DISENTANGLED FEATURE-GUIDED MULTI-EXPOSURE HIGH DYNAMIC RANGE IMAGING

Keuntek Lee, Yeong Il Jang, Nam Ik Cho

Dept. of Electrical & Computer Eng., INMC, Seoul National University, Seoul, Korea.

ABSTRACT

Multi-exposure high dynamic range (HDR) imaging aims to generate an HDR image from multiple differently exposed low dynamic range (LDR) images. It is a challenging task due to two major problems: (1) there are usually misalignments among the input LDR images, and (2) LDR images often have incomplete information due to under-/over-exposure. In this paper, we propose a disentangled feature-guided HDR network (DFGNet) to alleviate the above-stated problems. Specifically, we first extract and disentangle exposure features and spatial features of input LDR images. Then, we process these features through the proposed DFG modules, which produce a high-quality HDR image. Experiments show that the proposed DFGNet achieves outstanding performance on a benchmark dataset. Our code and more results are available at `https://github.com/KeuntekLee/DFGNet`.

Index Terms— High dynamic range imaging, multi-exposed imaging, convolutional neural network, feature disentanglement.

1. INTRODUCTION

While the human visual system (HVS) perceives scenes with high dynamic luminance ranges adaptively, standard digital cameras generally have narrower dynamic ranges than the HVS. A common approach to capture HDR scenes with such traditional cameras is to take the scenes with several different exposures and then merge them into an HDR image, which is called multi-exposure HDR imaging. There are two significant problems in this approach, which have long been addressed in the literature. One is the misalignment between LDR inputs, which leads to ghosting artifacts on the reconstruction results. The other is the insufficient image information of LDR inputs due to their saturated regions, especially for the LDR images taken with short or long exposure times.

Meanwhile, deep convolutional neural networks (CNN) have improved the performance of various computer vision tasks, including HDR imaging [1–6]. In developing the CNN-based HDR imaging, the researchers mainly focused on exploiting the common structures between the LDR inputs through the explicit image alignment or implicit feature attention.

Regarding the misalignment problem, conventional methods (before CNN-based approaches) attempted to align the LDR images before the merging process to alleviate the ghost artifacts [7–11]. For example, the optical flow has been adopted in [10, 11] to align the pixels between the LDR images. The early CNN-based method also adopted the optical flow as a pre-processing step. For example, [1] aligned the inputs by optical flow and then forwarded them to a merging network. Afterward, researchers focused on designing neural network structures that implicitly align LDR images in the feature space [2–6]. For example, AHDRNet [2] proposed an attention-guided deep neural network that learns the structural relationships between input LDR images and HDR output. It generates soft attention maps to measure the importance of different image regions for producing HDR images. NHDRRNet [3] constructs non-local blocks [12] in the merging process to obtain global features from aggregated LDR image features. By fusing the global and local features, it effectively merges LDR images at the feature level and achieves better-aligned results than the optical flow-based approaches. On the contrary to the misalignment problem, exposure information has not been well addressed in the existing works, which can enhance the resulting HDR quality when appropriately utilized.

To address the above-stated major issues in HDR imaging, *i.e.,* aligning the structure and exploiting the exposure information, we propose a disentangle network that extracts representative exposure features and spatial features separately from LDR images. Further, we design a disentangled feature-guided network (DFGNet), which consists of DFG modules that align LDR image features with the guidance of disentangled spatial and exposure features. We exploit the disentangled feature-aware attentions during the merge of LDR image features, which helps the network extract more image-specific exclusive features from each LDR image.

2. PROPOSED METHOD

2.1. Disentangle Network for Feature Extraction

Our disentangle network extracts global exposure features and local spatial features from LDR images. The extracted features have representative attributes of input LDR image,

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

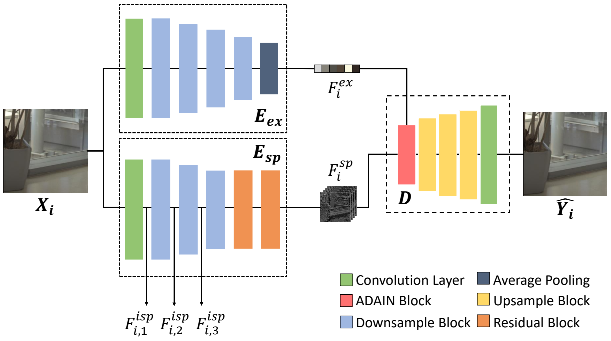

Fig. 1. Overall architecture of disentangle network.

