DEEP PRIOR GUIDED NETWORK FOR HIGH-QUALITY IMAGE FUSION

Jia-Li Yin[1], *Bo-Hao Chen*[1], *Yan-Tsung Peng*[2], *and Chung-Chi Tsai*[3]

[1]Department of Computer Science and Engineering, Yuan Ze University, Taoyuan, Taiwan
[2]Department of Computer Science, National Chengchi University, Taipei, Taiwan
[3]Qualcomm Technologies Inc., San Diego, USA

ABSTRACT

High dynamic range imaging requires fusing a set of low dynamic range (LDR) images at different exposure levels. Existing works combine the LDRs by either assigning each LDR a weighting map based on texture metrics at the pixel level or transferring the images into semantic space at the feature level while neglecting the fact that both texture calibration and semantic consistency are required. In this paper, we propose a novel encoder-decoder network consisting of a content prior guided (CPG) encoder and a detail prior guided (DPG) decoder for fusing the images at both the pixel level and feature level. Explicitly, the encoder constructed by the CPG layers includes the pyramid content prior to blend at the pixel level to transform the feature maps in the encoding layers. Correspondingly, the decoder comprises the DPG layers incorporated with the Laplacian pyramid detail prior to further boost the fusion performance. As the content and the detail priors are added to the network in a pyramid-structure manner, which provides fine-grained control to the features, both semantic consistency and texture calibration can be assured. Extensive experiments demonstrated the superiority of our method over existing state-of-the-art methods.

Index Terms— High dynamic range imaging, image fusion, neural convolution network

1. INTRODUCTION

High dynamic range imaging (HDRI) aims at producing informative and visually-pleasant photos by fusing a set of photos captured at different exposure levels. It can benefit various applications, such as photo restoration, and panoramic photography. Accordingly, this task has been widely researched and become an active topic of research in recent years [1].

High-quality HDRI usually requires not only texture calibration but also semantic consistency for fused HDR images. Existing approaches can be divided into two groups. The first one inspired by texture fusion techniques attempts to fuse the input LDRs at the pixel level [2, 3, 4, 5]. That is, such approaches usually compute the weights for each input LDR pixel-wise, and the fused image would be the weighted sum of these input images. While the details of images could be preserved, the lack of a high-level understanding of the image content often makes these approaches hard to maintain tone, brightness, and semantic consistency of the original input images. To solve this problem, the second group of approaches proposes to extract features from the semantic context of input images by deep convolutional neural networks (CNN) and then does feature fusion by stacked convolution operations [6, 7, 8, 9]. However, it is challenging to generate high-quality results from compact latent features, as image edges and details would be filtered out or removed by convolutions and pooling operations. An example of typical approaches for image fusion is given in Figure 1.

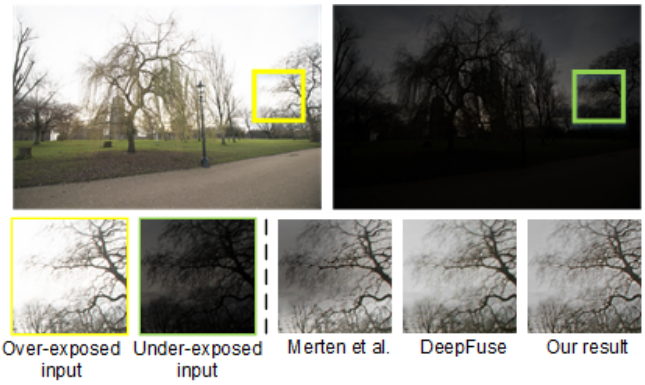

Fig. 1. Comparisons of image fusion results obtained by different methods. The first row displays two input images at different exposure levels for fusion. The second row shows the zoom-in regions of input images and fused results produced by different methods.

