Information Fusion 79 (2022) 248–262

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

Multi-exposure image fusion via deep perceptual enhancement

Dong Han [a], Liang Li [a], Xiaojie Guo [a,*], Jiayi Ma [b]

[a] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[b] Electronic Information School, Wuhan University, Wuhan 430072, China

ARTICLE INFO

Keywords:
Multi-exposure image fusion
Perceptual enhancement
Contrast enhancement
Color correction
Illumination adjustment

ABSTRACT

Due to the huge gap between the high dynamic range of natural scenes and the limited (low) range of consumer-grade cameras, a single-shot image can hardly record all the information of a scene. Multi-exposure image fusion (MEF) has been an effective way to solve this problem by integrating multiple shots with different exposures, which is in nature an enhancement problem. During fusion, two perceptual factors including the informativeness and the visual realism should be concerned simultaneously. To achieve the goal, this paper presents a deep perceptual enhancement network for MEF, termed as DPE-MEF. Specifically, the proposed DPE-MEF contains two modules, one of which responds to gather content details from inputs while the other takes care of color mapping/correction for final results. Both extensive experimental results and ablation studies are conducted to show the efficacy of our design, and demonstrate its superiority over other state-of-the-art alternatives both quantitatively and qualitatively. We also verify the flexibility of the proposed strategy on improving the exposure quality of single images. Moreover, our DPE-MEF can fuse 720p images in more than 60 pairs per second on an Nvidia 2080Ti GPU, making it attractive for practical use. Our code is available at https://github.com/dongdong4fei/DPE-MEF.

1. Introduction

When one takes pictures of natural scenes using digital cameras, it is difficult to acquire ideally exposed images, no matter how to adjust the exposure time and aperture. In other words, some areas of over-exposure and/or under-exposure often appear in a single image, hardly presenting all the contents of scene. Please see Fig. 1(a) and (b) for example. The reason is that the scene dynamic range spans much wider than the camera can record.

To mitigate the above issue, *multi-exposure image fusion* (MEF for short), as a cost-effective solution, has been drawing significant attention from the community, since barely MEF methods involve professional imaging equipment and additional knowledge about, for example, the camera response function, making them general and attractive for practical use. Formally, the task of MEF aims to reassemble a given sequence of *low dynamic range* (LDR) images under different exposures into a *high dynamic range* (HDR) one with **rich information** and **visual realism**, which is in nature an enhancement problem.

Towards this purpose, a variety of algorithms have been proposed, which can be roughly divided in to traditional methods and deep learning methods. Specifically for the traditional category, existing methods are either spatial-domain based [1,2] or transform-domain based [3]. Because all these methods employ handcraft features to accomplish the fusion process, the performance is limited and lack of robustness to dealing with complex and diverse scenes. Recently, with the emergence of deep learning, the improvement over traditional strategies have been witnessed by many computer vision tasks, like [4–7], due to the strong feature extraction ability of deep networks. As a consequence, a number of deep learning based attempts have been designed to solve the MEF problem. The key factor restricting the performance of deep learning methods is no ground-truth real data available for MEF. To ameliorate the situation, for instance, Prabhakar et al. designed a deep network called DeepFuse [8] by adopting the metric MEF-SSIM [9] in an unsupervised fashion. However, the MEF-SSIM term merely concerns the structure and contrast of source images to be preserved in the fused image, which is insufficient to produce satisfied results because of neglecting other aspects. Alternatively, Xu et al. [10] converted the unsupervised setting into a supervised one by choosing the best fusion results from existing methods to serve as the pseudo ground truths. Despite the improvement made by the conversion, the performance is inevitably restricted by the existing methods where those pseudo ground truths come from (please recall the enhancement nature).

We notice that, different from other image fusion scenarios, like infrared and visible fusion that concentrates more on the informativeness and acts as a pre-processing step to improve downstream tasks [11,12], MEF should take into account, besides the informativeness, the visual

* Corresponding author.
E-mail addresses: dongdong4fei@tju.edu.cn (D. Han), liangli@tju.edu.cn (L. Li), xj.max.guo@gmail.com (X. Guo), jyma2010@gmail.com (J. Ma).

https://doi.org/10.1016/j.inffus.2021.10.006
Received 24 June 2021; Received in revised form 31 August 2021; Accepted 18 October 2021
Available online 4 November 2021

realism of the fused results. Among factors affecting the aesthetics of an image, the color is arguably the most pivotal one. Most, if not all, of existing MEF approaches first switch the RGB color-space of images to a luminance and chrominance separation color-space, *e.g.* YCbCr, then apply fusion strategies only on the luminance (Y channel). However, in such a way, the color of fused images often turns out to be relatively pale and distorted, because the color information of (extremely) under-exposed images may be (seriously) ruined due to the limited quality of sensors.

Based on the above considerations, this study presents a deep neural network to accomplish the MEF task. To be concrete, the main contributions of this paper can be summarized as follows:

1. For obtaining informative and visually striking fused results, we design a deep perceptual enhancement net for MEF, namely DPE-MEF. The DPE-MEF is consisted of two functional subnets, which are responsible to collect important details from multiple inputs and guarantee the aesthetic, respectively.
2. Driven by the nature of fusion, the detail enhancement module attempts to fully explore details from source inputs. Enhanced images are efficiently formed through seeking best local exposures, which perform as references to guide the detail enhancement module.
3. To ensure the visual quality, the color enhancement module is introduced. It is able to refine the appearance by learning the relationship between color and brightness in natural images of the same scene, so as to fit more realistic and vivid color for fused images, significantly improving the visual perception.
4. Extensive experiments are conducted to demonstrate the efficacy of our design, and reveal its advantages in comparison with state-of-the-art methods. We further verify the proposed strategy can be used for boosting the exposure quality for single images.

2. Related work

Over past decades, MEF has been attracting much attention from the community due to its wide applicable range. Existing schemes can be roughly grouped into traditional and deep learning methods [15]. This section will briefly review the classic and contemporary approaches that are closely related to ours.

Traditional methods generally contain three major components, including image transformation, activity level (informativeness) measurement, and fusion strategy designing [16], which can be mainly split into spatial-domain based and transform-domain based techniques. The former ones directly perform fusion strategies on either pixel level or patch level. Methods work on pixel-level usually make effort to calculate proper weight maps for the source images, and then perform fusion by the weighted addition. As a representative, Liu et al. [17] proposed a method based on dense scale invariant feature transform, and apply dense SIFT descriptor as activity level measurement to compute weight maps. Patch-based works typically first evaluate information amounts of patches within source images from different aspects, then combine those with richest information to compose the fused image, with [18] as one of the earliest patch-based approaches. Ma et al. customized a structural patch decomposition method for MEF [1,19], and utilized this decomposition strategy as an optimization index to further propose an optimization-based method [20]. The above methods have the limitations, *i.e.*, pixel-based approaches often suffer from the brightness transition problem due to the lack of global information, while patch-based ones very likely introduce (halo) artifacts around boundaries. As for transform-domain based methods, they typically perform fusion strategies on the coefficients and then inverse transform back to the original domain. Burt et al. [21] designed a pyramid decomposition scheme for MEF, which is arguably the first transform-based attempt on the MEF task. Moreover, Mertens et al. [22] tried to combine the contrast, saturation and well-exposedness to measure the quality of source images, then generated weight maps to fuse the source images in a pyramid manner. Li et al. [23] employed a guided filter to decompose source images into their base and detail layers, and then merge them by weighted average for obtaining final fused images. Many follow-ups alternatively adopt other ways to do the job, such as wavelet [24], gradient [25], and PCA [26]. Though showing somewhat reasonable results, these traditional methods usually rely on hand-craft features and manually designed fusion strategies. Due to the inadequate abilities in feature extraction and integration, most of them require a long sequence of source images with small exposure intervals to generate relatively good fusion results, requiring heavy computational load and limiting the applicable scenarios. Notice that with less source images having large exposure variation, the quality of fused results by these approaches will be dramatically degraded.

Deep learning methods have become dominant in MEF recently, which considerably relieve the demand on the quantity and quality of source images, and achieve better fusion results [27]. The first DL-based attempt may trace back to DeepFuse [28], which builds a convolutional network to directly merge the luminance components of the source images by optimizing the unsupervised metric MEF-SSIM [9], and fuse the chrominance parts via a weighted fusion strategy. However, MEF-SSIM itself is insufficient to assure the fusion quality, and the absence of detailed chrominance treatment frequently results in color distortion. Following DeepFuse, Qi et al. [29] leveraged the multi-channel MEF-SSIM [20] as the optimization objective to avoid the conversion of color-space. Ma et al. [13] designed a network called MEF-Net, which follows the weighted fusion line. The MEF-Net generates weight maps through feeding down-sampled source images into the network, so as to reduce the computational cost. Because of its pixel-wise weight addition manner, it also has troubles as traditional pixel-based methods like overlooking global structure information. The above DL-based algorithms perform in different unsupervised ways. Alternatively, some works turn the MEF task into a supervised manner by taking fused images produced by existing fusion methods as pseudo ground truths. Xu et al. [10] employed generative adversarial networks for the MEF task, namely MEF-GAN. Zhang et al. proposed IFCNN [30], which uses two branches to extract features from each source image, then adopts element-wise fusion rules to fuse the deep features, and finally generates fused image from the fused features by two convolution layers. Obviously, the performance of these supervised approaches is restricted by the involved existing methods. It is not difficult to see artifacts in reference images and thus in final fused results. Besides, there are several unified deep learning based methods proposed to serve a variety of image fusion tasks. These methods usually use common image attributes as a measure of informativeness, then carry out fusion to gather the informative parts. For instance, DIF-Net [31] adopts structure tensors to evaluate the structure intensity of source images. In PMGI [32] and SDNet [33], the gradient and intensity are both used, while U2Fusion [14] utilizes the gradient of deep features to preserve the similarity between the fused result and source images. Although this kind of methods expands the scope of application, they inevitably lose specific considerations for different fusion scenarios. In addition, they also suffer from color distortion since the operation is performed only on the luminance channel to achieve multi-task versatility.

