{I}}_{\mathbb{F}}$:

$$\hat{\mathbf{I}}_{\mathbb{F}} = f_{\mathbb{F}}\left(\hat{\mathbf{I}}_{\mathbb{U}}, \hat{\mathbf{I}}_{\mathbb{S}}\right). \tag{9}$$

The fusion network applies two 3×3 convolution layers with 64 feature channels to $\hat{\mathbf{I}}_{\mathbb{U}}$ and $\hat{\mathbf{I}}_{\mathbb{S}}$ separately, concatenates the feature maps together, and then applies two 3×3 convolution layers with 32 and 3 feature channels to produce the final predicted output $\hat{\mathbf{I}}_{\mathbb{F}}$.

4.1. Loss Functions

Our loss functions consist of two intermediary losses $\mathcal{L}_{\mathbb{U}}$ and $\mathcal{L}_{\mathbb{S}}$ which we apply to the intermediate outputs of the residual splitting network and the highlight reconstruction network respectively. We also apply a final loss $\mathcal{L}_{\mathbb{F}}$ to the final output of our network. The total loss that we minimize when training our network is given by

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\mathbb{U}} + \mathcal{L}_{\mathbb{S}} + \mathcal{L}_{\mathbb{F}}. \tag{10}$$

Loss on Unsaturated Regions We adopt a perceptual loss as a metric of difference between the intermediate unsaturated image prediction $\hat{\mathbf{I}}_{\mathbb{U}}$ and the ground truth unsaturated image $\mathbf{I}_{\mathbb{U}}$. Our perceptual loss is defined using the pre-trained VGG-19 network and is given by

$$\mathcal{L}_{\text{VGG}}(\hat{x}, x) = \sum_l \nu_l \|\phi_l(\hat{x}) - \phi_l(x)\|_1, \tag{11}$$

where $\{\nu_l\}$ are loss balancing weights and ϕ_l are the feature maps from the l-th layer of pre-trained VGG-19. Specifically, we use the "conv2_2", "conv3_2", and "conv4_2" layers of the VGG-19 network.

To better separate the unsaturated region prediction from the residual prediction, we apply an exclusion loss [67] $\mathcal{L}_{\text{excl}}$ during network fine-tuning. We assume that the edges of the unsaturated image and the edges of the residual encoding are unlikely to overlap, and we apply this assumption through an exclusion loss that penalizes correlation between the predicted unsaturated image and the residual in the gradient domain. The exclusion loss is defined to be

$$\mathcal{L}_{\text{excl}} = \|\tanh(\eta_{\mathbb{U}} \mid \nabla\hat{\mathbf{I}}_{\mathbb{U}} \mid) \odot \tanh(\eta_r \mid \nabla\hat{\mathbf{I}}_r \mid)\|_F, \tag{12}$$

where $\eta_{\mathbb{U}} = \sqrt{\|\hat{\mathbf{I}}_r\|_F / \|\hat{\mathbf{I}}_{\mathbb{U}}\|_F}$ and $\eta_r = \sqrt{\|\hat{\mathbf{I}}_{\mathbb{U}}\|_F / \|\hat{\mathbf{I}}_r\|_F}$ represent normalization factors, and $\|\cdot\|_F$ represents the Frobenius norm.

In conclusion, the loss that is applied to the intermediate output of the residual splitting network is

$$\mathcal{L}_{\mathbb{U}} = \alpha_1 \mathcal{L}_{\text{VGG}}(\hat{\mathbf{I}}_{\mathbb{U}}, \mathbf{I}_{\mathbb{U}}) + \alpha_2 \mathcal{L}_{\text{excl}}(\hat{\mathbf{I}}_{\mathbb{U}}, \hat{\mathbf{I}}_r). \tag{13}$$

Loss on Saturated Regions To extract information and perform deconvolution from the residual artifacts we use the same VGG loss given in Eq 11 for the intermediate prediction of the saturated highlights:

$$\mathcal{L}_{\mathbb{S}} = \beta \mathcal{L}_{\text{VGG}}(\hat{\mathbf{I}}_{\mathbb{S}}, \mathbf{I}_{\mathbb{S}}). \tag{14}$$

