Kernel Prediction Network for Detail-Preserving High Dynamic Range Imaging

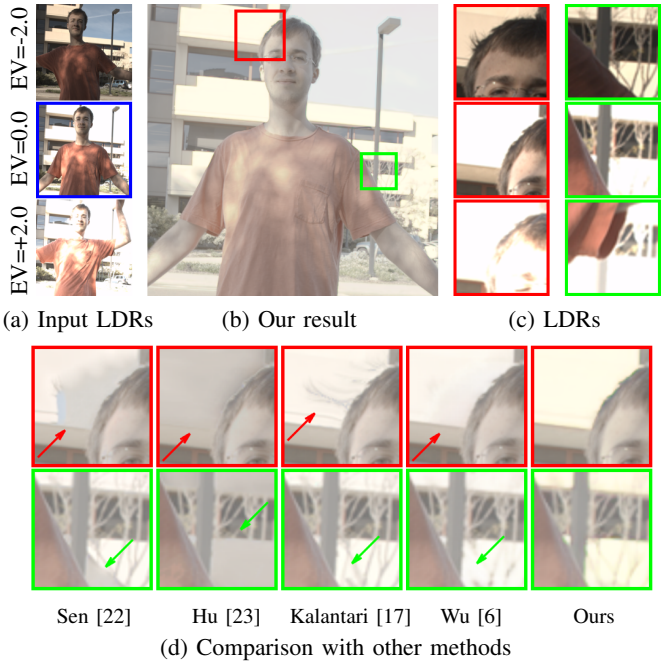
Haesoo Chung*, Yoonsik Kim*, Junho Jo*, Sang-hoon Lee*, and Nam Ik Cho*
* Department of ECE, INMC, Seoul National University, Seoul, Korea
E-mail: {reneeish, terryoo, jottue, tkd1088}@ispl.snu.ac.kr, nicho@snu.ac.kr

***Abstract*—Generating a high dynamic range (HDR) image from multiple exposure images is challenging in the presence of significant motions, which usually causes ghosting artifacts. To alleviate this problem, previous methods explicitly align the input images before merging the controlled exposure images. Although recent works try to learn the HDR imaging process using a convolutional neural network (CNN), they still suffer from ghosting or blurring artifacts and missing details in extremely under/overexposed areas. In this paper, we propose an end-to-end framework for detail-preserving HDR imaging of dynamic scenes. Our method employs a kernel prediction network and produces per-pixel kernels to fully utilize every pixel and its neighborhood in input images for the successful alignment. After applying the kernels to the input images, we generate a final HDR image using a simple merging network. The proposed framework is an end-to-end trainable method without any preprocessing, which not only avoids ghosting or blurring artifacts but also hallucinates fine details effectively. We demonstrate that our method provides comparable results to the state-of-the-art methods regarding qualitative and quantitative evaluations.**

I. Introduction

High dynamic range (HDR) imaging can significantly enhance the viewing experiences by generating an image that has a broad dynamic range of natural luminance and thus closely resembles what humans see. While human visual systems can perceive from very dark to very bright levels of light, standard digital cameras can only capture the restricted dynamic range due to limitations of sensor capacity. Therefore, the resulting low dynamic range (LDR) images inevitably lose information in severely underexposed or overexposed (saturated) regions. In order to alleviate this problem, several approaches [1]–[3] have been proposed to produce HDR images using specialized camera hardware, but these devices are highly expensive and not easily accessible. To address these difficulties, more practical HDR imaging methods that do not require high-priced devices have been proposed.

The most common approach is to take a series of LDR images at bracketed exposures and merge them into a single HDR image [4]. Although this strategy produces satisfactory results when the LDR images are perfectly registered without any motion difference, it usually generates ghosting or blurring artifacts when there is a shift between the images. Specifically, these undesirable artifacts mainly result from two factors that make it challenging to find corresponding pixels among the LDR images: occlusion of moving objects and image misalignment due to camera motions, which are unavoidable in the real world. While the artifacts caused by global image misalignment can be quite mitigated using homography transformation [5], [6], foreground motions are much more challenging to handle. Especially, large-scale foreground motions introduce severe ghosting artifacts in the final HDR image.



