Single-Photon Camera Guided Extreme Dynamic Range Imaging

Yuhao Liu Felipe Gutierrez-Barragan Atul Ingle Mohit Gupta Andreas Velten
{liu697,fgutierrez3,ingle,mgupta37,velten}@wisc.edu
University of Wisconsin-Madison

Abstract

Reconstruction of high-resolution extreme dynamic range images from a small number of low dynamic range (LDR) images is crucial for many computer vision applications. Current high dynamic range (HDR) cameras based on CMOS image sensor technology rely on multi-exposure bracketing which suffers from motion artifacts and signal-to-noise (SNR) dip artifacts in extreme dynamic range scenes. Recently, single-photon cameras (SPCs) have been shown to achieve orders of magnitude higher dynamic range for passive imaging than conventional CMOS sensors. SPCs are becoming increasingly available commercially, even in some consumer devices. Unfortunately, current SPCs suffer from low spatial resolution. To overcome the limitations of CMOS and SPC sensors, we propose a learning-based CMOS-SPC fusion method to recover high-resolution extreme dynamic range images. We compare the performance of our method against various traditional and state-of-the-art baselines using both synthetic and experimental data. Our method outperforms these baselines, both in terms of visual quality and quantitative metrics.

1. Introduction

Emerging computer vision applications require imaging systems capable of capturing brightness levels with extreme dynamic range (DR), where the brightest point in the image can be more than 6 orders of magnitude brighter than the dimmest point [52]. Fig. 1 shows an example of such extreme dynamic range scenario. Unfortunately, conventional image sensors, based on charged-coupled device (CCD) or complementary semiconductor metal oxide (CMOS) technology, have a limited DR. Various computational and hardware approaches have been developed over several decades to deal with this limitation [53, 4, 42], and continues to be an active area of research [51].

Exposure bracketing, where a sequence of images with different exposure times are fused into a single high dynamic range (HDR) image [11, 40], is one of the most widely used approaches. However, in dynamic applications this technique can lead to ghosting [60] and light flicker

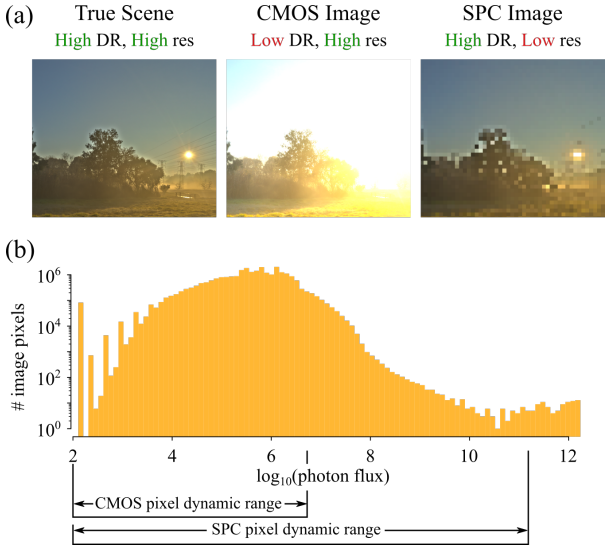

Figure 1. **Need for extreme dynamic range (DR):** (a) Many real world scenes span a wide range of brightness levels. CMOS cameras provide a high spatial resolution but limited dynamic range. A single-photon camera (SPC) has extreme dynamic range but not enough spatial resolution. (b) A histogram showing true scene pixel values from (a) that span photon flux levels over 10 orders of magnitude from 10^2 to 10^{12} photons/s. Fusing CMOS and SPC images can potentially provide extreme dynamic range while simultaneously providing high spatial resolution.

[61, 28, 13] artifacts. To mitigate these artifacts, commercial HDR sensors are limited to fusing only a few (2–4) exposures acquired through sequential capture [61] or with dual-pixel architectures [28, 2]. Recovering an extreme dynamic range image from only a few exposure stops results in spatially non-uniform signal-to-noise-ratio (SNR) dip artifacts [65, 44, 61] throughout the image. Large SNR dips are a challenge because fine image features can end up buried in the noise, which may be hard to denoise. Overall, spatially non-uniform SNR drops are a fundamental limitation of exposure bracketing in extreme DR scenarios when only a small number of exposures are available.

More recently, single-photon cameras (SPCs) based on single-photon avalanche diode (SPAD) detector technology have gained popularity for various image sensing applica-

tions [8, 7, 45, 35, 67, 23, 38]. Their extreme sensitivity to light down to individual photons and high timing resolution have been exploited to achieve extreme dynamic range from a *single capture* [30, 29]. These demonstrations, however, have been limited to either single-pixel detectors or very low-resolution SPAD arrays. Recently, the first megapixel SPAD arrays have been demonstrated [45]. Unfortunately, the per-pixel bit-depth of these sensors is only 1-bit, which means that thousands of binary frames need to be read off the sensor to reconstruct a single image. This design has prohibitively high power consumption, long acquisition times, and current algorithms are based on offline processing [39, 56]. Fortunately, multi-bit-depth SPAD arrays are becoming available, albeit, at lower spatial resolutions [14, 48]. Therefore, for the time being, users of SPAD arrays must resort to mechanical scanning techniques [30] or computational super-resolution methods [9] to increase the spatial resolution of captured images.

In this work we propose a learning-based sensor fusion approach that uses high-resolution, low dynamic range (LDR) information captured by a conventional CMOS camera and low-resolution but extremely high dynamic range image captured by a SPC to reconstruct a high spatial resolution and extreme dynamic range image (Fig. 1(a)). Our work is motivated by the observation that such multi-camera modules consisting of CMOS image sensors co-located with an SPC are already commercially available [66]. We show that our method of fusing a single SPC image and a single CMOS camera image can outperform dual-exposure bracketing fusion methods that rely on two images, especially in situations where the dynamic range is too large to be covered by two CMOS exposures (Fig. 1(b)).

2. Related Work

Multi-image HDR Imaging: Conventional HDR imaging methods capture multiple LDR images of the scene with different exposures and merge them into a single HDR image [11, 25, 32]. Although these methods work well for static scenes, they suffer from "ghosting" artifacts caused by motion [60]. This can be alleviated using spatially varying exposure image sensors [46, 28, 2], but introduces additional hardware complexity if more than two exposures are needed to cover the dynamic range. We show that by using just two image sensors (a CMOS and an SPC) we can capture extremely DR content, beyond the capability of conventional methods including CMOS-CMOS fusion. Although our method is restricted to static scenes with extreme DR, it is complementary to recent burst photography HDR methods [26] that can compensate for motion.

