Towards Practical and Efficient High-Resolution HDR Deghosting with CNN

K. Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh,
Balraj Ashwath, and R. Venkatesh Babu

Video Analytics Lab, Indian Institute of Science, Bangalore, INDIA

Abstract. Generating High Dynamic Range (HDR) image in the presence of camera and object motion is a tedious task. If uncorrected, these motions will manifest as ghosting artifacts in the fused HDR image. On one end of the spectrum, there exist methods that generate high-quality results that are computationally demanding and too slow. On the other end, there are few faster methods that produce unsatisfactory results. With ever increasing sensor/display resolution, currently we are very much in need of faster methods that produce high-quality images. In this paper, we present a deep neural network based approach to generate high-quality ghost-free HDR for high-resolution images. Our proposed method is fast and fuses a sequence of three high-resolution images (16-megapixel resolution) in about 10 seconds. Through experiments and ablations, on different publicly available datasets, we show that the proposed method achieves state-of-the-art performance in terms of accuracy and speed.

1 Introduction

Natural scenes have a wide range of illumination that exceeds the dynamic range of standard digital camera sensors. The resulting image has undesirable saturated regions (too bright or too dark) while capturing an HDR scene. Widely followed software solution is to merge multiple low dynamic range (LDR) images into a single HDR image. Each LDR image in the input stack captures part of the brightness spectrum by varying exposure time (or aperture, ISO). Then, the HDR image is generated by combining the best regions of each LDR image. The generated HDR has a wider dynamic range than each LDR input image.

HDR fusion is a simple and straightforward process for static scenes: scenes without any camera or object motion [27]. However, in the presence of camera or object motion, such naive static fusion techniques result in artifacts. While the camera motion can be corrected using homography-based alignment procedures, the harder challenge is to address the object motion. If uncorrected, the moving objects from all the input images appear mildly in the final result, resulting in a ghost-like perception, hence known as *ghosting artifact*. Several methods have been proposed in the literature to generate results without ghosting artifacts, a.k.a., HDR deghosting methods. The initially proposed rejection-based methods [4, 5, 8, 9, 11, 23, 24, 26, 28, 41, 45] are fast and easy to implement. Despite

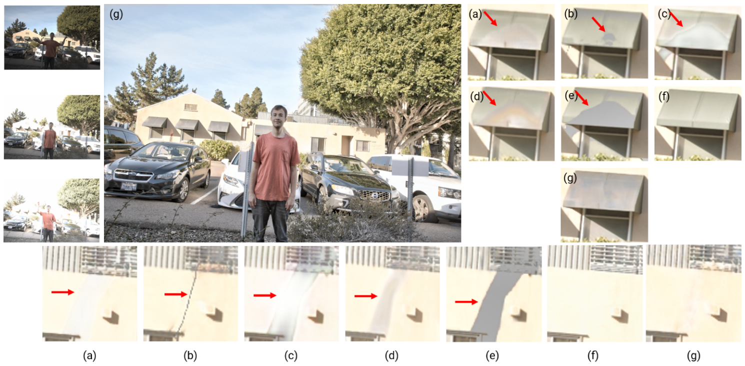


Fig. 1. The input exposure sequence with motion is shown on the leftmost column. The result by our proposed method is shown in the second column. The zoomed regions of different methods are highlighted in (a) to (g). (a) Wu18 [40], (b) Kalantari17 [15], (c) SCHDR [25], (d) AHDR [42], (e) guide image generated by N_g, (f) ground truth and (g) proposed method. Best viewed in color monitor.

their better performance for mostly static scenes, they suffer from having only LDR content for moving objects. Alignment-based methods register input images to a selected reference image using rigid [36, 39] or non-rigid [6, 12, 16, 46] registration techniques. While the rigid registration techniques find complex object motion difficult to handle, the non-rigid techniques (such as optical flow) are inaccurate for deformable motions and occluded pixels. On the other hand, patch-based optimization methods [13,34] synthesize a static sequence from the dynamic input sequence. Despite their high-quality results, they suffer from huge computational complexity, making them not suitable for portable devices with limited computational resources.

Recently proposed deep learning-based HDR deghosting methods [15,40,42–44] generate visually pleasing results for the majority of the contents. However, they still suffer from artifacts in heavily saturated regions (see Figure 1). Another limitation of the existing CNN-based methods is the need for more computational resources to process high-resolution images (more than 5 megapixels). For example, Yan *et al.*'s [43] approach can process a maximum of 2 megapixels (MP) on a GPU with 11GB memory (RTX 2080Ti); their method requires GPU with more memory to process high-resolution images. Similarly, [40] can only process images up to 5 MP on a GPU with 11GB memory.

To address the issues mentioned above, we propose a robust CNN based HDR deghosting method that can generate artifact-free results and can process up to 16 MP images within 11 seconds on a GPU with 11GB memory. Unlike previous CNN based methods [15,40,42,43], in our approach, we avoid processing full resolution images with CNN. Alternatively, we process low-resolution images and upscale the result to the original full resolution. However, a simple bilinear or bicubic upsampling may introduce blur artifacts. Hence, we make use of bilateral guided upsampling [2,7] to generate full resolution artifact-free output. The main contributions of our work is summarized as follows,

- We present a CNN based HDR deghosting algorithm that is robust to significant object motion and saturation.
- We propose an efficient HDR deghosting approach to process high-resolution images with Bilateral Guided Upsampler (BGU). We also demonstrate the use of BGU for a fusion task where the guide image is not readily available.
- We provide motion segmentation masks for UCSD dataset [15] to benefit HDR deghosting research community.
- We perform an extensive quantitative and qualitative evaluation on publicly available datasets. We also perform various ablation experiments on different choices in our model.

2 Related works

In general, we can classify most of the deghosting algorithms into four categories: alignment-based, rejection-based, patch-based, and learning-based methods[1].
Alignment-based methods: The first class of algorithms register all input images to a chosen reference image using rigid or non-rigid registration techniques. Rigid registration methods handle global camera motion by matching feature descriptors such as SIFT [36], SURF [8] and Median Threshold Bitmap [39]. In [36], Tomaszewska and Mantuik perform RANSAC after SIFT to refine the matches. Non-rigid registration methods like Bogoni *et al.* [1], Kang *et al.* [16], Gallo *et al.* [6] and Zimmer *et al.* [46] make use of optical flow to align images. The non-rigid registration based methods have the advantage of producing results with moving HDR content. However, as some of the features used for rigid and non-rigid registration are not robust against huge brightness changes, they are more prone to fail for saturated regions or occluded regions.

