Enhanced Bidirectional Motion Estimation Using Feature Refinement for HDR Imaging

An Gia Vien, Truong Thanh Nhat Mai, Seonghyun Park, Gahyeon Kim, and Chul Lee
Department of Multimedia Engineering, Dongguk University, Seoul, Korea
E-mail: {viengiaan,mtntruong,seonghyun,2019112529}@mme.dongguk.edu, chullee@dongguk.edu

***Abstract*—We propose a high dynamic range (HDR) image synthesis algorithm based on enhanced bidirectional motion estimation using feature refinement. First, we extract multiscale features from input low dynamic range (LDR) images and then estimate accurate motion vector fields between them in a coarse-to-fine manner via progressive refinement. Then, we estimate adaptive local kernels to merge only valid information in the spatio-exposed neighboring pixels for synthesis. Finally, we refine the initially merged image by exploiting global information to further improve synthesis performance. Experimental results show that the proposed algorithm outperforms state-of-the-art algorithms in quantitative and qualitative comparisons.**

I. Introduction

High dynamic range (HDR) imaging techniques have been actively developed to overcome the dynamic range limitation of conventional cameras by extending the range of intensity levels. A common approach to HDR image acquisition is merging multiple low dynamic range (LDR) images captured with different exposure times [1], called multi-exposure fusion (MEF). However, if the LDR images contain camera or object motions, ghosting artifacts appear in the synthesized HDR images due to misalignment among the images. Therefore, a lot of research has been conducted on HDR image synthesis without ghosting artifacts [2].

Early attempts for ghost-free HDR imaging employed mathematical models based on the properties of motions. For example, in [3], [4], the correspondences between input LDR images were first estimated, and then the aligned images using the estimated correspondences were merged to obtain an HDR image. In [5]–[7], the rank minimization framework was employed to exploit the linearity of the background scene with respect to exposure times. Sen *et al*. [8] formulated a joint optimization problem of alignment and fusion for HDR image synthesis. Further, Hu *et al*. [9] constrained the consistencies of textures and radiance values among the LDR images for alignment. However, modeling inaccuracies in these algorithms may degrade the quality of the synthesized HDR images. Further, they often demand high computational resources to solve optimization problems.

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00011, Video Coding for Machine) and in part by the National Research Foundation of Korea (NRF) grant funded MSIP (No. NRF-2022R1F1A1074402).

Recently, deep learning-based approaches have exhibited significant synthesis performance improvement by learning semantic information using convolution neural networks (CNNs). A common approach to using deep learning is estimating the correspondences between LDR images to suppress the effects of motions on an HDR image. For example, Kalantari and Ramamoorthi [10] employed CNNs to refine aligned images obtained using optical flows and then merged them using a weighted sum with the learned weights. In [11], instead of pre-alignment, alignment and HDR image synthesis were jointly performed using an encoder-decoder architecture. In [12], [13], attention mechanism and deformable convolution [14] were employed to better exploit essential information in the LDR images for fusion. Mai *et al.* [15] employed SIFT-Flow [16] to align the input images then compensated for incorrect alignment by developing a rank minimization-based network. Note that, because the input LDR images contain invalid information caused by occlusions and poor exposures, estimated correspondences contain errors, resulting in degraded synthesis performance. Vien *et al.* [17] proposed the cyclic cost volume to improve the performance of the correspondence estimation for HDR imaging. However, because their algorithm has separate subnetworks for correspondence estimation and HDR synthesis, it may fail to fully exploit motion information during HDR image synthesis.

To address the aforementioned issues, we develop an HDR imaging algorithm based on enhanced bidirectional motion estimation using feature refinement. First, multiscale features are extracted from the input LDR images, and then the motion vector fields between the LDR image pairs are estimated in a coarse-to-fine manner via progressive refinement. Next, we estimate adaptive local kernels to merge only valid information in the local neighboring pixels for synthesis. Finally, we refine the synthesized HDR image using global residual learning to improve the synthesis performance further by exploiting global information. Experimental results show that the proposed algorithm provides higher synthesis performance compared with state-of-the-art algorithms [12], [13], [17].

