HDR Video Reconstruction with Tri-Exposure Quad-Bayer Sensors

Yitong Jiang Inchang Choi Jun Jiang Jinwei Gu
SenseBrain
{jiangyitong,inchangchoi,jiangjun,gujinwei}@sensebrain.ai

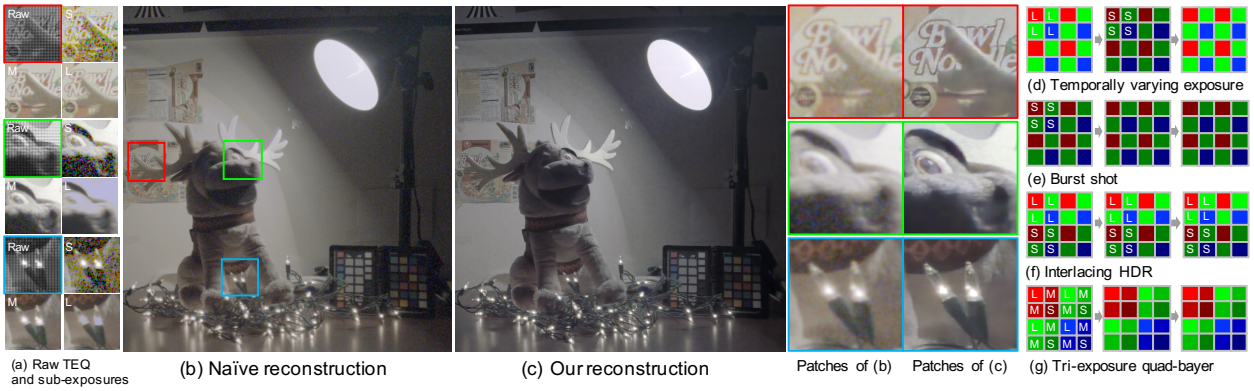


Figure 1: Our HDR reconstruction of a challenging scene is shown in (c). The real tri-exposure quad-bayer (TEQ) input of the scene, described in (g), and its sub-exposure images are shown in (a)[1]. Saturation, noise, spatial artifact, and motion blur in the sub-exposure images are attributed to the quality degradation of an engineered-interpolation-based naive reconstruction in (b). From (d) to (g), four exposure strategies for the HDR video are described.

Abstract

We propose a novel high dynamic range (HDR) video reconstruction method with new tri-exposure quad-bayer sensors. Thanks to the larger number of exposure sets and their spatially uniform deployment over a frame, they are more robust to noise and spatial artifacts than previous spatially varying exposure (SVE) HDR video methods. Nonetheless, the motion blur from longer exposures, the noise from short exposures, and inherent spatial artifacts of the SVE methods remain huge obstacles. Additionally, temporal coherence must be taken into account for the stability of video reconstruction. To tackle these challenges, we introduce a novel network architecture that divides-and-conquers these problems. In order to better adapt the network to the large dynamic range, we also propose LDR-reconstruction loss that takes equal contributions from both the highlighted and the shaded pixels of HDR frames. Through a series of comparisons and ablation studies, we show that the tri-exposure quad-bayer with our solution is more optimal to capture than previous reconstruction methods, particularly for the scenes with larger dynamic range and objects with motion.

[1]L, M, and S on the color filter stand for long, middle, and short exposures, respectively.

1. Introduction

Digital image sensors have a limited dynamic range (e.g., 60 dB for mobile phone cameras), which is determined by the full-well capacity, dark current, and read noise. The constrained dynamic range often gives us unsatisfying portrait photographs with either excessively dark and noisy faces or completely saturated backgrounds, and it is one of the major issues that make photographing less enjoyable. To mitigate the limit, *high dynamic range* (HDR) imaging was introduced [6, 14] and has been under significant attention for a couple of decades. In addition, *HDR video* [24, 22, 5] is getting closer to our real life. It is not difficult to find TVs that support HDR video, and video sharing platforms also have started to stream HDR contents.

For the acquisition of HDR video, there have been three major approaches. *Temporally varying exposure* (TVE), also known as *exposure bracketing*, takes multiple shots of a scene with different exposure settings as shown in Figure 1(d), and fuses one HDR frame by weighted-averaging well-exposed pixels from adjacent frames. Unfortunately, it suffers from *ghosting artifacts* [30, 11] for dynamic objects, and shows an intrinsic trade-off between motion and dynamic range. This could be alleviated by a *burst shot*

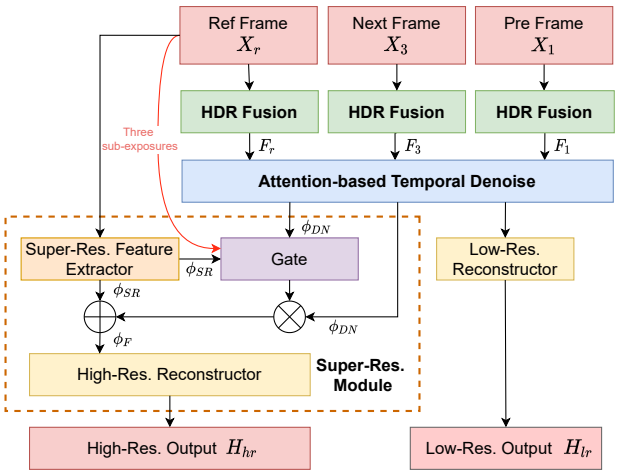


Figure 2: The overview of our HDR reconstruction network architecture.

(BS) [16, 26, 27] approach, described in Figure 1(e), that takes a considerably short exposure to minimize the motion between the frames. However, it undergoes severe quantization and noise in the dark region of scenes and requires expensive denoising algorithms [39, 19]. Finally, there are *spatially varying exposure* (SVE) methods that are the least hindered by the ghosting artifacts and the quantization problem. In modern image sensors that support SVE, a number of the sets of pixels within one frame can take different exposures. But their HDR reconstructions have problems of reduced resolution and interlacing artifacts [34, 14, 4, 44].