Rejection-based methods: The second class of algorithms assume the input images to be fairly static and detects the pixels affected by motion [4,18,29], after which they combine only static images in those regions. To identify the moving objects in the registered images, many techniques have been proposed that make use of illumination constancy criteria, linear relationship between images [17,35,41], prediction and thresholding [9,23,29], thresholding background probability map [18], etc. Gallo *et al.* [5] perform patch wise comparison in logarithmic domain to identify moving regions. Raman and Chaudhuri [28] perform a comparison in super-pixels to improve motion segmentation accuracy along edges. Although this class of algorithms is fast, a major drawback is that they generate low dynamic range content in the moving object regions.

Patch-based optimization methods: The third class of algorithms handles both camera and object motion jointly by performing patch-based registration between the varying exposure images [13,21,34]. They pick one of the input images as the structural reference and find dense correspondences between the reference image and all other input source images in the stack. Using these correspondences, they synthesize modified source images that are structurally similar

[1] Elaborate literature review can be found in [32,33,37].

to the original reference image but resemble the corresponding source images in terms of the brightness. The synthesized images are fused using static fusion techniques like Debevec and Malik [3]. Despite the impressive performance, the patch-based methods [13, 34] suffer from high computation time.

Deep learning methods: The fourth class of algorithms uses deep learning methods to fuse input images into an HDR image. In [15], a simple CNN network is trained to correct artifacts introduced by the optical flow method. Though [15] generates visually pleasing results for some images, it fails to correct warping errors in saturated regions, especially on moving regions. Wu *et al.* [40] treat HDR deghosting as an image translation problem. In their method, authors have shown that it is possible to fuse images without performing optical flow. In [43], Yan *et al.* use a multiscale CNN to extract features at different scales and fuse them. More recently, Yan *et al.* proposed an attention-guided method in [42]. In this approach, the authors identify regions with motion and exclude them during the fusion process. Each of the above-discussed methods addresses an important issue. However, existing CNN based methods require heavy computational resources to process high resolution images.

3 Proposed method

Motivation for using BGU: Existing CNN based HDR deghosting methods require a substantial amount of computational resources to process high-resolution images. The bottleneck is to process the images in their full resolution with CNNs. Hence these methods are not easily scalable to ever increasing image resolution. To circumvent this issue, we propose a method that performs all operations in low resolution and upscales the result to the required full resolution. For efficient and artifact-free upscaling, we utilize deep Bilateral Guided Upsampling (BGU) used by Gharbi *et al.* [7]. For upscaling, the BGU algorithm requires a guide image for structural reference (see Figure 2). Such a guide image should satisfy the following two criteria: (1) must be in full high-resolution as the expected final HDR result, (2) should have similar structural details as the expected final HDR result. In other words, the guide image should be a high-resolution fused image without any ghosting artifacts. However, predicting the high-resolution guide image through a CNN will have the same challenges as discussed before.

To address this challenge, we propose to generate a guide image through weight-map based strategy. The core idea is to predict weight maps at low resolution for downsampled inputs and upscale them using bicubic interpolation. The upscaled weight maps can be used to combine original high-resolution input images and generate guide image. As the upscaling is performed on the weight maps instead of the image domain, the predicted guide image will be void of any blurring artifacts. In this step, CNN processes only the low-resolution images, but the generated guide image is in full high-resolution. Finally, the generated guide image is used by BGU to produce the deghosted HDR result.

Pipeline: The objective of our HDR deghosting method is to fuse input varying exposure LDR input images ($\{I_1, \cdots, I_N\}$) into an HDR image (H_f)

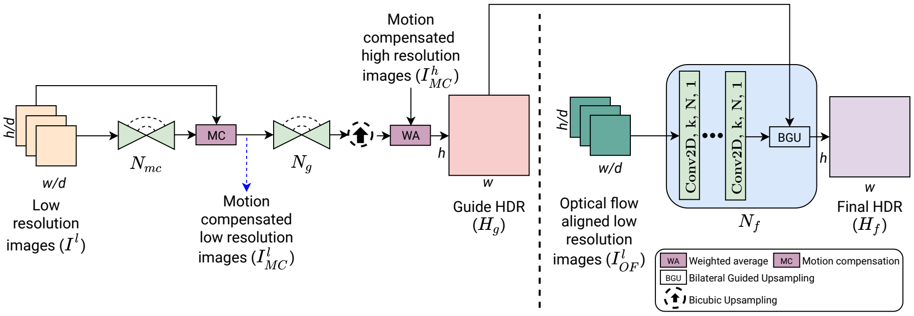

h/d
w/d
Low resolution images (I^l)
N_{mc}
MC
Motion compensated low resolution images (I^l_{MC})
N_g
Motion compensated high resolution images (I^h_{MC})
WA
h
w
Guide HDR (H_g)
h/d
w/d
Optical flow aligned low resolution images (I^l_{OF})
Conv2D, k, N, 1
Conv2D, k, N, 1
BGU
N_f
h
w
Final HDR (H_f)
WA Weighted average
MC Motion compensation
BGU Bilateral Guided Upsampling
Bicubic Upsampling

Fig. 2. Our method comprises of two components: guide and final HDR prediction. The input images in full-resolution (I^h) is downsampled by a factor of d to generate I^l. The optical flow aligned low-resolution images, I^l_{OF}, are fused by N_f model to generate final HDR result (H_f) in full-resolution using Bilateral Guided Upsampler (BGU). For efficient artifact-free upsampling, BGU requires a guide image (H_g) for structual guidance. H_g is generated with N_{mc} and N_g with I^l as input. For more details, see Section 3.

without any ghosting artifacts. Our approach is a reference-based HDR deghosting method, meaning, the generated HDR image will have identical structural details as that of the chosen reference image. Similar to [15,40,42,43], our method takes three different exposure images (I_-, I_0, I_+) as input with the middle image (I_0) as the reference, where (I_-, I_0, I_+) denotes the underexposed, normally exposed, and overexposed images.

