MEF-GAN: Multi-Exposure Image Fusion via Generative Adversarial Networks

Han Xu, Jiayi Ma, Member, IEEE, and Xiao-Ping Zhang, Senior Member, IEEE

Abstract—In this paper, we present an end-to-end architecture for multi-exposure image fusion based on generative adversarial networks, termed as MEF-GAN. In our architecture, a generator network and a discriminator network are trained simultaneously to form an adversarial relationship. The generator is trained to generate a real-like fused image based on the given source images which is expected to fool the discriminator. Correspondingly, the discriminator is trained to distinguish the generated fused images from the ground truth. The adversarial relationship makes the fused image not limited to the restriction of the content loss. Therefore, the fused images are closer to the ground truth in terms of probability distribution, which can compensate for the insufficiency of single content loss. Moreover, aiming at the problem that the luminance of multi-exposure images varies greatly with spatial location, the self-attention mechanism is employed in our architecture to allow for attention-driven and long-range dependency. Thus, local distortion, confusing results, or inappropriate representation can be corrected in the fused image. Qualitative and quantitative experiments are performed on publicly available datasets, where the results demonstrate that MEF-GAN outperforms the state-of-the-art, in terms of both visual effect and objective evaluation metrics. Our code is publicly available at https://github.com/jiayi-ma/MEF-GAN.

Index Terms—Image fusion, multi-exposure, generative adversarial network, self-attention.

I. INTRODUCTION

In the natural scene, the luminance difference of objects in the same scene can be considerable, presented as wide dynamic range. However, due to the limitations of digital cameras, the image produced by the sensors can only capture the details within very limited range. To solve this problem, the most common way is to take a series of photographs under different exposure settings of low dynamic range (LDR) and merge them into a well-exposed image of high dynamic range (HDR) for better visual perception. In LDR images, the regions that can be clearly represented may vary with different lighting conditions. While in the HDR image, the detailed information lost in under-exposed and over-exposed regions will be restored and enhanced. Thus, the fused image will contain better illumination for all regions, as shown in Fig. 1. Due to the greater dynamic range and perceptually appealing details, multi-exposure image fusion (MEF) have wide ranging applications in the fields of remote sensing, medical imaging, HDTV and so on [1], [2].

The existing MEF methods can be divided into four categories according to the theory: i) pixel-based methods, including methods based on weight and gradient, and multi-scale transform (e.g., pyramid, wavelet and contourlet, etc.); ii) methods based on sparse representation; iii) tone mapping-based methods; and iv) deep learning-based methods.

These above-mentioned methods work well by manually designing complex feature extraction methods or fusion rules. However, it is a time-consuming, laborious and difficult problem to consider the extracted features and fusion rules thoroughly. In existing deep learning-based methods, the application of convolutional neural networks (CNNs) are limited. The shortcomings of existing methods will be described concretely later in Sec. II-A. In this paper, based on deep learning, we propose a novel end-to-end model for MEF, named as multi-exposure image fusion via generative adversarial networks (MEF-GAN). In our MEF-GAN, given the over-exposed and under-exposed images, the fused image can be generated by a generator network, and a discriminator network is applied to assess whether the fused images are indistinguishable from the ground truth. In this adversarial framework,
more information can be retained to meet this high-level goal. The generator network consists of three blocks: a self-attention block to allow for attention-driven and long-range dependency by introducing the self-attention mechanism, a local detail block to preserve and fuse local details, and a merge block to merge features obtained by the above-mentioned two blocks. As shown in Fig. 1, the result of our MEF-GAN contains better illumination for all regions and thus, represents the details of source images more completely. Moreover, compared with the result of Deepfuse [3], our result is enriched with colors and is able to represent clearer texture details with higher intensity and saturation.

The contributions of our work include the following three aspects:

- To the best of our knowledge, it is the first time that the generative adversarial networks (GANs) are employed to fuse multi-exposure images.
- The proposed MEF-GAN is an end-to-end deep learning-based method, which releases the focus of research from the feature extraction and fusion rule design.
- We also introduce the self-attention mechanism to the MEF problem for better fusion performance.

