This WACV 2022 paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

High Dynamic Range Imaging of Dynamic Scenes with Saturation Compensation but without Explicit Motion Compensation

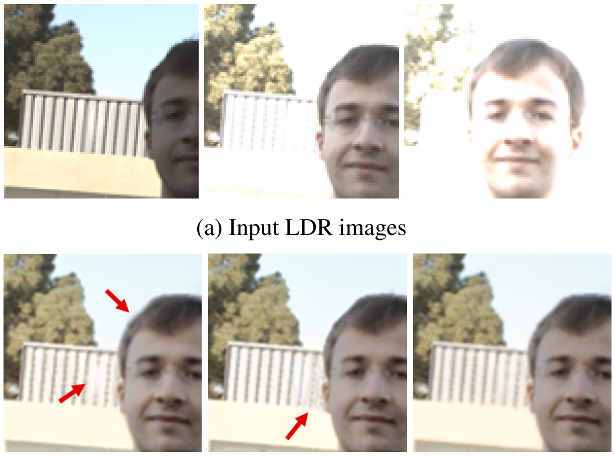
Haesoo Chung Nam Ik Cho
Department of ECE, INMC, Seoul National University, Korea
reneeish@ispl.snu.ac.kr, nicho@snu.ac.kr

Abstract

High dynamic range (HDR) imaging is a highly challenging task since a large amount of information is lost due to the limitations of camera sensors. For HDR imaging, some methods capture multiple low dynamic range (LDR) images with altering exposures to aggregate more information. However, these approaches introduce ghosting artifacts when significant inter-frame motions are present. Moreover, although multi-exposure images are given, we have little information in severely over-exposed areas. Most existing methods focus on motion compensation, i.e., alignment of multiple LDR shots to reduce the ghosting artifacts, but they still produce unsatisfying results. These methods also rather overlook the need to restore the saturated areas. In this paper, we generate well-aligned multi-exposure features by reformulating a motion alignment problem into a simple brightness adjustment problem. In addition, we propose a coarse-to-fine merging strategy with explicit saturation compensation. The saturated areas are reconstructed with similar well-exposed content using adaptive contextual attention. We demonstrate that our method outperforms the state-of-the-art methods regarding qualitative and quantitative evaluations.

1. Introduction

With the development of high dynamic range (HDR) display, the demand for HDR content is rapidly increasing. HDR content can provide the viewers rich perceptual experiences and also enhance the performance of subsequent computer vision tasks. Since the direct acquisition of HDR images is practically tricky and requires expensive imaging devices, HDR imaging techniques using low dynamic range (LDR) images are drawing considerable attention. There have been many methods to generate an HDR image from a single LDR input for this reason, where earlier methods just stretched the dynamic range of the LDR input [24, 1, 31, 10, 5, 6, 41, 16, 22, 23], and some recent methods learned LDR-to-HDR mapping using convolutional neural networks (CNNs) [11, 12, 30, 50, 26, 27]. However, these single-image HDR reconstruction methods usually suffer from information loss in under-exposed or over-exposed areas. Specifically, a large portion of the LDR image content is washed out in a scene with large lighting variations, which is hard to be recovered.


(a) Input LDR images

(b) Yan [47] (c) Prabhakar [38] (d) Ours

Figure 1: Input images with large-scale motions in the over-exposed areas provide insufficient information for HDR reconstruction. Our method successfully hallucinates details in the saturated regions by aggregating similar well-exposed content. The results are visualized after tonemapping.

Hence, there have also been many methods exploiting multi-exposure images to collect more information in dark or very bright areas [9, 28, 20, 13, 14, 4, 2, 3, 54, 3, 25, 35]. But, these methods deliver satisfying performances only when the multi-exposure images are perfectly aligned or have slight movements, which is a hardly realistic situation.

To deal with more dynamic scenes with foreground and background motions, many researchers attempted to align the input images and integrate the aligned LDR images into an HDR result [19, 57, 18, 42, 15, 37, 18, 45, 48, 39]. For example, several works [18, 36] used optical flow for aligning input images and passed the aligned images to a fusion network. Wu *et al*. [45] adopted homography transformation for background alignment and the U-Net for HDR reconstruction. Prabhakar *et al*. [38] used both homography and optical flow. Meanwhile, Yan *et al*. [47, 49] utilized attention modules and non-local blocks, respectively, to implicitly align the input features. Most of these methods concentrated on accurate alignment, and then they used relatively simple techniques in merging the aligned LDR images or features. Niu *et al*. [34] adopted GAN to synthesize missing content but did not perform any explicit hallucination process. In contrast, we present a coarse-to-fine HDR reconstruction strategy with consideration of the saturated areas. Since details in the over-exposed parts are hardly preserved, we employ the hallucination method. Also, we do not explicitly use optical flow or alignment but transfer the brightness of the multi-exposure images to a reference LDR image so that we can obtain multi-exposure features having the same structure as the reference.

More specifically, we propose an end-to-end framework with two sub-networks for HDR imaging of dynamic scenes. First, we present a set of brightness adjustment networks (BANs) that takes the multi-exposure inputs and generates multi-exposure features aligned to those of the reference image. While most of the existing methods transform the pixel position and value of multi-exposure images with respect to a reference, our method transfers the brightness of multi-exposures to the reference to have perfectly aligned multi-exposure images. To this end, each BAN adjusts the exposure of the reference image while retaining its structure using pixel-adaptive deformable convolutions. In addition, we introduce a coarse-to-fine merge-and-hallucination network (MAHN) to integrate the set of multi-exposure features into an HDR image and hallucinate details in the saturated regions. The hallucination is needed because we still have insufficient information in challenging areas, even with the multi-exposure images. For example, when all the images are over-exposed, or occlusions exist in the highlighted areas, naively merging the images leads to poor results, as shown in Fig. 1. To address this problem, we first coarsely generate an HDR image and then hallucinate content in the saturated regions subsequently. Specifically, we estimate the long-range correlations between the saturated patch and the well-exposed ones and then fill the saturated area with the correlated well-exposed content at the feature level. Extensive quantitative and qualitative evaluations demonstrate that our method generates a high-quality HDR image from LDR images in dynamic scenes.