Single-image HDR Imaging: State-of-the-art methods for single-image HDR use deep learning techniques to recover saturated regions from a single CMOS image [15, 55, 43, 37]. These methods perform quite well when the image contains only a few overexposed regions. However, in the case of extreme dynamic range scenes where large regions of the scene are overexposed, these methods can introduce significant hallucination artifacts, which are not appropriate in safety-critical applications. Here we show that providing an additional source of information in the form of a low resolution but high dynamic range SPC image, enables reconstructing extremely bright and saturated regions that conventional single-image HDR methods struggle to recover.

Other Emerging Sensors: Recently, event-based vision sensors have been used in conjunction with a CMOS image sensor [24, 63] for HDR imaging. Unlike an event-camera that only captures changes in brightness, our method uses an SPC that directly captures scene intensity with extreme DR. Quanta image sensors (QIS) [16] are also sensitive down to individual photons and can provide much higher dynamic range than conventional CMOS cameras [20]. Nonetheless, due to the lack of precise timing information, the dynamic range achievable by the QIS is still lower than what could be achieved with a SPAD-based SPC [29]. Additionally, there is already an industry trend towards adopting SPAD-based SPC's in commercial devices [66], mainly because of their additional application to LiDAR imaging [34].

3. Image Formation Model

The photon irradiance received at an image sensor pixel is proportional to the true brightness (radiance) of the scene point [58]. Assuming a fixed pixel size the photon irradiance is converted to total incident photon flux (photons per second) which can be used as a proxy for scene brightness. We therefore use the photon flux and brightness interchangeably in this paper. Consider a fixed scene point with a brightness of Φ photons/second. The response curve of the image sensor pixel determines the relationship between the incident photon flux and the pixel output. This response curve is an intrinsic property of the pixel and is different for a conventional CMOS camera pixel and an SPC pixel.

3.1. CMOS Response Function

A conventional CMOS camera pixel has a linear response curve where the photoelectric charge accumulated in the pixel is directly proportional to the incident photon flux Φ. Camera manufacturers often apply a proprietary non-linear compression curve called the camera response function (CRF) to the raw pixel measurement. We assume that the camera gives access[1] to the raw (linear) pixel values directly, where the pixel output, N_T^{CMOS}, is a linear function of Φ. The average number of photoelectrons accumulated in a CMOS pixel over an exposure time, T, is given by:

$$\mathsf{E}[N_T^{\text{CMOS}}] = q_{\text{CMOS}}\Phi T \tag{1}$$

[1] If raw linear pixel values are not available, existing CRF estimation methods can be used to linearize them [22, 3].

and the variance due to Poisson noise is given by:

$$\mathsf{Var}[N_T^{\text{CMOS}}] = q_{\text{CMOS}}^2 \Phi^2 T^2 \tag{2}$$

where $0 < q_{\text{CMOS}} < 1$ is the pixel sensitivity. Thanks to recent advances in CMOS technology, electronic read noise sources are approaching or achieving sub-electron levels in normal illumination conditions [10], making them negligible in the high-flux regime studied in this paper and are therefore ignored. We use a Gaussian approximation and assume that each CMOS pixel generates an output $\widehat{N}_T^{\text{CMOS}}$ that follows a normal distribution with mean and variance given by Eqs. (1) and (2), and rounded to the nearest integer. Additionally, we impose a full well capacity limit such that $\widehat{N}_T^{\text{CMOS}}$ is clamped at a maximum of N_{FWC}. We estimate the incident per-pixel photon flux using:

$$\widehat{\Phi}^{\text{CMOS}} = \frac{\widehat{N}_T^{\text{CMOS}}}{q_{\text{CMOS}} T} \tag{3}$$

provided the pixel is not saturated, i.e., $\widehat{N}_T^{\text{CMOS}} < N_{\text{FWC}}$.

3.2. SPC Response Function

We assume that each pixel in our SPC is a passive free-running SPAD—after each photon detection event, the SPAD enters a dead-time during which it does not capture any photons. Assuming that the photons incident on each pixel follow Poisson statistics, the number of detected photons (N_T^{SPC}) over a fixed exposure time (T) follows a renewal process [21] with mean and variance given by [30]:

$$\mathsf{E}[N_T^{\text{SPC}}] = \frac{q_{\text{SPAD}} \Phi T}{1 + q_{\text{SPAD}} \Phi \tau_{\text{d}}} \tag{4}$$

$$\mathsf{Var}[N_T^{\text{SPC}}] = \frac{q_{\text{SPAD}} \Phi T}{(1 + q_{\text{SPAD}} \Phi \tau_{\text{d}})^3} \tag{5}$$

where $0 < q_{\text{SPAD}} < 1$ is the pixel sensitivity, and τ_{d} is the dead-time. We use a Gaussian approximation and assume that each pixel generates normally distributed photon counts with the same mean and variance as Eqs. (4) and (5), and rounded to the closest integer [30]. From the measured photon counts, $\widehat{N}_T^{\text{SPC}}$, we estimate the per-pixel brightness by inverting Eq. (4):

$$\widehat{\Phi}^{\text{SPC}} = \frac{\widehat{N}_T^{\text{SPC}} / q_{\text{SPAD}}}{T - \tau_{\text{d}} \widehat{N}_T^{\text{SPC}}}. \tag{6}$$

Note that due to its dead-time, the SPAD pixel's estimated brightness in Eq. (6) is a non-linear function of the measured photon counts. We apply this linearization step to each pixel's photon counts in our SPC simulations.

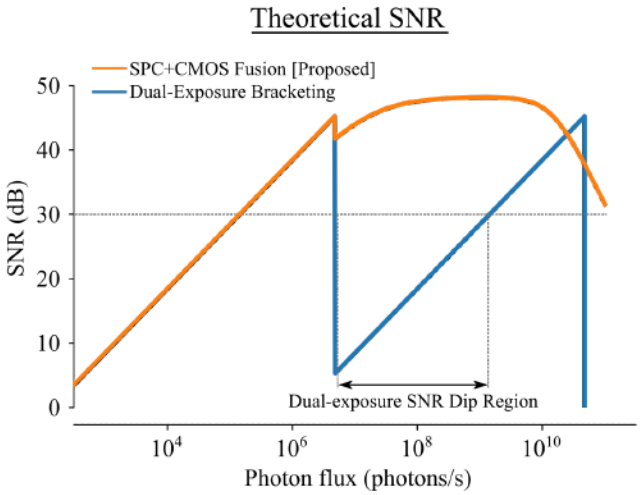

Figure 2. **SNR dip problem of dual-exposure bracketing:** To cover an extreme DR with brightness levels spanning many orders of magnitude, dual-exposure bracketing requires two widely spaced exposure times. This large exposure difference causes a drop in the per-pixel SNR as seen in this theoretical SNR plot. CMOS+SPC fusion does not experience this SNR dip. The fusion assumes a 10 ms exposure for each sensor, whereas the dual-exposure bracketing uses 10 ms and 1 µs exposure times for the configuration shown in this figure.

