Lightweight Attention-Guided Network with Frequency Domain Reconstruction for High Dynamic Range Image Fusion

박재현 이근택 조남익
서울대학교 전기정보공학부, 인공지능협동과정, 뉴미디어통신공동연구소
jaep0805@ispl.snu.ac.kr

Lightweight Attention-Guided Network with Frequency Domain Reconstruction for High Dynamic Range Image Fusion

Park, Jae Hyun Lee, Keuntek Cho, Nam Ik
Dept. of Electrical and Computer Eng., Interdisciplinary Program in Artificial Intelligence, Institute of New Media and Communications, Seoul National University

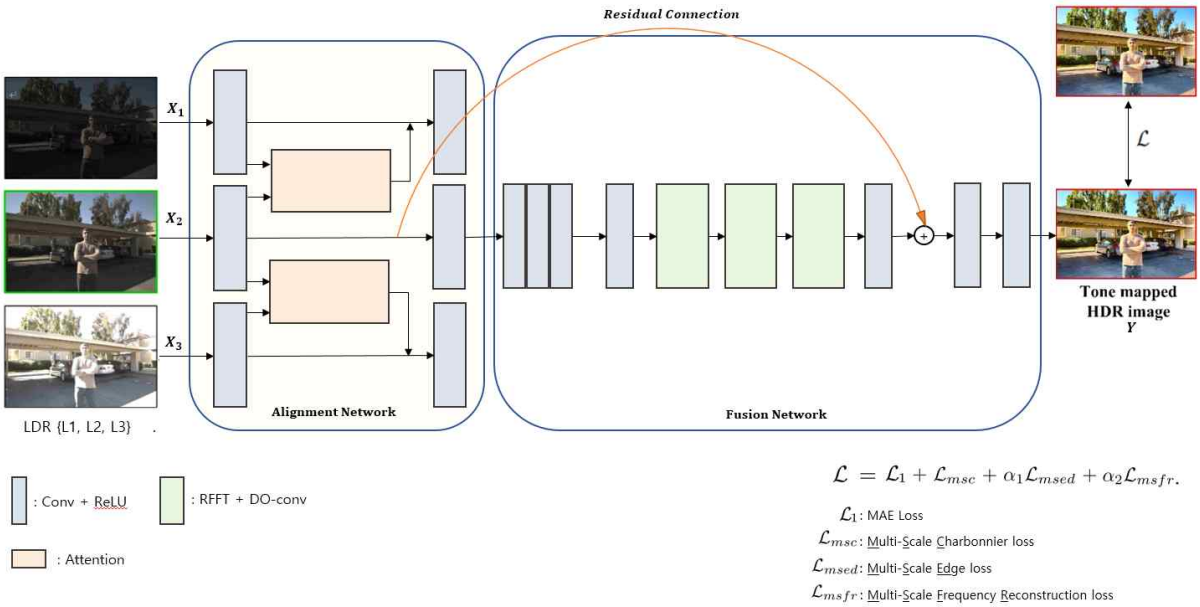
요약

Multi-exposure high dynamic range (HDR) image reconstruction, the task of reconstructing an HDR image from multiple low dynamic range (LDR) images in a dynamic scene, often produces ghosting artifacts caused by camera motion and moving objects and also cannot deal with washed-out regions due to over or under-exposures. While there has been many deep-learning-based methods with motion estimation to alleviate these problems, they still have limitations for severely moving scenes. They also require large parameter counts, especially in the case of state-of-the-art methods that employ attention modules. To address these issues, we propose a frequency domain approach based on the idea that the transform domain coefficients inherently involve the global information from whole image pixels to cope with large motions. Specifically we adopt Residual Fast Fourier Transform (RFFT) blocks, which allows for global interactions of pixels. Moreover, we also employ Depthwise Overparametrized convolution (DO-conv) blocks, a convolution in which each input channel is convolved with its own 2D kernel, for faster convergence and performance gains. We call this LFFNet (Lightweight Frequency Fusion Network), and experiments on the benchmarks show reduced ghosting artifacts and improved performance up to 0.6dB tonemapped PSNR compared to recent state-of-the-art methods. Our architecture also requires fewer parameters and converges faster in training.

1. Introduction

Given a scene with a dynamic range of luminance values, humans have the ability to perceive detail in the bright and dark regions of the scene simultaneously. Yet, when we take a photo of the scene with a digital photographic instrument, it is only able to capture a limited range of luminance values, leading to under/over-exposed regions in the LDR image. To reflect the visual perception of the human eye, HDR imaging that can represent a wider range of illumination has been actively researched through the lens of deep learning.