Loss on Fused Output We applied a Huber loss with $\gamma = 1/2$ to the ground truth HDR image $\mathbf{I}$ and final network prediction $\hat{\mathbf{I}}_{\mathbb{F}}$:

$$\mathcal{L}_{\mathbb{F}} = \mathcal{L}_{\text{Huber}}\left((\hat{\mathbf{I}} + \epsilon)^\gamma, (\mathbf{I} + \epsilon)^\gamma\right). \tag{15}$$

4.2. Implementation and Training

We implement our rank-1 DOE height map model and reconstruction network in TensorFlow. Our network assumes inputs are in the range $[0, 1]$, and outputs are in the range $[0, 2^8]$. The model is jointly optimized using the Adam optimizer with polynomial learning rate decay. For more details, please refer to the Supplemental Material.

Table 1: Quantitative comparison across single-shot HDR methods.

Methods	PSNR	HDR-VDP 2
Ours	**48.26**	**74.47**
Deep Optics [43]	40.30	67.96
Glare-HDR [55]	32.23	56.76
HDR-CNN [9]	34.06	54.34
LDR	33.57	52.43

4.3. Dataset

To ensure that our model accurately reconstructs highlights, we gathered HDR images that contain large dynamic ranges with small saturated regions. These include a mix of urban and rural scenes at night as well as indoor scenes from 19 different sources, see Supplemental Material for a complete list of dataset sources. To accommodate different image sizes we manually took 512×512 crops of the images specifically located at where the saturated regions were. After preprocessing, we had a total of 2039 images for training and 36 images for testing.

As part of the sensor simulation, we saturate a random percentage of pixels during training. That is, we multiply images by a scale factor such that 1% to 3% of pixels are larger than 1. After the scaling, we clip extreme pixel values, any pixel values larger than 2^8 are set to 2^8. We also augment the images using random rotations and flips. During testing, we saturate exactly 1.5% of the pixels in all test images and again clip pixel values larger than 2^8.

5. Evaluation and Comparisons

We evaluate our approach in simulation against recent state-of-the-art single-shot HDR methods [43, 55, 9]. For HDR-CNN we used their pre-trained model. For Rouf et al.'s glare filter method, we applied an 8-point star PSF to the image using their experimentally obtained glare filter. For Deep Optics, we used the authors' PSF and trained their network on our dataset. Table 1 displays quantitative results on the test set. PSNR is calculated in the linear domain with a maximum value of 2^8. HDR-VDP Version 2.2.1 was used with default settings except for pixels per degree which was computed using 24 inch diagonal display size, 512×512 resolution, and 1 m viewing distance. We report the Quality Correlation score. Figure 4 shows qualitative comparisons of our approach against others.

5.1. Ablation Study

We performed an ablation study to illustrate the benefits of our proposed optical design and reconstruction network. Table 2 compares performance when using different reconstruction networks. We found that our network was best suited to HDR recovery with our learned PSF. Table 2 also shows performance when using different PSFs with our reconstruction network. For these experiments, only the network was optimized, and the PSF remains fixed. We observed that our PSF outperforms alternative PSF designs.

Table 2: Ablation study with different PSFs and reconstruction networks.

PSF	Network	PSNR	HDR-VDP 2
Ours	**Ours**	**48.26**	**74.47**
Ours	Deep Optics [43]	37.91	61.30
Ours	HDR-CNN [9]	33.51	52.66
Dual Peak PSF [43]	Ours	43.08	70.25
Star PSF [55]	Ours	42.62	68.03
Dirac PSF	Ours	37.25	63.45

6. Experimental HDR Captures

We fabricate the optimized DOE using multilevel photolithography techniques [51, 22]. Due to fabrication limitations, we first slice the continuous phase mask into four layers with $2^4 = 16$ phase levels. This results in a high diffraction efficiency (theoretically more than 90%) [11]. By repeating the photolithography and reactive ion etching (RIE) for four iterations, we fabricated the phase mask on a 0.5 mm fused silica substrate with aperture size 9.16 mm and feature size 6 μm. Please refer to the Supplemental Material for further fabrication details.