The remainder of this paper is arranged as follows. Section II gives a brief introduction of existing MEF methods, GANs and the variants, as well as their applications. In Section III, we describe our MEF-GAN with the loss functions, architectures and other implementations in detail. Qualitative and quantitative comparisons and ablation study are performed in Section IV. Conclusion is presented in Section V.

II. RELATED WORK

This section provides a brief review of existing multi-exposure image fusion methods and the basic theory behind GANs. As an extension, we give a brief introduction of the variants of GANs and some typical applications of them in the image processing community.

A. Existing Multi-Exposure Image Fusion Methods

The existing MEF methods can be divided into four categories according to the theory:

i) Pixel-based methods. This kind of methods contain two categories. One is weight and gradient-based methods. The key of weight-based methods is the calculation of the weight for pixel-wise fusion [4], [5]. Gradient-based methods are devoted to retaining the maximum gradients in source images [6], [7]. The other one is multi-scale transform-based methods, including pyramid [8], wavelet [9], contourlet, shearlet, principal component analysis [10], dense scale invariant feature transform [11], etc. These methods mainly contain three stages: image transform, fusion of coefficients and inverse transform. These methods can obtain better visual perception than methods in the first category. As a representative, GFF [12] decomposes the image into a base layer and a detail layer. A novel guided filtering-based weighted average technique is proposed to make full use of spatial consistency for fusion of these layers. Besides, based on Gaussian pyramids transform, FLER [13] firstly synthesizes a virtual image with a medium exposure to brighten the under-exposed regions and darken the over-exposed regions. Then the virtual image and two source images are fused using Gaussian pyramid. However, a common limitation of these methods is that they cannot well represent the curves and edges of images [14]. Besides, the results of pixel-based methods may suffer from dark regions because they are unable to adjust the fused results according to the information in other regions of the source images;

ii) Methods based on sparse representation. Sparse representation is a novel representation theory which is widely applied due to the advantage of sharp components and textures without artifacts [15], [16]. First, according to the same over-complete dictionary, the source images can be represented by the corresponding sparse coefficients. Then, the fusion process is performed on the coefficients. Finally, the fused image can be obtained through the coefficients and the dictionary. However, how to obtain the coefficients and how to construct the dictionary are two general problems of these sparse representation-based methods [14].

iii) Tone mapping-based methods. By compressing high dynamic range through HDR reconstruction techniques over several LDR images, HDR scenes are able to be displayed on common devices. In the past few years, many different mapping methods have been proposed [17], [18].

iv) Deep learning-based methods. For the first time, deep learning is introduced into the field of multi-exposure image fusion by Deepfuse [3]. Deepfuse employs the metric MEF-SSIM [19] as the loss function and builds a novel CNN architecture to realize unsupervised learning. The definition of MEF-SSIM is based on SSIM. In MEF-SSIM, source images are transformed into a desired result. Then, MEF-SSIM is calculated by measuring the similarity between the desired result and the real result. To obtain the desired result, patches in source images are modelled into three components as in the SSIM framework: contrast, structure and luminance. Since the luminance comparison in the local patches is insignificant [3], the luminance component is discarded. Only the structures and contrasts of source images are obtained and preserved in the desired result by using the weight sum of structures of input patches and the highest contrast value of contrasts of input patches. Thus, simply depending on MEF-SSIM to guide the generation of fused images leads to the loss of other key information, as it only focuses on the structure and contrast distortion. Besides, in Deepfuse, the CNN is only applied in the Y channel for feature extraction and reconstruction while the fusion rules of the chrominance channels are still designed manually. However, the manually designed way may still fail to fully retain the information in the chrominance channels.

B. Generative Adversarial Networks

GAN was initially proposed for estimating generative models [20]. In GAN framework, a generative network $G$ and a discriminator network $D$ are trained simultaneously and form an adversarial process. Given the noise variable $z$ sampled from a latent space, the generative network $G$ is supposed to generate a sample $x = G(z)$. The purpose of training $G$ is to
learn a probability distribution $P_G(x)$ as an estimation of the real distribution $P_{\text{data}}(x)$ from real samples $\{x^1, x^2, \cdots, x^m\}$. Then the discriminator $D$ is applied to determine whether a sample is from $P_G(x)$ or $P_{\text{data}}(x)$. Through the constant adversarial process, the samples generated by $G$ will gradually approximate the real samples.

