Bidirectional Motion Estimation with Cyclic Cost Volume for High Dynamic Range Imaging

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

Abstract

We propose a high dynamic range (HDR) imaging algorithm based on bidirectional motion estimation. First, we develop a motion estimation network with the cyclic cost volume and spatial attention maps to estimate accurate optical flows between input low dynamic range (LDR) images. Then, we develop the dynamic local fusion network that combines the warped and reference inputs to generate a synthesized image by exploiting local information. Finally, to further improve the synthesis performance, we develop the global refinement network that generates a residual image by exploiting global information. Experimental results on the dataset from the NTIRE 2022 HDR Challenge Track 1 (Low-complexity constrain) demonstrate the effectiveness of the proposed HDR image synthesis algorithm.

1. Introduction

Despite recent advances in digital imaging technologies, the dynamic ranges of cameras are still limited compared with those of natural scenes. A common approach to capture scenes with a wide range of illuminance is to synthesize high dynamic range (HDR) images by merging multiple low dynamic range (LDR) images captured with varying exposure times [25]. However, camera or object motions across the LDR images may cause misalignments among the images, thereby causing ghosting artifacts in the synthesized HDR images and degrading image quality. To mitigate the impact of motions on HDR image quality, various approaches have been developed to handle misalignments [26].

Early attempts of HDR imaging exploited the properties of motions to establish mathematical models, which are categorized into three groups based on how they handle motions. The first category of algorithms computes the correspondences between the input LDR images and merges the aligned images [1, 2, 24]; however, these algorithms may fail to estimate accurate correspondences for large under-/over-exposed regions, producing ghosting artifacts. The second category of algorithms attempts to determine object movement regions [9, 10, 30]; despite their capability for handling motions in poorly exposed regions, these algorithms may fail when the scenes contain complex motions. The algorithms of the third category attempt to detect ghost regions and estimate correspondences simultaneously by solving joint optimization problems [5, 15, 21]; however, they require high computational costs for correspondence estimation and numerical optimization.

Recently, deep learning-based algorithms have been actively developed for HDR imaging [6, 11–14, 20, 27, 28]. They have shown superior performance against model-based algorithms thanks to the capabilities of convolutional neural networks (CNNs) in restoring complex textures and details by learning abstract visual features from large data. The deep learning-based algorithms are constructed on the basis of a common principle: the input LDR images are transformed to the feature domain by CNNs; then, the feature maps are merged and transformed back to the image domain by another CNN. Despite their superior performance, deep learning-based algorithms often require large computational resources because of their large numbers of computing operations, thus limiting their versatility. Therefore, developing computationally efficient algorithms while retaining the synthesis performance is essential.

In this work, we develop a lightweight deep network for HDR imaging based on bidirectional motion estimation consisting of the motion estimation network (MENet), dynamic local fusion network (DLFNet), and global refinement network (GRNet). First, MENet predicts optical flows between the reference and target LDR images. To enhance the accuracy of optical flows, we develop the cyclic cost volume with spatial attention maps. Second, we extract multiscale feature maps to exploit contextual information and then warp the target images and their corresponding feature maps to those of the reference image using the estimated optical flows. Next, DLFNet combines the warped results with the reference image and its feature maps to obtain synthesized outputs by exploiting local neighboring information. Finally, GRNet refines the outputs of DLFNet by exploit-

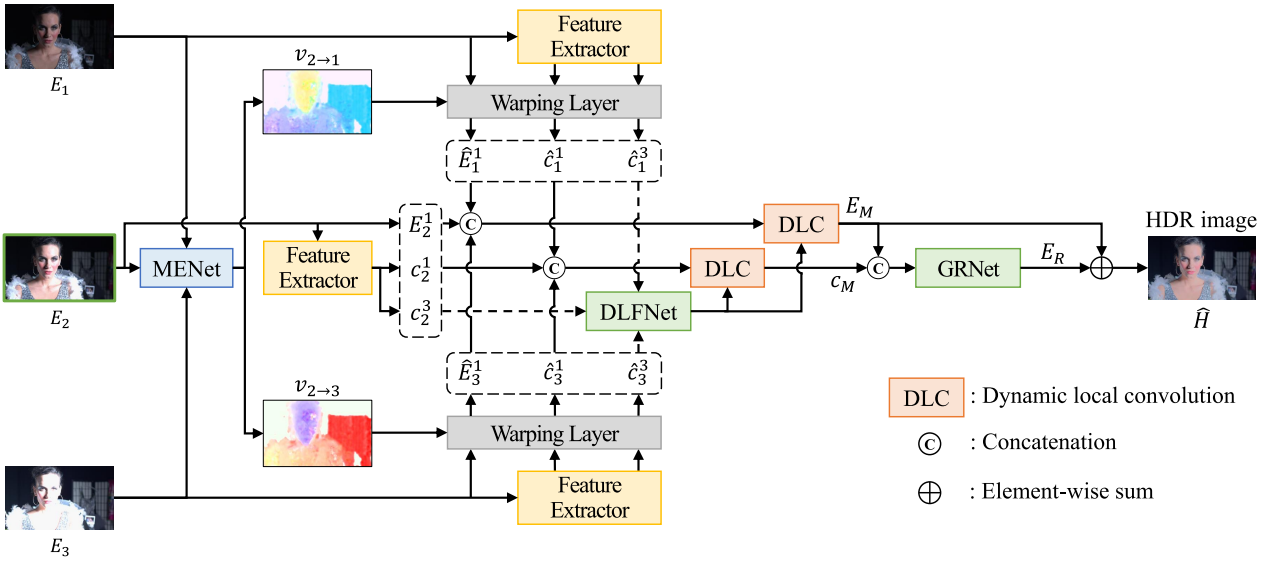

Figure 1. Overview of the proposed algorithm. MENet, DLFNet, and GRNet are detailed in Figures 2, 3, and 4, respectively.

ing global information. We demonstrate the effectiveness of the proposed algorithm on the NTIRE 2022 HDR Imaging Challenge dataset [19].