We propose a novel HDR video reconstruction method for new SVE sensors: *tri-exposure quad-bayer sensors* [36, 12, 7]. As shown in Figure 1(g), the quad-bayer sensors spatially extend each color filter of the bayer to four neighboring pixels. Their tri-exposure mode enables to set three different exposure settings within each color. Comparing to the interlacing HDR methods [44, 5] shown in (f), which has represented the SVEs, it samples more exposure sets uniformly over the image. Therefore it is more robust to the noise and the spatial artifacts. Nonetheless, a naive HDR video reconstruction[2] method (b) would suffer from three problems described in (a): the motion blur from longer exposures, the noise from short exposures, and inherent spatial artifact and resolution degradation. In addition, the temporal coherence must be taken into account for the better stability of the video.

To tackle the problem, we introduce novel network architecture, shown in Figure 2, that divides-and-conquers three problems. It is modularized by the *HDR feature fusion* module that performs HDR fusion in the feature space to address the motion blur, the *attention-based temporal denoising* module that performs the multi-frame noise reduction and maintains the temporal coherence, and the *super-resolution* module that alleviates the remaining spatial artifact and resolution problems as shown in Figure 1(c). To better adapt the network to the large dynamic range, we also propose *LDR-reconstruction loss* that takes equal contributions from highlighted and shaded pixels in output HDR frames. We provide a thorough comparison with other HDR video methods, including TVEs, BSs, and SVEs, and ablation studies. They show that the tri-exposure quad-bayer with our proposed solution is more optimal to capture HDR video than the previous HDR reconstruction methods, particularly for the scenes with larger dynamic range and dynamic objects. All the code and the data will be available upon publication.

[2] A traditional reconstruction method using engineered interpolations.

2. Related Work

HDR using Temporally Varying Exposure An exposure strategy of alternating different exposures called *exposure bracketing* was introduced by [6, 14] to capture HDR images. Kang et al. [24] extended it to video by utilizing optical flow to align neighborhood frames to a reference frame. In the succeeding research, more robust motion estimation methods relying on block-based motion [30, 31] and patch-based synthesis [23] were proposed. Subsequently, Gryaditskaya et al. [11] performed motion-aware exposure bracketing by considering the perceptual importance of motion and dynamic range, and Kalantari et al. [22] presented a CNN-based HDR video reconstruction. Nonetheless, none of these methods were completely free from the ghosting artifact caused by fast large motion and the extensive loss of rigidity. In contrast, our HDR video reconstruction is more robust to motion for the inherent robustness of the SVEs and our attention-based temporal denoising that does not rely on explicit motion estimation.

HDR from Burst Shots With recent emerging interests in low-light imaging [3, 41, 29], the burst-shot-based HDR image algorithms have obtained great attention [27, 16, 26]. It takes multiple shots of images with the same short exposure. The fixed exposure makes the motion estimation more robust since the intensity level and the noise level do not change over the burst sequence. Hasinoff et al. [16] proposed an efficient system for the burst-shot-based HDR, and Liba et al. [26] improved it by introducing a superior exposure scheduler and a new merging algorithm. Extending the burst imaging algorithm to video is straightforward, but it suffers severe noise and quantization problems for the dark region in the video. Therefore, strong and computationally demanding video denoising algorithms [28, 10, 19, 38, 39] must be accompanied as post-processing. On the other hand, the HDR video from our tri-exposure quad-bayer alleviates the noise and the quantization by taking short, middle, and long exposures together in each frame while maintaining the robustness to the ghosting artifact.

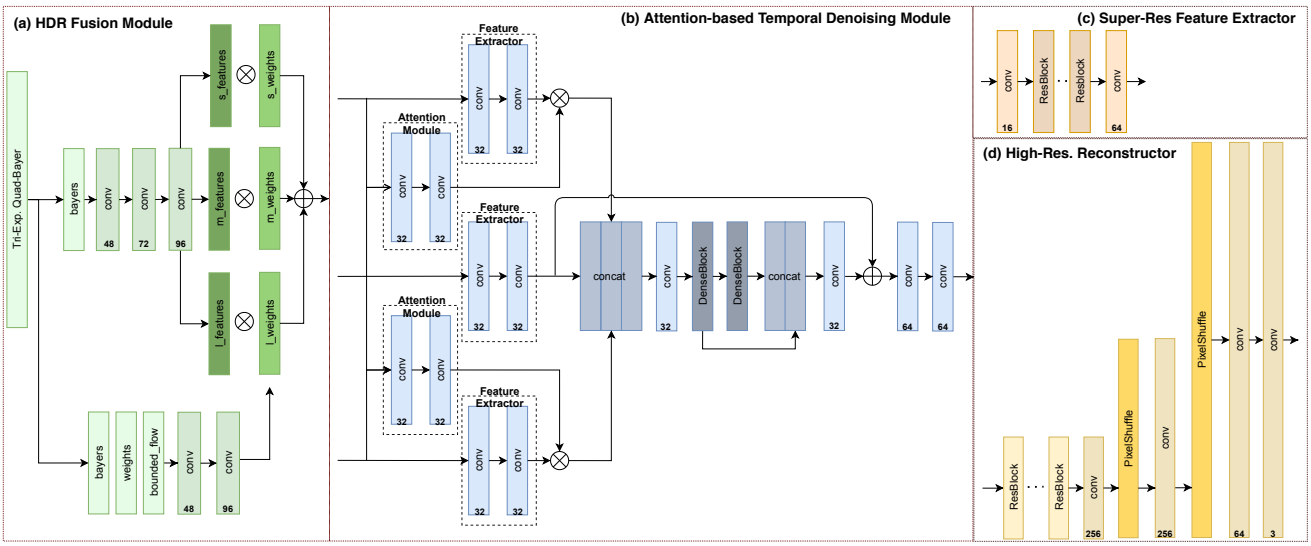


Figure 3: The modules in our reconstruction network are described here. The HDR fusion module is shown in (a). The attention-based temporal denoising module is depicted in (b). The super-resolution feature extractor and the high-resolution reconstructor are described in (c) and (d), respectively.

HDR using Spatially Varying Exposure The other HDR image/video acquisition is to utilize the spatially varying exposure that takes multiple different exposures within a single frame [34, 13]. Although this approach is less suffered from the ghosting artifacts, it suffers from spatial artifacts on under-/over-exposed pixels and high contrast regions. Heide et al. [17] and Cho et al. [4] addressed it by global optimization with image priors, and an adaptive filter was used in [15]. Serrano et al. [37] proposed to use learned image priors using a convolutional sparse coding model. Choi et al. [5] and çoğalan and Akyüz [44] introduced an HDR video reconstruction algorithm for interlacing sensors with two exposure sets using joint sparse coding and deep learning, respectively. But they were vulnerable to artifacts and resolution degeneration in the vertical direction. Conversely, three exposure sets of the tri-exposure quad-bayer, which are uniformly distributed over the image, and our novel reconstruction network produce better HDR video with less noise and spatial artifacts.