II. Proposed Algorithm

Fig. 1 shows an overview of the proposed algorithm. Given three input LDR images $\{I_1, I_2, I_3\}$ an HDR image aligned with the middle exposure image I_2 is synthesized. First, the input LDR images $\{I_1, I_2, I_3\}$ are transformed into irradiance

978-616-590-477-3 ©2022 APSIPA

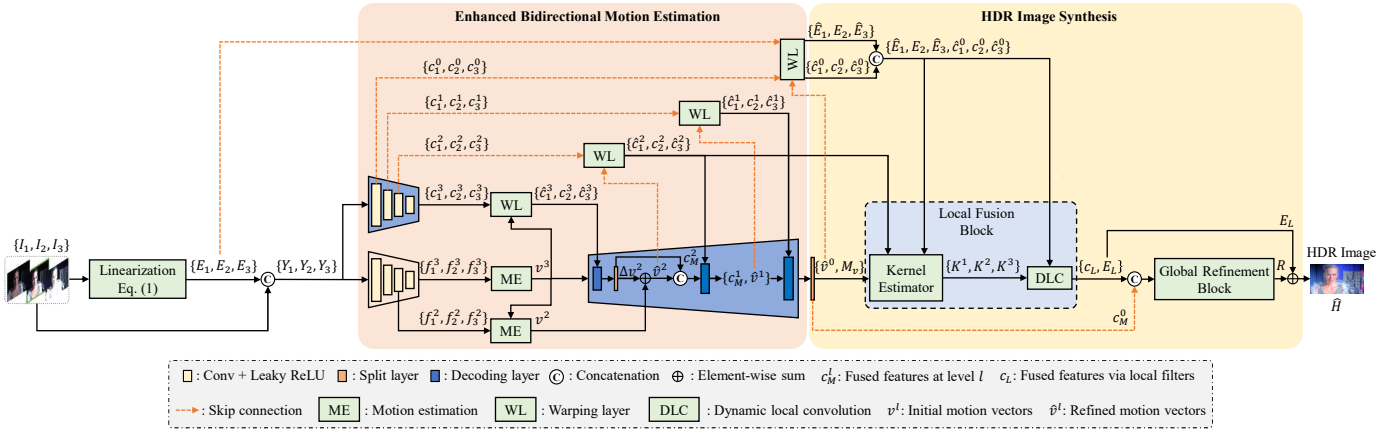

Fig. 1. Overview of the proposed HDR synthesis algorithm. The proposed algorithm first extracts contextual features from the input images using two encoders having the same architecture and then progressively refines bidirectional motion vector fields. The local fusion block (LFB) and global refinement block (GRB) synthesizes an HDR image and then refines it to improve the visual quality, respectively.

maps $\{E_1, E_2, E_3\}$ using the camera response function

$$E_k = \frac{(I_k)^\gamma}{\Delta t_k}, \tag{1}$$

where $k \in \{1, 2, 3\}$ is the exposure index and Δt_k is the exposure time for I_k. We set $\gamma = 2.24$ as in [18].

Then, the proposed algorithm takes $\{Y_1, Y_2, Y_3\}$, where $Y_k = \{I_k, E_k\}$, as the input and reconstructs an HDR image $\widehat{H}$. The proposed algorithm comprises two stages: enhanced bidirectional motion estimation and HDR image synthesis. In the first stage, two motion vector fields between E_2 and E_1 and between E_2 and E_3 are estimated. Then, two estimated motion vector fields are refined with warped features. Next, the local fusion block (LFB) learns to generate adaptive kernels by exploiting local neighboring information for merging images $\{E_1, E_2, E_3\}$. Finally, the global refinement block (GRB) generates a residual image R by considering global information to refine the synthesized HDR image E_L.