In traditional methods, the generative models normally consist of known distributions, such as Gaussian mixture models. Then, the maximum likelihood estimation can be performed conveniently. However, these predefined distributions limit the fitting ability of the generative models. If $P_{\text{data}}(x)$ is much more complicated, the likelihood function will be more difficult to calculate and the estimation will be more challenging to perform thereupon. As the promise of the deep learning is to discover rich and hierarchical models, a deep network is more applicable to capture the data distribution. Thus, in the framework of GAN, the generator network $G$ is employed to generate new samples. Since the objective of $G$ is to enable $P_G(x)$ to be as close as possible to $P_{\text{data}}(x)$, the optimization formulation is defined as follows:

$$G^* = \arg \min_{G} \text{Div}(P_G(x), P_{\text{data}}(x)), \quad (1)$$

where Eq. (1) denotes that the divergence between $P_G$ and $P_{\text{data}}$ is supposed to be as small as possible.

Due to the fact that $P_{\text{data}}(x)$ is the distribution to be solved and $P_G(x)$ is determined by the network $G$, the divergence $\text{Div}(P_G(x), P_{\text{data}}(x))$ cannot be expressed and solved concretely. In this condition, a network $D$ is applied to estimate the probability that a sample is from $P_{\text{data}}$ rather than $P_G$, which is denoted as $D(x)$. Thus, the Jensen-Shannon (JS) divergence of $P_G$ and $P_{\text{data}}$ can be expressed as:

$$E_{x \sim P_{\text{data}}} [\log D(x)] + E_{x \sim P_G} [\log (1 - D(x))]. \quad (2)$$

Defining the divergence in Eq. (2) with $V(G, D)$, the objective function of $D$ can be expressed as:

$$D^* = \arg \max_{D} V(G, D). \quad (3)$$

A larger $V(G, D)$ means a greater divergence and it is easier to be discriminated. Thus, Eq. (1) can be converted to:

$$G^* = \arg \min_{G} \max_{D} V(G, D). \quad (4)$$

When we are training the generator $G$, the discriminator $D$ is fixed and similarly, $G$ is fixed when we are training $D$. Then, $G$ and $D$ form an adversarial relationship and make up the two-player min-max game where $G$ is trained to fool $D$ while $D$ tries to distinguish the generated samples. With the advance of the adversarial process, the data generated by $G$ will be more and more indistinguishable from the real data.

**C. Variants of GAN and Their Applications**

The adversarial loss which forces $P_G(x)$ to be as close as $P_{\text{data}}(x)$ enables GAN to become more suitable for image generation tasks. Inspired by GAN, Ledig et al. [21] introduced a super-resolution GAN (SRGAN) for image super-resolution, which augments the content loss function with an adversarial loss by training a GAN. To improve the quality of generated images and improve the stability of training process, the least squares GAN (LSGAN) [22] adopts the least squares loss function for the discriminator. Accordingly, Ma et al. [23] proposed a novel LSGAN-based method named FusionGAN, which is an end-to-end model for dealing with infrared and visible image fusion tasks. By feeding some extra information as additional input layers to both the generator and the discriminator, GANs are extended to conditional GAN (cGAN) [24]. Based on cGAN, by introducing the VGG feature and L1-regularized gradient prior to modify it, [25] can restore a clear image from a hazy image. Instead of using the JS divergence to compare $P_G(x)$ and $P_{\text{data}}(x)$, wasserstein GAN (WGAN) [26] introduces the Wasserstein distance to improve the stability and get rid of problems like mode collapse. By applying WGAN and perceptual similarity, Yang et al. [27] proposed a contemporary deep neural network for LDCT image denoising. By applying adversarial losses for distribution matching and cycle consistency losses to prevent the learned mappings, Zhu et al. [28] proposed cycle-consistent GAN (CycleGAN) to learn the translation from a source domain to a target domain in the absence of paired examples. To solve the problem that traditional GANs generate high-resolution details as a function of only spatially local points, Zhang et al. [29] proposed the self-attention GAN which allows attention-driven, long-range dependency modeling. Thus, details can be generated using cues from all feature locations. Furthermore, to generate high-resolution and diverse samples from complex datasets, Brock et al. trained GANs at the largest scale and studied the instabilities specific to such scale [30].

