M.S. THESIS

High Dynamic Range Imaging by Feature Disentanglement of Multi-Exposure Inputs

다중 노출 입력의 피쳐 분해를 통한 하이 다이나믹 레인지 영상 생성 방법

BY

LEE KEUNTEK

AUGUST 2022

INTERDISCIPLINARY PROGRAM IN
ARTIFICIAL INTELLIGENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

M.S. THESIS

High Dynamic Range Imaging by Feature Disentanglement of Multi-Exposure Inputs

다중 노출 입력의 피쳐 분해를 통한 하이 다이나믹 레인지 영상 생성 방법

BY

LEE KEUNTEK

AUGUST 2022

INTERDISCIPLINARY PROGRAM IN
ARTIFICIAL INTELLIGENCE
COLLEGE OF ENGINEERING
SEOUL NATIONAL UNIVERSITY

High Dynamic Range Imaging by Feature Disentanglement of Multi-Exposure Inputs

다중 노출 입력의 피쳐 분해를 통한 하이 다이나믹 레인지 영상 생성 방법

지도교수 조 남 익

이 논문을 공학석사 학위논문으로 제출함

2022년 8월

서울대학교 대학원

협동과정 인공지능전공

이 근 택

이근택의 공학석사 학위 논문을 인준함

2022년 8월

위 원 장: _____
부위원장: _____
위 원: _____

Abstract

Multi-exposure high dynamic range (HDR) imaging aims to generate an HDR image from multiple differently exposed low dynamic range (LDR) images. Multi-exposure HDR imaging is a challenging task due to two major problems. One is misalignments among the input LDR images, which can cause ghosting artifacts on result HDR, and the other is missing information on LDR images due to under-/over-exposed region. Although previous methods tried to align input LDR images with traditional methods(e.g., homography, optical flow), they still suffer undesired artifacts on the result HDR image due to estimation errors that occurred in aligning step.

In this dissertation, disentangled feature-guided HDR network (DFGNet) is proposed to alleviate the above-stated problems. Specifically, exposure features and spatial features are first extracted from input LDR images, and they are disentangled from each other. Then, these features are processed through the proposed DFG modules, which produce a high-quality HDR image. The proposed DFGNet shows outstanding performance compared to previous methods, achieving the PSNR-ℓ of 41.89dB and the PSNR-μ of 44.19dB.

keywords: High Dynamic Range Imaging, Multi-Exposed Imaging, Feature Disentanglement

student number: 2020-29425

Contents

List of Tables

List of Figures

Chapter 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 capturing 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[5, 1, 8, 13, 16, 14]. 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 [4, 3, 2, 6, 7]. For example, the optical flow has been adopted in [6,

Figure 1.1: A brief explanation of multi-exposure HDR imaging task. Each LDR image on left side has different exposure time value(denoted in parenthesis) and not aligned.

7] to align the pixels between the LDR images. The early CNN-based methods also adopted the optical flow as a pre-processing step. For example, [5] 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 [1, 8, 13, 16, 14]. For example, AHDRNet[1] 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 [8] constructs non-local blocks [9] 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.

In this dissertation, to address the above-stated major issues in HDR imaging, *i.e.,* aligning the structure and exploiting the exposure information, a disentangle network is proposed that extracts representative exposure features and spatial features separately from LDR images. Further, a disentangled feature-guided network (DFGNet) is designed, which consists of DFG modules that align LDR image features with the guidance of disentangled spatial and exposure features. The main idea of this dissertation is to 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.

Chapter 2

Related Works

In this chapter, related works on HDR imaging task are briefly introduced into two categories: single-frame HDR imaging, multi-frame HDR imaging with dynamic scenes.

2.1 Single-frame HDR imaging

The single-frame HDR imaging aims to produce an HDR image from a single LDR image. Since there is severe missing information in a single LDR image due to the under-/over-exposed region, reconstructed HDR images often show undesired artifacts and unsatisfying quality. One popular strategy for addressing the above-stated issue is synthesizing multiple LDR images with a different exposure setting from a single input LDR image. Lee *et al.*[21] proposed a method that generates pseudo multi-exposed LDR images by modeling inverse tone-mapping with a deep neural network(DNN). Specifically, they adopt a generative adversarial network(GAN)[22] structure for producing more realistic pseudo LDR images. After generating pseudo multi-exposed LDR images, they merge generated LDR images to produce an HDR image. Endo *et al.*[23] proposed a network that synthesizes a set of up-exposed LDR images and a set of down-exposed images from a single LDR image input. They also construct a GAN

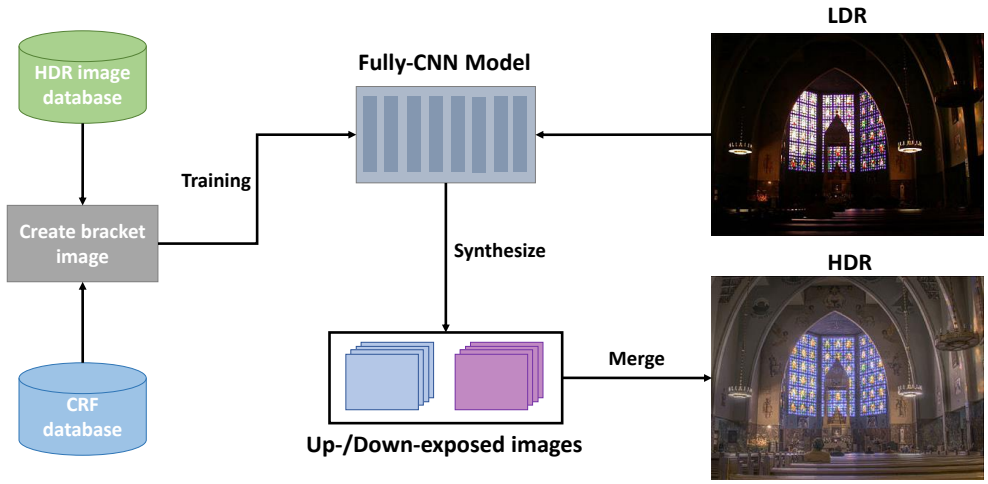

Figure 2.1: A brief explanation of single-frame HDR imaging method proposed by endo *et al.*[23]. The network is trained with external HDR image database and camera response function database to synthesize multiple LDR images from a single input LDR image.

based network structure and train the network with an external HDR image dataset and camera response function(CRF) dataset. The trained network can synthesize multiple LDR images that have differenent exposure-related attributes from a single LDR image. The synthesized multiple LDR images are merged to produce an HDR image. Fig. 2.1 displays a brief explanation of proposed HDR reconstruction process. Liu *et al.*[24] model the camera pipeline and reverse it to produce an HDR image from a single LDR image. First, they model the HDR-to-LDR image formation pipeline, which consists of dynamic range clipping, non-linear CRF, and quantization which are represented in Fig. 2.2. After simulating the HDR-to-LDR pipeline, three specialized CNNs proposed to reverse each step in the simulated pipeline. Further, the hallucination and refinement networks are also proposed to predict missing content in over-exposed regions and fine-tune the whole model, respectively.

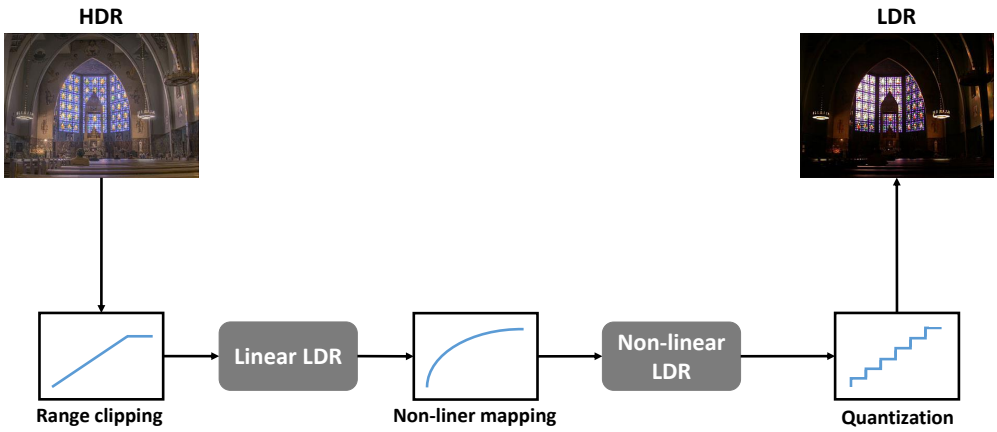

Figure 2.2: A general camera pipeline in HDR-to-LDR transformation. The pipeline includes range clipping, non-linear mapping(CRF), and quantization.