Our imaging pipeline uses a Sony A7 with a pixel pitch of 5.97 μm, and the phase mask is closely placed in front of a Zeiss 50 mm f/1.4 lens (recall that we do not model the propagation between DOE and standard lens). Figure 5 shows that the real-world PSF matches the simulated PSF with slight contrast loss due to manufacturing imperfections and model approximations. Therefore, we perform a PSF calibration step where we capture the real-world PSF and then use it to fine-tune our reconstruction network. The real-world PSF is obtained by placing a white point light source 5 m away from the sensor, taking multi-exposure (five) snapshots at 3 EV intervals, and then fusing the snapshots in linear space using MATLAB's HDR toolbox.

6.1. Results

Figures 1 and 6 show real-world captures and reconstructions with our setup and reconstruction procedure. Reference images were taken by the same camera without the DOE (same aperture and position) using exposure fusion as described above for the PSF capture. In Figure 1, we correctly reconstruct highlights in the illuminated letters while removing most of the encoding streaks. In Figure 6 the left images show a brick wall where details are lost due to the light sources. Our method recovers these details, including color and structure. Note that our method succeeds despite interference between the background image and our streak encodings. The middle images show that our method also works for dynamic scenes with of flashing

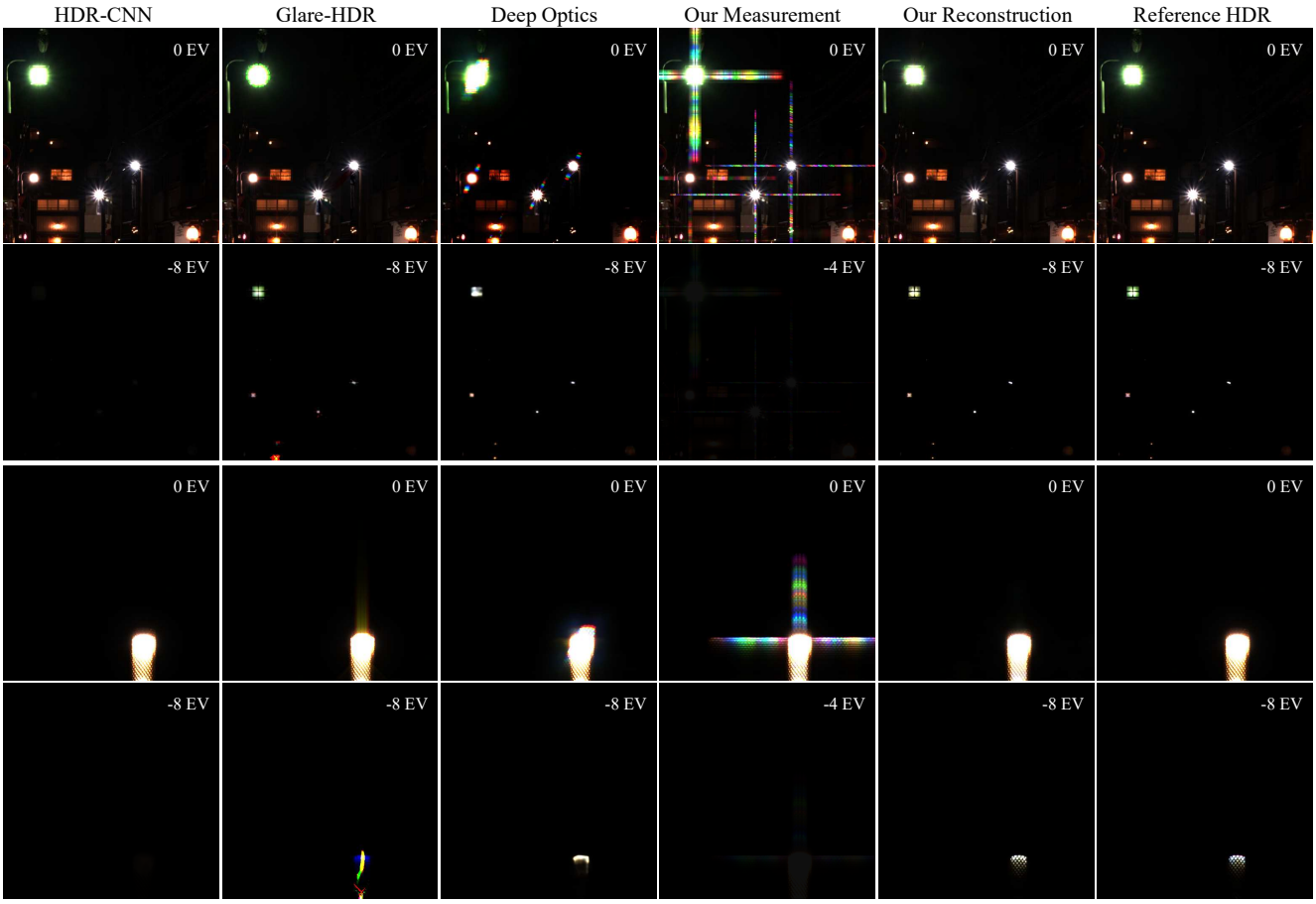

Figure 4: Visual comparison of different snapshot HDR methods in simulation. HDR-CNN [9] severely underestimates the intensity of saturated regions. Glare-HDR [55] often leaves artifacts and fails to estimate highlights correctly. The copied peaks for Deep Optics [43] sometimes overlap with the saturated areas and consequently are ineffective for highlight reconstruction. Please zoom in to view image details, such as the fine structures within the saturated areas.