The main contributions of this paper can be summarized as follows:

- We propose a brightness adjustment network (BAN) to generate the well-aligned features with different brightness. We reformulate the difficult image-alignment problem into an easier brightness-adjustment problem, which significantly alleviates the ghosting artifacts.
- We propose a merge-and-hallucination network (MAHN) to integrate the aligned multi-exposure features into an HDR result while hallucinating details in the saturated regions. The MAHN explicitly fills the saturated areas with similar well-exposed content.

2. Related Work

Single-image HDR reconstruction Single-image HDR reconstruction, also referred to as reverse tone mapping or inverse tone mapping, has been studied for decades. Early works apply global pixel transformations [24, 1], edit local areas [31, 10], or utilize an expand map to enhance the highlighted regions [5, 6, 41]. Recently, CNN-based methods [53, 30] are introducing end-to-end networks to learn the LDR-to-HDR mapping. Several works [12, 26] synthesize a multi-exposure stack and combine them into an HDR image, while Eilertsen *et al*. [11] only restore the saturated regions. More recent approaches [50, 27] generate an HDR image by reversing the LDR image formation procedure. These methods, however, struggle when severely under-/over-exposed areas exist.

Multi-image HDR reconstruction Multiple images with different exposures can provide richer information for HDR reconstruction. Some conventional methods [9, 28] capture a series of LDR images and merge them under the assumption that the scene is static. However, since camera and object motions are inevitable in the real world, the subsequent works propose various methods to handle the displacements. A number of approaches [20, 13, 14, 40, 54, 3, 25, 35] assume that input images are globally aligned and focus on detecting and rejecting moving pixels. However, these methods lose considerable information by dropping the pixels with motions. To perform an explicit alignment, several methods [19, 57, 18] exploit optical flow, but flow estimation error easily leads to distortions in the resulting image. Meanwhile, Sen *et al*. [42] and Hu *et al*. [15] rely on patch-based correspondences.

Recently, CNN-based multi-exposure HDR imaging methods have been developed. For some examples, Kalantari *et al*. [18] compensate for motions using an optical flow algorithm and merge the resulting images using a simple CNN. Wu *et al*. [45] first perform homography transformation to align background motions and pass the aligned images to an image translation network. Yan *et al*. [47] leverage spatial attention to exclude the misaligned components and construct a deep network for merging, while Yan *et al*. [49] use non-local blocks to align the input features. Also, Yan *et al*. [48] and Prabhakar *et al*. [36] align the images

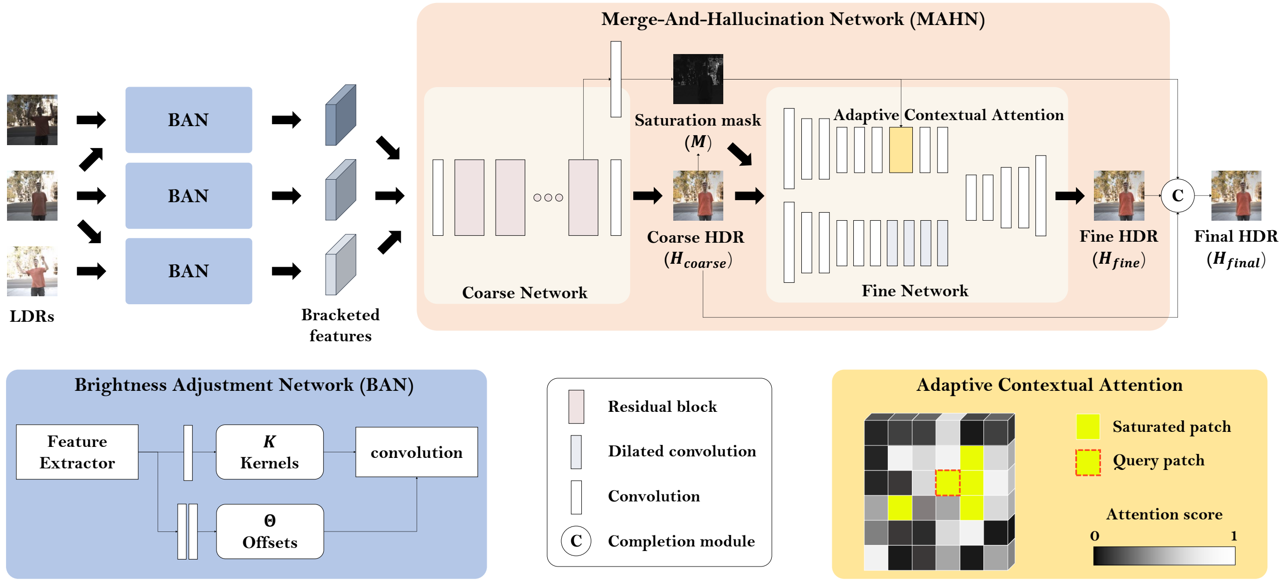


Figure 2: Overview of the proposed framework. The brightness adjustment network (BAN) adjusts the brightness of the reference image to be matched with the corresponding input. The BAN generates a unique kernel and offset value for each location and applies adaptive convolutions to the reference feature (bottom left). The well-aligned bracketed features obtained from a set of BANs are fed to the merge-and-hallucination network (MAHN). The MAHN first coarsely merges the bracketed features into the coarse HDR image and then hallucinates details in saturated areas using adaptive contextual attention in the fine network. The saturated patches are reconstructed with a weighted sum of the well-exposed content in the adaptive contextual attention layer according to the estimated attention scores. An example of the attention scores for a single query patch is illustrated in the bottom right corner, and the attention for all the saturated patches is computed in the same way.

