The Visual Computer
https://doi.org/10.1007/s00371-023-03032-4

ORIGINAL ARTICLE

Deep snapshot HDR imaging using multi-exposure color filter array

Yutaro Okamoto[1,2] · Masayuki Tanaka[1] · Yusuke Monno[1] · Masatoshi Okutomi[1]

Accepted: 2 July 2023

Abstract
In this paper, we propose a deep snapshot high dynamic range (HDR) imaging framework that can effectively reconstruct an HDR image from the RAW data captured using a multi-exposure color filter array (ME-CFA), which consists of a mosaic pattern of RGB filters with different exposure levels. To effectively learn the HDR image reconstruction network, we introduce the idea of luminance normalization that simultaneously enables effective loss computation and input data normalization by considering relative local contrasts in the "normalized-by-luminance" HDR domain. This idea enables the network to equally handle the errors in both bright and dark areas regardless of absolute luminance levels, which significantly improves the visual image quality. Experimental results using public HDR image datasets demonstrate that our framework outperforms other snapshot methods and produces high-quality HDR images with fewer visual artifacts, resulting in more than 4dB color peak signal-to-noise ratio improvement in the linear HDR domain.

Keywords High dynamic range (HDR) imaging · Multi-exposure color filter array · Demosaicking

1 Introduction

The dynamic range of a camera is determined by the ratio between the maximum and the minimum amounts of light that can be recorded by the image sensor in one shot. Standard digital cameras have a low dynamic range (LDR) and only capture a limited range of scene radiance. Consequently, they cannot capture a bright and a dark area outside the camera's dynamic range simultaneously. High dynamic range (HDR) imaging is a highly demanded computational imaging technique to overcome this limitation, which recovers the HDR scene radiance map from a single or multiple LDR images captured by a standard camera.

HDR imaging is typically performed by estimating a mapping from the sensor's LDR outputs to the scene radiance using multiple LDR images which are sequentially captured with different exposure levels [1]. Although this approach works for static situations, it is not suitable for dynamic scenes and video acquisition since ghost artifacts may occur because of target or camera motions while taking multiple images with different exposure settings. Recent learning-based methods [2–4] have successfully reduced ghost artifacts derived from target motions between input LDR images. However, those methods are limited to small target motions and the artifacts remain apparent for the areas with large motions.

Some studies have used only a single LDR image to realize one-shot HDR imaging [5–7]. They essentially inpaint or hallucinate missing over- and under-exposed areas by exploiting an external database of LDR-HDR image pairs. Although this approach is free from ghost artifacts, it generates inpainting or hallucination artifacts for largely over- and under-exposed areas.

As another one-shot approach, the methods based on a snapshot HDR sensor have also been investigated [8–11]. One way to realize a snapshot HDR sensor is to use a single-image sensor with spatially varying exposure levels [10, 11]. This can be achieved by using what we call a multi-exposure color filter array (ME-CFA), which consists of a mosaic pattern of RGB filters combined with neutral density filters with different attenuation levels (see Fig. 1 for an example). The snapshot HDR sensor has the advantage of capturing multi-exposure information in one shot. Thus, it has great potential for HDR imaging of dynamic scenes and HDR video acquisition without suffering from

✉ Yusuke Monno
ymonno@ok.sc.e.titech.ac.jp

1 Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology, 2-12-1 Ookayama, Meguro-ku, Tokyo 152-8550, Japan

2 Advanced Technology Research Institute, Kyocera Corporation, Kanagawa, Japan

Published online: 04 August 2023

ghosting and inpainting artifacts. However, HDR image reconstruction from the snapshot measurement includes two challenging problems: color demosaicking (i.e., interpolation of missing RGB values) and HDR reconstruction (i.e., scene radiance estimation from LDR measurements). As we will experimentally show later, a simple combination of existing demosaicking/interpolation and HDR reconstruction methods cannot produce satisfactory results for this joint problem.

In this paper, we address the joint problem of color demosaicking and HDR reconstruction for snapshot HDR imaging, given mosaic RAW data captured using an ME-CFA. To address this problem, we propose a novel deep snapshot HDR imaging framework that can effectively reconstruct an HDR image from the RAW data captured using an ME-CFA. Figure 1 shows our framework, where we introduce the idea of luminance normalization and reconstruct the HDR image in the luminance-normalized domain.

In the training phase, we first train an LDR interpolation network (LDR-I-Net) to estimate tentative HDR luminance from interpolated LDR image generation. Then, we normalize the input ME-CFA RAW data by the tentative HDR luminance. Finally, we train a luminance-normalized HDR image reconstruction network (LN-Net) to reconstruct an HDR image based on the pair of the ME-CFA RAW data and the ground-truth HDR image data in the luminance-normalized domain. In the application phase, the HDR image is reconstructed through the learned two networks, where the final HDR image result is derived as the multiplication of the tentative HDR luminance and the luminance-normalized HDR image.

The proposed framework mainly has two benefits. The first one is effective loss computation in the luminance-normalized domain. The standard mean squared error (MSE) loss in the linear HDR domain has a problem of neglecting the errors in dark areas because they are quite small compared with the errors in bright areas. However, those errors in dark areas significantly affect the visual quality in a tone-mapped domain [12], which is commonly used to display HDR images. Based on this, some studies have computed the loss in a transformed domain, such as a log domain [6] and a global tone-mapped domain [2].

However, these signal-independent transformations do not reflect an actual signal component of each image. In contrast, by computing the loss in the luminance-normalized domain, we can equally handle the errors in bright and dark areas by considering the actual luminance of each image.

The second benefit of the proposed framework is effective input data normalization. In deep learning, the normalization of input data is important to extract effective features. Since a diverse range of scene radiance information is simultaneously encoded in the ME-CFA RAW data, we need to consider relative local contrasts, rather than absolute differences. Otherwise, features such as edges and textures in dark areas are prone to be ignored. In our framework, by normalizing the input RAW data by the tentative HDR luminance, we can naturally consider the relative local contrasts in both bright and dark areas regardless of absolute luminance.

Through the experiments using three public HDR image or video datasets, we validate the effectiveness of our framework by comparing it with other snapshot methods and state-of-the-art HDR imaging methods using multiple LDR images.

Main contributions of this paper are summarized as follows.

- We propose a novel deep learning framework that effectively solves the joint demosaicking and HDR reconstruction problem for snapshot HDR imaging.
- We propose the idea of luminance normalization that simultaneously enables effective loss computation and input data normalization by considering the relative local contrasts of each image.
- We demonstrate that our framework can outperform other snapshot methods and reconstruct high-quality HDR images and videos with fewer visible artifacts.

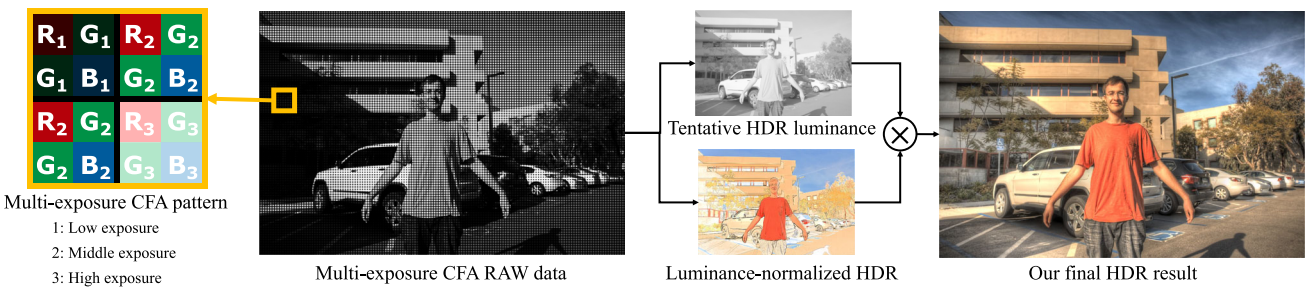


Fig. 1 We propose a deep snapshot HDR imaging method to produce a high-quality HDR image using a multi-exposure color filter array (ME-CFA). The ME-CFA RAW data contain three exposure levels with a regular Bayer pattern. In the proposed framework, we first estimate tentative HDR luminance and then reconstruct the HDR image in the luminance-normalized domain, where we can consider relative local contrasts in both bright and dark areas, regardless of the absolute luminance levels. The final HDR image result is derived as the multiplication of the tentative HDR luminance and the luminance-normalized HDR image

This paper is an extended version of our previous paper published in [13]. In the previous version, the loss of LDR-I-Net is computed only between interpolated LDR images and ground-truth LDR images. However, the previous version suffers from zipper artifacts in the final HDR image result. In this extended version, we have added a luminance loss for LDR-I-Net to update the network weights using both the errors of interpolated LDR images and the tentative luminance image, which significantly reduces visible zipper artifacts. We have also added experimental comparison using an HDR video dataset containing dynamic targets, which clearly demonstrates the advantage of our proposed snapshot method compared with existing methods using multiple LDR images.

2 Related work

Multiple-LDR-images-based methods have been studied for years. Their basic approaches include inverse radiometric function estimation [1] and exposure fusion [14–16]. In these methods, a weighting function is designed to fuse multiple LDR images. However, because input LDR images are assumed to be aligned, these methods are only applicable to static scenes, or some additional image alignment methods are needed. Some other methods consider the effect of dynamic objects. To reduce ghost artifacts, Hasinoff et al. and Hafner et al. compute optical flows to align input LDR images [17, 18]. Hu et al. and Sen et al. decompose input LDR images into small patches and apply a patch-match strategy to reconstruct a final HDR image [19, 20]. Lee et al. and Oh et al. divide input LDR images into a static background and moving objects by a rank minimization strategy [21, 22]. Learning-based methods have also actively been researched [2–4, 23–27], where recent studies adopt generative adversarial networks [28], zero- and few-shot learning [29], and transformers [30, 31]. Although their performance has continuously been improved (see [32, 33] for reviews), multiple-LDR-images-based methods are essentially difficult to handle fast and large target or camera motions, resulting in ghost artifacts or misalignment artifacts.

