Exposure Fusion Using a Relative Generative Adversarial Network

Jinhua WANG†, Member, Xuewei LI††, and Hongzhe LIU††, Nonmembers

SUMMARY At present, the generative adversarial network (GAN) plays an important role in learning tasks. The basic idea of a GAN is to train the discriminator and generator simultaneously. A GAN-based inverse tone mapping method can generate high dynamic range (HDR) images corresponding to a scene according to multiple image sequences of a scene with different exposures. However, subsequent tone mapping algorithm processing is needed to display it on a general device. This paper proposes an end-to-end multi-exposure image fusion algorithm based on a relative GAN (called RaGAN-F), which can fuse multiple image sequences with different exposures directly to generate a high-quality image that can be displayed on a general device without further processing. The RaGAN is used to design the loss function, which can retain more details in the source images. In addition, the number of input image sequences of multi-exposure image fusion algorithms is often uncertain, which limits the application of many existing GANs. This paper proposes a convolutional layer with weights shared between channels, which can solve the problem of variable input length. Experimental results demonstrate that the proposed method performs better in terms of both objective evaluation and visual quality.

key words: exposure fusion, relative generative adversarial network, high dynamic range image

1. Introduction

With the rise of deep learning in the fields of computer graphics and computer vision, deep models, which are based on the powerful processing and analysis capabilities of convolutional neural networks for digital images, have also been applied in the field of high dynamic range (HDR) image generation [1]–[3]. This type of algorithm uses multiple image sequences of the same scene with different exposures to synthesize an HDR image of the scene. However, existing general display devices (such as CRT displays) can only display brightness levels of about two orders of magnitude in dynamic range. This situation is difficult to change in the short term because of the constraints of hardware costs. Therefore, the problem of effectively displaying the obtained HDR image on a commonly used device with a low dynamic range is a problem of tone mapping the HDR image. Because of the need for real-time implementation, the practical application of such algorithms is limited. This paper hence proposes a multi-exposure image fusion algorithm based on a relative generative adversarial network (RaGAN), which can both directly fuse multiple images with different exposures and generate a high-quality image directly for display on a general device without subsequent processing.

Deep convolutional neural networks can obtain deeper high-frequency features in a data-driven way, obtain richer image details, and generate more accurate fusion result images. Although the deep convolutional neural network has further improved the quality of image restoration, there are still many issues. The general consensus in the field is that expanding the width of the network (the number of filters) and increasing the depth of the network (the number of layers) can enhance the visual quality of the fusion result image. However, a deeper and more complex network structure reduces the convergence speed of the network, increases the difficulty of model training, and often causes gradient disappearance/gradient explosion problems. Moreover, a wider and deeper network structure generally leads to a sharp increase in the number of parameters. Complex and huge network models require more storage space than simple and compact networks. To solve the problems caused by deep convolutional neural networks, this paper proposes a multi-exposure image fusion algorithm that is constructed by learning with a shallow network. A convolutional neural network (named VDSR) is used to construct the generator network for super-resolution [4]. This method uses multiple convolutional layers and only learns the residual function, which greatly accelerates the convergence speed.

The SGAN [5] ignores prior knowledge, that is, half of the batch samples are false. This makes it difficult for the training process of the discriminator to converge. As a result, the discriminator in the SGAN has a vanishing gradient and cannot be trained to an optimal state. In this paper, the relative GAN (RaGAN) [6] is used to design the loss function. It does not measure the “probability that the input data is true,” but instead measures the “the probability that the input data is more true than the opposite type of sampled data.” The RaGAN can retain more details in the source images.

In addition, the number of input image sequences of multi-exposure image fusion algorithms is often uncertain, which means many existing GANs cannot be used. This paper proposes a convolutional layer with weights that are shared between layers, which can solve the problem of uncertain input length. Regardless of how many images the image sequence contains as input, after the inter-layer
shared convolution operation, a fixed number of feature maps can be obtained, which are then used as the input of the subsequent network.

The remainder of this paper is organized as follows. Some related work is discussed in Sect. 2. Section 3 provides a detailed explanation of our proposed method. Experimental results and performance evaluations are presented in Sect. 4, and Sect. 5 concludes the paper.