To ensure both texture calibration and semantic consistency, we propose to fuse the input LDRs at both the pixel and feature levels in a multi-scale fashion. We adopt a U-Net-like encoder-decoder structure to fuse input LDR images. To further guide the fusion process, for each layer of the encoder and decoder, the contents and details of the input images are broken down in a Gaussian-Laplacian pyramid manner and attached to the corresponding layers in the encoder and decoder. To this end, we present a new network architecture

978-1-7281-1331-9/20/$31.00

with the content prior guided and detail prior guided layers for fusing LDR images into an HDR image. These layers are constructed based on the pyramid content prior and the pyramid detail prior of the LDR images. Our network adopts deep supervision to guide the training of the fusion network, whose loss function is multi-scale L2 losses and adversarial loss. Experiments on an HDR dataset, including several natural images, demonstrate that our network generates higher-quality HDR results for the cases of two-exposure image fusion than existing ones. An example result is shown in the last column of Figure 1. The main contributions of this paper can be summarized as follows:

- Presenting a encoder-decoder image fusion network that consists of content prior guided and detail prior guided layers to fuse LDR images at both the pixel level and feature level.
- Exploiting a hybrid loss function that considers the pixel and feature distance to assure fusion quality.

2. RELATED WORK

HDRI by pyramid-based methods. Pyramid-based fusion methods generally compute weight maps to fuse input images based on the potential contribution of each pixel. Burt *et al*. [10] used the Laplacian pyramid to compute the weight for each pixel based on local energy and correlation between the pyramid levels. Mertens *et al*. [2] proposed to compute the weight maps using quality metrics such as contrast, saturation, and well-exposedness. However, the fused results usually suffer from halo artifacts due to the weights. To overcome this problem, researchers have tried several approaches to improve fusion results by employing various filters to smooth the weighting maps or enhance the image details. Specifically, Li *et al*. [4] proposed to use the weighted guided image filter (WGIF) to smooth the weighting maps and apply a detail extraction module to refine the image details. Kou *et al*. [5] developed an edge-preserving gradient-domain guided image filter (GGIF) to preserve the edges in the images. Pyramid-based fusion methods for HDRI can generate sharp results with rich details. However, it is difficult to maintain semantic consistency by these methods due to the lack of a high-level understanding of images.

HDRI by deep learning methods. Deep learning methods for HDRI encode the input images into feature space, fusing these inputs at the feature-level, and decode the fused features back into an image. In recent years, impressive results have been achieved by deep learning methods. In [6], Prabhakar *et al*. first adopted deep convolutional neural networks (CNN) for two-extreme-exposure fusion, with a non-reference image quality metric defined in [11] as its loss function. Chen *et al*. [7] utilized generative adversarial network (GAN) framework and proposed context encoder and exposure encoder to capture the context and exposedness features for obtaining a transferred exposure image. In the HDR fusion GAN, the inputs are then fused into the final HDR image. These models can produce image with semantic consistency, however, the fused HDR images are not sharp enough as the image edges and details are largely lost during the convolutional and pooling operations.

3. METHOD

In this paper, we propose a novel image fusion framework based on both the pixel-level and feature-level for HDRI fusion. Specifically, we create the CPG and DPG layers in the framework for better fusing the input LDR images by incorporating the features of multi-scale image details. As shown in Figure 2, our image fusion model consists of three parts: a CPG encoder, a DPG decoder, and a discriminator. The whole network is built upon a U-Net-like structure [12], which can encode two input LDR images in the feature space and decode the features back to the HDR image. As the compact latent features are extracted from the semantics of input images in the CPG encoder, the DPG decoder refines the details to improve the fusion result. The feature maps at each scale in the decoder are used to compute the multi-scale L2 loss to guide the fusion process for a better result.

3.1. Network architecture

Content prior guided encoder. The proposed CPG layers are used in our encoder for semantic content fusion. Once the original images are transformed into the feature space, the proposed encoder strengthens the semantic contents by adding the image contents decomposed by the pyramid through the CPG layer at each scale. Concretely, the CPG layer learns a non-linear mapping function M that outputs semantic feature maps based on the content prior. Given an encoder of L layers, the features ϕ^l produced by lth CPG block based on given content prior φ_c^l are denoted as:

$$\phi^l = f\Big(\psi_c^l \oplus \phi^{l-1}\Big), \quad \psi_c^l = M(\varphi_c^l), \tag{1}$$

where f denotes the convolution operation, and $\oplus$ denotes feature concatenation. In CPG layers, the learned features adaptively influence the encoding process by integrating the pixel-wise content prior to each intermediate feature map in the fusion network, which substantially improves the semantic coherence in the final fused result.