and thus can be utilized effectively in the reconstruction step. As shown in Fig. 1, disentangle network consists of two encoders E_{ex}, E_{sp}, and one decoder D, where E_{ex} extracts the global exposure information F_i^{ex} from given $(H \times W \times 3)$ RGB input images $X_i, i = 1, 2, 3$. More precisely, it consists of 4 downsample convolutional blocks and an adaptive average pooling. E_{sp} has a similar architecture to E_{ex} but has 3 downsample convolutional blocks and 2 residual blocks [13] without pooling. Consequently, E_{sp} generates $(\frac{H}{8} \times \frac{W}{8} \times C_s)$ spatial feature F_i^{sp}, and E_{ex} generates $(1 \times 1 \times C_e)$ exposure feature vector. Decoder D is symmetric to the architecture of E_{ex}, which consist of 4 bilinear upsample convolutional blocks and AdaIN [14] module. The AdaIN module applies the encoded style information to image content, showing impressive performance on style transfer tasks. With the AdaIN module, our encoder-decoder structured disentangle network reconstructs input image effectively with given E_{ex} and E_{sp}. For network training, we use L_1 loss and perceptual loss as:

$$L_{recon1} = \sum_i \|X_i - D(E_{sp}(X_i), E_{ex}(X_i))\|_1, \quad (1)$$

$$L_{per1} = \sum_i \|\phi(X_i) - \phi(D(E_{sp}(X_i), E_{ex}(X_i)))\|_1, \quad (2)$$

where ϕ denotes the Gram matrix of VGG-19 [15] network intermediate features.

To boost disentanglement of the network, we use additional losses with $Y_1^{'}$ and $Y_3^{'}$, which have the same spatial features with ground-truth HDR image but have different exposure features. They are mapped from the ground-truth HDR image via the inverse version of gamma correction function as

$$Y_i^{'} = t_i Y_i^{\frac{1}{\gamma}}, i = 1, 3, \quad (3)$$

where γ= 2.2 denotes the gamma correction parameter and t_i denotes exposure time of Y_i. The additional losses for these spatial features are defined as follows:

$$L_{recon2} = \sum_{i=1,3} \|Y_i^{'} - D(E_{sp}(X_2), E_{ex}(X_i))\|_1, \quad (4)$$

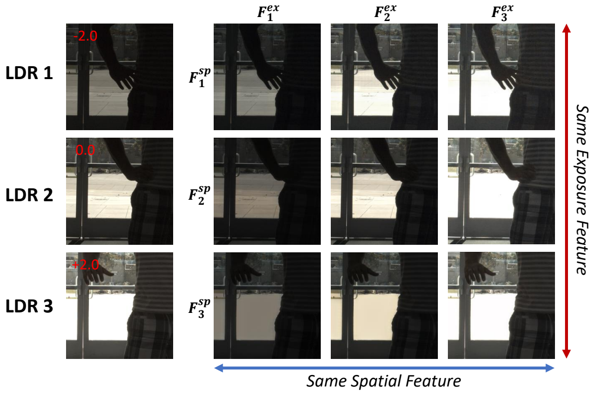

Fig. 2. Images generated by decoder D with various F_i^{sp} and F_i^{ex}. Images in the same row have the same spatial information, and images in the same column have the same global exposure information.

$$L_{per2} = \sum_{i=1,3} \|\phi(Y_i^{'}) - \phi(D(E_{sp}(X_2), E_{ex}(X_i)))\|_1. \quad (5)$$

Note that we only extract the spatial information from X_2, which has the same spatial information as the ground-truth HDR image. The total loss of the disentangle network is the weighted sum of reconstruction and perceptual losses as

$$L_{DIS} = L_{recon1} + L_{recon2} + \lambda(L_{per1} + L_{per2}), \quad (6)$$

where λ is a balance parameter.

Fig. 2 displays synthesized images with (F_i^{sp}, F_j^{ex}) pairs. Images generated with the same F_i^{sp} are spatially consistent (row), and images generated with the same F_j^{ex} have similar global exposure attribute (column). These images show that our disentangle network extracts meaningful features F_i^{sp} and F_i^{ex} from input images.