2. Related Work

Goodfellow et al. proposed the standard generative adversarial network (SGAN) algorithm [5]. It has been applied to many specific tasks, such as image generation, image super-resolution [7], text-to-image synthesis, and image-to-image translation. In a GAN, the aim of the discriminator is to distinguish between real data and generated data; whereas the aim of the generator is to generate fake samples that are as real as possible so that the discriminator believes the fake samples come from real data. The loss function is defined as:

\[
\begin{align*}
L^\text{GAN}_G &= -\mathbb{E}_r[\log(D(x_r))] \\
L^\text{GAN}_D &= -\mathbb{E}_r[\log(D(x_r))] - \mathbb{E}_g[\log(1 - D(x_g))]
\end{align*}
\]

where \(D(x) = \sigma(C(x))\) (\(\sigma\) is defined as sigmoid function) and \(C(x)\) denotes the non-transformed layer output of the discriminator, \(\mathbb{E}_x[\cdot]\) denotes the average value of the discriminant values of all false images in the batch. The loss function of SGAN is approximately equal to the Jensen-Shannon divergence (JSD) between the real data and the generated data. We can see that when minimizing the loss in SGAN, the method only increase \(D(x_f)\) but not decrease \(D(x_r)\). This makes it difficult for the training process of the discriminator to converge, and it is moreover difficult for the discriminator to make reasonable predictions.

A number of multi-exposure fusion methods have been proposed. Mertens et al. [8] presented a method to fuse multiple-exposure images based on processing separate Laplacian pyramids in the R, G, and B channels. The results contain brightness changes that are not consistent with the original source images. These are caused by significant changes in brightness among images with different exposure times. Goshtasby [9] proposed an exposure fusion method for multi-exposure images of a static scene. His approach blends the image blocks from a specific domain by selecting uniform image blocks that contain the most useful information. Because a block may span different objects, this approach cannot handle object boundaries. A generalized random walk framework has been proposed to calculate a globally optimal solution subject to certain quality measures, i.e., local contrast and color consistency, while combining the scene details revealed under different exposures [10]. This method does not incorporate a multi-resolution technique into the fusion process.

In addition, Gu et al. [11] proposed a gradient field multi-exposure image fusion method for HDR image visualizaton. The advantage of this method is its computational efficiency and robustness. Only two parameters are used, and they can generally be set to default values. However, the metric to measure the distance between intensities should be improved to reduce the need for tedious gradient modification. Song et al. [12] synthesized an exposure fusion image using a probabilistic model that preserves the luminance levels and suppresses reversals in the image luminance gradients. Hu et al. [13] presented a novel registration and fusion approach for exposure images in the presence of both dynamic scenes and camera motion. Their method tolerates a small amount of blur; however, sampling the whole irradiance range often requires fairly long exposures, and can therefore introduce a large amount of blur, thus potentially causing the method to fail. The authors do not provide any details for solving this problem. Shen et al. [14] proposed a novel hybrid exposure weight measurement that is guided not only by a single image’s exposure level, but also by the relative exposure level among different exposure images using a boosting Laplacian pyramid. Their method gives visually pleasing fusion images with more color and texture details. Their paper provides a comparison with some typical tone-mapping methods.

Deep learning was first adopted in the field of multi-exposure image fusion by Deepfuse [15]. Deepfuse is a deep unsupervised method which employs a metric called the multi-exposure fusion structural similarity index (named MEF-SSIM) [16] as the loss function and uses the convolutional neural network architecture to realize learning. Deepfuse limits the number of input images, and the model can only process two images as input. The method proposed in this paper is a supervised learning framework that uses GAN to build the network. Furthermore, this method has no limit on the number of input images.

3. Proposed RaGAN-EF Exposure Fusion Algorithm

The proposed end-to-end multi-exposure image fusion framework called RaGAN-EF is shown in Fig. 1. The fusion framework is mainly composed of a generator network, discriminator network, and feature extraction network. The generator network generates a fusion result image according to the input multi-exposure image sequence. The discriminator network is used to determine whether the input image is a real image or a generated fake image. The feature extraction network is used to extract image features and assist the generator network during training. The feature extraction network uses a pre-trained model, and there is no need to train the feature extraction network during the training process of the generator and discriminator networks.

As shown in Fig. 1, the multi-exposure image sequence is fed to the generator network after going through the inter-layer shared weight convolution layer. For multi-exposure image fusion, the number of input image sequences is often uncertain, which limits the use of many existing generator networks. To solve this problem, the convolutional layer with shared weights between layers is proposed to handle input with arbitrary lengths. Regardless of how many im-
age sequences are used as input, a fixed number of feature maps can be obtained after the shared convolution operation between layers, and the output feature maps are used as the input of the subsequent network. The formula for the shared weight convolution between layers is defined as follows:

$$F_1(Y) = \max(0, \sum_{i=1}^{N} W_i * Y_i)$$  \hspace{1cm} (3)$$

where $N$ represents the number of image sequences, $i$ represents the $i$th image of the input image sequence, and $W_i$ represents the filters, which are $f_1 \times f_1$ convolution kernels of size $n_1$. It can be seen that the role of $W_1$ is to perform $n_1$ convolution operations on the original image sequence. Each convolution operation uses an $f_1 \times f_1$ convolution kernel. The first layer of convolution outputs $n_1$ feature maps. This layer can be regarded as a part of the original image sequence. Note that “max” is a non-linear function. For example, if the input is an RGB color image, the size of the convolution kernel is $3 \times 3 \times 3 \times n_1$. When $n_1 = 1024$, the image sequence passes through a set of $3 \times 3 \times 3$ filters to obtain a feature image. After 1024 sets of filters, 1024 feature images are obtained, which can capture sufficient image feature data so that the subsequent network has sufficient information for learning. Convolution process of sharing weights between layers is shown in Fig. 2.

3.1 Generator and Discriminator Network Structure

If the image is constructed by learning the original mapping function through a shallow network, the convergence speed is slow, and the training time is long. To address these problems, Kim et al. proposed VDSR for super-resolution reconstruction in 2016 [4]. This method uses multiple convolutional layers and only learns the residual function, which substantially accelerates the convergence speed. In the deep network structure, small filters are cascaded multiple times to effectively use the context information of the large image area, so that the accuracy is significantly improved.

The VDSR method simply stacks multiple convolutional layers and adds the input image and the residual image at the end of the network to generate the final output image.
Except for the last layer, which uses one filter with a size of $3 \times 3$, each layer uses 64 filters with a size of $3 \times 3$ to learn the residual mapping. In the method proposed in our paper, we perform “0” padding before the convolution operation to keep the output image size of each layer the same. However, for training very deep networks, convergence speed is the key. If the convergence speed is only increased by increasing the learning rate, it may cause the gradient to disappear/explode. To solve this problem, Kim et al. proposed an adjustable gradient clipping method [4], that is, the gradient is clipped to $[-\theta/\eta, \theta/\eta]$, where $\eta$ is the current learning rate and $\theta$ is a value close to zero set in advance. In RaGAN-EF, VDSR is adopted to construct the generator network, its network architecture is shown in Fig. 3.

The discriminator network is used to determine whether the input image is a fake image generated by the generator network or a real image. Its output is a probability value. Its structure and details are shown in Fig. 4.

### 3.2 Loss Function

In this section, we present the definitions of generator loss and discriminator loss. The generator network generates a fusion result image according to the input multi-exposure image sequence. Its loss consists of four parts: adversarial loss, content loss, clarity loss, and feature loss.

1. **Definition of generator loss**

   The objective function of the generator network in the RaGAN-EF framework is defined as follows:

   $$
   \hat{\theta}_G = \arg \min_{\theta_G} \frac{1}{N} \sum_{n=1}^{N} \text{Loss}_G(G_{\theta_G}(I_{seq}^n), I^n) \tag{4}
   $$

   where $N$ represents the number of batch image samples, $I_{seq}^n$ represents the input multi-exposure image sequence, $I^n$ represents the real image, and $\theta_G$ represents the generator network parameters. Finally, $\text{Loss}_G$ represents the loss of the generator network, and its four parts are described in detail below.

   1. **Content loss**: This is the pixel-level MSE loss. Its formula is as follows:

   $$
   L_p = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} ((I^n_{xy}) - (G_{\theta_G}(I_{seq}))_{xy})^2 \tag{5}
   $$

   where $W$ and $H$ represent the width and height of the image. MSE loss can preserve the low-frequency information after image fusion, but the lack of high-frequency information may cause the image to be too smooth and the visual
effect is not very natural.

(2) Feature loss: This is the MSE value of the generated fake image feature and the real image feature. The features of the generated and real images are separately extracted by VGG19 [17]. Next, the MSE value of the two features is calculated, and this is regarded as the VGG feature loss. The definition $\phi_{ij}$ is the output of the $i$th convolutional layer before the $j$th maximum pooling layer in the VGG network. It is expressed as follows:

$$L_V = \frac{1}{W_iH_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (\phi_{ij}(I_r)_{xy} - \phi_{ij}(G_{\theta_g}(I_{seq}))_{xy})^2$$  \hspace{1cm} (6)$$

where $W_i$ and $H_i$ represent the width and height of the feature image respectively.