using optical flow and feed them into a fusion network. Pu *et al.* [39] use deformable convolution to align the dynamic input images, and Niu *et al.* [34] utilize residual merging blocks for alignment and expect the adversarial learning to help to restore missing details. These approaches mostly give weight to the sophisticated alignment and expect the merging network to combine the aligned features/images well. On the contrary, we present a coarse-to-fine HDR reconstruction strategy with supervision on the saturated regions and a brightness adjustment method to produce well-aligned bracketed features.

Flexible convolutions Jia *et al.* [17] propose a dynamic filter network to generate per-pixel filters conditioned on an input, which has been applied to various tasks dealing with motions [33, 32, 21]. Meanwhile, Dai *et al.* [8] present deformable convolution to enable flexible operations with learnable offsets. Zhu *et al.* [56] extend this work by introducing a modulation factor. Notably, the deformable convolution has been used in various fields related to videos [7, 55]. Especially, recent video super-resolution methods [43, 44, 46, 51] leverage deformable convolutions to align the multiple input frames. In this work, we predict spatially varying deformable kernels to deal with a challenging image pair with significant motion and exposure difference.

3. Proposed Method

Given a series of LDR images $\{I_{-N}, \ldots, I_0, \ldots, I_N\}$ sorted by their exposure biases, our goal is to generate an artifact-free HDR image H_{final}. The proposed method consists of two stages, as shown in Fig. 2. First, we place the proposed BANs to control the brightness of the reference image according to other exposure ones. Instead of aligning an exposure image to fit the structure of the reference, we adjust the brightness of the reference with respect to other exposures using the proposed BANs. Each BAN takes the reference and a differently-exposed image and applies pixel-adaptive deformable convolutions to the reference feature to adjust its brightness to that of the different exposure. The resulting multi-exposure features are stacked and fed into the MAHN, which first merges the given features into a coarse HDR image H_{coarse} and then hallucinates details in the saturated regions. The saturated areas are identified by the network and represented as a saturation mask M. The fine network then computes contextual attention [52] adaptively to find similar content from unsaturated regions and fills the saturated areas with the correlated well-exposed content according to the attention scores. The completion module outputs the final HDR result H_{final} by replacing the saturated parts in the coarse HDR image H_{coarse} with the corresponding ones in the fine HDR image H_{fine}.

Following a previous work [18], which provides a well-prepared dynamic multi-exposure dataset, we use three LDR images $\{I_{-1}, I_0, I_1\}$ and set the middle exposure image I_0 as the reference image in terms of structure. Here, we use static input images as well as original dynamic input images for training. The static image set $\{I^s_{-1}, I^s_0, I^s_1\}$ is generated by adjusting the exposure of the ground truth HDR image H and then applying gamma correction and clipping:

$$I^s_i = clip((Ht_i)^{1/\gamma}),\ i = -1, 0, 1, \tag{1}$$

where t_i denotes the exposure time of the corresponding dynamic input image I^d_i and γ denotes the gamma correction parameter. γ is set as 2.2 in our experiments. Since the proposed BAN aims to generate the bracketed features instead of learning motions, using static images as inputs does not hinder its training. The static images can serve as easy training samples. Then, we map the LDR images to the HDR images $\{H_{-1}, H_0, H_1\}$ using gamma correction:

$$H_i = \frac{I^\gamma_i}{t_i},\ i = -1, 0, 1, \tag{2}$$

We concatenate the LDR images with these HDR images along the channel dimension to obtain the 6-channel inputs $\{X_{-1}, X_0, X_1\}$. Our framework f is represented as follows:

$$H_{final} = f(X_{-1}, X_0, X_1). \tag{3}$$

3.1. Brightness Adjustment Network (BAN)

Given the reference image I_0 and the supporting image I_i, the BAN aims to adjust the brightness of the reference image I_0 in accordance with the exposure of the supporting image I_i. To this end, the BAN extracts features to predict spatially-varying deformable convolution kernels and applies the adaptive convolutions to deal with the input pair with motion and brightness difference. Note that we perform self-adjustment in the middle BAN. The generated features are free from the ghosting artifacts since we do not compensate for large motions between the input images but generate the adjusted features of I_0 that do not have structure-difference from the reference.

Feature extraction Our feature extractor has an individual branch for each input and fuses the information from two branches in a progressive manner. The features from the reference image are integrated into the supporting branch so that multi-level information is propagated. The detailed architecture of the feature extractor is illustrated in the supplementary material.

Adaptive convolution The extracted features are then passed to two separate paths to produce convolution kernels K and offsets Θ. The kernels K and the offsets Θ are unique for each position $\mathbf{p}_0$ on the feature map. Here, we set the kernel size as 3×3 and the regular grid as $\mathcal{R} = \{(-1,-1), (-1,0), \ldots, (1,1)\}$. With the pre-specified offset $\mathbf{p}_n \in \mathcal{R},\ n = 1, \ldots, |\mathcal{R}|$, the adaptive convolution is applied to each location $\mathbf{p}_0$ on the reference feature F_0 to generate the adjusted feature $\bar{F}_{0,i}$ whose brightness matches with the one of the supporting feature F_i:

$$\bar{F}_{0,i}(\mathbf{p}_0) = \sum_{\mathbf{p}_n \in \mathcal{R}} K(\mathbf{p}_0 + \mathbf{p}_n) F_0(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n), \tag{4}$$

where $\Delta\mathbf{p}_n \in \Theta$ is the learnable offset. Since $\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n$ can be fractional, bilinear interpolation is used to compute the value $F_0(\mathbf{p}_0 + \mathbf{p}_n + \Delta\mathbf{p}_n)$.