2.2. Disentangle Features Guided Network

The overall structure of DFGNet is shown in Fig. 3, which aims at producing high-quality HDR image $\hat{H}$ from given multiple LDR images X_i. We follow the existing practice in [1], which maps the given LDR images X_i to the HDR images $\tilde{X}_i$ by gamma correction.

$$\tilde{X}_i = X_i^{\gamma}/t_i, i = 1, 2, 3, \quad (7)$$

where $\gamma = 2.2$ denotes the gamma correction parameter and t_i denotes exposure time value of X_i. In this work, we concatenate X_i and $\tilde{X}_i$ along the channel dimension to obtain 6-channel input $L_i = [X_i, \tilde{X}_i]$. We adopt U-Net [16] structure as a base network architecture since many previous HDR imaging networks [1,4,6] show reliable results with the U-Net structure. To extract rich and diverse features from multiple LDR images, we construct individual encoders for each LDR image inputs. Each encoder contains three 3×3 convolutional layers with the stride of 2 and extracts multi-scale LDR

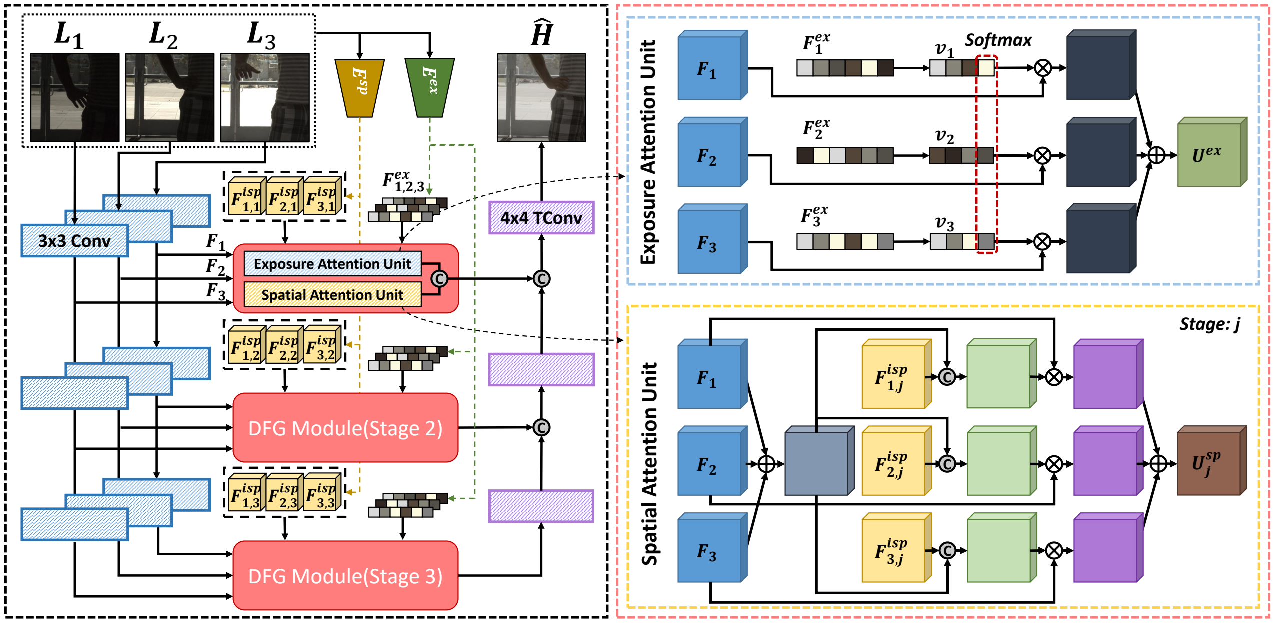

Fig. 3. The architecture of the proposed DFGNet and DFG Module. The DFGNet is described in the left box. DFG module consists of an *exposure attention unit* (top right) and a *spatial attention unit* (bottom right). The boxes with blue and purple boundary lines in DFGNet represent 3×3 Conv with a stride of 2 and 4×4 transposed Conv with a stride of 2, respectively.

image features of the size $(\frac{H}{2^{j-1}} \times \frac{W}{2^{j-1}} \times 2^{j-1}C)$, where j denotes the index of the stage. The decoder merges LDR image features of the same stage through a DFG module. The decoder contains two 4×4 transposed convolutional layers with the stride of 2 that upsample merged features from the previous stage to the same spatial size of the current stage. Every merged feature of each stage is concatenated with the upsampled features. Note that the DFG module is used to merge extracted features in the same stage instead of summing those features directly. The DFG module leverages key features of LDR images, which are extracted from pre-trained encoder E_{sp} and E_{ex}.

Fig. 3 also shows that our DFG module consists of two specific attention units. The spatial attention unit obtains an attention map from the sum of input features F_i and extracts intermediate feature $F^{isp}_{i,j}$ in the j stage. Attention maps are multiplied to each input feature, and we sum these features for merging. The whole merging process in the spatial attention unit is defined as

$$U^{sp}_j = \sum_i (F_i \otimes \sigma(Conv(Concat(F^{isp}_{i,j}, \sum_k F_k)))), \tag{8}$$

where U^{sp}_j denotes the merged feature of input features F_i in the stage j, $F^{isp}_{i,j}$ is the extracted intermediate feature of LDR image X_i, and $Concat(\cdot)$ is the channel-wise concatenation. Note that features in each stage have a different spatial size. Hence, as described in Fig. 1, we extract $F^{isp}_{i,j}$ from different intermediate layers of E_{sp}.