From the view of hardware designs, some methods have exploited a multi-camera/sensor system [34, 35] for the one-shot acquisition of multiple LDR images. Recently, a neuromorphic camera is used to guide HDR imaging from a single LDR image [36]. However, the systems using multiple cameras require image or sensor alignment, which is another challenging task.

Single-LDR-image-based methods, also called inverse tone mapping, have been actively studied in recent years. The representative approach trains the mapping from a single LDR image to an HDR image directly [5, 6, 37–40]. The other common approach trains the mapping from a single LDR image to multiple LDR images intermediately, from which the final HDR image is derived [7, 41, 42]. Another approach learns the camera imaging pipeline from HDR to LDR to inverse the process [43]. Other recent studies aim at converting an LDR image to an HDR image considering ultra-high-definition image quality [44] with novel dataset generation [45, 46]. Although single-LDR-image-based approaches realize one-shot HDR image acquisition, they are essentially difficult to reconstruct high-quality HDR images because there are no measurements obtained from different exposure levels. To handle saturated pixels, learning the optical design [47, 48] has also been investigated. However, these methods need an additional lens setup, which makes the hardware setting more complex.

Snapshot methods are based on a snapshot HDR imaging system with spatially varying exposure levels [10, 11]. Several hardware architectures or concepts have been proposed to realize a snapshot system, such as a coded exposure time [8, 49, 50], a coded exposure mask [51–53], a dual-ISO sensor [9, 54–56], and what we call an ME-CFA, which consists of the mosaic of RGB filters combined with neutral density filters with different attenuation levels [10, 11, 57–61]. The snapshot systems have great potential for HDR imaging in dynamic situations since it enables one-shot acquisition of multi-exposure information. However, HDR image reconstruction from the snapshot measurements is very challenging due to the sparse nature of each color-exposure component.

Some existing snapshot methods based on an ME-CFA first convert the snapshot LDR measurements to the sensor irradiance domain. By doing this, the problem reduces to the demosaicking problem in the sensor irradiance domain, for which several probability-based [57, 60] or learning-based [59, 62, 63] approaches have been proposed. However, this combined and sequential approach could not necessarily produce satisfactory results because the errors in the first step are propagated by the demosaicking step.

Joint approaches have also been proposed. Narasimhan and Nayer use the pair of ground-truth HDR images and ME-CFA RAW versions of them to train a simple polynomial function, which interpolates the missing pixel values using neighboring pixels [10]. Some other methods use image differentiation [61] and bilateral filter [8] to interpolate the pixel values. However, these simple interpolation approaches suffer from the artifacts such as blurring edges and false colors. Martel et al. apply a deep learning framework to HDR imaging from ME-CFA RAW data generated by a programmable shutter function called neural sensors [64]. Vien and Lee use a robust loss function, which considers luminance, chrominance, and contrast [65]. Xu et al. apply an exposure guidance mask to mask out the effect of over-exposed pixels to jointly learn the demosaicking and the HDR reconstruction [66].

However, these methods assume a specific exposure pattern, such as two exposure levels with a row-wise pattern.

To summarize, the methods based on multiple LDR images are susceptible to ghost artifacts in dynamic scenes, while the methods based on a single LDR image often produce inpainting or hallucination artifacts due to the lack of multi-exposure information. Although the snapshot methods overcome these limitations, the joint task of demosaicking and HDR reconstruction is challenging because of the sparse nature of each color-exposure component with many over-/under-exposed pixels in the mosaicked form. To address this challenge, we propose a novel over-/under-exposed pixel correction method as a pre-processing and develop a general and high-performance framework exploiting deep learning to jointly solve the demosaicking and the HDR reconstruction problems for snapshot HDR imaging.

3 Proposed deep snapshot HDR imaging

3.1 Framework overview

In this paper, we apply the ME-CFA pattern shown in Fig. 1, which consists of a 4×4 regular pattern with three exposure levels, assuming the mosaic of RGB filters combined with neutral density filters with three attenuation levels. Although our framework is general and not limited to this pattern, we use this pattern because (i) it is based on the common Bayer pattern [67], similar to existing ME-CFA patterns [10, 55], and (ii) it consists of three exposures, which are commonly used in recent HDR imaging studies [2, 3]. Those two conditions enable us to experimentally compare our framework with standard Bayer demosaicking methods [68, 69] and competitive HDR imaging methods using three LDR images [2, 3].

Figure 2 shows the overview of our framework, which mainly consists of two parts: (i) luminance estimation and (ii) luminance-normalized HDR image reconstruction. The first part estimates tentative HDR luminance based on the interpolated LDR images by the learned LDR-interpolation-network (LDR-I-Net). Then, based on the tentative HDR luminance, the second part estimates the luminance-normalized HDR image by the learned luminance-normalized HDR image reconstruction network (LN-Net). Each part is detailed in subsections 3.2 and 3.3, respectively. Finally, the HDR image is reconstructed by multiplying the tentative HDR luminance and the estimated luminance-normalized HDR image.

Throughout this paper, we use the term "irradiance" or "sensor irradiance [1]" to represent the irradiance of the light reaching the image sensor after going through the camera's optical elements such as a lens. Because we assume a linear optical system as in [1, 70], the sensor irradiance is assumed to be the same as the scene radiance in this paper. We also assume linear responses between the sensor irradiance and pixel values because we process the RAW data, which typically have linear camera responses.

3.2 Luminance estimation

In the luminance estimation, we first interpolate the missing RGB pixel values to generate interpolated LDR images. For this purpose, we train LDR-I-Net. Then, we apply Debevec's method [1] to the interpolated LDR images for tentative HDR

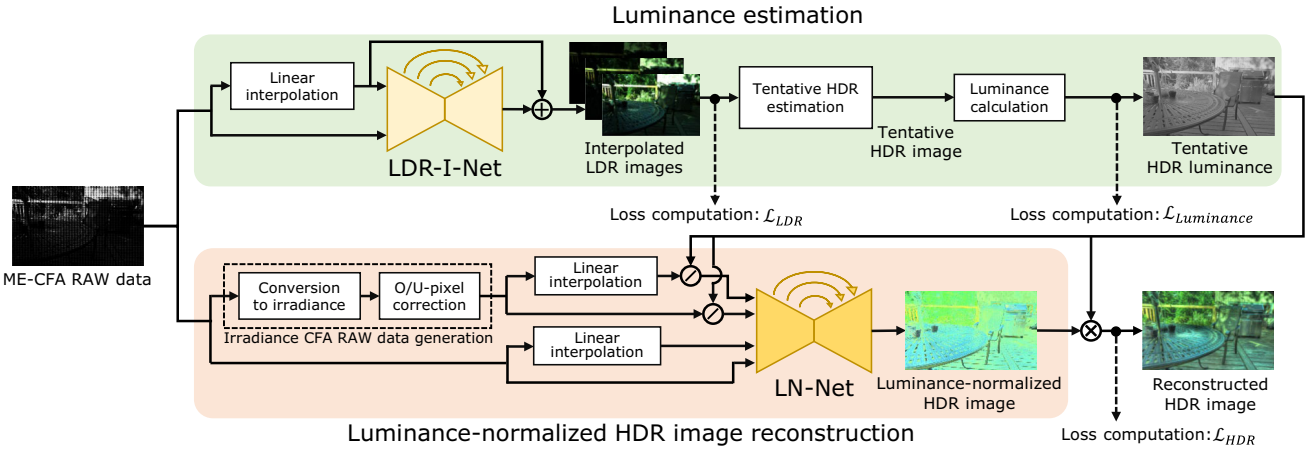


Fig. 2 The overview of our deep snapshot HDR imaging framework. It first estimates tentative HDR luminance and then estimates the HDR image in the luminance-normalized domain. The idea of luminance normalization enables us to consider relative local contrasts in both bright and dark areas, regardless of absolute luminance levels. The final HDR image is reconstructed by multiplying the tentative HDR luminance and the luminance-normalized HDR image

image estimation. Then, tentative HDR luminance is derived as the smoothed maximum value of the RGB sensor irradiance values, which corresponds to the value (V) in the HSV color space [71]. The smoothed tentative HDR luminance is calculated as

$$\hat{L}_{i,j} = \sum_{m=-d}^{d} \sum_{n=-d}^{d} \sigma_{m,n} \max_{c\in\{R,G,B\}} \left(\hat{E}^{c}_{i+m,j+n}\right) \tag{1}$$

where $\hat{L}_{i,j}$ is the calculated luminance at pixel (i, j), σ is the Gaussian kernel with the square size of $(2d+1) \times (2d+1)$ for smoothing, $\hat{E}^{c}_{i+m,j+n}$ is the value of the tentative HDR image at color channel c and pixel $(i+m, j+n)$, where (m, n) represents the relative pixel location in the Gaussian kernel.

For the training of LDR-I-Net, we input sub-mosaic ME-CFA RAW data and its linearly interpolated data (see Fig. 3 for illustrations). The loss function for LDR-I-Net is described as

$$\mathcal{L}_{\text{Total}} = \alpha \mathcal{L}_{\text{LDR}} + (1-\alpha)\mathcal{L}_{\text{Luminance}}, \tag{2}$$

where we take LDR interpolation loss $\mathcal{L}_{\text{LDR}}$ and luminance loss $\mathcal{L}_{\text{Luminance}}$ with balancing parameter α. The LDR interpolation loss is described as

$$\begin{aligned}\mathcal{L}_{\text{LDR}} = \sum_{k=1}^{N} (&||\boldsymbol{f}_k(\boldsymbol{y};\boldsymbol{\theta}) - \boldsymbol{Z}_k||_1 \\ &+\lambda_1||\nabla \boldsymbol{f}_k(\boldsymbol{y};\boldsymbol{\theta}) - \nabla \boldsymbol{Z}_k||_1),\end{aligned} \tag{3}$$

where $\boldsymbol{y} = [\boldsymbol{x}; h(\boldsymbol{x})]$ is the network input, $\boldsymbol{x}$ is the sub-mosaicked representation of the ME-CFA RAW data, as used in [72] (i.e., sparse 16-channel data for the 4×4 regular pattern, as shown in Fig. 3), $h()$ represents the linear interpolation for the sparse data, which can be performed by convolving the 7×7 linear interpolation kernel used in [72]. $\boldsymbol{f}_k(\boldsymbol{y};\boldsymbol{\theta})$ is the output of LDR-I-Net for k-th exposure LDR image, $\boldsymbol{\theta}$ represents the network weights, $\boldsymbol{Z}_k$ is the ground-truth k-th exposure LDR image, N is the number of exposure levels in the ME-CFA, ∇ represents the horizontal and vertical derivative operators, and λ_1 is a hyper-parameter. The first term directly evaluates the differences between the estimated and the ground-truth LDR images, while the second term, which we call the gradient term, evaluates the differences in the image gradients to suppress zipper artifacts.