2.2 Multi-frame HDR imaging with dynamic scenes

The target of multi-frame HDR imaging is producing an HDR image from multiple LDR images which are captured in the same scene. Multiple LDR images produce more rich information, which is important for producing fine HDR images. However, there are motion differences when moving objects exist in the scene since each frame is captured independently. The ghosting artifact which is displayed in Fig. 2.4 is most a common problem in the reconstructed HDR image, caused by misalignment of input multiple LDR images. To address misalignment problem, previous methods[3, 4, 5, 6, 7, 25] tried to align LDR images before merging process by using traditional aligning methods. For instance, [7, 6, 5] applied optical flow estimation to compensate motions of objects by pixel-level alignment. However, they still suffer from undesired artifacts in the result HDR image due to estimation errors caused in the alignment stage. To alleviate estimation errors in alignment stage, several methods[29, 30] proposed to ignore significant estimation error in alignments stage. However, they still suffer from unstable reconstruction results due to inaccurate pixel-level identification of moving

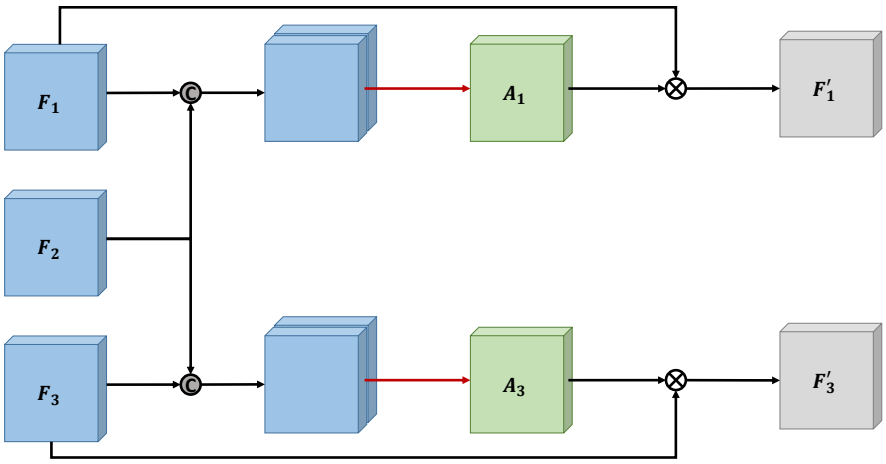

Figure 2.3: Illustration of spatial attention module in [1]. Input features(F_1, F_3) are concatenated with reference frame F_2, used to create attention masks A_1, A_3. Then, they multiplied with input features to produce aligned features. Red arrows denote the convolutional layer.

objects.

Recently, CNN-based methods have achieved notable success in the image restoration tasks[26, 27, 28, 36, 38, 35], including HDR imaging. Kalantari *et al.*[5] firstly proposed a CNN based merging model. They aligned LDR images with optical flow estimation in the pre-processing step. Aligned LDR images are merged through the CNN model to produce an HDR image. Wu *et al.*[16] proposed U-Net [12] based network, which consists of multiple encoders and a single decoder, including skip-connection. They align the background of LDR images with the homography transformation. After the alignment process, aligned LDR images are fed to the proposed network to produce an HDR image. Yan *et al.*[1] proposed attention based network to handle motions in scene. They construct spatial attention modules in the network which align input features by end-to-end training without any additional alignment process. As illustrated in Fig. 2.3, attention masks generated are from the concatenated feature, multiplied with input features for alignment. They also employ dilated residual dense

blocks(DRDB) for preserving image details and enlarging the receptive field of the network. Non-local[9] based network also proposed by Yan *et al.*[8]. They exploit non-local attributes of features and extract more rich and diverse global features to produce more realistic HDR images without any undesired artifacts. Liu *et al.*[32] employed pyramid cascading deformable(PCD) alignment module for feature alignment which proposed in video super-resolution task[34]. They apply the spatial attention module to input LDR images and the PCD module to gamma-corrected LDR images separately for more accurate pixel-level alignment. Niu *et al*[14] proposed GAN based method with multi-scale structure. The generator is designed with multi-scale individual encoders and one decoder for HDR image reconstruction. A patchGAN[33] discriminator with sphere loss[31] is adopted for adversarial training. Deep supervision is also proposed as an additional training objective, which helps the whole network to generate more fine intermediate features in the reconstruction process.

In this paper, the HDR reconstruction process is divided into a two-step. First, the spatial feature and the exposure feature are extracted from LDR images through the disentangle network. These features contain distinctive content information and global information of LDR images which can represent images effectively. After extracting features, a reconstruction network produces a realistic HDR image. Especially, the extracted spatial feature and exposure feature are leveraged in the feature merging process, which bring significant benefits in the reconstruction step.

Figure 2.4: Image samples that show ghosting artifacts and saturated artifacts. Regions close to the flame(red arrows) show severe ghosting artifacts and saturated artifacts caused by large motions and exposure setting.

Chapter 3

Proposed Method

3.1 Disentangle Network for Feature Extraction

The proposed disentangle network extracts global exposure features and local spatial features from LDR images. The extracted features have representative attributes of input LDR image, and thus can be utilized effectively in the reconstruction step. As shown in Fig. 3.1, 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[20] 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[10] module. The AdaIN module applies the encoded style information to image content, showing impressive performance on style transfer tasks which can be formulated as

$$AdaIN(F_i^{sp}, F_i^{ex}) = \left(\frac{F_i^{sp} - \mu(F_i^{sp})}{\sigma(F_i^{sp})} \right) FC_s(F_i^{ex}) + FC_b(F_i^{ex}), \tag{3.1}$$

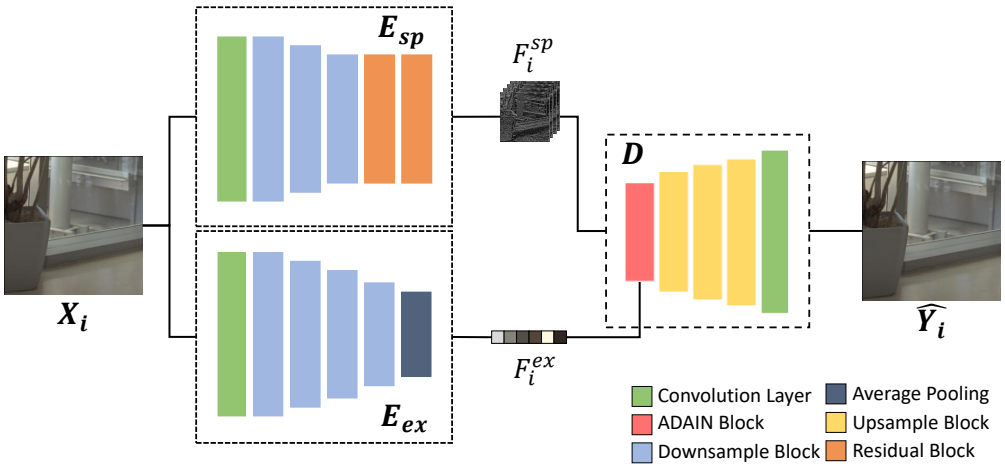

Figure 3.1: Overall architecture of disentangle network. The disentangle network consist of two encoder(E_{sp}, E_{ex}) and one decoder(D). Each layer denoted on the right down boxes.

where μ and σ denote mean and variance function, FC_s, FC_b denote fully connected layer for calculating scale and bias factor in AdaIN module. With the AdaIN module, our encoder-decoder structured disentangle network fuse E_{ex}, E_{sp} effectively and able to reconstruct the input image more realistic. For network training, L_1 loss and perceptual loss are used:

$$L_{recon1} = \sum_i \|X_i - D(E_{sp}(X_i), E_{ex}(X_i))\|_1, \tag{3.2}$$

$$L_{per1} = \sum_i \|\phi(X_i) - \phi(D(E_{sp}(X_i), E_{ex}(X_i)))\|_1, \tag{3.3}$$

where ϕ denotes the Gram matrix of VGG-19[11] network intermediate features.

To boost disentanglement of the network, additional losses with $Y_1^{'}$ and $Y_3^{'}$ are adopted, 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