Unlike previous works [43, 44, 46, 51] which adopt the deformable convolution, the proposed method applies the convolutions to the reference feature, and the supporting image is only involved in feature extraction. We prevent undesirable ghosting artifacts by converting the alignment problem into an easier brightness adjustment problem. We show the effect of this reformulation in Section 4.3. Furthermore, learning per-pixel kernels as well as offsets enables more flexible operations.

3.2. Merge-And-Hallucination Network (MAHN)

For generating a high-quality HDR result, the MAHN first merges the well-aligned bracketed features into a coarse HDR image H_{coarse} and then hallucinates details within the saturated areas by aggregating similar content in the unsaturated (*i.e.*, well-exposed) areas. The coarse network integrates the bracketed features using residual blocks. The following fine network is branched into two paths: a hallucination branch and a refinement branch. While the refinement branch refines the coarse result with dilated convolutions locally, the hallucination branch explicitly fills the saturated regions with similar valid content by estimating long-range adaptive contextual attention.

Our core idea is to find the correlated valid content to the saturated parts and assemble them to compensate for the lost content in the saturated areas. First, we construct a saturation mask M to indicate the saturated regions. We place a convolutional layer with a sigmoid activation function after the last residual block in the coarse network to obtain the sample-specific saturation mask. The saturation mask M represents how saturated each pixel is with values in range $[0, 1]$. An example is shown in Fig. 3. This adaptive mask generation strategy enables sample-specific hallucination, contrary to pre-defined rules such as thresholding. The

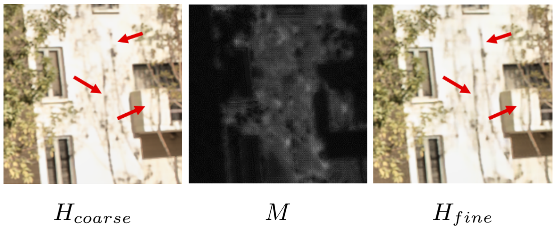


Figure 3: The saturation mask M represents the saturation level of the coarse HDR image H_{coarse}. As a result of hallucination using the adaptive contextual attention, the fine HDR image H_{fine} clearly retains richer details in the saturated parts.

hallucination branch fills the over-exposed regions (close to value 1 on the M) with the correlated well-exposed content (close to value 0 on the M) by measuring the patch-wise cosine similarity:

$$s_{x,y,x',y'} = \frac{o_{x,y}}{\|o_{x,y}\|} \cdot \frac{w_{x',y'}}{\|w_{x',y'}\|}, \tag{5}$$

where $o_{x,y}$ denotes the over-exposed patch at (x, y) and $w_{x,y}$ denotes the well-exposed patch at (x', y'). To obtain the attention scores, we apply softmax along the $x'y'$-dimension and then multiply $1 - M$ which represents well-exposedness so that valid patches can contribute more to the reconstruction. We replace the over-exposed patches with a combination of the well-exposed patches according to the estimated attention scores. The effectiveness of our adaptive contextual attention is validated in Section 4.3. Note that the whole process is implemented using convolution operations.

The saturation-compensated features are concatenated with the refined features from the refinement branch and go through additional layers to reconstruct the fine HDR image H_{fine}. The fine HDR image H_{fine} exhibits clear improvements in the saturated areas as shown in Fig. 3. Finally, the completion module generates the final HDR image H_{final} by replacing the saturated pixels of H_{coarse} with the ones of H_{fine}:

$$H_{final} = (1 - M) \odot H_{coarse} + M \odot H_{fine}, \tag{6}$$

where $\odot$ denotes the Hadamard product.

3.3. Training Loss

We propose a hybrid loss to enhance both fidelity and perceptual quality. Since HDR images are mostly tonemapped for displaying, we compute the loss functions between the tonemapped predicted HDR image $\mathcal{T}(\hat{H})$ and the tonemapped ground truth HDR image $\mathcal{T}(H)$. The HDR image H is tonemapped with the differentiable μ-law:

$$\mathcal{T}(H) = \frac{\log(1 + \mu H)}{\log(1 + H)}, \tag{7}$$

where μ is a parameter that defines the level of compression. μ is set as 5000.

Reconstruction loss We use a simple ℓ_1 reconstruction loss to minimize the distance between $\mathcal{T}(\hat{H})$ and $\mathcal{T}(H)$. The reconstruction loss is defined as:

$$\mathcal{L}_{recon} = \left\| \mathcal{T}(\hat{H}) - \mathcal{T}(H) \right\|_1. \tag{8}$$

Color loss To address the color shift problem, we also define color loss, which is based on the cosine similarity between the RGB vectors of $\mathcal{T}(\hat{H})$ and $\mathcal{T}(H)$. Formally, it is described as:

$$\mathcal{L}_{color} = 1 - \frac{1}{N} \sum_{n=1}^{N} \frac{\hat{\mathbf{v}}_n \cdot \mathbf{v}_n}{\|\hat{\mathbf{v}}_n\| \, \|\mathbf{v}_n\|}, \tag{9}$$

where N is the total number of pixels of the HDR image, and $\hat{\mathbf{v}}_n$ and $\mathbf{v}_n$ denote the RGB vectors at the n-th pixel of $\mathcal{T}(\hat{H})$ and $\mathcal{T}(H)$, respectively.