Since HDR images are usually visualized in a tone-mapped domain, taking a loss in a tone-mapped HDR domain improves the visual quality. Following the tone-mapped loss of [2], we calculate the luminance loss as

$$\mathcal{L}_{\text{Luminance}} = ||\gamma(\hat{\boldsymbol{L}}) - \gamma(\boldsymbol{L})||_1 + \lambda_1||\nabla\gamma(\hat{\boldsymbol{L}}) - \nabla\gamma(\boldsymbol{L})||_1, \tag{4}$$

where $\boldsymbol{L}$ and $\hat{\boldsymbol{L}}$ are ground-truth and tentative HDR luminance, respectively, and ∇ represents the horizontal and vertical derivative operator. The function γ is a global tone-mapping operator of [2], which is described as

$$\gamma(x) = \frac{\log(1+\mu x)}{\log(1+\mu)}, \tag{5}$$

where μ is a tone-mapping parameter. Similar to the LDR interpolation loss, the second term of the luminance loss represents the gradient term, which evaluates the gradients of the estimated luminance.

In this paper, we empirically set $\alpha = 0.01$, $\lambda_1 = 1$, and $\mu = 5000$ as the hyper-parameters. We validate the effects of the luminance loss and the gradient terms later in Sect. 4.5.

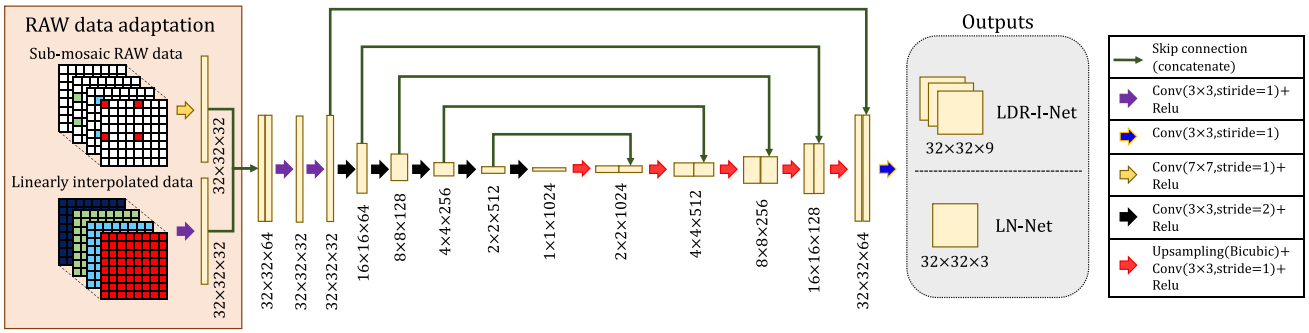


Fig. 3 The architecture of LDR-I-Net. We use U-net architecture with a depth of five. The inputs for LDR-I-Net are sub-mosaic RAW data and its linearly interpolated version. The number of output channels is nine for LDR-I-Net, which represents RGB channels for three LDR images. The architecture of LN-Net is similar, with the difference in input and output channels. As in Eq. (8), we use LDR domain data and luminance-normalized domain irradiance data for the inputs of LN-Net. The number of output channels of LN-Net is three, which represents RGB channels for one luminance-normalized HDR image

3.3 Luminance-normalized HDR image reconstruction

In the HDR image reconstruction, we train LN-Net to reconstruct the HDR image in the luminance-normalized domain. The inputs of the network are the sub-mosaicked representation of the ME-CFA RAW data, its sensor irradiance version as we will explain later, and linearly interpolated versions of them. The irradiance data are normalized by the tentative HDR luminance which is estimated by the previous step to consider relative local contrasts regardless of the absolute luminance levels. We detail each process to train LN-Net in the rest of this section.

We first convert the ME-CFA RAW data to the sensor irradiance domain by dividing the pixel value by the exposure time multiplied by the corresponding attenuation factor of the pixel associated with the ME-CFA. The irradiance is calculated as

$$\xi_{k,i} = \frac{x_i}{\rho_k \Delta t}, \tag{6}$$

where x_i is i-th pixel value of the ME-CFA RAW data, ρ_k is the attenuation factor for k-th exposure, Δt is the exposure time, and $\xi_{k,i}$ is the converted sensor irradiance of i-th pixel corresponding to k-th exposure. In the snapshot case using an ME-CFA, the attenuation factor ρ_k varies for each pixel according to the ME-CFA pattern, while the exposure time is constant for all the pixels. Thus, in what follows, we set to $\Delta t = 1$, without loss of generality. Also, we call the irradiance data converted by Eq. (6) "irradiance CFA-RAW data," in which different exposure levels are already corrected by converting the ME-CFA RAW data to the sensor irradiance domain. Figure 4a and b shows the examples of the ME-CFA RAW data and the converted irradiance CFA RAW data.

In the original ME-CFA RAW data, many pixels are over-exposed (saturated) or under-exposed (black-out) depending on the exposure level of each pixel compared to the scene radiance. Such over- or under-exposed pixels, which we denote as "O/U pixels," have no meaningful irradiance information even after the conversion by Eq. (6). Thus, we propose an effective O/U-pixel correction method, which replaces the irradiance of an O/U pixel with the linearly interpolated value using adjacent lower- or higher-exposure irradiance samples. For example, the irradiance of an over-exposed pixel is corrected as

$$\hat{\xi}_{k,i} = \begin{cases} \xi_{k,i} & (\xi_{k,i} \leq \tau_{O,k}) \\ h(\boldsymbol{\xi}_{k-1})_i & (\xi_{k,i} > \tau_{O,k}) \end{cases}, \tag{7}$$

where the suffix k represents k-th exposure, the suffix i represents i-th pixel, $\hat{\xi}_{k,i}$ is the irradiance after the over-exposed pixel correction, $\tau_{O,k}$ is the over-exposure threshold, $\boldsymbol{\xi}_{k-1}$ is the one-step lower-exposure sparse irradiance samples in the irradiance CFA RAW data, $h()$ is the linear interpolation operator, and $h(\boldsymbol{\xi}_{k-1})_i$ is i-th pixel value of the linearly interpolated irradiance of $\boldsymbol{\xi}_{k-1}$. We empirically set $0.995/(\rho_k \Delta t)$ to the over-exposure threshold $\tau_{O,k}$, where the range of the irradiance CFA RAW data is [0, 1]. This over-exposure correction is applied from the lower exposure data to the higher exposure data. The under-exposed pixel correction is performed in the same manner, where the under-exposure threshold is set to $\tau_{U,k} = 0.005/(\rho_k \Delta t)$. Figure 4b and c shows examples of the irradiance CFA RAW data before and after the proposed O/U-pixel correction, respectively.

We then apply linear interpolation to the corrected irradiance CFA RAW data to prepare the network input. The corrected irradiance CFA RAW data $\hat{\boldsymbol{\xi}}$ and its linearly interpolated version $h(\hat{\boldsymbol{\xi}})$ can be considered as the HDR domain data, in which local contrasts in dark areas are very low compared with those in bright areas. Thus, we normalize the HDR domain data by the estimated tentative HDR luminance. This luminance normalization converts the absolute local contrasts to the relative local contrasts. We also input the LDR domain data of the sub-mosaicked ME-CFA RAW data $\boldsymbol{x}$ and its linearly interpolated version $h(\boldsymbol{x})$. For these LDR domain data, we do not perform the luminance normalization because the range of the absolute local contrasts is limited. The input to LN-Net $\boldsymbol{\eta}$ is described as

$$\boldsymbol{\eta} = \left[\boldsymbol{x}, h(\boldsymbol{x}), \hat{\boldsymbol{\xi}}/\hat{\boldsymbol{L}}, h(\hat{\boldsymbol{\xi}})/\hat{\boldsymbol{L}}\right], \tag{8}$$

where $\hat{\boldsymbol{L}}$ is the tentative HDR luminance and the division operation is performed in a pixel-by-pixel manner.

With these inputs, LN-Net generates the luminance-normalized HDR image, which is multiplied by the tentative HDR luminance to generate the final HDR image as

$$\boldsymbol{g}(\boldsymbol{\eta}; \boldsymbol{\psi}) \circ \hat{\boldsymbol{L}}, \tag{9}$$

where $\boldsymbol{g}(\boldsymbol{\eta}; \boldsymbol{\psi})$ represents the luminance-normalized HDR image estimated by LN-Net, $\boldsymbol{\eta}$ is the network input defined

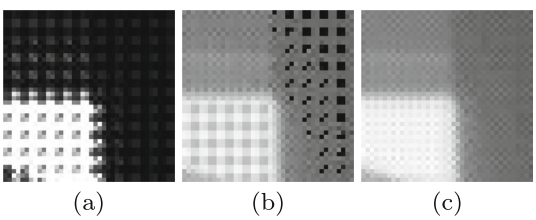


Fig. 4 Examples of **a** the ME-CFA RAW data in the LDR domain, **b** the irradiance CFA RAW data converted by Eq. (6), and **c** the irradiance CFA RAW data after our proposed O/U-pixel correction

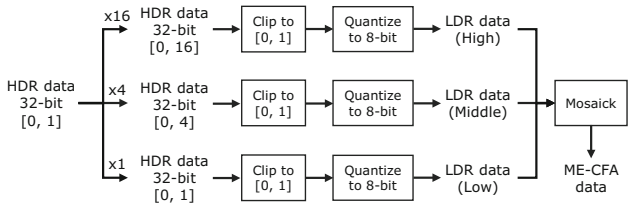


Fig. 5 The ME-CFA data simulation pipeline. High, middle, and low represent the corresponding exposure levels

in Eq. (8), $\boldsymbol{\psi}$ represents the weights for the network, $\hat{\boldsymbol{L}}$ is the tentative HDR luminance, and $\circ$ represents the pixel-wise product. The loss function of the network is an L1 loss in the HDR domain normalized by the ground-truth luminance as

$$\begin{aligned}\mathcal{L}_{\mathrm{HDR}} = &||\boldsymbol{g}(\boldsymbol{\eta};\boldsymbol{\psi}) \circ \hat{\boldsymbol{L}}/\boldsymbol{L} - \boldsymbol{E}/\boldsymbol{L}||_1 \\ &+\lambda_2||\nabla\left(\boldsymbol{g}(\boldsymbol{\eta};\boldsymbol{\psi}) \circ \hat{\boldsymbol{L}}/\boldsymbol{L}\right) - \nabla\left(\boldsymbol{E}/\boldsymbol{L}\right)||_1,\end{aligned} \tag{10}$$

where $\boldsymbol{E}$ is the ground-truth HDR image, $\boldsymbol{L}$ is the ground-truth luminance, ∇ represents the horizontal and vertical gradient operator, $\circ$ represents the pixel-wise product, and λ_2 is a hyper-parameter. The luminance normalization operation (i.e., the division by $\boldsymbol{L}$) is performed in a pixel-by-pixel manner. The second term represents the gradient term, and we empirically set it to $\lambda_2 = 1$ in this paper.