3. Problem analysis

We emphasize that the goal of MEF is to generate an informative and visually pleasant result from multiple images of a certain scene. The MEF task is in nature an enhancement problem, for which the ground truth is generally unavailable. In this situation, several schemes have been studied, which make efforts to train deep networks using non-reference metrics in unsupervised settings. The non-reference image quality metric MEF-SSIM is the most commonly used unsupervised MEF evaluation metric [9]. It is used as the loss term in many methods like [13,28,29]. Let us here take a closer look at MEF-SSIM. Inspired

(a) Under-exposed input (b) Over-exposed input (c) Corrected from (a) (ours) (d) Corrected from (b) (ours)

(e) Fused by MEF-Net [13] (f) Fused by U2Fusion [14] (g) Fused by MEF-GAN [10] (h) Fused by DPE-MEF (ours)

Fig. 1. An enhancement example. (a) and (b) depict a pair of under-exposed and over-exposed images of the same scene. (c) and (d) show the processed results by our method on single images (a) and (b), respectively. (e)–(h) give the fused results by MEF-Net, U2Fusion, MEF-GAN, and our DPE-MEF, respectively. By jointly considering the informativeness and the visual realism, our DPE-MEF shows its clear advantages, such as sharp details and vivid colors, over other competitors. Please zoom in the pictures to see more details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

by the structural similarity (SSIM) index, MEF-SSIM is a patch-based metric, which decomposes each image patch x_k into its signal strength c_k, signal structure s_k and mean intensity l_k in the following way:

$$\begin{aligned} x_k &= \left\| x_k - \mu_{x_k} \right\|_2 \cdot \frac{x_k - \mu_{x_k}}{\left\| x_k - \mu_{x_k} \right\|_2} + \mu_{x_k} \\ &= \left\| \tilde{x}_k \right\|_2 \cdot \frac{\tilde{x}_k}{\left\| \tilde{x}_k \right\|_2} + \mu_{x_k} = c_k \cdot s_k + l_k, \end{aligned} \tag{1}$$

where $\|\cdot\|_2$ denotes the ℓ_2 norm, μ_{x_k} stands for the mean value of x_k, and $\tilde{x}_k$ designates the mean-removed patch. The desired contrast for the target patch is determined by the highest contrast of K corresponding patches in source images, *i.e.* $\hat{c} = \max_{1 \le k \le K} c_k$. The final structure can be constructed by $\hat{s} = \frac{\bar{s}}{\|\bar{s}\|_2}$ with

$$\bar{s} = \frac{\sum_{k=1}^{K} \|\tilde{x}_k\|_p \cdot s_k}{\sum_{k=1}^{K} \|\tilde{x}_k\|_p}, \tag{2}$$

where $\|\cdot\|_p$ means the ℓ_p norm. As can be seen, the fusion is carried out by a simple weighted sum operation.

Unfortunately, such a fusion way would fail to form reasonable results when the source images under two extreme exposure conditions as shown in Fig. 1(e). Also, when the source images or some regions are all under poor exposure conditions, it will be not able to recall rich details, as only information from the original source image is computed. In other words, the techniques along this line do not make full use of the information existing in the images. Besides, several works [10,30] attempt to convert the unsupervised setting into a supervised one[1] by employing the best fused images from different existing methods as the pseudo ground truths. The shortcomings inherited from candidate methods are not fundamentally resolved, and thus the performance is restricted although improved, as shown in Fig. 1(g). Moreover, none of the mentioned methods can perform correction or enhancement from single images like ours (see Fig. 1(c) and (d)).

[1] Here, we argue that, due to the enhancement nature, it is a bit problematic to handle the MEF task with somehow constructed pseudo ground truth, which becomes a restoration task.

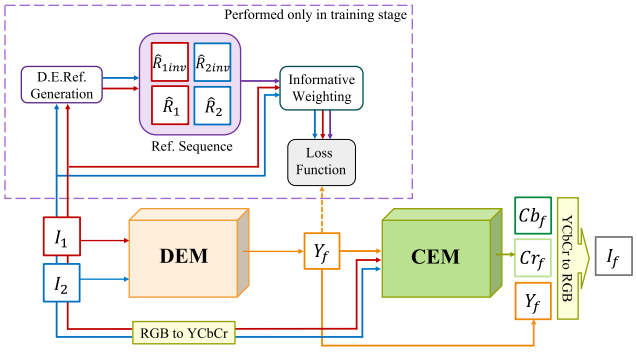


Fig. 2. The pipeline of our proposed DPE-MEF. D.E.Ref. denotes the detail enhanced reference, DEM and CEM represent the detail enhancement module and color enhancement module, respectively.

Another point that needs to be concerned is the visual aesthetics of fused images. Among diverse factors, we observe that the color plays a crucial role. Most of existing MEF approaches first convert the RGB color-space of images to a luminance and chrominance separation one, then execute fusion strategies only on the luminance channel. But, in such a way, the color of fused images is often distorted and unrealistic, because of ruined color information of badly-exposed images and/or the nonlinear relationship between color appearances and different exposures. As analyzed above, two key questions to generating high-quality multi-exposure image fusion results arise:

1. *How to exploit contents as detailed as possible from given source images?*
2. *How to recover visually pleasant and realistic colors for fused results?*

This study is to answer the above questions.

4. Deep perceptual enhancement for MEF

This section introduces a network called DPE-MEF to enhance the perceptual quality of fused images in an unsupervised setting. The blueprint of DPE-MEF is schematically illustrated in Fig. 2. As

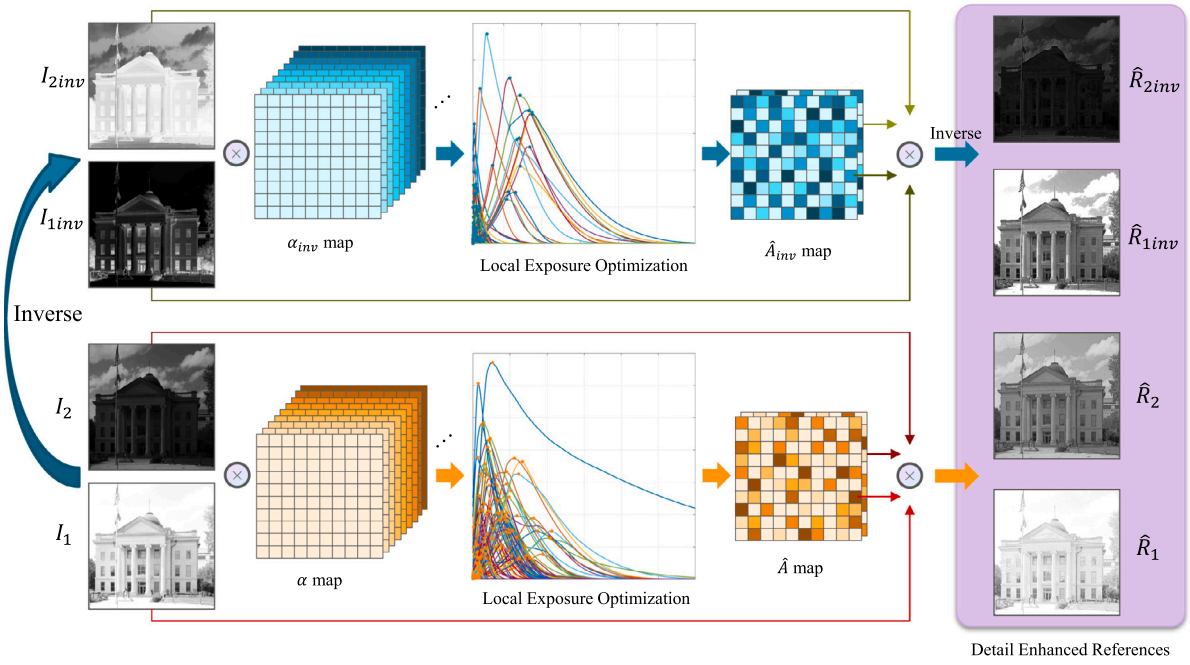


Fig. 3. The processing procedure for detail enhanced reference image generation.

discussed, the network should simultaneously take care of the informativeness and the visual aesthetics. From the perspective of functionality, we logically partition the entire network into two subnets, *i.e.*, detail enhancement module and color enhancement module, according to the two key questions. By the partition, the original problem is decoupled into two smaller ones, and thus the complexity is considerably reduced. In addition, it is natural to convert the working color-space from RGB to YCbCr, as, compared to the RGB color-space, the YCbCr can effectively separate the brightness and texture (Y channel, the luma component) from the color (Cb and Cr channels, the chroma component). The detail enhancement and color enhancement can be respectively executed on the luma and chroma components, which well fit our design. In what follows, we will detail the two modules.

4.1. DEM: Detail enhancement module

Let us now concentrate on the unsatisfied detail issue of under-exposed and over-exposed images. For under-exposed images, the high dynamic range information is squeezed in the limited ranges, while for over-exposed ones, the information is upper shifted and somewhere truncated, both of which lead to low contrast and ruined details. Since there is no ground-truth image can serve as the optimization target in multi-exposure fusion task, the core mission is to find a way to fully mine the information within the source images, enhance contrast and save details, so as to provide guidance for the optimization of DEM network.

4.1.1. Detail enhancement rule

Given an image I, a global gain can be easily obtained by $\hat{I}^{\alpha} = \alpha \cdot I$ with the exposure adjustment ratio α. The image is upgraded into a higher exposure (brighter) level with $\alpha > 1$, while being degraded into a lower (darker) one with $\alpha < 1$. Please note that the illumination of different areas in the same scene could greatly vary in an image. There may exist over-exposed, proper-exposed and under-exposed regions at the same time. Through globally tuning $\alpha > 1$, although the under-exposed areas will be brightened, the originally proper-exposed parts will turn out to be over-exposed due to the expression limit of digital images. In other words, hardly an optimal α can be found under the circumstances. Hence, a locally adaptive rule is desired.

Inspired by the Retinex theory [34], an image can be decomposed into two layers, say albedo and shading, or reflectance and illumination. In this work, we alternatively decompose I in the form of $I = R \circ E$, where R and E represent scene detail and exposure[2] components, respectively. The operator $\circ$ means element-wise product. By a simple algebraic transformation, we arrive at $R = \frac{1}{E} \circ I$, where $\frac{1}{E}$ is the element-wise reversed E. For ease of explanation, we denote $\frac{1}{E}$ by A. Since A (or equivalently E) is spatially variant, so is the adjustment. Notice that if all the elements in A are of the same value, the adjustment degenerates to the global one.

