Joint Demosaicing and Deghosting of Time-Varying Exposures for Single-Shot HDR Imaging

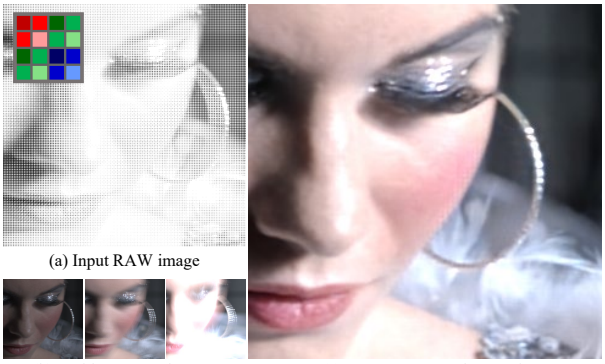
Jungwoo Kim Min H. Kim
KAIST

Abstract

The quad-Bayer patterned image sensor has made significant improvements in spatial resolution over recent years due to advancements in image sensor technology. This has enabled single-shot high-dynamic-range (HDR) imaging using spatially varying multiple exposures. Popular methods for multi-exposure array sensors involve varying the gain of each exposure, but this does not effectively change the photoelectronic energy in each exposure. Consequently, HDR images produced using gain-based exposure variation may suffer from noise and details being saturated. To address this problem, we intend to use time-varying exposures in quad-Bayer patterned sensors. This approach allows long-exposure pixels to receive more photon energy than short- or middle-exposure pixels, resulting in higher-quality HDR images. However, time-varying exposures are not ideal for dynamic scenes and require an additional deghosting method. To tackle this issue, we propose a single-shot HDR demosaicing method that takes time-varying multiple exposures as input and jointly solves both the demosaicing and deghosting problems. Our method uses a feature-extraction module to handle mosaiced multiple exposures and a multiscale transformer module to register spatial displacements of multiple exposures and colors. We also created a dataset of quad-Bayer sensor input with time-varying exposures and trained our network using this dataset. Results demonstrate that our method outperforms baseline HDR reconstruction methods with both synthetic and real datasets. With our method, we can achieve high-quality HDR images in challenging lighting conditions.

1. Introduction

Recent advancements in image sensor technology, such as interlaced and quad-Bayer patterned image sensors, have facilitated the development of multi-exposure color filter arrays, which enable single-shot high-dynamic-range (HDR) imaging. Single-shot HDR imaging takes only one image to generate an HDR image as input. This imaging technology has successfully expanded the dynamic range of captured images beyond conventional low-dynamic-range (LDR) images. However, multi-exposure color filter array sensors must share pixel budgets to capture not only different colors but also different exposures, which can result in severe degradation of spatial resolution. The reconstructed HDR images may experience a reduction in spatial resolution by half due to the use of color filters and multiple exposures. Current single-shot methods have focused on enhancing spatial resolution by addressing multi-exposure sampling artifacts, such as mosaicing [1, 17, 37] or interlacing artifacts [12, 10, 35, 4].



(a) Input RAW image
(b) Multi-exposure images
(c) HDR reconstruction result

Figure 1. (a) Input quad-Bayer patterned RAW image with different exposure times and colors. (b) Three multi-exposure images. (c) Our method result. We jointly solve demosaicing and deblurring problems to achieve a high-quality single-shot HDR image from the quad-Bayer pattern.

Current single-shot HDR imaging techniques use gain-based exposure variation to capture different levels of exposures simultaneously, but this approach provides only a minor improvement in dynamic range compared to traditional static HDR imaging. Gain-based exposure variation does not capture different amounts of photo energy, even with a long exposure input, which limits the dynamic range of reconstructed HDR images. Traditional HDR imaging captures time-varying multi-exposures, which allows for a wider range of photon energy and high signal-to-noise ratio in reconstructed HDR images. However, this approach is not applicable to single-shot HDR imaging of dynamic scenes or camera motions due to motion blur. Recent HDR

deghosting methods [18, 40, 42, 43, 32, 28, 25] have addressed the motion blur problem in HDR video input, but they rarely consider the varying amounts of motion blur with time-varying multiple exposures.

Two main challenges must be addressed to achieve high-quality demosaicing of time-varying multiple exposures in single-shot HDR imaging. The first challenge is the double mosaicing architecture of multiple exposure times and color filters, which results in sparse input pixel observations that severely degrade the spatial resolution in reconstructed HDR images. The second challenge arises from spatially-varying motion blur, which is particularly prevalent in long-exposure pixels on the array. This leads to spatially inconsistent ghost artifacts that degrade the edge details of moving objects or scenes in the resulting images.

In this work, we propose a novel single-shot HDR imaging technique that tackles the demosaicing and deghosting challenges of time-varying multiple exposures captured by quad-Bayer patterned sensors (Figure 1). To accomplish this, we created a dataset of time-varying quad-Bayer patterned HDR sensor inputs from existing HDR video datasets. Our demosaicing network comprises feature-extraction and U-net style multi-scale transformer modules. For each exposure level observation, we use a residual architecture to convert them into channel-wise attention features with a doubled spatial resolution. We also address spatial misalignment and motion blur in longer exposure channels through the transformer architecture by acquiring query, key, and value vectors from various sources in the transformer block. In contrast to gain-based single-shot approaches, our method takes into account the physical differences in exposure times in the quad-Bayer array, which helps mitigate the varying degrees of motion blur in input images. Our approach learns the spatial relationship among differently-blurred multiple exposures and registers them with high precision to prevent ghosting artifacts. We also design the transformer block to be configured coarse-to-fine, taking into account pixel displacement that ranges from the nearest neighboring pixels to further ones.