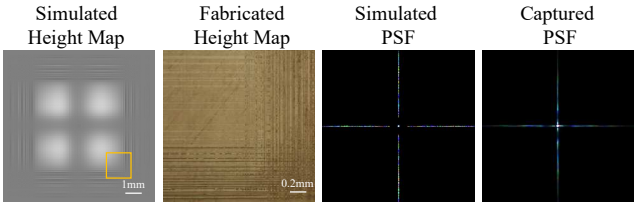

Figure 5: Comparison of simulation versus the real-world for the heightmap and PSF. Fine-tuning was performed with the captured PSF.

strobe lights, which are challenging for burst HDR methods as the bursts would not be synchronized with the strobe. The right side shows that detailed reconstructions can be obtained for high-intensity lamp regions.

The presented reconstruction results and additional results in the Supplemental Material validate the proposed method for various scene types, including high-contrast night time urban environments and indoor settings. However, it is important to use a low exposure time as our method fails when the streaks are overexposed.

To evaluate real-time applicability we benchmarked the reconstruction latency. Our unoptimized network in TensorFlow takes 530 ms on an Nvidia Titan RTX to process a single LDR capture. After TensorRT optimization and network pruning our network takes 85 ms with fp32 precision and 44 ms with fp16 precision on the same GPU.

7. Grating Optics In-the-Wild

This section explores reconstruction without a designed optic, but with grating-like optics in the wild. As such, front-facing automotive cameras suffer from glare induced by thin lines of dust and dirt remaining on the windshield after wiping [2], see Fig. 7. These thin streaks of dust are oriented perpendicular to the windshield wiping orientation on a typically curved windshield. As a result, they scatter light along streaks with varying orientation, which negatively impacts the imaging systems of autonomous vehicles during night time driving. Removing these streaks could