In image fusion, GANs have been applied successfully to multi-modal image fusion such as visible and infrared image fusion [23]. To the best of our knowledge, it is the first time that GANs are applied to solve the problem of multi-exposure image fusion in our work. Furthermore, in existing GAN-based image fusion methods, the techniques used are traditional GAN, cGAN, LSGAN and WGAN. It is the first time that the self-attention module is applied to solve the image fusion problem, especially multi-exposure image fusion because it allows for attention-driven and long-range dependency which are suitable to solve the regional over- or under-exposure issues. Generally speaking, GANs are crucial to many different state-of-the-art image generation and manipulation systems. In the future, GANs have the potential to enable many other applications [31]–[33].

**III. PROPOSED METHOD**

In this section, we provide the design and the formulation of losses, including the adversarial loss and the content loss of the generator, and the adversarial loss of the discriminator. Then, we give a description of the network architecture in detail. At the end, implementations are provided, including the publicly available dataset used for training and testing, some training details, and the settings of hyperparameters.

**A. Loss Function**

The generator $G$ is trained to learn a mapping from the source images, i.e., the under-exposed image $I^u$ and
the over-exposed image \( I^o \) to the well-exposed image \( I^f \). To capture the true probability distribution of the ground truth \( I^{gt} \), the adversarial loss of \( G \) is defined as follows:

\[
\mathcal{L}_{Adv} = \mathbb{E} \left[ \log \left( 1 - D \left( I^f \right) \right) \right].
\]  

(5)

By trying to fool the simultaneously trained discriminator \( D \) which differentiates between \( I^f \) and \( I^{gt} \), \( \mathcal{L}_{Adv} \) forces the fused images generated by \( G \) to reside on the manifold of the ground truth.

In previous research, the success and applications of GAN were limited due to the unstable behavior, the artifacts and the error or the noisy results [34]. Prior work of image generation controls the trade-off between \( \mathcal{L}_{Adv} \) and \( \mathcal{L}_{Con} \), e.g., \( L_1 \), \( L_2 \) norms or other losses, can benefit and boost the generator performance. Therefore, in our work, in additional to \( \mathcal{L}_{Adv} \), we also augment the generator loss function with the content loss \( \mathcal{L}_{Con} \). So the generator loss function \( \mathcal{L}_G \) can be defined as:

\[
\mathcal{L}_G = \mathcal{L}_{Adv} + \lambda \mathcal{L}_{Con},
\]  

(6)

where \( \lambda \) controls the trade-off between \( \mathcal{L}_{Adv} \) and \( \mathcal{L}_{Con} \).

To measure and constrain the similarity between \( I^f \) and \( I^{gt} \) in content in pixel-wise, we include the mean square error (MSE) loss as the content loss, which is expressed as follows:

\[
\mathcal{L}_{MSE} = \mathbb{E} \left[ \frac{1}{WHC} \sum_{i=1}^{W} \sum_{j=1}^{H} \sum_{k=1}^{C} \left( I_{i,j,k}^{gt} - I_{i,j,k}^f \right)^2 \right],
\]  

(7)

where \( W \) and \( H \) denote the width and height of the image, respectively, and \( C \) is the channel number of the image.

It is generally known that the MSE-based optimization may result in the lack of high-frequency content in the solution. The solutions are perceptually unsatisfying due to the smooth textures. In order to compensate for the deficiency of the MSE loss and retain the content with high frequency, which is generally expressed as the gradient information, we augment \( \mathcal{L}_{MSE} \) with a gradient-based loss \( \mathcal{L}_{grad} \), which is defined as:

\[
\mathcal{L}_{grad} = \mathbb{E} \left[ \frac{1}{WH} \sum_{i=1}^{W} \sum_{j=1}^{H} \left[ \nabla \Gamma \left( I^{gt} \right)_{i,j} - \nabla \Gamma \left( I^f \right)_{i,j} \right] \right],
\]  