3.4 Network architecture

In our framework, we use two networks: LDR-I-Net and LN-Net. In this paper, we adopt the U-Net architecture [73] for both networks, though one can use any network architecture.

The detailed network architecture of LDR-I-Net is shown in Fig 3. The network inputs are a pair of the sparse sub-mosaicked RAW data and the dense interpolated data. To adapt the data sparseness difference, we insert RAW data adaptation blocks. The RAW data adaptation for the sparse data consists of a convolution layer with the ReLU activation whose kernel size is 7×7, while the adaptation for the interpolated data is a convolution layer with the ReLU activation whose kernel size is 3×3. The outputs of both adaptations are concatenated and then fed into the U-Net architecture with a depth of five. The output channels of LDR-I-Net are $32 \times 32 \times 9$, which include RGB channels of three LDR images.

The architecture of LN-Net is similar, with the difference in input and output channels. As described in Eq. (8), we use both LDR domain data and luminance-normalized domain irradiance data for the inputs of LN-Net. The output channels of LN-Net are $32 \times 32 \times 3$, which include RGB channels of one luminance-normalized HDR image.

4 Experimental results

4.1 Setups

We evaluated our proposed framework using three public HDR image datasets: Funt's dataset [74], Kalantari's dataset [2], and Froehlich's dataset [75]. The details of each dataset will be described in each experimental part.

4.1.1 ME-CFA data simulation

We simulated the ME-CFA data from the ground-truth HDR data, which is provided by 32-bit RGBE image format or 12-bit OpenEXR format. Figure 5 illustrates the ME-CFA data simulation pipeline from 32-bit HDR data. We first scaled the ground-truth HDR data, which is normalized to the range of $[0, 1]$, by 1, 4, and 16 times as $[0, 1]$, $[0, 4]$, and $[0, 16]$, according to the assumed three exposure levels. Then, we clipped each scaled data by $[0, 1]$ and quantized the clipped data by 8-bit depth. By these clipping and quantization processes, three full-resolution LDR data corresponding to high, middle, and low exposures are generated. Finally, we mosaicked the quantized 8-bit LDR data according to the ME-CFA pattern to generate the ME-CFA data. By this pipeline, quantization errors are properly taken into account.

4.1.2 Training setups

In the training phase, we randomly sampled 32×32-sized patches from each training image set and randomly applied each of a horizontal flip, a vertical flip, and swapping of horizontal and vertical axes (transpose) for data augmentation. The used optimizer is Adam [76], where the learning rate was set to 0.001 and the parameters (β_1, β_2) were set to $(0.9, 0.999)$.

Our training phase consists of three steps: (i) We first trained LDR-I-Net only with the inputs of the sub-mosaic RAW data and its linearly interpolated version (see Fig. 3) using the loss function of Eq. (2). (ii) We then trained LN-Net only with the inputs of Eq. (8) using the loss function of Eq. (10). For these two steps, we applied 25,000 times mini-batch updates, where the mini-batch size was set to 32. (iii) We finally trained the two networks jointly with additional 25,000 times mini-batch updates using the both loss functions of Eqs. (2) and (10). During the application phase, we first applied LDR-I-Net to produce tentative HDR luminance for luminance normalization and then applied LN-Net to produce the final reconstructed HDR image.

4.1.3 Evaluation metrics

We used the following five metrics: color PSNR (CPSNR) in the linear HDR domain, CPSNR in the global tone-mapped

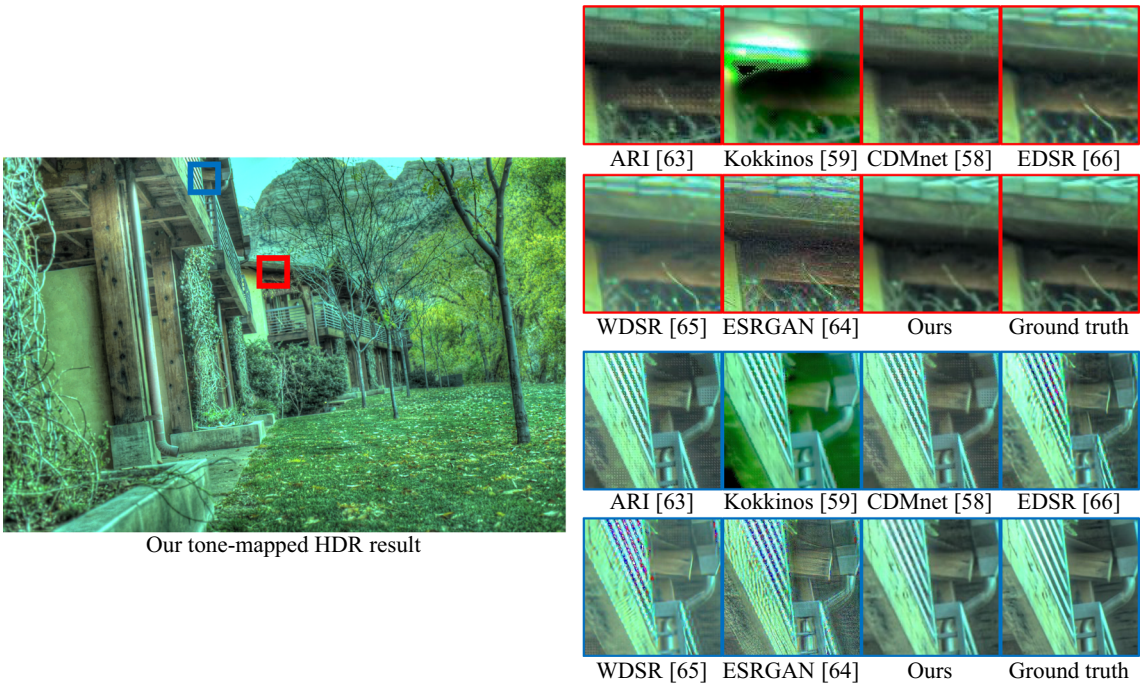


Fig. 6 Visual comparisons on Funt's dataset. The left image is our final HDR image result. The comparisons for a dark region and a bright region are shown in the red box and the blue box, respectively, where we can see that our framework produces better results with fewer visible artifacts

domain (G-CPSNR), CPSNR in the local tome-mapped domain (L-CPSNR), HDR-VDP-2 [77], and luminance-normalized MSE (LN-MSE), which is the MSE normalized by the true luminance. We used the same global tone-mapping function as [2] for G-CPSNR and the MATLAB local tone-mapping function for L-CPSNR. For each dataset, the average metric value of all test images is presented for numerical comparison. For subjective evaluation, we used commercial software, Photomatix,[1] to apply local tone mapping for effective visualization.

4.2 Comparison with other snapshot methods

4.2.1 Dataset

We used Funt's dataset [74] to compare our framework with other snapshot methods. Funt's dataset consists of static scenes with 2140 × 1420 resolution. Among various HDR image datasets as listed in [38], we used Funt's dataset because it contains relatively a large number of HDR images (224 images) generated using the same camera. Each HDR image was generated by Debevec's method [1] using nine LDR images with the EV set of {−4, −3, −2, −1, 0, 1, 2, 3, 4}. We randomly selected 211 images for training and the remaining 13 images for testing.

[1] https://www.hdrsoft.com.

4.2.2 Compared methods

We compared our framework with two combination frameworks. The first framework is a demosaicking-based framework. It first converts the ME-CFA RAW data to the irradiance CFA RAW data and then applies an existing Bayer demosaicking method to the irradiance CFA RAW data. To generate the irradiance CFA RAW data, we applied the same processes as in subsection 3.3, including our proposed O/U-pixel correction since it was confirmed that our pixel correction significantly improves the numerical performance of existing methods. We used state-of-the-art interpolation-based (ARI [78]) and deep learning-based (Kokkions [69] and CDMNet [68]) demosaicking methods for comparison.

The second framework is an LDR-interpolation-based framework. It first interpolates (up-samples) the sub-mosaicked ME-CFA RAW data by an existing super-resolution (SR) method with a scaling factor of four and then performs HDR reconstruction from the interpolated LDR images. We used existing competitive SR methods (ESRGAN [79], WDSR [80], and EDSR [81]) for SR and Debevec's method [1] for the HDR reconstruction.