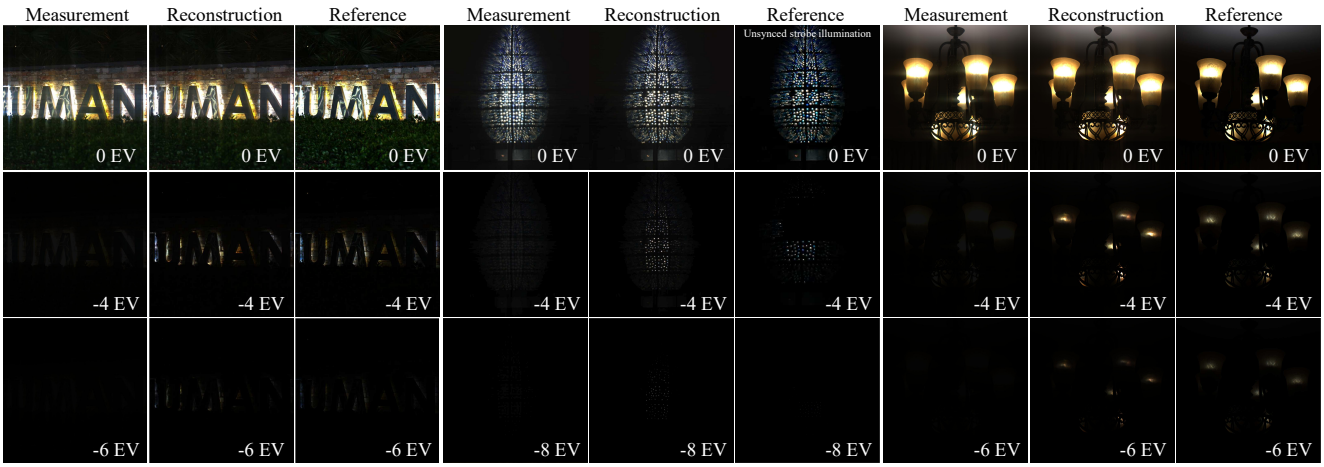

Figure 6: Real-world captures using fabricated DOE and reconstruction results. Note that the middle image is of a strobe light array with 50 Hz frequency. The reference images of -4 EV, -6 EV and -8 EV are taken by the same camera without the DOE (same aperture and position) by reducing the exposure time to $1/2^4$, $1/2^6$ and $1/2^8$ respectively. Please zoom in to view image details.

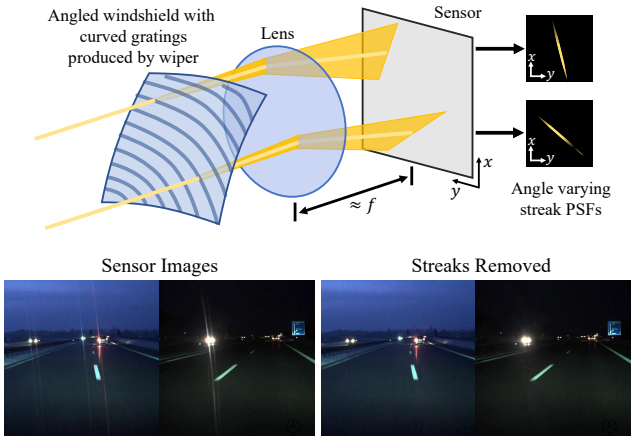

Figure 7: Automotive streaks are caused by grating-like patterns on the windshield. Applying our residual splitting network to the corresponding PSFs allows us to remove them.

improve performance for display applications, such as digital mirrors, as well as downstream computer vision tasks.

Although the PSFs corresponding to these streaks are different from our learned PSF, we can still apply our residual splitting network for removing these streaks. To demonstrate this, we collected several night time driving video sequences. We modeled the streak PSF in these videos using a 2-point star PSF, and we trained the residual splitting network using the same dataset from Section 4.3 and the unsaturated loss from Eq 13. To account for variations in the rotation angle of the 2-point star PSF, we uniformly sampled the rotation angle within $(-8, -2.5) \cup (2.5, 8)$ radians where 0 radians refers to the 2-point star PSF being parallel to a vertical line. Example snapshots along with removed glare results can be seen in Figure 7. For additional qualitative results, please refer to the Supplemental Material.

8. Discussion

Limitations Like other optical encoding methods, our method requires that the encoding streaks themselves are not saturated. While we can ensure this for applications where small, saturated regions are expected (e.g night time driving and indoor navigation) our method does struggle with larger saturated regions. Please refer to the Supplemental Matrial for further discussion and failure examples.

In the design phase, our simulation models how the DOE affects narrow RGB bands, and using a model with finer wavelength sampling would reduce the disparity between the simulated PSF and the real-world captured PSF. However, note, that such a model would be more time consuming to train, difficult to optimize, and requires a large corpus of HDR multispectral training data that does not exist today.

State-of-the-art GPUs allow us to achieve real-time latencies, requiring multiple GPUs for high sensor resolutions, but are impractical for low-power consumer applications. Porting to dedicated hardware, such as power-efficient ASICs or FPGAs, is an important next step.