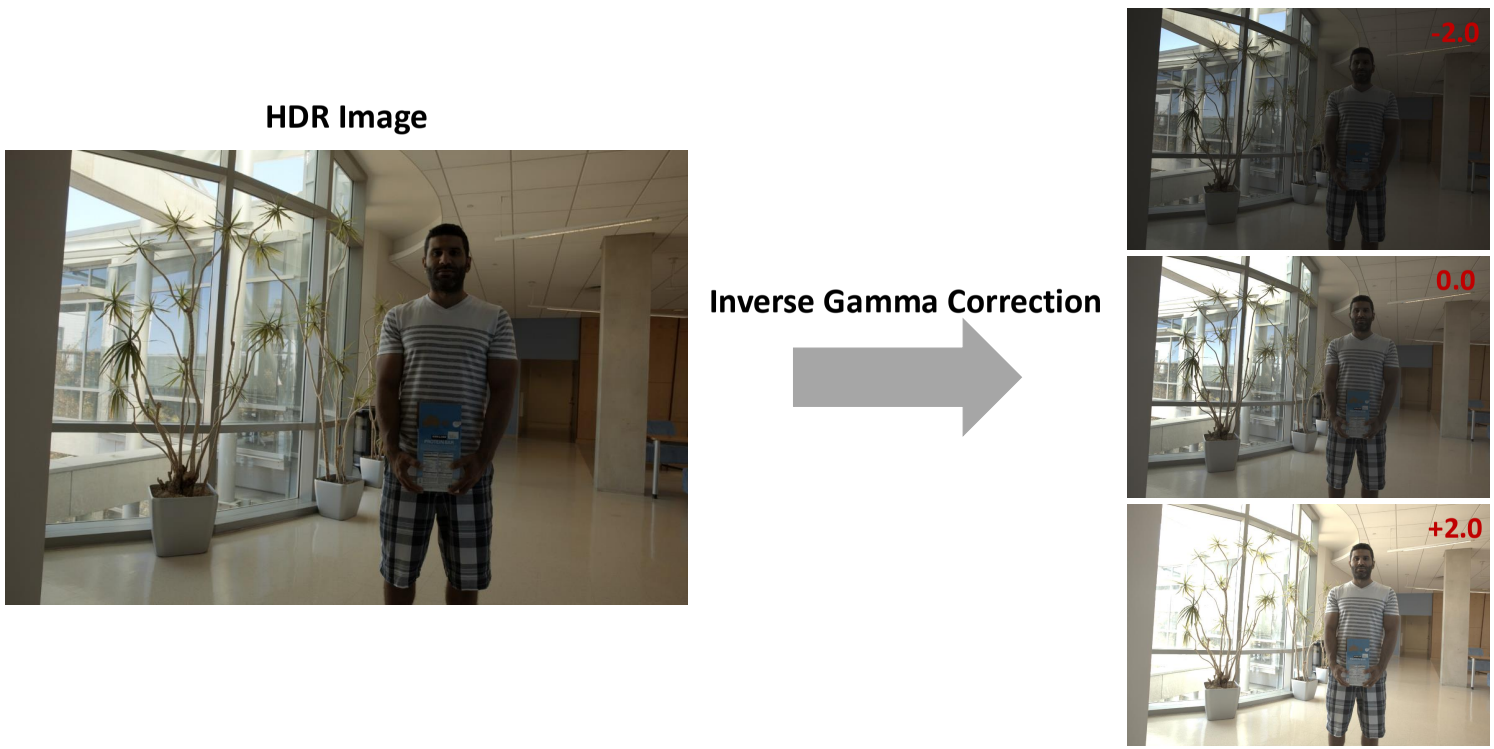

Figure 3.2: An example of synthesized LDR images with a single HDR image via inverse gamma correction. The inverse gamma correction manipulates HDR image to the corresponding exposure value.

the inverse version of gamma correction function as

$$Y_i^{'} = t_i Y^{\frac{1}{\gamma}}, i = 1, 3, \tag{3.4}$$

where γ= 2.2 denotes the gamma correction parameter and t_i denotes exposure time value. Fig. 3.2 illustrates synthesized LDR images from an HDR image via inverse gamma correction. Synthesized images show matching global attributes which are related to exposure level. Each LDR image shows corresponding exposure-related attributes. The additional losses for these spatial features and exposure features are defined as follows:

$$L_{recon2} = \sum_{i=1,3} \|Y_i^{'} - D(E_{sp}(X_2), E_{ex}(X_i))\|_1, \tag{3.5}$$

$$L_{per2} = \sum_{i=1,3} \|\phi(Y_i^{'}) - \phi(D(E_{sp}(X_2), E_{ex}(X_i)))\|_1. \tag{3.6}$$

Note that only the spatial information is only extracted from X_2, which has the same spatial information as the ground-truth HDR image. Fig. 3.3 displays additional training objective for the disentangle network. The total loss of the disentangle network is the weighted sum of reconstruction and perceptual losses as

$$L_{DIS} = L_{recon1} + L_{recon2} + \lambda(L_{per1} + L_{per2}), \tag{3.7}$$

where λ is a balance parameter. Fig. 3.4 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}, F_i^{ex} from input images.

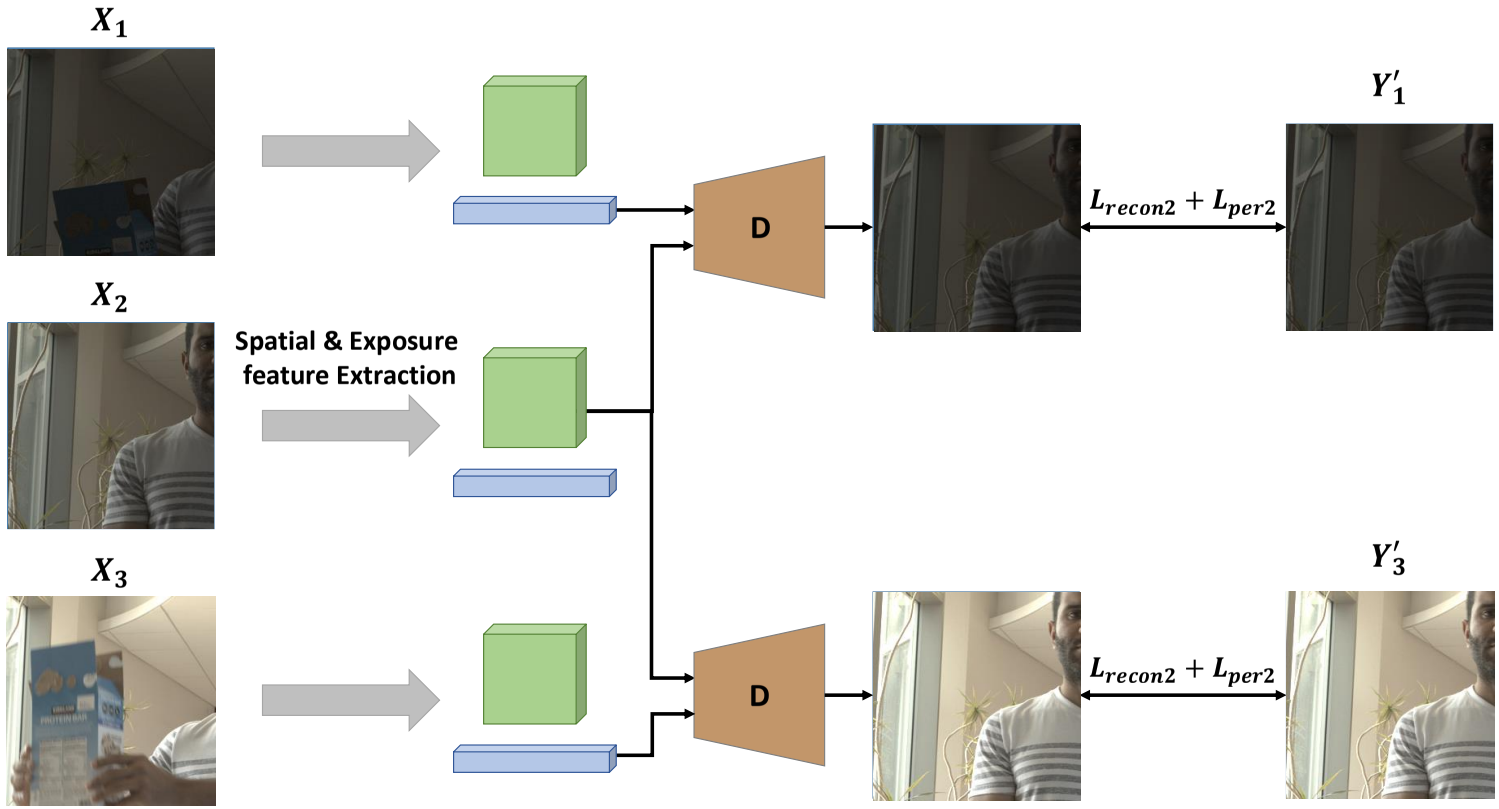

Figure 3.3: The description of additional training objective for the disentangle network.

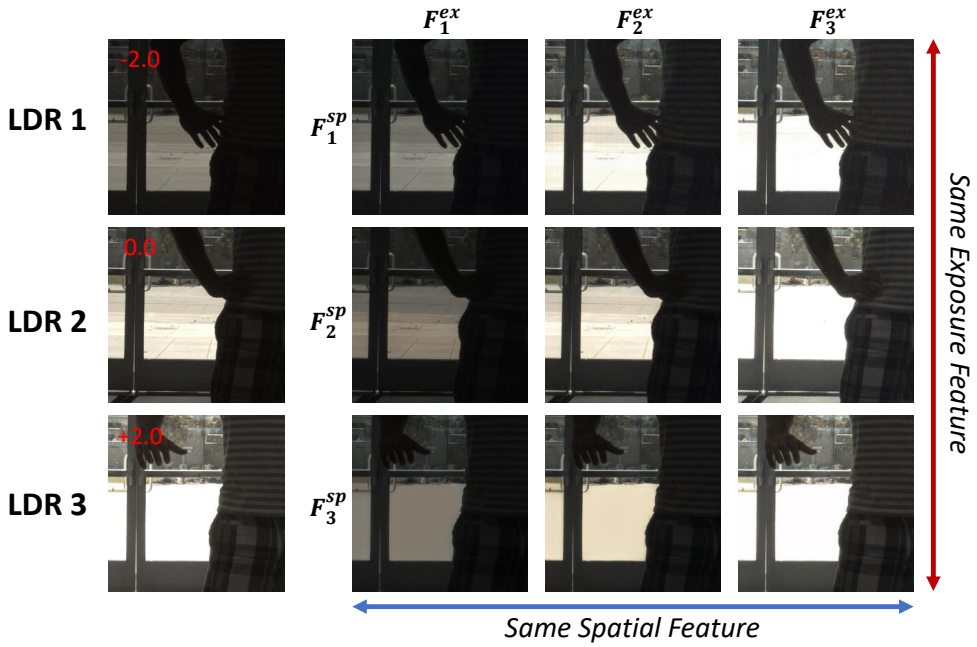

Figure 3.4: Images generated by decoder D with various F_i^{sp} and F_i^{ex}. Images in the same row have the same spatial information, and images in the same column have the same global exposure information.

