The Visual Computer (2019) 35:997–1011
https://doi.org/10.1007/s00371-019-01669-8

ORIGINAL ARTICLE

Deep binocular tone mapping

Zhuming Zhang[1,2] · Chu Han[1,2] · Shengfeng He[3] · Xueting Liu[4] · Haichao Zhu[5] · Xinghong Hu[1,2] · Tien-Tsin Wong[1,2]

Published online: 14 May 2019

Abstract
Binocular tone mapping is studied in the previous works to generate a fusible pair of LDR images in order to convey more visual content than one single LDR image. However, the existing methods are all based on monocular tone mapping operators. It greatly restricts the preservation of local details and global contrast in a binocular LDR pair. In this paper, we proposed the first binocular tone mapping operator to more effectively distribute visual content to an LDR pair, leveraging the great representability and interpretability of deep convolutional neural network. Based on the existing binocular perception models, novel loss functions are also proposed to optimize the output pairs in terms of local details, global contrast, content distribution, and binocular fusibility. Our method is validated with a qualitative and quantitative evaluation, as well as a user study. Statistics show that our method outperforms the state-of-the-art binocular tone mapping frameworks in terms of both visual quality and time performance.

Keywords Tone mapping · Binocular tone mapping · Binocular perception · Convolutional neural network

1 Introduction

High-dynamic-range (HDR) images can be acquired with daily used digital cameras or smartphones. The color depth precision can be as high as 14 or even 16 bits. By merging multiple images at different exposures, even higher color depth precision can be achieved [6]. HDR images are often transformed to low dynamic range (LDR) to satisfy the demands of monitors and projectors. However, because of the limited color depth precision of LDR images, *local details* and *global contrast* cannot be well preserved simultaneously [42]. Local details are the high-frequency components of the local texture, while global contrast is the ratio between the brightness of the brightest and the darkest areas in the image. To well preserve local details, global contrast becomes low making it difficult to understand the brightness difference in the real scene (Fig. 1a). In contrast, to well preserve global contrast, bright areas are saturated at the cost of losing the local details (Fig. 1b). To improve the perception of local details and global contrast, besides developing monitor with higher color depth precision, another effective way is to utilize stereoscopic devices which are widely used for movies, computer games, and augmented reality (AR).

Stereoscopic devices, which consist of two display channels, can help to increase the perceived color depth precision. This is because human visual system can preserve different visual contents separately perceived by two eyes and form a single vision. To utilize the binocular perception, *binocular tone mapping* is first proposed by Yang et al. [51] to generate two different LDR images. Given an HDR image and any tone-mapped LDR image, this method generates another LDR image that has the largest visual difference with the given one, while preserving the *binocular fusibility* for the LDR image pair. However, large visual difference between the image pair is not equivalent to more visual content. To resolve this issue, recently, Zhang et al. [54] proposed binocular perception metrics to measure the total visual content of a binocular LDR pair. Moreover, the two LDR images of

✉ Tien-Tsin Wong
ttwong@cse.cuhk.edu.hk

1 The Chinese University of Hong Kong, Shatin, NT, Hong Kong SAR, China

2 Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

3 South China University of Technology, Guangzhou, China

4 Caritas Institute of Higher Education, Tseung Kwan O, NT, Hong Kong, China

5 Rokid Corporation Ltd, Hangzhou, China

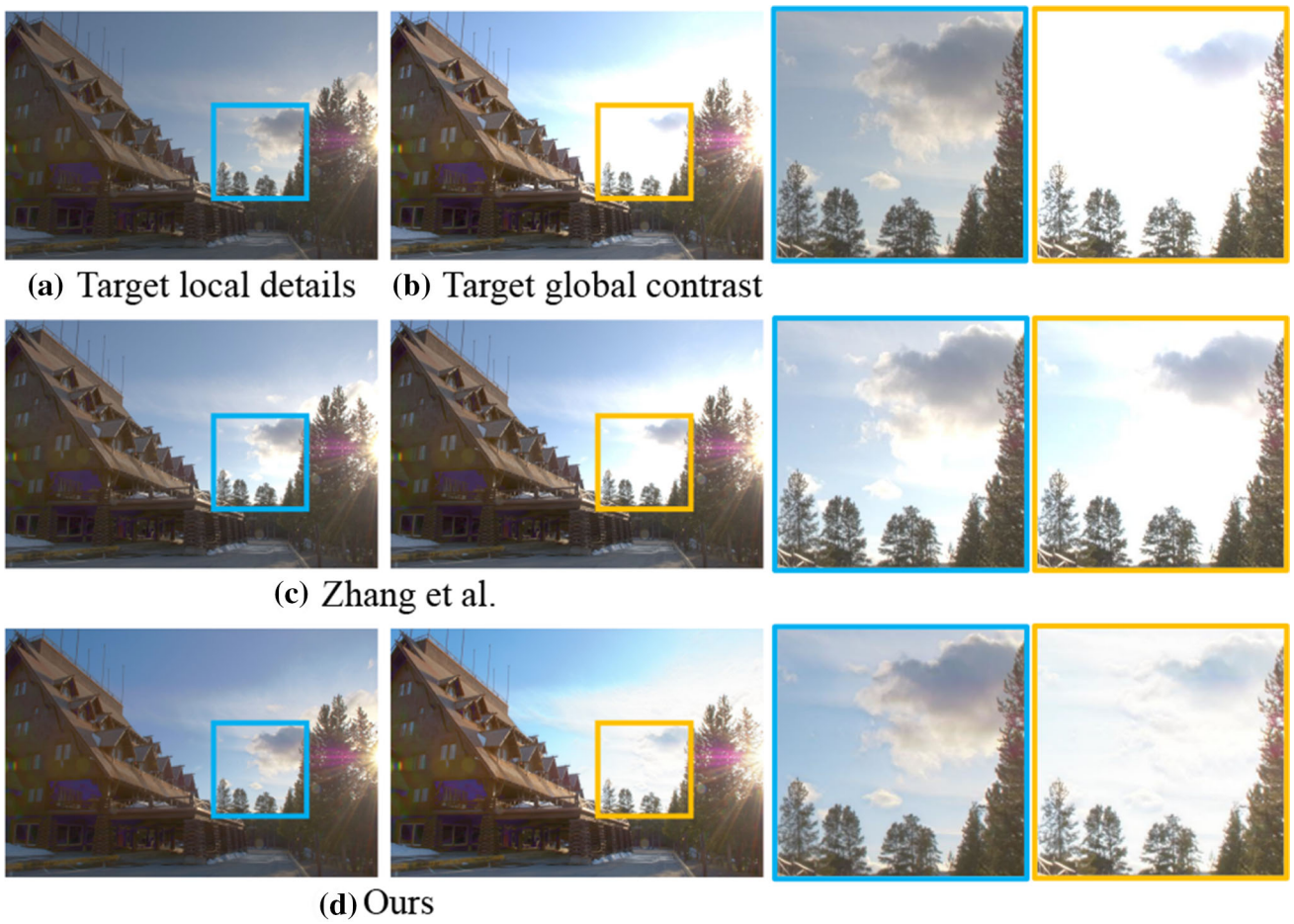


Fig. 1 **a** Tone-mapped LDR preserving local details but with low global contrast. **b** Tone-mapped LDR preserving global contrast but losing local details in the bright areas. **a** and **b** cannot form a feasible binocular pair since they are not binocularly fusible. **c** Binocular pair generated with Zhang at al's method [54]. It is fusible and with high global contrast, but cannot preserve all the details in the bright areas. **d** Our generated binocular pair well preserving both local details and global contrast. The sparse details in the bright areas of the right image not only helps binocular fusibility but also improve visual quality

a pair can be simultaneously optimized and generated with their method.

However, both Yang et al.'s and Zhang et al.'s methods generate LDR images relying on monocular tone mapping operators, and it leads to two major problems: First, it is not effective to optimize a binocular pair by tuning the parameters of existing monocular operators. It is because monocular operators are originally designed for the trade-off between local details and global contrast in one single image without considering how to distribute the visual content into a binocular pair. Moreover, the capacity of the two channels cannot be fully made use of, since the solution space is not only restricted by the model of the adopted monocular operator but also constrained by binocular fusibility. A stably fusible pair cannot be obtained with a simple combination of a image with good local detail preservation (Fig. 1a) and one with good global contrast preservation (Fig. 1b). Because local details in the bright areas are very different from each other in two views. On the other hand, the pair generated by Zhang et al.'s method is binocularly fusible because local details are not too different in two views (Fig. 1c). But the local details in the bright areas are not well preserved. Second, both of the existing methods are computationally expensive because of the iterative optimization. They optimize the output pair by iteratively updating the parameters of the adopted monocular operators. Tone mapping, visual content evaluation, and fusibility evaluation should be conducted in each iteration.