(3) Clarity loss: To determine this loss, the sharpness features of the generated fake and real images are first extracted and their MSE value is then obtained. The loss in sharpness can be determined by calculating the gradient. The gradient image is obtained by convolution with a filter for a grayscale image, thus, separate gradient images for the vertical and horizontal directions can be obtained. The clarity loss is calculated as follows:

$$L_C = \frac{1}{2W_iH_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (S_h(L_r) - S_h(G_{\theta_g}(I_{seq})))^2$$

$$+ \frac{1}{2W_iH_i} \sum_{x=1}^{W_i} \sum_{y=1}^{H_i} (S_v(L_r) - S_v(G_{\theta_g}(I_{seq})))^2$$  \hspace{1cm} (7)$$

where $S_h$ represents the horizontal gradient image, $S_v$ represents the vertical gradient image, $L_r$ represents the grayscale image corresponding to the real image, $G_{\theta_g}(I_{seq})$ represents the grayscale image corresponding to the generated fake image, and $W_i$ and $H_i$ represent the size of the gradient images.

(4) Adversarial loss: The objective function of the discriminant network in the RaGAN-EF framework is defined as follows: the discriminator loss does not measure “the probability that the input data is real,” but instead measures “the probability the input data is more real than the opposite type of random sampling data.” We can see that the relative discriminator uses the missing attributes required in the SGAN. To make the relative discriminator more global, the relative discriminator is defined in the average sense [6]; hence, the adversarial loss of the generator network is defined by the cross entropy loss function as follows:

$$L_G = -E_L[\log(1-D_R(I_r,I))] - E_L[\log(D_R(I_f,I))]$$  \hspace{1cm} (8)$$

where $D_R$ is denoted as $D_R(I_r,I) = \sigma(C(I_r) - E_l[C(I_f)])$.

2. Definition of discriminator loss

The discriminator loss is the following symmetrical form of formula (8):

$$L_D = -E_L[\log(D_R(I_f,I))] - E_L[\log(1-D_R(I_f,I))]$$  \hspace{1cm} (9)$$

Algorithm 1 RaGAN-EF Training

1: Suppose the number of iterations of the discriminator is represented by $n_d$;
2: Initialize the parameters of the discriminator $D$ and generator $G$, which are $\theta_d$ and $\theta_g$, respectively;
3: Each iteration
4: ◦ Training process of the discriminator (updated $n_d$ times):
5: Sample $N$ images from the real image, denoted as $\{I_1^r, I_2^r, \ldots, I_N^r\}$;
6: Define $C(I_r) = \frac{1}{N} \sum_{i=1}^{N} C(I_r^i) = E_l[C(I_r)]$ (C() denotes the current discriminator output probability value);
7: Obtain the multi-exposure image sequence corresponding to $N$ image samples, denoted as $\{I_1^{seq}, I_2^{seq}, \ldots, I_N^{seq}\}$;
8: According to the generator $G$, obtain the fusion result image $\{I_1^f, I_2^f, \ldots, I_N^f\}$, where $I_i^f = G_{\theta_g}(I_i^{seq})$;
9: Define $C(I_f) = \frac{1}{N} \sum_{i=1}^{N} C(I_f^i) = E_l[C(I_f)]$;
10: According to the formula (7), use SGD to update the discriminator parameters $\theta_d$: $\nabla_{\theta_d} \frac{1}{N} \sum_{i=1}^{N} \text{Loss}_D$;
11: ◦ Training generator process (update one time):
12: Sampling $N$ images from the real image, denoted as $\{I_1^r, I_2^r, \ldots, I_N^r\}$;
13: Define $C(I_f) = \frac{1}{N} \sum_{i=1}^{N} C(I_f^i) = E_l[C(I_f)]$ (C() denotes the current discriminator output probability value);
14: Obtain the multi-exposure image sequence corresponding to $N$ image samples, denoted as $\{I_1^{seq}, I_2^{seq}, \ldots, I_N^{seq}\}$;
15: According to the generator $G$, obtain the fusion result image $\{I_1^f, I_2^f, \ldots, I_N^f\}$, where $I_i^f = G_{\theta_g}(I_i^{seq})$;
16: Define $C(I_f) = \frac{1}{N} \sum_{i=1}^{N} C(I_f^i) = E_l[C(I_f)]$;
17: Define $\text{Loss}_G = L_V + L_P + L_D + L_C$, according to formulas (3–6), use SGD to update the discriminator parameters $\theta_d$.

It can be seen that the discriminator loss estimates the probability that the given real data is more real than the average fake data. Based on the above analysis, we give the training process of the RaGAN-EF, as shown in Algorithm 1.