A. Enhanced Bidirectional Motion Estimation

Because of different exposures, the input LDR images contain different poorly exposed regions among them, thereby making accurate motion estimation difficult. To address this challenge, Vien *et al.* [17] developed bidirectional motion estimation with cyclic matching costs. Although their algorithm improves the accuracy of motion estimation, it estimates motion vector fields and synthesizes HDR images independently, providing suboptimal HDR image synthesis performance. To address this issue, in this work, we develop an encoder-decoder framework to jointly refine bidirectional motion vector fields and synthesize high-quality HDR images. Specifically, the proposed algorithm first extracts feature maps from each input using an encoder to estimate motion vectors. Then, it progressively refines the estimated motion vectors in a coarse-to-fine manner. The detailed description of each component of this module is given below.

Encoders: To extract effective contextual information from input images for motion estimation and refinement, we design two encoders with the same architecture. As shown in Fig. 1, each encoder extracts four-level pyramid feature maps—f_k^l for motion estimation and c_k^l for motion refinement and HDR image synthesis, where $l \in \{0, 1, 2, 3\}$ is the level index. Each level of the encoder comprises two convolution layers with a 3×3 kernel and another convolution layer with a 3×3 kernel and a stride of 2 for downsampling of f_k^l and c_k^l.

Motion estimation: We employ the motion estimator (ME) in [17] to estimate two initial motion vectors—$v_{2\to1}(\mathbf{x})$ between E_2 and E_1 and $v_{2\to3}(\mathbf{x})$ between E_2 and E_3—at each pixel location $\mathbf{x}$ in E_2. Specifically, each non-reference feature map f_1^l or f_3^l and the reference feature map f_2^l are used to generate a spatial attention map [12], which determines regions of motions and poor exposures. Then, the attention maps and their corresponding input features are used to estimate motion vector fields by building a cyclic cost volume [17] composed of two bidirectional matching costs and a cyclic matching cost. The initial vector fields are estimated only at $l \in \{2, 3\}$.

Warping layer: The proposed algorithm synthesizes an HDR image by merging the aligned images. To this end, we employ the warping operation [19], which aligns the input images using the estimated motion vector fields. More specifically, both the input images E_1 and E_3 and their corresponding features c_1^l and c_3^l are warped toward the reference image to obtain warped images $\widehat{E}_1$ and $\widehat{E}_3$ and features $\hat{c}_1^l$ and $\hat{c}_3^l$.

Decoder: In contrast to Vien *et al.*'s algorithm [17], which directly uses the estimated motion vectors for HDR image synthesis, we further refine the estimated motion vector fields $\{v^2, v^3\}$ using contextual information $\{\hat{c}_1^l, c_2^l, \hat{c}_3^l\}$ to improve the synthesis performance. To this end, we develop a coarse-to-fine approach for motion refinement. Specifically, we design a decoding layer D^l similar to that in [20], which takes an initial motion vector field and feature maps and outputs a

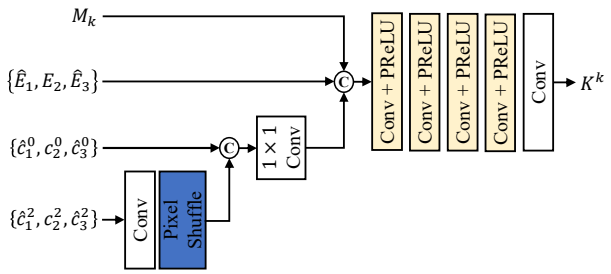

Fig. 2. Architecture of the kernel estimator in LFB with a mask M_k for $k \in \{1, 2, 3\}$.