We denote the full high-resolution input as $I^h = \{I^h_-, I^h_0$, and $I^h_+\}$ and the corresponding downsampled (by a factor d) images as $I^l = \{I^l_-$, I^l_0, and $I^l_+\}$ (see Figure 2)[2]. As I^l may have camera and object motion, we align them using optical flow method by Liu *et al.* [20]. We align the non-reference images (I^l_- and I^l_+) to reference image (I^l_0) after exposure correction. The generated optical flow aligned sequence (I^l_{OF}) is fed to fusion (N_f) network, to generate a fused feature map at low resolution. Then, the fused feature map is passed to BGU to generate H_f in high-resolution. However, BGU requires a guide image for structural reference.

The guide image is generated in two steps: motion compensation followed by weight map based fusion. In the first step, I^l is passed to a CNN network (N_{mc}) to segment moving regions between reference and other non-reference images. The predicted segmentation maps are used to generate motion compensated sequence, I^l_{MC} in low-resolution. The corresponding high-resolution motion maps are obtained by upscaling the low-resolution segmentation maps using bicubic interpolation. Then, the upsampled motion maps are used to compensate motion in I^h and generate high-resolution motion compensated sequence, I^h_{MC}. Naively fusing I^h_{MC} using classical fusion methods like triangle function [3] may have

[2] Throughout the rest of the paper, the full high-resolution images are denoted by superscript h, i.e., $\square^h$ and the downsampled low-resolution images by $\square^l$.

artifacts in the guide image. Hence, we correct such artifacts using a trainable CNN, N_g. With I^l_{MC} as input, N_g generates three weight maps, one for each of the three input images. The weighted sum of upscaled weight maps (using bicubic interpolation) and I^h_{MC} results in the guide image (H_g) at high-resolution. It should be noted that, throughout this process, both inputs and predictions of N_{mc} and N_g are in low-resolution.

Finally, H_f is generated by BGU with H_g as structural guidance. We provide elaborate details of each step in further sections.

Guide Image Generation: The input images (I^l) to N_{mc} are assumed to be aligned for camera motion. If not, the low resolution (I^l) images can be aligned with simple homography based methods. The aligned images are passed to a CNN model, N_{mc}, to segment moving regions between reference image (I^l_0) and non-reference images (I^l_-, I^l_+) individually. To do so, we concatenate reference and a non-reference image, I^l_- (or I^l_+) together to form a six channel data and feed as input to N_{mc}. We use the U-net architecture [30] for N_{mc} model[3]. The model architecture consists of four encoders followed by four decoders with skip connections from encoders. The output of N_{mc} is a single channel binary segmentation map, M^l_- (M^l_+ for I^l_+ input) with same resolution as I^l. Each pixel in M^l_- has value between 0 and 1. Value '0' indicates the absence of motion at that pixel in either reference or the non-reference image, and '1' indicating the presence of motion in any one of them. Using the predicted motion segmentation maps, we perform motion compensation in the input sequence (MC box shown in Figure 2). The motion compensation is achieved by replacing the moving regions in non-reference image with regions from reference image at the same location.

$$I^l_{mc,-} = M^l_- \times \mathcal{E}(I^l_0, t_0, t_-) + (1 - M^l_-) \times I^l_- \tag{1}$$

$$I^l_{mc,+} = M^l_+ \times \mathcal{E}(I^l_0, t_0, t_+) + (1 - M^l_+) \times I^l_+ \tag{2}$$

In the above equation, $\mathcal{E}()$ denotes exposure normalization function. For example, $\mathcal{E}()$ modifies the exposure of image A with exposure time t_A to exposure of image B with exposure time t_B by:

$$\mathcal{E}(A, t_A, t_B) = \left(A^\gamma \times \frac{t_B}{t_A} \right)^{\frac{1}{\gamma}} \tag{3}$$

where, γ=2.2. At the end of this process, we generate motion compensated sequence $I^l_{MC} = \{I^l_{mc,-}, I^l_0, I^l_{mc,+}\}$. To generate motion compensated sequence in high-resolution, the predicted segmentation maps, M^l_- and M^l_+, are upscaled using bicubic interpolation to generate M^h_- and M^h_+. Then, we repeat the steps in equations 1 and 2, but with high resolution motion maps and high resolution input images to create $I^h_{MC} = \{I^h_{mc,-}, I^h_0, I^h_{mc,+}\}$.

The motion compensated low resolution sequence I^l_{MC} is fed as input to N_g network to generate fusion weight maps. The input to N_g is obtained by concatenating $\{I^l_{mc,-}, I^l_0, I^l_{mc,+}\}$ images in channel dimension,

$$(w^l_-, w^l_0, w^l_+) = N_g(I^l_{mc,-}, I^l_0, I^l_{mc,+}) \tag{4}$$

[3] More detailed model architecture is provided in supplementary material

The output of N_g consists of three weight maps: w^l_-, w^l_0, and w^l_+. Similar to N_{mc} model, we use U-net architecture for N_g network as well. Before using the weight maps for fusion, they are upscaled using bicubic interpolation to the required high resolution: w^h_-, w^h_0, and w^h_+. Then, the upscaled and normalized weight maps are used to combine motion corrected high-resolution images I^h_{MC} as,

$$H_g = w^h_- \times LH(I^h_{mc,-}) + w^h_0 \times LH(I^h_0) + w^h_+ \times LH(I^h_{mc,+}) \tag{5}$$

where, $LH()$ denotes the LDR to HDR conversion process. For a sample LDR image A, the corresponding HDR version is obtained by,

$$LH(A) = \frac{A^\gamma}{t_A} \tag{6}$$

N_f **network**: The input images (I^l) are aligned using optical flow to compensate for camera and object motion. Instead of applying the optical flow on high resolution images, we apply it on downsampled low-resolution images, I^l. We use optical flow model proposed by Liu *et al.* [20] to align the non-reference images to the reference image. The optical flow aligned images, $I^l_{OF} = (I^l_{OF,-}, I^l_0, I^l_{OF,+})$, are passed as input data to N_f model to predict final HDR (H_f) result,

$$H_f = N_f(I^l_{OF}) \tag{7}$$

As shown in Figure 3, three input images are passed to a small encoder with three convolution layers to extract individual image features. The extracted features of reference and a non-reference image ($I^l_{OF,-}$ or $I^l_{OF,+}$) is concatenated together and further processed by another convolutional layer. By doing so, the network has the ability to compare both feature maps and choose the properly exposed details from both. This operation is repeated for all three convolutional levels of the encoder to produce final feature maps for non-reference images. Then, the output of last layer in reference image encoder is concatenated with other two final feature maps of non-reference images. The concatenated feature maps are further processed by three blocks of Extended Stacked and Dilated Convolution (XSDC) blocks.