4.2.3 Results

Figure 6 and Table 1 show the visual and numerical comparisons using Funt's dataset. In Fig. 6, we can observe that the demosaicking-based methods (ARI [78] and CDMNet

Table 1 Comparison with two frameworks that combine existing methods for snapshot HDR imaging

Framework	Method	CPSNR	G-CPSNR	L-CPSNR	HDR-VDP-2	LN-MSE
Demosaicking-based framework: Irradiance CFA RAW data generation → Demosaicking	ARI [78]	46.02	38.04	36.70	75.76	0.072
	Kokkinos [69]	41.01	26.22	26.63	69.32	0.185
	CDMNet [68]	46.19	38.29	37.13	75.72	0.072
LDR-interpolation-based framework: LDR interpolation by SR → HDR reconstruction	ESRGAN [79]	32.31	27.28	24.80	58.54	0.224
	WDSR [80]	35.71	30.92	29.32	60.20	0.378
	EDSR [81]	39.29	32.55	30.03	66.13	0.119
Snapshot HDR reconstruction method	Ours	**50.43**	**43.41**	**42.14**	**81.97**	**0.053**

[68]) generate severe zipper artifacts in the dark area of the red box, while the LDR-interpolation-based methods (ESRGAN [79], WDSR [80], and EDSR [81]) generate severe aliasing artifacts for the high-frequency area of the blue box. These sequential approaches cannot suppress the zipper artifacts and the aliasing artifacts, which are particular to demosaicking and super-resolution, respectively, because these approaches accumulate the errors of the first step and the second step and thus the errors of the first step cannot be corrected in the second step. In contrast, our framework can produce a better result with fewer artifacts by jointly learning the demosaicking and the HDR reconstruction in end-to-end to produce the final reconstructed HDR image. Table 1 demonstrates that our framework can provide the best performance in all metrics, where we can confirm that our framework can preserve high quality at the global and the local tone-mapped domains (G-CPSNR and L-CPSNR). This indicates that our framework can effectively handle dark areas as well as bright areas by the proposed luminance normalization.

4.3 Comparison with state-of-the-art methods using a single or multiple LDR images

4.3.1 Dataset

We used Kalantari's dataset [2] for comparison with state-of-the-art HDR imaging methods using a single or multiple LDR images. The dataset contains 74 scenes for training and 15 scenes for testing with 1500 × 1000 resolution. We used these prepared sets for training and testing. For each scene, the ground-truth HDR image was generated using static LDR images taken with a reference human pose. In contrast, test LDR images were taken with a human motion, including the reference pose as the second exposure. The EV set of $\{-2, 0, 2\}$ or $\{-3, 0, 3\}$ was used for the LDR image acquisition, where EV $= 0$ is used to take the image with the reference pose.

4.3.2 Compared methods

We compared our snapshot framework with state-of-the-art HDR imaging methods using multiple LDR images (Sen [20], Kalantari [2], and Wu [3]) or a single LDR image (HDR-CNN [6], DrTMO [41], and ExpandNet [5]). We used all three LDR images for the multiple-LDR-images-based methods and the second-exposure LDR image for the single-LDR-image-based methods.

4.3.3 Results

Figure 7 shows the visual comparison. We can observe that the multiple-LDR-images-based methods (Kalantari [2], and Wu [3]) generate severe ghost artifacts around the shoulder of the human in the red box, which are due to the arm motions between input LDR images. The artifacts also can be observed in the error map which shows the MSE of RGB irradiance values. The single-LDR-image-based methods (HDRCNN [6], DrTMO [41], and ExpandNet [5]) fail to plausibly inpaint the texture-less areas in the blue box, which is an over-exposed area in the input second-exposure LDR image. In contrast, our framework can reconstruct visually pleasing results without suffering from both ghosting and inpainting artifacts by effectively reconstructing the HDR image from one-shot multi-exposure information obtained using an ME-CFA.

Table 2 shows the numerical comparison. Our framework provides the highest scores in metrics other than L-CPSNR, while the multiple-LDR-images-based methods present better performance for L-CPSNR. This is because these methods have the benefit of having all three-exposure information for each pixel and thus should provide better performance for static regions, which are dominant in each scene of Kalantari's dataset. However, as shown in the visual comparison, these methods are very susceptible to ghost artifacts, which significantly disturb visual perception and make the perceptual HDR-VDP-2 score much lower.

To numerically evaluate the existence of severe artifacts, in Fig. 8, we evaluated the ratio of error pixels whose MSE of

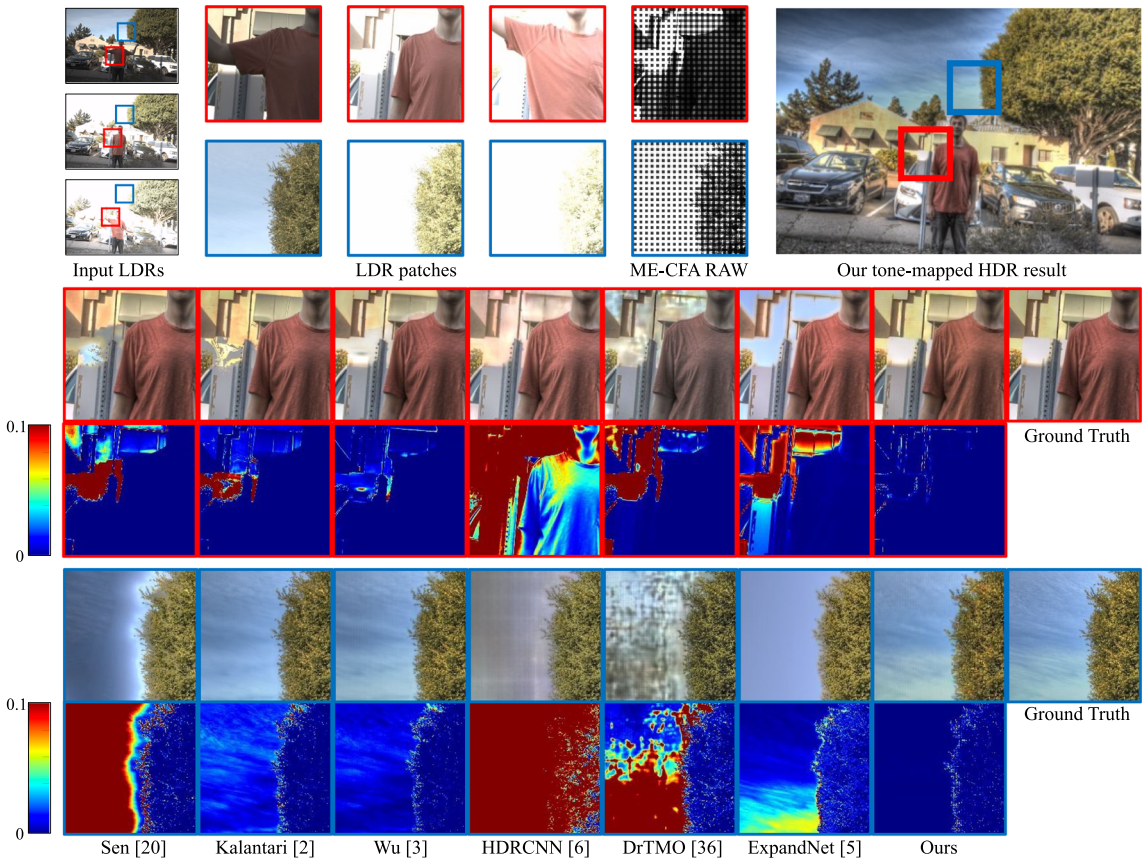


Fig. 7 Visual comparisons on Kalantari's dataset. The patches in the red box show a dynamic region in this scene, and the patches in the blue box show a static but saturated region in the second-exposure image. These patches are prone to cause ghosting or inpainting artifacts. Our proposed framework can generate the HDR image without such artifacts and produce the closest result to the ground truth. The color maps below the images show the error maps representing the MSE of RGB irradiance values. The compared methods generate large errors in the red or the blue box, while our framework generates smaller errors in both areas

Table 2 Comparison with state-of-the-art HDR imaging methods

Input sources	Methods	CPSNR	G-CPSNR	L-CPSNR	HDR-VDP-2	LN-MSE
Multiple LDR images	Sen [20]	35.88	38.43	36.16	60.22	0.067
	Kalantari [2]	38.67	40.24	**38.26**	63.35	0.059
	Wu [3]	38.33	40.15	37.95	64.49	0.062
Single LDR image (Second exposure)	HDRCNN [6]	13.03	14.39	34.83	54.35	3.771
	DrTMO [41]	17.61	13.93	24.65	56.59	14.027
	ExpandNet [5]	22.22	22.59	28.01	57.23	1.113
ME-CFA RAW data	Ours	**44.10**	**41.27**	36.26	**68.79**	**0.036**

RGB irradiance values is larger than the threshold of the horizontal axis. From the result, we can clearly observe that our snapshot framework can generate HDR images with much fewer error pixels that have significantly large errors, such as ghosting and inpainting artifacts appearing in existing methods.

4.4 Comparison using an HDR video dataset

4.4.1 Dataset

To evaluate our proposed framework for dynamic situations, we used Froehlich's HDR video dataset [75]. The videos in this dataset were taken by a camera system, in which a beam splitter is located to capture two images of different

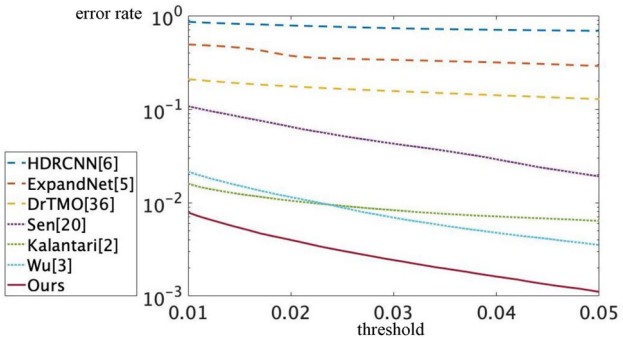


Fig. 8 The ratio of error pixels whose MSE of RGB irradiance values is larger than the threshold. At each of the threshold levels, the error rate of our framework is lower than that of the other compared methods

exposure levels at the same time. These two images were then processed to generate an HDR image. All video frames are provided in 12-bit OpenEXR format with 1920 × 1080 resolution. This dataset contains HDR scenes of five scene categories, and we selected 17 video clips for evaluation, which include various challenging situations such as a night scene with bright or flickering light sources, a scene with fast-moving objects, and a sunlight scene with substantially changing brightness levels.

