HDR Imaging with Spatially Varying Signal-to-Noise Ratios

Yiheng Chi Xingguang Zhang Stanley H. Chan
School of Electrical and Computer Engineering
Purdue University
chi14@purdue.edu zhan3275@purdue.edu stanchan@purdue.edu

Abstract

While today's high dynamic range (HDR) image fusion algorithms are capable of blending multiple exposures, the acquisition is often controlled so that the dynamic range within one exposure is narrow. For HDR imaging in photon-limited situations, the dynamic range can be enormous and the noise within one exposure is spatially varying. Existing image denoising algorithms and HDR fusion algorithms both fail to handle this situation, leading to severe limitations in low-light HDR imaging.

This paper presents two contributions. Firstly, we identify the source of the problem. We find that the issue is associated with the co-existence of (1) spatially varying signal-to-noise ratio, especially the excessive noise due to very dark regions, and (2) a wide luminance range within each exposure. We show that while the issue can be handled by a bank of denoisers, the complexity is high. Secondly, we propose a new method called the spatially varying high dynamic range (SV-HDR) fusion network to simultaneously denoise and fuse images. We introduce a new exposure-shared block within our custom-designed multi-scale transformer framework. In a variety of testing conditions, the performance of the proposed SV-HDR is better than the existing methods.

1. Introduction

Today's high dynamic range (HDR) image fusion algorithms have demonstrated remarkable performances in blending images across a wide range of luminance levels. Many algorithms are able to handle an interior room with a sunlit view, of which the overall dynamic range is in the order of 100000:1 or more. However, most of these algorithms are designed for well-illuminated scenes. Even in the shortest exposure frame, the noise is maintained at a modest level so that the algorithm can focus on the blending task. The question we ask in this paper is: What if we push the shortest exposure to a photon-starving condition?

Such an extreme HDR problem arises in many low-light scenarios. Figure 1 is a real image captured by a Sony ILCE-7M2 camera. The imaging condition is a night-time scenario in front of a building. The challenge of the problem is the co-existence of heavy noise in the darkest spots of the image and the high dynamic range. We refer to this as the spatially varying signal-to-noise ratio (SNR) problem where brighter pixels have higher SNR and darker pixels have lower SNR.

Figure 1. [Top] Real captures using a Sony ILCE-7M2 camera. Three low-dynamic range (LDR) images are captured. [Bottom] Image denoising and HDR fusion results.

The goal of this paper is to articulate the spatially varying SNR problem. We emphasize the difficulty of the problem by referring to the performance of three state-of-the-art HDR fusion algorithms, namely, Kalantari and Ramamoorthi [17], Wu et al. [36], and NHDRRNet [39]. As we can see from Figure 1, these methods produce disappointing results, mostly failing in denoising the dark regions.

The position of the paper can be visualized in Figure 2. While existing HDR fusion methods can handle the blending task, the individual exposures are sufficiently high so that the amount of noise is limited. Single image denoisers today seldom handle the dynamic range problem. They are mostly focusing on a tonemapped image normalized to $[0, 1]$. Therefore, when facing a wide dynamic range scene,

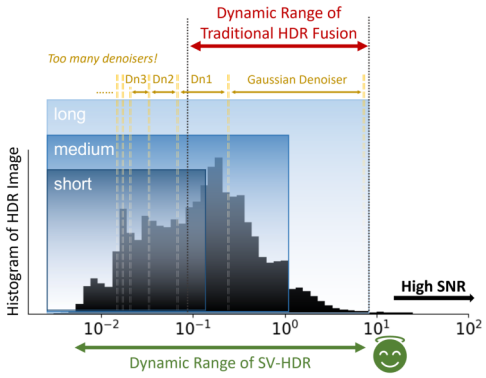


Figure 2. Traditional HDR algorithm can handle high SNR cases. Individual denoisers Dn1, Dn2, Dn3 have narrow operating regimes. SV-HDR offers a wide dynamic range coverage with denoising capability.

we need multiple denoisers to denoise the images before blending them. The proposed method solves both the noise problem and the dynamic range problem at once.

The main contribution of this paper is a new HDR fusion and denoising network called the spatially varying high dynamic range network (SV-HDR). SV-HDR simultaneously denoises the image and blends three exposures into a single HDR image. Our network is a transformer-based approach with three customized designs: (1) A multi-exposure transformer block to extract features. These transformers are adaptive to the varying SNRs. (2) We introduce an exposure-share block to blend the features coming from the three exposures. (3) We incorporate a multi-scale blending strategy to capture the local and global variations.

2. Problem Formulation and Related Work

In this section, we present the problem formulation and discuss the related work. We also briefly discuss a failure analysis of existing methods.

2.1. Realistic Image Sensor Model

We start by discussing the image formation model. Unlike the standard denoising problems where we can assume a Gaussian/Poisson model, in this paper, we need a precise model to emulate the actual image formation process.