The proposed method overcomes the degradation of spatial resolution caused by the mosaiced exposure and color Bayer patterns and successfully reconstructs high-resolution HDR images from time-varying multiple exposure inputs in single-shot HDR imaging. Results from various real and synthetic scenes demonstrate the superiority of our method over baseline HDR reconstruction methods.

2. Related Work

Multi-shot HDR reconstruction. Traditional HDR imaging methods capture input through exposure bracketing [6]. However, when objects or scenes move during exposure bracketing, *ghosting* artifacts occur. They mitigate these artifacts by estimating object or camera motion or measurement outliers. Jacobs et al. [16] estimate object movement with weighted variance. Pece et al. [31] reject moving pixel of exposure sequence using median threshold bitmap. Heo et al. [13] estimate ghost regions with joint probability density functions between the reference image and other images. Grosch et al. [9] align multiple images with median threshold bitmap and remove moving pixel with error map of color difference. Oh et al. [29] align LDR images and detect outliers exploiting the rank minimization algorithm.

Also, the flow-based approach has been extended to HDR video reconstruction. Optical flow is used by Bogoni et al. [3] and Kang et al. [19] to estimate motion between frames using affine transformation and gradient-based optical flow. Hu et al. [14] proposed a patch-based image alignment algorithm.

Recently, deep learning-based approaches have tackled the HDR registration problem. Kalantari et al. [18] exploit optical flow for frame alignment and merge images using CNN. Wu et al. [40] introduce CNN-based encoder-decoder architecture without optical flows. Yan et al. [42] propose a light-weight method that directly gets a global attention map for coarse alignment and increased receptive field with dilated convolution. Yan et al. [43] extract various local feature vectors via three different size of CNN kernel and get a global weight matrix with the non-local module [38]. For high-resolution HDR imaging, Prabhakar et al. [32] generate a coarse guide HDR image in low resolution and reconstruct the final HDR image with bilateral guided upsampling. Niu et al. [28] adopt the GAN framework and introduce a CNN-based multi-scale generator with the residual connection of the reference frame. Liu et al. [25] propose a transformer-based approach that extracts local context features through channel attention and captures global information with multi-head self-attention. However, as shown in the previous studies, the flow-based deghosting algorithm is necessary to reconstruct an HDR image. Also, these multi-shot input images differ from Bayer-patterned or interlaced single-shot HDR input.

Single-shot HDR reconstruction. Nayar et al. [27] propose capturing multiple exposures at a single shot with a spatially-varying image sensor and then interpolating multiple exposures to reconstruct HDR images. Heide et al. [12] propose a computational imaging system that jointly performs several image processing steps: demosaicing, denoising, and deconvolution. For optimization, they introduce several regularization terms for better optimization. Hajisharif et al. [10] predict each pixel value with a local noise-aware polynomial model and adaptive filter kernel. From a single-shot coded-exposure input, Serrano et al. [35] reconstruct HDR images utilizing convolutional sparse coding. Gain-interlaced readout-based methods [4, 1] are also proposed to capture HDR videos. Choi et al. [4] apply sparse

representation for denoising and deinterlacing. Also, they propose temporal denoising by applying a multi-scale homography flow method.

Recently, Akyuz et al. [1] proposed a deep learning-based method for single-shot mosaic HDR imaging. They first restore two raw images with different exposures using a neural network of denoising and demosaicing, and then they reconstruct HDR images analytically in a conventional way. Jiang et al. [17] proposed an end-to-end deep neural network for HDR video reconstruction with triple-exposure quad-Bayer input. Suda et al. [37] also introduced an HDR reconstruction method that infers HDR images from the multi-exposure color filter array. These methods focus on the spatial-resolution degradation problem from the mosaiced input in HDR imaging. In contrast, we jointly solve both demosaicing and deghosting problems from quad-Bayer patterned input with *motion blur*.

HDR hallucination. Many works that hallucinate an HDR image from a single LDR image have been proposed in the recent decade. Banterle et al. [2] restore the original HDR image from the tone-mapped output, so-called inverse tone-mapping. Rempel et al. [33] use a Gaussian filter and edge-stopping function for a clear boundary between dark and bright areas. By means of deep learning, Endo et al. [7] predict bracketed images with multiple exposures by using an auto-encoder, then merge them into the HDR image. Lee et al. [21] exploit the conditional generative adversarial network (GAN) structure to infer the relation between exposure values. Liu et al. [23] approach this task by reversing the LDR image formation pipeline. Santos et al. [34] proposed variable feature masking to avoid artifacts and adopt image inpainting tasks for the saturated region. However, our method differs from HDR image hallucination because we reconstruct HDR images explicitly from the captured multiple-exposure input.

3. Joint Demosaicing and Deghosting

Our method initiates from a single quad-Bayer image, where each color filter covers four pixels with short/mid/mid/long exposure times. Our reconstruction comprises of three main components: (1) pre-processing stage, which involves exposure normalization and subsampling of the input, (2) HDR feature extraction utilized for demosaicing, and (3) HDR feature registration employed for deblurring purposes.

3.1. Preprocessing

Exposure normalization. Our network reconstructs an HDR image I^{HDR} from the quad-Bayer patterned RAW image I^{RAW} with different exposures and colors. Color filters are allocated to follow the traditional R/G/G/B patterns, while the multiple exposures are designed to have 0/2/2/4 stops in the different Bayer patterns.