Conclusion We present a novel approach tackling the challenge of estimating HDR images from single-shot LDR captures. To this end, we propose a rank-1 DOE encoding of HDR content and a catered reconstruction network, which when jointly optimized allow for snapshot HDR captures that outperform previous state-of-the-art methods. Going forwards, we envision making snapshot HDR capture truly practical by extending our optical model to handle greater scene information, such as depth and multispectral data, as well as designing our algorithms for specialized hardware for low-power processing at the edge.

Acknowledgements. This work was supported by KAUST baseline funding.

References

[1] Ahmet Ouz Akyüz, Roland W. Fleming, Bernhard E. Riecke, Erik Reinhard, and Heinrich H. Bülthoff. Do hdr displays support ldr content?: a psychophysical evaluation. In *SIGGRAPH 2007*, 2007. 2

[2] Merrill J Allen. Automobile windshields, surface deterioration. Technical report, SAE Technical Paper, 1970. 7

[3] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *GRAPHITE*, 2006. 2

[4] Ayan Chakrabarti. Learning sensor multiplexing design through back-propagation. *ArXiv*, abs/1605.07078, 2016. 3

[5] Julie Chang and Gordon Wetzstein. Deep optics for monocular depth estimation and 3d object detection. *ArXiv*, abs/1904.08601, 2019. 3

[6] Arnaud Darmont and Society of Photo-optical Instrumentation Engineers. High dynamic range imaging: sensors and architectures. SPIE Washington, 2012. 1

[7] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *SIGGRAPH '08*, 1997. 2

[8] Piotr Didyk, Rafal Mantiuk, Matthias Hein, and Hans-Peter Seidel. Enhancement of bright video features for hdr displays. *Comput. Graph. Forum*, 27:1265–1274, 2008. 2

[9] 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. 1, 2, 6, 7

[10] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of SIGGRAPH Asia)*, 36(6):177, 2017. 2

[11] Robert Edward Fischer, Biljana Tadic-Galeb, Paul R Yoder, Ranko Galeb, Bernard C Kress, Stephen C McClain, Tom Baur, Richard Plympton, Bob Wiederhold, and Bob Grant Alastair J. *Optical system design*, volume 1. Citeseer, 2000. 6

[12] Konstantina Fotiadou, Grigorios Tsagkatakis, and Panagiotis Tsakalides. Snapshot high dynamic range imaging via sparse representations and feature learning. *IEEE Transactions on Multimedia*, 2019. 1, 2

[13] Orazio Gallo, Natasha Gelfandz, Wei-Chao Chen, Marius Tico, and Kari Pulli. Artifact-free high dynamic range imaging. *2009 IEEE International Conference on Computational Photography (ICCP)*, pages 1–7, 2009. 2

[14] Miguel Granados, Kwang In Kim, James Tompkin, and Christian Theobalt. Automatic noise modeling for ghost-free hdr reconstruction. *ACM Trans. Graph.*, 32:201:1–201:10, 2013. 2

[15] Michael D. Grossberg and Shree K. Nayar. High dynamic range from multiple images: Which exposures to combine? 2003. 2

[16] William Guicquero, Antoine Dupret, and Pierre Vandergheynst. An algorithm architecture co-design for cmos compressive high dynamic range imaging. *IEEE Transactions on Computational Imaging*, 2:190–203, 2016. 2

[17] Harel Haim, Shay Elmalem, Raja Giryes, Alex Bronstein, and Emanuel Marom. Depth estimation from a single image using deep learned phase coded mask. *IEEE Transactions on Computational Imaging*, 4:298–310, 2018. 3

[18] Saghi Hajisharif, Joel Kronander, and Jonas Unger. Adaptive dualiso hdr reconstruction. *EURASIP Journal on Image and Video Processing*, 2015:1–13, 2015. 2

[19] Samuel W. Hasinoff, Frédo Durand, and William T. Freeman. Noise-optimal capture for high dynamic range photography. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560, 2010. 2

[20] 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. 1, 2

[21] Lei He, Guanghui Wang, and Zhanyi Hu. Learning depth from single images with deep neural network embedding focal length. *IEEE Transactions on Image Processing*, 27:4676–4689, 2018. 3