Learning-based HDR Reconstruction Recently, a series of works that utilizes deep convolutional neural networks to reconstruct HDR images from the temporally varying exposure has been presented. Their focus was mainly on solving the ghosting artifacts. Kalantari et al. [21] proposed to use learned optical flow for the alignment, and it was extended to video in [22]. Larger network architecture and a spatial attention module mitigated the motion problem by Wu et al. [40] and Yan et al. [42]. For the spatially varying exposure, An et al. [1] and Coğalan and Akyüz [44] proposed a CNN-based reconstruction method for an interlacing sensor. All of these methods apply a simple global tone mapping with a fixed parameter [21, 22, 40, 42] on output and label HDR images to make the network adapt to a large dynamic range. This not optimal for preserving the local contrast and removing the noise in the dark, since relatively dark pixels cannot contribute enough in the computation of the loss. In contrast, our LDR reconstruction loss that uses simulated LDR images in the loss function leads to superior HDR video reconstruction.

3. Our Method

As shown in Figure 2, our HDR reconstruction network takes three consecutive raw frames of tri-exposure quad-bayers $\{X_1, X_2, X_3\}$ as inputs and generates a clean HDR frame $\widetilde{H}$. To clarify, we denote X_2 as X_r to indicate the reference frame and keep the same convention for other variables as well. The network consists of three modules. First, the *HDR fusion* module produces HDR features by fusing the sub-exposures[3] in each tri-exposure quad-bayer X_i. Second, the *attention-based temporal denoising* module takes the HDR features of the three input frames to remove the noise in the feature space. Note that the outputs of the HDR fusion module and the denoising module have the half resolution of the raw inputs. Finally, the *super-resolution* module removes the spatial artifacts and recovers the original resolution. This module utilizes the high-resolution feature extracted from the raw input X_r. It is merged with the denoised low-resolution by a learned gate operation [43], and the high-resolution reconstructor generates a final HDR frame. The following subsections describe the details of each module.

[3]Three sub-sampled images with the different exposures. They are one-fourth of the input raw in size since the sub-sampling is with respect to the exposure and the color.

3.1. HDR Fusion Module

This module performs the HDR fusion of three different exposures in a tri-exposure quad-bayer in the feature space. As shown in Figure 3(a), one branch of the module produces features from three exposures: f_S, f_M, and f_L. In the other branch, the weights, w_S, w_M, and w_L are estimated. This estimation includes the conventional trapezoidal intensity weight map [6], which is widely used in the HDR image reconstruction, and the bounded flow map [26] as inputs. Having the bounded flow helps the module learn to reject the motion blur from the long exposures. The HDR feature F is the weighted sum of each exposure feature.

$$F = \sum_i f_i \circ w_i \tag{1}$$

where $\circ$ denotes the Hadamard product, and i is a variable for indicating one of three exposures $\{S, M, L\}$. Here, we produce three HDR features $\{F_1, F_r, F_3\}$ for $\{X_1, X_r, X_3\}$ using a shared HDR fusion module.

3.2. Attention-based Temporal Denoising Module

Our temporal denoising utilizes attention modules [42, 39] to address the misalignment between the frames instead of explicitly estimating the motion. The attention modules extract useful features from different frames to refine the reference frame. As shown in Figure 3(b), we compute the attention A_j on F_j with respect to the reference HDR feature F_r as follows:

$$A_j = a(F_j, F_r), \tag{2}$$

where $a(\cdot)$ is the attention module that consists of two convolutional layers and $j \in \{1, 3\}$. Then the attention A_j is used to attend the refined non-reference HDR feature $\bar{F}_j$ that is produced from F_j by the feature extractor.

$$Z_j = A_j \circ \bar{F}_j \tag{3}$$

Note that the attention modules share the weight, and so do the feature extractor. Then, we stack the attended feature Z_1 and Z_3 with the refined reference HDR feature $\bar{F}_r$.

$$Z_t = concat(Z_1, \bar{F}_r, Z_3) \tag{4}$$

Z_t is passed through a convolutional layer followed by dilated residual dense blocks [42]. After one more convolutional layer, it goes through a skip connection that adds $\bar{F}_r$ and two more convolutional layers to make the denoised HDR feature ϕ_{DN}.

As shown in Figure 2, the temporal denoising module engages with the low-resolution reconstructor. The low-resolution reconstructor consists of three convolutional layers that reconstruct a low-resolution HDR frame with the quarter size of the raw input. As described in Section 3.4, we compute a loss on this low-resolution reconstruction to make the denoising module focus on learning its own task.

3.3. Super-Resolution Module

The super-resolution module consists of the super-resolution feature extractor and the high-resolution reconstructor. In the super-resolution feature extractor, shown in Figure 2(c), we use eight ResBlocks to extract the super-resolution feature ϕ_{SR} from an input raw quad-bayer. This is intended to fully utilize spatial information. The gate module performs a feature fusion to form an input to the high-resolution reconstructor, which is as shown in Figure 3(d):

$$\phi_F = G_{gate}\left(\phi_{SR}, \phi_{DN}, X_r\right) \circ \phi_{DN} + \phi_{SR}. \tag{5}$$

ϕ_F is fed into the high-resolution reconstructor. First, it passes through eight ResBlocks followed by a convolutional layer. Then, two pixel-shuffle-layers are deployed to upsample the features. Each pixel-shuffle is followed by a convolutional layer. Finally, the last convolutional layer reconstructs a high-resolution HDR image with RGB colors.

3.4. Training Loss

In the previous learning-based HDR image/video reconstructions [22, 40, 42], it has been prevalent to perform global tone mapping before computing the loss function. Because of the wide intensity range of the HDR images, the loss for the dark pixels has less effect on the total loss value, and it was not able to reconstruct the dark part of the images without boosting the dark by the tone mapping. The following global tone mapping function, called μ-law, was used in the previous methods:

$$T(H) = \frac{\log\left(1 + \mu H\right)}{\log\left(1 + \mu\right)}, \tag{6}$$

where H is a reconstructed HDR image, and μ is a parameter adjusting the amount of the dynamic range compression.