A realistic image sensor model involves several parameters as determined by the following equation

$$z(\mathbf{x}) = \text{ADC}\left\{\text{Clip}\left\{\alpha \times \mathcal{P}(\tau \times \text{QE} \times (\theta(\mathbf{x}) + \mu_{\text{dark}}))\right\}\right\} + \mathcal{N}(0, \sigma^2_{\text{read}}). \quad (1)$$

Here, the underlying scene flux is denoted by $\theta(\mathbf{x})$ where $\mathbf{x} \in \mathbb{R}^2$ is the spatial coordinate. The observation is $z(\mathbf{x})$, which is pixel dependent. Other parameters such as the dark current and quantum efficiency, are listed in Table 1.

Table 1. Image Sensor Model Parameters

Symbol	Meaning	Typical Value
ADC	Analog-digital	14-bit
Clip	Full well limit	5000 e-
τ	Exposure	50ms
μ_{dark}	dark current	0.02 e-/s
QE	Quantum efficiency	50%
α	Conversion gain	1 at ISO 500
σ_{read}	Read noise	10 ADU at ISO 500
$\mathcal{P}$	Poisson distribution	
$\mathcal{N}$	Gaussian distribution	

The above equation is sufficient for the problem we are interested in. The model does not take into consideration of other "high-order" effects such as instability of the ADC thresholds, dead pixels (including non-responsive, overly sensitive, random telegraph signal, cross-talk, etc), pixel response non-uniformity, underflow offsets, and more. It also assumes a perfect color filter array, i.e., the electric and optical cross-talks are negligible.

2.2. Related Work

Classical HDR Fusion Methods. Debevec et al. [5] is one of the earliest papers that proposes combining multiple low dynamic range (LDR) images to construct an HDR image. Extending the idea to scenes that contain motion, people began to estimate the dynamic partial images and reject pixels [11, 13, 16, 20, 22, 23, 28, 30, 43]. Since pixels are dropped, these methods often suffer from loss of information. Another family of HDR fusion is registration-based, where pixels are first aligned with respect to a reference frame. The alignment can be done using optical flow [1, 19, 47], energy optimization [14, 33], and rank minimization [44]. These methods can handle small motions, but they still suffer from large foreground-object motions and they cannot manage noise.

Deep HDR Fusion Methods. Deep learning approaches have been used on single-image HDR reconstructions such as [8, 9, 32] and HDR video reconstructions such as [18]. There are also attempts at solving HDR fusion with large foreground motion. Kalantari and Ramamoorthi [17] propose a two-step fusion consisting of an optical flow estimation network and a fusion network. Wu et. al. [36] propose a single step method based on U-Net [31] and ResNet [12], where LDR frames are processed in different branches. Yan [37] brings the work forward by proposing a redesigned merging network and attention modules. Using attention to extract features is effective. The idea is adopted by various groups, e.g., [7, 40, 45]. Apart from attention, [39] uses a non-local network to solve the HDR fusion problem, and [38] uses optical flow with a multi-scale dense network.

Low light HDR with New Sensors. With the prolif-

eration of Quanta Image Sensor (QIS) and Single Photon Avalanche Diodes (SPAD), people began to explore new sensors for HDR in low light. [10] and [2] showed the theoretical performance of QIS, whereas [6, 15, 25, 42] showed the corresponding result using SPAD. There is also a new dual-exposure sensor reported by [3] which demonstrates denoising and deblurring for HDR video.

2.3. Why do Denoisers Fail?

As illustrated in Figure 1, existing HDR fusion algorithms fail to blend images captured in the nighttime. Our hypothesis is that it is the co-existence of the shot noise and the dynamic range that causes the difficulty. In this subsection, we confirm this speculation through two ablation studies.

Impact of noise on HDR fusion. The first ablation study is shown in Figure 3 which contains two sets of images simulated using our realistic image sensor model. The first set of images is *clean* inputs simulated at 100 photons per pixel on average. Two HDR fusion algorithms [17, 36] were applied to blend the frames. The results are reasonable despite some minor tone differences. The second set of images is simulated at 10 photons per pixel, which is substantially noisier than the first case. The result is very bad as we can see from the figure. It also confirms that when everything remains the same, the strength of the noise will have an immediate impact on the fusion algorithm.

(a) Exposures (b) Kalantari [17] (c) Wu [36]

Figure 3. Existing HDR methods for spatially varying SNR. [Top] When the inputs are clean, existing HDR methods perform well. [Bottom] When the inputs are noisy, the methods fail even if they are fine-tuned using the noisy data.

Impact of dynamic range on denoising. The second ablation study aims to verify the impact of the dynamic range on a denoiser. To this end, we build a toy denoiser that contains four sub-denoisers as shown in Figure 4. The denoisers we use are REDNet models [27]. In this toy design, we send the input image to the four sub-denoisers to handle the four different luminance ranges. The denoised images are then concatenated and sent to an attention module. Five weight masks are generated to combine the denoised images and the input to form the final image. We remark that this four-denoiser system is used to handle just *one* exposure in an exposure bracket. For a bracket that contains three exposures, we need three denoising systems followed by an HDR fusion module.