To resolve the above issue, we propose a binocular tone mapping operator that directly optimizes the visual content distribution to an image pair, instead of optimizing two sets of parameters to generate two monocular tone-mapped images. However, it is still difficult to handcraft an efficient and effective visual content distribution model. So we propose to utilize the representability and interpretability of deep convolutional neural network (CNN), to learn the distribution model from a large number of images of different contents.

A series of loss functions are proposed to optimize the visual contents of the generated pairs without introducing visual discomfort. Our proposed method has two advantages over the two existing methods: First, our generated LDR pair is of better visual quality since local details and global contrast are more effectively distributed to the pair. Compared with Zhang et al.'s result (Fig. 1c), our generated pair (Fig. 1d) better preserves local details while achieving similar global contrast. As shown in Fig. 1d, in the extremely bright areas of the right image, there are sparse details. Although this kind of local details seldom appears in monocularly tone-mapped images, they not only help binocular fusibility but also improve visual quality of binocular perception. Second, our proposed CNN-based tone mapping operator is end-to-end and with fast computational speed.

However, training the CNN-based binocular tone mapping operator is challenging. The network that simply regresses on existing results (Fig. 1c, with limited visual content) or the infusible pairs (combination of Fig. 1a, b, with more visual content) cannot generate desirable pairs with good visual quality while maintaining binocular fusibility. So we propose to optimize the binocularly perceived visual content of the output LDR pairs. But how to directly compare the visual content of an LDR pair with that of an HDR image is still an open problem. Instead, our use two reference images (Fig. 1a, b) to respectively represent the target local detail (Fig. 1a) and the target global contrast (Fig. 1b), and learn visual content from them utilizing binocular perception models. Besides binocular visual content evaluation, two more loss functions are also proposed to ensure the effective distribution of visual content and the binocular fusibility of the output pair. To demonstrate the performance of our method, a qualitative and quantitative evaluation, as well as a user study, were conducted on HDR images of different genres and content. Statistics show that our method outperforms the existing binocular tone mapping frameworks in terms of both visual quality and time performance. Our contributions are summarized as follows:

- We propose the first binocular tone mapping operator to effectively distribute visual content to an LDR image pair, leveraging the great representability and interpretability of deep convolutional neural network.
- Novel loss functions are proposed to optimize image pairs in terms of local details, global contrast, visual content distribution, and binocular fusibility.
- Our method can deliver more visual content than the state-of-the-art binocular tone mapping methods while maintaining visual comfort.
- Compared to the existing methods, our method can achieve faster computational speed and is easier to adopt GPU acceleration.

2 Related works

2.1 Monocular tone mapping operators

Many tone mapping operators (TMO) have been proposed to transform the HDR images to LDR images for better compatibility to LDR display devices. According to the comprehensive surveys done by Reinhard et al. [37] and Banterle et al. [2], existing TMOs can be categorized into global and local operators.

Global operators transform HDR to LDR in a spatially invariant manner by applying the same compressive curve over the whole image [7,38,44,48]. In particular, Tumblin et al. [44] proposed to generate LDR images whose displayed brightness is as close to real-world sensation as possible. Ward [48] developed global mapping functions based on the results in psychophysics on brightness and contrast perception. Inspired by photographic practices, Reinhard et al. [38] proposed a photographic tone reproduction technique. To imitate the human response to light, Drago et al. [7] proposed to logarithmically compress the luminance values.

Different from global operators, local operators transform HDR to LDR in a spatially variant manner [1,8,13,34]. These methods usually decompose the HDR image into layers, adjust the layers independently, and re-combine them into an LDR image. In particular, Durand et al. [8] leverages the bilateral filter to decompose an image into a base layer and a detail layer and adjusts the global contrast by manipulating the base layer. Farbman et al. [13] proposed to decompose the HDR using a weighted least squares (WLS) framework and manipulate both global contrast and local details in a multi-scale manner. Paris et al. [34] and Aubry et al. [1] effectively and efficiently performs layer decomposition and tone manipulation leveraging Local Laplacian Filter.

The above monocular TMOs focus on adjusting the trade-off between global contrast and local details in one single image. In other words, monocular operators are not capable of effectively distributing visual content to a binocular pair. So existing binocular tone mapping methods that rely on monocular operators cannot fully utilize the potential of binocular perception. In contrast, our proposed binocular tone mapping operator learns visual content distribution on a large dataset, and can effectively distribute local details and global contrast to an LDR pair.

2.2 Binocular perception

Von Helmholtz and Southall [46] presented an interesting fact that our visual system is able to fuse two different images from our two eyes into a single vision, which is called binocular single vision. Many existing works study the binocular perception of brightness, contour, color and local contrast (detail). For binocular tone mapping, local details and global

contrast are the most important kinds of visual content [54]. Local details perception can be estimated with local contrast perception models. Local contrast perception in one monocular view generally inhibits the other [47]. In other words, binocular local detail perception is actually dominated by the view with more local details. Existing works [26,29,30,32] provided different functions and parameters to model the domination phenomenon. On the other hand, global contrast perception can be regarded as the overall understanding of brightness and estimated utilizing brightness perception models. As suggested in many psychological studies [3,5,11,25,27], binocular brightness perception is approximately a linear combination of the brightness in two views, as long as the two views are not too different with each other. The above two perception mechanisms make it possible to convey more visual content within an image pair, as long as local details and global contrast is effectively distributed among the image pair.

To utilize binocular single vision, there exist some works that generate image pair or video pair. In particular, Yang et al. [51] first proposed a binocular tone mapping framework. To generate a pair of LDR images, it first generates and fixes one monocular LDR image and then optimizes the other such that these two images keep large visual difference [4]. But the major drawback is that large visual difference always brings more visual content. To resolve this issue, Zhang et al. [54] proposed binocular perception metrics to measure the total visual content of an image pair. Base on their proposed metrics, the two images in a pair can be simultaneously optimized in terms of local details, global contrast, and binocular fusibility. Feng and Loew [14] extended Yang et al.'s [51] work to video tone mapping, taking temporal coherency into account. Besides, binocular luster effect (the salient shininess in fused vision) was explored by Chua et al. [18] for enhancing visual information.

However, although based on binocular perception models, the existing binocular tone mapping methods [14,51,54] all use monocular tone mapping operators to generate binocular image pairs. Since monocular operators are not designed for distributing visual content to an image pair, the potential of binocular perception cannot be fully utilized. On the contrary, our proposed CNN-based operator aims at effectively distributing visual content to an image pair based on the existing binocular perception models. To meet this target, we propose to learn the distribution model from images of different contents, utilizing the representability and interpretability of CNN.

2.3 Convolutional neural network

Convolutional neural networks (CNNs) have been shown its superior performance in computer vision and graphics. Many kinds of image reconstruction and generation tasks can be effectively conducted with CNN by learning the hierarchical image features. Iizuka et al. [19] proposed an end-to-end network to predict color information from grayscale images. Xie et al. [49] proposed a contour prediction model that leverages fully convolutional neural networks [28] and deeply-supervised nets. Gharbi et al. [15] approximated the desired image transformation by training a network to predict the coefficients of a locally-affine model in bilateral space. Ledig et al. [24] presented a generative adversarial network [16] for image super-resolution. More works such as image inpainting [36,50], HDR image reconstruction [9], image style transfer [21] and texture synthesis [41] also proved the effectiveness of CNN in image reconstruction and generation applications. Endo et al. [10] proposed a deep-learning-based approach to infer a set of LDR images at different exposures from one single input LDR image. Although multiple output images were simultaneously generated, they were independently optimized.

However, the exiting CNN methods only focused on generating one single image or multiple independent monocular images. Different from them, our target is to generate an LDR pair for binocular single vision. So the existing methods cannot be directly applied to our application. Our target LDR pair should preserve as many local details and global contrast as possible while under the constraint of binocular fusibility. So we propose to optimize the output pair based on the binocular perception models, such that the visual content can be automatically and effectively distributed to two images.

3 Overview

As overviewed in Fig. 2, our proposed method takes an HDR image I_{HDR} (Fig. 2a) as input, and then generates a tone-mapped LDR pair $\{L, R\}$ (Fig. 2c). The output pair $\{L, R\}$ preserves the visual content of I_{HDR} since local details and global contrast are effectively distributed to the pair. L contains more local details but with less global contrast, while R has more global contrast but less local details. When $\{L, R\}$ is viewed with a stereoscopic display, audiences can perceive the visual content of both images with binocular perception.