To weight the middle exposure more, we double the sample number of the middle exposure against lower/higher exposures, analogous to double green samples in the conventional Bayer color patterns. To estimate the scene radiance level at different levels of exposure, we generate a new RAW image $I^{\text{RAW}}_{\text{Norm}}$ normalized by the exposure time of each pixel:

$$I^{\text{RAW}}_{\text{Norm}} = \frac{I^{\text{RAW}}_i}{t_i}, \quad i = 1, 2, 3, \tag{1}$$

where I^{RAW}_i is a set of pixels that has i-th exposure level and t_i is the amount of i-th exposure time. Since the pixel value of our RAW image I^{RAW} is on a linear domain, we are free from the gamma correction. We then concatenate the RAW image I^{RAW} and the normalized RAW image $I^{\text{RAW}}_{\text{Norm}}$ into two-channel tensor X and pass them to the network as input, following Kalantari et al. [18]:

$$I^{\text{HDR}} = f(X; \theta), \tag{2}$$

where $f(\cdot)$ is our network and θ is the network parameter.

Subsampling. We formulate the demosaicing problem as a super-resolution problem from low-resolution input. We first subsample the multi-exposure color filter array image into three Bayer-pattern sub-images of each exposure level in half resolution of the original:

$$X_i = \text{subsample}(X), \quad i = 1, 2, 3, \tag{3}$$

where subsample$(\cdot)$ denotes a subsampling function, which has X with the size of $2 \times H \times W$, and X_i has the size of $2 \times H/2 \times W/2$ when i is 1 and 3. Otherwise, since we have twice the number of pixels for the middle exposure, the second X_i has the size of $4 \times H/2 \times W/2$ by concatenate two $2 \times H/2 \times W/2$-size images.

Moreover, the spatial resolution of each color of X_i at each exposure level i is four-time smaller in total than that of the original pixel resolution of I^{RAW}.

After subsampling, we reshape the subsampled color Bayer patterns as $\hat{X}_i$ before converting them into latent neural features. The height and width dimensions are halved while the channel number is multiplied by four. Refer to the left-top part of Figure 2.

3.2. HDR Feature Extraction

We design a feature extraction model that can demosaic and upscale each exposure level effectively. Since our task is to mitigate the degradation of the spatial resolution in HDR images later, we adopt the latent features of the residual channel-based representation from a widely used super-resolution method [45] that consists of multiple residual channel attention blocks through skip connections. See Figure 2(a). We convert the reordered color patterns of each

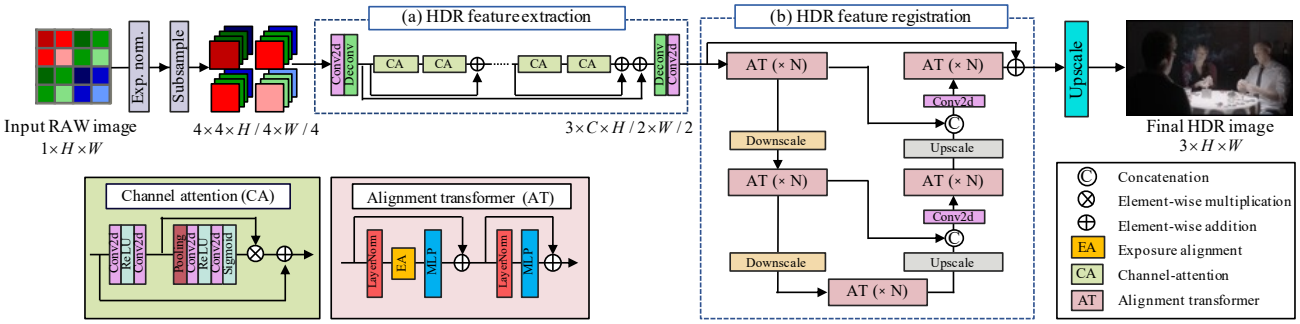


Figure 2. Network architecture. Input is a quad-Bayer RAW image with double mosaic patterns of different exposures and colors. We first subsample the input RAW image to three different exposure layers and then subsample the Bayer patterns of each exposure to each color channel. Then, the HDR feature extraction module (§ 3.2) produces the upscaled and enhanced features from the low-resolution RGB images with three different exposure levels. The HDR feature registration module (§ 3.3) recovers each exposure level feature, aligning pixel-wise displacements of the quad-Bayer patterns and removing motion blur. Lastly, we combine the aligned features into a single feature vector. After upscaling, our network finally produces an HDR image with three color channels in the original RAW resolution.

exposure level into intermediate features as follows:

$$F_i = \mathcal{E}^{\text{extract}}(\hat{X}_i), \quad i = 1, 2, 3, \tag{4}$$

where $\mathcal{E}^{\text{extract}}(\cdot)$ denotes the denoising module and the feature vector F_i has the size of $C \times H/2 \times W/2$, where C is set to 32. Note that we share the weight of the feature extraction module for every exposure level.

The first convolutional layer extracts the RGB channel-wise feature. To design this module, we adopt the channel attention and residual learning approach that consists of the aforementioned channel-attention (CA) block. Channel attention helps the network to represent rich features well, and residual learning enlarges the receptive field of the network. Practically, channel attention provides representations suitable for super-resolution and demosaicing problems [45, 41]. In this module, we regard the connection of two channel-attention blocks in a series as a single group and obtain features by connecting three groups in a series. See Figure 2(a). Since a relatively large number of parameters are required, recent super-resolution methods mainly use the post-upscaling method. To save memory, we also upscale the feature vectors using the deconvolution layer with stride number 2 at the end of the network.