A higher receptive field is necessary for the network to gather information over a vast image region to faithfully reconstruct occluded and saturated regions. We achieve such a high receptive field by using XSDC blocks. XSDC is an simple extention to SDC block proposed by Schuster *et al.* [31]. The original SDC block has four dilated convolutions with dilation rate from 1 to 4, stacked parallely. The output of stacked convolutional layers is concatenated to produce output feature maps. We make small extention to SDC by increasing the number of consecutive convolutional layers from one (in SDC) to three and process the concatenated features with a 1×1 convolutional layer to aggregate the feature maps. In our implementation, we use three stacks of three convolutional layers (see Figure 3) with dilation rate of 1,2 and 4, and kernel size 3×3 for stacked convolutional layers. Through ablation experiments, we observed that our changes perform better than original SDC (see Table 4) by having better ability to transfer details in saturated regions. In addition, we also have dense connections among each of the three XSDC blocks. The output low-resolution feature map of final XSDC block is passed to the BGU, which takes high-resolution H_g generated from N_g model as the guide image to predict final fused HDR image, H_f in high-resolution.

Loss functions: The total loss to train the model consists of three sub-losses at output of N_{mc}, N_g and N_f. The predicted segmentation maps (M_-^h and M_+^h) by N_{mc} is compared with ground truth segmentation maps ($\widetilde{M}_-^h$ and $\widetilde{M}_+^h$) using Binary Cross Entropy (BCE). The motion compensation loss ($\mathcal{L}_{mc}$) is computed by taking the sum of two individual BCE losses,

$$\mathcal{L}_{mc} = BCE(M_-^h, \widetilde{M}_-^h) + BCE(M_+^h, \widetilde{M}_+^h) \tag{8}$$

The ℓ_2 loss between guide image (H_g) and ground truth HDR ($\widetilde{H}$), final image (H_f) and $\widetilde{H}$ is used to train N_g and N_f.

$$\mathcal{L}_g = \ell_2(T(H_g), T(\widetilde{H})) \tag{9}$$

$$\mathcal{L}_f = \ell_2(T(H_f), T(\widetilde{H})) \tag{10}$$

where, $T(\Box)$ denotes the tonemapping operation applied to HDR images with commonly used μ-law function,

$$T(H) = \frac{log(1 + \mu H)}{log(1 + \mu)} \tag{11}$$

where $\mu = 5000$. The total loss is taken as the weighted sum of three losses: $\mathcal{L}_{mc}$, $\mathcal{L}_g$ and $\mathcal{L}_f$.

$$\mathcal{L}_{total} = \alpha_1 \times \mathcal{L}_{mc} + \alpha_2 \times \mathcal{L}_g + \alpha_3 \times \mathcal{L}_f \tag{12}$$

where α_1, α_2 and α_3 denote the weight values assigned to individual loss functions. We found that assigning equal weight values from beginning takes longer time to stabilize the training and converge to optimum solution. Hence, we assign low value of $1e^{-4}$ for α_2 and α_3 for initial 25 epochs, thus making N_{mc} network learn better segmentation maps that would improve N_g. From 26^{th} epoch till 50, we set α_2 as 1 and for epochs after 75, all three weights are assigned to one.

4 Experiments

4.1 Implementation

Datasets: We trained our model on the dataset provided by Kalantari and Ramamoorthi [15]. The dataset consists of 74 training and 15 testing images of 1500 × 1000 resolution, with ground truth for evaluation. Additionally, we also tested our model on datasets provided by Prabhakar *et al.* [25], Sen *et al.* [34] and Tursun *et al.* [38].

Groundtruth for N_{mc}: For training N_{mc} model, we have generated ground

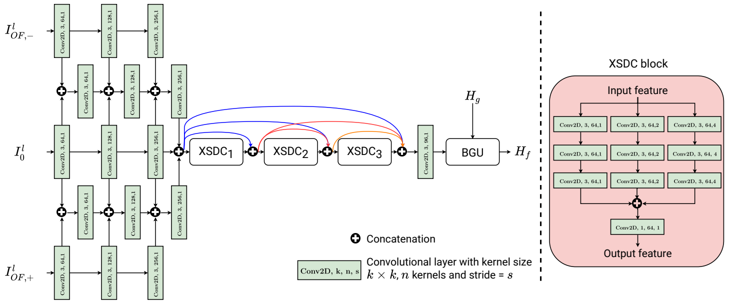


Fig. 3. Illustration of N_f architecture. The input to N_f is optical flow corrected low-resolution sequence $\{I^l_{OF}\}$ and H_g. The output is the final HDR image H_f in full-resolution.

truth by annotating the moving objects in each input image. For a sequence with three images: (I^h_-, I^h_0, I^h_+), we generated three segmentation maps, (S^h_-, S^h_0, S^h_+), by manually annotating only the moving regions (see Figure 4). The ground truth is obtained by taking the union of reference and non-reference image segmentation mask. For example, $\widetilde{M}^h_-$ is generated by performing a union of S^h_- and S^h_0. Similarly, $\widetilde{M}^h_+$ is generated by taking a union of S^h_+ and S^h_0. In total, we have annotated 89 images of UCSD dataset including train and test set.

Data pre-processing: To train our model, we extracted 50K patches of size 256 × 256 from the 74 training images and applied eight augmentations (rotation, flip, and noise) to increase the training sample size.