Our method consists of an end-to-end network (Fig. 2b) trained with a series of loss functions (Fig. 2f) optimizing $\{L, R\}$. $\{L, R\}$ are not evaluated by referring to I_{HDR}, since how to directly compare the visual content of LDR images with that of HDR images is still an open problem. Instead, we generate two reference LDR images I_d and I_c (Fig. 2d, e), respectively, representing the target local details and global contrast:

$$I_d = \mathbf{T}(I_{\mathrm{HDR}},\ \beta_d), \tag{1}$$

$$I_c = \mathbf{T}(I_{\mathrm{HDR}},\ \beta_c), \tag{2}$$

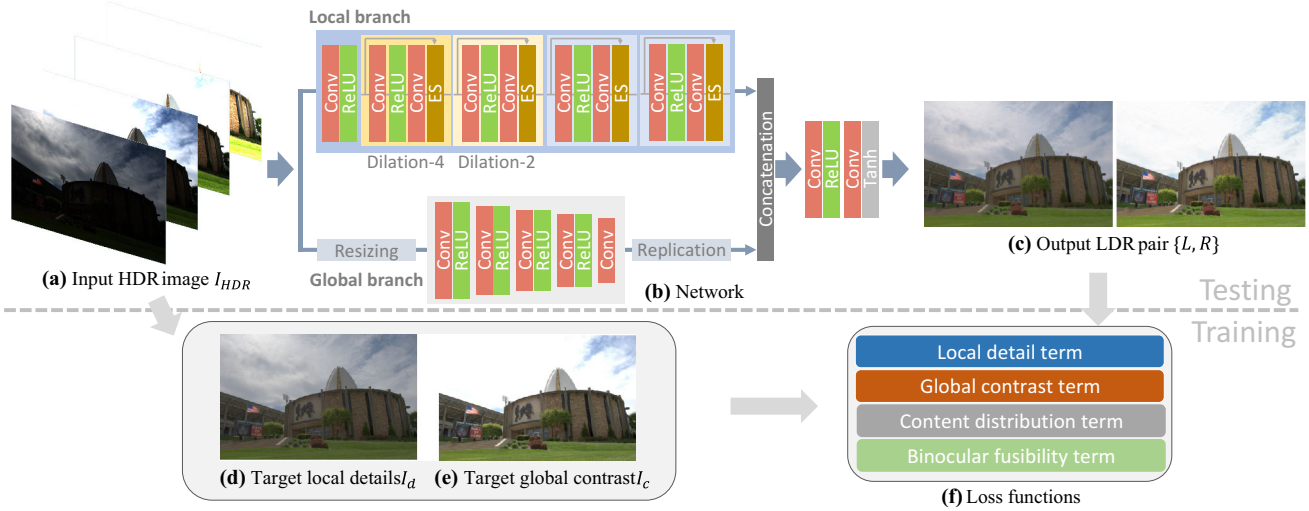


Fig. 2 System overview and network structures. Our proposed method takes I_{HDR} as input and generates an LDR pair $\{L, R\}$ with an end-to-end network. In the training phase, $\{L, R\}$ is optimized with the proposed loss functions. Two tone-mapped LDR images I_d and I_c are regarded as target local details and target global contrast. In the network structures, ES in the residual blocks denotes element-wise addition. Convolution filters in the Dilation-4 and dilation-2 blocks operate with dilation of 4 and 2. The size of all convolution kernels is 3×3 except for the 4×4 kernel in the last convolutional layer of the global branch. All the convolution layers generate feature maps with 64 channels

where $\mathbf{T}(I_{\text{HDR}}, \beta)$ is the bilateral tone mapping operator [8] with the parameter β controlling the contrast of the base layer of the tone-mapped image. A high value of β means global contrast. In our setting, I_d is generated with $\beta_d = 4$ preserving most of the local details. On the other hand, I_c is generated with $\beta_c = 5.5$ resulting in high global contrast at the cost of losing the local details in the bright areas. Two loss functions based on binocular perception models are proposed to evaluate the local details and global contrast of $\{L, R\}$ by referring to I_d and I_c. Also, we further design two more loss terms to improve visual content distribution and binocularly fusibility of $\{L, R\}$.

How to normalize HDR images to deal with different value ranges is introduced in Sect. 4.1. Our proposed network consists of a local branch and a global branch to process the local details and global contrast of the output $\{L, R\}$. The network structures are illustrated in Fig. 2b and detailed in Sect. 4.2. The design of the loss functions and the training details are elaborated in Sects. 4.3 and 4.4.

4 Approach

4.1 HDR normalization

Considering the various value range of HDR images, we first transform the values to the normalized logarithmic domain before tone mapping. Although the most straightforward way is to apply logarithmic transform independently on three color channels, unfortunately, it will cause color shift. Instead, we conduct the transformation only on luminance channel while maintaining the original color ratios. Inspired by the existing works [8,40,45], the luminance value l and the color ratio (r^*, g^*, b^*) for each pixel i are separated as:

$$l = 0.299r + 0.587g + 0.114b, \tag{3}$$

$$(r^*, g^*, b^*) = \frac{(r, g, b)}{\max(r, g, b)}, \tag{4}$$

where r, g and b are the red, green, blue values of a pixel i. The luminance value l is transformed to normalized log-scale luminance l' as:

$$l' = \frac{\log(l) - l_{\min}}{l_{\max} - l_{\min}}, \tag{5}$$

where $l_{\max}$ and $l_{\min}$ the maximal and minimal log-scale luminance values over the whole image. Thus, the range of l' is [0,1]. Finally, the original color ratio (r^*, g^*, b^*) and the normalized luminance l' are re-combined as:

$$(r', g', b') = \frac{l'}{0.299r^* + 0.587g^* + 0.114b^*}(r^*, g^*, b^*) \tag{6}$$

where (r', g', b') are the three input color channels to the network.

4.2 Network structures

A major purpose of our proposed method is to distribute local details and global contrast to the generated LDR pair. So we adopt a two-branch network structure to capture the local and

global information of the input HDR. Also, networks with multiple branches are effective to extract multi-scale features [43]. As shown in Fig. 2b, our network consists of a *local branch*, and a *global branch*.

The local branch extracts and reconstructs the local details. It consists of one convolutional layer and four residual blocks. Similar structures have been adopted in various image processing tasks, such as image super-resolution [23,39]. To increase the receptive field without increasing the complexity of the network, dilated convolutional layers [52] are adopted in the first two residual blocks. The dilation sizes are set to 4 and 2 respectively. Intuitively speaking, larger dilation size means large receptive field. Comparing with traditional strided convolution, dilated convolution can preserve the original resolution for the extracted feature map, so the extracted mid-level features are more precise and robust.

The global branch extracts and reconstructs global features that helps optimizing global contrast. It consists of a fixed number of strided convolution layers and requires a fixed input resolution. So the input image is first resized to the resolution of 64×64 to feed to the global branch. Then an $1 \times 1 \times 64$ feature vector is obtained after 4 convolutional layers with striding size of 2 and a convolutional layer with a 4×4 kennel without padding. This feature vector is further replicated to the image size as the output feature map of the local branch.

The output feature maps of local and global branches are then concatenated and processed with two convolutional layers followed by one tanh layer. With this two-branch network design, our network is able to capture multi-scale features of the input HDR images.

4.3 Loss functions

To optimize visual content distribution to the LDR pair generated by the network, we propose a series of loss functions. A local details term $\mathscr{L}_d$ and a global contrast term $\mathscr{L}_c$ are proposed based on binocular perception models, taking the reference LDR images I_d and I_c as targets, respectively optimizing local details and global contrast of the pair $\{L, R\}$. But as shown in Fig. 3a, training only with $\mathscr{L}_d$ and $\mathscr{L}_c$ cannot guarantee effective distribution into the pair. Thus we further propose a content distribution term $\mathscr{L}_{cd}$ that provide a guidance for distributing different visual content to L and R (Fig. 3b). Furthermore, an extra binocular fusibility term $\mathscr{L}_{bf}$ is proposed to improve the visual comfort of $\{L, R\}$ (Fig. 3c).