3.3. HDR Feature Registration

We reconstruct a final HDR image with an HDR feature registration module. Since our synthesized image sensor adjusts different exposure levels with exposure time, the amount of motion blur is different for each exposure level. Later, we design a feature alignment module $\mathcal{R}^{\text{regist}}$ by arranging multiple transformer blocks:

$$I^{\text{HDR}} = \mathcal{R}^{\text{regist}}(F_1, F_2, F_3) = f(X), \tag{5}$$

where $\mathcal{R}^{\text{regist}}(\cdot)$ takes three HDR feature vectors from the feature extraction module with each vector size of $C \times H/2 \times W/2$. See Figure 2(b).

The transformer, originating from natural language processing, can solve many vision problems by dividing an input image into a sequence of small patches and exploiting the transformer as an encoder network. It can capture a global attention map with a purpose [38, 15]. It also overcomes other CNN-based methods with large-scale computer vision datasets. Recently, Zamir et al. [44] apply self-attention along feature dimension, and Liang et al. [22] extract key, value, and query vectors from different frames for video restoration. To mitigate the different amounts of motion blur in the quad-Bayer array, we adopt the transformer based on a U-net shape multi-scale architecture [44, 22] that helps extract valuable information and large motions from the different sizes of patches. For each scale, a transformer block restores feature vectors of each exposure level with exposure alignment. See Figure 2(b).

Multi-exposure feature alignment. Our goal is to restore all feature vectors F_1, F_2, F_3 in a single transformer block. Meanwhile, the attention layer in the general transformer receives the key, query, and value vectors as input, calculates the similarity between the key and query vector, and multiplies it by the value vector. We obtain the key and value vectors from the other feature vector and the query vector from itself for exposure alignment. We first define the key K_i, value V_i, and query Q_i of F_i as:

$$K_i = F_i P_i^K, V_i = F_i P_i^V, Q_i = F_i P_i^Q, \quad i = 1, 2, 3, \tag{6}$$

where $P_i^K, P_i^V, P_i^Q \in \mathbb{R}^{C \times D}$ are the projection matrices of F_i, and D is the channel number of the projected features, and C and D are set to be the same, and i indicates the exposure level. The attention layer is formulated as:

$$\text{att}(F_i, F_j) = \text{SoftMax}(Q_i (K_j)^\intercal / \sqrt{D}) V_j, \quad i \neq j, \tag{7}$$

where $\text{SoftMax}(\cdot)$ means the row softmax operation and j indicates the remaining exposure level except i. The multiplication between the transpose of the key and query reflects the similarity between elements in the i-th feature and the j-th feature. While the similarity matrix multiplies by the value vector of j-th feature, the result of this attention

layer becomes a feature alignment of j-th feature to the i-th feature based on the similarity we calculate.

The exposure alignment (EA) shown in Figure 2 is formulated as:

$$\mathrm{EA}(F_i) = \mathrm{concat}(\mathrm{att}(F_i, F_j), \mathrm{att}(F_i, F_k)), i = 1, 2, 3, \tag{8}$$

where $F_{\{i,j,k\}}$ is the input feature vector of multiple exposures, and i, j, k are not the same with each other, i.e., we permute (i, j, k) as $(1, 2, 3)$, $(2, 1, 3)$, and $(3, 1, 2)$. Through the exposure alignment, the structure of the j, k-th features follows the i-th feature, which leads to the restoration of each exposure level. By employing this explicit permutation, each feature vector of the exposure level effectively addresses spatial misalignment and motion blur.

As a result, the whole process of each transformer block can be formulated as:

$$\begin{aligned} F_i &= \mathrm{MLP}(\mathrm{EA}(\mathrm{LN}(F_i))), \quad i = 1, 2, 3, \\ F &= \mathrm{Concat}(F_1, F_2, F_3) + F, \\ F_1, F_2, F_3 &= \mathrm{Split}(\mathrm{MLP}(\mathrm{LN}(F)) + F), \end{aligned} \tag{9}$$

where F represents the concatenation of three features in the channel dimension: F_1, F_2, and F_3. This F is then passed into the network using the residual skip connection mechanism. $\mathrm{LN}(\cdot)$ is layer normalization and $\mathrm{MLP}(\cdot)$ is a multi-layer perceptron. Following the feed-forward layer, the result is once again split into F_1, F_2, and F_3 based on the channel dimension. Here $\mathrm{Split}(\cdot)$ is a feature-channel splitter. However, since the time complexity of the transformer operation is proportional to the square of the number of elements, it is expensive to perform the transformer operation for the entire patch. Therefore, following the existing transformer-based methods [22, 24], we divide the $H \times W$-sized patch into the number of HW/M^2 non-overlapping $M \times M$-sized spatial windows in the EA.

3.4. Training Loss

We compute the HDR loss through the following tone-mapping function $\mathcal{T}(\cdot)$ based on the μ-law [18]:

$$\mathcal{T}(\bar{I}^{\mathrm{HDR}}) = \frac{\log(1 + \mu \bar{I}^{\mathrm{HDR}})}{\log(1 + \mu)}, \tag{10}$$

where $\bar{I}^{\mathrm{HDR}}$ denotes the normalized HDR image in the range of $[0, 1]$, μ denotes parameter deciding the amount of compression, and μ is set to 5,000 in our experiments. Here, our network is optimized by minimizing the $\mathcal{L}_1$ loss function after going through the tone mapping process of both the HDR image result and the ground-truth HDR image:

$$\mathcal{L}_1 = \left\| \mathcal{T}(\hat{I}^{\mathrm{HDR}}) - \mathcal{T}(I^{\mathrm{HDR}}) \right\|_1. \tag{11}$$

4. Experimental Results

4.1. Dataset Generation

HDR video dataset. For creating the training and testing dataset, we use the Stuttgart HDR Video Database [8]. The dataset consists of 22 HDR videos, and each frame is an HDR image. We select 321 HDR images with different scenes and viewpoints for creating our ground-truth HDR image set. For the test dataset, we randomly choose 3 HDR videos (40 HDR images), and we train our network with the remainder. Exploiting this HDR image dataset, we created a set of RAW images by simulating the quad-Bayer sensor.