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed algorithm. Section 4 discusses the experimental results. Finally, Section 5 concludes the paper.

2. Related Work

Model-based HDR imaging algorithms formulate multi-exposure fusion problems as mathematical models that are categorized into three subgroups based on how they handle motions. The first category of algorithms computes correspondences between the input LDR images and merges the aligned images. For example, Bogoni [1] employed optical flows estimated from the input images. Tomaszewska and Mantiuk [24] searched for feature points in the LDR images and used them for global alignment. Instead of computing correspondences between individual images, Gupta *et al.* [2] estimated them between two synthesized images obtained by summing successive LDR images. Algorithms in this category may fail to estimate correspondences accurately when the reference image has large poorly exposed regions. The second category of algorithms attempts to determine the object movement regions to alleviate their contributions. Zimmer *et al.* [30] integrated optical flow estimation into the energy minimization model for image fusion. Lee *et al.* [10] incorporated constraints of sparsity and connectivity across exposures on moving objects into the rank minimization model. Lee and Lam [9] formulated HDR imaging as an optimization problem by exploiting the low-rankness of irradiance maps from LDR images and the sparseness of moving objects. Despite specifically addressing movement regions, these algorithms may fail when the scene contains complex motions. The algorithms of the third category simultaneously detect ghost regions and estimate correspondences via joint optimization. For example, Sen *et al.* [21] integrated alignment and reconstruction into patch-based energy minimization. Hu *et al.* [5] exploited radiance and texture consistencies to align LDR images. Oh *et al.* [15] also exploited the low-rankness of LDR images by integrating alignment with the robust principal component analysis model. These algorithms often require high computational resources for correspondence estimation and numerical optimization.

Deep learning-based HDR imaging algorithms have been developed actively and show superior performances against model-based algorithms. Kalantari and Ramamoorthi [6] aligned target images using optical flows and developed two networks—one for refining the aligned images and the other for merging. Lee *et al.* [11] employed different encoder-decoder architectures for alignment and merge. Yan *et al.* [28] developed a joint alignment-fusion network that can suppress undesired information by exploiting hierarchical features. Prabhakar *et al.* [20] proposed an efficient algorithm that synthesizes an HDR image in a low resolution then restores the full resolution. To avoid pre-alignment, Wu *et al.* [27] performed LDR image alignment and HDR image synthesis via a single encoder-decoder network. Niu *et al.* [14] employed a generative adversarial network to synthesize missing regions caused by occlusions. Mai *et al.* [13] formulated the HDR imaging problem as a rank minimization model and developed an unrolling approach to leverage the strength of deep learning. In [12], Liu

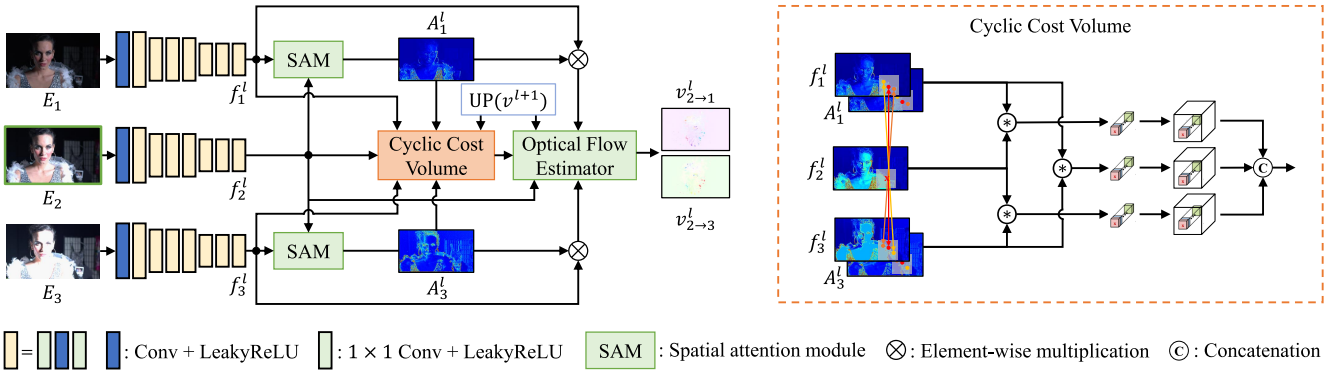

Figure 2. Architecture of MENet and illustration of the cyclic cost volume layer.

et al. aligned the input images in the feature domain then determined the best regions for fusion. Despite their superior performances, deep learning-based algorithms generally demand high computational resources because of their significant numbers of operations. In this work, we develop a lightweight yet effective joint alignment-fusion network for HDR imaging based on bidirectional motion estimation with cyclic cost volume.