[22] Felix Heide, Qiang Fu, Yifan Peng, and Wolfgang Heidrich. Encoded diffractive optics for full-spectrum computational imaging. *Scientific reports*, 6:33543, 2016. 6

[23] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (TOG)*, 33(6):231, 2014. 1, 2

[24] Keigo Hirakawa and Paul M. Simon. Single-shot high dynamic range imaging with conventional camera hardware. *2011 International Conference on Computer Vision*, pages 1339–1346, 2011. 2

[25] Roarke Horstmeyer, Richard Y. Chen, Barbara Kappes, and Benjamin Judkewitz. Convolutional neural networks that teach microscopes how to image. *ArXiv*, abs/1709.07223, 2017. 3

[26] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1163–1170, 2013. 2

[27] Hanbyol Jang, Kihun Bang, Jinseong Jang, and Dosik Hwang. Inverse tone mapping operator using sequential deep neural networks based on the human visual system. *IEEE Access*, 6:52058–52072, 2018. 2

[28] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36:144:1–144:12, 2017. 2

[29] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. *Comput. Graph. Forum*, 38:193–205, 2019. 2

[30] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B. Goldman, and Pradeep Sen. Patch-based high dynamic range video. *ACM Trans. Graph.*, 32:202:1–202:8, 2013. 2

[31] Sing Bing Kang, Matthew Uyttendaele, Simon A. J. Winder, and Richard Szeliski. High dynamic range video. *ACM Trans. Graph.*, 22:319–325, 2003. 2

[32] 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. 3

[33] Erum Arif Khan, Ahmet Oguz Akyüz, and Erik Reinhard. Ghost removal in high dynamic range images. *2006 International Conference on Image Processing*, pages 2005–2008, 2006. 2

[34] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep chain hdri: Reconstructing a high dynamic range image from a single low dynamic range image. *IEEE Access*, 6:49913–49924, 2018. 2

[35] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *The European Conference on Computer Vision (ECCV)*, September 2018. 2

[36] Ce Liu. Exploring new representations and applications for motion analysis. 2009. 2

[37] Steve Mann and Rosalind W. Picard. Being 'undigital' with digital cameras: extending dynamic range by combining differently exposed pictures. 1994. 2

[38] Julio Marco, Quercus Hernandez, Adolfo Muñoz, Yue Dong, Adrián 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.*, 36:219:1–219:12, 2017. 3

[39] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. *CoRR*, abs/1803.02266, 2018. 2

[40] Belén Masiá, Sandra Agustin, Roland W. Fleming, Olga Sorkine-Hornung, and Diego Gutierrez. Evaluation of reverse tone mapping through varying exposure conditions. *ACM Trans. Graph.*, 28:160, 2009. 2

[41] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. *Comput. Graph. Forum*, 28:161–171, 2009. 2

[42] Tom Mertens, Jan Kautz, and Frank Van Reeth. Exposure fusion: A simple and practical alternative to high dynamic range photography. In *Computer graphics forum*, volume 28, pages 161–171. Wiley Online Library, 2009. 1

[43] Christopher A Metzler, Hayato Ikoma, Yifan Peng, and Gordon Wetzstein. Deep optics for single-shot high-dynamic-range imaging. *arXiv preprint arXiv:1908.00620*, 2019. 1, 3, 6, 7

[44] Laurence Meylan, Scott J. Daly, and Sabine Süsstrunk. The reproduction of specular highlights on high dynamic range displays. In *Color Imaging Conference*, 2006. 2

[45] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018. 1, 2

[46] Atsushi Morimitsu, Isao Hirota, Sozo Yokogawa, Isao Ohdaira, Masao Matsumura, Hiroaki Takahashi, Toshio Yamazaki, Hideki Oyaizu, Yalcin Incesu, Muhammad Atif, et al. A 4m pixel full-pdaf cmos image sensor with 1.58 μm 2$\times$ 1 on-chip micro-split-lens technology. In *ITE Technical Report 39.35*, pages 5–8. The Institute of Image Information and Television Engineers, 2015. 1