Noise simulation. To make the synthetic sensor response more realistic, we also simulate a pixel measurement model which converts scene brightness to the sensor value, following Hasinoff et al. [11]:

$$I = \min\{\Phi t/g + I_0 + \eta, I_{\max}\}, \tag{12}$$

where I denotes a 14-bit integer number of the measured pixel value, Φ is the scene brightness, t is the exposure time, g is the sensor gain energy, I_0 is the dark current, η is overall noise, and $I_{\max}$ is the saturation level.

The noise η is an essential part in the real world. Hasinoff et al. treat noise as the zero-mean random variable, which has three different independent factors. For the pixels below the saturation level, the variance σ^2 of η is:

$$\sigma^2 = \Phi t/g^2 + \sigma^2_{\mathrm{read}}/g^2 + \sigma^2_{\mathrm{ADC}}, \tag{13}$$

where σ_{read} denotes the variance of read noise, and σ_{ADC} is the variance of ADC and quantization noise. The first term is a photon noise which is signal and gain dependent. The second term is a gain-dependent readout noise. At high signal levels, noise per pixel is dominated by photon noise, and at the lowest, other sources contribute. Since the pixel measurement in Equation (12) is a linear function, we can consider $\Phi t/g + \eta$ as a real camera sensor value.

Synthetic dataset generation. Each pixel value from the ground-truth HDR image can be regarded as the scene radiance value. As a result, following Hasinoff et al. [11], we can obtain raw Bayer image values with multiple exposure times from a single HDR image by adding consecutive frames. We generate a multi-level exposure image set with three exposure times, with an adjacent exposure time ratio of 4. Note that we first normalize the ground-truth HDR image to [0, 1] and obtain the camera sensor values of the other two exposure levels by summing the following number of frames and clipping to [0, 1] (1, 4, and 16 frames, respectively). See Equation (12) and Figure 3 for an example. This sum operation is actually the same as how the camera does for the different exposure times. Next, we add our simulated camera noise to each pixel value independently,

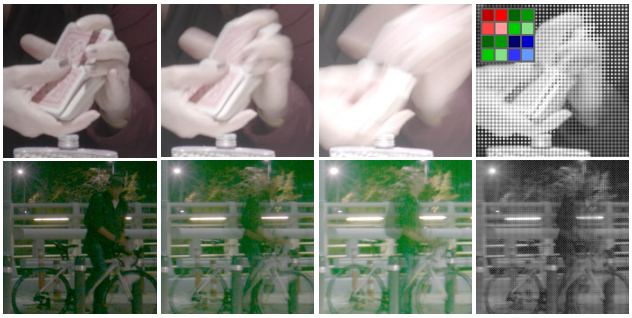


Figure 3. Closeup examples of the input Bayer images and cropped multi-exposure images. The four darkest R/G/G/B pixels in the inset represent a 1-level exposure, and the four brightest ones represent a 3-level exposure. We can observe that the 1-level exposure Bayer (the one with the lowest exposure time) in this dynamic scene produces the least motion blur. All images in this figure are gamma-corrected for visualization. Top row: synthetic dataset [8]. Bottom row: real-world dataset captured by ourselves.

based on an experimental measurement of noise [5]. Lastly, we discretize the normalized pixel value to a 14-bit integer number to simulate a real RAW pixel value. Let I_f^{HDR} as HDR image at frame number f. The i-level exposure image I_i^t at frame t is:

$$I_i^t = \mathrm{pre}\left(\sum_{n=t}^{t+k} I_n^{HDR}\right), \quad i = 1, 2, 3, \qquad (14)$$

where $\mathrm{pre}\,(\cdot)$ denotes the preprocessing steps that include pixel measurement, clipping, and adding simulated noise, and k is $2^{2(i-1)} - 1$ which means two-step intervals. We generate the desired dataset by sampling the pixel values of these four images as a 4×4 Bayer-pattern image (2×2 for color and another 2×2 for exposures). See Figure 3.

Real-world dataset generation. Currently, we do not have a time-varying exposure image sensor on the market yet. However, to validate the performance of our network in the real-world scenario, we create a real-world test dataset combining burst-shot images with different numbers by a DSLR camera (Canon EOS 5 Mark III) to mimic time-varying exposures in an HDR image sensor. Note that there is no ground-truth image for the real-world test dataset. We produce the real-world test dataset in the same way that we create the synthetic one, except that raw images captured by the DSLR burst shot are used instead of continuous HDR video frames. Since the actual camera already contains noise, we do not add simulated noise in the preprocessing step. An example of the real-world dataset is shown in Figure 3.

