AUTHOR GUIDELINES FOR ICME 2020 PROCEEDINGS

Anonymous ICME submission

ABSTRACT

The abstract should appear at the top of the left-hand column of text, about 0.5 inch (12 mm) below the title area and no more than 3.125 inches (80 mm) in length. Leave a 0.5 inch (12 mm) space between the end of the abstract and the beginning of the main text. The abstract should contain about 100 to 150 words, and should be identical to the abstract text submitted electronically along with the paper cover sheet. All manuscripts must be in English, printed in black ink.

***Index Terms*—** One, two, three, four, five

1. INTRODUCTION

High dynamic range imaging (HDRI) techniques aim at producing informative and visually-pleasant photo by fusing a set of photos captured at different light conditions. This task has been widely researched and become an active topic since HDRI techniques can benefit various applications such as photo restoration and enhancement.

High-quality HDRI usually requires not only texture calibration but also semantic consistency for fused HDR image. Existing approaches can be roughly divided into two groups. The first group inspired by texture fusion techniques attempts to fuse the input LDRs at pixel level [X-X]. Specifically, such approaches usually compute the weights for each input LDR in pixel wise and the fused image would be the weighted sum of these input images. While the details of images can be well-preserved, the lack of high-level understanding for the image content makes such approaches often fail in maintaining brightness and semantic consistency. To solve this problem, the second group of approaches proposes to encode the semantic context of input images into feature space by deep convolutional neural networks (CNN) and then conduct feature fusion by stacked convolution operations [X-X]. However, it remains challenging to generate high-quality results from a compact latent feature, as the image edges and details can be usually smoothed by convolutions and pooling operations. An example of typical approaches for image fusion is given in Figure 1.

To ensure both texture calibration and semantic consistency, we propose to fuse the input LDRs at both pixel and feature levels in a progressive multi-scale fashion. First, we adopt a U-Net-like encoder-decoder network as backbone to encode the input LDRs into high-level features and decode the features back into an image. Second, once the features have been encoded from images, the contents and details of images are added to the features in a pyramid pathway during the encoding and decoding process, respectively. To this end, we present a new approach known as Content Prior Guided (CPG) and Detail Prior Guided (DPG) that are capable of altering the behavior of a fusion network through transforming the features of some intermediate layers of the whole network. Specifically, the CPG and DPG layers are conditioned on pyramid content prior and pyramid detail prior of the image fused at pixel-level, respectively. Third, our proposed network architecture naturally accommodates deep supervision to guide the training of fusion network. The network is under deep supervision and optimized by pyramid L2 losses and adversarial loss.

To the best of our knowledge, the proposed fusion network is the first work that is able to fuse images at both pixel-level and feature-level. The main contributions of this paper can be summarized as follows:

- A novel image fusion network which consists of a con-

Fig. 1. Typical approaches for image fusion. 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 results produced by different methods. Pixel-level based method of Mertens *et al*. ensures texture details but fails in maintaining semantic consistency, while feature-level based method of Prabhakar *et al*. produces semantic-coherent image with smoothed details. Compared with these methods, our approach can satisfy both texture calibration and semantic consistency.

tent prior guided encoder and a detail prior guided involved decoder is proposed to fuse the images at both pixel level and feature level.

- A hybrid loss that measures the pixel distance and feature distance is employed to train our network on pixel level and feature level.

2. RELATED WORK

HDRI by pyramid-based methods. Pyramid-based methods aim to compute the proper weights for each input image based on the potential contribution of each pixel. Burt et al. [X] used Laplacian pyramid to compute the weights for each pixel by local energy and correlation between the pyramids levels. Mertens et al. [X] proposed to compute the weights using simple quality metrics such as contrast, saturation, and well-exposedness. However, this suffers from halo artifacts due to the weights. To overcome this problem, a number of approaches try to improve the fusion performance by employing various filters to smooth the weighting maps or enhance the image details. Specifically, Li et al. [X] proposed to use weighted guided image filter (WGIF) to smooth the weighting maps and apply a detail extraction module to refine the image details. Kou et al. [X] developed an edge preserving gradient domain guided image filter (GGIF) to preserve the edges in the images. Pyramid-based methods for HDRI are able to generate sharp results with rich details. However, it is difficult to maintain semantic consistency by these methods due to the lack of 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 [X], Prabhakar et al. first adopted deep convolutional neural networks (CNN) for two-extreme-exposure fusion, with a non-reference image quality metric defined in [X] as loss function. Chen et al. [X] utilized generative adversarial network (GAN) framework and proposed context encoder and exposure encoder to capture the context and exposed ness 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 an image fusion framework based on both pixel-level and feature-level for the task of multiple exposure fusion. Specifically, we construct CPG layers and DPG layers to alter the convolutional operations in the fusion network through transforming the features of multi-scale image details. As shown in Fig. 2, our image fusion model consists of three parts: a CPG layer involved encoder, a DPG layer involved decoder, and a discriminator. The whole network is built upon a U-Net structure, which can encode the input LDRs into feature space and decode the fusion features back into the final HDR image. As the compact latent features encode the semantics of the input image in the content encoder, the pyramid-detail decoder further improves the fusion effectiveness by refining the details. The feature maps at each scale in the decoder are used to compute multi-scale L_2 loss for further refine the prediction output.