In this paper, we discuss Multi-Exposure HDR Image reconstruction, a task in which we reconstruct an HDR image given a series of LDR images at different exposure levels. More specifically, we deal with a dynamic scene, where the series of LDR images contain object motion. The goal is to reconstruct an HDR image that shares the same spatial information as the reference LDR image (commonly the medium exposure LDR image) while compensating the under/over-exposed regions of the LDR images with information from other LDR images at various exposure levels. However, significant ghosting and



$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{msc} + \alpha_1 \mathcal{L}_{msed} + \alpha_2 \mathcal{L}_{msfr}.$$

$\mathcal{L}_1$: MAE Loss
$\mathcal{L}_{msc}$: Multi-Scale Charbonnier loss
$\mathcal{L}_{msed}$: Multi-Scale Edge loss
$\mathcal{L}_{msfr}$: Multi-Scale Frequency Reconstruction loss

blurring artifacts arise when LDR images with motions are fused into a single HDR image. Hence, research in this domain largely focused on learning to align the LDR images; and learning to fuse LDR images to produce a realistic HDR image.

Among the various methods, attention-based methods have been a popular choice to guide the alignment process. Attention modules are primarily applied to evaluate the importance of different LDR image regions in obtaining the required realistic HDR image.

Yet the use of attention was only limited to the alignment process. On the other hand, the fusion process, where the HDR reconstruction takes place, was confined by a small receptive field leading to the lack of reconstruction ability for fine details on obscured and washed-out regions. Additionally, attention-based methods require heavy computations and large parameters, especially when inputs consist of multiple frames. In this paper, we propose a lightweight HDR reconstruction framework called LFFnet, that retains the use of attention in alignment and fuses the resulting feature maps in the frequency domain, making the network robust to large object motion and image saturation.

Our contribution can be summarized as the following.

1) Propose a frequency domain approach to provide global interaction to the fusion process, allowing the network to preserve rich image details and hallucinate fine details in obscured or saturated regions.

2) Significantly reduce parameter counts and convergence speed by using lightweight residual connections and DO-conv layers.

3) Experiments on LFFnet show SotA results on the Kalantari dataset, exceeding previous networks in both reconstruction ability (PSNR-μ) and parameter counts.

2. Method

Following the prescribed setting of [1], task of Multi-Exposure HDR Imaging can be formulated as follows: Given a set of three LDR images {L1, L2, L3} and a set of corresponding exposure values {E1, E2, E3}, the goal is to learn parameters θ for HDR Imaging network H, such that the output of the network produces a realistic HDR image Y. Among the three input images, the medium exposure LDR image L2 is preset to share the same spatial information as the ground-truth HDR image.

Morover, we apply gamma correction to each LDR images with its exposure values, to produce X: a set of the 3 LDR images, each concatenated with their HDR domain counterparts.

$$Y = H(\{X1,\ X2,\ X3\}\ ;\ \theta)$$

2.1 Model Architecture

Our HDR Imaging network consists of two

sub-networks: 1) The alignment network, which uses attention to guide the LDR's alignment process. 2) The fusion network, which merges the feature maps in the frequency domain, to hallucinate the lost details and reconstruct the HDR image.

Alignment Network

The goal of the alignment network is to account for the dynamics and the difference in exposures of the LDR images, and produce feature maps that aid in the reconstruction of the HDR image. We design the network to learn the attention maps of LDR images X1 and X3 in reference to the feature map of the reference image X2, and give weight to clear, visible regions of the LDR image that are saturated in the reference image X2. On the other hand, we give less weight to regions that are obscure and blurry while non-existent in the reference image X2. This allows the network to learn which regions of the LDR images to focus on to produce the best possible HDR image.

Fusion Network

The goal of the fusion network is to hallucinate obscured details and reconstruct the HDR image. The fusion network takes the concatenated feature maps from the Alignment Network as input, and merges them through a series of modules composed of RFFT blocks and DO-conv layers. Our contribution is significant in the fusion network, learning feature fusion in the frequency domain to better hallucinate fine details, significantly reducing parameter cost, and also improving the convergence speed in the training step.

2.2 Residual Fast Fourier Transform block

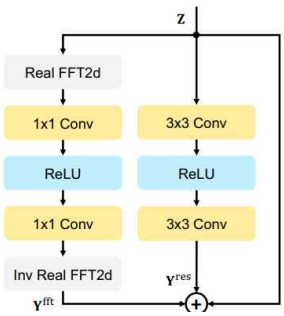


Figure 3 : RFFT block

Residual Fast Fourier Transform [3] blocks integrates high and low frequency residual information, and allows information to flow both locally and globally to aid the hallucination of details. As shown in Figure 2, the RFFT block adds a frequency domain convolutional branch to the Resblock. Additional Multi-scale Charbonnier loss, Edge loss and Frequency Reconstruction loss functions are used to train the RFFT block. As a result the loss function is defined as

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_{msc} + \alpha_1 \mathcal{L}_{msed} + \alpha_2 \mathcal{L}_{msfr}.$$

Relying on the local area information of the LDR images could be insufficient in reconstructing certain regions of the HDR image contaminated by saturation or moving objects. The spatial-invariant characteristic of the frequency domain residual connection allows the network to refer to different non-local regions of the LDR image complementary in the reconstruction of the detail of certain regions of the HDR image. Additionally, RFFT blocks allow propagation of features at a much lower parameter cost compared to state of the art fusion methods.