The exposure attention unit gives weight along the channel dimension of input features F_i. The exposure feature vector F^{ex}_i is used to obtain the channel-wise weight vector v_i. Similar to the spatial attention unit, the size of the channel dimension in each stage is different. Thus, we adopt one dense layer to match the channel size of v_i to F_i, and the softmax function is applied along channel dimension, described as

$$U^{ex} = \sum_i (F_i \otimes Softmax(v_i)). \tag{9}$$

We use tone-mapped HDR image and predicted HDR image in loss function for more effective training according to [1]. Given the ground-truth HDR image H, we compress the range of the image using μ-law:

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

where μ is the parameter of the tone-mapping function and $\mathcal{T}(H)$ is a tone-mapped HDR image. Finally, the DFGNet is trained with L_1 loss between the tone-mapped HDR image $\mathcal{T}(H)$ and tone-mapped predicted HDR image $\mathcal{T}(\hat{H})$:

$$L_{DFG} = \|\mathcal{T}(H) - \mathcal{T}(\hat{H})\|_1. \tag{11}$$

3. EXPERIMENTAL RESULTS

3.1. Dataset and Evaluation Metrics

We train and evaluate our DFGNet on Kalantri's dataset [1], which includes 74 training samples and 15 testing samples.

Each sample consists of three LDR images captured with a different exposure value of {-3,0,+3} or {-2,0,+2}. We compute the PSNR for images after tone-mapping using μ-law (PSNR-μ) and linear (PSNR-ℓ). Further, we also conduct quantitative evaluation by computing HDR-VDP-2 [17].

3.2. Implementation and Details

We implement our network using PyTorch and evaluate it on an NVIDIA Tesla V100 GPU. Parametric ReLU [18] is adopted in DFGNet for more flexible feature activation. We sample overlapping patches of size 256×256 with a stride of 64. To alleviate overfitting, we apply flip and rotation on sampled patches. The network is optimized by Adam [19] optimizer with an initial learning rate of 1e-4 and decay rate of 0.1. During the test step, we infer a set of LDR images at 1440×960 resolution.

3.3. Comparison with State-of-the-art Methods

We evaluate our method and compare it with state-of-the-art methods on Kalantari's dataset [1]. The state-of-the-art methods include patch-based model [20] and deep learning-based models [1–3, 5, 6]. Especially, Kalantari *et al.* [1] applied the optical flow in the pre-processing step, and DeepHDR [5] used homography transformation to align the background of the input image. Table 1 shows the result of the experiments. Model 1 and Model 2 are ablations of components in DFGNet to see their roles. Specifically, Model 1 excludes the exposure attention unit (EAU), and Model 2 is without the spatial attention unit (SAU). The proposed DFGNet outperforms all the previous methods on all evaluation metrics without using optical flow or additional data. Fig. 4 shows a qualitative comparison of several models. Note that HDR-GAN is excluded here since the authors do not provide the code. Sen *et al.*, Kalantari *et al.*, and DeepHDR fail to produce an image with similar global brightness to ground-truth and cannot preserve the details. AHDRNet keeps more details than previous models but shows saturating artifacts due to failure in recovering over-exposed regions. Our DFGNet shows HDR image with matching brightness and color to ground-truth, recovering saturated region and details of image successfully.

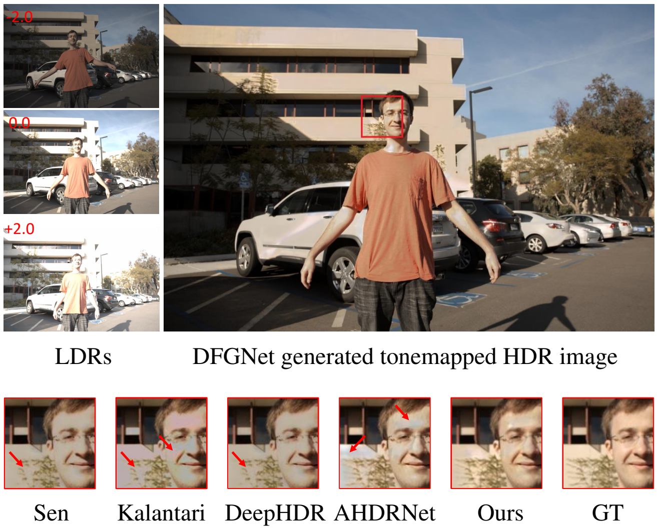

Fig. 4. An example from the test dataset [1]. We compare patches of predicted tone-mapped HDR images from various methods. Our DFGNet shows brightness matched to the ground truth and recovers saturated region and details.