However, for the scenes with an extensively large dynamic range, no matter how large μ[4] we set, the contribution of the dark pixels in the total loss is not amplified significantly enough. This will cause artifacts in the dark region. To address this, we propose a novel *LDR-reconstruction loss*. For the LDR-reconstruction loss, we first simulate the LDR images from the reconstructed HDR image:

$$I_i^{LDR}(H) = \left(H \cdot t_i \cdot g_i\right)^{\frac{1}{2.2}}, \tag{7}$$

where t_i and g_i is the exposure time and the gain[5], and i is a variable for indicating one of three exposures $\{S, M, L\}$ in the tri-exposure quad-bayer. Then, the LDR-reconstruction loss $\mathcal{L}^{LDR}(\widetilde{H}, H)$ is defined as:

$$\sum_i \mathcal{L}(\,\omega \circ I_i^{LDR}(\widetilde{H}),\ \omega \circ I_i^{LDR}(H)\,), \tag{8}$$

[4]The fixed value ($\mu = 5000$) has been widely adopted.

[5]t_i and g_i are known from the moment of the image capture

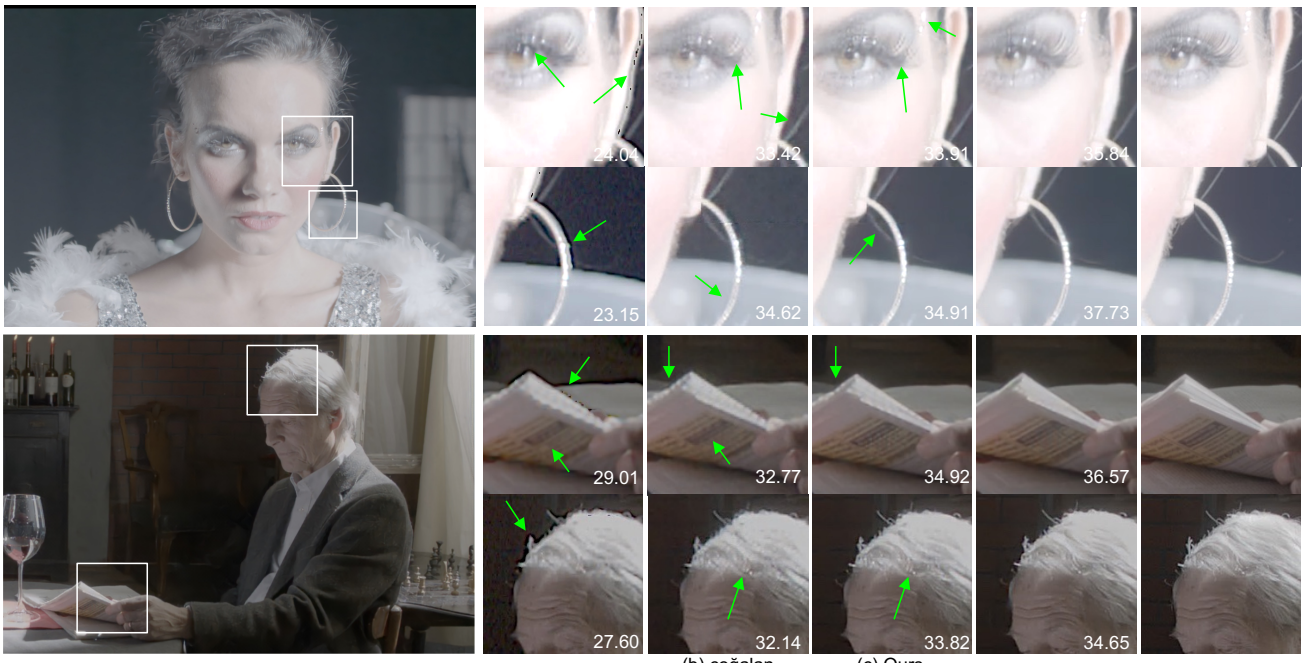


Figure 4: We compare the proposed method to state-of-the-art interlacing HDR video methods. The methods of Choi et al. [5] and çoğalan and Akyüz [44] are shown in (a) and (b). Our proposed model trained on interlaced inputs in (c) better removes interlace artifacts. Our results from the tri-exposure quad-bayer inputs are demonstrated in (d), and (e) is the ground-truth. The PSNR values are located in each patch.

where $\widetilde{H}$ is the ground-truth HDR image, and ω is a mask indicating well-exposed pixels in the simulated LDR image. $\mathcal{L}$ could be any loss function that measures the difference of images, but we use the ℓ^1 loss and the perceptual loss [20].

Our network produces a high-resolution HDR image H_{hr} as well as a low-resolution HDR image H_{lr} as mentioned in Section 3.2. By computing the loss on the low-resolution HDR image, we enforce the temporal denoising module to focus on denoising tasks, and we achieve better HDR reconstruction as shown in Section 5.2. Our final loss function is formulated as follows:

$$\mathcal{L}^{LDR}(\widetilde{H}_{hr}, H_{hr}) + \alpha\mathcal{L}^{LDR}(\widetilde{H}_{lr}, H_{lr}), \qquad (9)$$

where $\widetilde{H}_{hr}$ and $\widetilde{H}_{lr}$ are the corresponding ground-truth images, and α is set to 0.6.

3.5. HDR Dataset and Raw Simulation

We simulated tri-exposure quad-bayers from 28 HDR footages from three public datasets: Froehlich et al. [8], Kronander et al. [25], and Azimi et al. [2]. From the selected HDR footages, we generate three LDR images with short, middle, and long exposure time. The ratio between adjacent exposures is set to 4. Then, we merge the three LDR images to be a tri-exposure quad-bayer image.

We added Gaussian noise to the LDR images. The standard deviation was randomly selected from between 4×10^{-3} to 1.6×10^{-2} for the short exposure time LDR image. The standard deviations for the middle and the long exposure are computed by multiplying the exposure ratio on it. In addition to the noise, we also simulated motion blur in the LDR images using Super Slomo [18].

3.6. Implementation Details

Our model is implemented in PyTorch [35]. For training, we used Adam optimizer and set the batch size and learning rate as 12 and 1×10^{-4}, respectively. We break down the training images into the overlapping patches of 256×256 with a stride of 120 pixels. The network was initialized by Xavier initialization [9]. Our method takes 0.008 seconds to generate an HDR image from a 1920×1080 quad-bayer raw with NVIDIA 2070 Super GPU.