4.4.2 Compared methods

We compared our framework with HDR imaging methods using multiple LDR images (Debevec [1], Sen [20], Kalantari [2], and Wu [3]). For the learning methods of Kalantari, Wu, and Ours, we applied the same trained models as Sect. 4.3, which were trained using Kalantari dataset. Because the multi-frame methods of Sen, Kalantari, and Wu are alignment-based methods and they assume the middle-exposure image (i.e., the second-exposure image for three inputs) as a reference image for the alignment in their default implementation and usage, we converted every three adjacent frames in the HDR video to three input LDR images with the EV set of {0, 2, 4}. In this case, because every successive three frames were used to generate one HDR image aligned to the second-exposure frame, the frame per second (fps) of the generated video by these methods reduces to 10fps, which is one-third of the 30fps of the original video.

For our proposed framework, each frame in the HDR video was used to generate ME-CFA RAW data as described in Sect. 4.1.1 and then HDR images were generated by the proposed framework trained using Kalantari's dataset. The fps of the generated video by the proposed framework is 30fps, which is the same as that of the original video because one input frame is used to generate one output frame.

As a consequence, the HDR videos generated by the compared methods are 10fps, while ours and ground-truth HDR videos are 30fps. To compare each method in both 10pfs and 30fps domains, we raised the fps of the videos generated by the compared methods from 10fps to 30fps by duplicating the second-exposure reference frames, for which aligned HDR image results were generated, to the other frames. We also reduced the fps of the ground truth and the video generated by our framework from 30fps to 10fps by sampling every three frames corresponding to the second-exposed reference frames.

4.4.3 Results

Tables 3 and 4 show the numerical comparison on 17 clips in the HDR video dataset. The averaged evaluation values are shown in the tables. In the 10fps evaluation of Table 3, our framework shows the best scores for CPSNR, L-CPSNR, and HDR-VDP-2, while Sen's method shows the highest scores for G-CPSNR and LN-MSE. The higher scores by Sen's method can be expressed with the same reason as discussed in Sect. 4.3, i.e., multiple-LDR-images-based methods have the advantage of having all three exposure information in the pixels of static regions. However, in the 30fps evaluation of Table 4, the scores of the compared multiple-LDR-images-based methods tend to be significantly lower than that of 10fps, while our framework provides almost the same scores and the highest scores for all the metrics in the 30fps evaluation. This clearly indicates the advantage of our snapshot framework that produces the same fps video as the ground truth without the necessity of any alignment and duplication between the frames.

Figure 9 shows the visual comparison for some scenes. We can see that the compared methods generate severe ghost artifacts around the hand, the fireworks, and the bonfire in the scenes. In contrast, our framework can generate higher-quality HDR video frames without such artifacts, owing to the one-shot nature of our framework. The video results can be seen in the supplemental video.

Regarding the computational time, under the computational environment of Intel Core i7-6850K CPU and NVIDIA GeForce GTX1080 GPU, our non-optimized implementation currently takes 0.38 s per frame with 1920 × 1080 resolution, which is slightly faster than the second-best Wu's method (0.48 s per frame). Although Kalantari's method is faster (0.16 s per frame) than ours, it provides much lower numerical performance as shown in Tables 3 and 4. The computational time is expected to become faster if we use more latest GPUs such as NVIDIA GeForce 4000 series. As a reference, the non-learning-based methods of Debevec and Sen take 0.70 s and 69.35 s per frame, respectively, using CPU implementation with MATLAB.

Table 3 Numerical comparison on HDR video dataset with 10fps evaluation

Input frame	Methods	CPSNR	G-CPSNR	L-CPSNR	HDR-VDP-2	LN-MSE
Every successive three frames with three exposures	Debevec [1]	22.78	31.52	25.12	49.32	0.989
	Sen [20]	29.81	**41.75**	33.55	54.12	**0.005**
	Kalantari [2]	20.19	22.73	21.08	46.05	2.430
	Wu [3]	30.05	39.73	32.50	58.01	0.009
Every frame with ME-CFA RAW	Ours	**34.31**	39.30	**34.12**	**65.59**	0.006

Table 4 Numerical comparison on HDR video dataset with 30fps evaluation

Input video	Methods	CPSNR	G-CPSNR	L-CPSNR	HDR-VDP-2	LN-MSE
Every successive three frames with three exposures	Debevec [1]	22.99	31.43	25.38	49.89	1.120
	Sen [20]	26.79	34.12	27.89	50.76	0.750
	Kalantari [2]	19.74	22.09	20.49	45.57	3.100
	Wu [3]	27.39	33.54	27.66	53.14	0.779
Every frame with ME-CFA RAW	Ours	**34.31**	**39.31**	**34.11**	**65.58**	**0.006**

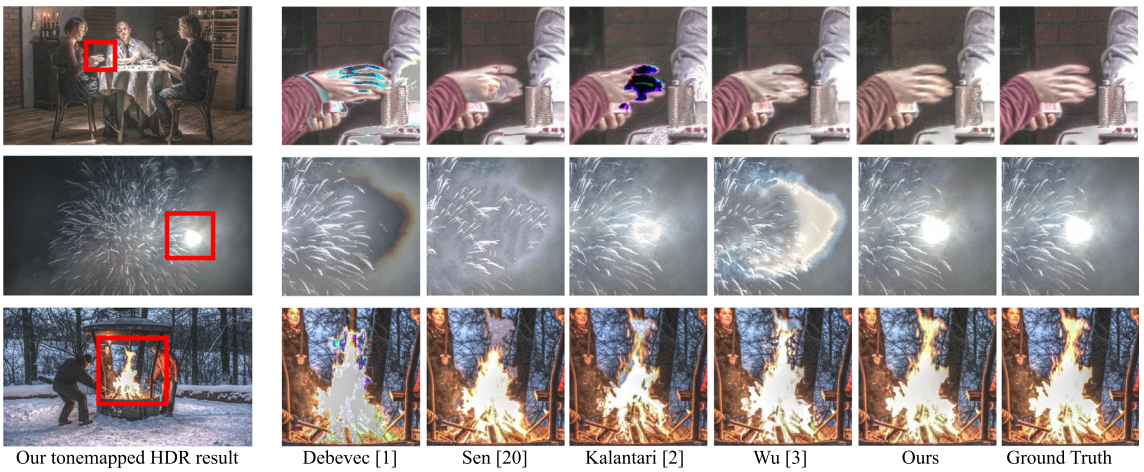


Fig. 9 Visual comparison using an HDR video dataset. The scene POKER FULLSHOT on the top row includes the fast motion of hands shuffling cards. The scenes CAROUSEL FIREWORKS and FIRE PLACE on the second and the last rows include drastically changing luminance levels for each frame. We can see that compared methods using multiple LDR images generate severe ghost artifacts, while our framework can produce the images closest to the ground truths

4.5 Validation studies of our framework

To confirm the effectiveness of each proposed component, we performed validation studies using Funt's dataset.

4.5.1 Ablation study on HDR image reconstruction

Table 5 shows the result of an ablation study on HDR image reconstruction. In the case without input data normalization, we do not perform luminance estimation and luminance normalization in the HDR image reconstruction. From the comparison with this case, we can see that the input data normalization by the estimated luminance contributes to effective feature extraction and the generation of plausible HDR image results with higher numerical scores. We can also confirm that the proposed O/U-pixel correction certainly contributes to the performance improvements in all the evaluated metrics.

4.5.2 Comparison of loss computation domains

Table 6 shows the comparison of loss computation domains. The loss in the standard linear HDR domain presents relatively a high value for CPSNR, but lower G-CPSNR and L-CPSNR values. This is because the loss in the linear domain tends to disregard the errors in dark areas, which

Table 5 Ablation study on HDR image reconstruction

	CPSNR	G-CPSNR	L-CPSNR	HDR-VDP-2	LN-MSE
Ours without input normalization	47.56	39.82	39.14	77.88	0.067
Ours without O/U-pixel correction	42.85	40.66	37.14	77.25	0.062
Ours	**50.07**	**42.94**	**41.74**	**81.71**	**0.054**

Table 6 Comparison of loss computation domains

	CPSNR	G-CPSNR	L-CPNSR	HDR-VDP-2	LN-MSE
Linear HDR	49.57	41.48	40.55	80.71	0.059
Log HDR	48.32	42.28	40.88	79.01	0.058
Global tone-mapped HDR	48.90	42.45	41.02	79.83	0.056
Luminance-normalized HDR (ours)	**50.07**	**42.94**	**41.74**	**81.71**	**0.054**

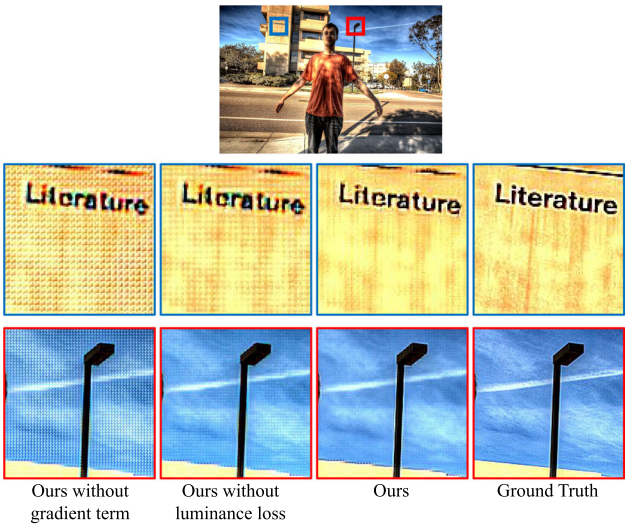


Fig. 10 The visual effect using the luminance loss and the gradient terms

are significantly enhanced in a global or local tone-mapped domain, lowering the G-CPSNR and L-CPSNR values. The losses in the log and the global tone-mapped domains improve the G-CPSNR and the L-CPSNR performance compared to the linear domain. The loss in our proposed luminance-normalized domain provides further better performance in G-CPSNR and L-CPSNR by considering the relative local contrasts. Furthermore, the higher HDR-VDP-2 score shows that the luminance-normalized domain successfully produces visually higher-quality HDR images.