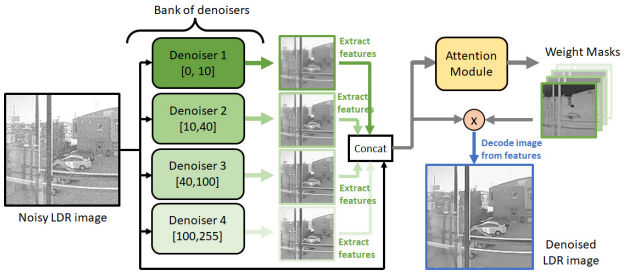


Figure 4. A toy denoising (system) that uses multiple denoisers to handle the luminance range. Note that this system handles only *one* exposure input. For an exposure bracket containing three or more exposures, each exposure would require one such denoising system. To alleviate the issue of requiring many denoisers, an *alternative* design is presented in this paper.

Figure 5 shows a comparison between using this (gigantic) 12-denoiser system (4 denoisers per exposure for 3 exposures) and a 3-denoiser system (1 denoiser per exposure for 3 exposures.) As we can see in Figure 5, if we have divided the exposure into a fine-enough sub-dynamic range, the denoiser is able to perform the required task. Therefore, we confirm that it is the dynamic range that limits the performance of the denoiser.

Figure 5. HDR fusion of noisy inputs. (a) Using 3 denoisers and the vanilla fusion algorithm by [17]. (b) Using the denoiser in Figure 4, followed by [17].

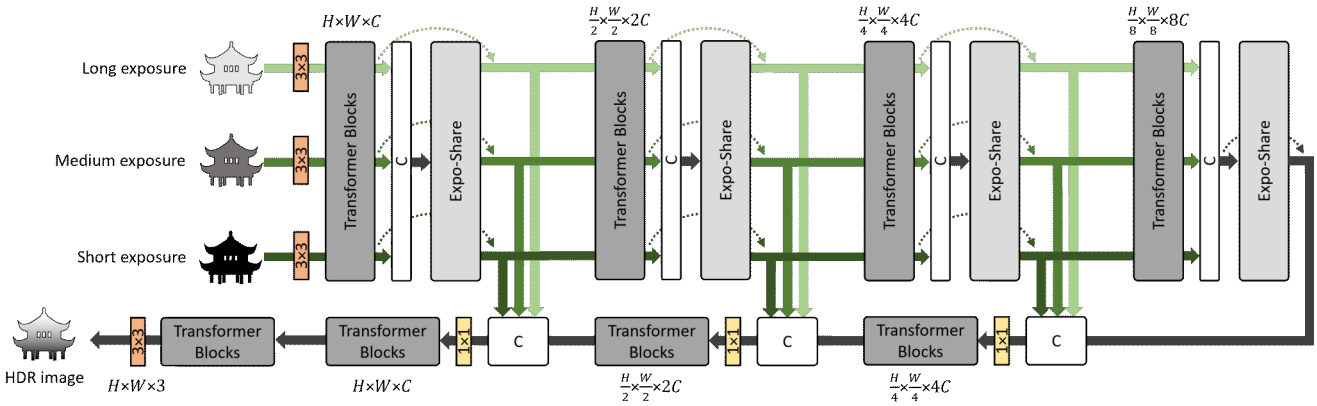


Figure 6. Structure of the proposed SV-HDR network. SV-HDR has a multi-scale design, spatially adaptive multi-exposure transformer blocks to extract and fuse features, and Expo-Share blocks to exchange and blend features across exposures.

3. Spatially Varying HDR (SV-HDR) Network

In this section, we present the proposed algorithm to simultaneously denoise and perform the multi-exposure fusion. Our goal is to develop a computationally efficient neural network that achieves the goal of Figure 4 while being much simpler. The overall architecture is shown in Figure 6.

3.1. Network architecture

The proposed SV-HDR is an encoder-decoder architecture following the U-shape structure [31]. Given a set of low dynamic range (LDR) images, the encoder extracts features from each of them, whereas the decoder generates the HDR results. Previous HDR fusion methods use different branches to process short, medium, and long exposure. In SV-HDR, we propose a single encoder to process *all* exposures to reduce the network capacity when handling a spatially varying SNR scene $\theta(\mathbf{x})$. This is achieved by using self-attention in the transformer block, which distinguishes the difference in exposure, gain, and noise variance. All LDR images are processed individually with the same encoder. In addition, we propose a new Expo-Share module to allow information exchange among the exposures. At the end of the encoder, three sets of features are concatenated. The decoder merges them and performs the HDR fusion.