Perceptual loss To generate a more realistic texture, we adopt the VGG loss $\mathcal{L}_{vgg}$ as our perceptual loss. We use three feature maps after the first, second, and third block of VGG-16 network for the VGG loss. The total variation (TV) loss $\mathcal{L}_{tv}$ is also included for smoothness.

Total loss From the above losses, the overall loss is defined as $\mathcal{L} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{color}\mathcal{L}_{color} + \lambda_{vgg}\mathcal{L}_{vgg} + \lambda_{tv}\mathcal{L}_{tv}$, which is applied to the coarse output H_{coarse}, the fine output H_{fine}, and the final output H_{final} with different weights. We empirically set the corresponding weights as specified in Table 1.

Table 1: The weights for each loss constituting our total loss.

	λ_{recon}	λ_{color}	λ_{vgg}	λ_{tv}
H_{coarse}	1	1	0.001	0.1
H_{fine}	1	1	0.001	0.1
H_{final}	1	0	0	0

3.4. Implementation Details

We sample patches of size 128×128 from the training images and apply augmentations: random rotation and flipping. We use Adam optimizer with a learning rate of 10^{-4} and set the batch size as 16. Each batch is composed of the dynamic images and the static images with the ratio of $3 : 1$. We train our model on a single NVIDIA RTX 2080 Ti GPU.

4. Experiments

4.1. Experimental Settings

Datasets We use the dataset constructed by Kalantari *et al.* [18] for both training and testing. This dataset consists of 74 scenes for training and 15 scenes for testing, each of which

Table 2: Quantitative comparisons of our method with state-of-the-art methods. † indicates that the values are taken from their original papers. O.F. and Homo. refer to the optical flow-based alignment and the homography transformation, respectively.

Methods	Pre-alignment		Boundary	PSNR_T	SSIM_T	PSNR_L	SSIM_L	HDR-VDP-2
	O.F.	Homo.	Cropping					
Sen [42]				41.11	0.9815	38.82	0.9749	57.43
Hu [15]				34.87	0.9698	31.72	0.9511	55.20
AHDRNet [47]				42.22	0.9904	41.26	0.9862	61.54
NHDRRNet† [49]				42.41	0.9887	-	-	-
Prabhakar† [36]	✓			42.82	-	41.33	-	-
HDR-GAN† [34]				43.92	0.9905	41.57	0.9865	-
Ours				**44.48**	**0.9917**	**42.45**	**0.9880**	**61.76**
Kalantari [18]	✓		✓	42.83	0.9877	41.49	0.9858	59.82
Ours			✓	**43.42**	**0.9892**	**41.68**	**0.9866**	**61.81**
Wu [45]		✓	✓	42.49	0.9889	42.06	0.9870	61.30
Prabhakar [38]	✓	✓	✓	41.95	0.9873	41.82	**0.9879**	61.23
Ours		✓	✓	**43.11**	**0.9901**	**42.37**	**0.9879**	**61.70**

contains three dynamic LDR images with different exposures. We also conduct qualitative evaluations on Sen *et al.*'s [42] dataset. Both datasets contain LDR images which have large-scale motions and severe saturation.

Evaluation metrics We use five evaluation metrics for the quantitative evaluation. We compute the PSNR and SSIM values between the predicted and ground truth HDR images after tonemapping (PSNR_T and SSIM_T) and in the linear domain (PSNR_L and SSIM_L). We also calculate the HDR-VDP-2 score [29] to measure the visual quality of HDR images.