Implementation details. Our network is implemented with PyTorch [30]. For training, we use the Adam optimizer [20] with $\beta_1 = 0.9, \beta_2 = 0.999$, and $\epsilon = 10^{-8}$. We set the batch size and learning rate as 6 and 1×10^{-4}, respectively. For the training dataset, we randomly crop 256×256 patches. We conduct data augmentation by randomly rotating 90°, 180°, 270° and flipping horizontally. The noise level is also augmented by randomly selecting different standard deviations $\sigma = [0.9\sigma_0, 1.1\sigma_0]$. σ_0 is the standard deviation value obtained by using Equation (13) for the actual camera [5]. Our model takes 0.017s to produce an HDR image of the 256×256 resolution from a RAW image with a single NVIDIA TITAN RTX GPU and Intel CPU i9-12900k. We have trained our model with 100 epochs. It takes about 72 hours on the machine. We set channel number C as 32 and the attention window size of the transformer block as 8×8. For the HDR feature registration module, from the smallest scale to the highest scale, the numbers of transformer blocks are [2,4,2], attention heads are [2,4,8], and channel numbers are [32,64,128]. Also, we use the Pixelshuffle method [36] for downscaling and upscaling.

Table 1. Quantitative comparison of our method with three baseline HDR reconstruction methods. The bold numbers indicate the highest image quality.

Method	HDR-VDP-2 ↑	PSNR ↑	PSNR-μ ↑	SSIM-μ ↑
Suda et al.	64.66	40.00	28.46	0.8133
Yan et al.	73.41	45.50	38.99	0.9548
Liu et al.	72.78	45.53	39.89	0.9604
Ours	**75.57**	**47.89**	**40.10**	**0.9619**

4.2. Baseline Comparison

We evaluate our proposed method compared with other baseline HDR reconstruction methods. Suda et al. [37] use the multi-exposure color filter array, which has similar patterns but different orders of colors and exposures. We use their pre-trained model, which is trained with Kalantari et al.'s HDR dataset [18]. We generate a new input color filter array same as the original quad-Bayer patterns [37] by sampling color pixels differently for a fair comparison. The HDR deghosting task has been continuously researched and has many aspects in common with our problem. As we aim to eliminate the ghosting artifacts caused by combining exposure levels with different motion blur, the HDR deghosting task also targets eliminating ghost artifacts. Moreover, both methods reconstruct HDR images with three different exposure levels. Therefore, we perform a comparison between two state-of-the-art HDR deghosting methods (an attention-based method [42] and a transformer-based method [25]) and our model. For a fair comparison, we retrain their network based on their official source code using the same hyperparameters as our network. We subsample our input into three images according to the exposure levels, put them into their network, and upscale the result to restore the original resolution in the same manner.

Table 1 quantitatively compares the accuracy of the reconstructed HDR images by four methods. 40 HDR images reconstructed by each method are compared with ground-truth HDR images in the test dataset (Figure 5). The accuracy of the results is evaluated with four standard HDR image quality evaluation metrics: HDR-VDP-2 [26], PSNR,

Figure 4. Synthetic dataset comparison. Qualitative comparison of our method with three baseline methods using our synthetic dataset. Two test scenes are presented in this figure. All HDR images in this paper are tone-mapped adaptively by the μ-law for visualization. Our model outperforms baseline methods in terms of color reconstruction and denoising, particularly in the area with strong motion blur.

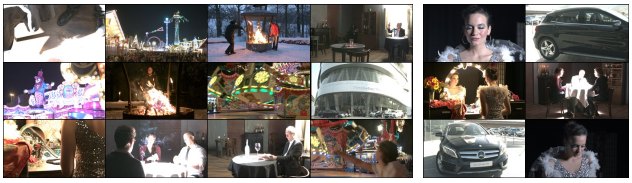


Figure 5. Examples of the training quad-Bayer HDR RAW image dataset (left group) and test dataset (right group), created using [8].

PSNR-μ [18], and SSIM-μ [39] based on the μ-law [18] (μ is set to 5000 in our experiments). In all four metrics, our method outperforms the baseline methods. In particular, our method's performance is significantly better than others in terms of HDR-VDP-2 and PSNR. It validates that our HDR demosaicing method is effective in achieving fine details and robust against motion blur.

Figure 4 compares the reconstruction results of the HDR test dataset qualitatively. Suda et al. [37]'s method takes a multi-exposure filter array as input. They perform HDR reconstruction well in the region without motion. Since the method assumes that multiple exposures have the same integration time, their method does not account for spatially varying motion blur when solving a demosaicing problem. And thus, their method cannot avoid mosaic artifacts due to the motion blur. Yan et al. [42]'s method is less affected by motion blur than the other two methods. However, this method does not recover color well (see the red sleeve in the top row). Also, we can observe that the image structure is destroyed and does not perform denoising well by looking at the white car license plate and the tire wheel in the second scene. Liu et al. [25]'s method shows overall high-quality reconstruction results compared to the previous two methods. However, if we look at the arms of the first scene and the letters of the second scene, we can observe that this method cannot eliminate the motion blur well. In contrast, our model effectively removes artifacts from different motion blur in bright and dark regions within the input image and performs well in demosaicing and debluring.

4.3. Real Camera-based Comparison

Figure 6 compares the reconstruction results of the real-world dataset qualitatively. In the real-world test dataset, Suda et al.'s HDR reconstruction results suffer from severe mosaic artifacts due to the large motion of the objects. Yan et al.'s method still produces HDR images with overall color degradation as a synthetic dataset. Liu et al.'s HDR reconstruction results are vulnerable to the motion of bright objects. Especially when a car's headlight moves, the grid-shaped ghost artifact remains. Our method generates HDR images with consistently high quality compared to other methods.