Multi-scale. SV-HDR adopts a multi-scale design as shown in 6. The operation goes as follows. Given three LDR images $\mathbf{I}_i \in \mathbb{R}^{H\times W\times 3}$ at different exposures, the images are first mapped to the HDR scale by an inverse gamma correction and exposure normalization $\hat{\mathbf{I}}_i = \frac{\mathbf{I}_i^\gamma}{t_i}$, where t_i is the exposure time factor representing the multiplicative factor of the exposures $\theta(\mathbf{x}) + \mu_{\text{dark}}$. $\hat{\mathbf{I}}_i$ is then appended to $\mathbf{I}_i$ as additional channels, forming $\mathbf{J}_i \in \mathbb{R}^{H\times W\times 6}$ as the input to SV-HDR. SV-HDR first extracts shallow features from $\mathbf{J}_i$ by convolution, then extracts the deep features by a series of transformer blocks. The encoder uses an Expo-Share block at the end of each level to aid the feature exchange and motion mitigation. Details of the transformer blocks and Expo-Share blocks are discussed in the following sections. Down-sampling between the levels in the encoder is done by a 3×3 convolution and a pixel-unshuffling layer [34], symmetrically by convolutional and pixel-shuffling operations for up-sampling in the decoder. Skip connections are used to assist the information flow from the encoder to the decoder. Finally, the transformer blocks and a 3×3 convolution reconstruct the output HDR image $\hat{\mathbf{H}} \in \mathbb{R}^{H\times W\times 3}$ from the features. In our implementation, we set the number of multi-scale levels to 4.

Loss function. HDR imaging is about aggregating information from a wide range of signal levels. Standard loss function such as mean squared error is biased toward the bright pixels. In dark regions, subtle variations of pixel value will not be reflected in such losses but can be significantly influential compared to the signal level. Therefore, we need to compute the loss function between the tonemapped HDR image $\mathcal{T}(\hat{\mathbf{H}})$ and tonemapped ground truth HDR $\mathcal{T}(\mathbf{H})$. We use the differentiable tonemapping following [17, 36].

$$\mathcal{T}(\mathbf{H}) = \frac{\log(1+\mu\mathbf{H})}{\log(1+\mu)}, \tag{2}$$

where we set μ to be 5000. Our final loss function is:

$$\mathcal{L} = \|\mathcal{T}(\hat{\mathbf{H}}) - \mathcal{T}(\mathbf{H})\|_2. \tag{3}$$

3.2. Transformer Block

We use the visual transformers as the basic components of SV-HDR. A typical transformer block consists of a multi-head self-attention (MSA) module and a feed-forward module. Among several candidates of MSA and feed-forward, we experimentally found the best performing transformer block is a Shifted Window-based MSA (SW-MSA) proposed in [24] followed by a multi-layer perception (MLP)

of two fully connected layers with a GELU activation in between. We also use residual connections to skip over each of the SW-MSA and MLP modules. The experiment details are shown in Section 4.5 with results in Table 3.

SV-HDR has a serial of transformer blocks at each down-sample level. We use 4, 6, 6, and 8 transformer blocks for the four levels with down-sample ratios $1\times$, $2\times$, $4\times$, and $8\times$, respectively. The numbers of heads of MSA are 1, 2, 4, and 8 in the four levels. In the final refinement stage, four additional transformer blocks are used. The number of channels of features, C, is selected to be 48.

3.3. Expo-Share Block

To promote temporal information sharing across the exposures, we design an Expo-Share Block that jointly processes features from all exposures at the end of each level. At level l, the features $\mathbf{F}_i \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times 2^l C}$ are concatenated in the channel dimension before Expo-Share and are split after it, denoted $\tilde{\mathbf{F}}_i \in \mathbb{R}^{\frac{H}{2^l}\times\frac{W}{2^l}\times 2^l C}$. Considering that the scene can be dynamic across the frames, we use deformable convolutions [4, 46] to augment the spatial sampling region and implicitly align the features. Each Expo-Share Block contains three 3×3 deformable convolution layers, followed by three 1×1 convolutions to aggregate pixel-wise information across channels. GELU activation is used between each two consecutive layers. The overall operation of an Expo-Share block is shown in Figure 7. Features processed by the Expo-Share blocks are added back to itself with residual connections $\tilde{\mathbf{F}}_i + \mathbf{F}_i$ as illustrated in Figure 6, and go through the remaining encoder computations individually.

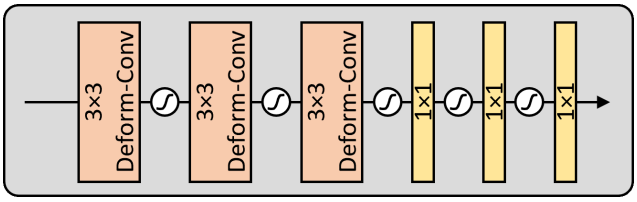


Figure 7. The proposed Expo-Share Block. Each block contains three sections of deformable convolutions and three 1 convolutions. The activation functions are the GELUs.