Detail prior guided decoder. After the latent features from the CPG encoder are obtained, the image reconstruction is carried out by the proposed DPG decoder. Similarly, we progressively integrate feature maps derived from the detail prior through DPG layers to provide fine-grained control to the features. Moreover, the skip connection is also adopted in our network to ensure structural stability. The lth DPG layer takes as the input the features from the last layer ϑ^{l-1}, the features

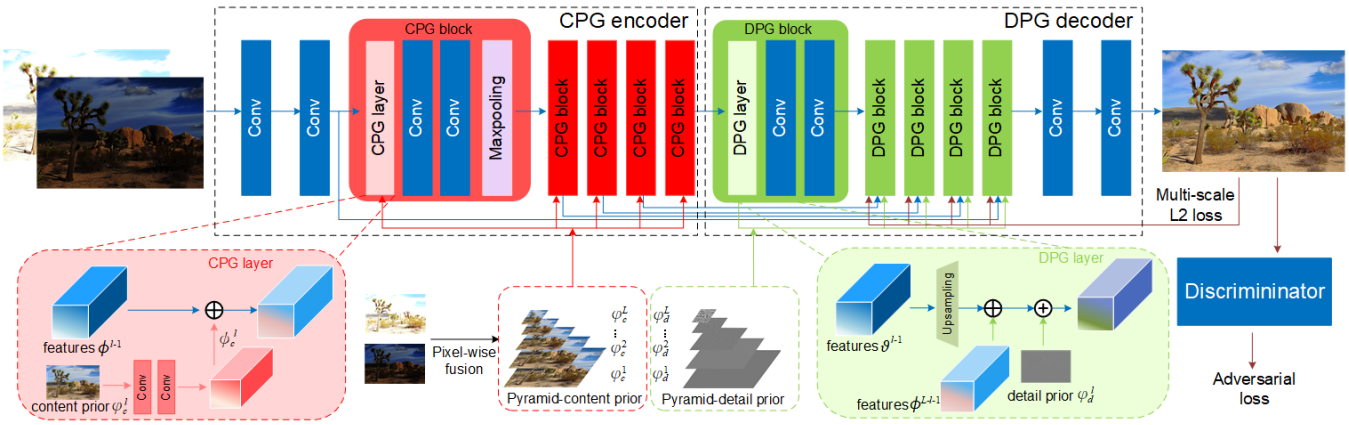

Fig. 2. Our proposed fusion network is composed of three components: a CPG encoder, a DPG decoder and a discriminator. First, the input images are encoded into feature space by several convolutions. The CPG encoder improves the encoding effectiveness by progressively adding content prior fused at pixel-level through CPG layers. Second, the DPG decoder takes as input the compact features and provide fine-grained control to features by DPG layers during the decoding process. The whole network is under deep-supervised learning and optimized using a multi-scale L2 loss and an adversarial loss.

from the encoder ϕ^{L-l+1}, and the features by detail prior φ_d^l, and operates as follows:

$$\vartheta^l = f\Big(\big(u(\vartheta^{l-1}) \oplus \phi^{L-l+1}\big) + \varphi_d^l\Big), \tag{2}$$

where u denotes the upsampling operation, and L denotes the total number of layers in the encode part. On one hand, the features generated by the DPG layer have low-level information for detail refinement. On the other hand, the features obtained from each the DPG block are leveraged for computing the deep-supervised multi-scale L2 loss to guide training of the model.