3.2 Disentangle Features Guided Network

DFGNet aims at producing high-quality HDR image $\hat{H}$ from given multiple LDR images X_i. Following the existing practice in [5], which maps the given LDR images X_i to the HDR images $\tilde{X}_i$ by gamma correction.

$$\tilde{X}_i = X_i^{\gamma}/t_i, i = 1, 2, 3, \tag{3.8}$$

where $\gamma = 2.2$ denotes the gamma correction parameter and t_i denotes exposure time value of X_i. In this work, X_i and $\tilde{X}_i$ are concatenated along the channel dimension to obtain 6-channel input $L_i = [X_i, \tilde{X}_i]$. The U-Net [12] structure is employed as a base network architecture since many previous HDR imaging networks [5, 14, 13] show reliable results with the U-Net structure. To extract rich and diverse features from multiple LDR images, individual encoders are used for each LDR image inputs. Each encoder contains three 3×3 convolutional layers with the stride of 2, except the first stage. These convolutional layers extracts multi-scale LDR 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 encoding stage in each stage can be represented as

$$\begin{aligned} F_{i,1} &= E_{i,1}(L_i) \in \mathbb{R}^{H \times W \times C}, \\ F_{i,2} &= E_{i,2}(E_{i,1}(L_i)) \in \mathbb{R}^{H/2 \times W/2 \times 2C}, \\ F_{i,3} &= E_{i,3}(E_{i,2}(E_{i,1}(L_i))) \in \mathbb{R}^{H/4 \times W/4 \times 4C}, \end{aligned} \tag{3.9}$$

where i refers to the index of the frame number of input LDR image. $E_{i,j}$ denotes encoding block for i^{th} input LDR frame in j^{th} stage which consists of convolutional layer and parametric ReLU[17] activation. 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

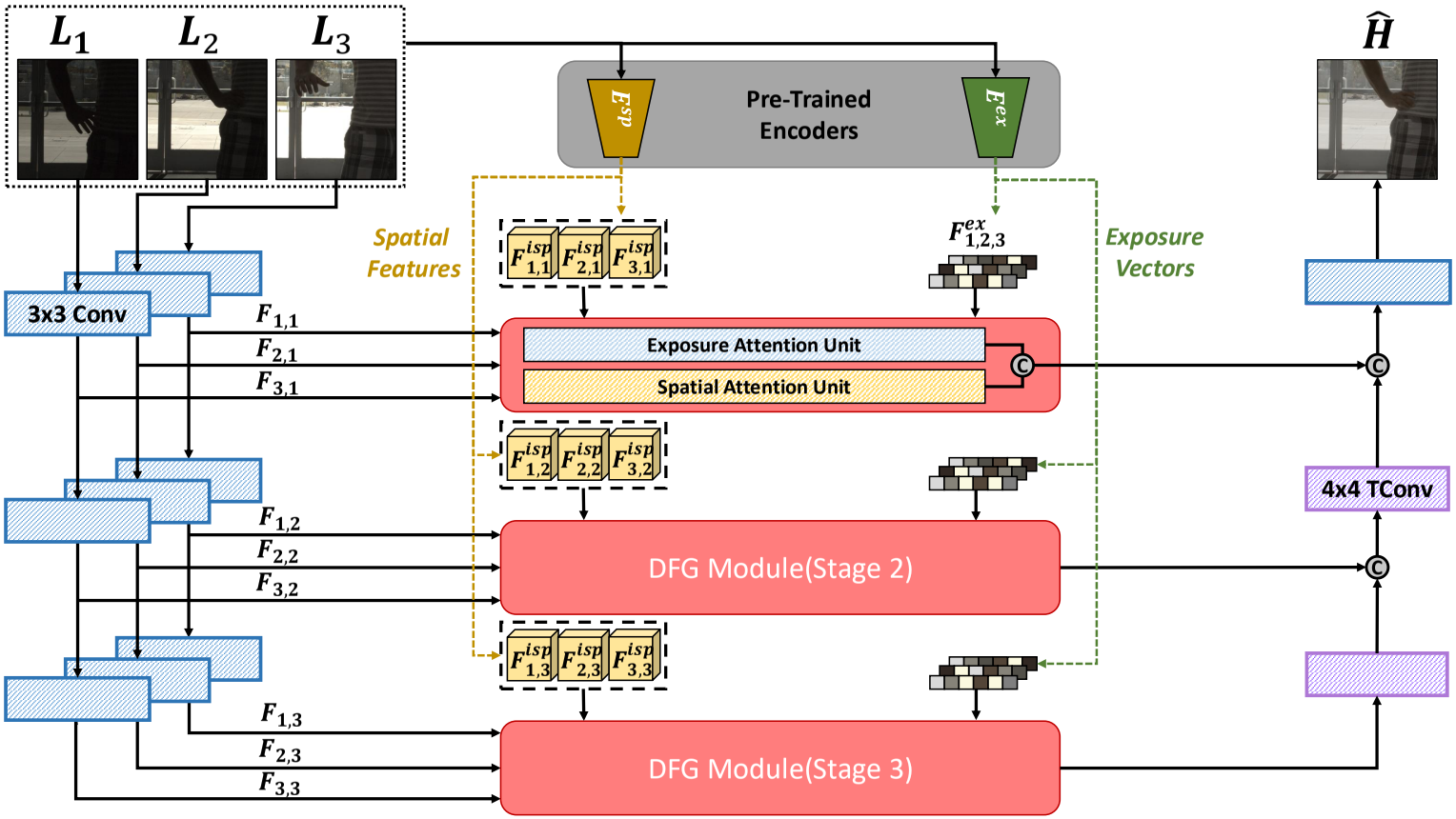

Figure 3.5: The architecture of the proposed DFGNet and DFG Module. The DFGNet is described in the left box. DFG module consists of an *exposure attention unit* and a *spatial attention unit*. Each blue-lined box and purple-lined box in DFGNet represent 3×3 Conv with a stride of 2 and 4×4 transposed Conv with a stride of 2, respectively.

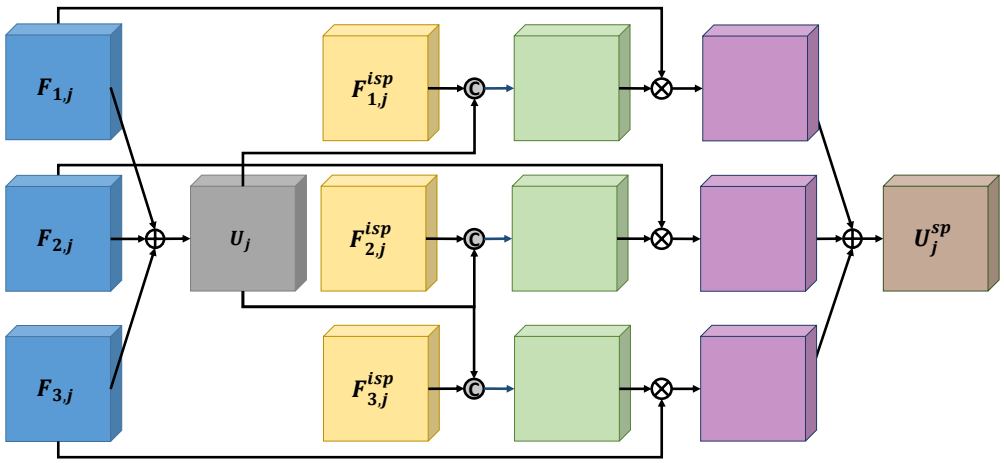

Figure 3.6: The details of a spatial attention unit in j stage. All input features are summed to create a unified feature U_j. The unified feature is concatenated with extracted spatial feature $F_{i,j}^{isp}$, producing attention maps. Each attention map is multiplied with the corresponding input feature and summed to produce a merged feature U_j^{sp}. Blue arrows denote the 3×3 convolutional layer.

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}.

As shown in Fig. 3.5, our DFG module consists of two specific attention units. Fig. 3.6 illustrates detail structure of spatial attention unit. The spatial attention unit obtains an attention map from the sum of input features $F_{i,j}$ and extracts intermediate feature $F_{i,j}^{isp}$ in the j stage. Attention maps are multiplied to each input feature, and these features are summed for merging. The whole merging process in the spatial attention unit is defined as

$$U_j = \sum_i F_{i,j}, i = 1, 2, 3 \tag{3.10}$$

$$U_j^{sp} = \sum_i (F_{i,j} \otimes \sigma(Conv(Concat(F_{i,j}^{isp}, U_j)))), \tag{3.11}$$

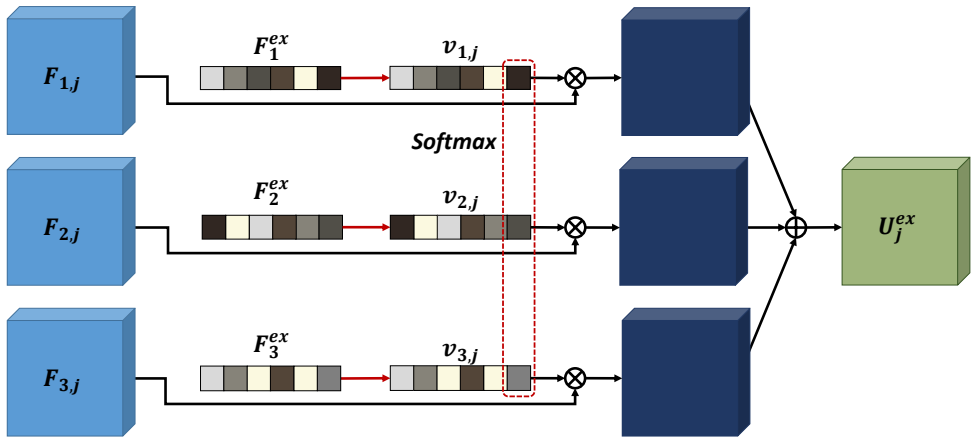

Figure 3.7: The details of an exposure attention unit. Each extracted exposure feature F_i^{ex} is used to create attention vector $v_{i,j}$ by applying a single fully connected layer and softmax function. Attention vectors are multiplied with the corresponding input feature and summed to produce a merged feature U_j^{ex}. Red arrows denote the fully connected layer.