4. SPC-Guided Extreme HDR Imaging

The dynamic range of SPC's is sufficient for many applications. Similarly, the spatial resolution of CMOS sensors is also sufficient. Therefore, a practical solution to the limited resolution of SPC's and limited dynamic range of CMOS sensors is the fusion of these two sensors. One advantage of this design is that it prevents SNR dip artifacts, because at extreme brightness levels SPC's can sustain high SNR by using sufficiently long (but practical) exposure times. This results in a photon flux vs. SNR curve[2] with a minimal dip, as shown in Figure 2. Moreover, existing commercial devices such as the recent iPhone 12 Pro, already incorporate a high-resolution CMOS and Sony's 30-kilo-pixel SPC [66, 1], suggesting that this is a practical design. In this section we present our proposed model to expand the dynamic range of a CMOS image with an SPC image to reconstruct an extreme HDR image.

4.1. SPC-Guided HDR Network

Similar to previous learning-based HDR models we adopt a U-net architecture [54]. Different from traditional U-nets, our design uses two encoders that extract features from the CMOS and SPC images, as shown in Figure 3. The network architecture shown here assumes the SPC image is

[2] $\text{SNR} := 20 \log_{10} \left(\Phi / \sqrt{\mathsf{E}[(\widehat{\Phi} - \Phi)^2]} \right)$. Closed form expressions for this plot are taken from [30].

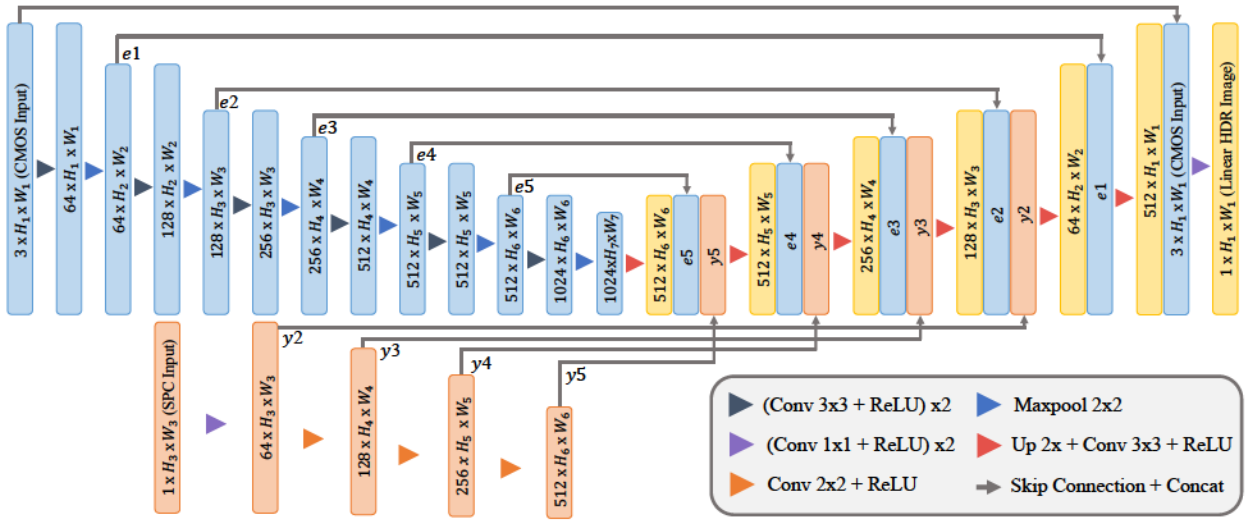

Figure 3. **SPC-Guided HDR network architecture:** A block diagram of the dual-encoder U-net implemented in this paper is shown here. The encoders (blue and orange blocks) sequentially filter and downsample the input CMOS and SPC images by $2\times$ (e.g., $H_2 = \frac{H_1}{2}$, $H_3 = \frac{H_1}{4}$) to extract multi-scale features. The sensor fusion decoder further filters and up-samples the feature maps. The last layer applies a blending operation of the input CMOS image and the learned up-sampling of the SPC image.

$4\times$ smaller in resolution than the CMOS image. Performance limits with other resolution factors will be explored in future work. Our model operates on the linearized brightness images, requiring a pre-processing step that applies Eqs. (3) and (6) to the CMOS and SPC pixel outputs.

Sensor Fusion Decoder: The goal of the decoder is to reconstruct the HDR image given the multi-scale features extracted from both image sources. The long-skip connections from the CMOS encoder help preserve the high-resolution spatial information available in the non-saturated regions. The long-skip connections from the SPC encoder provide low-resolution spatial information of the full image, and in particular, of the saturated regions from the CMOS image, which guides the decoder to upsample the CMOS saturated regions with extended dynamic range. Moreover, the deep CMOS-SPC features provide the decoder contextual information which helps hallucinate the higher-resolution details of the saturated regions that may not be available in the SPC image. In Section 6.2, we show that despite our simple network design, our model performs comparably to other more complex architectures that use attention gates [49, 24].

4.2. Loss Functions

The proposed SPC-guided HDR network with parameters θ, reconstructs a linear photon flux HDR image:

$$\widehat{\Phi}^{\text{Fused}} = F_\theta(\widehat{\Phi}^{\text{CMOS}}, \widehat{\Phi}^{\text{SPC}}). \tag{7}$$

As discussed in previous data-driven HDR reconstruction works [24, 32, 15], computing a loss directly on the linear HDR values results in the loss function being dominated by the larger pixel values. Therefore, we adopt a similar strategy where we compute the loss functions on the tone-mapped domain. We use μ-compression, proposed in [32], as the differentiable tone-mapping operator:

$$\widehat{\Phi}^{\text{Fused}}_\mu = \frac{\log(1+\mu\widehat{\Phi}^{\text{Fused}})}{\log(1+\mu)}. \tag{8}$$

For all the models presented in this paper we set $\mu = 2000$.