The training data uses the ILSVRC 2012 verification set with ImageNet from [18]. The data set has a total of 50,000 images, and each image can be regarded as a natural scene image with better exposure. A random number generation mechanism is used to generate a value ranging between 0.4 and 1.0 Multiplying the original image in this way can change the brightness value of the image and obtain the corresponding low exposure image. In addition, the corresponding high-exposure image must be generated, and the range of the random number is set to 1.2–1.8. In this way, corresponding low-exposure and high-exposure images can be obtained from each original image. The original images with better exposure are used as the ground truth images (as the targets for the GAN framework), and the obtained low-exposure and high-exposure images are used as the input of the network. Image blocks with size $33 \times 33$ were randomly extracted from the original image and the corresponding low/high exposure images, and 744,175 matching pairs were subsequently obtained as training data.

3.3 Obtaining Chrominance Information

At present, there are some exposure fusion methods that treat the R, G, and B channels separately to preserve the chrominance information of the scene. This processing strategy leads to high time complexity, especially when mul-
tiple images are used as input. To solve this problem, we adopt a way to obtain the chrominance information of the scene, as described in [18]. It consists of the following steps:

Step 1: Make the multiple exposure source images equal in size.

Step 2: Chose the median two images to preserve the chrominance information of the real scene. If the number of input images \( K \) is even, the images are denoted as \( \{I_1, I_2, \ldots, I_K\} \), the median two images \( I_{K/2} \) and \( I_{K/2+1} \) are chosen. Otherwise, if \( K \) is odd, the images \( I_{[K/2]} \) and \( I_{[K/2]+1} \) are chosen, where \([K/2]\) denotes the nearest integer greater than \( K/2 \). After the two images are chosen, we apply the average scheme, i.e., the mean value of each corresponding pixel in the two images is kept for final reconstruction. The scheme is applied to the R, G and B channels separately to obtain the chrominance values \( R_s, G_s, \) and \( B_s \) of the scene.

Step 3: Using the \( R_s, G_s, \) and \( B_s \) obtained in step 2, we compute the luminance value \( I_s \) as follows: \( I_s = 0.299 \times R_s + 0.587 \times G_s + 0.114 \times B_s \). Then, \( R_s' = (R_s/I_s)\lambda, G_s' = (G_s/I_s)\lambda, \) and \( B_s' = (B_s/I_s)\lambda \) are calculated, where \( \lambda \) is used to adjust the color saturation of the final image. In this study, we found that for most images, pleasing results can be obtained when \( \lambda = 0.8 \).

Step 4: Using RaGAN-EF, we obtain the fused luminance image \( I_f \).

Step 5: The new luminance image \( I_f \) and the chrominance information \( R_s', G_s', \) and \( B_s' \) are combined to generate the final fused color image. The combination formulas are \( R_{out} = R_s' \times I_f, G_{out} = G_s' \times I_f, \) and \( B_{out} = B_s' \times I_f \).

4. Experiments

To verify the effectiveness of the RaGAN-EF algorithm, we used 24 sets of exposure sequences for testing, which are shown in Fig. 5. The data set was taken from [19]. Currently, there are two ways to evaluate the performance of multi-exposure fusion algorithms: one is subjective evaluation (visual), and the other is objective evaluation (quantitative). In the experiment, we used both methods to evaluate the performance of various algorithms. The experimental software environment was as follows: Python 3.6 and the PyTorch deep learning framework running on a Windows 10 operating system. The hardware environment had the following specifications: Intel Core i7 4.2 GHz with 16 G memory and a GTX 2080ti 12G GPU.

In our paper, four objective criteria were used to quantitatively evaluate the performance of the exposure fusion methods. The first criterion was mutual information (MI), defined as the sum of the mutual information between each input image and the fused image. We calculated this between each input image and fused image, and used the sum of their values as the overall MI. This reflects the total amount of information that the fused image \( F \) retains from the input source images. The second criterion was EG, which measures the amount of edge information transferred from the source images to the fused image. The third criterion was entropy, which measures the overall information in the fused image. The fourth is standard difference (SD) which reflects the degree of dispersion of the distribution of image pixel values. The larger the SD of the image, the greater the contrast of the image, and the better the visibility of the image. Reference [18] provides more details on these criteria.

First, to evaluate the influence of the relative average loss on the fusion result, we compared standard GAN (SGAN), least-squares GAN (LSGAN [25]), Relative GAN (RGAN), Relative average GAN (RaGAN), Relative average LSGAN (RaLSGAN) [6] using four network architectures (VDSR, ResNet [27], DRRN [28], and DenseNet [29]) with the 24 sets of multi-exposure image sequences, more details on these GANs definition and network architectures can be seen in the corresponding paper.