4. Experiments

4.1. Implementation Details

We train our model using the Kalantari dataset [17]. This dataset contains a large number of LDR images with a ground truth HDR image aligned with the middle exposure. We first augment the data by 8 times by randomly rotating and flipping the images, and we down-sample the images by $2\times$ and randomly crop the centers of images to 128×128 patches, where the down-sampling and cropping from center ensure these patches contain dynamic foreground objects. We then synthesize the noisy LDR image following our sensor model. To make our model applicable to a wide range of lighting conditions, we randomly sample τ such that the maximum photon count per pixel has a triangular distribution between 4 and 256 with a mode of 8. This is roughly equivalent to 0.005 to 0.323 lux at a typical exposure time of 1/50 sec and a pixel pitch of 6 microns for light with 400 to 700 nm wavelengths [29]. We fix QE at 50% and α at 1. The read noise σ_{read} is set to 0.0292, 0.1798, 1.4384 for short, medium, and long exposures, respectively, matching the characteristics of a Sony ILCE-7M2 camera at ISO of 200, 1600, and 12800 to enable testing on real images.

We train our model 1184 iterations per epoch for 300 epochs with a batch size of 3. We use the loss function in Eq. 3. We use the Adam optimizer [21] with parameters β_1 = 0.9 and β_2 = 0.999. Our learning rate is scheduled by a cosine scheduler with an initial rate of 0.0001. The training of our model takes 32 hours using an RTX 2080 Ti GPU.

4.2. Experiment Setup

We test our model on the test set of [17]. The noise synthesis process again follows 1. We fix τ at specific levels to conduct the qualitative and quantitative comparisons. We plot the test performance as curves PSNR versus τ.

We also conduct denoising and HDR fusion experiments on real images. We capture images with a Sony ILCE-7M2 camera and a Sigma Art 24-70mm F2.8 DG DN lens. We conduct two sets of real experiments. (I) We capture three images at a fixed ISO of 3200 and an aperture of f/2.8. The exposure time is 1/1250 sec, 1/160 sec, and 1/20 sec for the short, medium, and long exposure images. The setup of this experiment ensures the Poisson shot noise characteristics match our training setup. We further conduct a generalization experiment. (II) At each scene, we capture three images at a fixed exposure of 1/50 second, an aperture of f/5.0, and at varying ISO of 200, 1600, and 12800. Varying ISO is preferred to varying exposure time in certain imaging conditions, but a higher ISO amplifies the Poisson shot noise variance and corresponds to a higher read noise, so real experiment (II) is considered more challenging. SV-HDR can handle this challenging problem because it is adaptive to the non-uniform noise variance.

We compare the performances with Kalantari [17], Wu [36], AHDRNet [37], NHDRRNet [39]. As we demonstrate in Section 2.3, fine-tuning Kalantari's and Wu's models under our setup do not converge, nor do their fine-tuned models generate high-fidelity HDR reconstruction. Therefore, we test their model as-is. [37] and [39] have more sophisticated designs, so we fine-tune their models for 50,000 additional iterations in 50 epochs with batches of 8 and 32 samples sized 256×256. The learning rates start at 0.0001 and are scheduled using a polynomial decay with a power of 0.9.

(a) Input / **0.644 lux** (b) Kalantari [17] (c) Wu [36] (d) AHDRNet [37] (e) NHDRRNet [39] (f) **SV-HDR [Ours]** (g) Ground Truth

(h) Input / **0.080 lux** (i) Kalantari [17] (j) Wu [36] (k) AHDRNet [37] (l) NHDRRNet [39] (m) **SV-HDR [Ours]** (n) Ground Truth

Figure 8. Qualitative comparison using synthetic data. The input images of this figure are synthesized using the realistic sensor model and parameters. The corresponding photon level is marked in the input image. Dataset: [17].

4.3. Synthetic Experiments

Qualitative comparisons. We visually compare the results. Figure 8 shows the tone-mapped reconstruction results at a scene brightness roughly equivalent to 0.644 lux and 0.080 lux. Kalantari's and Wu's methods produce noisy HDR results. From Figure 8 (b), we also find that the optical flow aligning mechanism fails when noise is strong. This causes the fusion results to have ghosting effects. Although fine-tuned on the realistic sensor model, NHDRRnet cannot adapt sufficiently to the spatially varying SNR, making the fusion output over-smoothed at low-light regions. For example, the cars in the top image are over-smoothed. The proposed SV-HDR successfully removes noise and retains details at various lighting conditions.

Quantitative comparisons. We quantitatively analyze the performances using PSNR and MS-SSIM metrics. We also calculate PSNR and MS-SSIM on tonemapped images (PSNR-μ and MS-SSIM-μ) following 2. We set exposure τ such that the peak scene illuminance is roughly 0.4, 0.2, 0.1, and 0.05 lux. Table 2 shows the average PSNR and MS-SSIM scores. SV-HDR outperforms all competing methods in all metrics. We consider HDR-VDP-2 [26] a less suitable metric, because it is less sensitive to distortions in dark regions while our experiments simulate images captured at photon-limited scenes. We further plot and compare the PSNR curves versus the illuminance in Figure 9. The proposed SV-HDR has a consistently higher PSNR for an illumination level from 0.005 lux to 0.644 lux. We observe that NHDRRnet has a PSNR peak at about 0.01 lux where the majority of training samples lie, and its PSNR quickly drops as the illuminance decreases or increases. This phenomenon also suggests the difficulty of using a single network to handle a wide range of noise variance.