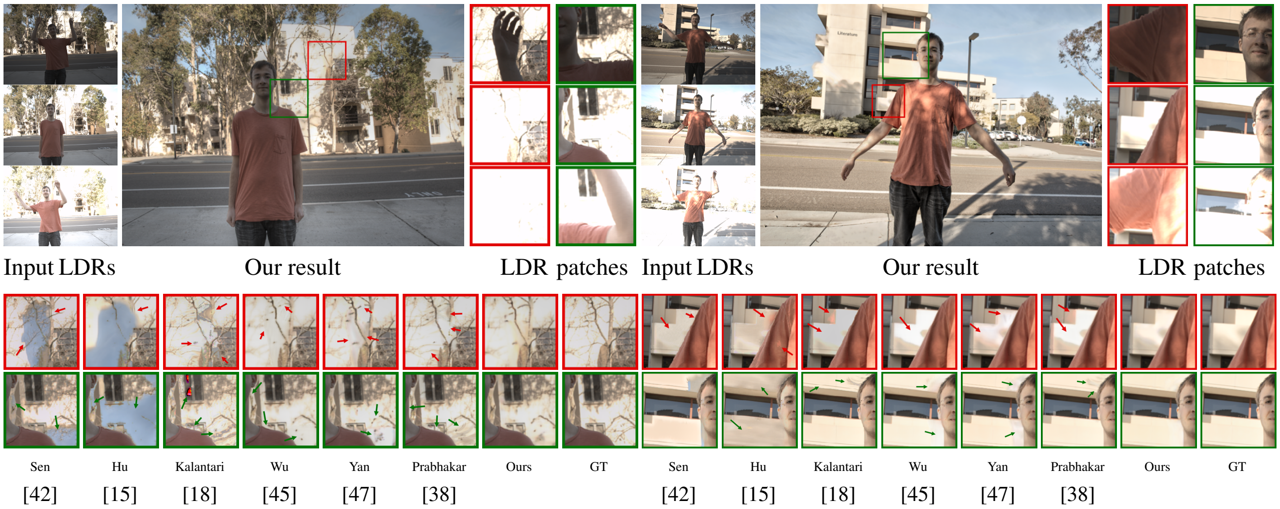
4.2. Comparisons

We compare our results with previous state-of-the-art methods, including two patch-based methods [42, 15] and seven CNN-based approaches [18, 45, 47, 49, 36, 34, 38]. Note that Kalantari *et al.* [18] and Prabhakar *et al.* [36] first align the input images using optical flow and Wu *et al.* [45] apply homography transformation. Also, Prabhakar *et al.* [38] use both of them for pre-alignment. We used the official codes if they are provided. Otherwise, we re-implemented their methods according to their papers except three methods [49, 36, 34]. We used the quantitative results reported in their papers, since we could not reproduce their results due to absence of the necessary data [36] or insufficient implementation details [49, 34]. HDR-VDP-2 score is not taken because it changes depending on the evaluation setting, which is not specified in their papers.

Quantitative evaluations We compute five metrics mentioned above in Table 2 for the quantitative evaluations. Note that Kalantari *et al.*'s method [18] needs to crop 6 pixels near image boundary, thus we compare with this method separately after boundary cropping to evaluate on the same input. Also, the methods of Wu *et al.* [45] and Prabhakar *et al.* [38] lose boundary content irregularly due to homography transformation, thus we apply homography transformation and pass the cropped images to our framework for fair comparisons with these methods. But, we do not use homography transformation during training. The evaluations are conducted on full images without losing image boundary for our method and the other six methods [42, 15, 47, 49, 36, 34]. It can be seen that the proposed method achieves the best performance in terms of all metrics, which validates that the results produced by our method are visually pleasing both in the linear domain and after tonemapping. Note that our method is an end-to-end framework which does not require any pre-alignment process such as homography transformation and optical flow algorithm.

Qualitative evaluations We compare our visual results with state-of-the-art methods. Fig. 4 shows a challenging case where all the input images are over-exposed and large-scale motion exists. While the other methods fail to reconstruct detailed texture, our method successfully hallucinates details in the severely saturated regions. Our method also generates realistic details even when the given information is insufficient due to occlusions, while the other methods introduce ghosting artifacts or color distortions within the saturated areas.

We also demonstrate our generalization ability by evaluating on Sen *et al.* 's dataset [42]. Fig. 7 shows the results on the case where the input images contain large saturated areas and especially the reference image provides little information. It can be shown that the proposed method successfully hallucinates details and texture in the saturated areas. More visual results are presented in the supplementary material.



(a) (b)

Figure 4: Qualitative comparisons of our method with state-of-the-art methods.

Table 3: Comparisons of the proposed brightness adjustment strategy with the motion compensation scheme.

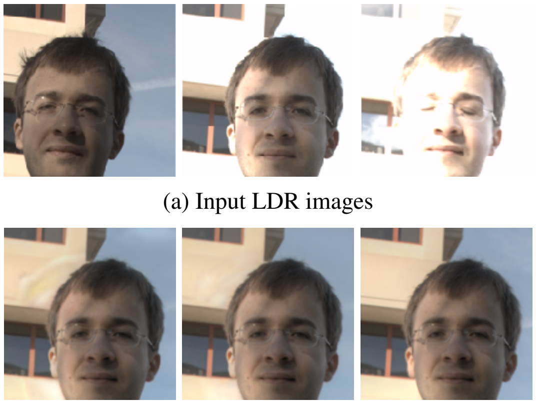
Methods	PSNR$_T$	SSIM$_T$	PSNR$_L$	SSIM$_L$
Motion Compensation	44.23	0.9913	42.21	0.9878
Brightness Adjustment	**44.48**	**0.9917**	**42.45**	**0.9880**

4.3. Ablation Study

In this section, we evaluate the contributions of the proposed components.

Brightness adjustment The proposed approach generates well-aligned multi-exposure features by reformulating a motion alignment problem into a simple brightness adjustment problem. We demonstrate the effectiveness of this reformulation in Table 3. We compare the proposed brightness adjustment method with the motion compensation method which uses the same architecture as the proposed one but aligns the supporting image to have the same structure as the reference as most previous works do. The motion compensation method needs to align the images with large motions and thus are prone to artifacts. Fig. 5 shows that the motion compensation scheme fails to handle the ghosting artifacts when the foreground motion exists in the over-exposed areas. It can be seen that the proposed reformulation strategy greatly eases the task and results in favorable performance.

Coarse-to-fine MAHN Our MAHN first reconstructs a coarse HDR image and then hallucinates content in saturated regions in the fine network. We validate the effectiveness of our coarse-to-fine reconstruction strategy and contribution of each sub-network in Table 4. The coarse-to-fine architecture shows even better performances than the one-stage coarse network, even only with the hallucination branch. The proposed architecture including both the refinement branch and the hallucination branch produces the



(a) Input LDR images

(b) M.C. (c) B.A. (d) GT

Figure 5: Effectiveness of the proposed brightness adjustment network. M.C. and B.A. denote motion compensation and brightness adjustment, respectively.

Table 4: Analysis on the MAHN architecture. Hall. and Refine. denote the hallucination branch and the refinement branch, respectively.

Methods			PSNR$_T$	SSIM$_T$	PSNR$_L$	SSIM$_L$
Coarse	Fine					
	Hall.	Refine.				
✓			41.58	0.9871	38.30	0.9786
✓	✓		44.17	0.9911	41.96	0.9869
✓	✓	✓	**44.48**	**0.9917**	**42.45**	**0.9880**

Table 5: Analysis on the soft adaptive contextual attention. a adjusts the softness of the saturation mask M. The proposed method adopts $a = 3$.