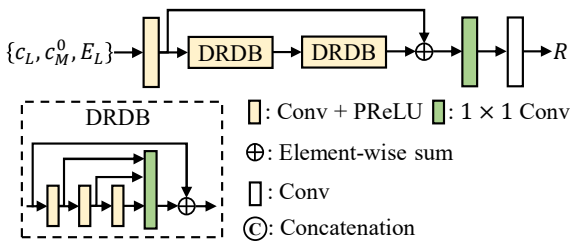

Fig. 3. Architecture of GRB.

fused feature map c_M^{l-1} and refined motion vector field $\hat{v}^{l-1}$. Since the fused feature map c_M^{l-1} is obtained by merging input features $\{\hat{c}_1^l, c_2^l, \hat{c}_3^l\}$, it can effectively restore lost information in poorly exposed regions, thereby improving the accuracy of motion estimation. Simultaneously, the improved motion vectors align c_1^l and c_3^l more accurately, thus improving the representation ability of c_M^{l-1}. Therefore, the proposed approach jointly refines motion vector fields and the fused feature maps.

As the initial motion vector fields v^l are estimated only at levels $l \in \{2, 3\}$, D^3 generates a residual field Δv^2 to refine the motion vector field as $\hat{v}^2 = v^2 + \Delta v^2$, as shown in Fig. 1. Additionally, the last decoding layer D^1 generates the motion reliability mask M_v that indicates the reliability of the motion vector at the corresponding pixel location in the range $[0, 1]$, which will be used for HDR image synthesis. To summarize, the decoder can be compactly represented as

$$\{\Delta v^2, c_M^2\} = D^3\left(\hat{c}_1^3, c_2^3, \hat{c}_3^3, v^3\right), \tag{2}$$

$$\{\hat{v}^1, c_M^1\} = D^2\left(c_M^2, \hat{c}_1^2, c_2^2, \hat{c}_3^2, \hat{v}^2\right), \tag{3}$$

$$\{\hat{v}^0, c_M^0, M_v\} = D^1\left(c_M^1, \hat{c}_1^1, c_2^1, \hat{c}_3^1, \hat{v}^1\right). \tag{4}$$

Each decoding layer D^l comprises six convolution layers with a 3×3 kernel and a stride of 1 along with a single deconvolution layer with a 4×4 kernel. Each convolution layer is followed by a PReLU activation function [21].

B. HDR Image Synthesis

As shown in Fig. 1, an HDR image $\widehat{H}$ is synthesized by merging three images $\{\widehat{E}_1, E_2, \widehat{E}_3\}$. $\widehat{E}_1$ and $\widehat{E}_3$ are warped by the warping layer using the refined motion vector fields $\hat{v}_{2\to1}^0$ and $\hat{v}_{2\to3}^0$, respectively. In this work, to synthesize high-quality HDR images, we develop two blocks—LFB and GRB—to exploit local and global information, respectively. LFB is composed of the kernel estimator, which learns to generate adaptive convolution kernels $K_{x,y}^k$ for $k \in \{1, 2, 3\}$ for each pixel (x, y), and the dynamic local convolution (DLC) block, which synthesizes the intermediate image E_L and feature map c_L. GRB refines E_L by learning a residual image R. Finally, the synthesized HDR image is obtained as $\widehat{H} = E_L + R$.

LFB: As the contextual information in the input images improves synthesis performance [17], we exploit the features $\{\hat{c}_1^0, c_2^0, \hat{c}_3^0\}$ as well as the input images $\{\widehat{E}_1, E_2, \widehat{E}_3\}$ as input to the kernel estimator. Moreover, each input image contains regions with invalid information, caused by poor exposures or motion errors, thereby degrading the synthesis performance. In this work, we consider poor exposures in E_2 and unreliable warping due to motion errors in $\widehat{E}_1$ and $\widehat{E}_3$. First, we define the binary mask M_2 for E_2, where a value of 1 indicates that the corresponding pixel in I_2 is well exposed, *i.e.*, $\tau < I_2(x, y) < 255 - \tau$ with a threshold τ. We set $\tau = 15$ in this work. Next, we define the soft masks for $\widehat{E}_1$ and $\widehat{E}_3$ using the motion reliability mask as $M_1 = M_v$ and $M_3 = 1 - M_v$, respectively.