2.3 Depthwise Overparametrized Convolution block

In addition to the RFFT blocks, we use Depthwise Over-parametrized convolutional layer [4] as our replacement to the conventional convolutional layers. DO-Conv convolves each of the input channels separately, each with a different 2D kernel. The extra layer of depthwise convolution augments the network with overparameterization[4] and collapses to a single layer at inference time, maintaining the same computation time as a conventional convolution layer.

The extra linear transformation layer of DO-conv introduces more channel intra-dependencies to our network, allowing for active intra-channel feature

Table 1 : Comparison using PSNR-μ with previous SotA models on the Kalantari dataset

	PSNR-μ
Hu	32.1872
Kalantari	42.7423
Oh	27.351
Sen	40.9453
Wu	41.6377
AHDRnet	43,6172
Ours (LFFnet)	**44.2693**

Table 2 : Evaluation of the parameter count of the model at inference time & convergence epoch. Baseline model excludes the use of DO-conv and RFFT blocksand is instead identical to Yan et al [2].

	#Parameter	Convergence Epoch	PSNR-μ
1) Baseline	1515139	17290	43.6172
2) Baseline w/ RFFT	1038211	22320	44.2074
3) (Proposed Model) Baseline w/ RFFT + DO-conv	**1069315**	**17340**	**44.2693**

fusion. Additionally, DO-conv has optimized architecture for its learnable parameters, and hence leads to faster convergence for our computation-heavy network that uses attention to guide the alignment process.

3. Experiment

Dataset and Evaluation Metric

We evaluate LFFnet on the HDR Kalantari Dataset[1] and compare it to other state-of-the-art HDR models. The Kalantari dataset is composed of 74 training samples and 15 testing samples in a dynamic scene. We compare the reconstructed HDR image with ground truth HDR image and assess its quality with PSNR-μ averaged across all 15 testing samples.

Performance

As seen from Table 1, we make a quantitative comparison of our network with state-of-the-art networks on the Kalantari dataset. Previously, Yan et al [2] proposed an attention guided network with the best performance at 43.6172 db on PSNR-μ. We improve upon this PSNR-μ score by ~0.6dB through the use of RFFT and DO-conv blocks.

Parameter Count

Next we evaluate the effect of RFFT and DO-conv on the convergence speed and parameter count of the LFFnet. As seen from Table 2, replacing the baseline model's DDRB[2] with RFFT blocks, we use only 2/3 of the parameters, while also improving on the network's reconstruction ability through its global interactions. However, learning the network through complex Fourier and Laplacian domain, MSC, MSED, MSFR losses leads to difficulties on network convergence. We solve this through the adoption of DO-conv. As seen from the third model, we significantly reduce the time taken for convergence at a slight cost of increase in the number of parameters.

4. Conclusion

In this paper, we have proposed the LFFnet, a lightweight attention-guided network with frequency-domain reconstruction for HDR image fusion. Through additional HDR reconstruction in the frequency domain and optimization through DO-conv, we have overcome the stated limitations, and presented a lightweight HDR framework capable of generating high-quality HDR images with rich hallucinated details.

5. Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF), grant funded by the Korea government (MSIT) (2021R1A2C2007220).

Additionally this work was supported by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.2021-0-01343, Artificial Intelligence Graduate School Program (Seoul National University)]

Finally, this work was supported by the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University in 2022

6. References

[1] Kalantari, Nima Khademi, and Ravi Ramamoorthi. "Deep high dynamic range imaging of dynamic scenes." ACM Trans. Graph. 36.4 (2017): 144-1.

[2] Yan, Qingsen, et al. "Attention-guided network for ghost-free high dynamic range imaging." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019.

[3] Mao, Xintian, et al. "Deep Residual Fourier Transformation for Single Image Deblurring." arXiv preprint arXiv:2111.11745 (2021).

[4] Cao, Jinming, et al. "Do-conv: Depthwise over-parameterized convolutional layer." arXiv preprint arXiv:2006.12030 (2020).

[5] Wang, Lin, and Kuk-Jin Yoon. "Deep Learning for HDR Imaging: State-of-the-Art and Future Trends." IEEE Transactions on Pattern Analysis and Machine Intelligence (2021).