Methods	a	PSNR$_T$	SSIM$_T$	PSNR$_L$	SSIM$_L$
hard attention		43.47	0.9905	41.70	0.9863
soft attention	0.5	43.61	0.9909	41.68	0.9868
	1	**44.52**	0.9915	42.34	0.9870
	3	44.48	**0.9917**	**42.45**	**0.9880**
	5	43.45	0.9906	41.32	0.9861

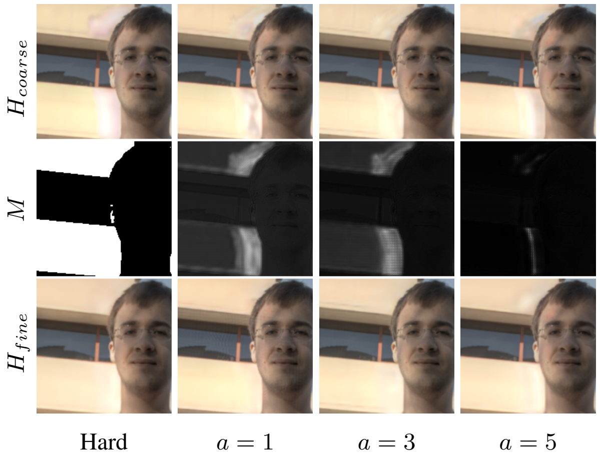


Figure 6: Visual comparisons for different choices of the softness parameter a of the saturation mask M. The proposed method adopts $a = 3$.

best results. It can be observed that the proposed coarse-to-fine architecture is effective and the hallucination branch greatly contributes to the HDR reconstruction process by performing explicit restoration for the saturated parts.

Adaptive contextual attention We compensate for saturation using the soft adaptive contextual attention in the MAHN. To generate the saturation mask M representing the saturation level with values in range $[0, 1]$, we use a sigmoid function: $sigmoid(x) = 1/(1 + e^{-ax})$, where a is a parameter that controls the steepness. We can adjust the softness of the saturation mask M by changing the parameter a. As a increases, the resulting mask becomes closer to a hard (binary) mask as shown in Fig. 6. When a is small, the mask becomes too soft to clearly identify the saturated regions which need to be restored. On the other hand, when a is too big, the value should be almost 0 or 1, thus it is difficult to reconstruct smooth and natural images. Table 5 shows the results with different softness parameter a. The best performances are achieved with the modest softness $a = 3$. We also compare the proposed soft adaptive attention with hard attention which is similar to the original contextual attention. The hard attention method generates the saturation mask by thresholding the coarse HDR image H_{coarse} with a threshold of τ: $M(x, y) = \mathbb{1}[H_{coarse}(x, y) \geq \tau]$, where (x, y) denotes a pixel location and $\mathbb{1}$ is an indicator function. τ is set as 0.9 empirically. We observe that the soft adaptive contextual attention enables sample-specific mask generation and hallucination, which leads to detailed HDR image generation.

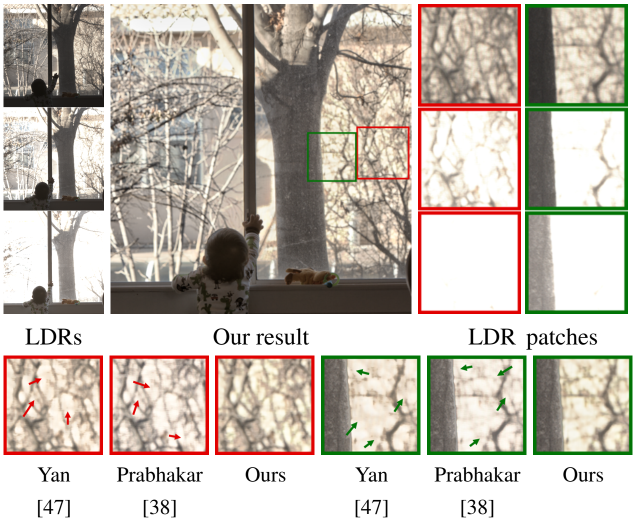


Figure 7: Qualitative comparisons on the image from Sen *et al.*'s [42] dataset.

5. Conclusion

We have proposed an end-to-end HDR imaging CNN, which takes multi-exposure inputs with dynamic motions and generates ghost-free HDR images with some hallucinations in washed-out regions. For this, we have introduced the BAN which adjusts the brightness of the reference feature using adaptive convolutions so that the well-aligned multi-exposure features are generated. Then, the bracketed features are integrated into a clean HDR image with supervision on the saturated regions. We have also proposed the MAHN, which reconstructs details in the saturated areas by aggregating valid content from the unsaturated regions. Experiments show that the proposed system delivers high-quality HDR results even in the presence of severe saturation and large displacement. Our code is available at `https://github.com/haesoochung/hdri-saturation-compensation`.

Acknowledgments

This work was supported in part by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (2021R1A2C2007220), and in part by Samsung Electronics Co., Ltd.

References

[1] Ahmet Oğuz Akyüz, Roland Fleming, Bernhard E Riecke, Erik Reinhard, and Heinrich H Bülthoff. Do hdr displays support ldr content? a psychophysical evaluation. *ACM Transactions on Graphics (TOG)*, 26(3):38–es, 2007.

[2] Jaehyun An, Seong Jong Ha, and Nam Ik Cho. Reduction of ghost effect in exposure fusion by detecting the ghost pixels in saturated and non-saturated regions. In *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1101–1104. IEEE, 2012.