Local Detail Term Inspired by the existing work [53], we use the local gradient to evaluate the local details of the image. The gradient amplitude $G(I)$ over an LDR image I is locally calculated with the Scharr gradient operator [20]:

$$G_x(I) = \frac{1}{16}\begin{bmatrix} 3 & 0 & -3 \\ 10 & 0 & -10 \\ 3 & 0 & -3 \end{bmatrix} \circledast I, \tag{7}$$

$$G_y(I) = \frac{1}{16}\begin{bmatrix} 3 & 10 & 3 \\ 0 & 0 & 0 \\ -3 & -10 & -3 \end{bmatrix} \circledast I, \tag{8}$$

$$G(I) = \sqrt{G_x^2(I) + G_y^2(I)}, \tag{9}$$

where $\circledast$ is convolution operation. $G(I)$ is calculated in a monocular manner. Then utilizing the local gradient amplitude $G(L)$ and $G(R)$, we further estimate the local detail perception of the pair $\{L, R\}$ in a binocular manner. Local details in one monocular view generally inhibit details in the other view [47]. In other words, binocular detail perception is dominated by the view with more local details. It is approximated with the existing detail perception model [26]:

$$G_b(L, R) = \frac{(G(L)^s + G(R)^t)^{s/t}}{z + G(L)^s + G(R)^t}, \tag{10}$$

where s, t, and z are parameters of the model. In practice, we set $s = 3$, $t = 3$ and $z = 4.76$, which is similar to the values suggested by [31]. Then the estimated detail perception is used to evaluate how the output $\{L, R\}$ preserves the local details of input HDR image I_{HDR}. But how to compare the visual content of LDR images with that of HDR image is still an open problem. Instead, we generate a reference tone-mapped LDR image I_d to represent the target local details. So the proposed local detail term evaluates $\{L, R\}$ referring to $\{I_d, I_d\}$ with the detail perception model $G_b(L, R)$ as:

$$\mathscr{L}_d(L, R) = ||G_b(L, R) - G_b(I_d, I_d)||_1, \tag{11}$$

where $||\cdot||_1$ represents l_1 norm. The detail evaluation is conducted in a binocular manner.

Global Contrast Term Global contrast is the overall perception of local brightness. Local brightness can be estimated by averaging local image intensities [51,54]. We estimate the local brightness by smoothing the image I:

$$\mu(I) = M_k \circledast I \tag{12}$$

where M_k is a Gaussian kernel of size k, and $\circledast$ is convolution operation. We set $k = 11$ in practice. With $\mu(L)$ and $\mu(R)$, we can then further estimate the local brightness perception of pair $\{L, R\}$ in a binocular manner. Many existing psychological studies [3,5,11,25,27] suggest that the binocular brightness of an image pair can be approximated as a linear combination, as long as the two views are not too different with each other. Since our target output pair is binocular fusible, it fulfills the above assumption. So the global contrast term based on binocular brightness perception is defined as:

(a) $\mathscr{L}_d + \lambda_1 \mathscr{L}_c$

(b) $\mathscr{L}_d + \lambda_1 \mathscr{L}_c + \lambda_2 \mathscr{L}_{cd}$

(c) $\mathscr{L}_d + \lambda_1 \mathscr{L}_c + \lambda_2 \mathscr{L}_{cd} + \lambda_3 \mathscr{L}_{bf}$

Fig. 3 LDR pairs generated by networks trained with different combinations of loss functions (the local details term $\mathscr{L}_d$, the global contrast term $\mathscr{L}_c$, and the content distribution term $\mathscr{L}_{cd}$, and the binocular fusibility term $\mathscr{L}_{bf}$). **a** Only with $\mathscr{L}_d$ and $\mathscr{L}_c$. Visual content is not well distributed to the two images. Moreover, the two images cannot form a binocular pair. **b** With $\mathscr{L}_d$, $\mathscr{L}_c$, and $\mathscr{L}_{cd}$, but without $\mathscr{L}_{bf}$. The two images well preserve local details and global contrast but are not binocular fusible because the details in the bright areas differ too much from each other. **c** With all the proposed loss terms. The generated pair not only well preserves local details and global contrast but also is binocular fusible

$$\mathscr{L}_c(L, R) = ||\frac{1}{2}\mu(L) + \frac{1}{2}\mu(R) - \mu(I_c)||_1, \tag{13}$$

where $||\cdot||_1$ represents l_1 norm. I_c is the reference image representing the target global contrast.

Content Distribution Term The local detail term and the global contrast can evaluate the visual content. But training only with these two terms results in a feasible binocular pair (Fig. 3a). So we propose a content distribution term to improve the distribution ability of the network, such that the two LDR images can differently but effectively preserve visual contents. The target of this term is to preserve more local details but less global contrast in L while more global contrast but less local details in R. But designing a loss term to guide the distribution is not a straightforward task. Fortunately, inspired by the bilateral tone mapping operator [8], we found that providing target global contrast is efficient to guide the distribution. Moreover, the reference images I_d and I_c can be regarded as reasonable target levels of global contrast for L and R to respectively preserve visual content of two different kinds. Similar to Eq. 13, we evaluate global contrast based on local brightness $\mu(\cdot)$. So the content distribution is defined as minimizing the local brightness between the $\{L, R\}$ and $\{I_d, I_c\}$:

$$\mathscr{L}_{cd}(L, R) = ||\mu(L) - \mu(I_d)||_1 + ||\mu(R) - \mu(I_c)||_1, \tag{14}$$

where $||\cdot||_1$ represents l_1 norm. The global contrast is separately guided for the L and R in a monocular manner.

Binocular Fusibility Term. In our application, visual discomfort usually happens in the extremely bright areas where the difference between L and R is large. As shown in Fig. 3b, R preserves the brightness of these areas at the cost of losing

local details. On the other hand, L preserve most of the local details. As studied in [51], visual conflict of contour happens if and only if the details in one view are larger than the obvious color difference (OCD) threshold, while that in the other view is smaller than the just noticeable color difference (JND) threshold. For a single pixel, the visual conflict condition can be expressed as $G(L) > \text{OCD}$ and $G(R) < \text{JND}$. Region-based conflict condition is estimated by counting the pixels with visual conflict. To avoid visual conflict, we can conservatively encourage $G(R)$ to be larger than $\frac{\text{JND}}{\text{OCD}}\ G(L)$ in every pixel. Thus, our fusibility term is defined as:

$$\mathscr{L}_{bf}(L, R) = \max(\alpha_f G(L) - G(R), 0) \tag{15}$$

where α_f is parameter for fusibility enhancement and should be larger than $\frac{\text{JND}}{\text{OCD}}$. α_f is set to 0.6 in practice.

Overall Loss Function. The overall loss function is defined as the weighted sum of the local detail term, the global contrast term, the content distribution term and the binocular fusibility term:

$$\mathscr{L}(L, R) = \mathscr{L}_d + \lambda_1 \mathscr{L}_c + \lambda_2 \mathscr{L}_{cd} + \lambda_3 \mathscr{L}_{bf} \tag{16}$$

where $\lambda_1 = 0.2$, $\lambda_2 = 0.1$ and $\lambda_3 = 0.5$. Figure 3 shows results generated by networks trained with different combinations of loss terms. The network trained with the complete overall loss function can generate binocularly fusible image pair with good visual quality (Fig. 3c).

4.4 Training details

Thanks to the HDR normalization and our network design, our method is robust to various value ranges and different resolutions of the input HDR images. To make our method able to effectively distribute visual content to the output LDR pair, the network should be trained on a large set of HDR images of different genres and content. The HDR+ Burst Photography Dataset [17], which contains 3620 HDR images, is used as our training dataset.

During the training, all images in the training dataset are resized to 320×320, and then randomly cropped to 256×256. The network is trained on a PC equipped with two Nvidia GTX 1080Ti GPU, using the Adam optimizer [22] with β_1=0.999 and β_2=0.999. The batch size is 32, and the learning rate is set to $1e^{-5}$. The whole training process takes about 32 h for 800 epochs.

5 Results

To validate our method, we conducted a qualitative and a quantitative evaluation, as well as a user study, on HDR images of different genres and content. The Fairchild dataset [12] and the HDR-Eye dataset [33] were used for the evaluations. They respectively contain 105 and 46 HDR images and are independent of the training dataset (the HDR+ Burst Photography Dataset [17] introduced in Sect. 4.4). The images were resized to 960×640.

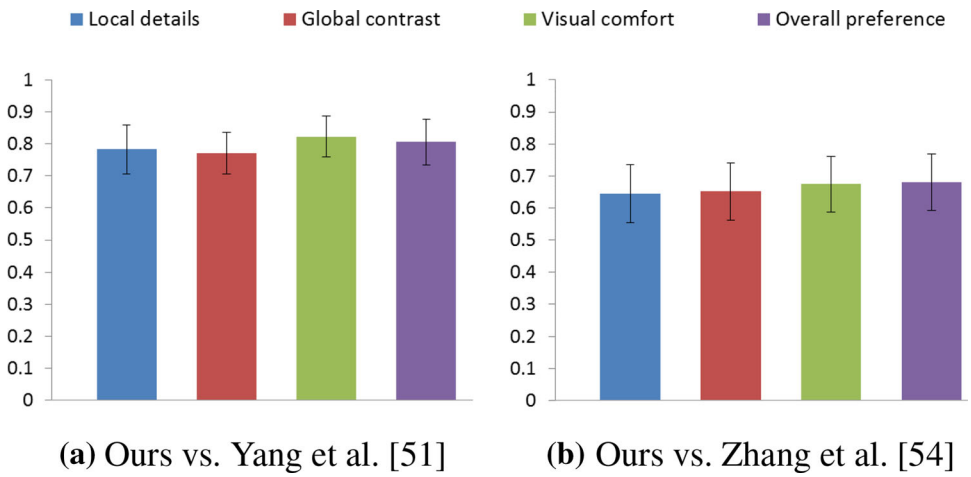
We compared our method with two state-of-the-art existing methods [51,54] which utilized monocular bilateral tone mapping operator [8] and optimized the parameters of the base layer contrast β_L and β_R. β_L and β_R can directly control the global contrast of left and right views. Yang et al.'s method [51] first generates an image pair with a fixed β_L. Then β_R is optimized such that the generated LDR image pair was with the maximal visual difference between them while satisfying fusibility. In our experiment, β_L was set to be 5, the suggested default value by [8]. On the other hand, Zhang et al.'s method [54] simultaneously optimizes β_L and β_R in terms of local details, global contrast, and binocular fusibility. In the following sections, results generated by Yang et al.'s, Zhang et al.'s and our methods are respectively denoted as $\{L^*_{\text{Yang}}, R^*_{\text{Yang}}\}$, $\{L^*_{\text{Zhang}}, R^*_{\text{Zhang}}\}$, and $\{L, R\}$.