By assuming that R contains the richest details, the adjustment should work towards seeking an optimal $\hat{A}$ from I. For each location (i, j), we determine $\hat{A}_{ij}$ based on local area statistics-the local mean μ_{ij} and the standard deviation σ_{ij}-within a window of radius r surrounding I_{ij} (denoted by P_{ij}). The standard deviation can be viewed as a measurement of detail richness, which is computed by:

$$\sigma_{ij} = \frac{1}{2r+1} \sqrt{\sum_{-r \le p,q \le r} (I_{i+p,j+q} - \mu_{ij})^2}. \tag{3}$$

As the value of $\alpha > 1$ gradually increases in a pre-defined set $\{\alpha_1, \ldots, \alpha_T\}$, the standard deviation of patch $\hat{P}_{ij}^{\alpha_t} = \alpha_t \cdot P_{ij}$ will accordingly increase to $\sigma_{ij}^{\alpha_t} = \alpha_t \sigma_{ij}$ until the appropriate exposure is reached. Then, as we continue to increase α, the details will be progressively truncated due to the over exposure, and thus $\sigma_{ij}^{\alpha_t}$ drops. Please see Fig. 3 for illustration. In the sequel, $\hat{A}_{ij}$ can be filled out in a similar manner to Eq. (2), that is $\frac{\sum_{t=1}^{T} (\sigma_{ij}^{\alpha_t})^p \cdot \alpha_t}{\sum_{t=1}^{T} (\sigma_{ij}^{\alpha_t})^p}$ with $p \ge 0$ having many options. In this work, we simply choose $p = \infty$, which is to pick the value of α_t corresponding to the maximal σ_{ij}^{α}. Having the exposure adjustment map $\hat{A}$, the detail component $\hat{R}$ can be immediately obtained by $\hat{A} \circ I$. Compared with previous methods only using original source images themselves to accomplish the fusion, like [13,28,29], our proposed rule implicitly generates T virtual images of different exposures from each source image. In other words, our adjustment explores much more

[2] Different from the concept of camera exposure, E reflects how strong the detail is enhanced.

information (T times referred images). Moreover, we do not need to explicitly process T times inputs, thus saving the computational cost.

Bi-directional detail enhancement. As may be noticed, the above enhancement is merely upward, mainly stretching the contrast for under-exposed areas. It cannot deal with over-exposed regions, because reducing the exposure by setting $\alpha < 1$ will always lead to a smaller standard deviation than the original image. Therefore, the areas will be maintained as it is. To make better use of the content in relatively over-exposed areas, we invert the source image by $I_{inv} = 1 - I$. In the inverted image, the originally over-exposed regions would appear like underexposed ones. Then, we apply exactly the same enhancement rule on I_{inv}. After calculating the corresponding adjustment map $\hat{A}_{inv}$, the downward enhancement version is captured as $\hat{R}_{inv} = 1 - \hat{A}_{inv} \circ I_{inv}$. By the bi-direction detail enhancement, each source image has two enhanced references, *i.e.*, $\hat{R}$ and $\hat{R}_{inv}$. Fig. 3 summarizes the whole procedure of generating enhanced references from source images. As can be seen, the squeezed details of under-exposed areas are stretched significantly by the upward enhancement, while those of relatively over-exposed are effectively amplified from the downward process.

4.1.2. Architecture & loss function

Taking two source images I_1 and I_2 with different exposures as input, the DEM is expected to generate a "good" luma component Y_f with richer details for the fused image, while the chrominance component will be taken care by the CEM. The function of DEM can be formulated as follows:

$$Y_f = \mathcal{N}_{\text{DEM}}(I_1, I_2, \theta_{\text{DEM}}), \tag{4}$$

where $\mathcal{N}_{\text{DEM}}$ denotes the DEM network with the parameter θ_{DEM} to learn. The detailed network architecture of DEM is shown in Fig. 4. We simply employ a UNet-like [35] encoder–decoder architecture as our backbone. Specifically, the DEM consists of two encoders, one of which, called the joint encoder, receives both two source images as input and aims at extracting the correlation features between two source images, while the other one encodes each source image separately, tends to discover the representative information from each input, namely the discriminative encoder. Then, a decoder takes the output features from the two encoders as input, and receive skip connections of the encoders at each scale, to generate the final fused luma component.

To guide the training of DEM (in this part, two-exposure fusion is considered) for producing desired results, the loss function takes into account the luminances of source images (*i.e.*, Y_1 and Y_2) together with the detail enhanced references from each source luminance via the bi-direction enhancement rule (*i.e.*, $\hat{Y}_1$, $\hat{Y}_{1inv}$, $\hat{Y}_2$, $\hat{Y}_{2inv}$). It involves $Q = 6$ references in total, say $\hat{R}_q \in \{\hat{Y}_1, \hat{Y}_{1inv}, \hat{Y}_2, \hat{Y}_{2inv}, Y_1, Y_2\}$, which can regularize the learning from different angles. As a result, the loss function can be written in the following shape:

$$\mathcal{L}_{\text{DEM}} = \sum_{q=1}^{Q} \gamma_q \left(\ell_{pix}(\hat{R}_q, Y_f) + \ell_{per}^{\phi}(\hat{R}_q, Y_f) \right). \tag{5}$$

The ℓ_{pix} represents the normalized Manhattan distance between each $\hat{Y}_q$ and Y_f as:

$$\ell_{pix}(\hat{R}_q, Y_f) = \frac{1}{HW} \|\hat{R}_q - Y_f\|_1, \tag{6}$$

where H and W are the height and width of the inputs, same as of the outputs. Further, the ℓ_{per}^{ϕ} term denotes the perceptual loss [36] defined as:

$$\ell_{per}^{\phi}(\hat{R}_q, Y_f) = \sum_{l} \frac{1}{C_l H_l W_l} \|\phi_l(Y_f) - \phi_l(\hat{R}_q)\|_1, \tag{7}$$

where ϕ_l represents the lth layer in the perceptual network. C_l, H_l, W_l are the dimensions of the tensor feature map of the lth layer. This work adopts pre-trained VGG-19 Network [37] for perceptual feature extraction, where l indicates the layer index of {$conv1\_1$, $conv2\_1$, $conv3\_1$, $conv4\_1$, $conv5\_1$}. As can be seen from Eq. (5), it considers the guidance from both deep-feature and original-image domains.

One may tune the hyper-parameters γ_qs to select a satisfied learning configuration of DEM for multiple trials. For making the training free of tuning, we design an automatic manner to determine the values of γ_q. Suppose the total energy or informativeness of the given references can be evaluated by $\Gamma = \sum_{q=1}^{Q} \|\nabla \hat{R}_q\|_1$, where ∇ means the Laplacian operator. The importance/weight of each $\hat{R}_q$ can be reflected by the proportion it occupies from Γ, *i.e.* $\gamma_q = \|\nabla \hat{R}_q\|_1 / \Gamma$. Note that there have other metrics for measuring informativeness, and thus for determining the values of γ_qs. In our experiments, this auto way performs sufficiently well.

4.2. CEM: Color enhancement module

As aforementioned, color information plays an important role in the subjective evaluation of image quality. Different from other fusion tasks, such as infrared and visible image fusion and medical image fusion, the goal of MEF is to produce visually pleasant fusion results. Therefore, enforcing fusion images to have vivid and realistic colors can significantly promote the visual performance of an MEF algorithm. However, in most of existing MEF techniques, the color factor has been paid little attention. For example, in [13,14,22,28], sources images are converted to the YCbCr color-space, then fusion strategies are performed only on the Y (luminance) channel, while the fusion rules of Cb and Cr (chrominance) channels are still designed in a straightforward form. The most commonly adopted rule is the weighted summation suggested in [28] as follows:

$$C_f = \frac{C_1(|C_1 - \tau|) + C_2(|C_2 - \tau|)}{|C_1 - \tau| + |C_2 - \tau|}, \tag{8}$$

where C_1 and C_2 denote the Cb (or Cr) channels of the input image pair, and C_f is the corresponding fused chrominance channel. The value of τ is often set to 128.

However, when the source images are badly exposed, the color information may be interfered or even ruined due to the limited quality of digital devices. In addition, the color under different lighting conditions is not consistent. Under the circumstances, the color obtained directly by weighted summation will be unreasonable. To mitigate this issue, we customize a module called color enhancement module (CEM). It aims to learn the color mapping from the target (fused) luminance together with the source images to a suitable chrominance for the fused image.

The CEM is expected to infer the chrominance most suitable for the fused luminance (generated from DEM), by taking the whole information of two source images (both the luminance and chrominance components), and the target luminance as input, as follows:

$$[Cb_f, Cr_f] = \mathcal{N}_{\text{CEM}}(I_1, I_2, Y_f, \theta_{\text{CEM}}), \tag{9}$$

where $\mathcal{N}_{\text{CEM}}$ denotes the CEM network with the parameter θ_{CEM} to learn. The CEM is set as a joint encoder–decoder structure with 4-layers in each, to explore the color mapping relationship between the input images, as shown in Fig. 5.

In real situations, there are no color ground truths for the fused images. This is to say, we cannot execute the training by this means. But notice that we have multiple images with different exposures for a scene. These images are genuinely captured by cameras, which provide relatively proper and realistic color information at corresponding luminance conditions. For the sake of training the CEM to infer colors for given luminances, we alternative resort to the existing real data. To be concrete, three images of each sequence are randomly selected, then feed two of the selected images together with the luminance of the third image into the CEM. The chrominance component of the third image naturally serves as the reference (Cb_3 and Cr_3). By this means, the CEM can be trained by minimizing the gap between the estimate and the reference. In this work, we simply adopt the ℓ_1 (Manhattan) distance to measure the difference as:

$$\mathcal{L}_{\text{CEM}} = \|Cb_f - Cb_3\|_1 + \|Cr_f - Cr_3\|_1. \tag{10}$$

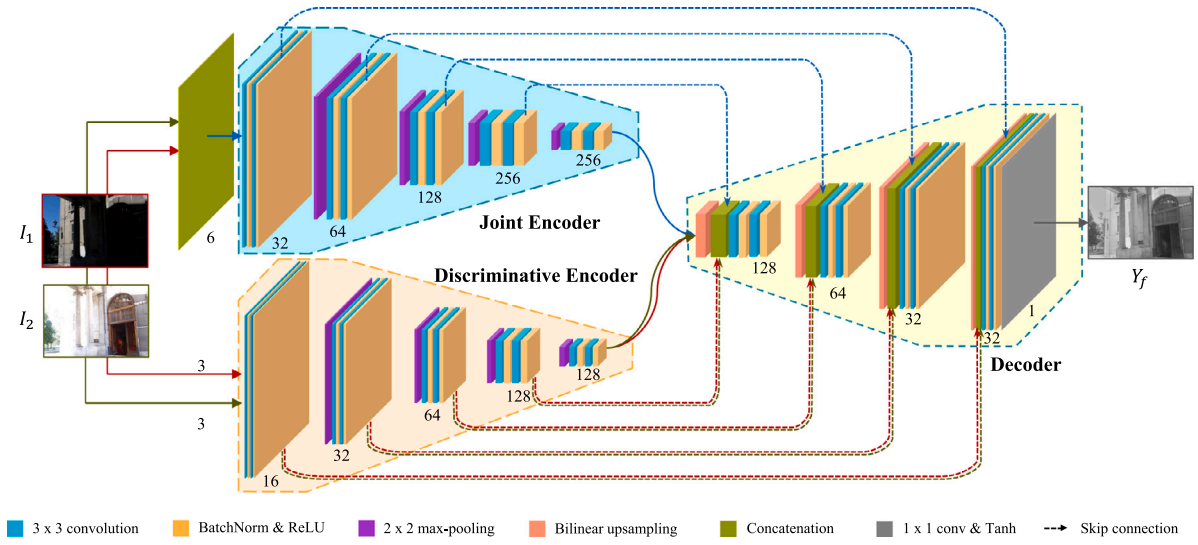


Fig. 4. The architecture of detail enhancement module. The numbers indicate the channel amounts.