where U_j^{sp} denotes the merged feature of input features $F_{i,j}$ in the stage j, $F_{i,j}^{isp}$ 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. 3.1, $F_{i,j}^{isp}$ is extracted from different intermediate layers of E_{sp}. Fig. 3.7 displays detail structure of exposure attention unit. The exposure attention unit gives weight along the channel dimension of input features $F_{i,j}$. The exposure feature vector $F_{i,j}^{ex}$ is used to obtain the channel-wise weight vector $v_{i,j}$. Similar to the spatial attention unit, the size of the channel dimension in each stage is different. Thus, one fully connected layer is added to match the channel size of $v_{i,j}$ to $F_{i,j}$, and the softmax function is applied along channel dimension.

$$v_{i,j} = FC(F_i^{ex}), \tag{3.12}$$

$$U_j^{ex} = \sum_i (F_{i,j} \otimes \mathit{Softmax}(v_{i,j})), \tag{3.13}$$

where j refers to the index of stage. Note that we construct spatial attention unit and exposure attention unit in the DFG module parallelly. With this parallel structure, each unit focus on creating a distinctive merged feature that can compensate for the feature from the other unit.

$$M_j = \mathit{Concat}(U_j^{sp}, U_j^{ex}), \tag{3.14}$$

where j refers to the index of current stage in network. Finally, decoder concatenate merged feature in each stage sequentially, which can be represented as

$$\begin{aligned} D_3 &= TC_3(M_j) \in \mathbb{R}^{H/2 \times W/2 \times 2C}, \\ D_2 &= TC_2(\mathit{Concat}(M_2, D_3)) \in \mathbb{R}^{H \times W \times C}, \\ D_1 &= \mathit{Concat}(M_1, D_2) \in \mathbb{R}^{H \times W \times 2C}, \end{aligned} \tag{3.15}$$

where j refers to the index of the network stage and TC_j denotes decoding block that consists of 4×4 transposed convolutional layer and parametric ReLU activation. The predicted HDR image $\hat{H}$ is obtained by applying a single 3×3 convolutional layer to D_1.

$$\hat{H} = \mathit{Conv}(D_1). \tag{3.16}$$

The HDR image is used after ton-mapping and predicted HDR image in loss function for more effective training according to [5]. Given the ground-truth HDR image H, the range of the image is compressed with μ-law:

$$\mathcal{T}(H) = \frac{log(1 + \mu H)}{log(1 + \mu)}, \tag{3.17}$$

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 $\mathcal{T}(\hat{H})$:

$$L_{DFG} = \|\mathcal{T}(H) - \mathcal{T}(\hat{H})\|_1. \tag{3.18}$$

Chapter 4

Experimental Results

4.1 Implementation and Details

The proposed network is implemented with PyTorch[37] and evaluate it on an NVIDIA Tesla V100 GPU. Parametric ReLU [17] is adopted in DFGNet for more flexible feature activation. Patches of size 256×256 with a stride of 64 are sampled for training. To alleviate overfitting, flip and rotation are applied on sampled patches. The network is optimized by Adam [18] optimizer with an initial learning rate of 1e-4 and decay rate of 0.1. During the test step, a set of LDR images at 1440×960 resolution are used.

4.2 Comparison with State-of-the-art Methods

In this section, proposed method are evaluated and compared with state-of-the-art methods on Kalantari's dataset [5]. The state-of-the-art methods include patch-based model [19] and deep learning-based models [5, 16, 8, 1, 14]. Especially, Kalantari *et al.* [5] applied the optical flow in the pre-processing step, and DeepHDR [16] used homography transformation to align the background of the input image. Table 4.1 shows the result of the experiments. The proposed DFGNet outperforms all the previous methods on all evaluation metrics without using optical flow or additional data.

Table 4.1: 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[15]
Sen *et al.*[19]	40.80	38.11	59.38
Kalantari *et al.*[5]	42.74	41.23	65.05
DeepHDR[16]	41.64	40.91	64.90
AHDRNet[1]	43.63	41.14	64.61
NHDRRNet[8]	42.41	-	61.21
HDR-GAN[14]	43.92	41.57	65.45
Proposed	**44.19**	**41.89**	**66.84**

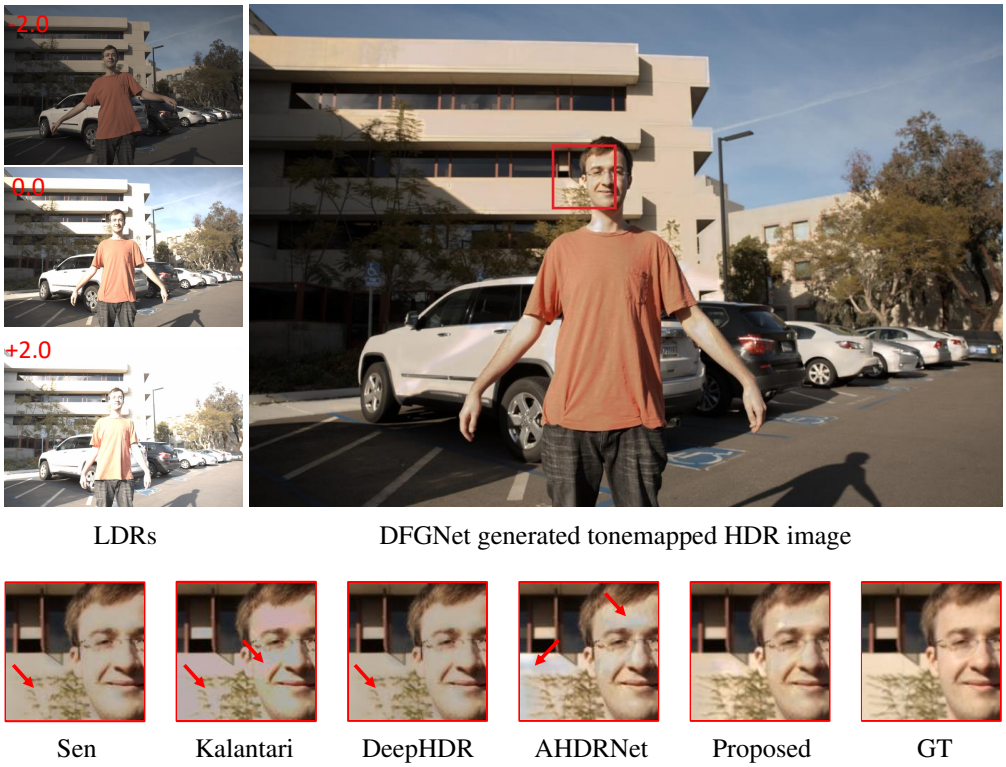

Figure 4.1: An example from the test dataset[5]. Patches of predicted tone-mapped HDR images from various methods are compared. HDR images are tone-mapped with Photomatix[39] for visualization. Proposed DFGNet shows brightness matching to the ground truth and recovers saturated region and details.

Fig. 4.1 shows a qualitative comparison between several models. Note that HDR-GAN

is excluded in this comparison 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 of the image. AHDRNet preserves more details than previous models but shows saturating artifacts due to failure in recovering over-exposed regions. The proposed DFGNet shows HDR image with matching brightness and color to ground-truth, recovering saturated region and details of image successfully.