4.5.3 Ablation study on loss computation

One of the main challenges of snapshot HDR imaging is to reduce zipper artifacts, which are caused by a very sparse sampling of each color-exposure component and many saturated/blacked-out pixels in ME-CFA RAW data. Furthermore, zipper artifacts may occur even for uniform areas without textures because of the differences in the quantization levels of the three exposure images, meaning that converted sensor irradiance values in the uniform area do not match completely among the three exposure levels.

Our framework suppresses zipper artifacts by the gradient terms and the luminance loss in the loss functions. Figure 10 shows the visual comparison of zipper artifacts. In the case without the gradient terms, we removed the gradient terms in Eqs. (3), (4), and (10), i.e., the parameters λ_1 and λ_2 are set to 0. In the case without the luminance loss, we set the parameter α of the LDR loss in Eq. (2) to 1. In this case, the errors of luminance are not considered in the loss computation, which corresponds to the method in our previous version [13].

From Fig. 10, we can see that the case without the gradient terms generates severe zipper artifacts. Although the case without the luminance loss reduces the zipper artifacts, they are still apparent. In contrast, our proposed loss computation can generate more accurate tentative HDR luminance and obtain the HDR image result with much fewer zipper artifacts. Table 7 shows the numerical comparison of these cases. From the table, we can confirm that both the gradient terms and the luminance loss contribute to the improvement of all the evaluated metrics.

4.6 Limitation

In our results, zipper artifacts still remain in some areas. This is because of the very challenging nature of snapshot HDR reconstruction with a very sparse sampling of each color-exposure component and many saturated or blacked-out pixels. Furthermore, in the snapshot HDR problem, zipper artifacts may occur even for uniform areas without textures because of the differences in the quantization levels of three exposure images, meaning that converted sensor irradiance

Table 7 Ablation study on loss computation

	CPSNR	G-CPSNR	L-CPSNR	HDR-VDP-2	LN-MSE
Ours without gradient terms	38.90	39.80	37.38	68.84	0.129
Ours without luminance loss	49.60	42.59	41.44	80.86	0.055
Ours	**50.07**	**42.94**	**41.74**	**81.71**	**0.054**

values in the uniform area do not match completely among the three exposure levels.

5 Conclusion

In this paper, we have proposed a novel deep learning-based framework that can effectively address the joint demosaicking and HDR reconstruction problem for snapshot HDR imaging using an ME-CFA. We have introduced the idea of luminance normalization that simultaneously enables effective loss computation and input data normalization to learn the HDR image reconstruction network by considering relative local image contrasts.

In the experimental comparison with other snapshot methods, our framework achieves more than 4dB CPSNR improvement in the evaluation of the linear HDR domain and demonstrates that it can produce HDR images with much fewer visual zipper artifacts. In the experimental comparison with existing HDR imaging methods using multiple LDR images, we have demonstrated that our framework can produce HDR images without severe ghost artifacts, which are apparent for the existing methods.

Our snapshot method is particularly useful for dynamic scenes requiring HDR imaging technology. A typical example is a driving application that encounters situations with large intensity variations, e.g., the entrance and the exit of tunnels and night scenes with car headlights. Experimental validations on these situations are one of our future works. Another future work is a joint design of the ME-CFA and the reconstruction network to simultaneously optimize the snapshot sensor design and HDR imaging network to further reduce zipper artifacts.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00371-023-03032-4.

Data availability The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

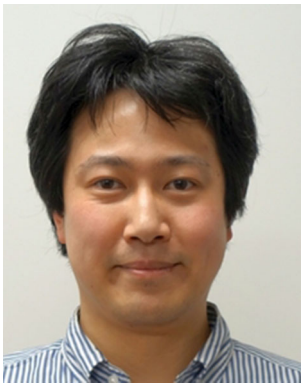
1. Debevec, P., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: Proceedings of SIGGRAPH, pp. 1–10 (1997)
2. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Trans. Graph. **36**(4), 144 (2017)
3. Wu, S., Xu, J., Tai, Y.-W., Tang, C.-K.: Deep high dynamic range imaging with large foreground motions. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 121–135 (2018)
4. Yan, Q., Gong, D., Shi, Q., van den Hengel, A., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1751–1760 (2019)
5. Marnerides, D., Bashford-Rogers, T., Hatchett, J., Debattista, K.: ExpandNet: a deep convolutional neural network for high dynamic range expansion from low dynamic range content. Comput. Graph. Forum **37**(2), 37–49 (2018)
6. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: HDR image reconstruction from a single exposure using deep CNNs. ACM Trans. Graph. **36**(6), 1–15 (2017)
7. Lee, S., An, G.H., Kang, S.-J.: Deep chain HDRI: reconstructing a high dynamic range image from a single low dynamic range image. IEEE Access **6**, 49913–49924 (2018)
8. Cho, H., Kim, S.J., Lee, S.: Single-shot high dynamic range imaging using coded electronic shutter. Comput. Graph. Forum **33**(7), 329–338 (2014)
9. Choi, I., Baek, S.-H., Kim, M.H.: Reconstructing interlaced high-dynamic-range video using joint learning. IEEE Trans. Image Process. **26**(11), 5353–5366 (2017)
10. Narasimhan, S.G., Nayar, S.K.: Enhancing resolution along multiple imaging dimensions using assorted pixels. IEEE Trans. Pattern Anal. Mach. Intell. **27**(4), 518–530 (2005)
11. Nayar, S.K., Mitsunaga, T.: High dynamic range imaging: spatially varying pixel exposures. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)
12. Eilertsen, G., Mantiuk, R.K., Unger, J.: Real-time noise-aware tone mapping. ACM Trans. Graph. **34**(6), 1–15 (2015)
13. Suda, T., Tanaka, M., Monno, Y., Okutomi, M.: Deep snapshot HDR imaging using multi-exposure color filter array. In: Proceedings of the Asian Conference on Computer Vision (ACCV)
14. Ma, K., Duanmu, Z., Yeganeh, H., Wang, Z.: Multi-exposure image fusion by optimizing a structural similarity index. IEEE Trans. Comput. Imaging **4**(1), 60–72 (2017)
15. Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., Zhang, L.: Robust multi-exposure image fusion: a structural patch decomposition approach. IEEE Trans. Image Process. **26**(5), 2519–2532 (2017)
16. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion: a simple and practical alternative to high dynamic range photography. Comput. Graph. Forum **28**(1), 161–171 (2009)
17. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Trans. Graph. **35**(6), 1–12 (2016)

18. Hafner, D., Demetz, O., Weickert, J.: Simultaneous HDR and optic flow computation. In Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 2065–2070 (2014)
19. Hu, J., Gallo, O., Pulli, K., Sun, X.: HDR deghosting: how to deal with saturation? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1163–1170 (2013)
20. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based HDR reconstruction of dynamic scenes. ACM Trans. Graph. **31**(6), 203 (2012)
21. Lee, C., Li, Y., Monga, V.: Ghost-free high dynamic range imaging via rank minimization. IEEE Signal Process. Lett. **21**(9), 1045–1049 (2014)
22. Oh, T.-H., Lee, J.-Y., Tai, Y.-W., Kweon, I.S.: Robust high dynamic range imaging by rank minimization. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1219–1232 (2014)
23. Kalantari, N.K., Ramamoorthi, R.: Deep HDR video from sequences with alternating exposures. Comput. Graph. Forum **38**(2), 193–205 (2019)
24. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: Proceedings of the IEEE International Conference on Computational Photography (ICCP), pp. 170–177 (2019)
25. Ram Prabhakar, K., Sai Srikar, V., Venkatesh Babu, R.: DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 4724–4732 (2017)
26. Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I., Zhang, Y.: Multi-scale dense networks for deep high dynamic range imaging. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 41–50 (2019)
27. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep HDR imaging via a non-local network. IEEE Trans. Image Process. **29**, 4308–4322 (2020)
28. Niu, Y., Wu, J., Liu, W., Guo, W., Lau, R.W.: HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. IEEE Trans. Image Process. **30**, 3885–3896 (2021)
29. Prabhakar, K.R., Senthil, G., Agrawal, S., Babu, R.V., Gorthi, R.K.S.S.: Labeled from unlabeled: exploiting unlabeled data for few-shot deep HDR deghosting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4875–4885 (2021)
30. Song, J.W., Park, Y.-I., Kong, K., Kwak, J., Kang, S.-J.: Selective TransHDR: transformer-based selective HDR imaging using ghost region mask. In: Proceedings of the European Conference on Computer Vision (ECCV),pp. 288–304 (2022)
31. Liu, Z., Wang, Y., Zeng, B., Liu, S.: Ghost-free high dynamic range imaging with context-aware transformer. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 344–360 (2022)
32. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: The state of the art in HDR deghosting: a survey and evaluation. Comput. Graph. Forum **34**(2), 683–707 (2015)
33. Wang, L., Yoon, K.-J.: Deep learning for HDR imaging: state-of-the-art and future trends. arXiv:2110.10394 (2021)
34. Ogino, Y., Tanaka, M., Shibata, T., Okutomi, M.: Super high dynamic range video. In: Proceedings of the International Conference on Pattern Recognition (ICPR), pp. 4208–4213 (2016)
35. Tocci, M.D., Kiser, C., Tocci, N., Sen, P.: A versatile HDR video production system. ACM Trans. Graph. **30**(4), 1–9 (2011)
36. Han, J., Zhou, C., Duan, P., Tang, Y., Xu, C., Xu, C., Huang, T., Shi, B.: Neuromorphic camera guided high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),pp. 1730–1739 (2020)
37. Yang, X., Xu, K., Song, Y., Zhang, Q., Wei, X., Lau, R.W.: Image correction via deep reciprocating HDR transformation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1798–1807 (2018)
38. Moriwaki, K., Yoshihashi, R., Kawakami, R., You, S., Naemura, T.: Hybrid loss for learning single-image-based HDR reconstruction. arXiv:1812.07134 (2018)
39. Kim, S.Y., Kim, D.-E., Kim, M.: ITM-CNN: learning the inverse tone mapping from low dynamic range video to high dynamic range displays using convolutional neural networks. In: Proceedings of the Asian Conference on Computer Vision (ACCV), pp. 395–409 (2018)
40. Santos, M.S., Ren, T.I., Kalantari, N.K.: Single image HDR reconstruction using a CNN with masked features and perceptual loss. ACM Trans. Graph. (TOG) **39**(4), 80–1 (2020)
41. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping. ACM Trans. Graph. **36**(6), 177 (2017)
42. Lee, S., Hwan An, G., Kang, S.-J.: Deep recursive HDRI: inverse tone mapping using generative adversarial networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 613–628 (2018)
43. Liu, Y.-L., Lai, W.-S., Chen, Y.-S., Kao, Y.-L., Yang, M.-H.,Chuang, Y.-Y., Huang, J.-B.: Single-image HDR reconstruction by learning to reverse the camera pipeline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2020)
44. Zheng, Z., Ren, W., Cao, X., Wang, T., Jia, X.: Ultra-high-definition image HDR reconstruction via collaborative bilateral learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 4449–4458 (2021)
45. Chen, X., Zhang, Z., Ren, J.S., Tian, L., Qiao, Y., Dong, C.: A new journey from SDRTV to HDRTV. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV),pp. 4500–4509 (2021)
46. Cheng, Z., Wang, T., Li, Y., Song, F., Chen, C., Xiong, Z.: Towards real-world HDRTV reconstruction: a data synthesis-based approach. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 199–216 (2022)
47. Sun, Q., Tseng, E., Fu, Q., Heidrich, W., Heide, F.: Learning rank-1 diffractive optics for single-shot high dynamic range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2020)
48. Metzler, C.A., Ikoma, H., Peng, Y., Wetzstein, G.: Deep optics for single-shot high-dynamic-range imaging. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), (2020)
49. Gu, J., Hitomi, Y., Mitsunaga, T., Nayar, S.: Coded rolling shutter photography: flexible space-time sampling. In: Proceedings of the IEEE International Conference on Computational Photography (ICCP), pp. 1–8 (2010)
50. Uda, S., Sakaue, F., Sato, J.: Variable exposure time imaging for obtaining unblurred HDR images. IPSJ Trans. Comput. Vis. Appl. **8**(1), 1–7 (2016)
51. Alghamdi, M., Fu, Q., Thabet, A., Heidrich, W.: Reconfigurable snapshot HDR imaging using coded masks and inception network. In: Proceedings of the Vision, Modeling, and Visualization (VMV), pp. 1–9 (2019)
52. Nagahara, H., Sonoda, T., Liu, D., Gu, J.: Space-time-brightness sampling using an adaptive pixel-wise coded exposure. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1834–1842 (2018)
53. Serrano, A., Heide, F., Gutierrez, D., Wetzstein, G., Masia, B.: Convolutional sparse coding for high dynamic range imaging. Comput. Graph. Forum **35**(2), 153–163 (2016)
54. Go, C., Kinoshita, Y., Shiota, S, Kiua, H.: Image fusion for single-shot high dynamic range imaging with spatially varying exposures. In: Proceedings of the Asia-Pacific Signal and Information Process-