Illuminance (lux)	0.4	0.2	0.1	0.05	0.4	0.2	0.1	0.05
Method	PSNR / PSNR-μ (dB)				MS-SSIM / MS-SSIM-μ			
Kalantari [17]	30.04/20.33	30.08/19.40	30.10/18.11	30.11/16.53	0.8762/0.8661	0.8768/0.8202	0.8769/0.7540	0.8788/0.6749
Wu [36]	27.92/20.28	27.79/19.56	27.56/18.44	27.17/16.98	0.8215/0.7936	0.8151/0.7569	0.8055/0.7042	0.7914/0.6422
AHDRNet [37]	36.46/29.40	36.61/30.77	36.77/31.92	36.87/31.94	0.9664/0.9475	0.9687/0.9513	0.9711/0.9530	0.9721/0.9511
NHDRRNet [39]	32.60/22.46	33.45/23.78	34.59/25.35	35.99/27.24	0.9282/0.8481	0.9398/0.8784	0.9510/0.9050	0.9614/0.9231
SV-HDR [Ours]	**41.75/37.30**	**41.61/37.38**	**41.36/36.73**	**40.98/35.70**	**0.9835/0.9826**	**0.9831/0.9801**	**0.9822/0.9740**	**0.9805/0.9616**

Table 2. Reconstruction quality metric comparisons of various HDR fusion algorithms on a synthetic testing dataset.

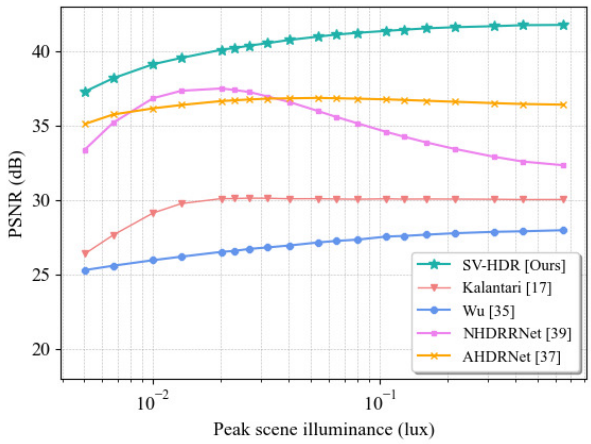


Figure 9. PSNR Curves of various HDR fusion algorithms. We note that SV-HDR stays at the top across all illumination levels.

Encoder	Decoder	PSNR (dB)
MDTA	W-MSA	25.48
MDTA	SW-MSA	25.48
MDTA	MDTA	39.68
W-MSA	W-MSA	38.91
W-MSA	SW-MSA	39.33
W-MSA	MDTA	39.37
SW-MSA	W-MSA	39.39
SW-MSA	**SW-MSA**	**40.28**
SW-MSA	MDTA	39.90

Feed-Forward	LeFF	DeLeFF	GDFN	**MLP**
PSNR (dB)	39.65	39.88	40.09	**40.34**

Table 3. Ablation study of different choices of encoders and decoders, and choices of feed-forward models.

4.4. Real Experiments

Visual comparisons for the real experiments are shown in Figure 11, where the top images show real experiment (I) and the bottom showing (II). We further show gamma-corrected raw images to highlight the presence of spatially varying noise. The reconstructions in (I) resonate with those of synthetic experiments. The competing methods, however, fail to fuse the low-light foreground to their reconstruction outputs in (II). In contrast, SV-HDR recovers these details and generates a wider dynamic range.

4.5. Ablation Study

MSA Module. We conducted an ablation study on types of multi-head self-attention (MSA) modules to build our SV-HDR network. Each MSA module is followed by an MLP. The candidates of MSA are: Window-based MSA (W-MSA) [35], Shifted Window-based MSA (SW-MSA) [24], and multi-Dconv head transposed attention (MDTA) [41]. We further experimented with different MSA choices for encoder and decoder. We test all configurations on the Kalantari test dataset [17] at a fixed τ of 4. Results are shown in Table 3 (Top). We empirically found that SW-MSA performs the best among all candidates in both the encoder and decoder. We also found using channel attention as a feature extractor can lead to performance degrades.