3. Proposed Algorithm

Given three input LDR images $\{I_1, I_2, I_3\}$ of $H \times W \times 3$, where H and W are the height and width of the images, respectively, captured with three different exposures, we recover an HDR image aligned to a reference image I_2. First, the input LDR images $\{I_1, I_2, I_3\}$ are linearized to obtain images $\{E_1, E_2, E_3\}$ using the gamma function as the camera response function

$$E_i = \frac{(I_i)^\gamma}{\Delta t_i}, \tag{1}$$

where $i \in \{1, 2, 3\}$ denotes the exposure index, Δt_i is the exposure time for I_i, and $\gamma = 2.24$ is the gamma parameter.

Figure 1 shows an overview of the proposed algorithm, which takes the linearized images $\{E_1, E_2, E_3\}$ as input and reconstructs an HDR image $\widehat{H}$. The proposed algorithm is composed of three subnetworks: MENet, DLFNet, and GRNet. MENet estimates two motion fields between E_2 and E_1 and between E_2 and E_3. Then, the target images E_1 and E_3 are warped using the estimated motion fields to the reference image E_2. Next, DLFNet merges these warped images and reference image by learning the dynamic local filters. Finally, GRNet improves the synthesis performance by learning a residual image to refine the filtered image.

3.1. Motion Estimation with Cyclic Cost Volume

MENet: We estimate two motion vectors $v_{2\to1}(\mathbf{x})$ and $v_{2\to3}(\mathbf{x})$ between E_2 and E_1 and between E_2 and E_3, respectively, at each pixel location $\mathbf{x}$ in E_2. Assuming linear motion between the exposures, we have

$$v_{2\to1}(\mathbf{x}) = -v_{2\to3}(\mathbf{x}). \tag{2}$$

However, since the input images contain poorly exposed regions, the linear constraint is invalid in these regions. To address this issue, we develop MENet, which estimates $v_{2\to1}(\mathbf{x})$ and $v_{2\to3}(\mathbf{x})$, and cyclic cost volume to handle poorly exposed regions. Figure 2 shows the architecture of MENet employing PWC-Net [23] as the baseline with modification for bidirectional motion estimation using the proposed cyclic cost volume. Specifically, multiscale feature maps f_i^l are first generated from each input image E_i via a feature pyramid extractor, where i and l denote the exposure and pyramid level indices, respectively. At each level l, attention maps A_1^l and A_3^l are estimated for the target images, which identify motion and poorly exposed regions, using the spatial attention module (SAM) [28]. The estimated attention maps are multiplied with the corresponding target images to enhance the useful features and exclude motion and poorly exposed regions. Next, we use three feature maps extracted from the three input images with their corresponding attention maps to build a cyclic cost volume that stores two bidirectional costs and a cyclic matching cost to handle occlusions or poorly exposed regions. Finally, motion vector fields at level l are estimated using an optical flow estimator. Each component of MENet is described in detail below.

Feature extractor: We employ a feature extractor in PWC-Net [23] for each input image to extract 3-level feature pyramids f_i^l, where $i \in \{1, 2, 3\}$ and $l \in \{1, 2, 3\}$. Specifically, the numbers of feature channels are 16, 32, and 64 from the first to third levels, respectively. Further, we use a convolution layer with a 3×3 kernel and a stride of 2 to downsample f_i^l to the $(l+1)$th level.

SAM: This generates attention maps to highlight useful information to the reference image as well as exclude motion and poorly exposed regions. Specifically, each target feature map at the lth level is concatenated with the reference

feature map f_2^l as input to SAM [28], which consists of two convolution layers and a sigmoid function to generate the attention map A_i^l in the range of $[0, 1]$. Then, the spatially attenuated feature map $\tilde{f}_i^l$ is computed as

$$\tilde{f}_i^l = f_i^l \otimes A_i^l, \tag{3}$$

where $\otimes$ denotes the element-wise multiplication.

Cyclic cost volume: A cost volume records the matching costs between a pixel in a reference image with its corresponding pixels in a target image [4]. However, unlike the conventional cost volume [16, 23] that uses two input images, we compute a cost volume from three input LDR images—a single reference and two target images. Furthermore, the input images contain invalid pixels caused by under-/over-exposures as well as occlusions. This is especially problematic for the reference image, especially alignment to invalid pixels. To address this issue, we develop the cyclic cost volume for bidirectional motion estimation, where the cyclic matching cost between pixels in the two target images is useful when a pixel in the reference image is invalid. In addition, the matching cost in the proposed cyclic cost volume is computed with the spatial attention maps, which identify motion and poorly exposed regions.