Network details: For both N_{mc} and N_g, we use U-net model with four convolutional layers for encoder. The features are downsampled using the max pool after each block before passing to the next convolution layer. The decoder part consists of the same number of layers as encoders with the upsampling layer to increase the feature resolution. We use Leaky ReLu in all the layers except the final layer of N_{mc} and N_g, where we use sigmoid activation to predict weight maps in [0-1] range. Adam optimizer [19] is used to train our model with a learning rate of $2e^{-4}$ for 200 epochs.

4.2 Quantitative evaluation

We compare our proposed method against seven state-of-the-art methods: 1. Hu13 [13], 2. Sen12 [34], 3. Kalantari17 [15], 4. Wu18 [40], 5. MSDN-HDR[4] [43], 6. AHDR [42], and 7. SCHDR [25]. In Table 1, we report the quantitative evaluation on two datasets: [15] and [25] using three full-reference image quality assessment metrics: PSNR computed in linear HDR domain (P_L), PSNR computed in tonemapped domain (P_T) after applying tonemapping (equation 11),

[4] HDR-VDP-2 metric and scores for [25] dataset is not reported for Yan20 [44] and MSDN [43] as the codes are not publicly available. The numbers reported are taken from their paper.

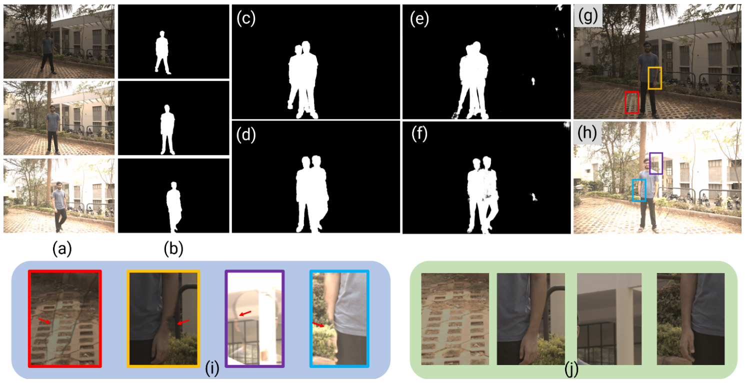


Fig. 4. (a) Input sequence, (b) ground truth motion masks (S^h_-, S^h_0, S^h_+), (c) $\widetilde{M}^h_-$, obtained by combining S^h_- and S^h_0, (d) $\widetilde{M}^h_+$, obtained by combining S^h_+ and S^h_0, (e) predicted motion map, M^h_-, for inputs I^l_- and I^l_0, (f) predicted motion map M^h_+ for inputs I^l_+ and I^l_0, (g) motion corrected image $I^h_{mc,-}$, (h) motion corrected image $I^h_{mc,+}$, (i) zoomed regions in motion corrected images highlighting artifacts, and (j) corresponding regions at guide image (H_g) after refining with N_g.

and HDR-VDP-2 [22]. HDR-VDP-2 is a full-reference metric specifically designed for quantifying degradations in HDR images.

From results in Table 1, we observe that our proposed method with BGU (denoted as *Ours (w/ BGU)*) for downsampling factor of 8 performs better than almost all methods in both datasets. Additionally, we present results for another variation of our model without using BGU (*Ours (w/o BGU)*). By setting d=1, means passing the input without downsampling, in full resolution to the network. Since the output of the XSDC block is already in full resolution, using BGU is redundant. Thus, the network is modified to predict three-channel RGB output at the end of the last XSDC module, excluding BGU. From the results reported in Table 1, we see that the network trained with d=1 shows better performance than all methods, in all three metrics for both [15, 25] datasets.

4.3 Ablation experiments

Guide-image ablations: In Table 2, we present results for different simple baseline approaches to generate guide image. We report the PSNR-L (P_L) and PSNR-T (P_T) on test set of UCSD dataset.

[1] B1+ [3]: A simple baseline to generate guide image would be to detect moving pixels by taking difference between images. We detect the moving pixels by thresholding (with 0.1) the image difference after brightness normalization. The threshold maps are used as M^h_- and M^h_+ in equation 1 and 2 to generate motion compensated sequence. Then, the motion corrected HDR sequence is fused by using triangle weighing strategy [3].

[2] N_{mc}+ [**3**]: Motion correction by N_{mc} followed by fusion with [3].

[3] N_g: Direct input without motion correction is passed to N_g model.

[4] B1+N_g: Motion compensation by simple difference and fusion with N_g model.

Table 1. Quantitative comparison of our proposed method against eight state-of-art HDR deghosting algorithms on [15] and [25] datasets. For [15] dataset, we report the scores averaged over all 15 test sequences, and for [25], over 116 test sequences with PSNR-L (P_L) and PSNR-T (P_T) metrics. The best performing method is highlighted in **bold** and the second best in blue color. P_T (in dB) is computed after tonemapping with equation 11.

Methods\ Datasets	Metrics	Sen12 [34]	Hu13 [13]	Kalantari17 [15]	Wu18 [40]	Yan20 [44]	MSDN [43]	AHDR [42]	SCHDR [25]	Ours (w/o BGU)	Ours (w/ BGU)
[15]	P_L	38.57	30.83	41.07	40.91	-	41.01	41.01	39.68	**41.68**	41.33
	P_T	40.94	32.18	42.74	41.65	42.41	42.22	42.10	40.47	**43.08**	42.82
	HDR-VDP-2	61.98	60.12	66.64	**67.96**	-	-	66.67	66.80	67.21	66.94
[25]	P_L	29.57	28.87	32.08	30.72	-	-	31.83	31.44	**32.52**	32.11
	P_T	32.09	30.82	35.34	31.31	-	-	33.72	30.57	**35.84**	35.46
	HDR-VDP-2	62.43	60.47	64.47	64.03	-	-	64.32	62.20	**64.76**	64.57

[5] N_{mc}+N_g: Motion compensation by N_{mc} and fusion with N_g model. This denotes the accuracy of H_g images.

From Table 2, we observe that while the individual performance of N_{mc} and N_g is low, using them together can result in significant boost in PSNR. Similarly, using N_{mc} for motion correction achieves a P_T score approximately 1.5 dB higher than B1. This is due to the fact that simple difference can produce false segmentation in heavily saturated regions. Whereas, N_{mc} network has learnt reliable features that can produce accurate motion segmentation maps.