The kernel estimator is composed of three subnetworks with the same architecture, each of which estimates dynamic local kernels for each input image. Fig. 2 shows the architecture of a subnetwork. The feature maps $\{\hat{c}_1^2, c_2^2, \hat{c}_3^2\}$ are $4\times$ upsampled and then merged with the feature maps $\{\hat{c}_1^0, c_2^0, \hat{c}_3^0\}$. Then, the merged feature maps, input images $\{\widehat{E}_1, E_2, \widehat{E}_3\}$, and each mask M_k are fed into four convolution layers with a 3×3 kernel and a PReLU activation function [21]. The last convolution layer with a 3×3 kernel generates dynamic local kernel $K_{x,y}^k$ for each pixel (x, y) and exposure $k \in \{1, 2, 3\}$. The coefficients are normalized as $\sum_i \sum_j \sum_k K_{x,y}^k(i, j) = 1$, where (i, j) denotes local coordinates around (x, y) and $k \in \{1, 2, 3\}$ is the exposure index.

Subsequently, the synthesized HDR image E_L is obtained via DLC with the learned coefficients $K_{x,y}^k$ as

$$E_L(x, y) = \sum_{k=1}^{3} K_{x,y}^k * \widetilde{E}_k(x, y), \tag{5}$$

where $\widetilde{E}_i(x, y)$ is a local patch centered at (x, y) in $\widehat{E}_i$ with $\widehat{E}_2 = E_2$. The feature map c_L is similarly obtained.

GRB: In (5), the LFB merges multiple images using only local neighbors. Therefore, if the local neighbors contain invalid information caused by motion errors or poor exposures, the LFB may yield visible artifacts in those regions. To address this issue and further improve the synthesis performance, we develop a GRB by adopting global residual learning, as shown in Fig. 3. Specifically, as the feature maps c_L and c_M^0 convey contextual information and global information, respectively, the GRB uses those two feature maps to improve the synthesis performance. More specifically, to facilitate

(a) (b) (c) (d) (e) (f) (g) (h)

Fig. 4. Qualitative comparison of synthesized HDR images. (a)–(c) Input LDR images and the synthesized results of (d) AHDRNet [12], (e) ADNet [13], (f) Vien *et al.* [17], (g) the proposed algorithm, and (h) ground-truth. The second and fourth rows show the magnified parts of the red rectangles in the first and third rows, respectively.

global information by increasing the receptive field, we adopt two dilated residual dense blocks (DRDBs) as in [13], [17].

C. Training

To train the proposed network, we adopt the HDR image reconstruction loss $\mathcal{L}_r$, as in [10]–[13], between the output $\widehat{H}$ and ground-truth $H_{\rm gt}$ in the tone-mapped domain [10] as

$$\mathcal{L}_r = \|\mathcal{T}(\widehat{H}) - \mathcal{T}(H_{\rm gt})\|_1, \tag{6}$$

with

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

where $\mu = 5000$ controls the amount of compression.

We use the Adam optimizer [22] with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 200 epochs with an initial learning rate of 10^{-4} and decay rate of 0.1. The training took approximately four days using a PC with an Intel® Core™ i9-7900X @3.30GHz CPU, 64GB RAM, and four Nvidia RTX™ 3090 GPUs.

Training dataset: We use the dataset from [18], which contains 1,494 HDR images with their corresponding LDR images of resolution 1900×1060. For testing, we randomly selected 210 HDR images from the dataset. Thus, the training set contains 1,284 images out of 1,494. Further, we crop 128×128 patches with a stride of 128 to generate training samples.

III. Experimental Results

A. Quantitative and Qualitative Comparison

We compare the HDR image synthesis performance of the proposed algorithm with those of state-of-the-art algorithms: AHDRNet [12], ADNet [13], and Vien *et al.*'s algorithm [17]. We retrained these algorithms using the new training set described in Section II-C. We use three quality metrics: μ-PSNR, HDR-VDP (Q), and HDR-VDP (P) [23]. μ-PSNR is an extension of PSNR—computed in the tone-mapped domain—to consider the human perception of luminance values.