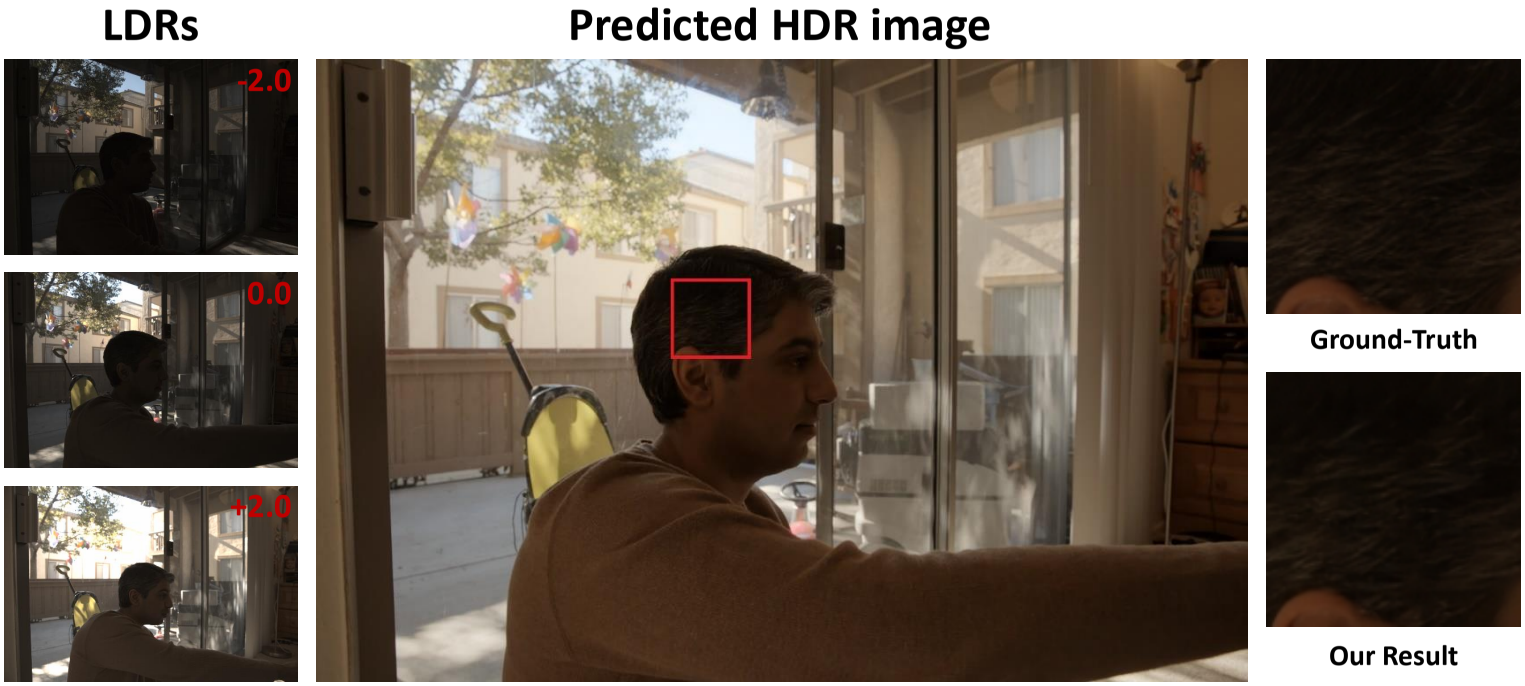

Figure 4.2: Another example from test dataset[5]. Our proposed network can reconstructs detailed region of image more realistic.

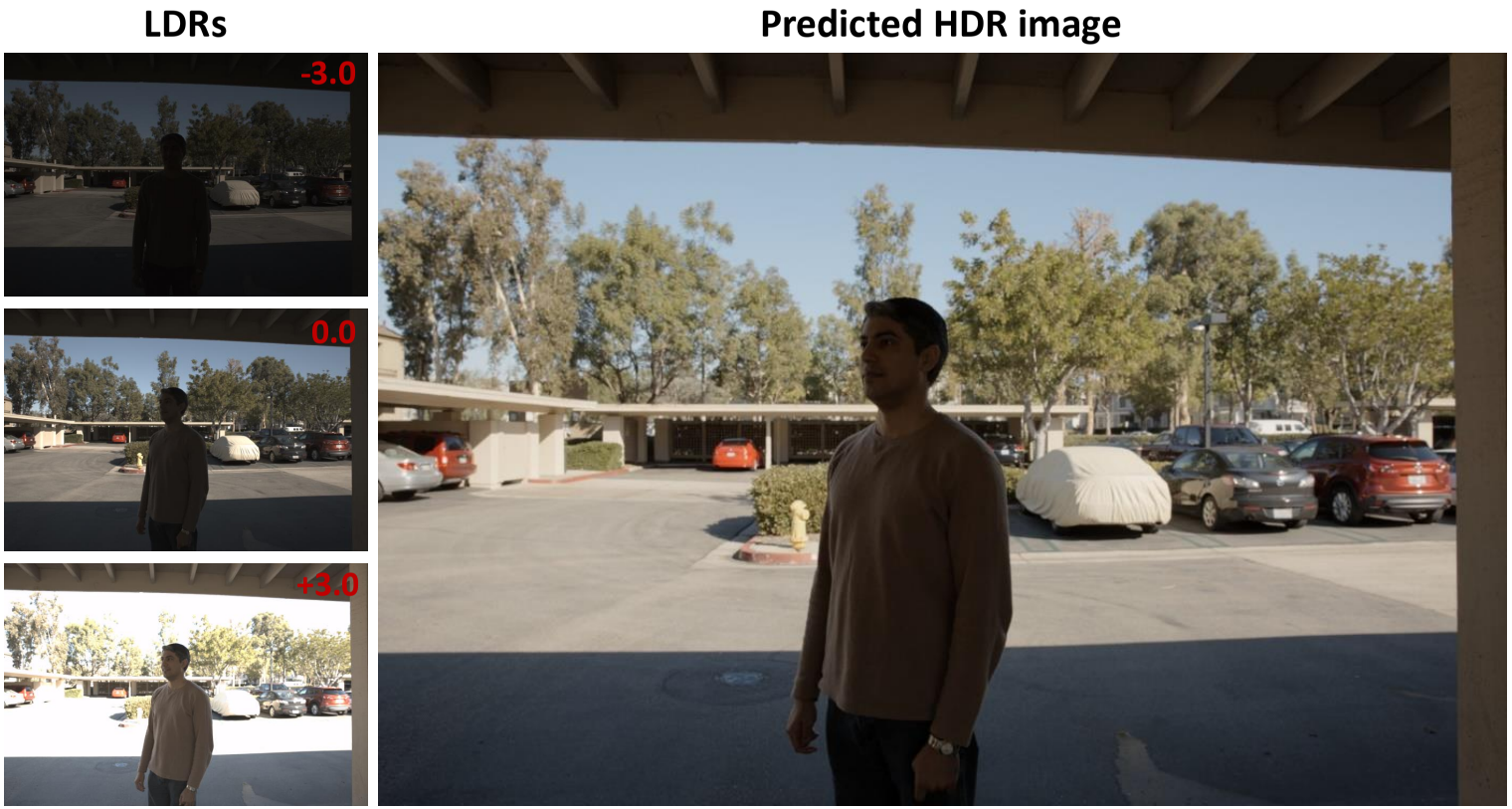

Figure 4.3: An example from test dataset[5] which has [-3,0,+3] range of exposure values.

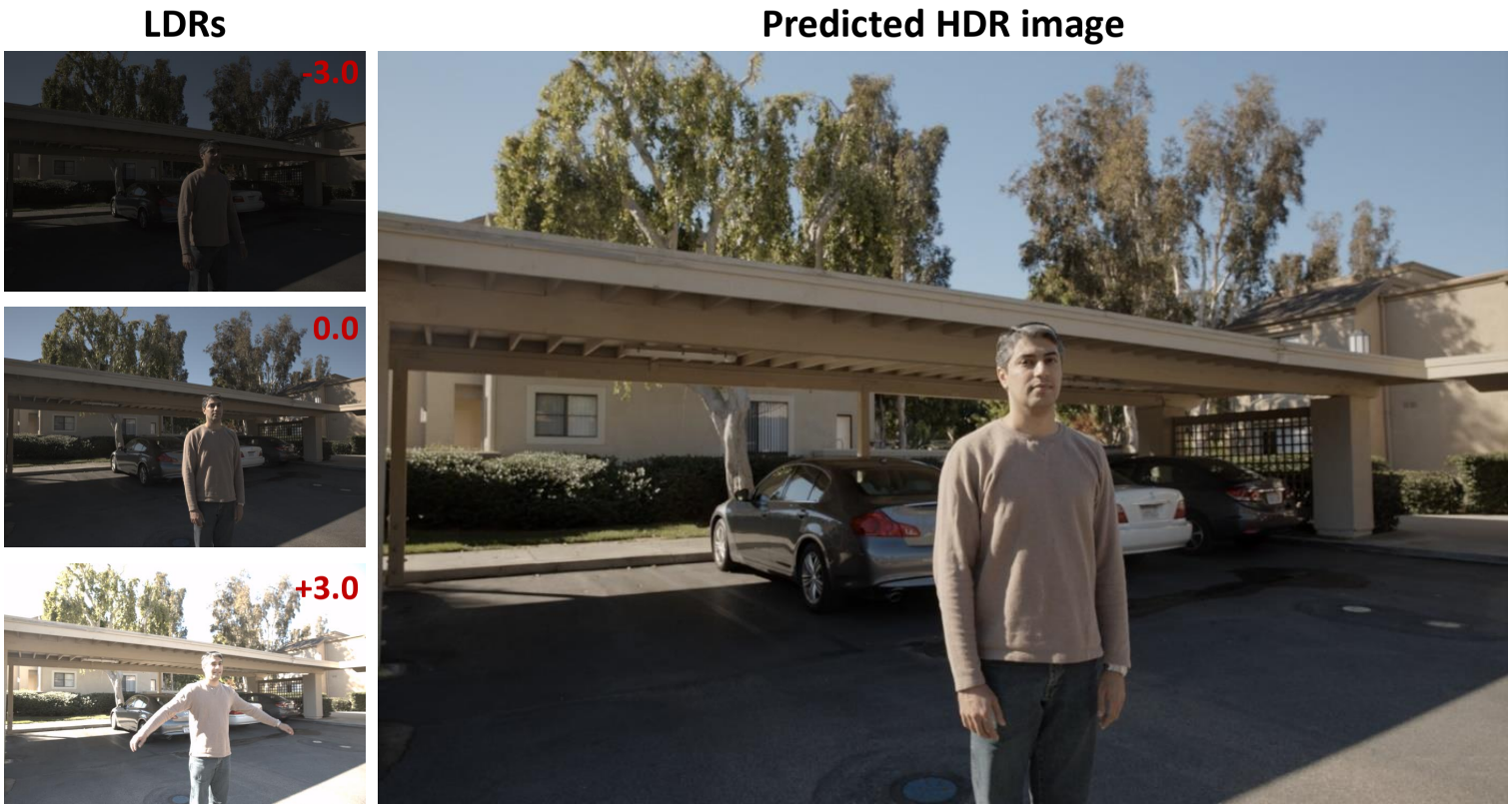

Figure 4.4: Another example from test dataset[5] which has [-3,0,+3] range of exposure values and the large motion.