Choice of inputs: In Table 3, we present the accuracy obtained with different type of inputs to the fusion model (N_f). We report scores for the d=1 setting.

[1] N_f: We pass direct input images without optical flow correction to N_f model.
[2] N_{mc}+N_f: We pass motion corrected sequence (I^l_{MC}) as input to N_f.
[3] **[20]**+N_f: optical flow aligned input is passed to N_f.

The low score for passing input images without optical flow correction is because of the fact that N_f has to perform both motion correction as well as fusion. For saturated regions in I_0, N_f has to choose details from under or over exposed image. However, in moving regions, it has to choose details from reference image. Hence, the network gets confused in heavily saturated regions, as to whether it is a saturated region or motion affected region. Most often, it chooses to identify those regions as motion affected, hence losing HDR content in those pixels. By passing I_{MC} images to N_f, the fusion network has to focus only on merging them. However, as motion corrected sequence has low HDR content for moving objects and saturated regions, the final result has lower PSNR. Finally, passing optical flow corrected sequence results in higher PSNR as it can transfer details from neighbouring pixels for motion affected and saturated regions.

Architecture ablation: In Table 4, we show results for different choices of

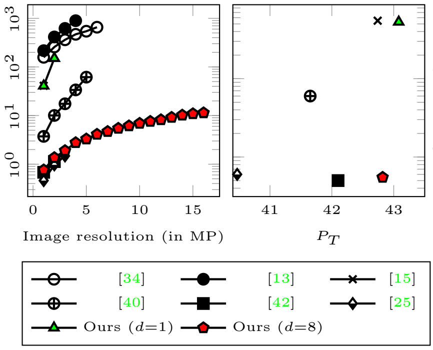
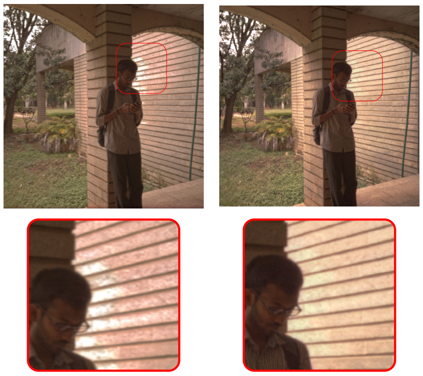


Fig. 5. Our proposed method can fuse images up to 16MP on a GPU with 11GB memory for PSNR higher than state-of-the-art HDR deghosting algorithms. The y-axis for both the plots is in the logarithmic time scale in seconds. See section 4.5 for details.



(a) (b)

Fig. 6. A qualitative comparison between (a) Wu18 [40], (b) Proposed method on a test sequence from [25] dataset.

Table 2. Ablation experiments for different choice of guide image generation. See section 4.3 for more details.

Methods	P_L	P_T
B1 + [3]	34.40	33.50
N_{mc} + [3]	37.55	37.98
N_g	36.38	37.24
B1 + N_g	38.32	39.06
N_{mc} + N_g	**40.73**	**41.56**

Table 3. Ablation on different type inputs to N_f model. See section 4.3 for more details.

Methods	P_L	P_T
N_f	40.88	41.16
N_{mc} + N_f	41.02	41.54
[20] + N_f	**41.33**	**42.82**

Table 4. Base architecture ablation. See section 4.3 for more details.

Methods	P_L	P_T
Vanilla	39.60	41.08
U-Net [30]	41.13	41.76
ResNet18 [10]	41.24	42.45
DenseNet [14]	41.03	41.78
SDC [31]	40.76	41.94
Ours (w/o dense)	41.21	42.34
Ours (w/ dense)	**41.33**	**42.82**

network architecture. The vanilla network denotes using ten convolutional layers instead of XSDC blocks.

4.4 Qualitative evaluation

In Figures 1 and 6, we show qualitative comparison against Wu18 [40] and AHDR [42] methods. Both Wu18 and AHDR fail to reconstruct details in saturated regions of the reference image that is occluded in other input images. As a result, the occluded regions appear as distinct structure in the result (zoomed regions in Figure 1). Sen12 [34] suffers from color bleeding and over-smoothing artifacts introduced in heavily saturated regions of reference image (see Figure 7). In Figure 1 and 7, we show the results by Kalantari17 [15] and proposed method. Kalantari17 method introduces structural distortions for moving objects in saturated regions and hallucinates details in heavily saturated regions.

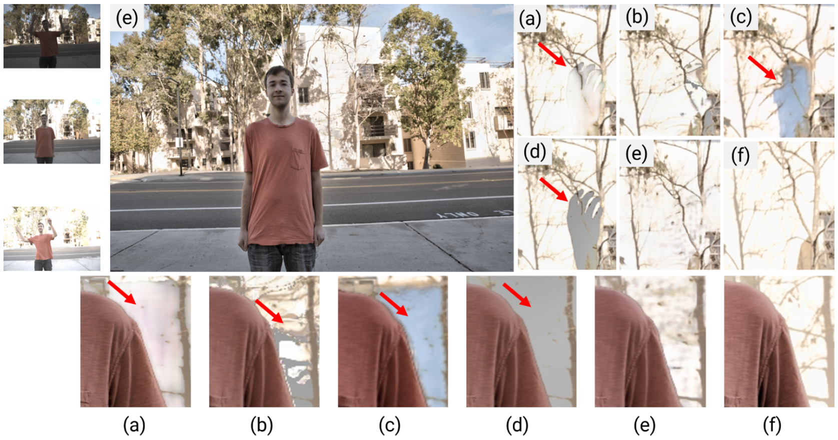


Fig. 7. Qualitative comparison between proposed method and state-of-the-art methods in a test set from UCSD dataset [15]. The input sequence is shown on the left column. (a) SCHDR [25], (b) Kalantari17 [15], (c) Sen12 [34], (d) H_g, (e) H_f and (f) ground truth.

Comparatively, our proposed method generates visually pleasing results without any artifacts[5]. All the results shown in the paper are generated with d=8.