(a) Input LDRs (b) Our result (c) LDRs

(d) Comparison with other methods

Fig. 1: Our goal is to generate an HDR image that avoids undesirable artifacts while preserving details. Input LDR images are shown in (a) where the blue box indicates the reference image. Our HDR result after tonemapping is displayed in (b). In (c), the regions with both large foreground motions and severe saturation in LDR images are shown. The proposed method handles both issues successfully, compared to other state-of-the-art methods, as shown in (d).

In order to address the ghosting artifacts, a number of methods [7]–[14] focused on finding moving objects, under the assumption that the input images are nearly static. These approaches detect moving pixels and then simply reject or downweight them when merging the images. Hence, they sometimes suffer from the lack of LDR contents available to reconstruct an HDR image, especially in the objects in a large motion. To deal with scenes with larger displacement, many approaches have been proposed to carry out more sophisticated alignment of input images using optical

flow [15], [16] before merging them into an HDR image. However, optical flow methods basically assume brightness constancy which is always violated in the case of images with different exposure levels, thus additional steps [17]–[19] to meet the constancy assumption are executed. Besides, inevitable optical-flow estimation error leads to color artifacts and distortions in the final HDR image.

Recent approaches have adopted deep neural networks (DNNs) to develop the HDR imaging procedure. Kalantari *et al.* [17] divided the HDR imaging process into two stages, *i.e.*, alignment and HDR merge, and used optical flow algorithm for the alignment stage and a convolutional neural network (CNN) for the HDR merge stage. However, the CNN-based merging method still fails to resolve the aforementioned artifacts resulting from the optical flow, as shown in Fig. 1. Wu *et al.* [6] modeled the HDR imaging as translation problem and implemented the whole process using a CNN to cope with large foreground motions, but the input images are first registered using homography transformation and then fed to the CNN. Although this approach has shown improvement in dealing with ghosting artifacts, it often fails to avoid blurring artifacts and hallucinate fine details in under/overexposed regions. These DNN-based methods need a pre-alignment process, which not only disenables end-to-end training but also corrupts information at the boundaries of input images and takes considerable time.

To overcome the above-mentioned challenges, we propose an end-to-end framework for HDR imaging of dynamic scenes. In contrast to previous methods which directly synthesize HDR images, our model builds upon a kernel prediction network that can generate per-pixel kernels to be applied to input LDR images. These kernels are expected to determine each pixel's importance in creating HDR images, thus suppress pixels in undesirable regions. The resulting kernels are applied to input images, followed by the merging network to produce a final HDR image. The merging network is a straightforward CNN which performs weighted averaging of the aligned images. The proposed HDR imaging network learns to suppress undesirable artifacts and preserve fine details in an end-to-end manner.

In summary, the main contributions of the paper can be summarized as:

- We propose an end-to-end framework for HDR imaging of scenes with foreground or background motions. Our method does not require any pre-alignment procedure that rather introduces unavoidable artifacts and decreases training efficiency.
- Our method can avoid ghosting or blurring artifacts and hallucinate plausible details effectively even when large object motion or severe saturation is present, achieving comparable performance to the state-of-the-art.
- We propose a kernel prediction network for HDR imaging, which generates a unique kernel for each pixel in input images. The generated kernels can fully utilize the contents of input images and perform an accurate alignment.

II. Related Work

HDR imaging has been the subject of extensive research area over the past decades, thus we restrict ourselves to the HDR imaging with dynamic scenes. We categorize the relevant approaches into three classes.

A. Pixel Rejection Based Methods

These approaches assume that all the input images are globally registered so that pixels with motion can be detected and rejected directly. Various algorithms are utilized to identify moving pixels. The algorithms in [7], [12]–[14] compute continuous or binary weights based on the probability that a pixel belongs to the static/moving parts. Several works [8], [9] compare the predicted pixel colors to the original ones. Heo *et al.* [10] detect motion regions using the global intensity transfer functions. Zhang and Cham [11] analyze the image gradient to generate a weighting map. Lee *et al.* [20] and Oh *et al.* [21] use rank minimization to reject moving pixels and reconstruct the HDR image. However, these methods lose available pixel information by rejecting or downweighting pixels and thus struggle to successfully reconstruct the HDR image.