4. Results

4.1. Comparison to Interlacing HDR

We conducted a comparison between our reconstruction with the TEQ and the state-of-the-art interlacing HDR video reconstruction method. In this comparison, we included a model of our proposed architecture trained with simulated interlacing bayers to verify the effectiveness of it. Figure 4 shows the results of the comparison. The joint sparse coding method of Choi et al. [5] in (a) suffers from noise and spatial interlacing artifacts on the over-exposed and under-

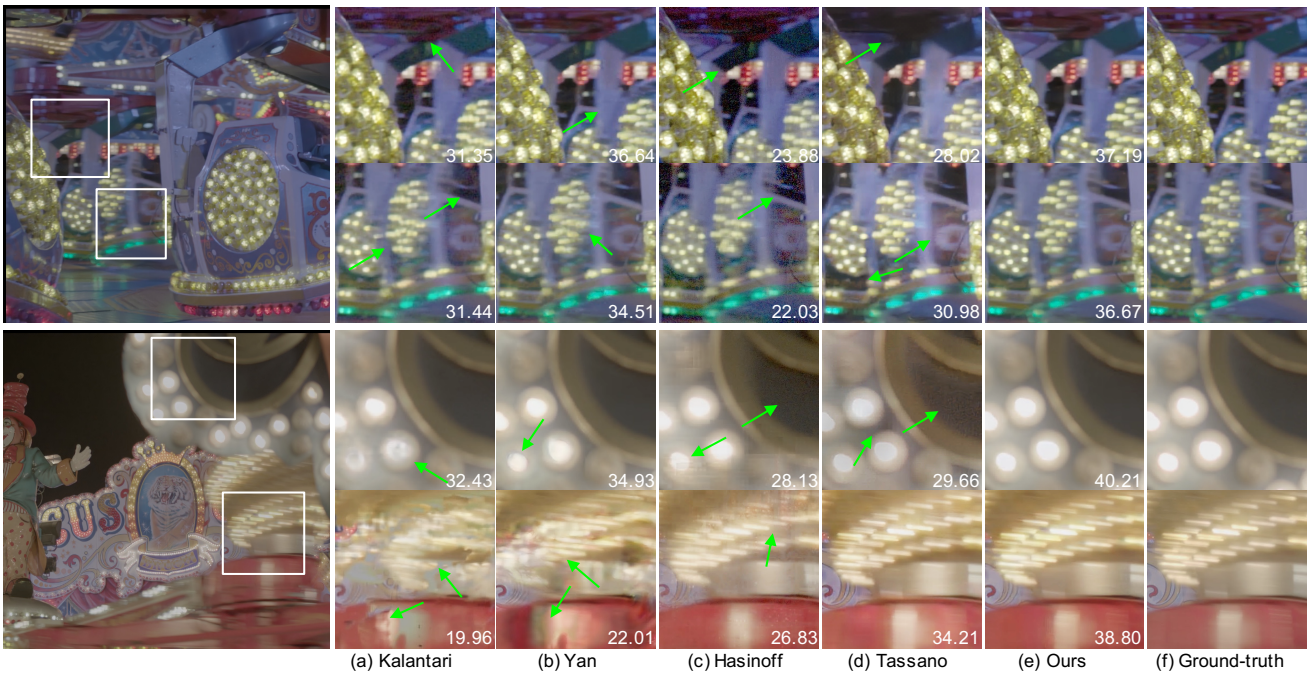


Figure 5: We compare the proposed method to state-of-the-art HDR video methods. The temporally varying exposure methods of Kalantari et al. [22] and Yan et al. [42] are shown in (a) and (b). The burst shot methods of Hasinoff et al. [16] and Tassano et al. [39] are shown in (c) and (d). Our results are demonstrated in (e), and (f) is the ground-truth. The PSNR values are located in each patch.

	PSNR-μ	PuPSNR	HDR-VDP	HDR-VQM
Choi	28.32	36.16	40.21	0.7703
çoğalan and Akyüz	34.71	40.91	44.69	0.8802
Ours (interlace)	35.90	41.56	45.02	0.9194
Ours (TEQ)	36.06	41.64	46.16	0.9275

Table 1: This table shows the quantitative evaluation of state-of-the-art HDR video methods with interlacing inputs.

exposed region. The CNN-based method of çoğalan and Akyüz [44] in (b) better removes the noise, but it cannot recover the detail of textures. In contrast, our model trained with interlacing inputs in (c) shows better reconstruction than the previous two methods. However, our reconstruction using TEQ in (d) is more close to the ground-truth in (e). This is attributed to the uniform deployment of the exposure samples and the extra medium exposure samples in TEQ as shown in Figure 1.

We conducted the qualitative evaluations on twelve test scenes. The averaged quality evaluations are shown in Table 1. Our method produces the best scores of PSNR-μ, PuPSNR, HDR-VDP [32], and HDR-VQM [33]. Note that HDR-VQM is a metric to assess the HDR video quality including the notion of temporal coherence.

4.2. Comparison to TVEs and BS

We compared our TEQ HDR video reconstruction to previous state-of-the-art methods on challenging HDR video scenes. Specifically, the comparison includes two temporally varying exposure (TVE) HDR video methods [22, 42] and two burst shot HDR video methods [16, 39]. The same dataset and exposure settings are used for simulating the inputs with the different exposure strategies.

	PSNR-μ	PuPSNR	HDR-VDP	HDR-VQM
Kalantari	33.27	38.48	44.79	0.8632
Yan	35.67	40.75	45.01	0.9226
Hasinoff	23.73	31.62	39.42	0.7645
Tassano	26.15	34.10	41.45	0.8096
Ours	36.06	41.64	46.16	0.9275

Table 2: This table shows the quantitative evaluation of state-of-the-art HDR video methods.