4.5 Running time

In Figure 5 left graph, we compare the running times of seven methods for image resolutions from 1 to 16 MP. The reported numbers are obtained from a machine with NVIDIA RTX 2080 Ti GPU, 32GB RAM and an i7-8700 CPU. For each resolution, we report the average of 15 runs with three varying exposure images. The non-deep methods Sen12 [34] and Hu13 [13] take up to 468 seconds and 893 seconds respectively to fuse a 4MP sequence. Among the deep learning-based methods, Kalantari17 [15] fuses a 1MP sequence in 42 seconds. For any other higher resolutions, it throws an out-of-memory error. Similarly, Yan19 [42] can support a maximum of 2MP resolution in 0.83 seconds. Wu18 [40] method takes up to 10 seconds to process 1MP images and can only support up to 5MP. Comparatively, our proposed method with d=8 can fuse 16MP images in 11 seconds on a 11GB GPU card. In Figure 5 right graph, we report the average PSNR-T score for 15 test images of size 1.5MP from dataset provided by Kalantari17 [15]. It should be noted that, for a fair comparison, we have used SIFT followed by the RANSAC method to align images for both Wu18 and our method. However, as the input images are downsampled by a factor of 8 before image alignment, it does not have significant overhead on our approach.

5 Discussion

Guide vs Final HDR: As shown in Figures 1, 7 and 8, H_g image lacks details in moving regions affected by saturation. This effect is because H_g generation

[5] For more results, please check supplementary material.

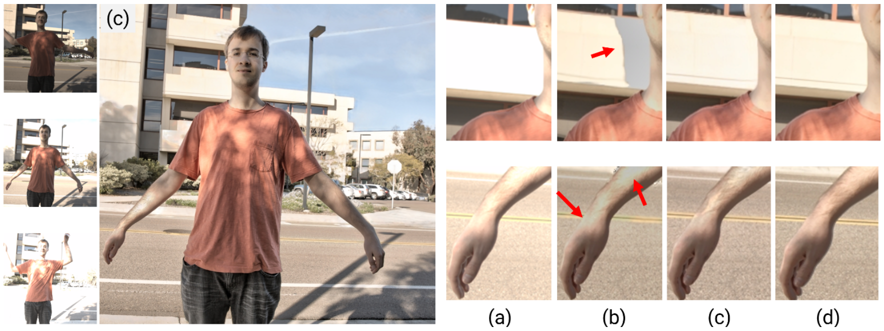


Fig. 8. H_g vs H_f: (a) Patches from I_0 image, (b) H_g, (c) H_f and (d) ground truth. The regions affected by motion and saturation is not corrected in the guide image (highlighted by the red arrows). In contrast, N_f network can generate plausible details in those regions as well.

process is a weight map-based method; it does not have the freedom to move details from neighboring pixels. In contrast, as input to N_f are optical flow corrected images, the H_f has proper contents in those regions as well.
Limitations: Similar to [15,40,42], our method is also trained for a fixed number of three images. Hence, extending our method (including [15,41,42]) requires retraining with the new number of images. To overcome this issue, one can use the max-mean fusion strategy instead of feature concatenation [25].

6 Conclusion

In the current era of high-resolution imaging with smartphones and DSLRs, it is difficult to find the best performing HDR deghosting method that can support higher image resolutions with limited computational resources. We address this issue by performing all heavy processing in low resolution (including optical flow computation) and upscale the low-resolution output to full resolution using a guide image. The guide image is generated using a simple weight map based fusion of original full-resolution inputs. We have shown that our approach improves the state-of-the-art in generating high-quality artifact-free HDR images. Our approach can fuse 16-megapixel images in about 10 seconds on a GPU with 11GB memory and achieves state-of-the-art results in publicly available datasets. We also have demonstrated the use of BGU for tasks where the guide image is not readily available. We hope that our paper will motivate the research community to explore BGU for other fusion tasks.

Acknowledgements: This work was supported by a project grant from MeitY (No.4(16)/2019-ITEA), Govt. of India and WIRIN.

References

1. Bogoni, L.: Extending dynamic range of monochrome and color images through fusion. In: Proceedings. 15th International Conference on Pattern Recognition. (2000) 3
2. Chen, J., Adams, A., Wadhwa, N., Hasinoff, S.W.: Bilateral guided upsampling. ACM Transactions on Graphics (TOG) **35**(6), 203 (2016) 3
3. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH 2008 classes. p. 31. ACM (2008) 4, 5, 10, 12
4. Eden, A., Uyttendaele, M., Szeliski, R.: Seamless image stitching of scenes with large motions and exposure differences. In: Conference on Computer Vision and Pattern Recognition. IEEE (2006) 1, 3
5. Gallo, O., Gelfandz, N., Chen, W.C., Tico, M., Pulli, K.: Artifact-free high dynamic range imaging. In: ICCP. pp. 1–7. IEEE (2009) 1, 3
6. Gallo, O., Troccoli, A., Hu, J., Pulli, K., Kautz, J.: Locally non-rigid registration for mobile HDR photography. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 49–56 (2015) 2, 3
7. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Transactions on Graphics (TOG) **36**(4), 1–12 (2017) 3, 4
8. Granados, M., Kim, K.I., Tompkin, J., Theobalt, C.: Automatic noise modeling for ghost-free HDR reconstruction. ACM Transactions on Graphics (TOG) **32**(6), 201 (2013) 1, 3
9. Grosch, T.: Fast and robust high dynamic range image generation with camera and object movement. Vision, Modeling and Visualization, RWTH Aachen pp. 277–284 (2006) 1, 3
10. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (2016) 12
11. Heo, Y.S., Lee, K.M., Lee, S.U., Moon, Y., Cha, J.: Ghost-free high dynamic range imaging. In: Asian Conference on Computer Vision. pp. 486–500. Springer (2010) 1
12. Hu, J., Gallo, O., Pulli, K.: Exposure stacks of live scenes with hand-held cameras. In: ECCV, pp. 499–512. Springer (2012) 2
13. Hu, J., Gallo, O., Pulli, K., Sun, X.: HDR deghosting: How to deal with saturation? In: IEEE Conference on Computer Vision and Pattern Recognition (2013) 2, 3, 4, 9, 11, 12, 13
14. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017) 12
15. Kalantari, N.K., Ramamoorthi, R.: Deep high dynamic range imaging of dynamic scenes. ACM Transactions on Graphics (Proceedings of SIGGRAPH 2017) **36**(4) (2017) 2, 3, 4, 5, 8, 9, 10, 11, 12, 13, 14
16. Kang, S.B., Uyttendaele, M., Winder, S., Szeliski, R.: High dynamic range video. In: ACM Transactions on Graphics (TOG). vol. 22, pp. 319–325. ACM (2003) 2, 3
17. Kao, W.C., Hsu, C.C., Chen, L.Y., Kao, C.C., Chen, S.H.: Integrating image fusion and motion stabilization for capturing still images in high dynamic range scenes. IEEE Transactions on Consumer Electronics **52**(3), 735–741 (2006) 3