B. Sophisticated Alignment Based Methods

These approaches perform more detailed alignment to find correspondence among the LDR images. A number of works adopted optical flow for sophisticated alignment of LDR images before merging. In order to deal with the brightness constancy assumption of optical flow algorithms, Kang *et al.* [19] use a hierarchical homography and optical flow after mapping the images to the luminance domain. Zimmer *et al.* [18] compute optical flow in the gradient domain. These flow-based approaches often generate undesirable artifacts. Meanwhile, Sen *et al.* [22] and Hu *et al.* [23] rely on patch-based dense correspondence. Hu *et al.* [23] propose a patch-based method to synthesize a set of aligned images and reconstruct an HDR image using intensity and gradient information. Sen *et al.* [22] do not separate alignment and reconstruction but rather formulate them as a joint problem, and let the information from the merging stage help with the alignment stage. These approaches, however, are slow and often fail when large motion or large under/overexposed region exists.

C. Deep Learning Based Methods

Recently, several deep learning based methods have been developed. Kalantari *et al.* [17] first align the input images using the optical flow algorithm [15] and pass them to the merging stage implemented with CNN. Wu *et al.* [6] perform homography transformation on the input images for background alignment and train an image translation network to learn a mapping from the registered LDR images to a ghost-free HDR image. However, even these approaches still suffer from non-negligible artifacts and missing details and require an additional pre-alignment process. To alleviate these problems, we present an end-to-end trainable kernel prediction network.

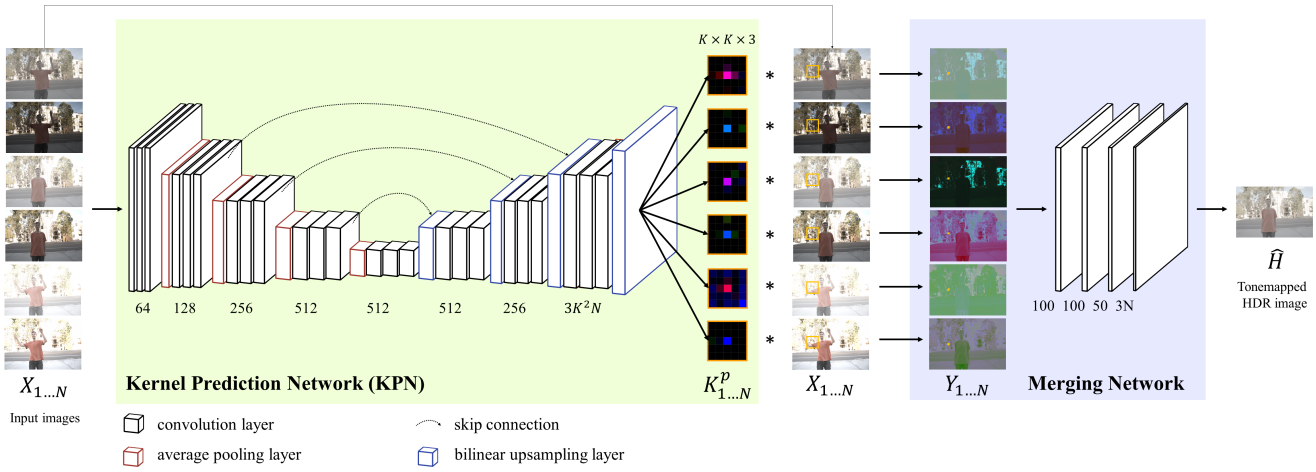


Fig. 2: Our framework for HDR imaging is composed of the kernel prediction network (KPN) and the merging network. The KPN generates per-pixel kernels for alignment, and the merging network merges the aligned images into an HDR image. The final HDR image is visualized after tonemapping.

III. Proposed Method

Given a series of LDR images $\{I_1, \ldots, I_k\}$ sorted by their exposure time, our goal is to generate a ghost-free HDR image H that is aligned to the pre-defined reference image $I_r,\ r \in \{1, \ldots, k\}$. In our experiments, we use three LDR images $\{I_1, I_2, I_3\}$ and set the middle exposure image I_2 as the reference image I_r, but our method can be applied to more input images.

Before feeding the LDR images into the network, we map them to $\{H_1, H_2, H_3\}$ in the HDR domain using gamma correction:

$$H_i = \frac{I_i^\gamma}{t_i},\ \gamma > 1, \tag{1}$$

where t_i denotes the exposure time of the i^{th} image I_i and γ denotes the gamma correction parameter. The LDR images help to identify the under/overexposed regions, while the HDR images facilitate detection of misalignments. Since the LDR and HDR images have different properties, we treat them as independent images, unlike the previous works [6], [17]. We now denote the six input images as $\{X_1, X_2, X_3, X_4, X_5, X_6\}$ instead of $\{I_1, H_1, I_2, H_2, I_3, H_3\}$. We then represent our HDR imaging process as:

$$H = f(X_1, X_2, X_3, X_4, X_5, X_6), \tag{2}$$

where $f(\cdot)$ denotes our network. Our end-to-end network outputs the HDR image H, given the six input images without any additional process.