[3] Jaehyun An, Seong Jong Ha, and Nam Ik Cho. Probabilistic motion pixel detection for the reduction of ghost artifacts in high dynamic range images from multiple exposures. *EURASIP Journal on Image and Video Processing*, 2014(1):1–15, 2014.

[4] Jaehyun An, Sang Heon Lee, Jung Gap Kuk, and Nam Ik Cho. A multi-exposure image fusion algorithm without ghost effect. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1565–1568. IEEE, 2011.

[5] 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.

[6] Francesco Banterle, Patrick Ledda, Kurt Debattista, Alan Chalmers, and Marina Bloj. A framework for inverse tone mapping. *The Visual Computer*, 23(7):467–478, 2007.

[7] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018.

[8] 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.

[9] 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*, pages 369–378. ACM Press/Addison-Wesley Publishing Co., 1997.

[10] Piotr Didyk, Rafal Mantiuk, Matthias Hein, and Hans-Peter Seidel. Enhancement of bright video features for hdr displays. In *Computer Graphics Forum*, volume 27, pages 1265–1274. Wiley Online Library, 2008.

[11] 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):178, 2017.

[12] Yuki Endo, Yoshihiro Kanamori, and Jun Mitani. Deep reverse tone mapping. *ACM Transactions on Graphics (Proc. of SIGGRAPH ASIA 2017)*, 36(6), Nov. 2017.

[13] Thorsten Grosch et al. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling and Visualization, RWTH Aachen*, 277284, 2006.

[14] 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.

[15] 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.

[16] Yongqing Huo, Fan Yang, Le Dong, and Vincent Brost. Physiological inverse tone mapping based on retina response. *The Visual Computer*, 30(5):507–517, 2014.

[17] Xu Jia, Bert De Brabandere, Tinne Tuytelaars, and Luc Van Gool. Dynamic filter networks. In *NIPS*, 2016.

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

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

[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.

[21] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. Spatio-temporal transformer network for video restoration. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 106–122, 2018.

[22] Rafael Pacheco Kovaleski and Manuel M Oliveira. High-quality brightness enhancement functions for real-time reverse tone mapping. *The Visual Computer*, 25(5):539–547, 2009.

[23] Rafael P Kovaleski and Manuel M Oliveira. High-quality reverse tone mapping for a wide range of exposures. In *2014 27th SIBGRAPI Conference on Graphics, Patterns and Images*, pages 49–56. IEEE, 2014.

[24] Hayden Landis. Production-ready global illumination. In *Siggraph 2002*, volume 5, pages 93–95, 2002.

[25] 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.

[26] 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.

[27] 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.

[28] S Mann and R Picard. Beingundigital'with digital cameras. *MIT Media Lab Perceptual*, 1:2, 1994.

[29] Rafat 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 Transactions on graphics (TOG)*, volume 30, page 40. ACM, 2011.

[30] Demetris Marnerides, Thomas Bashford-Rogers, Jonathan Hatchett, and Kurt Debattista. Expandnet: A deep convolutional neural network for high dynamic range expansion from low dynamic range content. In *Computer Graphics Forum*, volume 37, pages 37–49. Wiley Online Library, 2018.

[31] Laurence Meylan, Scott Daly, and Sabine Süsstrunk. The reproduction of specular highlights on high dynamic range displays. In *Color and Imaging Conference*, volume 2006, pages 333–338. Society for Imaging Science and Technology, 2006.

[32] Ben Mildenhall, Jonathan T Barron, Jiawen Chen, Dillon Sharlet, Ren Ng, and Robert Carroll. Burst denoising with kernel prediction networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2502–2510, 2018.

[33] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017.

[34] 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.

[35] 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, 2015.

[36] K Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R Venkatesh Babu. Towards practical and efficient high-resolution hdr deghosting with cnn.

[37] K Ram Prabhakar and R Venkatesh Babu. Ghosting-free multi-exposure image fusion in gradient domain. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1766–1770. IEEE, 2016.

[38] K. Ram Prabhakar, Gowtham Senthil, Susmit Agrawal, R. Venkatesh Babu, and Rama Krishna Sai S Gorthi. Labeled from unlabeled: Exploiting unlabeled data for few-shot deep hdr deghosting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4875–4885, June 2021.

[39] Zhiyuan Pu, Peiyao Guo, M Salman Asif, and Zhan Ma. Robust high dynamic range (hdr) imaging with complex motion and parallax. In *Proceedings of the Asian Conference on Computer Vision*, 2020.

[40] Shanmuganathan Raman and Subhasis Chaudhuri. Reconstruction of high contrast images for dynamic scenes. *The Visual Computer*, 27(12):1099–1114, 2011.

[41] 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.

[42] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based hdr reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203–1, 2012.

[43] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020.

[44] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[45] 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.

[46] Xiaoyu Xiang, Yapeng Tian, Yulun Zhang, Yun Fu, Jan P Allebach, and Chenliang Xu. Zooming slow-mo: Fast and accurate one-stage space-time video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3370–3379, 2020.

[47] 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. *arXiv preprint arXiv:1904.10293*, 2019.

[48] 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.

[49] 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.

[50] Xin Yang, Ke Xu, Yibing Song, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Image correction via deep reciprocating hdr transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[51] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters*, 27:1500–1504, 2020.

[52] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018.

[53] Jinsong Zhang and Jean-François Lalonde. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4519–4528, 2017.

[54] Wei Zhang and Wai-Kuen Cham. Gradient-directed multiexposure composition. *IEEE Transactions on Image Processing*, 21(4):2318–2323, 2012.

[55] Yue Zhao, Yuanjun Xiong, and Dahua Lin. Trajectory convolution for action recognition. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 2208–2219, 2018.

[56] 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.

[57] 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.