Figure 5 shows the comparison on challenging scenes with significant darkness/saturation and complex motions. The TVE methods of Kalantari et al. [22] in (a) and that of Yan et al. [42] in (b) suffer from ghosting artifacts for fast-moving objects. The ghosting is more severe when the reference frame is the long exposure frame among the alternating exposures, since it contains more motion blur. Also, Kalantari's method shows noise shown in the second row of (a), because the model does not have any denoising module. The burst shot methods of Hasinoff et al. [16] in (c) and Tassano et al. [39] in (d) are not able to recover the details in the dark area, since the burst shot methods focus on the motion with the expense of the severe quantization in the dark part. In contrast, our reconstruction of the tri-exposure

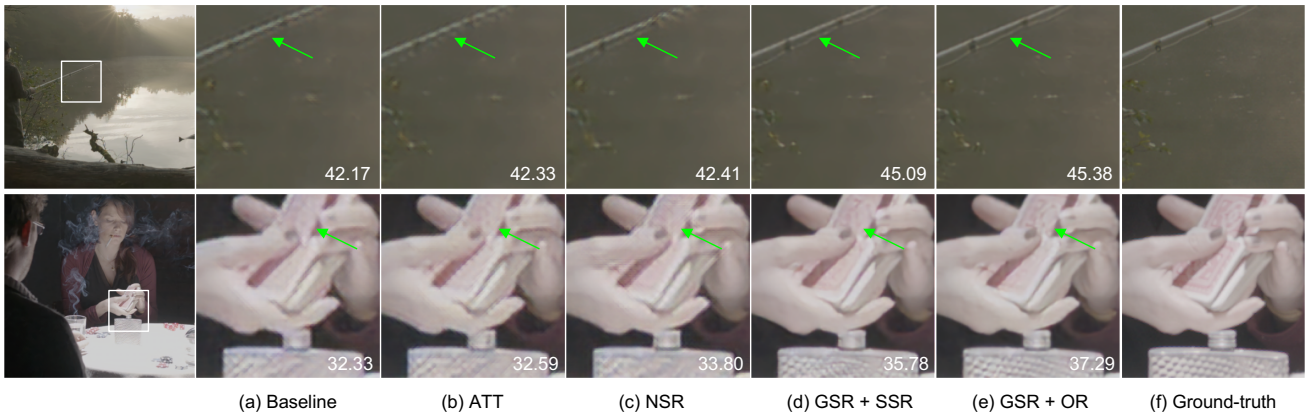


Figure 6: We investigated the effect of each module in the proposed network model. The network, GSR + OR, that has all of the proposed modules shows the best quality. Refer to Section 5.1 for the details of each network.

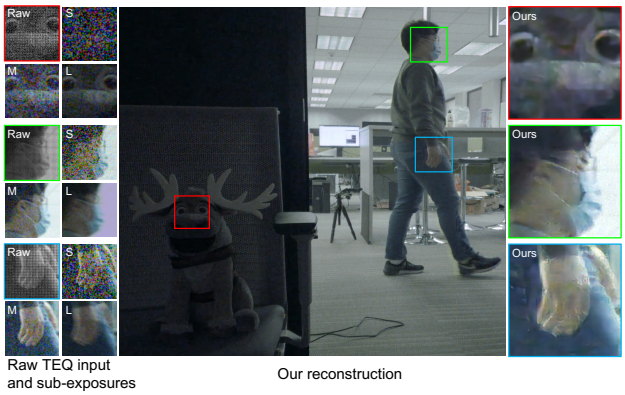


Figure 7: Our HDR reconstruction on a real sensor (Sony IMX708) output. Together with our reconstruction, the raw input and its sub-exposures are shown on the left. Note that the motion of the person induced the severe blur in L.

quad-bayer in (e) outperforms other methods by avoiding the ghosting artifacts and the noise in the dark area. In the quantitative evaluation shown in Table 2, our method shows the best scores among state-of-the-art HDR video methods.

4.3. HDR Reconstruction on Real TEQ Inputs

We reconstructed HDR from real tri-exposure quad-bayer (TEQ) inputs. A Sony IMX708 sensor was used to capture challenging real-life scenes with a large dynamic range and complex motion. Two results are shown in Figure 1 and Figure 7. From the raw patches and the sub-exposure patches extracted from them, shown on the left of the figure, saturation, severe noise, spatial artifact, and motion blur are observed in the inputs. For all of the magnified sample patches, our reconstruction is able to produce clean and sharp HDR images from severe corruption including noise and blur. More real results and videos will be available in the supplemental material.

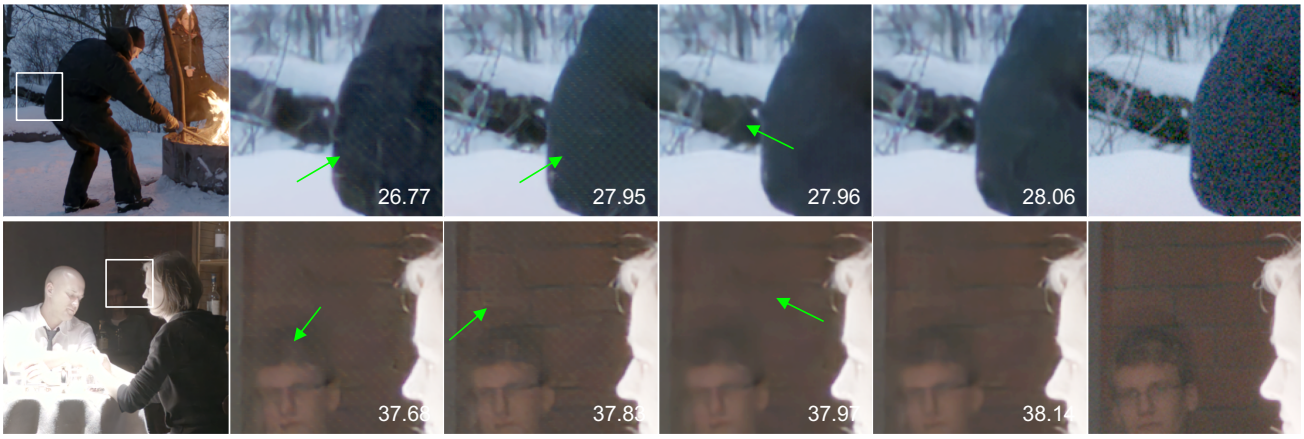
4.4. Model and Computation Complexity

Our network consists of 2.41 M parameters, which is smaller than 4.95 M of Tassano's and 11.76 M of Kalantari's. But Yan's model has the smallest number of parameters, 1.00 M. However, our computation complexities are 14.56 GFLOPS and 29.04 GMADD (Giga Multiply-Add), and they are less than 25.45 GFLOPS and 80.81 GMADD of Yan's. Kalantari's method is computationally lightest by having 9.46 GFLOPS and 22.00 GMADD. Tassano's method requires 12.32 GFLOPS and 24.59 GMADD. Note that the complexity is measured for an image patch of 256×256.