The loss function used to train our proposed SPC-guided HDR network is composed of a pixel-wise loss and a perceptual loss [31]. The pixel-wise loss is the ℓ^1 distance between the tone-mapped output and tone-mapped target images:

$$L_{\ell^1} = \left\| \Phi_\mu - \widehat{\Phi}^{\text{Fused}}_\mu \right\|_1 \tag{9}$$

where Φ is the true photon flux image. The perceptual loss is computed using a pre-trained VGG-19 [57] in PyTorch [59, 50] as follows:

$$L_{\text{vgg}} = \sum_{i=1}^{N} \omega_i \left\| g_i(\Phi_\mu) - g_i(\widehat{\Phi}^{\text{Fused}}_\mu) \right\|_1 \tag{10}$$

where $g_i(\cdot)$ is the ith layer output of the VGG model, and ω_i is a constant weighing factor that assigns larger weights to deeper layers. Putting Eqs. (9) and (10) together we get our loss function:

$$L = L_{\ell^1} + \alpha L_{\text{vgg}} \tag{11}$$

where $\alpha = 0.1$.

5. Datasets and Implementation

In this section, we introduce our simulator for CMOS and SPC images, and the HDR datasets used for training and testing our model. We also describe the real data used for evaluation. Lastly, we discuss the implementation details of the proposed model.

5.1. Simulator and Datasets

Since SPCs are an emerging technology, there are no readily available real-world datasets. To overcome this challenge we implemented a simulation pipeline that leverages existing HDR image datasets to generate a large-scale paired CMOS+SPC image dataset. Moreover, current SPC sensors are monochrome, so for the remainder of this paper we restrict our analysis to monochrome images.

Simulator: The simulators take as input ground truth photon flux images, Φ.

- **CMOS Simulation:** The CMOS image is simulated from Φ using the Gaussian approximation described in Section 3.1 with the pixel sensitivity and exposure parameters set to $q_{\text{CMOS}} = 0.75$ and $T = 0.01$s. Pixel outputs are clipped at $N_{\text{FWC}} = 33400$. Thus, the final simulated CMOS images contain linear digitized pixel intensities (i.e., $\widehat{N}_T^{\text{CMOS}}$), with approximately 15 bits per pixel.
- **SPC Simulation:** We begin by spatially downsampling the ground-truth Φ by $4\times$ using OpenCV's `cv2.INTER_AREA` interpolation [5]. The SPC image is simulated from Φ using the Gaussian approximation (Sec. 3.2) with $q_{\text{SPAD}} = 0.25, T = 0.01$s, $\tau_{\text{d}} = 150$ns. The simulated SPC image contains the photon counts measured by each pixel ($\widehat{N}_T^{\text{SPC}}$).

Synthetic Dataset: We gathered a total of 667 high-resolution HDR images from Poly Haven [27] (469, 4096×2048), Laval et al. [19] (93, 2048×1024), and Funt et al. [17, 18] (105, 2142×1422). For each dataset, we analyzed the distribution of its irradiance values to determine an appropriate scaling factor that would make the distribution span a wide range of realistic photon flux values [64] (please see Sec. S. 4 for details). For the models that require monochrome inputs, we dropped the R/B color channels.

Real Images: We use the CMOS-SPC image pairs acquired in [30]. Unfortunately, the images are not aligned, the physical distance between the sensors is significant, and each sensor is subject to its own optical parameters (e.g., focal length and aberrations). To alleviate these limitations, we manually select small overlapping crops from each image and register them by estimating an affine transformation using MATLAB's `imregtform` function. The approximately aligned crops shown in the first and second rows of Figure 8 are 84×73 and 71×71, respectively. Finally, we bilinearly re-sample the CMOS and SPC crops such that their dimensions are 256×256 and 64×64.

5.2. Training and Implementation

Data Pre-processing: To guarantee that the CMOS and SPC images have similar distributions in non-saturated regions, we estimate photon flux from the pixel intensities using Eqs. (3) and (6). We normalize the CMOS ($\widehat{\Phi}^{\text{CMOS}}$), SPC ($\widehat{\Phi}^{\text{SPC}}$), and the ground truth (Φ) images by dividing by the CMOS photon flux saturation limit (i.e., N_{FWC}/T), and multiplying by 255.

Augmentation & Patch Selection: During each training step, we randomly select patches from the CMOS and SPC images of size 512×256 and 128×64, respectively. To promote a balanced dataset that contains sufficient examples of saturated CMOS image regions, when selecting the random patch, we sample 10 patches and select the patch where at least 10% of the pixels are saturated. If none of them satisfy this criteria, we simply return one of them. We found that this patch selection strategy prevents the network from only learning to output a copy of the CMOS image. Finally, we apply a random horizontal and vertical flip to the patch.

Network Architecture: Figure 3 shows the detailed dual-encoder U-net architecture. To avoid checkerboard artifacts, we use the resize-convolution upsampling operator [47] in the sensor fusion decoder.

Training Details: We split our synthetic dataset into training, validation, and testing subsets. The training and validation sets are composed of the simulated images from [27, 19] using an 80/20 split. The test set is Funt et al. HDR dataset [17]. The weights of the CMOS encoder are initialized to pre-trained VGG-16 weights and the SPC encoder and sensor fusion decoder use PyTorch's default initialization. We train all models using the ADAM optimizer [33] with its default parameters and a batch size of 16. We train the model for 2000 epochs using a multi-step learning rate schedule where the learning rate starts at 10^{-3}, and every 500 epochs it is reduced by a factor of 0.8.

6. Experiments and Results

6.1. Baselines

To illustrate the need of both SPC and CMOS images for extreme HDR, we compare against the following baselines:

- **DHDR [55]**: State-of-the-art single-image HDR imaging model trained on a larger dataset.
- **ExpandNet [43]:** A single-image HDR CNN model.
- **ESRGAN [62]:** A single-image super-resolution model based on generative adversarial networks.
- **Laplacian Blending [6]:** Algorithmic blending of an ESRGAN super-resolved SPC image and a CMOS image.

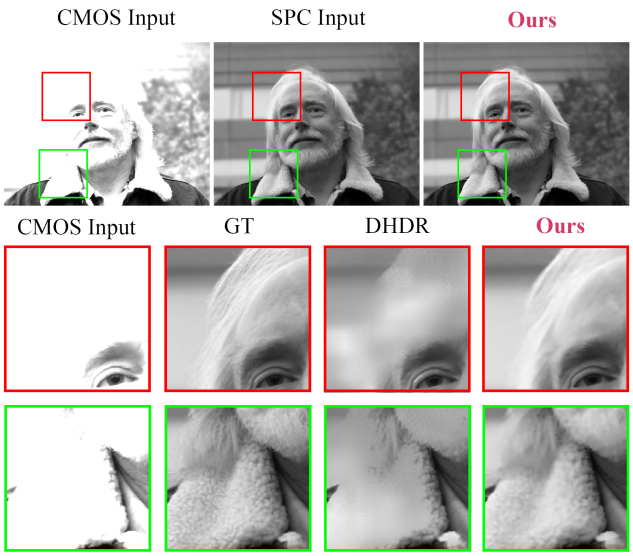

Figure 4. **Importance of low-res. HDR information:** As the area of over-exposed regions increases, single-image HDR networks like DHDR [55], can no longer recover details in these regions due to the lack of contextual information. DHDR produces images with missing texture (e.g., hair in red crop) or incorrect texture (e.g., beard in green crop). In contrast, our method uses low-resolution HDR information to reproduce plausible texture details that closely resemble the physical appearance of the subjects, even in extreme dynamic range.