Pixel-wise fused contents and details. We follow the state-of-the-art approach [2] to conduct pixel-wise fusion and generate the content and detail prior by the Gaussian-Laplacian pyramid decomposition. Given an image pair $\{I^1,\ I^2\}$, the weighting maps $W^i,\ i \in \{1, 2\}$ are first calculated by considering quality metrics including contrast, saturation and well-exposedness, as described in [2]. Next, the input images and weight maps are decomposed using the Gaussian-Laplacian pyramids.

Let us denote the Guassian pyramids of weighting maps as $G\{W^i\}$ and the Laplacian pyramids of input images as $L\{I^i\}$, the lth level of fused pyramid detail prior is then computed as follows:

$$\varphi_d^l = \sum_{\forall i} G\{W^i\}^l L\{I^i\}^l. \tag{3}$$

For the pyramid content prior, we use the Gaussian pyramids to decompose the input images and the pyramid content prior is then computed as the weighted sum of each level, which can be represented as follows:

$$\varphi_c^l = \sum_{\forall i} G\{W^i\}^l G\{I^i\}^l. \tag{4}$$

3.2. Loss function

Inspired by [13], we apply a deep-supervised multi-scale L2 loss and an adversarial loss to train our network. The multi-scale L2 loss is the MSE [14] between the output at each scale of the DPG decoder and the corresponding resized target HDR image, which can be described as follows:

$$L_p = \sum_{l=1}^{L} \left\| f\left(\vartheta^l\right) - I_{\text{target}} \right\|_2. \tag{5}$$

Similar to [13], we use a 1×1 convolutional operation to merge the intermediate features into an image, and the target HDR image is resized to the same size as that merged image at each scale. The adversarial loss defined from the GAN [15] is also employed to reinforce our network to favor the results in the HDR manifold. The loss is represented as:

$$\begin{aligned} L_D = &-\mathbb{E}_{I_{\text{target}} \sim p_{HDR}} \log D\big(I_{\text{target}}\big) + \\ &\mathbb{E}_{I \sim p_{LDR}} \log\Big(1 - D\big(F(I)\big)\Big). \end{aligned} \tag{6}$$

4. EXPERIMENTS

The authors from the universities in Taiwan completed the experiments on the datasets.