Since all the outputs are binocular image pairs which should be perceived via stereoscopic devices, we have put the left and right image pairs of all the results shown in our paper in supplementary materials. Readers are strongly recommended to view the image pairs via a stereoscopic display for the best visual effect.

5.1 Qualitative evaluation

Figure 4 shows the comparison of our results against those generated by Yang et al. [51] and Zhang et al. methods [54]. Both of the existing methods were based on the monocular operator and generated two monocular LDR images. As shown in Fig. 4a, b, their results cannot simultaneously well preserve local details and global contrast. This is because the solution space is limited by the monocular operator and constrained by binocular fusibility, and visual content cannot be effectively distributed to two images. On the contrary, as shown in Fig. 4c, our results well preserve both of local details and global contrast. Compared with the results by the existing methods, ours contains much more details in the extremely bright areas while maintaining the high global contrast, because visual content is effectively distributed to the pair. Moreover, by keeping sparse details in R, our results are visually more comfortable than those generated the existing methods.

We further compare our results (Fig. 5b) with Yang et al.'s (Fig. 5a). Yang et al.'s method first generates a left view with fixed parameters and then optimizes the other view. It maximizes large visual difference between two views while fulfilling the fusibility constrain. However, a large visual difference, which Yang et al. encourage, is not equivalent to more visual content. When optimizing the



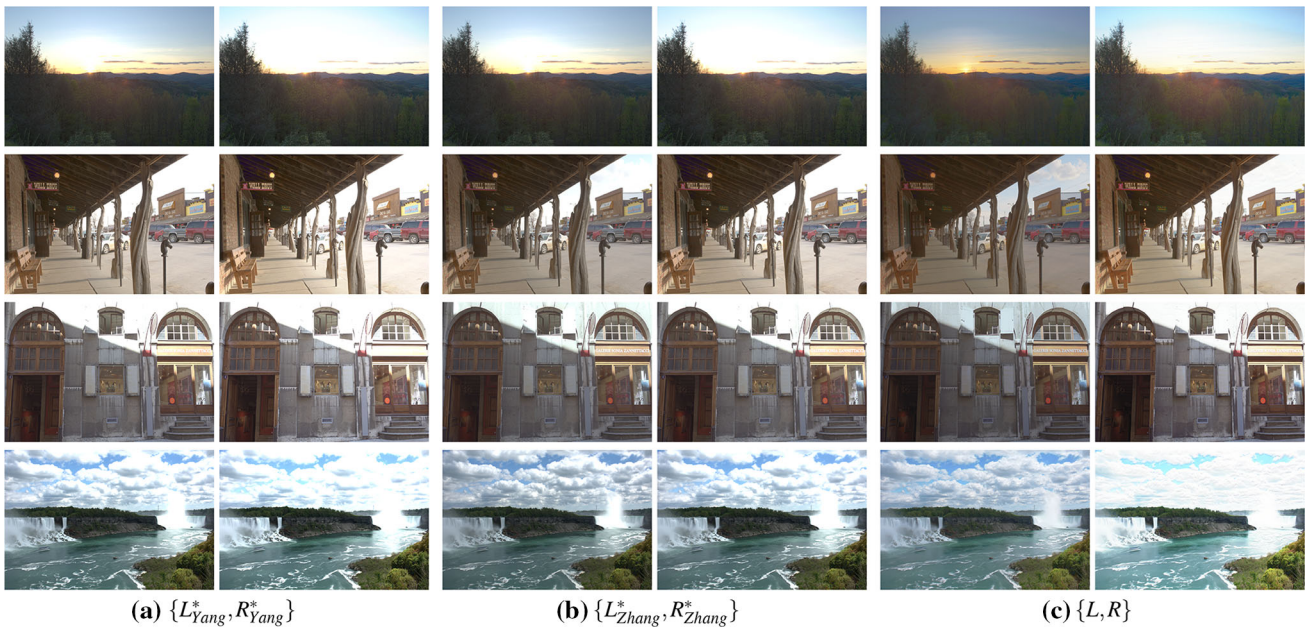


Fig. 4 **a** Yang et al.'s results; **b** Zhang et al.'s results; **c** our results. Our generated LDR pairs can better preserve both local details and global contrast

Fig. 5 **a** Yang et al.'s results; **b** our results. Our generated LDR pairs can better preserve local details in bright areas

unfixed view, large visual difference is usually achieved by over-exposure in the bright areas. So their results achieve large global contrast but cannot well preserve local details in the bright areas. On the contrary, our results can well preserve local details in the bright areas while achieving large global contrast for the whole image.

Also, we compare our results (Fig. 6b) with Zhang et al.'s (Fig. 6a). Both of their and our methods optimize the output pair in terms of global contrast, local details, and fusibility. Zhang et al's results achieve a trade-off between global contrast and local details. But they cannot simultaneously well preserve both kinds of visual content because the adopted monocular tone mapping operator cannot well distribute visual content to the pair. In the cases of the upper two rows in Fig. 6b, high global contrast is achieved, but local details in the bright areas are not well preserved. On the other hand, in the cases of the lower two rows in Fig. 6b, local details are well preserved but the global contrast is low. On the contrary, it can be easily observed that our results in Fig. 6a deliver more visual content by preserving more local details while maintaining high global contrast. Sparse details in R not only helps local detail preservation but also make the pair visually more comfortable.

5.2 Quantitative evaluation

We conducted a quantitative evaluation on the totally 151 HDR images from the Fairchild dataset [12] and the HDR-Eye dataset [33]. To fairly compare our results with Yang et al.'s [51] and Zhang et al.'s [54], we adopted binocular perception metrics (E_c, E_d, and E) proposed by Zhang et al. [54] and Visible Difference Predictor (VDP) [4] used in