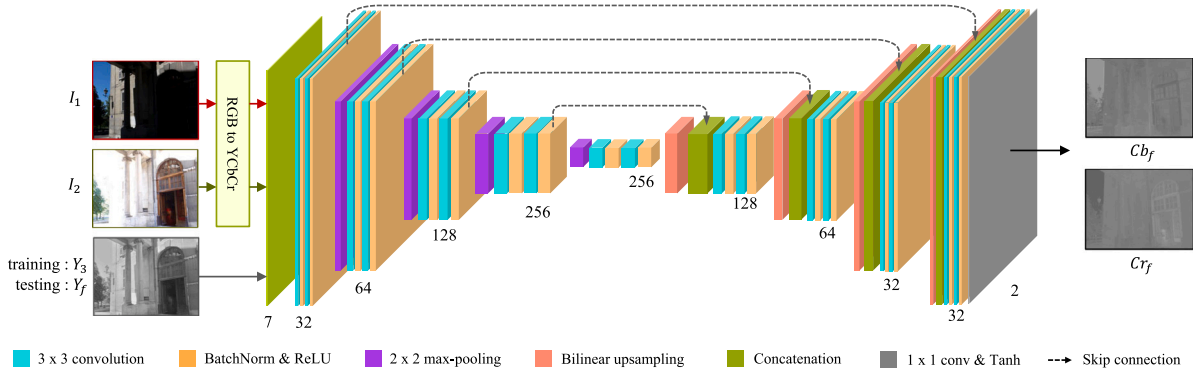


Fig. 5. The architecture of color enhancement module. The numbers indicate the channel amounts. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

As can be observed, the training of CEM can be also free of parameter tuning. Of course, one can adjust the weights between the two terms. But, considering the definitions of Cb (blue relative to green) and Cr (red relative to green), we appeal to treat them equally.

5. Experimental validation

5.1. Implementation details

The training and evaluation of the proposed DPE-MEF are carried out on the SICE dataset [38], which provides 589 multi-exposure image sequences of both indoor and outdoor scenes. The images in each sequence are taken by consumer grade cameras and are well-aligned. We randomly select 489 sequences for training, while the rest 100 sequences for testing. We choose image pairs with large exposure difference from each sequence to form the test set, since they are more challenging and could better evaluate the ability of an MEF algorithm in extracting details and maintaining the global structure. Our framework is implemented in PyTorch. The training and testing for involved learning-based competitors are carried out all on an Nvidia 2080Ti GPU. Our DEM and CEM are trained separately thanks to the logical partition.

We randomly select 2 and 3 pictures from each exposure sequence to train DEM and CEM, respectively. The training images are resized to the size of 512 × 512. For the two modules, the batch size is set to 32 and the patch size is set to 128 × 128 with data augmentation performed (random flipping, rotating, resizing and cropping). We use the ADAM solver [39] to optimize the network, with default parameters and fixed learning rate $1e^{-4}$. In the training phase, we set the radius r of window to 5, and uniformly employ the pre-defined set $\{\alpha_1, \ldots, \alpha_T\}$ via varying ζ_t in the manner of $\alpha_t = 2^{\zeta_t}$. Please note that in some areas with extreme exposures, the information has been completely lost, even contains only noise information. For this kind of area, setting a large exposure enhancement ratio may still not reach a desired exposure, and will significantly increase the noise. Therefore, we empirically appoint $\zeta_t \in [0, 6]$ with an interval of 0.01, so as to avoid the negative impact of excessive enhancement. The bi-directional enhancement process is no longer required in the testing phase.

5.2. Performance evaluation

To demonstrate the advantages of the proposed DPE-MEF, we compare it with 9 competitors, including Mertens [22], GFF [23], DSIFT [17], MEF-Net [13], FMMEF [3], DeepFuse [28], U2Fusion [14], IFCNN [30] and MEF-GAN [10]. Among these methods, Mertens, GFF, DSIFT and FMMEF are traditional methods; DeepFuse, MEF-Net and U2-Fusion are unsupervised deep learning methods; IFCNN and MEF-GAN are supervised deep learning approaches.

5.2.1. Qualitative comparison

Figs. 6–9 provide visual comparisons on several typical image sequences. As can be seen from the figures, though the source images have large exposure gaps, our results can still maintain promising

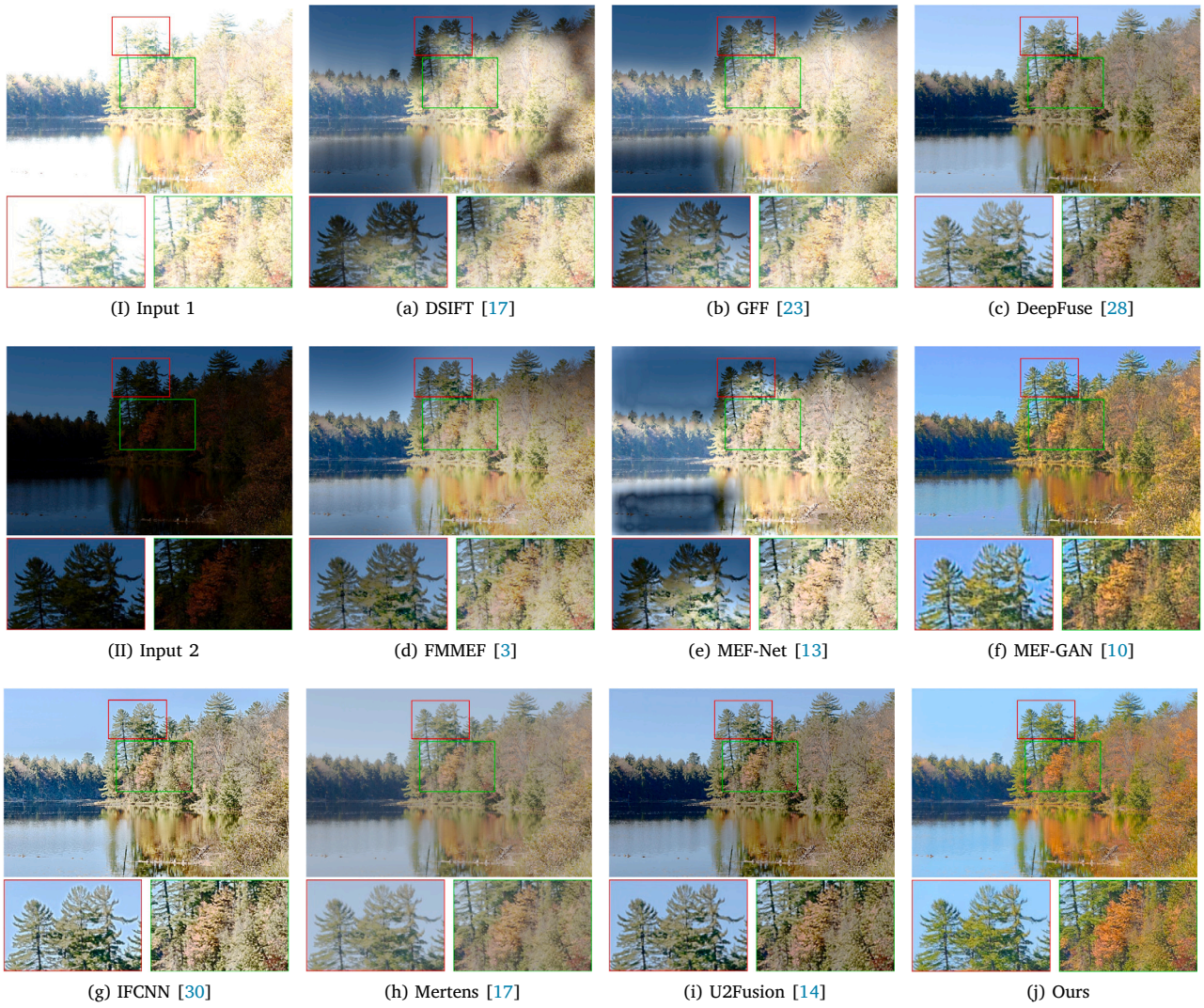


Fig. 6. Qualitative comparison on image pair 1. Please see zoomed-in patches for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

overall structures, and are adjusted to proper exposure levels. However, results from competitive methods, especially the traditional and weight map based ones, lose the consistency of the global structure and appear obvious artifacts. For example, in Fig. 7, results of (a), (b) and (e) suffer from noticeable brightness transition artifacts, while DL-based methods including (c), (f) and (i) fail to find a proper brightness for the grassland region.

Our fusion results are of high quality with enriched details, while some comparison methods, especially the deep learning-based methods like (f), are relatively blurry. The blur effect of MEF-GAN can also be found in quantitative comparisons. For example, its results in the metrics related to detail and gradient such as average gradient (AG) and edge intensity (EI), are obviously lower than other comparison methods. On the contrary, the contrast of the results of (f) is somewhat over-enhanced, and introduced some textures that do not exist in the source images (e.g. the artifact texture around the edge of branches in Fig. 6(g)).