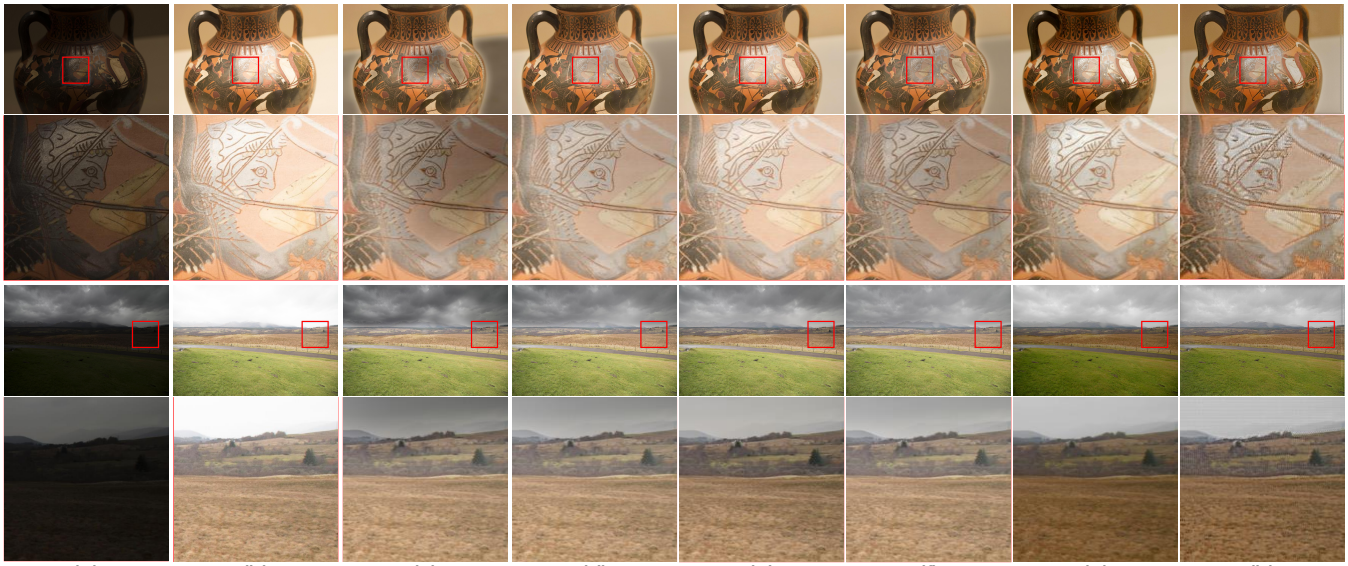

Fig. 3. Qualitative comparisons with other state-of-the-art methods on the SICE database. (a) Under-exposed images, (b) over-exposed images, (c)-(g) the results produced by the methods of Mertens *et al*., Kou *et al*., Li *et al*., Yang *et al*., and Prabhakar *et al*., respectively, and (h) our results.

4.1. Implementation details

We collect 871 image pairs (i.e., an over-exposed image and an under-exposed image) from the SICE dataset [16], where 712 pairs are chosen as the training dataset, and the other 159 pairs are used as the validation dataset. For network training, we used Adam optimizer with a fixed learning rate of 0.0001. We alternatively optimized our fusion network and the discriminator for 500 epochs. The training process takes about 72 hours on a computer with a single NVIDIA TITAN RTX GPU.

In the following, we evaluate the proposed method for two-exposure image fusion for HDR rendering. We first conduct qualitative and quantitative assessments for state-of-the-art methods and the proposed method. Next, we discuss several variants of our method.

4.2. Comparisons with other state-of-the-art methods

The qualitative and quantitative comparisons are obtained on the SICE testing dataset [16]. The compared state-of-the-art methods include Mertens *et al*. [2], Li *et al*. [4], Kou *et al*. [5], Prabhakar *et al*. [6], and Yang *et al*. [3].

Qualitative comparison. Figure 3 shows the qualitative results of different image fusion methods. Figure 3 (a) and (b) are the over-exposed and under-exposed images. Figure 3 (c) shows Mertens *et al*.'s results, where the details of the image are preserved; however, halo artifacts also come out, and the results fail in maintaining brightness consistency, which is often inevitable for most existing pixel-level fusion algorithms. The Kou *et al*.'s and Li *et al*.'s methods employ filters to smooth the weighting maps to avoid the halo artifacts. However, Kou *et al*.'s results are under-exposed [see Figure 3 (d)] while Li *et al*.'s results are over-exposed [see Figure 3 (e)]. Yang *et al*.'s method generates an additional intermediate image at a medium exposure level for fusion. This method improves the brightness consistency; however, the results sometimes are unsatisfying because the intermediate image generated based on the intensity estimation function does not always lead to a better fusion result. The color distortions and artifacts occur in the results of Figure 3 (f). From Figure 3 (g), although we can see the improvement in terms of brightness consistency. Moreover, the details are blurred, and the color looks faded in the results. By contrast, the proposed fusion network can produce better results with consistent brightness and calibrated details, as shown in Figure 3 (h).

Table 1. Quantitative comparison of state-of-the-art algorithms for two-exposure fusion

Method	TMQI		
	Q	Fidelity	Naturalness
Mertens *et al*. [2]	0.7904	0.6311	0.4587
Kou *et al*. [5]	0.7989	0.6563	0.5497
Li *et al*. [4]	0.7916	**0.6702**	0.4367
Yang *et al*. [3]	0.8054	0.6665	0.5147
Prabhakar *et al*. [6]	0.7887	0.6600	0.4264
Ours	**0.8056**	0.6618	**0.5618**