3.1. Network architecture

Encoder. In order to improve the effectiveness of encoding, the CPG layer based encoder is proposed for semantic content fusion before decoding. Once the original images are mapped into the feature space, the proposed encoder strengthens the semantic contents by adding pyramid-decomposed image contents through CPG layer at each scale. Specifically, the CPG layer learns a mapping function M that outputs semantic feature maps based on 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:

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

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

Decoder. After the latent features from the encoder are obtained, the image reconstruction is carried out by the proposed decoder. Similarly, we progressively integrate feature maps derived from the detail prior through DPG layer to provide fine-grained control to the features. Moreover, the skip connection is also adopted in our network to ensure structure stability. Hence, the lth DPG layer takes as input the features from the last layer ϑ^{l-1}, the features from the encoder ϕ^{L-l+1}, and the features by detail prior ψ_d^l, and operates as follows:

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

Where u denotes the unsampling operation, and L denotes the total number of layers in the encode part. On one hand, the features generated by DPG layer encode more low-level information for detail refinement. On the other hand, the features obtained from each DPG block are leveraged for computing the deep-supervised pyramid L_2 loss to train the network.

Pixel-wise fused contents and details as prior. We follow the state-of-the-art approach [X] to conduct pixel-wise fusion

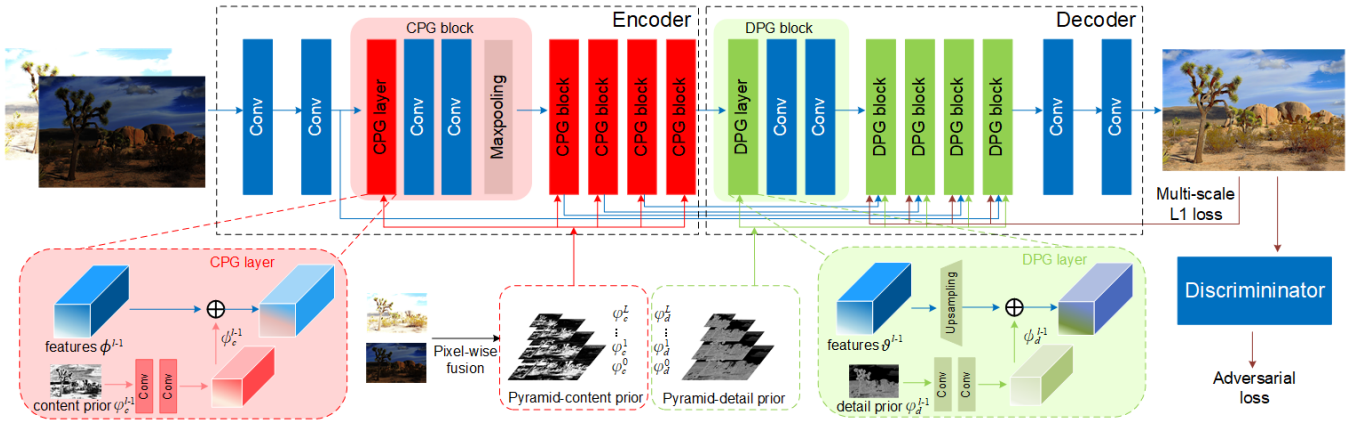

Fig. 2. Our proposed fusion network is composed of three components: encoder, decoder and discriminator. First, the input images are encoded into feature space by several convolutions. The encoder improves the encoding effectiveness by progressively adding content prior fused at pixel-level through CPG layers. Second, the 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 pyramid L2 loss and an adversarial loss.

and generate the content and detail prior through 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 [X]. Then the Gaussian pyramids and the Laplacian pyramids are employed to decompose the weighting maps and input images, respectively. 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 [X-X], we apply a deep-supervised pyramid L2 loss and an adversarial loss to train our network in pixel space and feature space, respectively. The pyramid L2 loss measures the MSE distance between the output from each DPG block 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 [X], we use a 1×1 convolutional operation to decode the features into an image, and the target HDR image is resized into the same size with the decoded image at each scale.