4.4. Ablation Study

We conduct an ablation study by adding each component module or method to the base module one by one. Table 2 compares the reconstruction performance for each case. The base model means the backbone network that includes only the single-scale HDR feature registration module, which performs self-attention (SA) instead of exposure alignment (EA) in the transformer block. Refer to Figure 2(b). For self-attention, to obtain a direct dependency

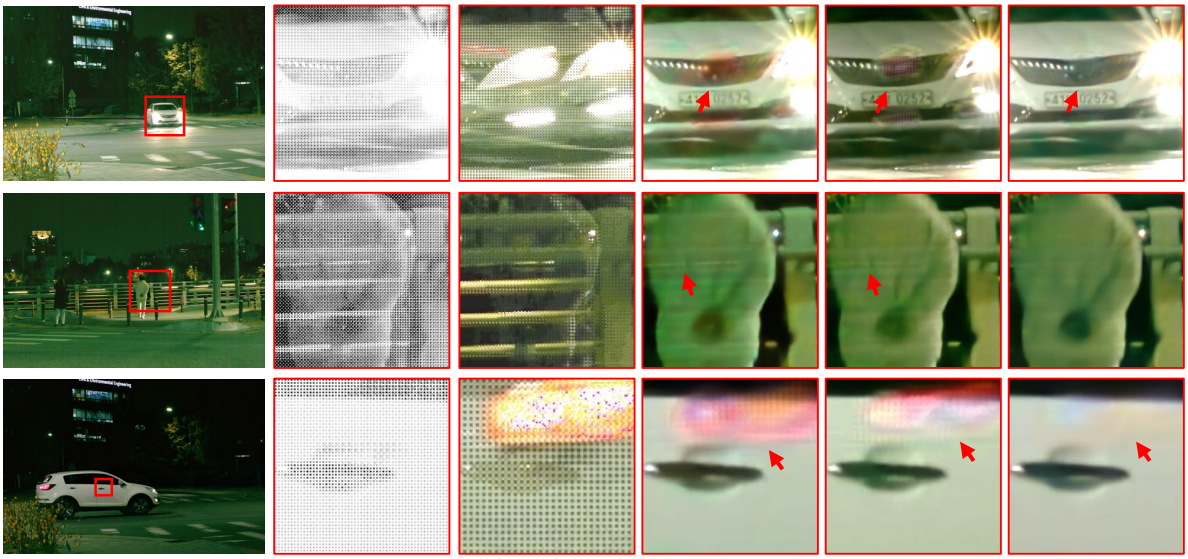


Figure 6. Real-world dataset comparison. We qualitatively compare our method with three baseline methods on our real-world dataset. Even in these extreme environments, our approach successfully minimizes motion blur artifacts.

Table 2. Ablation study of our model. We begin with the base model by gradually adding each component module (SA: self-attention, EA: exposure alignment, FE: feature extraction, MS: multi-scale). The overall model provides the best results in all metrics. The best results are **highlighted as bold**.

Method	HDR-VDP-2 ↑	PSNR ↑	PSNR-μ ↑	SSIM-μ ↑
Baseline (SA)	70.89	43.98	36.99	0.9446
SA + MS	71.31	44.26	37.61	0.9484
SA + FE	72.73	44.91	38.29	0.9529
SA + MS + FE	74.28	46.54	38.62	0.9581
EA	74.47	46.81	39.08	0.9553
EA + MS	75.53	47.57	39.85	0.9604
EA + FE	74.87	46.74	39.93	0.9604
Overall	**75.57**	**47.89**	**40.10**	**0.9619**

between exposures, we concatenate three feature vectors into a single one after the feature extraction module. In addition, in the transformer block, we divide the patch into $3\times M\times M$-sized windows as input to the transformer. First, the overall performance increases when we change attention operation SA to EA of the transformer block. This shows the highest metric increment among three options: EA, MS, and FE. When we adopt the transformer block's multi-scale (MS) arrangement, HDR-VDP-2 and PSNR highly increase. On the other hand, when we add the feature extraction (FE) module, PSNR-μ and SSIM-μ mainly increase. The full model that includes all the tested modules shows the best performance in this study.

5. Discussion and Conclusion

We have presented a learning-based demosaicing method of a time-varying exposures array for single-shot HDR imaging. Our method demosaics quad-Bayer RAW images, removing spatially-varying motion blur. It yields high-quality HDR images with high resolution. Our method leverages the transformer model with changing source of the query, key, and value vectors of the transformer block. Our ablation study evaluates the impact of each module on performance. Our complete model presents the best performance with all the designed components. Lastly, the results validate that the complete model outperforms the baseline HDR reconstruction methods. We anticipate that the proposed method can reconstruct consistent single-shot HDR images for the dynamic scene without compromising the dynamic range when enabling single-shot HDR imaging with the quad-Bayer sensor architecture.

Both the synthetic and real-world datasets we generate have limitations compared to the data from the actual exposure-time varying image sensor. In the case of scenes with motion blur, we obtain a long exposure time pixel value by summing each burst-shot frame with a short exposure time, but the fast motion often appears to be cut off (Figure 4).

We use the integrated burst shots as input to create synchronously triggered multiple exposures on the time-varying image sensor. However, each burst-shot image has a too-short exposure time, having dominant dark current noise. Also, each shot has read noise. This read noise level increases when we create each level of multi-exposure proportionally to the number of input burst shots.

Acknowledgements

Min H. Kim acknowledges the main support of Samsung Electronics, in addition to the additional support of the MSIT/IITP of Korea (RS-2022-00155620, 2022-0-00058, and 2017-0-00072), the Samsung Research Funding Center (SRFC-IT2001-04), and the NIRCH of Korea (2021A02P02-001).