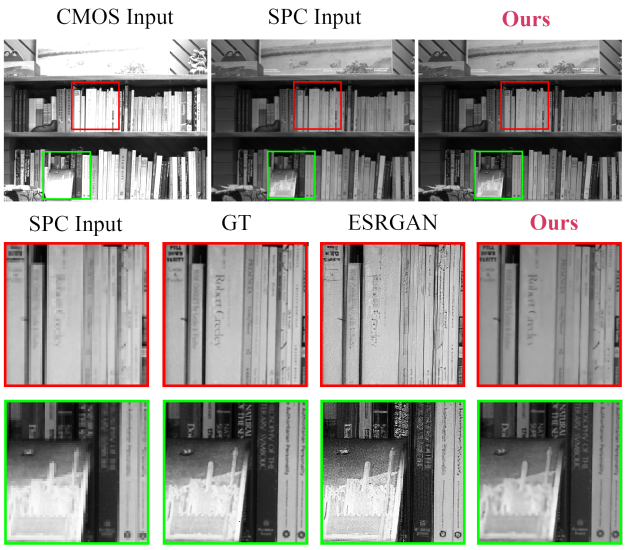

Figure 5. **Importance of high-res. LDR information:** Recent single-image super-resolution methods like ESRGAN [36] can be used to super-resolve the low-res HDR SPC image. Although this helps recover certain fine details like the book boundaries, it also introduces non-existent high-frequency textures (green crop). Moreover, it fails to recover structured details, such as text. Our method not only avoids these artifacts by using the information from the high-res CMOS image, but also super-resolve regions where the CMOS input is saturated.

- **Dual-Exposure Bracketing [12]:** Fusion of two CMOS images with a short and a long exposure (0.001ms and 10ms), using a last-sample-before-saturation approach.

DHDR and ExpandNet take 8-bit input images, therefore, we scale the estimated CMOS photon flux images ($\widehat{\Phi}^{\text{CMOS}}$) to [0,1], apply gamma-compression ($\gamma = 0.5$), and re-scale to an 8-bit image. Since DHDR and ExpandNet are trained with RGB data, they rely on inter-channel information, therefore, we do not drop the R/B channels for these models. ESRGAN takes as input tone-mapped images scaled between $[0, 1]$, so we first apply μ-compression to the estimated SPC photon flux images ($\widehat{\Phi}^{\text{SPC}}$) and then scale to $[0, 1]$. Finally, for the output images of DHDR, ExpandNet, and ESRGAN, we invert the aforementioned pre-processing steps to produce the corresponding linear photon flux images. This last step is necessary to ensure that all visual comparisons use the same visualization pipeline which operates on linear photon flux images.

6.2. Synthetic Dataset Evaluation

For visual comparisons between our model and the baselines (i.e., Figures 4, 5, and 6), we tone-map all images to 16-bit PNGs using OpenCV's `TonemapDrago` function with gamma and saturation parameters set to 1.0 [5]. We apply the same exposure and contrast adjustments to the crops shown to highlight the details.

Limitations of Single-image HDR: Figure 4 compares our proposed method with a state-of-the-art single-image HDR network, DHDR [55]. In the red crop, DHDR fails to recover both the contour and texture of the forehead and hair. As seen in the green crop, not only is the cotton-like texture on the collar missing, but the hallucinated texture of the beard also falsely mimics the pattern found on the edge of the collar. These hallucinated image segments are not acceptable in safety-critical applications. Figure 4 suggests that single-image HDR methods are unable to recover extreme DR images because of insufficient contextual information in the saturated regions that these models can use for in-painting, resulting in image patches that either lack texture or contain textures that deviate from the ground truth. In contrast, our proposed method uses the true low-resolution HDR information from the SPC sensor to guide the dynamic range extension, producing visually pleasing images consistent with the ground truth.

Limitations of Single-image Super-resolution: Figure 5 compares our method with ESRGAN [62], a recent single-image super-resolution model. At first glance, ESRGAN produces sharp, high-contrast images, but it does so at the cost of introducing non-existent high-frequency patterns and textures. For instance, ESRGAN introduces artificial film grain-like texture on the metal plate (green crops). More importantly, ESRGAN fails to recover structured fine

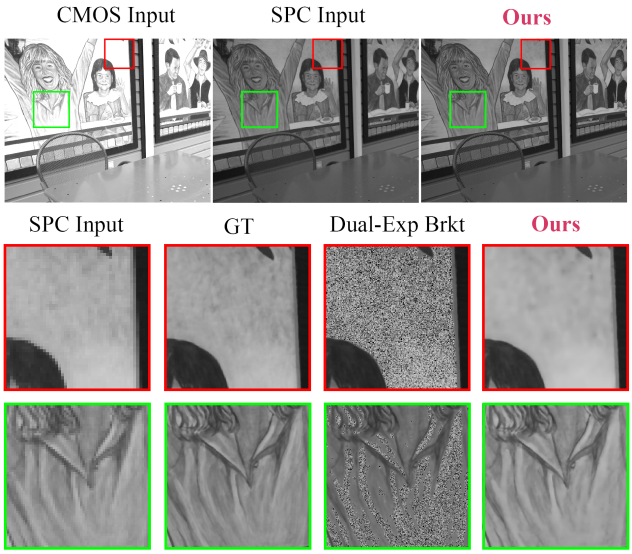

Figure 6. **Importance of spatially uniform SNR:** Merging two CMOS images at drastically different exposure times can increase dynamic range. However, in extreme dynamic range settings, dual-exposure bracketing produces image segments with discontinuous SNR levels (e.g., green crop). Denoising these regions can be challenging due to the spatial non-uniformity of the SNR dips, and in very low SNR cases, image features like the wall texture (red crop) appear extremely noisy. Our method produces clean images while maintaining image details across a wide range of brightness values.

details, such as text (red crops), which are essential features for downstream computer vision tasks. Our model uses the unsaturated high-resolution CMOS data to retain image details, such as the flower patterns at the bottom right of the green crop. Even in regions where the CMOS image is completely over-exposed, our network super-resolves the SPC image free of any hallucinated high-frequency artifacts, and effectively recovering structured details like text.