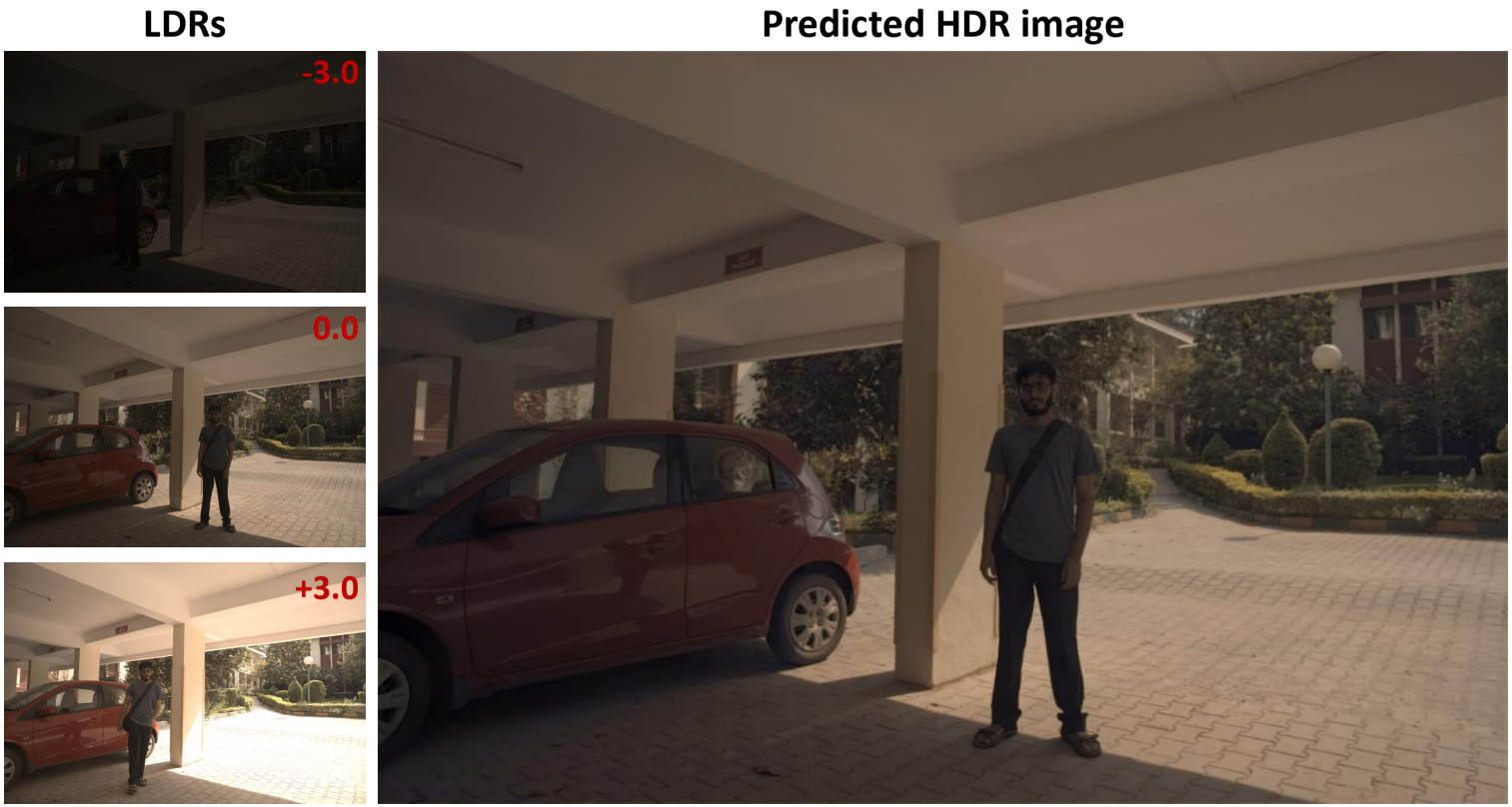

Figure 4.5: An example from a different prabhakar[40] test dataset which has [-3,0,+3] range of exposure values.

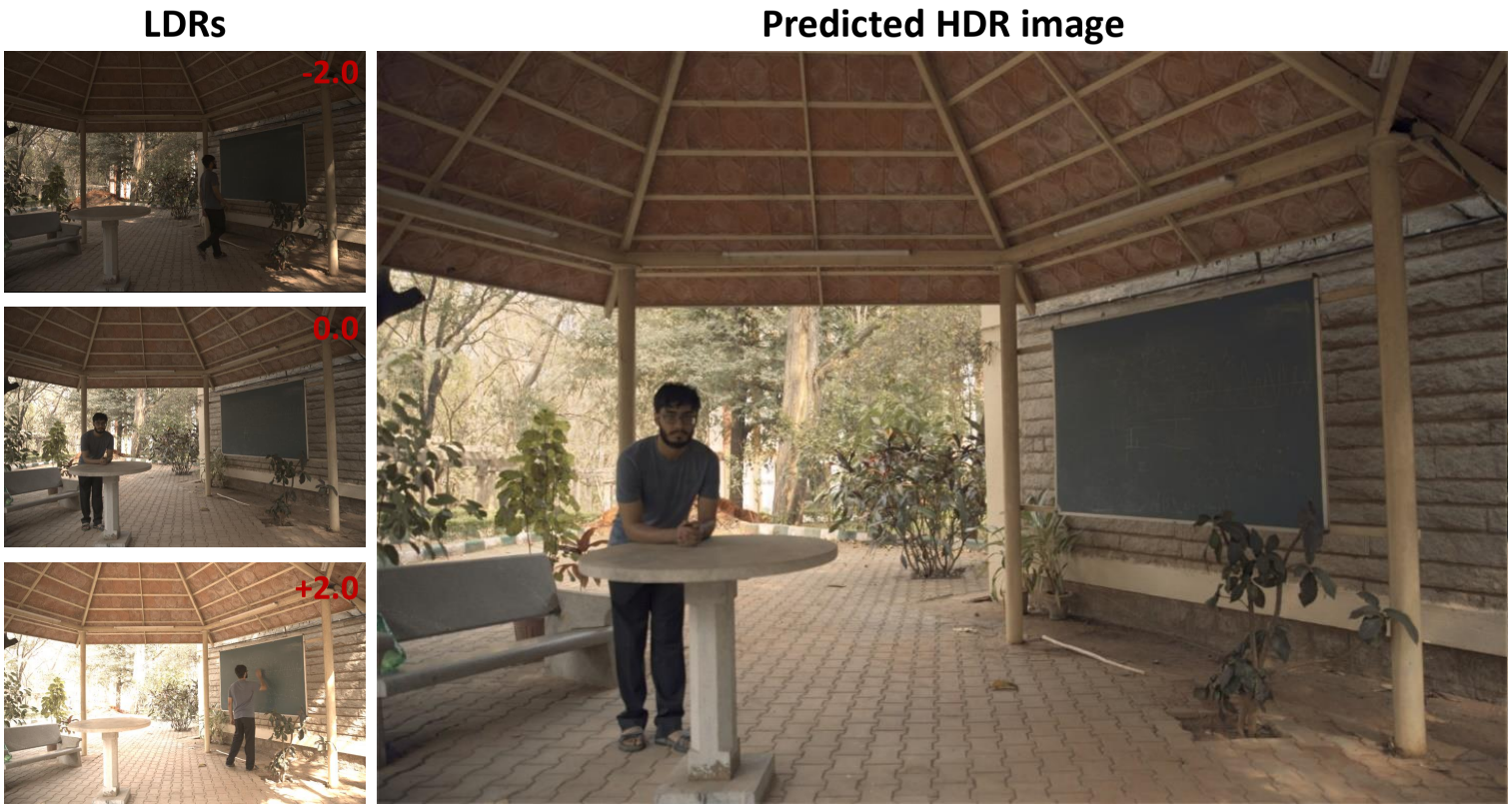

Figure 4.6: Another example from a different prabhakar[40] dataset which has [-2,0,+2] range of exposure values.

Chapter 5

Ablation Study

5.1 Impact of Proposed Modules

Table 5.1: The performance evaluation on network variants of DFGNet. The EAU denotes exposure attention unit and the SAU denotes spatial attention unit.

Method	PSNR-μ	PSNR-ℓ	HDR-VDP-2
Model 1(w/o EAU)	43.57	41.45	65.88
Model 2(w/o SAU)	43.92	41.59	**66.91**
Full	**44.19**	**41.89**	66.84

In this section, the effectiveness of the proposed modules is evaluated. The network without an exposure attention unit and the network without a spatial attention unit are trained and evaluated in the same scenario. Model 1 and Model 2 in Table 5.1 denote network variants of DFGNet and Full denotes original DFGNet. The exposure attention unit(EAU) is removed in Model 1, and the spatial attention unit(SAU) is removed in Model 2. For a fair comparison, the channel volume of each network variant is adjusted to make the network has the same parameter size. The psnr-μ of both variant Model 1 and Model 2 dropped compared to the original DFGNet. Especially, the performance dropped significantly when the exposure attention unit is removed. These

results prove that our proposed exposure attention unit and spatial attention unit in the DFG module boost HDR reconstruction performance of the network.