18. Khan, E.A., Akyiiz, A., Reinhard, E.: Ghost removal in high dynamic range images. In: IEEE International Conference on Image Processing (2006) 3
19. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) 9
20. Liu, C., et al.: Beyond pixels: exploring new representations and applications for motion analysis. Ph.D. thesis, Massachusetts Institute of Technology (2009) 5, 7, 11, 12
21. Ma, K., Li, H., Yong, H., Wang, Z., Meng, D., Zhang, L.: Robust multi-exposure image fusion: A structural patch decomposition approach. IEEE Transactions on Image Processing **26**(5), 2519–2532 (2017) 3
22. Mantiuk, R., Kim, K.J., Rempel, A.G., Heidrich, W.: Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. ACM Transactions on graphics (TOG) **30**(4), 40 (2011) 10
23. Min, T.H., Park, R.H., Chang, S.: Histogram based ghost removal in high dynamic range images. In: 2009 IEEE International Conference on Multimedia and Expo. pp. 530–533. IEEE (2009) 1, 3
24. Pece, F., Kautz, J.: Bitmap movement detection: HDR for dynamic scenes. In: 2010 Conference on Visual Media Production. pp. 1–8. IEEE (2010) 1
25. Prabhakar, K.R., Arora, R., Swaminathan, A., Singh, K.P., Babu, R.V.: A fast, scalable, and reliable deghosting method for extreme exposure fusion. In: 2019 IEEE International Conference on Computational Photography (ICCP). pp. 1–8. IEEE (2019) 2, 8, 9, 10, 11, 12, 13, 14
26. Prabhakar, K.R., Babu, R.V.: Ghosting-free multi-exposure image fusion in gradient domain. In: IEEE International Conference on Acoustics, Speech and Signal Processing (2016) 1
27. Prabhakar, K.R., Srikar, V.S., Babu, R.V.: Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In: IEEE International Conference on Computer Vision (ICCV). pp. 4724–4732. IEEE (2017) 1
28. Raman, S., Chaudhuri, S.: Reconstruction of high contrast images for dynamic scenes. The Visual Computer **27**(12), 1099–1114 (2011) 1, 3
29. Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., Myszkowski, K.: High dynamic range imaging: acquisition, display, and image-based lighting. Morgan Kaufmann (2010) 3
30. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 6, 12
31. Schuster, R., Wasenmuller, O., Unger, C., Stricker, D.: Sdc-stacked dilated convolution: A unified descriptor network for dense matching tasks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2556–2565 (2019) 7, 12
32. Sen, P.: Overview of state-of-the-art algorithms for stack-based high-dynamic range (HDR) imaging. Electronic Imaging **2018**(5), 1–8 (2018) 3
33. Sen, P., Aguerrebere, C.: Practical high dynamic range imaging of everyday scenes: photographing the world as we see it with our own eyes. IEEE Signal Processing Magazine **33**(5), 36–44 (2016) 3
34. Sen, P., Kalantari, N.K., Yaesoubi, M., Darabi, S., Goldman, D.B., Shechtman, E.: Robust patch-based HDR reconstruction of dynamic scenes. ACM Trans. Graph. **31**(6), 203 (2012) 2, 3, 4, 8, 9, 11, 12, 13
35. Sidibé, D., Puech, W., Strauss, O.: Ghost detection and removal in high dynamic range images. In: 17th European Signal Processing Conference (2009) 3

36. Tomaszewska, A., Mantiuk, R.: Image registration for multiexposure high dynamic range image acquisition. In: Proceedings of the International Conference on Computer Graphics, Visualization and Computer Vision (2007) 2, 3
37. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: The state of the art in HDR deghosting: a survey and evaluation. In: Computer Graphics Forum. vol. 34, pp. 683–707. Wiley Online Library (2015) 3
38. Tursun, O.T., Akyüz, A.O., Erdem, A., Erdem, E.: An objective deghosting quality metric for HDR images. In: Computer Graphics Forum. vol. 35, pp. 139–152. Wiley Online Library (2016) 8
39. Ward, G.: Fast, robust image registration for compositing high dynamic range photographs from hand-held exposures. Journal of Graphics Tools **8**(2), 17–30 (2003) 2, 3
40. Wu, S., Xu, J., Tai, Y.W., Tang, C.K.: Deep high dynamic range imaging with large foreground motions. In: European Conference on Computer Vision. pp. 120–135 (2018) 2, 4, 5, 9, 11, 12, 13, 14
41. Wu, S., Xie, S., Rahardja, S., Li, Z.: A robust and fast anti-ghosting algorithm for high dynamic range imaging. In: 2010 IEEE International Conference on Image Processing. pp. 397–400. IEEE (2010) 1, 3, 14
42. Yan, Q., Gong, D., Shi, Q., van den Hengel, A., Shen, C., Reid, I., Zhang, Y.: Attention-guided network for ghost-free high dynamic range imaging. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1751–1760 (2019) 2, 4, 5, 9, 11, 12, 13, 14
43. Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I., Zhang, Y.: Multi-scale dense networks for deep high dynamic range imaging. In: IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 41–50. IEEE (2019) 2, 4, 5, 9, 11
44. Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q., Zhang, Y.: Deep HDR imaging via a non-local network. IEEE Transactions on Image Processing **29**, 4308–4322 (2020) 2, 9, 11
45. Zhang, W., Cham, W.K.: Reference-guided exposure fusion in dynamic scenes. Journal of Visual Communication and Image Representation **23**(3), 467–475 (2012) 1
46. Zimmer, H., Bruhn, A., Weickert, J.: Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. In: Computer Graphics Forum. vol. 30, pp. 405–414. Wiley Online Library (2011) 2, 3