Figure 2 illustrates the proposed cyclic cost volume generation that takes feature maps f_1^l, f_2^l, and f_3^l of the three input images, upsampled motion fields $\hat{v}_{2\to1}^l = \mathrm{UP}(v_{2\to1}^{l+1})$ and $\hat{v}_{2\to3}^l = \mathrm{UP}(v_{2\to3}^{l+1})$ estimated at the $(l+1)$th level, and estimated attention maps A_1^l and A_3^l. Here, UP is the upsampling operator using bilinear interpolation [23]. Let $\mathbf{x}$ be a pixel location in the reference image E_2^l. Then, we define the matching costs $CV_{21}^l(\mathbf{x}, \mathbf{d})$ and $CV_{23}^l(\mathbf{x}, \mathbf{d})$ as the bidirectional correlation between feature maps $\{f_2^l, f_1^l\}$ and $\{f_2^l, f_3^l\}$ with their corresponding spatial attention maps $\{A_1^l, A_3^l\}$ in (4) and (5), respectively, at the bottom of the page, where $\mathbf{d}$ denotes the displacement vector within the search window $\mathcal{D} = [-d, d] \times [-d, d]$. Note that the reference image I_2 may contain occlusions or poorly exposed regions; in such cases, the matching costs are invalid, leading to inaccurate motion estimation. Therefore, in this work, in addition to the two bidirectional costs, we further define the cyclic matching cost between features indexed by the motion vector that passes through $\mathbf{x}$ in E_2 and both spatial attention maps $\{A_1^l, A_3^l\}$, given by (6) at the bottom of the page. The dimension of the cyclic cost volume at level l is $W^l \times H^l \times D^2 \times 3$, where W^l and H^l denote the width and height, respectively, of the feature maps at level l, $D = 2d + 1$, and 3 is the number of matching costs.

Optical flow estimator: We follow a common approach [28] to optical flow estimation that refines optical flows by exploiting contextual information from features. In particular, the optical flow estimator is implemented as a multilayer CNN. More specifically, at each level l, the cyclic cost volume, upsampled motion fields $\hat{v}_{2\to1}^l$ and $\hat{v}_{2\to3}^l$, feature map of the reference image f_2^l, and two spatially attentive feature maps $\tilde{f}_1^l$ and $\tilde{f}_3^l$ are used as input to generate $v_{2\to1}^l$ and $v_{2\to3}^l$.

However, since the optical flow estimator uses the same network architecture for different levels, it takes up most parameters and computational costs. To facilitate a computational complexity versus estimation accuracy trade-off, we employ a shuffle block decoder (SBD) [8] as an optical flow estimator. The SBD reforms standard convolutions as group convolutions by channel shuffle operations [29], which can effectively reduce computational costs while maintaining accuracy.

3.2. HDR Image Synthesis

In Figure 1, we synthesize an HDR image by merging three images, $\widehat{E}_1$, E_2, and $\widehat{E}_3$, two of which are warped by the warping layers. As the contextual information in the input images improves synthesis performance, in addition to the input images, we exploit contextual information as in [16]. Specifically, we extract multiscale feature maps c_i^l as contextual information in the input images using the feature extractor in Figure 1, which has the same architecture as that of the feature pyramid extractor in MENet. Furthermore, to reduce computational complexity, we only use the features at levels $l \in \{1, 3\}$. The feature extraction from the input frames is performed with shared parameters. Then, the inputs E_1 and E_3 are warped using $v_{2\to1}$ and $v_{2\to3}$ to generate the estimates $\widehat{E}_1$ and $\widehat{E}_3$, respectively. Similarly, the feature maps c_1^l and c_3^l are warped to generate $\hat{c}_1^l$ and $\hat{c}_3^l$ for $l \in \{1, 3\}$, respectively. These warped images and feature maps are merged to synthesize the HDR image. In this work, we develop two subnetworks—DLFNet and GRNet—to exploit local and global information, respectively.

Warping layer: Note that the HDR image is synthesized

$$CV_{21}^l(\mathbf{x}, \mathbf{d}) = f_2^l(\mathbf{x})^T \left[f_1^l\left(\mathbf{x} + \hat{v}_{2\to1}(\mathbf{x}) - \mathbf{d}\right) \times A_1^l\left(\mathbf{x} + \hat{v}_{2\to1}(\mathbf{x}) - \mathbf{d}\right) \right] \tag{4}$$

$$CV_{23}^l(\mathbf{x}, \mathbf{d}) = f_2^l(\mathbf{x})^T \left[f_3^l\left(\mathbf{x} + \hat{v}_{2\to3}(\mathbf{x}) + \mathbf{d}\right) \times A_3^l\left(\mathbf{x} + \hat{v}_{2\to3}(\mathbf{x}) + \mathbf{d}\right) \right] \tag{5}$$

$$CV_{31}^l(\mathbf{x}, \mathbf{d}) = \left[f_3^l\left(\mathbf{x} + \hat{v}_{2\to3}(\mathbf{x}) + \mathbf{d}\right) \times A_3^l\left(\mathbf{x} + \hat{v}_{2\to3}(\mathbf{x}) + \mathbf{d}\right) \right]^T \left[f_1^l\left(\mathbf{x} + \hat{v}_{2\to1}(\mathbf{x}) - \mathbf{d}\right) \times A_1^l\left(\mathbf{x} + \hat{v}_{2\to1}(\mathbf{x}) - \mathbf{d}\right) \right] \tag{6}$$

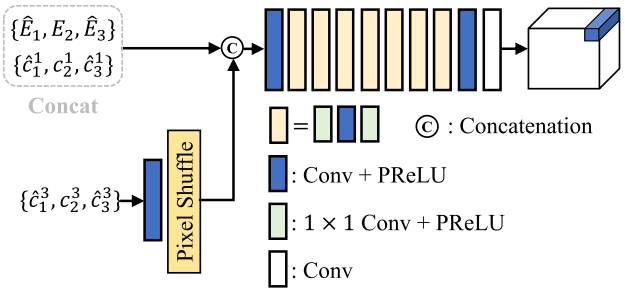

Figure 3. Architecture of DLFNet.