A. Overall Architecture

Unlike the previous CNN-based approaches [6], [17] that directly obtain an HDR image, the proposed framework produces and applies kernels to each input image and then merges the aligned images in the form of weighted averaging. As shown in Fig. 2, our model consists of two components: the kernel prediction network (KPN) for alignment and the merging network for reconstruction of an HDR image. The KPN generates a distinct kernel for each pixel in each input image. This per-pixel kernel is then applied to the corresponding pixel and its neighborhood, and decides how much to reflect their contents for precise alignment. The merging network then simply estimates weights for weighted averaging of the aligned images and reconstructs a final HDR image.

Kernel prediction network (KPN) The KPN is an encoder-decoder architecture with skip connections resembling the network in [24] which predicts kernels to jointly align and denoise bursts of images. The input images $X_i,\ i = 1, 2, \ldots, N$ are concatenated along the channel dimension and then fed into the network. The last feature map of the KPN has the same size as the input images and $3K^2N$ dimensions. This feature map is reshaped to generate N $K \times K \times 3$ kernels at each pixel. Then the per-pixel kernels are applied to the input images through the dot product. The pixel value at pixel p in the i^{th} resulting image Y_i can be represented as:

$$Y_i^p = \langle K_i^p, N^p(X_i) \rangle, \tag{3}$$

where K_i^p denotes the corresponding kernel and $N^p(X_i)$ denotes the $K \times K$ neighborhood around the pixel p in the image X_i. In our experiments, we set N as 6 and K as 5. The generated kernels extract rich contents in well-exposed regions from the corresponding images, while suppressing the contents in saturated or misaligned regions. As a result, we obtain the aligned images $Y_i,\ i = 1, 2, \ldots, N$, which are passed to the merging network.

Merging network This subnetwork takes the aligned images $Y_i,\ i = 1, 2, \ldots, N$ as input and performs channel-wise concatenation. The merging network then computes a weighted average of them. The architecture is a simple CNN that is composed of four convolutional layers with decreasing

kernel sizes similar to the model in [17]. Typical merging algorithms for HDR imaging compute a weighted average of only aligned HDR images, but here we use both aligned LDR and HDR images $Y_i,\ i = 1, 2, \ldots, N$ to generate the final HDR image:

$$\hat{H}(p) = \frac{\sum_{j=1}^{N} \alpha_j(p) Y_j(p)}{\sum_{j=1}^{N} \alpha_j(p)}, \tag{4}$$

where $\alpha_j(p)$ denotes the weight. While the weight $\alpha_j(p)$ is calculated from the input images using various algorithms [4], [25] in the previous works, our merging network learns to find the optimal weight $\alpha_j(p)$. The output of the last convolutional layer is reshaped into N weight maps with the same shape as the input images. We obtain the final HDR image by computing a weighted average of the aligned images $Y_i,\ i = 1, 2, \ldots, N$ using the generated weight maps.

B. Loss Function

Since HDR images are mostly displayed after tonemapping, we compute the loss function on the tonemapped HDR images. We use μ-law as our tonemapping function, as it is differentiable and thus suitable for training the network. The μ-law function is defined as:

$$\mathcal{T}(H) = \frac{\log(1 + \mu H)}{\log(1 + \mu)}, \tag{5}$$

where μ is a parameter that defines the level of compression, H is the HDR image in HDR domain, and $\mathcal{T}(H)$ is the tonemapped image. In this work, H is always in the range $[0, 1]$ and μ is 5000. We train the network by minimizing the squared ℓ_2 distance between the tonemapped estimated and ground truth HDR images. Our loss function is defined as:

$$\mathcal{L} = \left\| \mathcal{T}(\hat{H}) - \mathcal{T}(H) \right\|_2^2, \tag{6}$$

where $\hat{H}$ and H denote the estimated and ground truth HDR images, respectively.