References

[1] Ahmet Oğuz Akyüz et al. Deep joint deinterlacing and denoising for single shot dual-iso hdr reconstruction. *IEEE Transactions on Image Processing*, 29:7511–7524, 2020. 1, 2, 3

[2] Francesco Banterle, Patrick Ledda, Kurt Debattista, and Alan Chalmers. Inverse tone mapping. In *Proceedings of the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia*, pages 349–356, 2006. 3

[3] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proceedings 15th International Conference on Pattern Recognition. ICPR-2000*, volume 3, pages 7–12. IEEE, 2000. 2

[4] Inchang Choi, Seung-Hwan Baek, and Min H Kim. Reconstructing interlaced high-dynamic-range video using joint learning. *IEEE Transactions on Image Processing*, 26(11):5353–5366, 2017. 1, 2

[5] Roger N. Clark. Clarkvision.com. 6

[6] Paul E Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *ACM SIGGRAPH 2008 classes*, pages 1–10. 2008. 2

[7] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Trans. Graph*, pages 177:1–177:10, 2017. 3

[8] Jan Froehlich, Stefan Grandinetti, Bernd Eberhardt, Simon Walter, Andreas Schilling, and Harald Brendel. Creating cinematic wide gamut hdr-video for the evaluation of tone mapping operators and hdr-displays. In *Digital photography X*, volume 9023, page 90230X. International Society for Optics and Photonics, 2014. 5, 6, 7

[9] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen*, pages 277–284, 2006. 2

[10] Saghi Hajsharif, Joel Kronander, and Jonas Unger. Hdr reconstruction for alternating gain (iso) sensor readout. In *Eurographics, Strasbourg, France, April 7-11, 2014*, 2014. 1, 2

[11] 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. 5

[12] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid Pajkak, Dikpal Reddy, Orazio Gallo, Jing Liu, Wolfgang Heidrich, Karen Egiazarian, et al. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics (ToG)*, 33(6):1–13, 2014. 1, 2

[13] Yong Seok Heo, Kyoung Mu Lee, Sang Uk Lee, Youngsu Moon, and Joonhyuk Cha. Ghost-free high dynamic range imaging. In *Asian Conference on Computer Vision*, pages 486–500. Springer, 2010. 2

[14] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. Hdr deghosting: How to deal with saturation? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1163–1170, 2013. 2

[15] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 603–612, 2019. 4

[16] Katrien Jacobs, Celine Loscos, and Greg Ward. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics and Applications*, 28(2):84–93, 2008. 2

[17] Yitong Jiang, Inchang Choi, Jun Jiang, and Jinwei Gu. Hdr video reconstruction with tri-exposure quad-bayer sensors. *arXiv preprint arXiv:2103.10982*, 2021. 1, 3

[18] Nima Khademi Kalantari, Ravi Ramamoorthi, et al. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144–1, 2017. 2, 3, 5, 6, 7

[19] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics (TOG)*, 22(3):319–325, 2003. 2

[20] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[21] Siyeong Lee, Gwon Hwan An, and Suk-Ju Kang. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 596–611, 2018. 3

[22] Jingyun Liang, Jiezhang Cao, Yuchen Fan, Kai Zhang, Rakesh Ranjan, Yawei Li, Radu Timofte, and Luc Van Gool. Vrt: A video restoration transformer. *arXiv preprint arXiv:2201.12288*, 2022. 4, 5

[23] 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. 3

[24] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[25] Zhen Liu, Yinglong Wang, Bing Zeng, and Shuaicheng Liu. Ghost-free high dynamic range imaging with context-aware transformer. *arXiv preprint arXiv:2208.05114*, 2022. 2, 6, 7

[26] 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. 6

[27] 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. 2

[28] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau. Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions on Image Processing*, 30:3885–3896, 2021. 2

[29] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE transactions on pattern analysis and machine intelligence*, 37(6):1219–1232, 2014. 2

[30] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[31] Fabrizio Pece and Jan Kautz. Bitmap movement detection: Hdr for dynamic scenes. In *2010 Conference on Visual Media Production*, pages 1–8. IEEE, 2010. 2

[32] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution hdr deghosting with cnn. In *European Conference on Computer Vision*, pages 497–513. Springer, 2020. 2

[33] Allan G Rempel, Matthew Trentacoste, Helge Seetzen, H David Young, Wolfgang Heidrich, Lorne Whitehead, and Greg Ward. Ldr2hdr: on-the-fly reverse tone mapping of legacy video and photographs. *ACM transactions on graphics (TOG)*, 26(3):39–es, 2007. 3

[34] Marcel Santana Santos, Tsang Ing Ren, and Nima Khademi Kalantari. Single image hdr reconstruction using a cnn with masked features and perceptual loss. *ACM Transactions on Graphics*, 39(4), Aug 2020. 3

[35] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. Convolutional sparse coding for high dynamic range imaging. In *Computer Graphics Forum*, volume 35, pages 153–163. Wiley Online Library, 2016. 1, 2

[36] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1874–1883, 2016. 6

[37] Takeru Suda, Masayuki Tanaka, Yusuke Monno, and Masatoshi Okutomi. Deep snapshot hdr imaging using multi-exposure color filter array. In *Proceedings of the Asian Conference on Computer Vision*, 2020. 1, 3, 6, 7

[38] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 2, 4

[39] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[40] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 117–132, 2018. 2

[41] Wenzhu Xing and Karen Egiazarian. End-to-end learning for joint image demosaicing, denoising and super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3507–3516, 2021. 4

[42] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1751–1760, 2019. 2, 6, 7

[43] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang. Deep hdr imaging via a non-local network. *IEEE Transactions on Image Processing*, 29:4308–4322, 2020. 2

[44] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5728–5739, 2022. 4

[45] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 286–301, 2018. 3, 4