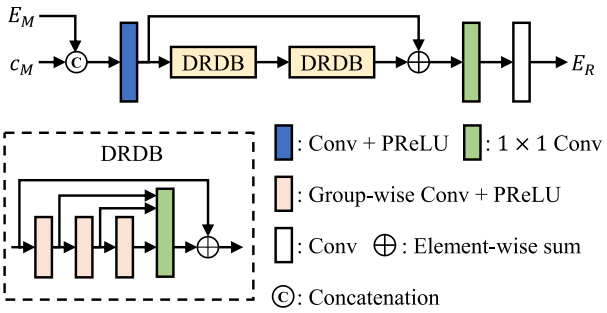

Figure 4. Architecture of GRNet.

by merging the warped frames obtained by the estimated motion vector fields. In this work, we warp both the input images E_1 and E_3 and their multiscale features c_1^l and c_3^l toward the reference image using the upsampled motion fields from the previous level. Specifically, the warped feature $\hat{c}_i^l(\mathbf{x})$ at $\mathbf{x}$ is given by

$$\hat{c}_i^l(\mathbf{x}) = c_i^l\left(\mathbf{x} + \mathrm{UP}(v_{2\to i}^{l+1})(\mathbf{x})\right), \tag{7}$$

where $i \in \{1, 3\}$.

DLFNet: This takes the warped and reference images as well as their corresponding feature maps and learns to generate dynamic local filters for merging three images $\{\widehat{E}_1, E_2, \widehat{E}_3\}$. Figure 3 shows the architecture of DLFNet. We first $4\times$ upsample the coarsest feature maps $\{\hat{c}_1^3, c_2^3, \hat{c}_3^3\}$ via a single convolution layer with a pixel shuffle operation [22] and then concatenate the upsampled feature maps with the inputs $\{\widehat{E}_1, E_2, \widehat{E}_3\}$ and their corresponding feature maps $\{\hat{c}_1^1, c_2^1, \hat{c}_3^1\}$. To effectively learn dynamic filters from multiple inputs, we modify GridNet in [17] using only 10 convolution layers. In addition, to reduce the computational complexity, based on [3], we replace each convolution layer with a sequence of three convolutions with filters of sizes 1×1, 3×3, and 1×1. For the first and last two convolution layers in DLFNet, we use a single convolution layer with a filter of 3×3. For each pixel (x, y), the last convolution layer of DLFNet generates three local filter coefficients $K_{x,y}(i, j, k)$, where (i, j) are local coordinates around (x, y) and $k \in \{1, 2, 3\}$ indexes the input image, to fuse 3×3 local neighboring pixels in $\{\widehat{E}_1, E_2, \widehat{E}_3\}$. The filter coefficients are normalized as $\sum_i \sum_j \sum_k K_{x,y}(i, j, k) = 1$.

Then, we obtain the synthesized HDR image E_M via dynamic local convolution (DLC) with the learned filter coefficients $K_{x,y}(i, j, k)$ as

$$E_M(x, y) = \sum_{k=1}^{3} \sum_{i=-1}^{1} \sum_{j=-1}^{1} K_{x,y}(i, j, k) \widehat{E}_k(x + i, y + j), \tag{8}$$

where $\widehat{E}_2 = E_2$. Similarly, by applying the same filters to $\{\hat{c}_1^1, c_2^1, \hat{c}_3^1\}$, we obtain the synthesized feature map c_M.

GRNet: DLFNet merges multiple inputs by considering only local neighbors. Thus, if the local neighbors do not contain valid information due to motion errors or poorly exposed regions, DLFNet may fail to produce valid information in these regions. To improve the synthesis performance using global information, we develop GRNet, as shown in Figure 4, to generate a residual output E_R to refine E_M. GRNet takes the synthesized feature maps c_M and images E_M as input. Two dilated residual dense blocks (DRDBs) [28] are employed to faithfully exploit global information and to increase the receptive field. In each DRDB, instead of convolution layers with a 3×3 kernel, we adopt group-wise convolutions [29] to reduce computational complexity. Finally, the synthesized HDR image is given as $\widehat{H} = E_M + E_R$.

3.3. Implementation Details

Loss function: We define the HDR reconstruction loss $\mathcal{L}_r$ as the L_2-norm of the difference between the output $\widehat{H}$ and ground-truth H_{gt}. Because of its higher dynamic range, the HDR reconstruction loss is defined in the tone-mapped domain using the μ-law function $\mathcal{T}$ [6] as

$$\mathcal{L}_r = \|\mathcal{T}(\widehat{H}) - \mathcal{T}(H_{\mathrm{gt}})\|_2^2, \tag{9}$$

with

$$\mathcal{T}(x) = \frac{\log(1 + \mu x)}{\log(1 + \mu)}, \tag{10}$$

where μ is a parameter to control the amount of compression. In this work, we set $\mu = 5000$.