The average results of the EG, MI, entropy, and SD of the five methods are given in Fig. 6. This figure shows that, for the VDSR, the average EG value obtained when using the SGAN loss is 0.564, and the average EG value obtained when using the RaGAN loss is 0.591, which verifies the ability of the RaGAN loss to retain edges. The average MI value when using the SGAN loss is 1.743 (1.855 for LSGAN) and the average MI value when using the RaGAN loss is 1.863 (1.888 for RaLSGAN), which indicates that the Relative GAN loss better retains the original image information. Similarly, the average entropy value when using the SGAN loss is 7.154 (7.217 for LSGAN) and the average entropy value when using the RaGAN loss is 7.226 (7.221 for RaLSGAN), which verifies the ability of the relative GAN loss to retain image information. Similar results are obtained when using other network architecture. In summary, it can be concluded that the algorithm using the relative GAN loss in VDSR obtains a fusion result image that has a much higher quality than that obtained using SGAN. Importantly, using different network structures as generators and obtaining the same conclusions further verify the relative GAN theory has better scalability for exposure fusion.

A set of subjective comparison results are shown in Fig. 7 and Fig. 8. To demonstrate the quality of the results of
the proposed method, we also present enlarged versions of the areas indicated by the red frames in the upper right corners of the results images. The result images were provided by [19], which states that the fusion results of the compared multi-exposure fusion algorithms are partly obtained from the original author and partly generated by publicly available code using the default parameters. Figure 7 (a) and Fig. 8 (a) present the source image sequence, and Fig. 7 (b)–6 (h) and Fig. 8 (b)–7 (h) were obtained by the algorithms in [8], [14], [20]–[23], and [24]. In addition, Fig. 7 (i) and 8 (i) were obtained by the proposed RaGAN-EF. It can be seen that the images in Fig. 7 (b), Fig. 7 (e), Fig. 7 (f), and Fig. 7 (g) have only had their detail enhanced, and the overall appearance of the source sequence has not been preserved. As a result, the fused images look very unnatural. In contrast, the images in Fig. 7 (c) and Fig. 7 (d) better retain the overall brightness of the input image sequence. Although the image in Fig. 7 (h) has vivid restored color, it looks more natural and warm. However, the details in the cave are not clear. The result in Fig. 7 (i) obtained by RaGAN-EF retains more details in the cave, such as the texture of the stone wall, which is clearly visible, and the image looks natural and visually attractive. Similarly, in Fig. 8, it can be seen that the images in Fig. 8 (b) and Fig. 8 (c) have severe halo artifacts near the edges as well as dramatic

brightness changes between the sun and cloud regions. The images in Fig. 8 (d), Fig. 8 (f), and Fig. 8 (g) display an obvious brightness reversal phenomenon. In the original image sequence, the brightness in the distance is brighter than the nearby shore. However, in these images, the brightness of the nearby shore is brighter than that of the distant clouds, making the image look unnatural. The image in Fig. 8 (e) looks whitish and loses the chroma information of the scene. Both the images in Fig. 8 (h) and Fig. 8 (i) retain the overall contrast of the input image sequence. However, the foreground area of the image in Fig. 8 (i) looks clearer.

To further verify the performance of RaGAN-EF, we also use an criterion called MEF-SSIM[16], which has a high correlation with the human perception of image quality. It mainly measures the fusion effect from the perspective of local structure similarity. Higher values of MEF-SSIM indicate that the local structure information extracted by the fusion image from the source image is more accurate.

In Table 1, the results of RaGAN-EF are compared with 9 existing algorithms using 24 sets of images: local average (LE) and global average (GE), Mertens09 (M09) [8], Raman09 (R09) [21], Shen11 (S11) [10], Shutao12 (S12) [20], Zhang12 (Z12) [26], Bruce14 (B14) [23], and Ma15 (M15) [24]. The MEF-SSIM results of the compared 9 algorithms are taken from [19]. As the results in Table 1
Fig. 7  Comparison of various fusion algorithms. (a) Source image sequence; (b) Shen [14]; (c) Li [20]; (d) Mertens [8]; (e) Raman [21]; (f) Li Shuto [22]; (g) Bruce [23]; (h) Ma [24]; (i) RaGAN-EF.
Fig. 8 Comparison of various fusion algorithms. (a) Source image sequence; (b) Shen [14]; (c) Li [20]; (d) Mertens [8]; (e) Raman [21]; (f) Li Shuto [22]; (g) Bruce [23]; (h) Ma [24]; (i) RaGAN-EF.
Table 1 MEF-SSIM comparison results with various multi-exposure image fusion algorithms.