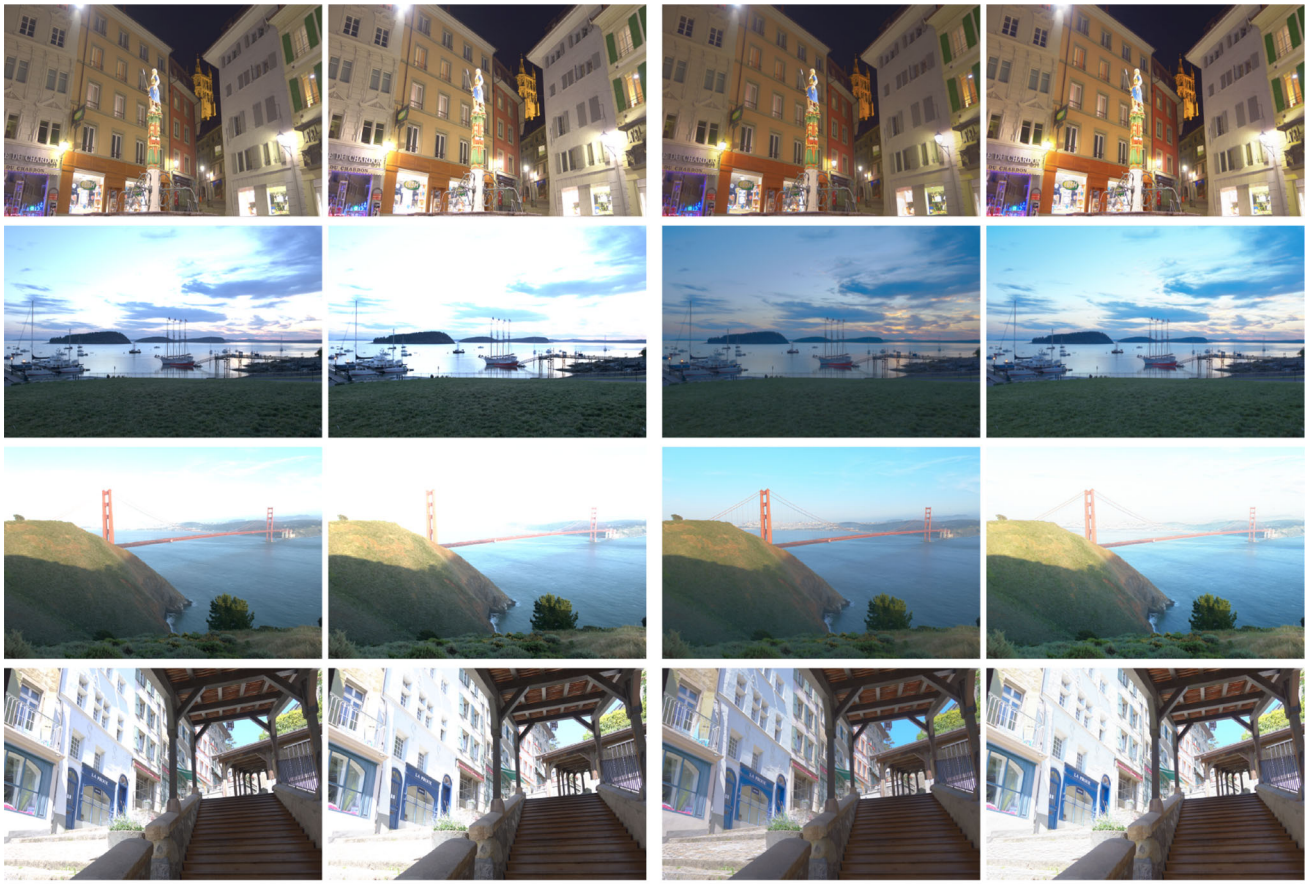


Fig. 6 **a** Zhang et al.'s results; **b** our results. Our generated LDR pairs can better preserve both local details and global contrast

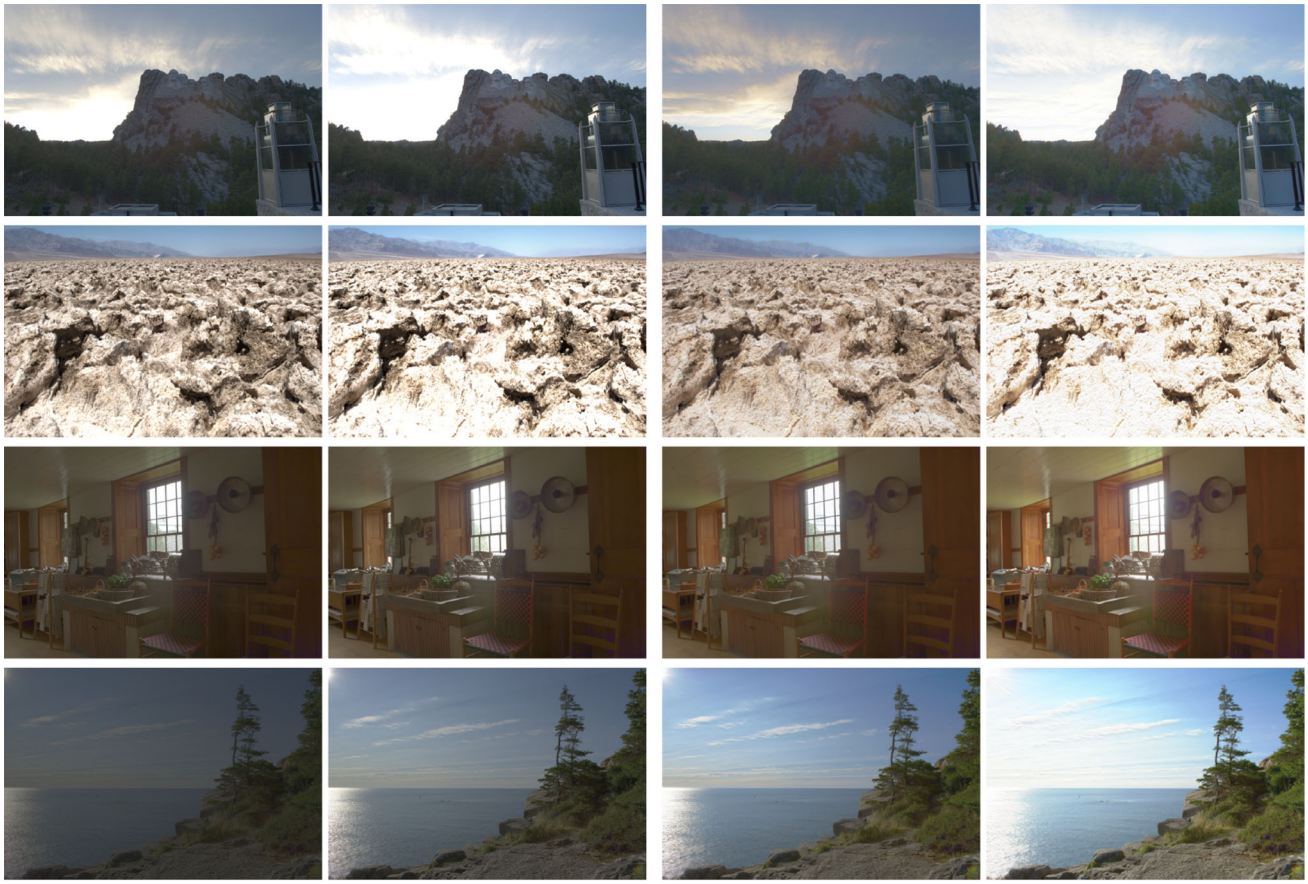
Yang et al. [51], under the same setting of their papers. E_c and E_d respectively measure the binocularly perceived global contrast and local details. A smaller E_c value indicates better global contrast. Similarly, a smaller value of E_d means more local details that have been preserved. E, which is the sum of E_c and E_d, intuitively represents the preservation of the total visual content. On the other hand, VDP indicates the visual difference between the left and right images. A larger VDP value means the image pair contains larger visual difference but does not always indicate better preservation of visual content.

As can be seen in Table 1, Yang et al. results achieves the highest VDP because their method is designed to maximize the VDP values of image pairs. High VDP can be usually achieved by over-exposure in one view while keeping some details in the other view. That is also the reason why they are with the best global contrast preservation E_c. But they sacrifice local details and lead to the worst local detail preservation E_d. Compared with Yang et al.'s results, Zhang et al.'s results achieve a better trade-off between global contrast and local details. So they have a better E than Yang et al.'s results. However, the same as Yang et al.'s method, Zhang et al.'s method generates images with monocular tone mapping operators, which cannot fully utilize the capacity of a binocular pair. Our local details preservation E_d is much better than Yang et al's and comparable to Zhang et al's, while our global contrast preservation E_c largely outperforms Zhang et al's. Moreover, the smallest E means our method is able to deliver much more visual content than the other two methods.

Table 1 Statistics of quantitative evaluation

Score	E_c	E_d	E	VDP
$\{L^*_{\text{Yang}}, R^*_{\text{Yang}}\}$	**0.1816**	0.4943	0.6759	**0.0245**
$\{L^*_{\text{Zhang}}, R^*_{\text{Zhang}}\}$	0.3233	**0.3435**	0.6568	0.0150
$\{L, R\}$	0.2254	0.3581	**0.5836**	0.0112

Mean values are tabled. The smaller the values of E_c and E_d are, the better the pair preserves global contrast and local details. $E = E_c + E_d$ measures the total visual content. Large VDP means large visual difference between the two images of a pair
Bold values indicate the best score of a particular metric among the different tested methods



(a) $\{L^*_{Zhang}, R^*_{Zhang}\}$ (b) $\{L, R\}$

Fig. 7 Statistics of user study for visual comparison. The scores represent ratios of the users who picked our results. Mean values and 95% confident intervals are illustrated

5.3 User study

We also conducted a user study to compare our method with the two existing methods [51,54] in terms of local details, global contrast, visual comfort, and overall preference. Totally 50 images were used in the user study. Among them, 25 images were used to compare our method with Yang et al.'s methods [51], while the other 25 images were used to compare our method with Zhang et al.'s methods [51] Totally 14 participants joined this user study, including 7 males and 7 females. All the setting were the same as [51,54]. We show the images on an ASUS G750JX laptop with a 3D display. Users were asked to sit half a meter away from the 3D display wearing the 3D glasses. The displaying luminance is set to 250 cd/m^2.