Halo artifact is a widespread problem in the fusion results of existing methods, which has a great impact on the visual realism of fused images. In our concern, the appearance of halo artifacts in existing methods is mainly caused by two reasons: (1) Patch-based fusion strategy. For either patch-based traditional fusion method or deep learning fusion methods optimized by patch-based evaluation metrics (such as MEF-SSIM), the inaccuracy caused by patches overlap at the edges of the image is the culprit of halo artifacts. Since the proposed method does not adopt patch-based fusion strategy, it can effectively alleviate the halo caused by this reason; (2) Inaccurate pseudo ground truths with halo artifacts. The existing supervised deep learning methods usually use the optimal fusion results generated by existing methods as ground truths to perform training. Obviously, there may exist halos in pseudo ground truths due to reason 1, and the halos may migrate to the results of these supervised methods. Moreover, because supervised methods actually turn the MEF problem into an image restoration problem, the better restoration performance will even aggravate this situation. Since the optimization goal of the proposed method is obtained by mining the information within source images themselves and do not need pseudo ground truths, our method can also basically avoid halo artifacts caused by this reason. Therefore, our method alleviates the halo issue to a great extent and brings higher visual quality.

Owe to the color enhancement module, our results have more realistic and vivid colors even when the original images lack of color information due to extreme exposures, while results of competitors may suffer from the pale or unreal color issue, because of the absence of specific treatment on the color information. For example, in Fig. 6, the color information of the forest in the source images is ruined due to improper exposure, but in fact it is possible to infer the color of the

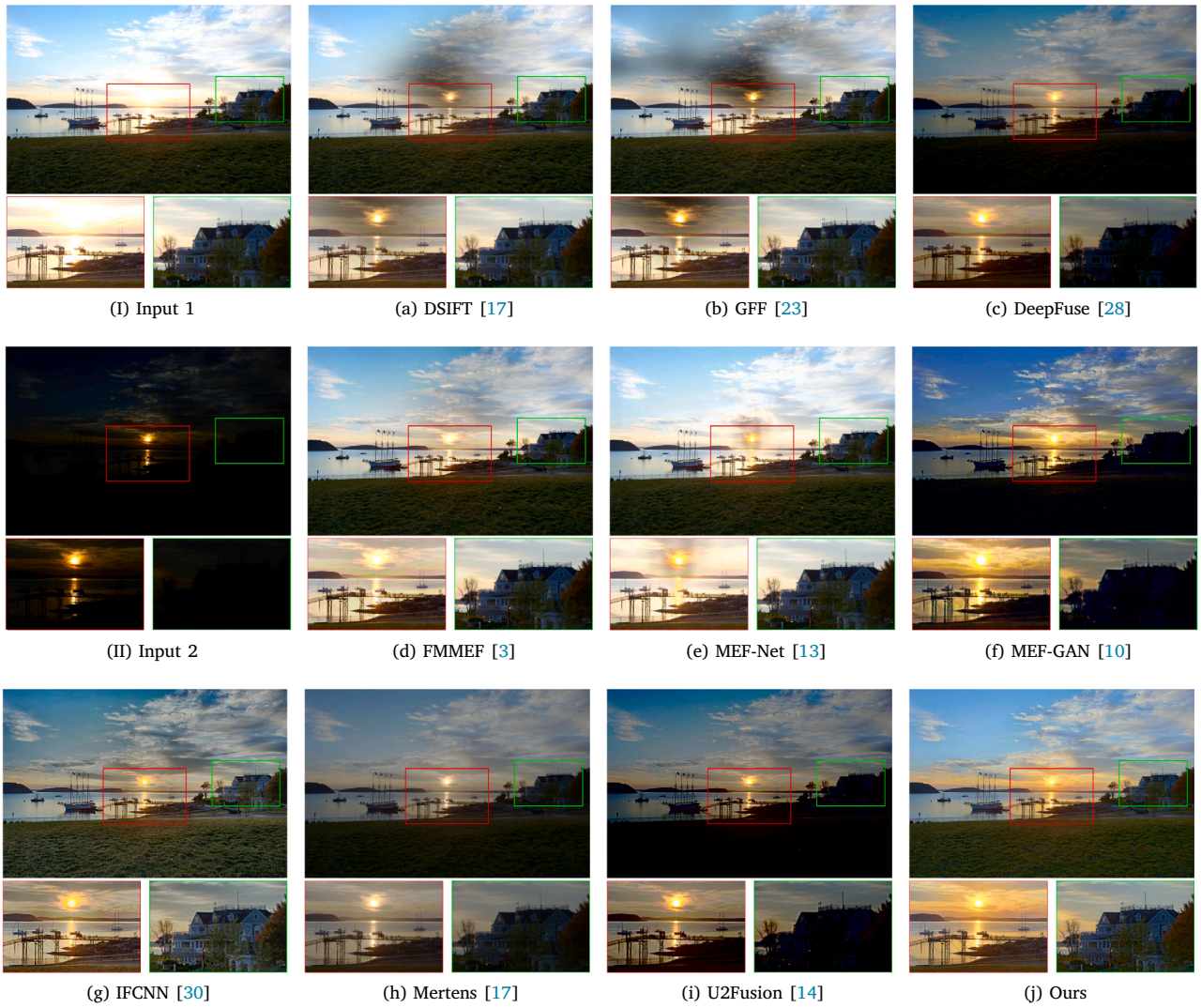


Fig. 7. Qualitative comparison on image pair 2. Please see zoomed-in patches for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

forest from the source images. Compared to the pale color obtained by traditional weighted summation method, the color corrected by CEM is more visually striking. The superiority of color can also be seen clearly from the sunset region in Fig. 7 and the sky in Fig. 8.

To further demonstrate the performance of our color enhancement module, we show more results in Fig. 10. The luminance of the fused images is obtained by DEM, while the color is obtained in different ways. Images in column (c) are results by the traditional weighted summation method, and in column (d) are results by CEM. As can be seen, due to the serious improper exposure of the original image, the color information in the image is very limited, thus the color obtained by the traditional weighted summation is relatively pale (please see color of the jungle in row 2, plate in row 3, sky in row 4) or deviated from the proper color (see sky color in row 1, palette in row 3). In contrast, the color fitted by our CEM is more vivid and realistic.

5.2.2. Quantitative comparison

Until now, there is no generally accepted optimal evaluation metric for MEF. In current works, MEF-SSIM is the most commonly used evaluation measurement, but as analyzed in Section 3 and in [10], it also has shortcomings, and is not sufficient to fully reflect the image quality of fusion images. In order to objectively evaluate the fusion algorithms more comprehensively, we adopt 8 metrics from different perspectives, including Average Gradient (AG) [40], fusion metric proposed by Chen et al. (Q_{CV}) [41], Edge intensity (EI) [42], Correlation Coefficient (CC) [43], Cross Entropy (CE) [44], Peak Signal-to-Noise Ratio (PSNR), Spatial Frequency (SF) [45], and MEF-SSIM [9]. Specifically, AG quantifies the gradient information of the fused image. Q_{CV} is a human vision system (HVS)-based fusion metric, which evaluates the quality of a fused image by dividing it into different local regions and transforming into the frequency domain, then measures the quality according to a human contrast sensitivity function (CSF). EI measures the sharpness of the edge in the fused image, CC cares the linear correlation between the fused image and source images, while CE considers the information differences between source images and fused image. PSNR is the ratio of peak value versus noise in the fused image, representing the distortion in the fusion process. SF measures the gradient distribution of an image. MEF-SSIM focuses on the structural similarity between the fused image and source images. For all the mentioned metrics except CE and Q_{CV}, larger values indicate better performance. As for CE and Q_{CV}, smaller values mean better performance.

We first offer the overall comparison in Table 1, and then give a more detailed per-image comparison on Fig. 11. It can be seen from the table that the results of our method achieve top-3 performance in terms

Fig. 8. Qualitative comparison on image pair 3. Please see zoomed-in patches for details. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of all the metrics from different aspects, which verifies the effectiveness of our proposed technique.

The proposed method achieves the best value in Q_{CV}, which says that our results are more consistent with human perception. The second best value is obtained by DPE-MEF in EI, SF and AG, indicating our method can generate fusion results with richer details and edge information. In the official implementation of IFCNN [30], the authors use CLAHE [46] as a post-processing step to further improve the performance of the algorithm. Here we give both quantitative results of IFCNN with and without CLAHE as post-processing. Note that these feature-based indicators may be also high due to the influence of noise, but by observing the visual quality of the fused images, we can see that the high values of our method are not caused by this case. The second best and third best values are obtained in other metrics, which verifies our results are close to the source images in content and structure. From detailed per-image comparison on different metrics in Fig. 11, we can see that our method reaches a stable and satisfactory performance on different images, which further reveal the robustness and wide applicability of our method.

Further, the comparison in terms of algorithm efficiency is given in Table 2. The DL-based methods are tested on an Nvidia GTX 2080Ti GPU, and the traditional methods are tested on Intel I7-8750H CPU. We can observe that our method is remarkably efficient (the second best result) among all competitors. We attribute the high time efficiency of the proposed method to two reasons. First, the backbone of DPE-MEF is a simple UNet-like encoder–decoder structure that operates on smaller-size features in intermediate layers. Compared to other network structures that stack more CNN layers to extract features at the original size of source images, the downsample–upsample structure usually takes less time in a single feedforward process. Second, DPE-MEF can perform end-to-end fusion without additional pre-/post-processing steps, which further improves the time efficiency. Although DPE-MEF expands the number of features, resulting in a larger model size than other methods, this setting leads to lower GPU utilization as the size of features in intermediate layers is reduced. Note that when resources are limited, the number of feature channels in the network can be further reduced, and satisfactory fusion results can still be obtained. The fastest method MEF-Net [13] promotes the efficiency by downsampling the images and then performing operations on the downsampled images, while DPE-MEF performs fusion at original size to ensure the fusion quality. Our method can achieve real-time (over 60 pairs per second) fusion for 720p image pairs, which significantly boosts the practicability.

Fig. 9. Qualitative comparison on image pair 4. Please see zoomed-in patches for details.