| Method  | LE     | GE     | M09    | R09    | S11    |
|---------|--------|--------|--------|--------|--------|
| Average | 0.9094 | 0.8467 | 0.9449 | 0.7726 | 0.9212 |
| Method  | S12    | Z12    | B14    | M15    | RaGAN-EF |
| Average | 0.9459 | 0.9000 | 0.8513 | 0.9579 | 0.9472 |

show, the average MEF-SSIM value calculated by RaGAN-EF is 0.9472, moreover, the average results of LE, GE, M09, R09, S11, S12, Z12, B14, and M15 are 0.9094, 0.8467, 0.9381, 0.7960, 0.9212, 0.9459, 0.9000, 0.8513, 0.9533, and 0.9579, respectively. The result of RaGAN-EF is slightly lower than that of M15, but is higher than the results of the other eight algorithms. This shows that the multi-exposure image fusion algorithm based on the RaGAN can achieve a performance comparable to some typical multi-exposure fusion algorithms. Moreover, the RaGAN-EF framework has good scalability, and the study of a more effective network architecture (or loss function) could further improve performance.

Last, to validate the performance of the proposed method with different numbers of inputs, we present fusion images with five and six exposure sequences as input respectively in Fig. 9 (a) and 9 (b). Because the multi-exposure image sequence is fed to the generator network after going through the inter-layer shared weight convolution layer, good fusion results are obtained, as these results show in the Fig. 9.

5. Conclusion

This paper proposed an end-to-end multi-exposure image fusion method that uses a RaGAN to fuse multi-exposure images through the generation network. Because the relative discriminator uses the missing attributes required in SGAN, it can make a more reasonable prediction. Hence, it was adopted to fuse the multiple exposure images in the GAN fusion framework. The results of subjective and objective experimental comparative analyses on 24 public multi-exposure image sequences, lead us to conclude that the proposed RaGAN-EF can obtain more stable results and is better than some classic multi-exposure image fusion algorithms. Because the research space of GANs is reasonably large, it will be necessary to explore better loss functions in the multi-exposure GAN image fusion framework, and this will be the focus of our future research.

Acknowledgements

This work was supported by National Nature Science Foundation of China (No.61572077, No.61872042, No.61871039), the projects (KM202111417009, KZ201911417048) supported by the Education Commission of Beijing Municipal, the Collaborative Innovation Center for Visual Intelligence (Grant No. CYXC2011), the Academic Research Projects of Beijing Union University (No. ZB10202003), Premium Funding Project for Academic Human Resources Development in Beijing Union University (BPHR2020AZ01), and the Academic Research Projects of Beijing Union University (No.ZK50202001).