Chapter 6

Conclusion

In this dissertation, a disentangled feature-guided network is proposed 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.

Bibliography

[1] Yan, Q., Gong, D., Shi, Q., Hengel, A., Shen, C., Reid, I. & Zhang, Y. Attention-guided network for ghost-free high dynamic range imaging. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 1751-1760 (2019)

[2] Srikantha, A. & Sidibé, D. Ghost detection and removal for high dynamic range images: Recent advances. *Signal Processing: Image Communication*. **27**, 650-662 (2012)

[3] Jacobs, K., Loscos, C. & Ward, G. Automatic high-dynamic range image generation for dynamic scenes. *IEEE Computer Graphics And Applications*. **28**, 84-93 (2008)

[4] Grosch, T. Fast and robust high dynamic range image generation with camera and object movement. *Vision, Modeling And Visualization, RWTH Aachen*. pp. 277-284 (2006)

[5] Kalantari, N., Ramamoorthi, R. & Others Deep high dynamic range imaging of dynamic scenes.. *ACM Trans. Graph.*. **36**, 144-1 (2017)

[6] Kang, S., Uyttendaele, M., Winder, S. & Szeliski, R. High dynamic range video. *ACM Transactions On Graphics (TOG)*. **22**, 319-325 (2003)

[7] Zimmer, H., Bruhn, A. & Weickert, J. Freehand HDR imaging of moving scenes with simultaneous resolution enhancement. *Computer Graphics Forum*. **30**, 405-414 (2011)

[8] Yan, Q., Zhang, L., Liu, Y., Zhu, Y., Sun, J., Shi, Q. & Zhang, Y. Deep HDR imaging via a non-local network. *IEEE Transactions On Image Processing*. **29** pp. 4308-4322 (2020)

[9] Wang, X., Girshick, R., Gupta, A. & He, K. Non-local neural networks. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 7794-7803 (2018)

[10] Huang, X. & Belongie, S. Arbitrary style transfer in real-time with adaptive instance normalization. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 1501-1510 (2017)

[11] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*. (2014)

[12] Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. *International Conference On Medical Image Computing And Computer-assisted Intervention*. pp. 234-241 (2015)

[13] Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R. & Unger, J. HDR image reconstruction from a single exposure using deep CNNs. *ACM Transactions On Graphics (TOG)*. **36**, 1-15 (2017)

[14] Niu, Y., Wu, J., Liu, W., Guo, W. & Lau, R. HDR-GAN: HDR image reconstruction from multi-exposed ldr images with large motions. *IEEE Transactions On Image Processing*. **30** pp. 3885-3896 (2021)

[15] Mantiuk, R., Kim, K., Rempel, A. & Heidrich, W. HDR-VDP-2: A calibrated visual metric for visibility and quality predictions in all luminance conditions. *ACM Transactions On Graphics (TOG)*. **30**, 1-14 (2011)

[16] Wu, S., Xu, J., Tai, Y. & Tang, C. Deep high dynamic range imaging with large foreground motions. *Proceedings Of The European Conference On Computer Vision (ECCV)*. pp. 117-132 (2018)

[17] He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. *Proceedings Of The IEEE International Conference On Computer Vision*. pp. 1026-1034 (2015)

[18] Kingma, D. & Ba, J. Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*. (2014)

[19] Sen, P., Kalantari, N., Yaesoubi, M., Darabi, S., Goldman, D. & Shechtman, E. Robust patch-based hdr reconstruction of dynamic scenes.. *ACM Trans. Graph.*. **31**, 203-1 (2012)

[20] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 770-778 (2016)

[21] Lee, S., An, G. & Kang, S. Deep recursive hdri: Inverse tone mapping using generative adversarial networks. *Proceedings Of The European Conference On Computer Vision (ECCV)*. pp. 596-611 (2018)

[22] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. & Bengio, Y. Generative adversarial nets. *Advances In Neural Information Processing Systems*. **27** (2014)

[23] Endo, Y., Kanamori, Y. & Mitani, J. Deep Reverse Tone Mapping. *ACM Transactions On Graphics (Proc. Of SIGGRAPH ASIA 2017)*. **36** (2017,11)

[24] Liu, Y., Lai, W., Chen, Y., Kao, Y., Yang, M., Chuang, Y. & Huang, J. Single-image HDR reconstruction by learning to reverse the camera pipeline. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 1651-1660 (2020)

[25] Yan, Q., Gong, D., Zhang, P., Shi, Q., Sun, J., Reid, I. & Zhang, Y. Multi-scale dense networks for deep high dynamic range imaging. *2019 IEEE Winter Conference On Applications Of Computer Vision (WACV)*. pp. 41-50 (2019)

[26] Dong, C., Loy, C., He, K. & Tang, X. Image super-resolution using deep convolutional networks. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **38**, 295-307 (2015)

[27] Zhang, K., Zuo, W., Chen, Y., Meng, D. & Zhang, L. Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions On Image Processing*. **26**, 3142-3155 (2017)

[28] Kim, J., Lee, J. & Lee, K. Accurate image super-resolution using very deep convolutional networks. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1646-1654 (2016)

[29] Oh, T., Lee, J., Tai, Y. & Kweon, I. Robust high dynamic range imaging by rank minimization. *IEEE Transactions On Pattern Analysis And Machine Intelligence*. **37**, 1219-1232 (2014)

[30] Raman, S. & Chaudhuri, S. Reconstruction of high contrast images for dynamic scenes. *The Visual Computer*. **27**, 1099-1114 (2011)

[31] Park, S. & Kwon, J. Sphere generative adversarial network based on geometric moment matching. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 4292-4301 (2019)

[32] Liu, Z., Lin, W., Li, X., Rao, Q., Jiang, T., Han, M., Fan, H., Sun, J. & Liu, S. ADNet: Attention-guided deformable convolutional network for high dynamic range imaging. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition*. pp. 463-470 (2021)

[33] Isola, P., Zhu, J., Zhou, T. & Efros, A. Image-to-image translation with conditional adversarial networks. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 1125-1134 (2017)

[34] Wang, X., Chan, K., Yu, K., Dong, C. & Change Loy, C. Edvr: Video restoration with enhanced deformable convolutional networks. *Proceedings Of The IEEE/CVF Conference On Computer Vision And Pattern Recognition Workshops*. pp. 0-0 (2019)

[35] Fu, X., Huang, J., Zeng, D., Huang, Y., Ding, X. & Paisley, J. Removing rain from single images via a deep detail network. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 3855-3863 (2017)

[36] Nah, S., Hyun Kim, T. & Mu Lee, K. Deep multi-scale convolutional neural network for dynamic scene deblurring. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition*. pp. 3883-3891 (2017)

[37] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. & Others Pytorch: An imperative style, high-performance deep learning library. *Advances In Neural Information Processing Systems*. **32** (2019)

[38] Lim, B., Son, S., Kim, H., Nah, S. & Mu Lee, K. Enhanced deep residual networks for single image super-resolution. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops*. pp. 136-144 (2017)

[39] Joffre, G., Puech, W., Comby, F. & Joffre, J. High dynamic range images from digital cameras raw data. *ACM SIGGRAPH 2005 Posters*. pp. 72-es (2005)

[40] Prabhakar, K., Arora, R., Swaminathan, A., Singh, K. & Babu, R. A fast, scalable, and reliable deghosting method for extreme exposure fusion. *2019 IEEE International Conference On Computational Photography (ICCP)*. pp. 1-8 (2019)

초 록

다중 노출(Multiple-exposure) 하이 다이나믹 레인지(High Dynamic Range, HDR) 이미징은 각각 다른 노출 정도로 촬영된 다수의 로우 다이나믹 레인지(Low Dynamic Range, LDR) 이미지를 사용하여 하나의 HDR 이미지를 생성하는 것을 목표로 한다. 다중 노출 HDR 이미징은 두 가지 주요 문제점 때문에 어려움이 있는데, 하나는 입력 LDR 이미지들이 정렬되지 않아 결과 HDR 이미지에서 고스트 아티팩트(Ghosting Artifact)가 발생할 수 있다는 점과, 또 다른 하나는 LDR 이미지들의 과소노출(Under-exposure) 및 과다노출(Over-exposure) 된 영역에서 정보 손실이 발생한다는 점이다. 과거의 방법들이 고전적인 이미지 정렬 방법들(e.g., homography, optical flow)을 사용하여 입력 LDR 이미지들을 전처리 과정에서 정렬하 여 병합하는 시도를 했지만, 이 과정에서 발생하는 추정 오류로 인해 이후 단계에 악영향을 미침으로써 발생하는 여러가지 부적절한 아티팩트들이 결과 HDR 이미지에서 나타나고 있다.

본 심사에서는 피쳐 분해를 응용한 HDR 네트워크를 제안하여, 언급된 문제들을 경감하고자 한다. 구체적으로, 먼저 LDR 이미지들을 노출 피쳐와 공간 피쳐로 분해하고, 분해된 피쳐를 HDR 네트워크에서 활용함으로써 고품질의 HDR 이미지를 생성할 수 있도록 한다. 제안한 네트워크는 성능 지표인 PSNR-ℓ과 PSNR-μ에서 각각 41.89dB, 44.19dB의 성능을 달성함으로써, 기존 방법들보다 우수함을 입증한다.

주요어: 하이 다이나믹 레인지 이미징, 다중 노출 이미징, 피쳐 분해
학번: 2020-29425