Table 1. Quantitative comparison of different models. Each score is the average across all testing images. The best score are indicated in boldface.

Method	PSNR-μ	PSNR-ℓ	HDR-VDP-2
Sen *et al.* [20]	40.80	38.11	59.38
Kalantari *et al.* [1]	42.74	41.23	65.05
DeepHDR [5]	41.64	40.91	64.90
AHDRNet [2]	43.63	41.14	64.61
NHDRRNet [3]	42.41	-	61.21
HDR-GAN [6]	43.92	41.57	65.45
Model 1(w/o EAU)	43.57	41.45	65.88
Model 2(w/o SAU)	43.92	41.59	**66.91**
Ours	**44.19**	**41.89**	66.84

4. CONCLUSION

We have proposed a disentangled feature-guided network for generating an HDR image from multiple LDR inputs. To alleviate the major problems in multi-exposure HDR imaging, namely the misalignments and the information losses in LDR inputs, we have extracted representative spatial/exposure features and leveraged those features in the proposed network. Experiments show that the proposed network successfully merges the LDR inputs with fewer artifacts and better brightness/color matching to the ground truth compared to state-of-the-art methods.

Acknowledgment This work was supported in part by the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by Korea government(MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)], and in part by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2021-0-02068, Artificial Intelligence Innovation Hub).

5. REFERENCES

[1] Nima Khademi Kalantari, Ravi Ramamoorthi, et al., “Deep high dynamic range imaging of dynamic scenes.,” *ACM Trans. Graph.*, vol. 36, no. 4, pp. 144–1, 2017.

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 1751–1760.

[3] Qingsen Yan, Lei Zhang, Yu Liu, Yu Zhu, Jinqiu Sun, Qinfeng Shi, and Yanning Zhang, “Deep hdr imaging via a non-local network,” *IEEE Transactions on Image Processing*, vol. 29, pp. 4308–4322, 2020.

[4] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger, “Hdr image reconstruction from a single exposure using deep cnns,” *ACM transactions on graphics (TOG)*, vol. 36, no. 6, pp. 1–15, 2017.

[5] Shangzhe Wu, Jiarui Xu, Yu-Wing Tai, and Chi-Keung Tang, “Deep high dynamic range imaging with large foreground motions,” in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 117–132.

[6] Yuzhen Niu, Jianbin Wu, Wenxi Liu, Wenzhong Guo, and Rynson WH Lau, “Hdr-gan: Hdr image reconstruction from multi-exposed ldr images with large motions,” *IEEE Transactions on Image Processing*, vol. 30, pp. 3885–3896, 2021.

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

[8] Katrien Jacobs, Celine Loscos, and Greg Ward, “Automatic high-dynamic range image generation for dynamic scenes,” *IEEE Computer Graphics and Applications*, vol. 28, no. 2, pp. 84–93, 2008.

[9] Abhilash Srikantha and Désiré Sidibé, “Ghost detection and removal for high dynamic range images: Recent advances,” *Signal Processing: Image Communication*, vol. 27, no. 6, pp. 650–662, 2012.

[10] Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski, “High dynamic range video,” *ACM Transactions on Graphics (TOG)*, vol. 22, no. 3, pp. 319–325, 2003.

[11] Henning Zimmer, Andrés Bruhn, and Joachim Weickert, “Freehand hdr imaging of moving scenes with simultaneous resolution enhancement,” in *Computer Graphics Forum*. Wiley Online Library, 2011, vol. 30, pp. 405–414.

[12] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, “Non-local neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 7794–7803.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] Xun Huang and Serge Belongie, “Arbitrary style transfer in real-time with adaptive instance normalization,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1501–1510.

[15] Karen Simonyan and Andrew Zisserman, “Very deep convolutional networks for large-scale image recognition,” *arXiv preprint arXiv:1409.1556*, 2014.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[17] Rafał Mantiuk, Kil Joong Kim, Allan G Rempel, and Wolfgang Heidrich, “Hdr-vdp-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions,” *ACM Transactions on graphics (TOG)*, vol. 30, no. 4, pp. 1–14, 2011.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Delving deep into rectifiers: Surpassing human-level performance on imagenet classification,” in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[19] Diederik P Kingma and Jimmy Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[20] 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.*, vol. 31, no. 6, pp. 203–1, 2012.