Limitations of Dual-Exposure Bracketing: Figure 6 compares our proposed method with dual-exposure bracketing [12]. For dual-exposure bracketing, an SNR dip visually translates to non-uniform regions in the merged image where the noise level suddenly increases, effectively reducing image quality. For instance, in the red crop, the smooth bright and dark spots on the wall are occluded by the noise, making these features hard to denoise. Moreover, the green crop shows an example where such SNR discontinuities can be spatially complex and fragmented, introducing additional denoising challenges. In the chosen configuration where the SPC and CMOS exposures are both 10ms, the SNR levels of the two sensors approximately match across the image, despite increases in brightness that may saturate the CMOS image. By maintaining high and uniform SNR, our method produces clean images with important details across brightness levels, suggesting that CMOS-SPC is a superior hybrid setup than CMOS-CMOS in extreme DR settings. Additional results can be found in Section S. 1.

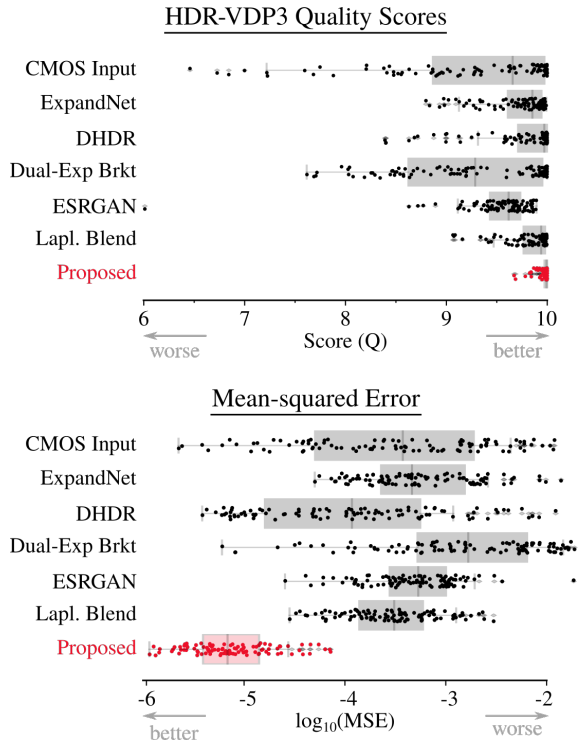

Figure 7. **Numerical quality metrics:** This table shows HDR-VDP3 quality scores [41] and MSE with respect to ground truth HDR images (computed on μ-compressed images with $\mu = 500$) for all methods. Our method outperforms state-of-the-art single-image HDR and super-resolution methods and provides the highest quality score and lowest MSE for the test set.

Quantitative Evaluations: Figure 7 shows HDR-VDP3 and mean-squared error (MSE) scores for each image in the test set (Funt et al. HDR dataset [17]). Images with very few saturated regions achieve high HDR-VDP3 scores in all methods that have CMOS as an input. Although, ExpandNet and DHDR improve the CMOS images with extremely low HDR-VDP3 scores, models with SPC inputs have fewer outliers and produce tightly distributed scores. The poor performance of dual-exposure bracketing in both metrics suggest that these metrics penalize low SNR heavier than saturation. Surprisingly, DHDR achieves comparable median HDR-VDP3 and MSE scores to Laplacian blending, despite only using a single-image. Nonetheless, Laplacian blending better prevents outliers with very poor image quality. Overall, our model consistently outperforms the evaluated baselines by a large margin.

Ablation Study: To evaluate the importance of each component of our network and inputs we compare the performance of the following ablation models: (M1) uses a single CMOS input image to generate an HDR output and relies

Figure 8. **Real-world evaluation:** We test our model on approximately aligned CMOS and SPC image crops (columns 1 and 2) from [30]. To circumvent the initial low-resolution of the CMOS image, we upsample it to make it compatible with our network (details in Section 5.1), which prevents our model from using true high-resolution CMOS information. Despite the non-idealities of the experimental data and approximate alignment, our model recovers structures like the thin lamp wire (1st row) and text (2nd row), that the baselines fail to recover. Although, using 6 exposures helps resolve SNR dips in the exposure bracketed image, it still contains over-exposed regions (1st row).

on the back-bone U-Net shown in Fig. 3. (M2) concatenates SPAD features to the decoder network. (M3) introduces attention gates in the decoder network in addition to the SPAD input of (M2). The median HDR-VDP3 quality scores for the three models computed on the validation set are 9.68, 9.99 and 9.98 and median μ-compressed MSE values are 5.25×10^{-4}, 2.95×10^{-5} and 4.14×10^{-5}. Therefore we choose model (M2) as our final proposed model. Additional metrics and qualitative visual comparisons supporting this choice are shown in Section S. 3.

6.3. Experimental Evaluation

As described in Section 5.1, we use data from [30] to demonstrate the effectiveness of the proposed method on real-world data. Figure 8 shows two pre-processed image crops with extreme DR (CMOS and SPC columns) that are used as inputs to our network. The SPC images have a $5\,\mathrm{ms}$ exposure. The CMOS inputs (first column) have a exposure times of $0.1\,\mathrm{ms}$ (first row) and $0.5\,\mathrm{ms}$ (second row). Different from our synthetic data evaluation, the exposure times between CMOS and SPC are chosen not to match because using higher exposures for CMOS would have led to fully saturated images. Similar to our synthetic data evaluation, DHDR fails to recover fine structures in the saturated CMOS regions (e.g., fire dept. letters). Moreover, exposure bracketing, despite using 6 exposures ranging from $0.005\,\mathrm{ms}$ to $5\,\mathrm{ms}$, is still unable to recover the thin wires of the lamp (first row). Due to imperfect alignment of the CMOS and SPC crops, our model blurs these fine structures in both images (lamp details and text). Nonetheless, these features are still visible and are not completely suppressed by the CMOS saturation limit. Note that the CMOS and SPC image crops are derived from similar spatial resolutions, therefore, the CMOS crop does not contain any additional spatial information that our model can exploit.

7. Discussion and Limitations

We present a model for extreme HDR imaging that fuses a high-resolution LDR CMOS image with a low-resolution HDR SPC image. These two imaging technologies address the fundamental limitations (low dynamic range and low spatial resolution) of each other, making their fusion a natural design choice. Our evaluation demonstrates that in extreme DR scenarios, CMOS and SPC sensors cannot overcome their limitations on their own, even with state-of-the-art models such as DHDR [55] and ESRGAN [62]. Moreover, we show that extending the CMOS dynamic range with a second low-exposure CMOS image is a sub-optimal design choice in an extreme HDR setting, due to the significant SNR dip artifacts. Overall, our proposed SPC-guided HDR model is a promising imaging modality for emerging computer vision applications that require HDR content.