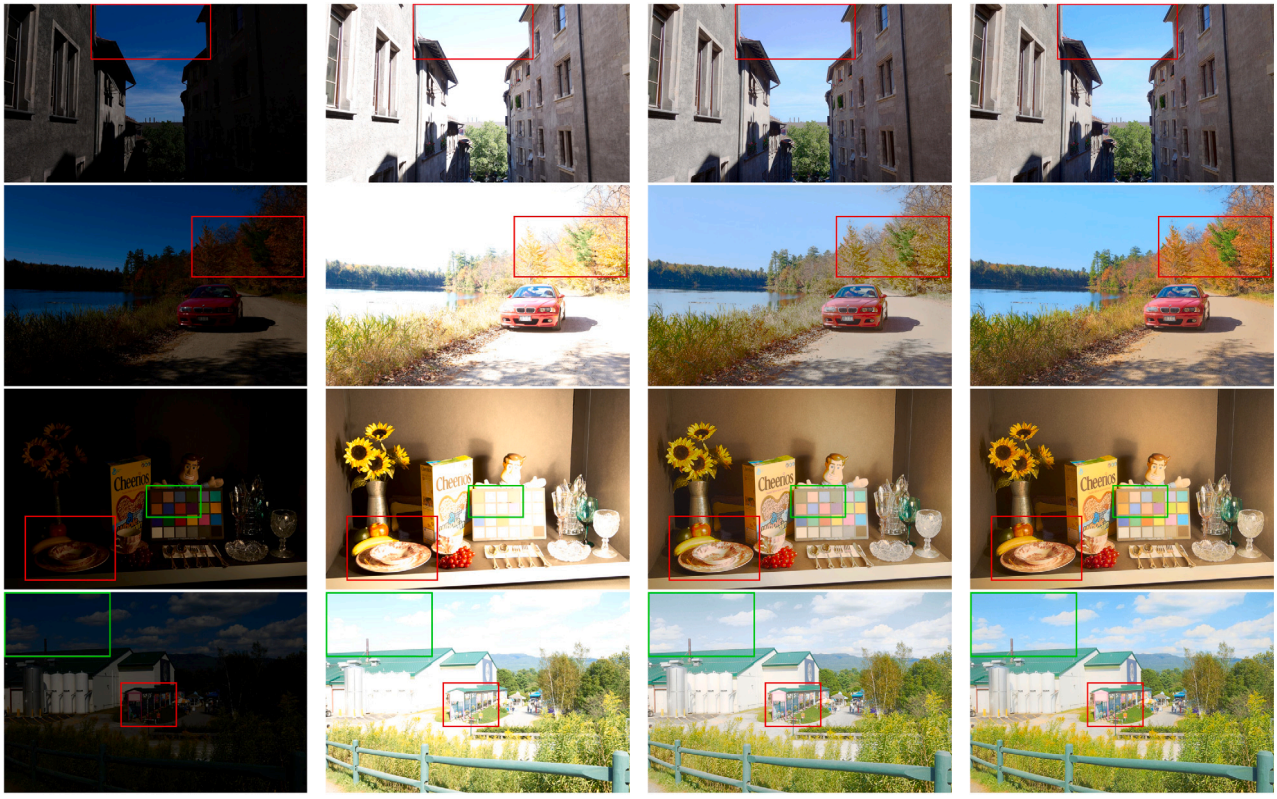
Table 1
Quantitative comparison on the SICE dataset in terms of various metrics. The sign (↓) indicates that the smaller the value, the better the performance. The top 3 results are highlighted in red, green and blue, respectively.

Methods	AG	Q_{CV}(↓)	EI	CC	CE (↓)	PSNR	SF	MEF-SSIM
Mertens [22]	4.8974	282.086	47.731	0.9438	3.2887	58.788	15.420	0.9003
DSIFT [17]	6.5190	1225.778	63.486	0.5769	2.3745	58.140	20.704	0.8301
GFF [23]	7.0395	1468.066	68.533	0.5659	3.5089	58.035	22.419	0.8266
FMMEF [3]	7.0923	912.832	69.151	0.7316	3.3653	58.430	22.642	0.8951
DeepFuse [28]	5.9092	228.444	57.729	0.9468	3.3518	58.685	19.094	0.8874
IFCNN [30]	11.4420	405.682	108.310	0.8874	3.4026	58.414	34.779	0.9017
IFCNN (w/o CLAHE)	*6.2467*	*295.671*	*58.669*	*0.9329*	*2.685*	*58.715*	*20.641*	*0.8928*
MEF-Net [13]	7.6236	590.068	74.281	0.7097	3.5324	58.266	24.547	0.9129
MEF-GAN [10]	5.9652	337.256	61.759	0.9235	3.0703	58.404	16.946	0.8712
U2Fusion [14]	7.1711	239.052	72.584	0.9394	3.6110	58.569	22.322	0.8199
Ours	8.5952	212.027	79.692	0.9436	2.8863	58.596	27.508	0.9058

5.3. Ablation study

In DEM module, we optimize the network parameters by computing the Manhattan distance between the fused luminance with the references on both original images and features. In this section, we give more results under different settings on the loss terms and used references, to verify the rationality of our design.

Ablation on different loss combinations. We present the results trained under different loss combinations in Fig. 12. For the pixel-level constraint ℓ_{pix}, we evaluated two specific variants, one is Manhattan distance (denoted as $\ell_{pix}^{\ell_1}$), the other is Euclidean distance(denoted as $\ell_{pix}^{\ell_2}$). As can be seen, in the result trained by adopting only ℓ_{pix}, due to the tight constraint, the details might be slightly over enhanced, and some artifacts appear around edges. In results trained by using only ℓ_{per}^{ϕ}, the overall structure and brightness are well preserved, but the fine-grained details are lost somewhere due to the loose restriction. Jointly optimizing ℓ_{pix} and ℓ_{per}^{ϕ} (finally used in DPE-MEF), as shown in (g), could balance the two terms well and capture the fusion result



(a) Input 1 (b) Input 2 (c) Colored by weighted-addition (d) Colored by CEM (ours)

Fig. 10. Visual results of CEM. Given the fusion luminance from DEM, the column (c) is colored by the traditional weighted-addition method in Eq. (8), and (d) is colored by our CEM. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Efficiency comparison on 720p test image pairs (unit: second for runtime, MB for model size and GPU usage). The best result in runtime is highlighted in red and the second best is in blue.

Method	Runtime	Platform	Model size	GPU usage
Mertens [22]	0.6547	MATLAB (CPU)	–	–
DSIFT [17]	2.2555	MATLAB (CPU)	–	–
GFF [23]	1.3554	MATLAB (CPU)	–	–
FMMEF [3]	1.1942	MATLAB (CPU)	–	–
DeepFuse [28]	0.1828	TensorFlow (GPU)	0.34	3137
IFCNN [30]	0.0475	PyTorch (GPU)	0.34	4291
MEF-Net [13]	0.0082	PyTorch (GPU)	0.33	757
MEF-GAN [10]	0.8484	TensorFlow (GPU)	20.0	7809
U2Fusion [14]	0.7462	TensorFlow (GPU)	2.52	7759
Ours	0.0164	PyTorch (GPU)	51.9	2507

with better details and global structure. For the specific setting of ℓ_{pix}, the result constrained using Manhattan distance will be sharper, while the result using Euclidean distance may occur relatively smooth details, which can be seen from the comparison between Fig. 12(d) and (c). Therefore, better fusion results can be obtained by using Manhattan distance as pixel-level constraint, then combining loose perceptual loss term to provide feature level constraint, which is our final setting. The qualitative metrics of ablation experiments are shown in Table 3. From the table, we can see that the result trained with only $\ell_{pix}^{\ell_1}$ has higher values in edge and gradient related metrics, while the result of only ℓ_{per}^{ϕ} is lower, which is consistent with the analysis above.

Ablation on different reference combinations. We evaluate the effect of setting $\hat{Y}_q$ with different combinations in Fig. 13. Remind that we set $Q = 6$ in the final version of DPE-MEF (result shown in (f)), we alternatively give results without considering upward enhanced references $\{\hat{Y}_1, \hat{Y}_2\}$ in (c), without downward enhanced references $\{\hat{Y}_{1inv}, \hat{Y}_{2inv}\}$ in (d), and without source images $\{Y_1, Y_2\}$ in (e).

The significance of upward enhancement can be viewed in the comparison between (c) and other results with upward enhancement. Without the upward enhancement references $\{Y_1, Y_2\}$, the details of the dark regions (*e.g.* the door area) are not handled well and the overall brightness is dim.

Table 3
Quantitative comparison on different settings of loss and reference sequence combinations. The sign (↓) indicates that the smaller the value, the better the performance. The best and second best results are highlighted in red and blue.

	AG	Q_{CV} (↓)	EI	CC	CE (↓)	PSNR	SF	MEF-SSIM
$\ell_{pix}^{\ell_2}$	6.2078	310.771	58.245	0.9386	2.4292	58.664	19.931	0.8891
$\ell_{pix}^{\ell_1}$	9.4484	261.877	86.816	0.9430	3.2552	58.434	31.860	0.9116
ℓ_{per}^{ϕ}	7.5455	209.729	70.588	0.9019	2.1503	58.470	24.245	0.9387
w/ $\ell_{pix}^{\ell_2} + \ell_{per}^{\phi}$	6.8884	283.636	64.706	0.9072	2.4602	58.623	22.346	0.9434
w/o $\{\hat{Y}_1, \hat{Y}_2\}$	7.5096	213.747	69.596	0.9270	2.6196	58.591	24.701	0.8358
w/o $\{\hat{Y}_{1inv}, \hat{Y}_{2inv}\}$	6.7154	234.166	63.281	0.9015	2.2251	58.429	21.608	0.9563
w/o $\{Y_1, Y_2\}$	7.8843	306.773	71.997	0.9381	2.8837	58.481	25.948	0.8603
ours	8.5952	212.027	79.962	0.9436	2.8863	58.596	27.508	0.9058