C. Implementation Details

In the KPN, we use 3×3 kernels with a stride of 1 in all the convolutional layers, which are followed by ReLU activations. In the merging network, we use four convolutional layers with decreasing kernel sizes from 7 to 1. The first three layers are followed by ReLU activations and the last layer is followed by a sigmoid activation.

IV. Experiments

A. Experimental Settings

Datasets We use the dataset provided by [17] for both training and testing. This dataset consists of 74 scenes for training and 15 scenes for testing with ground truth images. For each scene, there are three LDR images with exposure biases of $\{-2.0,\ 0.0,\ +2.0\}$ or $\{-3.0,\ 0.0,\ +3.0\}$.

Evaluation Metrics For the quantitative evaluation, we use five evaluation metrics. We compute the PSNR and SSIM values between the tonemapped estimated and ground truth HDR images (PSNR-T and SSIM-T), and the PSNR and SSIM values between the estimated and ground truth HDR images before tonemapping (PSNR-L and SSIM-L). We also compute the HDR-VDP-2 score [26] that measures the visual quality of HDR images.

(a) Example with large foreground motions and saturation

(b) Example with moving objects and under/overexposure

Fig. 3: A qualitative evaluation of our method on images in challenging cases. In both (a) and (b), the top half part shows the input LDR images with different exposures, our result after tonemapping, and the areas with large foreground motions (from left to right). The bottom half part compares the result produced by our method and other state-of-the-art methods.

B. Experimental Results

We perform both quantitative and qualitative evaluations on the proposed method. We also compare our results with previous state-of-the-art methods, including two patch-based

TABLE I: Quantitative comparisons of our method with other state-of-the-art methods. PSNR-T/SSIM-T is calculated on the tonemapped images and PSNR-L/SSIM-L is calculated on the linear images. HDR-VDP-2 scores evaluate the visual quality of HDR images.

	PSNR-T	SSIM-T	PSNR-L	SSIM-L	HDR-VDP-2
Sen [22]	41.11	0.9815	38.82	0.9749	57.43
Hu [23]	34.87	0.9698	31.72	0.9511	55.20
Kalantari [17]	**42.70**	0.9879	41.21	0.9845	59.68
Wu [6]	41.73	0.9854	**41.42**	0.9847	61.30
Ours	42.42	**0.9920**	38.60	**0.9850**	**66.20**

methods [22], [23], the method using optical flow based alignment and a DNN-based merger [17], and the DNN-based method with homography transform based pre-alignment [6]. Note that we used the codes provided by the authors for all methods.

Quantitative Evaluations We compute five aforementioned evaluation metrics and compare with the other methods. Table I shows quantitative comparison of the proposed method with other methods. The proposed method achieves the best performance in terms of SSIM-T, SSIM-L, and HDR-VDP-2, and produces comparable PSNR-T. Especially, our method achieves significantly higher HDR-VDP-2 scores than other methods, which shows our method can generate a high-quality HDR image.

Qualitative Evaluations We compare our qualitative results with state-of-the-art methods. The results are shown in Fig. 3. The test images contain moving objects, global misalignment and under/overexposed regions. Sen *et al.*'s method [22] sometimes produces geometric distortions and artifacts. Hu *et al.*'s method [23] often fails to reconstruct the HDR image. Kalantari *et al.*'s method [17] generates artifacts mainly due to the failure of optical flow based alignment. Wu *et al.*'s method [6] produces blurry regions and cannot hallucinate plausible details. Fig. 3 (b) shows that all the results except ours fail to reconstruct the details or texture successfully in the presence of occlusion, saturation, and underexposure. The overall results demonstrate the capability of our method to generate detail-preserving HDR results without apparent artifacts.

V. Conclusion

In this paper, we have presented an end-to-end framework for HDR imaging of dynamic scenes. We have proposed the kernel prediction network to fully utilize every pixel in input images and demonstrated the proposed method's superior ability to generate high-quality HDR images without additional preprocessing. We have successfully handled the ghosting or blurring artifacts and preserved fine details even in the presence of severe under/overexposure, displacement, and occlusion.

Acknowledgments

This paper was result of the research project supported by SK hynix Inc.

References

[1] S. K. Nayar and T. Mitsunaga, "High dynamic range imaging: spatially varying pixel exposures," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* vol. 1, pp. 472-479, 2000.

[2] J. Tumblin, A. Agrawal, and R. Raskar, "Why I want a gradient camera," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* vol. 1, pp. 103-110, 2005.