CMOS-SPC Prototype: As discussed in Section 6.3, the imperfect alignment between SPC and CMOS images can cause our model to blur fine details. Moreover, the low-resolution of the CMOS crop prevents our model from leveraging the high-resolution information that was available during training. Nonetheless, our experimental results suggest a reasonable degree of generalization by our model on real-world data, despite being trained solely on synthetic data. The next step is to build a hardware prototype that mitigates image mis-alignments between the CMOS and SPC images, and uses appropriate optics for HDR scenarios. This will enable extensive real-world evaluation of our model, and inform future designs.

Extreme HDR in Color: Here we restricted our analysis to extending the dynamic range of monochrome images. Although current SPCs are monochrome, future sensors will likely incorporate color filter arrays. Moreover, as discussed in the supplement, there is important inter-channel information that can be used to further extend the dynamic range of an image. In the future, our method can also leverage multiple color channels to further improve dynamic range.

Extreme HDR Video: Our proposed model only requires two images that can be acquired simultaneously, which is a practical design for dynamic applications that require extreme DR. Extending our model for extreme DR video reconstruction is a promising avenue for future work.

Acknowledgments: This work was supported in part by the U.S. Department of Energy/NNSA (DE-NA0003921), National Science Foundation (CAREER 1846884 and 1943149), and UW-Madison's Draper Technology Innovation Fund. The authors would like to thank the Computational Imaging Group at Rice University for providing conference travel funds for Yuhao Liu. U.S. DoE full legal disclaimer: `https://www.osti.gov/stip/about/disclaimer`.

References

[1] AIoT/AR/Auto. http://4da.tech 3DA: 3D/ToF/LiDAR. Apple LIDAR Demystified: SPAD, VCSEL, and Fusion, March 2021. [Online; accessed 01-08-2021].

[2] T Asatsuma, Y Sakano, S Iida, M Takami, I Yoshiba, N Ohba, H Mizuno, T Oka, K Yamaguchi, A Suzuki, et al. Sub-pixel architecture of cmos image sensor achieving over 120 db dynamic range with less motion artifact characteristics. In *Proceedings of the 2019 International Image Sensor Workshop*, volume 1, 2019.

[3] Abhishek Badki, Nima Khademi Kalantari, and Pradeep Sen. Robust radiometric calibration for dynamic scenes in the wild. In *2015 IEEE International Conference on Computational Photography (ICCP)*, pages 1–10. IEEE, 2015.

[4] Francesco Banterle, Alessandro Artusi, Kurt Debattista, and Alan Chalmers. *Advanced high dynamic range imaging*. AK Peters/CRC Press, 2017.

[5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[6] Peter J Burt and Edward H Adelson. A multiresolution spline with application to image mosaics. *ACM Transactions on Graphics (TOG)*, 2(4):217–236, 1983.

[7] Mauro Buttafava, Federica Villa, Marco Castello, Giorgio Tortarolo, Enrico Conca, Mirko Sanzaro, Simonluca Piazza, Paolo Bianchini, Alberto Diaspro, Franco Zappa, et al. SPAD-based asynchronous-readout array detectors for image-scanning microscopy. *Optica*, 7(7):755–765, 2020.

[8] Mauro Buttafava, Jessica Zeman, Alberto Tosi, Kevin Eliceiri, and Andreas Velten. Non-line-of-sight imaging using a time-gated single photon avalanche diode. *Optics express*, 23(16):20997–21011, 2015.

[9] Clara Callenberg, Ashley Lyons, Dennis den Brok, Areeba Fatima, Alex Turpin, Vytautas Zickus, Laura Machesky, Jamie Whitelaw, Daniele Faccio, and MB Hullin. Super-resolution time-resolved imaging using computational sensor fusion. *Scientific reports*, 11(1):1–8, 2021.

[10] William J. Claff. Input-referred Read Noise vs. ISO Setting. `https://www.photonstophotos.net/Charts/RN_e.htm`, 2021.

[11] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '97, page 369–378, USA, 1997. ACM Press/Addison-Wesley Publishing Co.

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

[13] Brian Deegan. Led flicker: Root cause, impact and measurement for automotive imaging applications. *Electronic Imaging*, 2018(17):146–1, 2018.

[14] Vinit Dhulla, Sapna S Mukherjee, Adam O Lee, Nanditha Dissanayake, Booshik Ryu, and Charles Myers. 256 x 256 dual-mode cmos spad image sensor. In *Advanced Photon Counting Techniques XIII*, volume 10978, page 109780Q. International Society for Optics and Photonics, 2019.

[15] 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):1–15, 2017.

[16] Eric R Fossum, Jiaju Ma, Saleh Masoodian, Leo Anzagira, and Rachel Zizza. The quanta image sensor: Every photon counts. *Sensors*, 16(8):1260, 2016.

[17] Brian Funt and Lilong Shi. The effect of exposure on MaxRGB color constancy. In *Human Vision and Electronic Imaging XV*, volume 7527, page 75270Y. International Society for Optics and Photonics, 2010.

[18] Brian Funt and Lilong Shi. The rehabilitation of MaxRGB. In *Color and imaging conference*, volume 2010, pages 256–259. Society for Imaging Science and Technology, 2010.

[19] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.

[20] Abhiram Gnanasambandam and Stanley H Chan. Hdr imaging with quanta image sensors: Theoretical limits and optimal reconstruction. *IEEE Transactions on Computational Imaging*, 6:1571–1585, 2020.

[21] Geoffrey Grimmett and David Stirzaker. *Probability and random processes*. Oxford university press, 2020.

[22] Michael D Grossberg and Shree K Nayar. Determining the camera response from images: What is knowable? *IEEE Transactions on pattern analysis and machine intelligence*, 25(11):1455–1467, 2003.

[23] Anant Gupta, Atul Ingle, and Mohit Gupta. Asynchronous single-photon 3d imaging. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7909–7918, 2019.

[24] Jin Han, Chu Zhou, Peiqi Duan, Yehui Tang, Chang Xu, Chao Xu, Tiejun Huang, and Boxin Shi. Neuromorphic camera guided high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1730–1739, 2020.

[25] Samuel W Hasinoff, Frédo Durand, and William T Freeman. Noise-optimal capture for high dynamic range photography. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 553–560. IEEE, 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):1–12, 2016.