TABLE I
Quantitative comparison of synthesis performance on the test set. For each metric, the boldface value indicates the best result.

	μ-PSNR	HDR-VDP (Q)	HDR-VDP (P)
AHDRNet [12]	36.19	53.92	0.4527
ADNet [13]	36.25	**54.20**	0.4569
Vien *et al.* [17]	35.39	53.19	0.4826
Proposed	**36.29**	53.67	**0.3874**

Table I quantitatively compares the synthesis performance of the algorithms on the test set. The μ-PSNR metric measures the fidelity of the synthesized images to the ground-truths, whereas two HDR-VDP metrics measure perceptual differences between them. The proposed algorithm achieves the highest μ-PSNR score, providing a 0.04 dB higher score than the second-best ADNet. In addition, the proposed algorithm provides the best synthesis performance in terms of the P-value of HDR-VDP, indicating that the synthesized images are the most similar to the ground-truth images. This indicates that the proposed enhanced bidirectional motion estimation using feature refinement is effective for HDR image synthesis.

Fig. 4 compares the synthesized HDR images qualitatively. Because the input images contain both large motions and a large number of poorly exposed pixels, the conventional algorithms fail to synthesize high-quality HDR images, providing visible artifacts. For example, AHDRNet and ADNet in Figs. 4(d) and (e), respectively, fail to reconstruct fine textures in the saturated regions, *e.g.*, the light rays in the second row. Vien *et al.*'s algorithm in Fig. 4(f) provides noise components and color distortions in the under-exposed regions, *e.g.*, the background of the hand in the fourth row. In contrast, the proposed algorithm in Fig. 4(g) synthesizes higher-quality images by reconstructing faithful details and

TABLE II
EFFECTS OF THE MASKS M_k IN THE KERNEL ESTIMATOR AND THE MULTIPLE ENCODERS ON THE SYNTHESIS PERFORMANCE.

M_k	Multiple encoders	μ-PSNR	HDR-VDP (Q)	HDR-VDP (P)
-	-	35.52	52.75	0.4676
✓	-	35.83	52.95	0.4334
✓	✓	**36.29**	**53.67**	**0.3874**

suppressing visible artifacts.

B. Ablation Studies

We analyze the effectiveness of the masks M_k in LFB and multiple encoders of the proposed algorithm.

We train the network with different settings to analyze the effectiveness of the masks M_k in LFB in Section II-B. Table II demonstrates that the proposed kernel estimator using the masks improves the synthesis performance by effectively handling regions with motion errors or poor exposures. Table II also shows that using multiple encoders further increases the performance by large margins. This is because the respective encoders for motion estimation and refinement can better extract contextual information for HDR image synthesis.

IV. CONCLUSIONS

We developed an MEF-based HDR image synthesis algorithm based on enhanced bidirectional motion estimation using feature refinement. First, we extracted multiscale features from the input LDR images and then estimated the optical flows between the LDR images in a coarse-to-fine manner. Next, we estimated adaptive local kernels to merge only valid information in the neighboring information for HDR image synthesis. Finally, the synthesized image was further refined by global residual learning, which improves the synthesis performance. Experimental results demonstrated that the proposed algorithm outperforms state-of-the-art HDR image synthesis algorithms.

REFERENCES

[1] O. T. Tursun, A. O. Akyüz, A. Erdem, and E. Erdem, "The state of the art in HDR deghosting: A survey and evaluation," *Comput. Graph. Forum*, vol. 34, no. 2, pp. 683–707, Jun. 2015.

[2] L. Wang and K.-J. Yoon, "Deep learning for HDR imaging: State-of-the-art and future trends," *IEEE Trans. Pattern Anal. Mach. Intell.*, Nov. 2021, Early access.