5. Ablation Studies

5.1. Study on the Model Architecture

We investigated the effect of the modules in our reconstruction network. We define a baseline reconstruction model that consists of an HDR fusion module and a temporal denoising module without any attention mechanisms. The HDR fusion module in the baseline does not perform the weight estimation, but it rather directly estimates an HDR feature. On the top of the baseline, we define a model called ATT by adding back the weight estimation and the attention mechanism to constitute exactly the same HDR fusion module and attention-based temporal denoising module as explained in Section 3.1 and 3.2. And NSR has a naive super-resolution module without the gate operation on top of ATT. In contrast, GSR+SSR network has a gated super-resolution module that uses a stacked image of subsampled color and exposure pixels as the input of the super-resolution feature extractor. The stacked image is of the quarter size of the original quad-bayer image. Finally, GSR+OR network is our proposed one with the gated super-resolution utilizing the original quad-bayer image.



(a) Single + TM (b) Multi + TM (c) Single + LDR (d) Multi + LDR (e) Ground-truth

Figure 8: We studied the effect of multi-frame inputs and our LDR-reconstruction loss. The configuration of our proposed network, Multi + LDR, shows the best quality. Refer to Section 5.2 for the details of each network.

	PSNR-μ	PuPSNR	HDR-VDP	HDR-VQM
Baseline	34.29	38.69	44.77	0.8663
ATT	34.83	39.53	45.40	0.8807
NSR	35.32	40.24	45.91	0.8924
GSR + SSR	36.07	41.43	45.97	0.9215
GSR + OR	36.06	41.64	46.16	0.9275

Table 3: The network, GSR + OR, that has all of the proposed module shows the best scores. Refer to Section. 5.1 for the details of each network.

The Weight Estimation and the Attention By comparing (a) and (b) in Figure 6, we observe ATT that has the weight estimation and the attention relatively less suffers from the artifacts in both dark and saturated region and helps to handle ghosting artifacts. The quantitative results in the first two rows of Table 3 agree with the observation.

The Explicit SR Module NSR that utilizes a naive explicit super-resolution module is clearly effective in improving the resolution and the blur problem as shown in Figure 6(c). This is also validated by the third row in Table 3.

The Gated SR Module Two networks, GSR+SSR and GSR+OR, that are equipped with the gated super-resolution module result in clear improvements over NSR with the naive super-resolution. In particular, for GSR+OR, we can further increase the resolution and the image quality by making use of the original quad-bayer input to extract super-resolution features. Compare (d) and (e) of Figure 6 and check out the quantitative results in Table 3.

5.2. Study on the Inputs and the Loss functions

We conducted an experiment to verify the effect of multi-frame inputs and the LDR-reconstruction loss proposed in Equation 9. Accordingly, we trained four different models shown in the first column of Table 4. The labels, Single and Multi, indicate the number of frames used as the inputs of the network. LDR stands out for the adoption of our LDR-reconstruction loss, and the naive tone-mapping loss function used in the previous works is indicated by TM.

	PSNR-μ	PuPSNR	HDR-VDP	HDR-VQM
Single+TM	35.89	41.58	46.06	0.9107
Multi+TM	36.06	41.64	46.16	0.9275
Single+LDR	35.99	42.01	48.43	0.9202
Multi+LDR	36.42	42.45	48.47	0.9247

Table 4: The network taking multi-frame inputs trained with our LDR-reconstruction loss shows the best scores. Refer to Section 5.2 for the details.

The Number of Input Frames The results of the multi-frame inputs in (b) and (d) of Figure 8 recover the dark region better than those of the single-frame inputs in (a) and (c). In the quantitative evaluation, we could validate this effect as shown in the second and the fourth rows of Table 4.

The LDR-recon. Loss By comparing Figure 8(b) and (d), it is shown that the network trained with our LDR-reconstruction loss is more robust to the noise. Also, our proposed model labeled by Multi+LDR in Table 4 shows the best scores of PSNR-μ, PuPSNR, and HDR-VDP.

6. Conclusion

We proposed a novel network-based HDR video reconstruction for tri-exposure quad-bayer (TEQ) sensors. Our network is able to produce high-quality HDR video. The experiments showed that HDR video with TEQ is more optimal than the temporally varying exposure, the burst shot strategy, and the interlacing HDR, especially for scenes with large dynamic range and moving objects. Our ablation studies validated the functionality of modules in our network and the proposed LDR-reconstruction loss function.

References

[1] V. G. An and C. Lee. Single-shot high dynamic range imaging via deep convolutional neural network. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017. 3

[2] Maryam Azimi, Amin Banitalebi-Dehkordi, Yuanyuan Dong, Mahsa T Pourazad, and Panos Nasiopoulos. Evaluating the performance of existing full-reference quality metrics on high dynamic range (hdr) video content. *arXiv preprint arXiv:1803.04815*, 2018. 5

[3] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE CVPR*, 2018. 2

[4] Hojin Cho, Seon Joo Kim, and Seungyong Lee. Single-shot high dynamic range imaging using coded electronic shutter. *Computer Graphics Forum*, 33(7), Oct. 2014. 2, 3

[5] Inchang Choi, Seung-Hwan Baek, and Min H. Kim. Reconstructing interlaced high-dynamic-range video using joint learning. *IEEE Transactions on Image Processing (TIP)*, 26(11), 2017. 1, 2, 3, 5

[6] Paul E. Debevec and Jitendra Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1997. 1, 2, 4

[7] DPREVIEW. *Redmi Note 7*, 2019 (accessed November 16, 2020). https://www.dpreview.com/news/8414546688/redmi-7-smartphone-offers-sony-48mp-quad-bayer-sensor-at-budget-price-point. 2

[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, 2014. 5

[9] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256, 2010. 5

[10] Clément Godard, Kevin Matzen, and Matt Uyttendaele. Deep burst denoising. In *ECCV*. Springer, 2018. 2

[11] Yulia Gryaditskaya, Tania Pouli, Erik Reinhard, Karol Myszkowski, and Hans-Peter Seidel. Motion aware exposure bracketing for hdr video. *Computer Graphics Forum*, 34(4), July 2015. 1, 2