For each group of questions, participants were shown two different images side by side. One was generated by our method and the other was generated by one of the existing methods. The positions of the images were all random. Also, the left and right views were randomly swapped. Participants were asked to select the better image in terms of better preservation of local details, better preservation of global contrast, less visual discomfort, and the one they prefer. The score was marked as 1 if the user picked our result. Otherwise, it was marked as 0. Each image is regarded as a sample whose scores for the four problems are calculated by averaging the scores marked by different users. So the final scores for each image indicate the ratios of the users who picked our result. Figure 7 shows the statistics of the scores of the images. As can be seen, our method greatly outperforms Yang et al.'s in terms of local details, global contrast, visual comfort, and preference. Also, our method slightly outperforms Zhang et al.'s in all terms.

5.4 Timing statistics

We implemented our method using Pytorch [35], and tested it on a PC with an Intel i7-6700K @ 4.0GHz CPU, 32GB RAM, and an Nvidia GeForce GTX 1070 GPU. Running times of the network were recorded for input images of different resolutions with and without GPU acceleration. For each image resolution, we tested 100 images and recorded

Table 2 Timing statistics (in seconds)

Image size	CPU only	With GPU
256 × 256	1.389	0.006
512 × 512	4.582	0.012
1024 × 1024	30.313	0.046

the average time. The timing statistics are shown in Table 2. With the GPU acceleration, our method achieves real-time speedup for the resolution of 1024 × 1024.

As reported in [54], for the resolution of 800 × 600, Zhang et al.'s method [54] took 2.36 s for each iteration, while Yang et al.'s method [51] took 22.24 s. It was tested on a PC with Xeon E5-1620 v2 CPU @ 3.70 GHz and 32 GB RAM. No GPU is used. A single iteration involves tone mapping and energy function evaluation for one output image pair. Yang et al.'s method was much slower since the calculation of VDP [4] is quite time-consuming. For optimizing one single parameter for global contrast with bilateral tone mapping operator [8], Zhang et al.'s method could be solved with the gradient descent solver which generally converged in 30 iterations.

Thanks to the end-to-end network, our method generates pairs without conducting iterative optimization. Compared with the existing methods, our method achieves faster computational speed and is easier to adopt GPU acceleration.

5.5 Limitations

Our method achieves an effective distribution of local details to the LDR pair, improving detail perception and fusibility. However, since the limited dynamic range of LDR images, the trade-off between global contrast and local details still exists. In order to contain more details, our method may lead to a slight decrease in local brightness in the extremely bright areas, compared with the other existing binocular tone mapping methods.

By separately providing referencing global contrast for the two images, constraining the detail similarity between them, and optimizing the perceived visual content, our proposed method generates image pairs which are better than the results generated by the state-of-the-art methods. But what is the optimal distribution of visual content to the image pair is still an open problem. So our method cannot guarantee to generate image pairs with the optimal visual content.

6 Conclusion

In this paper, we proposed a CNN-based binocular tone mapping method. We leverage the strong representability and interpretability of CNN to automatically distribute the local details and global contrast to a binocular image pair. Our proposed method makes full use of the capacity of the two images and can deliver more visual content than the existing methods. Loss functions in terms of local details, global contrast, content distribution, and binocular fusibility were designed for network training.

In the future, visual saliency models can be adopted into our method to improve the visual content distribution ability. Also, binocular tone mapping from a stereoscopic HDR image pair is another promising and useful extension. Moreover, although it involves far more complicated perception models, how to generate binocular LDR video sequences from HDR video sequences is worth keeping discovering.

Acknowledgements This project is supported by the Research Grants Council of the Hong Kong Special Administrative Region, under RGC General Research Fund (Project No. CUHK 14201017), and Shenzhen Science and Technology Programs (No. JCYJ20160429190300857, No. JCYJ20180507182410327, and No. JCYJ20180507182415428).

Compliance with ethical standards

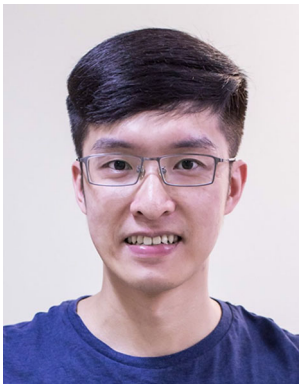
Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aubry, M., Paris, S., Hasinoff, S.W., Kautz, J., Durand, F.: Fast local laplacian filters: theory and applications. ACM Trans. Graph. (TOG) **33**(5), 167 (2014)
2. Banterle, F., Artusi, A., Debattista, K., Chalmers, A.: Advanced High Dynamic Range Imaging. Taylor & Francis, CRC Press (2017)
3. Curtis, D.W., Rule, S.J.: Binocular processing of brightness information: a vector-sum model. J. Exp. Psychol. Human Percept. Perform. **4**(1), 132 (1978)
4. Daly, S.J.: Visible differences predictor: an algorithm for the assessment of image fidelity. In: Human Vision, Visual Processing, and Digital Display III, vol. 1666, pp. 2–16. International Society for Optics and Photonics (1992). https://doi.org/10.1117/12.135952
5. De Weert, C.M., Levelt, W.J.M.: Binocular brightness combinations: additive and nonadditive aspects. Percept. Psychophys. **15**(3), 551–562 (1974)
6. Debevec, P.E., Malik, J.: Recovering high dynamic range radiance maps from photographs. In: ACM SIGGRAPH, pp. 369–378 (1997). https://doi.org/10.1145/258734.258884
7. Drago, F., Myszkowski, K., Annen, T., Chiba, N.: Adaptive logarithmic mapping for displaying high contrast scenes. In: Computer Graphics Forum, vol. 22, pp. 419–426. Wiley Online Library (2003). https://doi.org/10.1111/1467-8659.00689
8. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. In: ACM Transactions on Graphics (TOG), vol. 21, pp. 257–266. ACM (2002). https://doi.org/10.1145/566570.566574
9. Eilertsen, G., Kronander, J., Denes, G., Mantiuk, R.K., Unger, J.: Hdr image reconstruction from a single exposure using deep cnns. arXiv preprint arXiv:1710.07480 (2017)