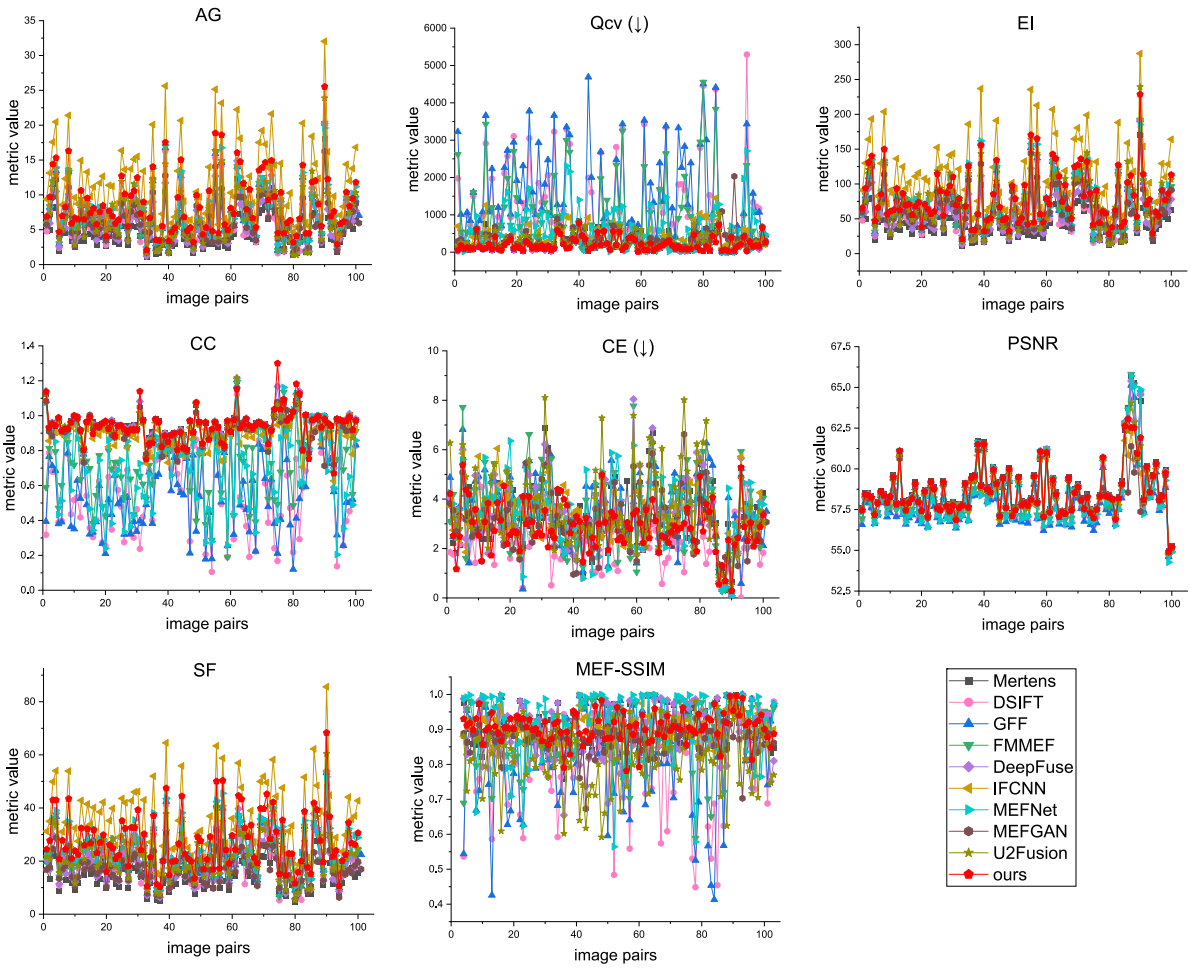
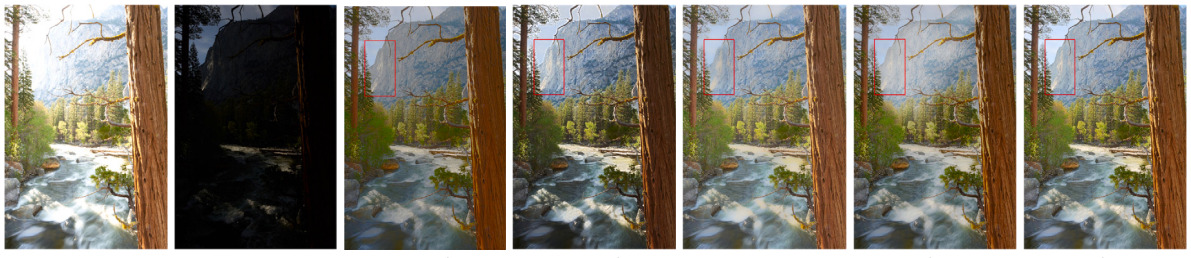


Fig. 11. Detailed per-image quantitative comparisons in terms of different metrics.



(a) Input 1 (b) Input 2 (c) $\ell_{pix}^{\ell_2}$ (d) $\ell_{pix}^{\ell_1}$ (e) ℓ_{per}^{ϕ} (f) $\ell_{pix}^{\ell_2} + \ell_{per}^{\phi}$ (g) $\ell_{pix}^{\ell_1} + \ell_{per}^{\phi}$ (ours)

Fig. 12. Ablation experiments on different loss combinations.

In contrast, the effectiveness of downward enhancement can be observed from the comparison between (d) and other results. Without the downward references $\{\hat{Y}_{1inv}, \hat{Y}_{2inv}\}$, the details in brighter areas (*e.g.* the white wall region) cannot be well corrected, resulting in a poor sense of hierarchy, and the whole fused image is slightly over-exposed.

To make full use of existing information, and considering the characteristics of the fusion task, we add source images to the reference sequence. The fusion results without the source images $\{Y_1, Y_2\}$ as reference are shown in (e). The difference between using the source images (f) and not using them (e) is mainly on the extent of enhancement. With the source images as guidance, the fusion results will not deviate from the exposure of source images too much while fully extracting the details of the source image, which is more consistent with the general expectation of fusion process, that is to gather information on the original scale of the source images. Without the source images, since the fusion process is only guided by the detail enhanced references, the fusion results (e) will be enhanced to a greater extent, which can be seen more clearly from the bottom row of Fig. 13.

5.4. Fusion under various conditions

Due to the high (extremely high in some situations) dynamic range of real scenes, the exposure ratio difference of the source images captured under different settings may vary in a very large range, making it difficult to set a proper exposure bracketing. In this part, we gives more results fused from source images with different exposure ratios

Fig. 13. Ablation experiments on different reference sequence combinations. The top row shows the fusion results in the case of one overexposed and one under-exposed, the middle row and bottom row show results in the case of both source images over-exposed and under-exposed, respectively.

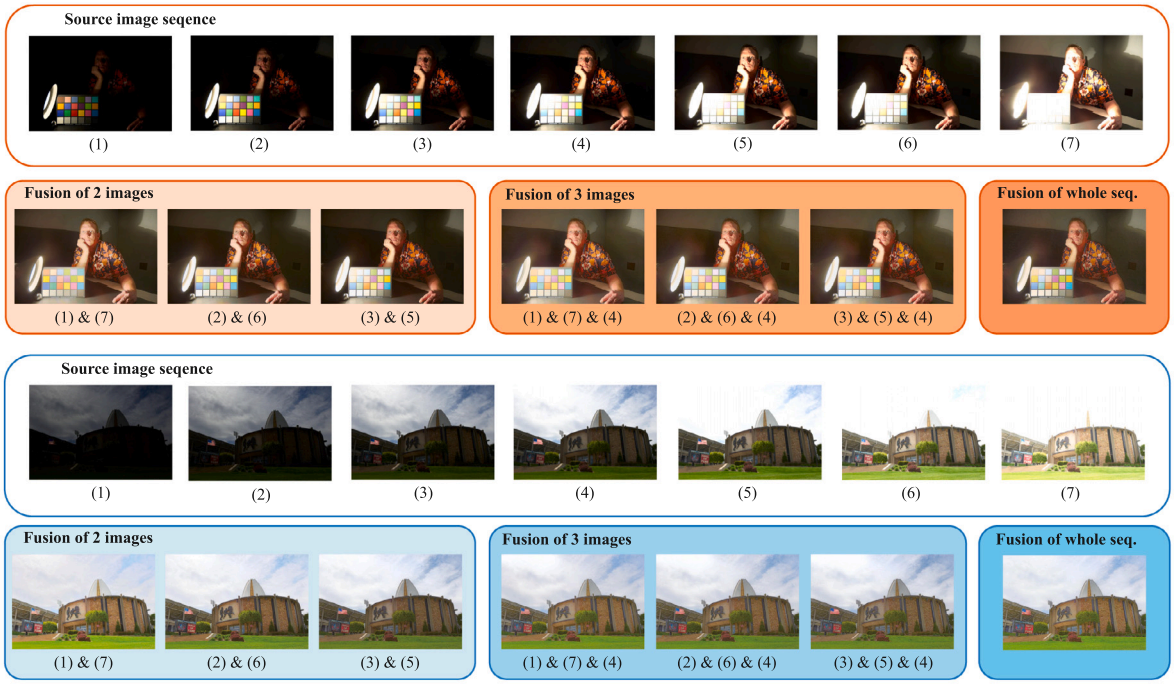


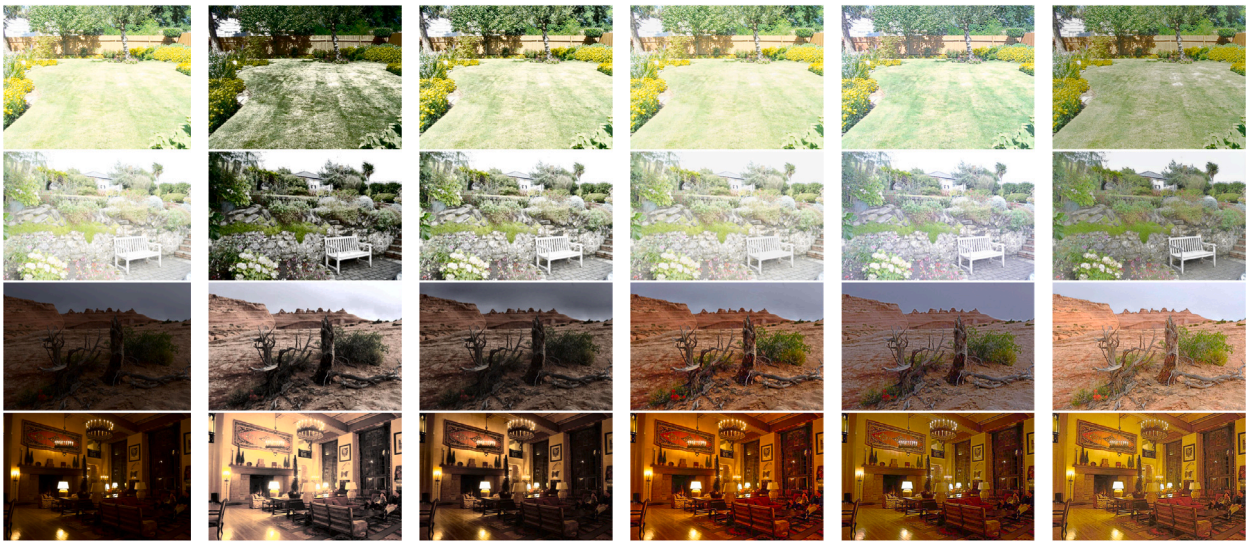
Fig. 14. Fusion results of source images under different exposure conditions, and results of multiple image fusion. Our method can produce satisfactory results under different exposure ratios, and can further improve the visual effect by fusing multiple images if available.

in Fig. 14. As can seen from the source image sequences, because of the large difference of brightness in the scene, it is difficult to obtain a dense image sequence with small exposure intervals, which may leads to extremely long exposure sequence. Also it is hard to determine which images could form the optimal exposure bracketing selection. In this case, we test different source images pairs, and perform fusion on them. The results show that DPE-MEF can perform high-quality fusion in different exposure cases, which verifies the robustness of our design. Our method also relaxes the strict requirement on the quality and quantity of source images, which further proves the applicability of our method. Moreover, if more source images are available, the proposed DPE-MEF could fuse them sequentially, and further improve the fusion quality, as shown in boxed regions in Fig. 14.