[47] Kenta Moriwaki, Ryota Yoshihashi, Rei Kawakami, Shaodi You, and Takeshi Naemura. Hybrid loss for learning single-image-based HDR reconstruction. *arXiv preprint arXiv:1812.07134*, 2018. 2

[48] Shree K Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 1, pages 472–479. IEEE, 2000. 1, 2

[49] Elias Nehme, Daniel Freedman, Racheli Gordon, Boris Ferdman, Tomer Michaeli, and Yoav Shechtman. Dense three dimensional localization microscopy by deep learning. 2019. 3

[50] Shiyu Ning, Hongteng Xu, Li Song, Rong Xie, and Wenjun Zhang. Learning an inverse tone mapping network with a generative adversarial regularizer. *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1383–1387, 2018. 2

[51] Yifan Peng, Qiang Fu, Hadi Amata, Shuochen Su, Felix Heide, and Wolfgang Heidrich. Computational imaging using lightweight diffractive-refractive optics. *Optics express*, 23(24):31393–31407, 2015. 6

[52] Erik Reinhard, Greg Ward, Summant Pattanaik, Paul E. Debevec, Wolfgang Heidrich, and Karol Myszkowski. High dynamic range imaging: Acquisition, display, and image-based lighting. 2010. 2

[53] Allan G. Rempel, Matthew Trentacoste, Helge Seetzen, H. David Young, Wolfgang Heidrich, Lorne Arthur Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. In *SIGGRAPH 2007*, 2007. 2

[54] 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. 4

[55] Mushfiqur Rouf, Rafal Mantiuk, Wolfgang Heidrich, Matthew Trentacoste, and Cheryl Lau. Glare encoding of high dynamic range images. *CVPR 2011*, pages 289–296, 2011. 1, 2, 3, 6, 7

[56] U Seger. Hdr imaging in automotive applications. In *High Dynamic Range Video*, pages 477–498. Elsevier, 2016. 1

[57] 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.*, 31:203:1–203:11, 2012. 2

[58] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belén Masiá. Convolutional sparse coding for high dynamic range imaging. *Comput. Graph. Forum*, 35:153–163, 2016. 2

[59] Yoav Shechtman, Lucien E Weiss, Adam S. Backer, Maurice Y. Lee, and W E Moerner. Multicolour localization microscopy by point-spread-function engineering. *Nature photonics*, 10:590–594, 2016. 3

[60] 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. 1, 3

[61] Shuochen Su, Felix Heide, Gordon Wetzstein, and Wolfgang Heidrich. Deep end-to-end time-of-flight imaging. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6383–6392, 2018. 3

[62] Michael D Tocci, Chris Kiser, Nora Tocci, and Pradeep Sen. A versatile hdr video production system. In *ACM Transactions on Graphics (TOG)*, volume 30, page 41. ACM, 2011. 1

[63] Chao Wang, Yang Zhao, and Ronggang Wang. Deep inverse tone mapping for compressed images. *IEEE Access*, 7:74558–74569, 2019. 2

[64] Trygve Willassen, Johannes Solhusvik, Robert Johansson, Sohrab Yaghmai, Howard Rhodes, Sohei Manabe, Duli Mao, Zhiqiang Lin, Dajiang Yang, Orkun Cellek, et al. A 1280× 1080 4.2 μm split-diode pixel hdr sensor in 110 nm bsi cmos process. In *Proceedings of the International Image Sensor Workshop, Vaals, The Netherlands*, pages 8–11, 2015. 1

[65] Yicheng Wu, Vivek Boominathan, Huaijin Chen, Aswin Sankaranarayanan, and Ashok Veeraraghavan. Phasecam3d learning phase masks for passive single view depth estimation. *2019 IEEE International Conference on Computational Photography (ICCP)*, pages 1–12, 2019. 3

[66] Jinsong Zhang and Jean-François Lalonde. Learning high dynamic range from outdoor panoramas. *CoRR*, abs/1703.10200, 2017. 2

[67] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 4, 5

[68] Hang Zhao, Boxin Shi, Christy Fernandez-Cull, Sai-Kit Yeung, and Ramesh Raskar. Unbounded high dynamic range photography using a modulo camera. *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10, 2015. 2