Dataset: We use the training dataset provided by the NTIRE 2022 HDR Imaging Challenge [19]. Because ground-truth images are not provided for testing, we randomly selected 210 HDR images from the training set for test. Thus, the new training set contains 1,284 images out of 1,494. We generate training patches by cropping 128×128 patches with a stride of 128.

Training: We use the Adam optimizer [7] using $\beta_1 = 0.9$ and $\beta_2 = 0.999$ for 200 epochs with an initial learning rate of 10^{-4} and a decay rate of 0.1. The proposed algorithm

Table 1. Quantitative comparison of the results on the NTIRE2022 High Dynamic Range Challenge dataset. For each metric, the boldface value indicates the best result.

	PSNR	μ-PSNR	Runtime (s)	# GMACCs	# Parameters
ADNet	**43.58**	**36.82**	2.34	6238.59	2,808,772
Proposed	39.44	35.39	**0.32**	**199.11**	**1,300,602**

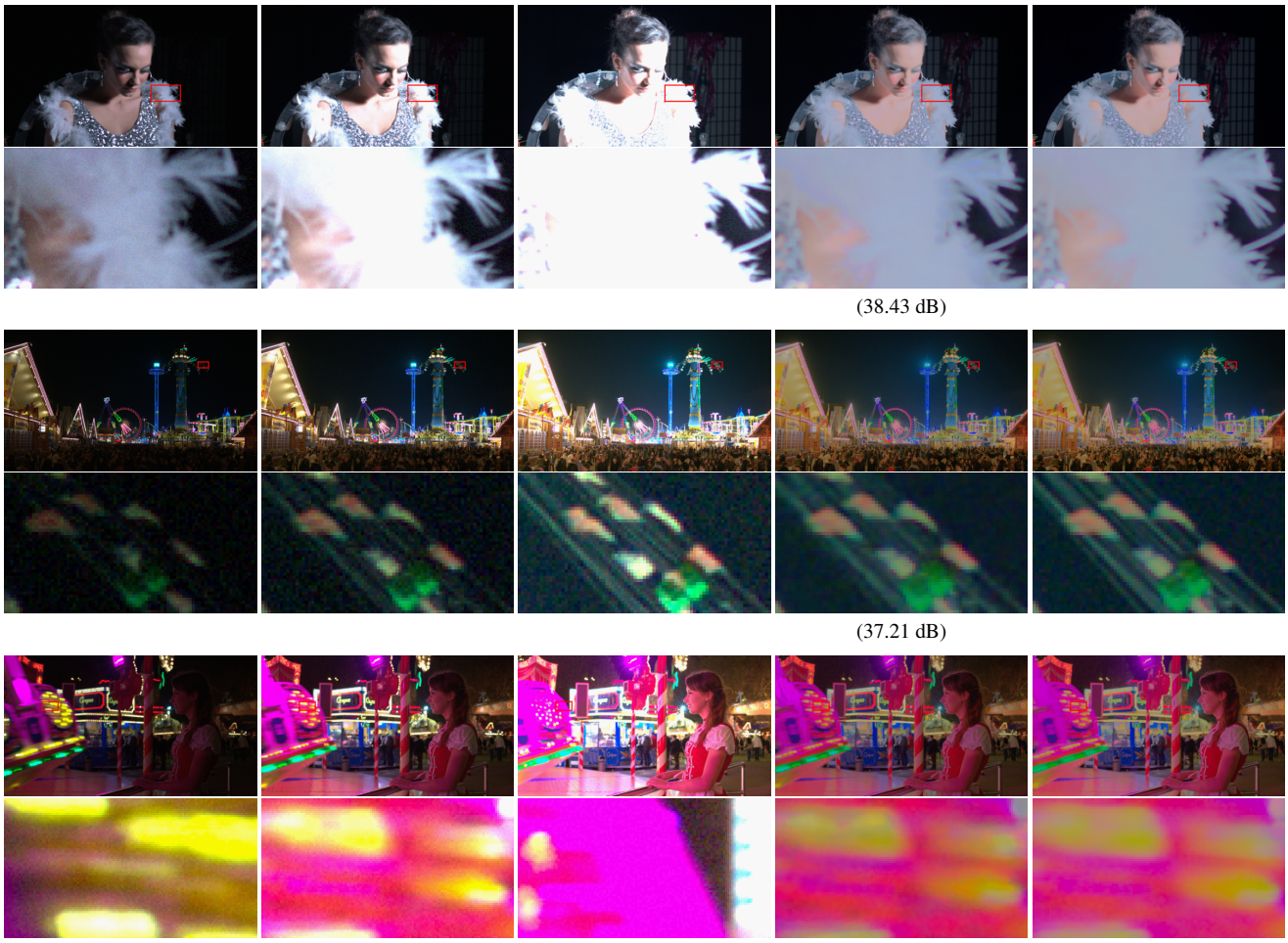

Figure 5. Qualitative comparison of the synthesized HDR images. (Top) Full-resolution images and (bottom) magnified parts. μ-PSNR scores are provided below each synthesized image.

is implemented using PyTorch [18]. The training took approximately three days using a PC with an Intel® Core™ i9-7900X @3.30GHz CPU, 64GB RAM, and four Nvidia RTX™ 3090 GPUs.