The adversarial loss defined from GAN [X] is also employed to force the fusion network to favor the results in the HDR manifold, the loss is represented as:

$$L_D = -\mathbb{E}_{I_{\text{target}} \sim p_{HDR}} \log D\left(I_{\text{target}}\right) + \mathbb{E}_{I \sim p_{LDR}} \log\left(1 - D\left(F(I)\right)\right). \tag{6}$$

4. EXPERIMENTS

The paper title (on the first page) should begin 1.38 inches (35 mm) from the top edge of the page, centered, completely capitalized, and in Times 14-point, boldface type. The authors' name(s) and affiliation(s) appear below the title in capital and lower case letters. Papers with multiple authors and affiliations may require two or more lines for this information.

5. TYPE-STYLE AND FONTS

To achieve the best rendering both in the proceedings and from the CD-ROM, we strongly encourage you to use Times-Roman font. In addition, this will give the proceedings a more uniform look. Use a font that is no smaller than nine point type throughout the paper, including figure captions.

If you use the smallest point size, there should be no more than 3.2 lines/cm (8 lines/inch) vertically. This is a minimum spacing; 2.75 lines/cm (7 lines/inch) will make the paper much more readable. Larger type sizes require correspondingly larger vertical spacing. Please do not double-space your paper. True-Type 1 fonts are preferred.

(a) Result 1

(b) Results 2 (c) Result 3

Fig. 3. Example of placing a figure with experimental results.

The first paragraph in each section should not be indented, but all following paragraphs within the section should be indented as these paragraphs demonstrate.

6. MAJOR HEADINGS

Major headings, for example, "1. Introduction", should appear in all capital letters, bold face if possible, centered in the column, with one blank line before, and one blank line after. Use a period (".") after the heading number, not a colon.

6.1. Subheadings

Subheadings should appear in lower case (initial word capitalized) in boldface. They should start at the left margin on a separate line.

6.1.1. Sub-subheadings

Sub-subheadings, as in this paragraph, are discouraged. However, if you must use them, they should appear in lower case (initial word capitalized) and start at the left margin on a separate line, with paragraph text beginning on the following line. They should be in italics.

7. PAGE NUMBERING

Please do **not** paginate your paper. Page numbers, session numbers, and conference identification will be inserted when the paper is included in the proceedings.

8. ILLUSTRATIONS, GRAPHS, AND PHOTOGRAPHS

Illustrations must appear within the designated margins. They may span the two columns. If possible, position illustrations at the top of columns, rather than in the middle or at the bottom. Caption and number every illustration. All halftone illustrations must be clear black and white prints. Do not use any colors in illustrations.

Table 1. Table caption

Column One	Column Two	Column Three
Cell 1	Cell 2	Cell 3
Cell 4	Cell 5	Cell 6

Since there are many ways, often incompatible, of including images (e.g., with experimental results) in a LaTeXdocument. Figure 3 shows you an example of how to do this.

9. TABLES AND EQUATIONS

Tables and important equations must be centered in the column. Table 1 shows an example of a table while the equation

$$\begin{aligned} y &= ax^2 + bx + c \\ &= (x+p)(x+q) \end{aligned} \tag{7}$$

shows an example of an equation layout.

Large tables or long equations may span across both columns. Any table or equation that takes up more than one column width must be positioned either at the top or at the bottom of the page.

10. FOOTNOTES

Use footnotes sparingly (or not at all!) and place them at the bottom of the column on the page on which they are referenced. Use Times 9-point type, single-spaced. To help your readers, avoid using footnotes altogether and include necessary peripheral observations in the text (within parentheses, if you prefer, as in this sentence).

11. CITATIONS AND REFERENCES

List and number all bibliographical references at the end of the paper. The references can be numbered in alphabetic order or in order of appearance in the document. When referring to them in the text, type the corresponding reference number in square brackets as shown at the end of this sentence [3]. All citations must be adhered to IEEE format and style. Examples such as [3], [4] and [5] are given in Section 12.

12. REFERENCES

[1] Authors, "The frobnicatable foo filter," ACM MM 2013 submission ID 324. Supplied as additional material `acmmm13.pdf`.

[2] Authors, "Frobnication tutorial," 2012, Supplied as additional material `tr.pdf`.

[3] Dennis R. Morgan, "Dos and don'ts of technical writing," *IEEE Potentials*, vol. 24, no. 3, pp. 22–25, Aug. 2005.

[4] J. W. Cooley and J. W. Tukey, "An algorithm for the machine computation of complex Fourier series," *Math. Comp.*, vol. 19, pp. 297–301, Apr. 1965.
[5] S. Haykin, "Adaptive filter theory," Information and System Science. Prentice Hall, 4th edition, 2002.