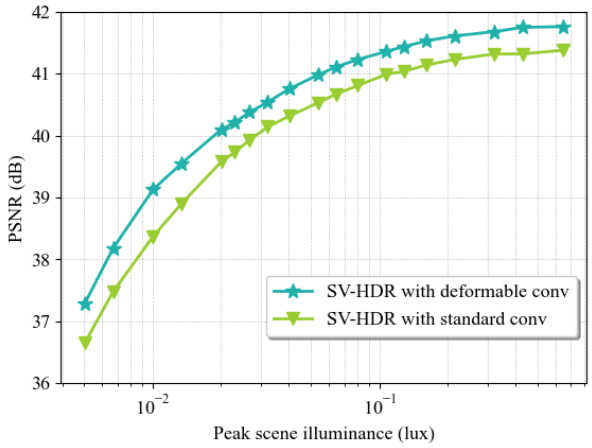
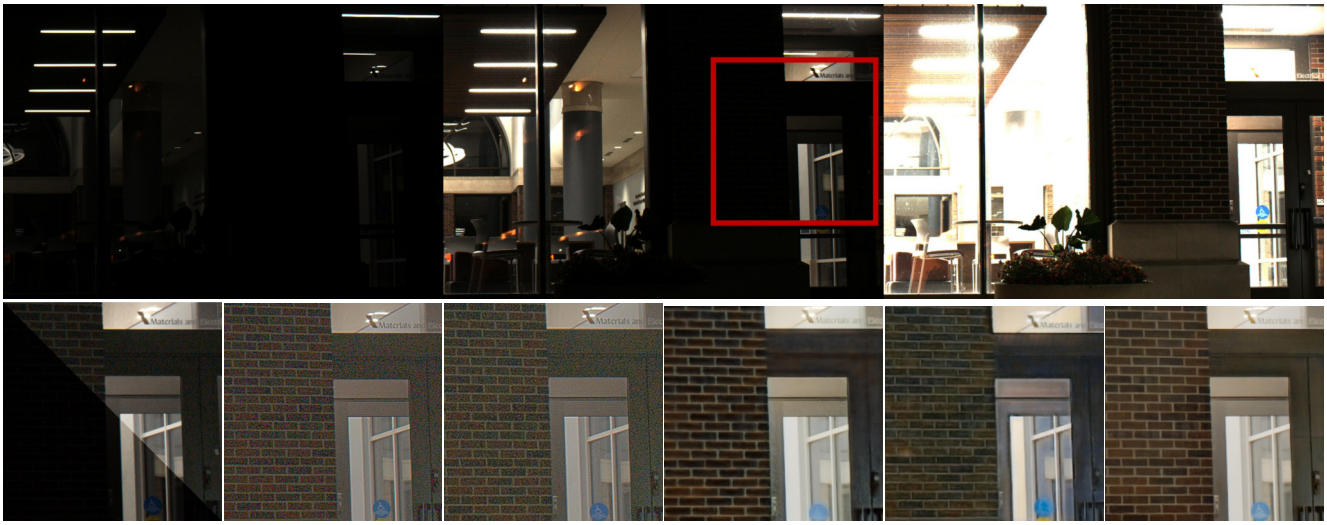


Figure 10. Ablation study where the deformable convolution is replaced by the standard convolution.

Feed-Forward Module. We also ablate on types of feed-forward modules that follow MSA, while fixing SW-MSA as the spatial attention. The candidates of feed-forward are MLP, Locally-enhanced Feed-Forward Network (LeFF) [35], Depth-wise Locally-enhanced Feed-Forward Network (DwLeFF), and Gated-Dconv feed-forward network (GDFN) [41]. Results are shown in Table 3 (Bottom). We found MLP outperforms all other feed-forward modules.



(a) Input / **1/160 sec** (b) Kalantari [17] (c) Wu [36] (d) AHDRNet [37] (e) NHDRRNet [39] (f) **SV-HDR [Ours]**

(g) Input / **ISO 12800** (h) Kalantari [17] (i) Wu [36] (j) AHDRNet [37] (k) NHDRRNet [39] (l) **SV-HDR [Ours]**

Figure 11. Qualitative comparison using real data. [Top] Real images captured at different exposure times. [Bottom] Real images captured with different ISO. Gamma-corrected (top right) and raw (bottom left) images are shown in (a) and (f) to highlight the presence of spatially varying noise.

Expo-Share Block. We further compared the performance of SV-HDR with the deformable convolutions in the Expo-Share blocks replaced by standard convolutions. We plot the average test PSNR against illuminance for both models in Figure 10. SV-HDR with deformable convolutions performs consistently better than with standard convolutions only.

5. Conclusion

HDR image fusion algorithms today typically lack the capability to handle noise. They tend to fail in low-light conditions, where signals will be contaminated heavily by the photon shot noise. In this paper, we provided evidence to show the existence of the problem. We found that the issue is caused by the co-presence of heavy noise and a wide dynamic range. A new HDR fusion and denoising network (SV-HDR) is presented as a solution to the problem. By introducing a customized multi-scale transformer and a new exposure-share block, we demonstrated the possibility of fusing noisy images with real cameras.

Acknowledgement. This work is supported, in part, by the US National Science Foundation under the grants ECCS-2030570, IIS-2133032, DMS-2134209, a gift from Google, and a gift from Intel Labs. The authors thank Dr. Vladlen Koltun for his continuous support of this project and valuable advice.