4. Experimental Results

4.1. Quantitative and Qualitative Evaluation

We compare the HDR image synthesis performance of the proposed algorithm with that of the state-of-the-art ADNet [12]. For fair comparison, we retrain ADNet using the training set with the same settings as the proposed algorithm.

Table 1 compares the synthesis performance quantita-

Table 2. Impacts of the cyclic cost volume and GRNet on the HDR image synthesis performance.

Cyclic cost volume	GRNet	PSNR	μ-PSNR
	✓	39.21	35.17
✓		34.89	31.13
✓	✓	**39.44**	**35.39**

Table 3. Analysis of computational complexity of each component in the proposed algorithm.

	MENet	Feature extractor	DLFNet	GRNet
# GMACCs	49.67	33.68	102.85	12.91

tively on 210 images in the test set described in Section 3.3 using the PSNR metric and its extension μ-PSNR computed in the tone-mapped domain. PSNR measures the accuracy of luminance value reconstruction, whereas μ-PSNR considers the human perception to the luminance values. Table 1 also compares the computational complexity in terms of the average execution times, the number of giga multiply-accumulate (GMACC) operations to process images of the resolution 1900×1060, and the number of network parameters. Even though the proposed algorithm yields 1.43 dB lower μ-PSNR score than ADNet, the proposed algorithm provides significantly higher computational efficiency than ADNet. Specifically, the proposed algorithm runs $7\times$ faster and requires $31\times$ and $2\times$ less GMACCs and parameters, respectively, than ADNet.

Figure 5 visually compares the synthesized HDR images. Although there are either camera motions or object motions across the three inputs, the proposed algorithm provides high-quality HDR images. For example, two input images I_2 and I_3 in the first row contain a large number of over-exposed pixels in the motion regions. The proposed algorithm restores textures faithfully without ghosting artifacts in these regions. In addition, even when the input images contain large motions and object deformation, where the linear motion constraint is invalid, in the last row, the proposed algorithm recovers the details in these regions with only a small amount of visible artifacts. This confirms the effectiveness of the cyclic cost volume with spatial attention maps.

4.2. Model Analysis

We conduct several ablation studies to analyze the contributions of the key components in the proposed algorithm: cyclic cost volume and GRNet. We also analyze the computational cost of each component by comparing GMACCs.

Cyclic cost volume: To analyze the effectiveness of the proposed cyclic cost volume, we train the proposed algorithm with the conventional cost volume [23]. Table 2 compares the average scores of different settings. The proposed cyclic cost volume provides higher scores than the conventional cost volume. This indicates that the cyclic cost volume with spatial attention maps is essential to handle motions in poorly exposed regions.

GRNet: We analyze the effectiveness of GRNet by training the proposed algorithm with and without GRNet. Table 2 compares the results. Using GRNet increases the scores with significantly large margins. This is because the output image is synthesized using only local neighboring information, which may contain invalid information from poorly exposed regions, to degrade the synthesis performance. In contrast, GRNet exploits global information, thereby improving the synthesis performance.

Computational cost: Finally, to analyze the computational complexity of each component in the proposed algorithm, we compare the number of GMACCs for MENet, feature extractor, DLFNet, and GRNet. Table 3 compares the results. First, because both MENet and the feature extractor generate 3-level pyramids of feature maps, they produce high computational costs. DLFNet combines multiple images and feature maps by learning local filters for each pixel; it requires the highest computational resource. Finally, GRNet processes the inputs without either multiscale feature representations or local filters; it provides the greatest efficiency of computational cost.

5. Conclusions

We developed a lightweight joint alignment-fusion algorithm for HDR imaging based on bidirectional motion estimation. First, in MENet, we developed the cyclic cost volume with spatial attention maps to predict optical flows from the reference image to the target images. Then, we warped the target images and their corresponding feature maps using the estimated optical flows and fed them to DLFNet. Subsequently, DLFNet learns local filter coefficients to generate a synthesized output. Finally, GRNet refines the synthesized output by exploiting global information. The experimental results on the NTIRE 2022 HDR Imaging Challenge demonstrated that the proposed algorithm can synthesize high-quality HDR images with significantly less demands on computational resources compared with the state-of-the-art algorithm.

Acknowledgements

This work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00011, Video Coding for Machine).

References

[1] Luca Bogoni. Extending dynamic range of monochrome and color images through fusion. In *Proc. ICPR*, pages 7–12, Sep. 2000. 1, 2

[2] Mohit Gupta, Daisuke Iso, and Shree K. Nayar. Fibonacci exposure bracketing for high dynamic range imaging. In *Proc. ICCV*, pages 1473–1480, Dec. 2013. 1, 2

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proc. ICCV*, pages 1026–1034, Dec. 2015. 5

[4] Asmaa Hosni, Christoph Rhemann, Michael Bleyer, Carsten Rother, and Margrit Gelautz. Fast cost-volume filtering for visual correspondence and beyond. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(2):504–511, Feb. 2013. 4

[5] Jun Hu, Orazio Gallo, Kari Pulli, and Xiaobai Sun. HDR deghosting: How to deal with saturation? In *Proc. CVPR*, pages 1163–1170, Jun. 2013. 1, 2