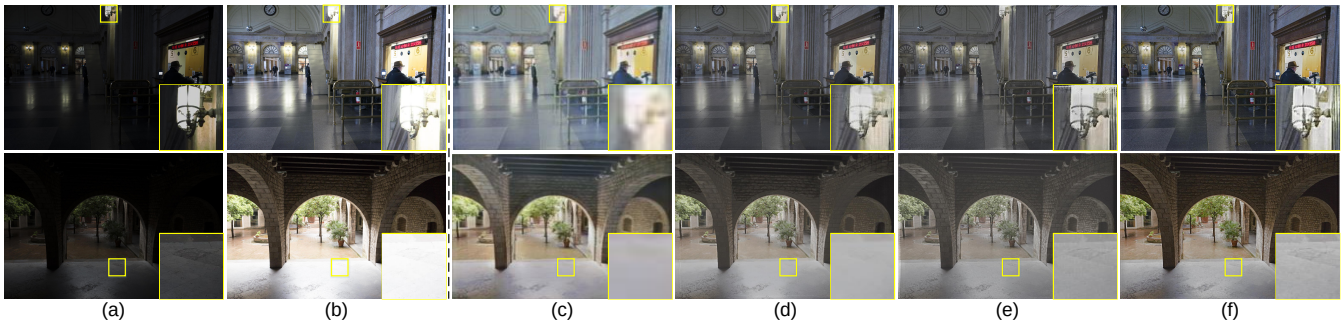

Fig. 4. Qualitative comparisons with other state-of-the-art methods on SICE database. (a) input under-exposed image, (b) input over-exposed image, (c)-(e) the results produced by architecture of a simple encoder-decoder, the CPG encoder-decoder, the encoder-DPG decoder, respectively, and (f) our results.

Table 2. Quantitative comparison of state-of-the-art algorithms for two-exposure fusion

Method	TMQI		
	Q	Fidelity	Naturalness
Encoder-decoder	0.7272	0.5755	0.3300
CPG encoder-decoder	0.7781	0.6105	0.4515
Encoder-DPG decoder	0.8014	0.6113	**0.5767**
CPG encoder-DPG decoder	**0.8056**	**0.6618**	0.5618

Quantitative comparison. We also conduct a quantitative comparison using the testing dataset. The Tone Mapping Quality Index (TMQI) [17], which evaluates an HDR image by computing the multi-scale signal fidelity and naturalness of the image, is used as the metric to quantify the results of each method. As we can see in Table 1, Prabhakar *et al.*'s method achieves the lowest Q value among all the compared methods. The low Q value is mainly because of the low naturalness value caused by the blurred results. Mertens *et al.*'s method has a higher Q value but has the lowest Fidelity due to the artifacts, as mentioned earlier. The Kou *et al.*'s and Li *et al.*'s methods improve the performance in terms of Q value by employing filters to avoid artifacts. In contrast, our method achieves the highest Q value by simultaneously satisfying semantic consistency and texture calibration.

4.3. Analysis

To further validate the superiority of the proposed method, we analyze the effectiveness of different components of the proposed network by visualizing the performance or quantitative comparison as follows.

Effectiveness of CPG encoder and DPG decoder. To verify the effectiveness of the proposed CPG and DPG layers, we experiment to isolate the effect of these two layers. Specifically, we construct a fusion network with a simple encoder-decoder structure, a CPG encoder-decoder structure, and an encoder-DPG decoder structure, respectively. We train the three networks using the same training dataset with the proposed loss function.

Fig. 5. Results produced by the DPG decoder at each scale. (a) The input images, (b) from the left to right and top to bottom: images produced by the first to the last DPG block in the decoder.

Figure 4 shows some examples of these variations. As we can see in Figure 4 (c), the simple Encoder-Decoder structure generates blurry results. Using the CPG layers in the encoder gives better results, but the results are still little blurry, and the image details are missing. For the structure with the DPG decoder, as can be seen in Figure 4 (e), adding the detail prior renders much sharper edges and finer details, but introduces undesired visual artifacts. In contrast, using both the CPG encoder and DPG decoder reduces these artifacts and produces more vivid results with rich and informative details [see Figure 4 (f)].