[27] Poly Haven. Poly haven. `https://polyhaven.com/`, 2021.

[28] S Iida, Y Sakano, T Asatsuma, M Takami, I Yoshiba, N Ohba, H Mizuno, T Oka, K Yamaguchi, A Suzuki, et al. A 0.68 e-rms random-noise 121dB dynamic-range sub-pixel architecture CMOS image sensor with LED flicker mitigation. In *2018 IEEE International Electron Devices Meeting (IEDM)*, pages 10–2. IEEE, 2018.

[29] Atul Ingle, Trevor Seets, Mauro Buttafava, Shantanu Gupta, Alberto Tosi, Mohit Gupta, and Andreas Velten. Passive inter-photon imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8595, 2021.

[30] Atul Ingle, Andreas Velten, and Mohit Gupta. High flux passive imaging with single-photon sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6760–6769, 2019.

[31] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[32] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017.

[33] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[34] Oichi Kumagai, Junichi Ohmachi, Masao Matsumura, Shinichiro Yagi, Kenichi Tayu, Keitaro Amagawa, Tomohiro Matsukawa, Osamu Ozawa, Daisuke Hirono, Yasuhiro Shinozuka, et al. A 189x600 Back-Illuminated Stacked SPAD Direct Time-of-Flight Depth Sensor for Automotive LiDAR Systems. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pages 110–112. IEEE, 2021.

[35] Martin Laurenzis. Single photon range, intensity and photon flux imaging with kilohertz frame rate and high dynamic range. *Optics express*, 27(26):38391–38403, 2019.

[36] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 136–144, 2017.

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

[38] Ashley Lyons, Francesco Tonolini, Alessandro Boccolini, Audrey Repetti, Robert Henderson, Yves Wiaux, and Daniele Faccio. Computational time-of-flight diffuse optical tomography. *Nature Photonics*, 13(8):575–579, 2019.

[39] Sizhuo Ma, Shantanu Gupta, Arin C Ulku, Claudio Bruschini, Edoardo Charbon, and Mohit Gupta. Quanta burst photography. *ACM Transactions on Graphics (TOG)*, 39(4):79–1, 2020.

[40] 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] Rafał 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 Transactions on graphics (TOG)*, 30(4):1–14, 2011. v.3.0.6.

[42] Rafał Mantiuk, Grzegorz Krawczyk, Dorota Zdrojewska, Radosław Mantiuk, Karol Myszkowski, and Hans-Peter Seidel. *High dynamic range imaging*. Wiley Encyclopedia of Electrical and Electronics Engineering, 2015.

[43] 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. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018.

[44] Mitsuhito Mase, Shoji Kawahito, Masaaki Sasaki, Yasuo Wakamori, and Masanori Furuta. A wide dynamic range CMOS image sensor with multiple exposure-time signal outputs and 12-bit column-parallel cyclic A/D converters. *IEEE Journal of Solid-State Circuits*, 40(12):2787–2795, 2005.

[45] Kazuhiro Morimoto, Andrei Ardelean, Ming-Lo Wu, Arin Can Ulku, Ivan Michel Antolovic, Claudio Bruschini, and Edoardo Charbon. Megapixel time-gated SPAD image sensor for 2D and 3D imaging applications. *Optica*, 7(4):346–354, 2020.

[46] 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.

[47] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[48] Jun Ogi, Takafumi Takatsuka, Kazuki Hizu, Yutaka Inaoka, Hongbo Zhu, Yasuhisa Tochigi, Yoshiaki Tashiro, Fumiaki Sano, Yusuke Murakawa, Makoto Nakamura, et al. A 250fps 124dB Dynamic-Range SPAD Image Sensor Stacked with Pixel-Parallel Photon Counter Employing Sub-Frame Extrapolating Architecture for Motion Artifact Suppression. In *2021 IEEE International Solid-State Circuits Conference (ISSCC)*, volume 64, pages 113–115. IEEE, 2021.

[49] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention U-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[50] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, pages 8026–8037, 2019.

[51] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Ales Leonardis, and Radu Timofte. NTIRE 2021 challenge on high dynamic range imaging: Dataset, methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 691–700, 2021.

[52] Ana Radonjić, Sarah R Allred, Alan L Gilchrist, and David H Brainard. The dynamic range of human lightness perception. *Current Biology*, 21(22):1931–1936, 2011.

[53] 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.

[54] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image com-*

puting and computer-assisted intervention, pages 234–241. Springer, 2015.

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

[56] Trevor Seets, Atul Ingle, Martin Laurenzis, and Andreas Velten. Motion adaptive deblurring with single-photon cameras. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1945–1954, 2021.

[57] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[58] Richard Szeliski. *Computer vision: algorithms and applications*. Springer Science & Business Media, 2010.

[59] PyTorch Team. PyTorch VGG-Nets. `https://pytorch.org/hub/pytorch_vision_vgg/`.

[60] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in hdr deghosting: a survey and evaluation. In *Computer Graphics Forum*, volume 34, pages 683–707. Wiley Online Library, 2015.

[61] Sergey Velichko, Scott Johnson, Dan Pates, Chris Silsby, Cornelis Hoekstra, Ray Mentzer, and Jeff Beck. 140 db dynamic range sub-electron noise floor image sensor. *Proceedings of the IISW*, 1, 2017.

[62] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks. *arXiv:1809.00219 [cs]*, Sep 2018. arXiv: 1809.00219.

[63] Zihao W Wang, Peiqi Duan, Oliver Cossairt, Aggelos Katsaggelos, Tiejun Huang, and Boxin Shi. Joint filtering of intensity images and neuromorphic events for high-resolution noise-robust imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1609–1619, 2020.

[64] Wikipedia contributors. Orders of magnitude (illuminance) — Wikipedia, the free encyclopedia, 2004. [Online; accessed 02-August-2021].

[65] David XD Yang and Abbas El Gamal. Comparative analysis of snr for image sensors with enhanced dynamic range. In *Sensors, cameras, and systems for scientific/industrial applications*, volume 3649, pages 197–211. International Society for Optics and Photonics, 1999.

[66] Junko Yoshida. Breaking Down iPad Pro 11's LiDAR Scanner, June 2020. [Online; accessed 01-08-2021].

[67] Vytautas Zickus, Ming-Lo Wu, Kazuhiro Morimoto, Valentin Kapitany, Areeba Fatima, Alex Turpin, Robert Insall, Jamie Whitelaw, Laura Machesky, Claudio Bruschini, et al. Fluorescence lifetime imaging with a megapixel SPAD camera and neural network lifetime estimation. *Scientific Reports*, 10(1):1–10, 2020.