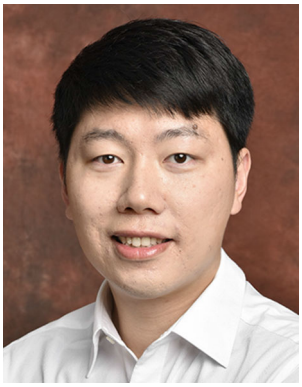
10. Endo, Y., Kanamori, Y., Mitani, J.: Deep reverse tone mapping. ACM Trans. Graph. (TOG) **36**(6), 177 (2017)
11. Engel, G.: The autocorrelation function and binocular brightness mixing. Vis. Res. **9**(9), 1111–1130 (1969)
12. Fairchild, M.D.: The hdr photographic survey. In: Color and Imaging Conference, vol. 2007, pp. 233–238. Society for Imaging Science and Technology (2007)
13. Farbman, Z., Fattal, R., Lischinski, D., Szeliski, R.: Edge-preserving decompositions for multi-scale tone and detail manipulation. In: ACM Transactions on Graphics (TOG), ACM, vol. 27, p. 67 (2008). https://doi.org/10.1145/1399504.1360666
14. Feng, M., Loew, M.H.: Video-level binocular tone-mapping framework based on temporal coherency algorithm. In: 2017 IEEE Applied Imagery Pattern Recognition Workshop (AIPR), IEEE, pp. 1–5 (2017)
15. Gharbi, M., Chen, J., Barron, J.T., Hasinoff, S.W., Durand, F.: Deep bilateral learning for real-time image enhancement. ACM Trans. Graph. (TOG) **36**(4), 118 (2017)
16. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in Neural Information Processing Systems, pp. 2672–2680 (2014). https://papers.nips.cc/paper/5423-generative-adversarial-nets
17. Hasinoff, S.W., Sharlet, D., Geiss, R., Adams, A., Barron, J.T., Kainz, F., Chen, J., Levoy, M.: Burst photography for high dynamic range and low-light imaging on mobile cameras. ACM Trans. Graph. (TOG) **35**(6), 192 (2016)
18. Hau Chua, S., Zhang, H., Hammad, M., Zhao, S., Goyal, S., Singh, K.: Colorbless: augmenting visual information for colorblind people with binocular luster effect. ACM Trans. Comput.-Hum. Interact. (TOCHI) **21**(6), 32 (2015)
19. Iizuka, S., Simo-Serra, E., Ishikawa, H.: Let there be color!: joint end-to-end learning of global and local image priors for automatic image colorization with simultaneous classification. ACM Trans. Graph. (TOG) **35**(4), 110 (2016)
20. Jähne, B., Haussecker, H., Geissler, P.: Handbook of computer vision and applications, vol. 2. Citeseer (1999)
21. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) European Conference on Computer Vision, pp. 694–711. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46475-6_43
22. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
23. Ledig, C., Theis, L., Huszar, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., Shi, W.: Photo-realistic single image super-resolution using a generative adversarial network. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
24. Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., Aitken, A.P., Tejani, A., Totz, J., Wang, Z., et al.: Photo-realistic single image super-resolution using a generative adversarial network. In: CVPR, vol. 2, p. 4 (2017)
25. Legge, G.E.: Binocular contrast summation-II. Quadratic Summ. Vis. Res. **24**(4), 385–394 (1984)
26. Legge, G.E., Rubin, G.S.: Binocular interactions in suprathreshold contrast perception. Percept. Psychophys. **30**(1), 49–61 (1981)
27. Levelt, W.J.: Binocular brightness averaging and contour information. Br. J. Psychol. **56**(1), 1–13 (1965)
28. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440 (2015)
29. Maehara, G., Goryo, K.: Binocular, monocular and dichoptic pattern masking. Opt. Rev. **12**(2), 76–82 (2005)
30. Meese, T.S., Georgeson, M.A., Baker, D.H.: Interocular masking and summation indicate two stages of divisive contrast gain control. In: Twenty-Eighth European Conference on Visual Perception (2005)
31. Meese, T.S., Georgeson, M.A., Baker, D.H.: Binocular contrast vision at and above threshold. J. Vis. **6**(11), 7–7 (2006)
32. Meese, T.S., Hess, R.F.: Low spatial frequencies are suppressively masked across spatial scale, orientation, field position, and eye of origin. J. Vis. **4**(10), 2–2 (2004)
33. Nemoto, H., Korshunov, P., Hanhart, P., Ebrahimi, T.: Visual attention in ldr and hdr images. In: 9th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM), EPFL-CONF-203873 (2015)
34. Paris, S., Hasinoff, S.W., Kautz, J.: Local laplacian filters: edge-aware image processing with a laplacian pyramid. ACM Trans. Graph. (TOG) **30**(4), 1–68 (2011)
35. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in pytorch. In: NIPS-W (2017)
36. Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., Efros, A.A.: Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2536–2544 (2016). https://doi.org/10.1109/CVPR.2016.278
37. Reinhard, E., Heidrich, W., Debevec, P., Pattanaik, S., Ward, G., Myszkowski, K.: High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting. Morgan Kaufmann, Burlington (2010)
38. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. ACM Trans. Graph. (TOG) **21**(3), 267–276 (2002)
39. Sajjadi, M.S.M., Schölkopf, B., Hirsch, M.: Enhancenet: Single image super-resolution through automated texture synthesis. In: IEEE International Conference on Computer Vision (ICCV) (2017)
40. Schlick, C.: Quantization techniques for visualization of high dynamic range pictures. In: Photorealistic Rendering Techniques, pp. 7–20. Springer (1995)
41. Sendik, O., Cohen-Or, D.: Deep correlations for texture synthesis. ACM Trans. Graph. (TOG) **36**(5), 161 (2017). https://doi.org/10.1145/3015461
42. Smith, K., Krawczyk, G., Myszkowski, K., Seidel, H.P.: Beyond tone mapping: enhanced depiction of tone mapped hdr images. In: Computer Graphics Forum, vol. 25, pp. 427–438. Wiley Online Library (2006)
43. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI, vol. 4, p. 12 (2017). https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14806
44. Tumblin, J., Rushmeier, H.: Tone reproduction for realistic images. IEEE Comput. Graph. Appl. **13**(6), 42–48 (1993). https://doi.org/10.1109/38.252554
45. Tumblin, J., Turk, G.: Lcis: A boundary hierarchy for detail-preserving contrast reduction. In: Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques, pp. 83–90. ACM Press/Addison-Wesley Publishing Co. (1999)
46. von Helmholtz, H., Southall, J.P.C.: Treatise on Physiological Optics, vol. 3. Courier Corporation, North Chelmsford (2005)
47. Wilson, H.R.: Binocular contrast, stereopsis, and rivalry: toward a dynamical synthesis. Vis. Res. **140**, 89–95 (2017)
48. Ward, G.: A contrast-based scalefactor for luminance display. Graph. Gems **IV**, 415–421 (1994)
49. Xie, S., Tu, Z.: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015)

50. Yang, C., Lu, X., Lin, Z., Shechtman, E., Wang, O., Li, H.: High-resolution image inpainting using multi-scale neural patch synthesis. arXiv preprint arXiv:1611.09969 (2016)
51. Yang, X., Zhang, L., Wong, T.T., Heng, P.A.: Binocular tone mapping. ACM Trans. Graph. (SIGGRAPH 2012 issue) **31**(4), 93:1–93:10 (2012)
52. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. arXiv preprint: arXiv:1511.07122 (2015)
53. Zhang, L., Zhang, L., Mou, X., Zhang, D., et al.: Fsim: a feature similarity index for image quality assessment. IEEE Trans. Image Process. **20**(8), 2378–2386 (2011)
54. Zhang, Z., Hu, X., Liu, X., Wong, T.T.: Binocular tone mapping with improved overall contrast and local details. Comput. Graph. Forum **37**(7), 433–442 (2018)

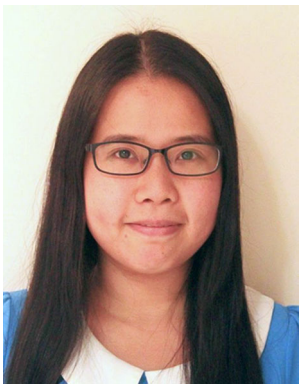


Zhuming Zhang is currently a Ph.D. student in the Department of Computer Science and Engineering at The Chinese University of Hong Kong. Before that, he received the B.E. degree from the Guangdong University of Technology, China, in 2010, and the M.E. degree from the South China University of Technology, China, in 2013. His research interests include image analysis, image/video recognition, and computer graphics.

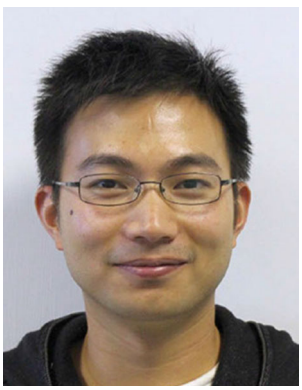
Chu Han graduated from the South China Agricultural University in 2011 with a B.Sc. degree in Computer Science. He received the M.Phil. degrees in computer science from the South China University of Technology in 2014, under the supervision of Prof. XU Xuemiao. Now he is pursuing Ph.D. in the Department of Computer Science and Engineering of the Chinese University of Hong Kong, under the supervision of Prof. WONG Tien-Tsin. His current research interests include computer graphics, image processing, computer vision and machine learning.



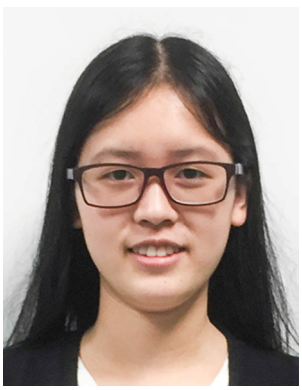
Shengfeng He is an Associate Professor in the School of Computer Science and Engineering, South China University of Technology. He was a Research Fellow at City University of Hong Kong. He obtained his B.Sc. degree and M.Sc. degree from Macau University of Science and Technology, and the Ph.D degree from City University of Hong Kong. His research interests include computer vision, image processing, computer graphics, and deep learning.



Xueting Liu received her B.Eng. degree from Tsinghua University and Ph.D. degree from The Chinese University of Hong Kong in 2009 and 2014 respectively. She is currently an Assistant Professor in the School of Computing and Information Sciences, Caritas Institute of Higher Education. Her research interests include computer graphics, computer vision, machine learning, computational manga and anime, and non-photorealistic rendering.



Haichao Zhu received his B.E. degree in Information Engineering from Beijing University of Posts and Telecommunications in 2010, and M.Sc. degree in Information Engineering and Ph.D. degree in Computer Science and Engineering both from The Chinese University of Hong Kong in 2011 and 2017 respectively. He is currently a senior computer vision research scientist at the R-Lab of Rokid. His research interests include computer vision, computer graphics, and augmented reality.



Xinghong Hu received her B.Eng. degree from Nanjing University in 2013. She is currently a Ph.D. student in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. Her research interests include computer graphics and binocular video applications.

Tien-Tsin Wong received his B.Sc., M.Phil. and Ph.D. degrees in Computer Science from The Chinese University of Hong Kong in 1992, 1994, and 1998 respectively. He is currently a professor in the Department of Computer Science and Engineering, The Chinese University of Hong Kong. His main research interests include computer graphics, computational manga, precomputed lighting, image-based rendering, GPU techniques, medical visualization, multimedia compression, and computer vision. He received the IEEE Transactions on Multimedia Prize Paper Award 2005 and the Young Researcher Award 2004.