[6] Nima Khademi Kalantari and Ravi Ramamoorthi. Deep high dynamic range imaging of dynamic scenes. *ACM Trans. Graph.*, 36(4):144:1–144:12, Jul. 2017. 1, 2, 5

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, May 2015. 5

[8] Lingtong Kong, Chunhua Shen, and Jie Yang. FastFlowNet: A lightweight network for fast optical flow estimation. In *Proc. ICRA*, pages 10310–10316, May/Jun. 2021. 4

[9] Chul Lee and Edmund Y. Lam. Computationally efficient truncated nuclear norm minimization for high dynamic range imaging. *IEEE Trans. Image Process.*, 25(9):4145–4157, Sep. 2016. 1, 2

[10] Chul Lee, Yuelong Li, and Vishal Monga. Ghost-free high dynamic range imaging via rank minimization. *IEEE Signal Process. Lett.*, 21(9):1045–1049, Sep. 2014. 1, 2

[11] Sang-Hoon Lee, Haesoo Chung, and Nam Ik Cho. Exposure-structure blending network for high dynamic range imaging of dynamic scenes. *IEEE Access*, 8:117428–117438, Jun. 2020. 1, 2

[12] Zhen Liu, Wenjie Lin, Xinpeng Li, Qing Rao, Ting Jiang, Mingyan Han, Haoqiang Fan, Jian Sun, and Shuaicheng Liu. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. In *Proc. CVPRW*, pages 463–470, Jun. 2021. 1, 2, 6

[13] Truong Thanh Nhat Mai, Edmund Y. Lam, and Chul Lee. Ghost-free HDR imaging via unrolling low-rank matrix completion. In *Proc. ICIP*, pages 2928–2932, Sep. 2021. 1, 2

[14] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson W. H. Lau. HDR-GAN: HDR image reconstruction from multi-exposed LDR images with large motions. *IEEE Trans. Image Process.*, 30:3885–3896, Mar. 2021. 1, 2

[15] Tae-Hyun Oh, Joon-Young Lee, Yu-Wing Tai, and In So Kweon. Robust high dynamic range imaging by rank minimization. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(6):1219–1232, Jun. 2015. 1, 2

[16] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Proc. ECCV*, page 109–125, Aug. 2020. 4

[17] Junheum Park, Chul Lee, and Chang-Su Kim. Asymmetric bilateral motion estimation for video frame interpolation. In *Proc. ICCV*, pages 14519–14528, Oct. 2021. 5

[18] Adam Paszke et al. PyTorch: An imperative style, high-performance deep learning library. In *Proc. NeurIPS*, pages 8026–8037, Dec. 2019. 6

[19] Eduardo Pérez-Pellitero, Sibi Catley-Chandar, Richard Shaw, Ales Leonardis, Radu Timofte, et al. NTIRE 2022 challenge on high dynamic range imaging: Methods and results. In *Proc. CVPRW*, Jun. 2022. 2, 5

[20] K. Ram Prabhakar, Susmit Agrawal, Durgesh Kumar Singh, Balraj Ashwath, and R. Venkatesh Babu. Towards practical and efficient high-resolution HDR deghosting with CNN. In *Proc. ECCV*, pages 497–513, Aug. 2020. 1, 2

[21] Pradeep Sen, Nima Khademi Kalantari, Maziar Yaesoubi, Soheil Darabi, Dan B Goldman, and Eli Shechtman. Robust patch-based HDR reconstruction of dynamic scenes. *ACM Trans. Graph.*, 31(6):203:1–203:11, Nov. 2012. 1, 2

[22] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P. Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proc. CVPR*, pages 1874–1883, Jun. 2016. 5

[23] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *Proc. CVPR*, pages 8934–8943, Jun. 2018. 3, 4, 7

[24] Anna Tomaszewska and Radoslaw Mantiuk. Image registration for multi-exposure high dynamic range image acquisition. In *Proc. WSCG*, pages 49–56, Jan./Feb. 2007. 1, 2

[25] Okan Tarhan Tursun, Ahmet Oğuz Akyüz, Aykut Erdem, and Erkut Erdem. The state of the art in HDR deghosting: A survey and evaluation. *Comput. Graph. Forum*, 34(2):683–707, Jun. 2015. 1

[26] Lin Wang and Kuk-Jin Yoon. Deep learning for HDR imaging: State-of-the-art and future trends. *IEEE Trans. Pattern Anal. Mach. Intell.*, Nov. 2021. Early access. 1

[27] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang. Deep high dynamic range imaging with large foreground motions. In *Proc. ECCV*, pages 120–135, Sep. 2018. 1, 2

[28] Qingsen Yan, Dong Gong, Qinfeng Shi, Anton van den Hengel, Chunhua Shen, Ian Reid, and Yanning Zhang. Attention-guided network for ghost-free high dynamic range imaging. In *Proc. CVPR*, pages 1751–1760, Jun. 2019. 1, 2, 3, 4, 5

[29] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. ShuffleNet: An extremely efficient convolutional neural network for mobile devices. In *Proc. CVPR*, pages 6848–6856, Jun. 2018. 4, 5

[30] Henning Zimmer, Andres Bruhn, and Joachim Weickert. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *Comput. Graph. Forum*, 30(2):405–414, Apr. 2011. 1, 2