The main task of this work is the fusion of multiple exposure images. But recall that the design of our detail enhancement rule is in fact not limited to multiple images, it can also be used to optimize the local exposure of a single image. Therefore, in this part, we demonstrate the potential of our proposed detail enhancement method for processing single images, which is similar to the "exposure correction" task. A concise visual comparison with existing exposure correction works, including histogram equalization (HE) [47], contrast-limited adaptive histogram equalization (CLAHE) [46], high-quality exposure correction (HQEC) [48] and zero-reference deep curve estimation (Zero-DCE) [49] is given in Fig. 15, the top two rows are to correct over exposure, and the bottom two rows are to correct under exposure. Thanks to the bi-directional detail enhancement rule, our method can use a same model to deal with both over exposure and under exposure situations. As can be seen from the results, our DPE-MEF can effectively correct the inappropriate exposure areas in the images, enlighten the dark areas and retrieve the details of the over exposed areas.

6. Conclusion and discussion

In this paper, we have proposed a novel multi-exposure image fusion method, namely DPE-MEF. We discussed the characteristics of



(a) Input images (b) HE [47] (c) CLAHE [46] (d) HQEC [48] (e) Zero-DCE [49] (f) Ours-single

Fig. 15. Results on single image exposure correction. The leftmost column shows the input image, the columns (b)–(e) are results of existing methods, while the column (f) shows our results.

multi-exposure image fusion (MEF) task in detail, and positioned the goal of MEF task is to generate fusion results with both rich information and pleasant visual perception. Aiming at the goal, we formed our DPE-MEF with two sub modules: a detail enhancement module (DEM), which ensures the detail and structure of the fused image by fully mining the information within the source images, and a color enhancement module (CEM), which learns the mapping relationship between color and brightness in various scenes, and could render more vivid and realistic color for the fused image. We carried out detailed experimental comparisons and ablation experiments to verify the effectiveness and rationality of our proposed method, and demonstrated that our method has promising time efficiency and strong robustness to various scenes.

The current DPE-MEF mainly works on static scenes, that is, the source images need to be strictly registered. However, due to camera and object motion, foreground and background misalignment sometimes occur in the exposure sequences, which may lead to the failure of the static fusion methods to produce satisfactory fusion results. Therefore, how to extend the proposed method to deal with dynamic scenes is an import future research direction. In addition, due to the characteristics of CNN structure, it is hard to adjust the number of input images during testing for a trained network. How to design a more flexible and effective model structure in order to fuse adjustable numbers of source images in a single fusion process, is another important aspect to consider for practical use.

CRediT authorship contribution statement

Dong Han: Conceived and designed the research, Performed the experiments, Wrote the draft. **Liang Li:** Provided insightful advices, Revised the manuscript. **Xiaojie Guo:** Conceived and designed the research, Provided insightful advices, Revised the manuscript. **Jiayi Ma:** Provided insightful advices, Revised the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant nos. 62072327 and 61772512.

References

[1] K. Ma, Z. Wang, Multi-exposure image fusion: A patch-wise approach, in: IEEE International Conference on Image Processing, IEEE, 2015, pp. 1717–1721.
[2] S.-h. Lee, J.S. Park, N.I. Cho, A multi-exposure image fusion based on the adaptive weights reflecting the relative pixel intensity and global gradient, in: IEEE International Conference on Image Processing, IEEE, 2018, pp. 1737–1741.
[3] H. Li, K. Ma, H. Yong, L. Zhang, Fast multi-scale structural patch decomposition for multi-exposure image fusion, IEEE Trans. Image Process. 29 (2020) 5805–5816.
[4] J. Xie, L. Xu, E. Chen, Image denoising and inpainting with deep neural networks, Adv. Neural Inf. Process. Syst. 25 (2012) 341–349.
[5] C. Dong, C.C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[6] B. Cai, X. Xu, K. Jia, C. Qing, D. Tao, DehazeNet: An end-to-end system for single image haze removal, IEEE Trans. Image Process. 25 (11) (2016) 5187–5198.
[7] K. Zhang, W. Zuo, Y. Chen, D. Meng, L. Zhang, Beyond a Gaussian Denoiser: Residual learning of deep CNN for image denoising, IEEE Trans. Image Process. 26 (7) (2017) 3142–3155.
[8] K.R. Prabhakar, R.V. Babu, Ghosting-free multi-exposure image fusion in gradient domain, in: IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2016, pp. 1766–1770.
[9] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Trans. Image Process. 24 (11) (2015) 3345–3356.
[10] H. Xu, J. Ma, X.S. Zhang, MEF-GAN: multi-exposure image fusion via generative adversarial networks, IEEE Trans. Image Process. 29 (2020) 7203–7216.
[11] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[12] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: A survey, Inf. Fusion 45 (2019) 153–178.
[13] K. Ma, Z. Duanmu, H. Zhu, Y. Fang, Z. Wang, Deep guided learning for fast multi-exposure image fusion, IEEE Trans. Image Process. 29 (2020) 2808–2819.
[14] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. (2020).
[15] X. Zhang, Benchmarking and comparing multi-exposure image fusion algorithms, Inf. Fusion 74 (2021) 111–131.
[16] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[17] L. Yu, Z. Wang, Dense sift for ghost-free multi-exposure fusion, J. Vis. Commun. Image Represent. 31 (2015) 208–224.
[18] A.A. Goshtasby, Fusion of multi-exposure images, Image Vis. Comput. 23 (6) (2005) 611–618.

[19] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, L. Zhang, Robust multi-exposure image fusion: A structural patch decomposition approach, IEEE Trans. Image Process. 26 (5) (2017) 2519–2532.
[20] K. Ma, Z. Duanmu, H. Yeganeh, Z. Wang, Multi-exposure image fusion by optimizing a structural similarity index, IEEE Trans. Comput. Imaging 4 (1) (2018) 60–72.
[21] P.J. Burt, R.J. Kolczynski, Enhanced image capture through fusion, in: Fourth International Conference on Computer Vision, 1993, pp. 173–182.
[22] T. Mertens, J. Kautz, F.V. Reeth, Exposure fusion: A simple and practical alternative to high dynamic range photography, Comput. Graph. Forum 28 (1) (2009) 161–171.
[23] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[24] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, C.N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inf. Fusion 8 (2) (2007) 119–130.
[25] S. Paul, I.S. Sevcenco, P. Agathoklis, Multi-exposure and multi-focus image fusion in gradient domain, J. Circuits Syst. Comput. 25 (10) (2016) 1650123:1–1650123:18.
[26] Y. Zheng, X. Hou, T. Bian, Z. Qin, Effective image fusion rules of multi-scale image decomposition, in: 5th International Symposium on Image and Signal Processing and Analysis, IEEE, 2007, pp. 362–366.
[27] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion (2021).
[28] K.R. Prabhakar, V.S. Srikar, R.V. Babu, DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: IEEE International Conference on Computer Vision, 2017, pp. 4724–4732.
[29] Y. Qi, S. Zhou, Z. Zhang, S. Luo, X. Lin, L. Wang, B. Qiang, Deep unsupervised learning based on color un-referenced loss functions for multi-exposure image fusion, Inf. Fusion 66 (2021) 18–39.
[30] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, IFCNN: a general image fusion framework based on convolutional neural network, Inf. Fusion 54 (2020) 99–118.
[31] H. Jung, Y. Kim, H. Jang, N. Ha, K. Sohn, Unsupervised deep image fusion with structure tensor representations, IEEE Trans. Image Process. 29 (2020) 3845–3858.
[32] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, no. 07, 2020, pp. 12797–12804.
[33] H. Zhang, J. Ma, SDNet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. (2021) 1–25.
[34] E.H. Land, The retinex theory of color vision, Sci. Am. 237 (6) (1977) 108–129.
[35] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.
[36] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[37] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.
[38] J. Cai, S. Gu, L. Zhang, Learning a deep single image contrast enhancer from multi-exposure images, IEEE Trans. Image Process. 27 (4) (2018) 2049–2062.
[39] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[40] G. Cui, H. Feng, Z. Xu, Q. Li, Y. Chen, Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition, Opt. Commun. 341 (2015) 199–209.
[41] H. Chen, P.K. Varshney, A human perception inspired quality metric for image fusion based on regional information, Inf. Fusion 8 (2) (2007) 193–207.
[42] B. Rajalingam, R. Priya, Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis, Int. J. Eng. Sci. Invent. 2 (Special issue) (2018) 52–60.
[43] P. Shah, S.N. Merchant, U.B. Desai, Multifocus and multispectral image fusion based on pixel significance using multiresolution decomposition, Signal Image Video Process. 7 (1) (2013) 95–109.
[44] D. Bulanon, T. Burks, V. Alchanatis, Visible and thermal images for fruit detection, in: Encyclopedia of Agrophysics, Springer Netherlands, 2014, pp. 944–954.
[45] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[46] K. Zuiderveld, Contrast limited adaptive histogram equalization, Graph. Gems (1994) 474–485.
[47] R.C. Gonzalez, R.E. Woods, et al., Digital image processing, Prentice hall Upper Saddle River, NJ, 2002.
[48] Q. Zhang, G. Yuan, C. Xiao, L. Zhu, W.-S. Zheng, High-quality exposure correction of underexposed photos, in: Proceedings of the 26th ACM International Conference on Multimedia, 2018, pp. 582–590.
[49] C. Guo, C. Li, J. Guo, C.C. Loy, J. Hou, S. Kwong, R. Cong, Zero-reference deep curve estimation for low-light image enhancement, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1780–1789.