ing Association Annual Summit and Conference (APSIPA ASC), pp. 1082–1086 (2018)
55. Hajisharif, S., Kronander, J., Unger, J.: Adaptive dualISO HDR reconstruction. EURASIP J. Image Video Process. **2015**(41), 1–13 (2015)
56. Heide, F., Steinberger, M., Tsai, Y.-T., Rouf, M., Pająk, D., Reddy, D., Gallo, O., Liu, J., Heidrich, W., Egiazarian, K., Kautz, J., Pulli, K.: FlexISP: a flexible camera image processing framework. ACM Trans. Graph. **33**(6), 1–13 (2014)
57. Aguerrebere, C., Almansa, A., Delon, J., Gousseau, Y., Musé, P.: A Bayesian hyperprior approach for joint image denoising and interpolation, with an application to HDR imaging. IEEE Trans. Comput. Imaging **3**(4), 633–646 (2017)
58. Aguerrebere, C., Almansa, A., Gousseau, Y., Delon, J., Muse, P.: Single shot high dynamic range imaging using piecewise linear estimators. In: Proceedings of the IEEE International Conference on Computational Photography (ICCP), pp. 1–10 (2014)
59. An, V.G., Lee, C.: Single-shot high dynamic range imaging via deep convolutional neural network. In: Proceedings of the Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), pp. 1768–1772 (2017)
60. Rouf, M., Ward, R.K.: High dynamic range imaging with a single exposure-multiplexed image using smooth contour prior. In: Proceedings of the IS&T International Symposium on Electronic Imaging (EI), pp. 440:1–6 (2018)
61. Cheng, C.-H., Au, O.C., Cheung, N.-M., Liu, C.-H., Yip, K.-Y.: High dynamic range image capturing by spatial varying exposed color filter array with specific demosaicking algorithm. In: Proceedings of the IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM), pp. 648–653 (2009)
62. Cogalan, U., Akyüz, A.O.: Deep joint deinterlacing and denoising for single shot dual-ISO HDR reconstruction. IEEE Trans. Image Process. **29**, 7511–7524 (2020)
63. Vien, A.G., Lee, C.: Exposure-aware dynamic weighted learning for single-shot HDR imaging. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 435–452 (2022)
64. Martel, J.N.P., Müller, L.K., Carey, S.J., Dudek, P., Wetzstein, G.: Neural sensors: learning pixel exposures for HDR imaging and video compressive sensing with programmable sensors. IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1642–1653 (2020)
65. Vien, A.G., Lee, C.: Single-shot high dynamic range imaging via multiscale convolutional neural network. IEEE Access **9**, 70369–70381 (2021)
66. Xu, Y., Liu, Z., Wu, X., Chen, W., Wen, C., Li, Z.: Deep joint demosaicing and high dynamic range imaging within a single shot. IEEE Trans. Circuits Syst. Video Technol. **32**(7), 4255–4270 (2022)
67. Bayer, B.E.: Color imaging array, US patent 3971065, (1976)
68. Cui, K., Jin, Z., Steinbach, E.: Color image demosaicking using a 3-stage convolutional neural network structure. In: Proceedings of the IEEE International Conference on Image Processing (ICIP), pp. 2177–2181 (2018)
69. Kokkinos, F., Lefkimmiatis, S.: Deep image demosaicking using a cascade of convolutional residual denoising networks. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 317–333 (2018)
70. Grossberg, M.D., Nayar, S.K.: What is the space of camera response functions? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1–8 (2003)
71. Kang, H.R.: Computational Color Technology. SPIE Press, Bellingham (2006)
72. Henz, B., Gastal, E.S., Oliveira, M.M.: Deep joint design of color filter arrays and demosaicing. Compu. Graph. Forum **37**(2), 389–399 (2018)
73. Ronneberger, O., Fischer, P., Brox, T.: U-net: convolutional networks for biomedical image segmentation. In: Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241 (2015)
74. Funt, B., Shi, L.: The rehabilitation of MaxRGB. In: Proceedings of the Color and Imaging Conference (CIC), pp. 256–259 (2010)
75. Froehlich, J., Grandinetti, S., Eberhardt, B., Walter, S., Schilling, A., Brendel, H.: Creating cinematic wide gamut HDR-video for the evaluation of tone mapping operators and HDR-displays. Proc. SPIE **9023**, 279–288 (2014)
76. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014)
77. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Trans. Graph. **30**(4), 1–13 (2011)
78. Monno, Y., Kiku, D., Tanaka, M., Okutomi, M.: Adaptive residual interpolation for color and multispectral image demosaicking. Sensors **17**(12), 2787 (2017)
79. Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., Qiao, Y., Change Loy, C.: ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European Conference on Computer Vision Workshops (ECCVW), pp. 1–16 (2018)
80. Fan, Y., Yu, J., Huang, T.S.: Wide-activated deep residual networks based restoration for BPG-compressed images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition workshops (CVPRW), pp. 2621–2624 (2018)
81. Lim, B., Son, S., Kim, H., Nah, S., Mu Lee, K.: Enhanced deep residual networks for single image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 1132–1140 (2017)

Yutaro Okamoto received his bachelor's and master's degrees from the Department of Control and Systems Engineering, Tokyo Institute of Technology, in 2020 and 2022, respectively. He is currently working at Advanced Technology Research Institute, Kyocera Corporation, in Japan.

Masayuki Tanaka received his bachelor's and master's degrees in control engineering and a doctoral degree from Tokyo Institute of Technology in 1998, 2000, and 2003. He was a software engineer at Agilent Technologies from 2003 to 2004. He was a Research Scientist at Tokyo Institute of Technology from 2004 to 2008. He was an Associate Professor at the Graduate School of Science and Engineering, Tokyo Institute of Technology, from 2008 to 2016. He was a Visiting Scholar at Department of Psychology, Stanford University, from 2013 to 2014. He was an Associate Professor at School of Engineering, Tokyo Institute of Technology, from 2016 to 2017. He was a Senior Researcher at National Institute of Advanced Industrial Science and Technology from 2017 to 2020. He was an Associate Professor at School of Engineering, Tokyo Institute of Technology, from 2020 to 2023. Since 2023, he has been a Professor at Graduate Major in Engineering Sciences and Design, Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology.



Yusuke Monno received the B.E., M.E., and Ph.D. degrees from Tokyo Institute of Technology, Tokyo, Japan, in 2010, 2011, and 2014, respectively. From Nov. 2013 to Mar. 2014, he joined the Image and Visual Representation Group at École Polytechnique Fédérale de Lausanne as a research internship student. He is currently a Specially Appointed Associate Professor with the Department of Systems and Control Engineering, School of Engineering, Tokyo Institute of Technology. His research interests are in both theoretical and practical aspects of image processing, computer vision, and biomedical engineering.

Masatoshi Okutomi received the B.Eng. degree from the Department of Mathematical Engineering and Information Physics, the University of Tokyo, Tokyo, Japan, in 1981, and the M.Eng. degree from the Department of Control Engineering, Tokyo Institute of Technology, Tokyo, in 1983. He joined the Canon Research Center, Canon Inc., Tokyo, in 1983. From 1987 to 1990, he was a Visiting Research Scientist with the School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA. He received the Dr.Eng. degree from Tokyo Institute of Technology, in 1993, for his research on stereo vision. Since 1994, he has been with Tokyo Institute of Technology, where he is currently a Professor with the Department of Systems and Control Engineering, the School of Engineering.