[12] GSMArena. *Quad Bayer Sensors*, 2019 (accessed November 16, 2020). https://www.gsmarena.com/quad_bayer_sensors_explained-news-37459.php. 2

[13] J. Gu, Y. Hitomi, T. Mitsunaga, and S. Nayar. Coded rolling shutter photography: Flexible space-time sampling. In *IEEE ICCP*, 2010. 3

[14] M. Gupta, D. Iso, and S. K. Nayar. Fibonacci exposure bracketing for high dynamic range imaging. In *IEEE ICCV*, 2013. 1, 2

[15] Saghi Hajisharif, Joel Kronander, and Jonas Unger. HDR Reconstruction for Alternating Gain (ISO) Sensor Readout. In *Eurographics 2014 - Short Papers*. The Eurographics Association, 2014. 3

[16] Samuel W. Hasinoff, Dillon Sharlet, Ryan Geiss, Andrew Adams, Jonathan T. Barron, Florian Kainz, Jiawen Chen, and Marc Levoy. Burst photography for high dynamic range and low-light imaging on mobile cameras. *ACM Transactions on Graphics*, 35(6), 2016. 2, 6

[17] Felix Heide, Markus Steinberger, Yun-Ta Tsai, Mushfiqur Rouf, Dawid PajÄ…k, Dikpal Reddy, Orazio Gallo, Jing Liu abd Wolfgang Heidrich, Karen Egiazarian, Jan Kautz, and Kari Pulli. Flexisp: A flexible camera image processing framework. *ACM Transactions on Graphics*, 33(6), December 2014. 3

[18] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik G. Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *CVPR*. IEEE Computer Society, 2018. 5

[19] Haiyang Jiang and Yinqiang Zheng. Learning to see moving objects in the dark. In *IEEE ICCV*, October 2019. 2

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711. Springer, 2016. 5

[21] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics*, 36(4), 2017. 3

[22] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep HDR Video from Sequences with Alternating Exposures. *Computer Graphics Forum*, 2019. 1, 2, 3, 4, 6

[23] Nima Khademi Kalantari, Eli Shechtman, Connelly Barnes, Soheil Darabi, Dan B Goldman, and Pradeep Sen. Patch-based High Dynamic Range Video. *ACM Transactions on Graphics*, 32(6), 2013. 2

[24] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High dynamic range video. *ACM Transactions on Graphics*, 22(3), 2003. 1, 2

[25] J. Kronander, S. Gustavson, G. Bonnet, and J. Unger. Unified hdr reconstruction from raw cfa data. In *IEEE ICCP*, 2013. 5

[26] Orly Liba, Kiran Murthy, Yun-Ta Tsai, Tim Brooks, Tianfan Xue, Nikhil Karnad, Qiurui He, Jonathan T. Barron, Dillon Sharlet, Ryan Geiss, Samuel W. Hasinoff, Yael Pritch, and Marc Levoy. Handheld mobile photography in very low light. *ACM Transactions on Graphics*, 38(6), Nov. 2019. 2, 4

[27] Ziwei Liu, Lu Yuan, Xiaoou Tang, Matt Uyttendaele, and Jian Sun. Fast burst images denoising. *ACM Transactions on Graphics*, 33(6), Nov. 2014. 2

[28] Matteo Maggioni, Giacomo Boracchi, Alessandro Foi, and Karen Egiazarian. Video denoising using separable 4D nonlocal spatiotemporal transforms. In *Image Processing: Algorithms and Systems IX*. SPIE, 2011. 2

[29] Henrik Malm, Magnus Oskarsson, Eric Warrant, Petrik Clarberg, Jon Hasselgren, and Calle Lejdfors. Adaptive enhancement and noise reduction in very low light-level video (pdf). *IEEE ICCV*, 01 2007. 2

[30] Stephen Mangiat and Jerry Gibson. High dynamic range video with ghost removal. In *Applications of Digital Image Processing XXXIII*. SPIE, 2010. 1, 2

[31] Stephen Mangiat and Jerry Gibson. Spatially adaptive filtering for registration artifact removal in hdr video. 09 2011. 2

[32] 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*, 30(4), July 2011. 6

[33] Manish Narwaria, Matthieu Perreira Da Silva, and Patrick Le Callet. Hdr-vqm. *Image Commun.*, 35(C):46–60, July 2015. 6

[34] Shree Nayar and Tomoo Mitsunaga. High dynamic range imaging: Spatially varying pixel exposures. volume 1, 02 2000. 2, 3

[35] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NIPS*, pages 8024–8035. Curran Associates, Inc., 2019. 5

[36] Lars Rehm. *Samsung Tetracell*, 2020 (accessed November 16, 2020). https://www.dpreview.com/news/8188876144/samsung-isocell-gn1-sensor-tetracell-tech-and-phase-detection-on-all-active-pixels. 2

[37] Ana Serrano, Felix Heide, Diego Gutierrez, Gordon Wetzstein, and Belen Masia. Convolutional sparse coding for high dynamic range imaging. In *Proceedings of the 37th Annual Conference of the European Association for Computer Graphics*. Eurographics Association, 2016. 3

[38] Matias Tassano, Julie Delon, and Thomas Veit. DVDNET: A fast network for deep video denoising. In *IEEE ICIP*, 2019. 2

[39] Matias Tassano, Julie Delon, and Thomas Veit. Fastdvdnet: Towards real-time deep video denoising without flow estimation. In *IEEE CVPR*, pages 1351–1360, 2020. 2, 4, 6

[40] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *ECCV*. Springer, September 2018. 3, 4

[41] Xuan Dong, Guan Wang, Yi Pang, Weixin Li, Jiangtao Wen, Wei Meng, and Yao Lu. Fast efficient algorithm for enhancement of low lighting video. In *IEEE International Conference on Multimedia and Expo*, 2011. 2

[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. *IEEE CVPR*, 2019. 3, 4, 6

[43] Xinyi Zhang, Hang Dong, Zhe Hu, Wei-Sheng Lai, Fei Wang, and Ming-Hsuan Yang. Gated fusion network for joint image deblurring and super-resolution. In *BMVC*, page 153. BMVA Press, 2018. 3

[44] U. çoğalan and A. O. Akyüz. Deep joint deinterlacing and denoising for single shot dual-iso hdr reconstruction. *IEEE Transactions on Image Processing*, 29:7511–7524, 2020. 2, 3, 5, 6