References

[1] N.K. Kalantari and R. Ramamoorthi, “Deep high dynamic range imaging of dynamic scenes,” ACM Transactions on Graphics, vol.36, no.4, pp.1–12, 2017.
[2] Y. Endo, Y. Kanamori, and J. Mitani, “Deep Reverse Tone Mapping,” ACM Transactions on Graphics, vol.36, no.6, Article 177, 2017.
[3] S. Lee, G.H. An, and S.-J. Kang, “Deep recursive hdri: Inverse tone mapping using generative adversarial networks,” Proceedings of the European Conference on Computer Vision (ECCV), vol.11206, pp.613–628, 2018.
[4] J. Kim, J.K. Lee, and K.M. Lee, “Accurate image super-resolution using very deep convolutional networks,” CVPR, pp.1646–1654, 2016.
[5] I.J. Goodfellow, et al., “Generative adversarial networks,” Proceedings of the International Conference on Neural Information Processing Systems, pp.2672–2680, 2014.
[6] A. Jolicoeur-Martineau, “The relativistic discriminator: a key element missing from standard GAN,” arXiv preprint arXiv: 1807.00734, 2018.
[7] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C.C. Loy, “Esrgan: Enhanced super-resolution generative adversarial networks,” ECCV Workshops, vol.11133, pp.63–79, 2018.
[8] T. Mertens, J. Kautz, and F. Van Reeth, “Exposure fusion: A simple and practical alternative to high dynamic range photography,” Comput. Graph. Forum, vol.28, no.1, pp.161–171, 2009.
[9] A.A. Goshishty, “Fusion of multi-exposure images,” Image and Vision Computing, vol.23, no.6, pp.611–618, 2005.
[10] R. Shen, I. Cheng, J. Shi, and A. Basu, “Generalized random walks for fusion of multi-exposure images,” IEEE Trans. Image Process., vol.20, no.12, pp.3634–3646, 2011.
[11] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, “Gradient field multi-exposure images fusion for high dynamic range image visualization,” Visual Communication and Image Representation, vol.23, no.4, pp.604–610, 2012.
[12] M. Song, D. Tao, C. Chen, J. Bu, J. Luo, and C. Zhang, “Probabilistic exposure fusion,” IEEE Transactions on Image Processing, vol.21, no.1, pp.341–357, 2011.
[13] J. Hu, O. Gallo, and K. Pulli, “Exposure stacks of live scenes with hand-held cameras,” ECCV, vol.7572, pp.499–512, 2012.
[14] J. Shen, Y. Zhao, S. Yan, and X. Li, “Exposure fusion using boosting Laplacian pyramid,” IEEE Trans. Cybern., vol.44, no.9, pp.1579–1590, 2014.
[15] K.R. Prabhakar, V.S. Srikan, and R.V. Babu, “DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs,” Proc. IEEE Int. Conf. Comput. Vis. (ICCV), pp.4724–4732, Oct. 2017.
[16] M. Kede, Z. Kai Zeng, and W. Zhou, “Perceptual Quality Assessment for Multi-Exposure Image Fusion,” IEEE Trans. Image Processing (TIP), vol.24, no.11, pp.3345–3356, 2015.
[17] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” CVPR, 2014.
[18] J. Wang, H. Liu, and N. He, “Exposure fusion based on sparse representation using approximate K-SVD,” Neurocomputing, vol.135, pp.145–154, 2014.
[19] K. Ma, Z. Duangmu, H. Yeganeh, and Z. Wang, “Multi-exposure image fusion by optimizing a structural similarity index,” IEEE Transactions on Computational Imaging, vol.4, no.1, pp.60–72, 2017.
[20] Z.G. Li, J.H. Zheng, and S. Rahardja, “Detail-enhanced exposure fusion,” IEEE Trans. Image Process., vol.21, no.11, pp.4672–4676, 2012.
[21] S. Raman and S. Chaudhuri, “Bilateral filter based compositing for variable exposure photography,” Eurographics (short papers), pp.1–4, 2009.
[22] S. Li, X. Kang, and J. Hu, “Image Fusion With Guided Filtering,” IEEE Transactions on Image Processing, vol.22, no.7, pp.2864–2875, 2013.
[23] N.D.B. Bruce, “Expoblend: Information preserving exposure blending based on normalized log-domain entropy,” Comput. & Graph., vol.39, pp.12–23, 2014.
[24] K. Ma and Z. Wang, “Multi-exposure image fusion: A patchwise approach,” IEEE International Conference on Image Processing (ICIP), pp.1717–1721, 2015.
[25] X. Mao, Q. Li, H. Xie, R.Y.K. Lau, Z. Wang, and S.P. Smolley, “Least squares generative adversarial networks,” ICCV, pp.2813–2821, 2017.
[26] W. Zhang and W.-K. Cham, “Gradient-directed multi-exposure composition,” IEEE Transactions on Image Processing, vol.21, no.4, pp.2318–2323, 2011.
[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” CVPR, pp.770–778, 2016.
[28] Y. Tai, J. Yang, and X. Liu, “Image super-resolution via deep recursive residual network,” CVPR, pp.2790–2798, 2017.
[29] G. Huang, Z. Liu, L. Van Der Maaten, and K.Q. Weinberger, “Densely connected convolutional networks,” CVPR, pp.2261–2269, 2017.

Jinhua Wang received the Ph.D. degree in applied computer applied technology from Beijing Jiaotong University, Beijing, China, in 2010. She is currently an associate professor with Smart City College, Beijing Union University, Beijing, China. Her research interests include computer vision and machine learning.

Xuewei Li received the B.E. degree from the Heilongjiang Institute of Business, Heilongjiang, China, in 1982, the M.A.Sc. degree from Southwest Jiaotong University, Sichuan, China, in 1987, and the Ph.D. degree from Central South University (Changsha Railway College), Hunan, China. He is currently a Professor with Beijing Jiaotong University. He is also the President of Beijing Union University, the Director of the President’s Union of China and Russia Transportation Universities, China, and the President of the Eurasian Transport University Federation. He is engaged in scientific research and technological innovation in the field of transportation, such as rail transit and high-speed rail, and familiar with the characteristics of disciplines and talent training system of engineering, management, economics, and science. His research interests include theoretical frontier of management science, big data application and model method, modern enterprise management theory and practice, high-speed rail operation safety theory, and risk control. He is a member of the 863 General Expert Group of Ministry of Science and Technology. He is the Vice President of Transportation Branch of the China Society of Systems Engineering.

Hongzhe Liu received the B.E. degree from Chinese Marine University, Shandong, China, in 1995, the M.A.Sc. degree from California State University, Long Beach, CA, USA, in 2000, and the Ph.D. degree from Beijing Jiaotong University, Beijing, China. She is currently the Leader of the Beijing Key Laboratory of Information Service Engineering and a Professor with Beijing Union University. Her research interests include artificial intelligence, visual intelligence, cognitive computing, and visual computing.