We also quantify these observations using the TMQI metric on the SICE testing dataset, as reported in Table 2. The simple encoder-decoder network achieves the lowest scores, indicating that the fused images include few image details. The CPG encoder-decoder network has higher scores due to the addition of content prior. Adding the DPG layers brings fine-grained control to the features, and the encoder-DPG decoder outperforms the encoder-decoder network and CPG encoder-decoder network in terms of the Q value. Combing

both CPG encoder and DPG decoder works the best with the highest Q value.

Effectiveness of multi-scale L2 loss. The multi-scale L2 loss is used to refine the output in the proposed decoder. We demonstrate the effectiveness of multi-scale loss by visualizing the predictions of the decoder in each scale. Figure 5 depicts a qualitative example of results produced by each block in the decoder. As can be seen in Figure 5 (b), the result of each layer looks more natural and sharper as it gets closer to the last layer.

5. CONCLUSIONS

In this paper, we propose a CPG encoder-DPG decoder network to generate results with calibrated texture and consistent semantic for image fusion. The proposed network improves the encoding and decoding effectiveness by adding the content and detail prior fused at the pixel-level in a pyramid pathway. The extensive experimental results show the efficiency and superiority of the proposed network.

6. ACKNOWLEDGEMENT

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant Nos. MOST 108-2221-E-155-034-MY3, and MOST 107-2221-E-155-052-MY2, and funded in part by the MOST AI Biomedical Research Center under Grant No. MOST 109-2634-F-019-001, in part by MOST 109-2634-F-004-001 through Pervasive Artificial Intelligence Research (PAIR) Labs, and in part by Qualcomm through a Taiwan University Research Collaboration Project.

7. REFERENCES

[1] 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 *CVPR*, 2019.

[2] T. Mertens, J. Kautz, and F. V. Reeth, "Exposure fusion," in *proceedings of Pacific Graphics*, 2007.

[3] Y. Yang, W. Cao, S. Wu, and Z. Li, "Multi-scale fusion of two large-exposure-ratio images," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1885–1889, Dec 2018.

[4] Z. Li, Z. Wei, C. Wen, and J. Zheng, "Detail-enhanced multi-scale exposure fusion," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1243–1252, March 2017.

[5] F. Kou, Z. Li, C. Wen, and W. Chen, "Multi-scale exposure fusion via gradient domain guided image filtering," in *ICME*, July 2017, pp. 1105–1110.

[6] K. Ram Prabhakar, V. Sai Srikar, and R. Venkatesh Babu, "Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *ICCV*, 2017.

[7] Y. Chen, M. Yu, K. Chen, G. Jiang, Y. Song, Z. Peng, and F. Chen, "New stereo high dynamic range imaging method using generative adversarial networks," in *ICIP*, Sep. 2019, pp. 3502–3506.

[8] K. R. Prabhakar, R. Arora, A. Swaminathan, K. P. Singh, and R. V. Babu, "A fast, scalable, and reliable deghosting method for extreme exposure fusion," in *ICCP*, May 2019, pp. 1–8.

[9] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, and Z. Wang, "Deep guided learning for fast multi-exposure image fusion," *IEEE Trans. Image Process.*, pp. 1–1, 2019.

[10] P. J. Burt and R. J. Kolczynski, "Enhanced image capture through fusion," in *ICCV*, 1993.

[11] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov 2015.

[12] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015.

[13] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo, "Learning pyramid-context encoder network for high-quality image inpainting," in *CVPR*, 2019.

[14] J. Yin, B. Chen, and Y. Li, "Highly accurate image reconstruction for multimodal noise suppression using semisupervised learning on big data," *IEEE Trans. Multimedia*, vol. 20, no. 11, pp. 3045–3056, Nov 2018.

[15] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017.

[16] J. Cai, S. Gu, and L. Zhang, "Learning a deep single image contrast enhancer from multi-exposure images," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 2049–2062, April 2018.

[17] H. Yeganeh and Z. Wang, "Objective quality assessment of tone-mapped images," *IEEE Trans. Image Process.*, vol. 22, no. 2, pp. 657–667, Feb 2013.