[3] L. Bogoni, "Extending dynamic range of monochrome and color images through fusion," in *Proc. Int. Conf. Pattern Recognit.*, Sep. 2000, pp. 7–12.

[4] A. Tomaszewska and R. Mantiuk, "Image registration for multi-exposure high dynamic range image acquisition," in *Proc. Int. Conf. Central Europe Comput. Graph. Vis. Comput. Vis.*, Jan./Feb. 2007, pp. 49–56.

[5] C. Lee, Y. Li, and V. Monga, "Ghost-free high dynamic range imaging via rank minimization," *IEEE Signal Process. Lett.*, vol. 21, no. 9, pp. 1045–1049, Sep. 2014.

[6] T.-H. Oh, J.-Y. Lee, Y.-W. Tai, and I. S. Kweon, "Robust high dynamic range imaging by rank minimization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1219–1232, Jun. 2015.

[7] C. Lee and E. Y. Lam, "Computationally efficient truncated nuclear norm minimization for high dynamic range imaging," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4145–4157, Sep. 2016.

[8] P. Sen, N. K. Kalantari, M. Yaesoubi, S. Darabi, D. B. Goldman, and E. Shechtman, "Robust patch-based HDR reconstruction of dynamic scenes," *ACM Trans. Graph.*, vol. 31, no. 6, pp. 203:1–203:11, Nov. 2012.

[9] J. Hu, O. Gallo, K. Pulli, and X. Sun, "HDR deghosting: How to deal with saturation?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1163–1170.

[10] N. K. Kalantari and R. Ramamoorthi, "Deep high dynamic range imaging of dynamic scenes," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144:1–144:12, Jul. 2017.

[11] S. Wu, J. Xu, Y.-W. Tai, and C.-K. Tang, "Deep high dynamic range imaging with large foreground motions," in *Proc. European Conf. Comput. Vis.*, Sep. 2018, pp. 120–135.

[12] Q. Yan, D. Gong, Q. Shi, A. van den Hengel, C. Shen, I. Reid, and Y. Zhang, "Attention-guided network for ghost-free high dynamic range imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 1751–1760.

[13] Z. Liu, W. Lin, X. Li, Q. Rao, T. Jiang, M. Han, H. Fan, J. Sun, and S. Liu, "ADNet: Attention-guided deformable convolutional network for high dynamic range imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2021, pp. 463–470.

[14] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 764–773.

[15] T. T. N. Mai, E. Y. Lam, and C. Lee, "Ghost-free HDR imaging via unrolling low-rank matrix completion," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2021, pp. 2928–2932.

[16] C. Liu, J. Yuen, and A. Torralba, "SIFT Flow: Dense correspondence across scenes and its applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 978–994, May 2011.

[17] A. G. Vien, S. Park, T. T. N. Mai, G. Kim, and C. Lee, "Bidirectional motion estimation with cyclic cost volume for high dynamic range imaging," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2022, pp. 1183–1190.

[18] E. Pérez-Pellitero, S. Catley-Chandar, R. Shaw, A. Leonardis, R. Timofte *et al.*, "NTIRE 2022 challenge on high dynamic range imaging: Methods and results," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2022, pp. 1009–1023.

[19] J. Park, C. Lee, and C.-S. Kim, "Asymmetric bilateral motion estimation for video frame interpolation," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2021, pp. 14 519–14 528.

[20] L. Kong, B. Jiang, D. Luo, W. Chu, X. Huang, Y. Tai, C. Wang, and J. Yang, "IFRNet: Intermediate feature refine network for efficient frame interpolation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 1969–1978.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1026–1034.

[22] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, May 2015.

[23] R. Mantiuk, K. J. Kim, A. G. Rempel, and W. Heidrich, "HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions," *ACM Trans. Graph.*, vol. 30, no. 4, pp. 40:1–40:14, Jul. 2011.