[3] M. D. Tocci, C. Kiser, N. Tocci, and P. Sen, "A versatile HDR video production system," in *ACM Transactions on Graphics (TOG),* vol. 30, no. 4, pp. 41:1-41:10, 2011.

[4] P. E. Debevec and J. Malik, "Recovering high dynamic range radiance maps from photographs," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques,* pp. 369-378, 1997.

[5] A. Tomaszewska and R. Mantiuk, "Image registration for multi-exposure high dynamic range image acquisition," 2007.

[6] S. Wu, J. Xu, Y. Tai, and C. Tang, "Deep high dynamic range imaging with large foreground motion,"in *Proceedings of the European Conference on Computer Vision (ECCV),* 2018.

[7] E. A. Khan, A. O. Akyuz, and E. Reinhard, "Ghost removal in high dynamic range images," in *International Conference on Image Processing,* pp. 2005-2008, 2006.

[8] T. Grosch, "Fast and robust high dynamic range image generation with camera and object movement," in *Vision, Modeling and Visualization*, pp. 277-284, 2006.

[9] S. Raman and S. Chaudhuri, "Reconstruction of high contrast images for dynamic scene," in *The Visual Computer,* vol. 27, no. 12, pp. 1099-1114, 2011.

[10] Y. Heo, K. Lee, S. Lee, Y. Moon, and J. Cha, "Ghost-free high dynamic range imaging," in *IEEE Asian Conference on Computer Vision (ACCV),* pp. 486-500, 2011.

[11] W. Zhang and W. Cham, "Gradient-directed multiexposure composition," in *IEEE Transactions on Image Processing,* vol. 21, no. 4, pp. 2318-2323, 2012.

[12] J. An, S. J. Ha, N. I. Cho, "Reduction of ghost effect in exposure fusion by detecting the ghost pixels in saturated and non-saturated regions," *IEEE International Conference on Image Processing (ICIP),* pp. 1101-11104, 2012.

[13] J. An, S. J. Ha, N. I. 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*, vol. 2014, no. 1, 2014.

[14] J. An, S. J. Ha, J. G. Kuk, N. I. Cho, "A multi-exposure image fusion algorithm without ghost effect," *IEEE International Conference on Acoustics, Speech, Signal Processing (ICASSP),* pp. 1565-1568, 2011.

[15] C. Liu, "Beyond pixels: exploring new representations and applications for motion analysis," PhD thesis, Massachusetts Institute of Technology, 2009.

[16] Z. Chen, H. Jin, Z. Lin, S. Cohen, and Y. Wu, "Large displacement optical flow from nearest neighbor fields," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2443-2450, 2013.

[17] N. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," in *ACM Transactions on Graphics (TOG),* vol. 36, no. 4, pp. 144:1-144:12, 2017.

[18] H. Zimmer, A. Bruhn, and J. Weichkert, "Freehand HDR imaging of moving scenes with simultaneous resolution enhancement," in *Computer Graphics Forum,* pp. 405-414, 2011.

[19] S. B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski, "High Dynamic Range Video," in *ACM Transactions on Graphics (TOG),* vol. 22, no. 3, pp. 319-325, 2003.

[20] 18) C. Lee, Y. Li, and V. Monga, "Ghost-Free High Dynamic Range Imaging via Rank Minimization," in *IEEE Signal Processing Letters (SPL),* vol. 21, no. 4, pp. 2318-2323, 2012.

[21] T. Oh, J. Lee, Y. Tai, and I. Kwon, "Robust high dynamic range imaging by rank minimization," in *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol.37, no. 6, pp. 1219-1232, 2015.

[22] P. Sen *et al.*, "Robust Patch-Based HDR Reconstruction of Dynamic Scenes," in *ACM Transactions on Graphics (TOG),* vol. 31, no. 6, pp. 203:1-203:11, 2012.

[23] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: how to deal with saturation?," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 1163-1170, 2013.

[24] B. Mildenhall *et al.*, "Burst denoising with kernel prediction networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 2502-2510, 2018.
[25] M. Granados *et al.*, "Optimal HDR reconstruction with linear digital cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 215-222, 2010.
[26] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heilrich, "HDR-VDP-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions," in *ACM Transactions on Graphics (TOG),* vol. 30, no. 4, pp. 40:1-40:14, 2011.