References

[1] 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

[2] Stanley H. Chan. What does a one-bit quanta image sensor offer? *IEEE Trans. Computational Imaging*, 8:770–783, Aug. 2022. 3

[3] Uğur Çoğalan, Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Hdr denoising and deblurring by learning spatio-temporal distortion models. *arXiv preprint arXiv:2012.12009*, 2020. 3

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 5

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

[6] Francescopaolo Mattioli Della Rocca, Tarek Al Abbas, Neale AW Dutton, and Robert K Henderson. A high dynamic range spad pixel for time of flight imaging. In *2017 IEEE SENSORS*, pages 1–3. IEEE. 3

[7] Yipeng Deng, Qin Liu, and Takeshi Ikenaga. Multi-scale contextual attention based hdr reconstruction of dynamic scenes. In *Twelfth International Conference on Digital Image Processing (ICDIP 2020)*, volume 11519, pages 413–419. SPIE, 2020. 2

[8] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM transactions on graphics (TOG)*, 36(6):1–15, 2017. 2

[9] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (TOG)*, 36(6):1–10, 2017. 2

[10] Abhiram Gnanasambandam and Stanley H. Chan. HDR imaging with Quanta Image Sensors: Theoretical limits and optimal reconstruction. *IEEE Trans. Computational Imaging*, 6:1571–1585, 2020. 3

[11] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. 2006. 2

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 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] Atul Ingle, Andreas Velten, and Mohit Gupta. High flux passive imaging with single-photon sensors. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6760–6769, 2019. 3

[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] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Transactions on Graphics (TOG)*, 36(4):1–12, 2017. 1, 2, 3, 4, 5, 6, 7, 8

[18] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep hdr video from sequences with alternating exposures. In *Computer graphics forum*, volume 38, pages 193–205. Wiley Online Library, 2019. 2

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

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[22] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE signal processing letters*, 21(9):1045–1049, 2014. 2

[23] Hui Li, Kede Ma, Hongwei Yong, and Lei Zhang. Fast multi-scale structural patch decomposition for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 29:5805–5816, 2020. 2

[24] Jingyun Liang, Jiezhang Cao, Guolei Sun, Kai Zhang, Luc Van Gool, and Radu Timofte. Swinir: Image restoration using swin transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1833–1844, 2021. 4, 7

[25] Sizhuo Ma, Shantanu Gupta, Arin C. Ulku, Claudio Brushini, Edoardo Charbon, and Mohit Gupta. Quanta

burst photography. *ACM Transactions on Graphics*, 39(4), Jul. 2020. 3

[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. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11, New York, NY, USA, 2011. Association for Computing Machinery. 6

[27] Xiao-Jiao Mao, Chunhua Shen, and Yu-Bin Yang. Image restoration using convolutional auto-encoders with symmetric skip connections. *arXiv preprint arXiv:1606.08921*, 2016. 3

[28] 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

[29] Russell J. Palum. How many photons are there? In *PICS*, 2002. 5

[30] 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

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2, 4

[32] 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 (TOG)*, 39(4):80–1, 2020. 2

[33] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Transactions on Graphics (TOG)*, 31, 11 2012. 2

[34] 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. 4

[35] Zhendong Wang, Xiaodong Cun, Jianmin Bao, Wengang Zhou, Jianzhuang Liu, and Houqiang Li. Uformer: A general u-shaped transformer for image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17683–17693, June 2022. 7

[36] 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. 1, 2, 3, 4, 5, 6, 7, 8

[37] 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, 5, 6, 7, 8

[38] Qingsen Yan, Dong Gong, Pingping Zhang, Qinfeng Shi, Jinqiu Sun, Ian Reid, and Yanning Zhang. Multi-scale dense networks for deep high dynamic range imaging. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 41–50. IEEE, 2019. 2

[39] 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. 1, 2, 5, 6, 7, 8

[40] Qingsen Yan, Song Zhang, Weiye Chen, Yuhang Liu, Zhen Zhang, Yanning Zhang, Javen Qinfeng Shi, and Dong Gong. A lightweight network for high dynamic range imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2022. 2

[41] 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. 7

[42] Majid Zarghami, Leonardo Gasparini, Matteo Perenzoni, and Lucio Pancheri. High dynamic range imaging with tdc-based cmos spad arrays. *Instruments*, 3(3), 2019. 3

[43] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2011. 2

[44] Jinghong Zheng, Zhengguo Li, Zijian Zhu, Shiqian Wu, and Susanto Rahardja. Hybrid patching for a sequence of differently exposed images with moving objects. *IEEE transactions on image processing*, 22(12):5190–5201, 2013. 2

[45] Lingkai Zhu, Fei Zhou, Bozhi Liu, and Orcun Göksel. Hdrfeat: A feature-rich network for high dynamic range image reconstruction. *arXiv preprint arXiv:2211.04238*, 2022. 2

[46] Xizhou Zhu, Han Hu, Stephen Lin, and Jifeng Dai. Deformable convnets v2: More deformable, better results. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9308–9316, 2019. 5

[47] Henning Zimmer, Andrés Bruhn, and Joachim Weickert. Freehand hdr imaging of moving scenes with simultaneous resolution enhancement. In *Computer Graphics Forum*, volume 30, pages 405–414. Wiley Online Library, 2011. 2