(8)

where \( \nabla \) is the gradient operation and the function \( \Gamma (\cdot) \) denotes the operation of transforming image in RGB channels to gray channel. Thus, by introducing a weight \( \xi \) to control the trade-off, the content loss can finally be written as:

\[
\mathcal{L}_{Con} = \mathcal{L}_{MSE} + \xi \mathcal{L}_{grad}.
\]  

(9)

The discriminator \( D \) in MEF-GAN is trained to discriminate between the generated image \( I^f \) and the ground truth \( I^{gt} \). The output of \( D \) is a scalar between 0 and 1, representing the probability that the input image is from the ground truth. We express the adversarial loss function of \( D \) as:

\[
\mathcal{L}_D = \mathbb{E} \left[ -\log D \left( I^{gt} \right) \right] + \mathbb{E} \left[ -\log \left( 1 - D \left( I^f \right) \right) \right].
\]  

(10)

### B. Network Architecture

1) **Generator Architecture**: The generator network consists of three sub-parts: the self-attention block, the local detail block, and the merge block, as shown in Fig. 2(a). The combination of the multiple networks are employed to capture features with multiple scales [36] and multiple receptive fields.

The self-attention block extracts features by introducing attention-driven and long-range dependency to take into account cues from all other feature locations, rather than the restricted receptive field around. However, because the max-pooling layers in this block will reduce the scale of feature maps and result in distortion, we add the local detail block to retain some details that might be lost. As the layers in this detail block are all with a limited respective field, we name this block local detail block. Then, the merge block is applied to merge features obtained by the self-attention block and the local detail block and generate the final fused image. The following subsections are a detailed introduction of the specific architectures of these blocks.
In testing phase, if tested source images are clipped into many patches and fed into the generator. The height and weight of these patches are much smaller than those of the original over-exposed or under-exposed images. During the training phase, images in the training dataset are clipped into many patches and fed into the generator. This will result in significant differences between patches, both in luminance and chrominance. Then the visual effect of fused images will be seriously affected. Considering the training and testing process comprehensively and trading off between performance and computational cost, the max-pooling layer and a longer stride in the convolutional layer are employed before the self-attention layer. To compensate for the loss of details caused by these operations, we add another block in parallel, which contains only a few simple layers. This block is named as local detail block and will be discussed later.

As shown in Fig. 2(a), the generator contains 5 layers. In the first layer, to accelerate training and avoid exploding/vanishing gradients, batch normalization [40] is added after the convolutional layer. The rectified linear unit (ReLU) [41] is used as the activation function. After the first convolutional layer, the max-pooling layer is applied with the kernel and the stride set to 2. In the next convolutional layer, we perform the spectral normalization on the convolutional kernels. In fact, in this paper, we introduce spectral normalization both in the generator and the discriminator architecture, which is abbreviated as sn. The spectral normalization was originally proposed by Miyato et al. [42] and applied to the discriminator network by restricting the spectral norm of each layer and has the advantage of not requiring extra hyper-parameter tuning.

Then, Zhang et al. [29] found that introducing the spectral normalization to the training of the generator $G$ can prevent the escalation of parameter magnitudes and avoid unusual gradients, which is in favor of the conditioning of $G$ in the performance of GAN [43]. After the first two layers, both the height and width of feature maps are one-eighth of those of the original input. These feature maps are then passed through the self-attention layer. The output feature maps are then upsampled using the nearest neighbor interpolation. Then passed through another two convolutional layers and upsampling layers shown in the figure, the feature maps of this block are finally obtained.

b) Local detail block: The local detail block contains five convolutional layers. All these layers are accompanied by batch normalization and the ReLU activate function. In each layer, regardless of the number of channels for the input of this layer, the layer always generates 40 feature maps by

---

**Fig. 3.** The self-attention mechanism. \([\ldots]\) expresses the shape of the maps of space with $B, H, W, C$ denoting the batch size, height, width, and channels, respectively. We set $c = C/4$. “flatten” represents the operation to change the shape of feature map from $[H, W]$ to $H \times W$. “reshape” represents the opposite operation. $\gamma$ is the scalar parameter learned during training.
3 × 3 filters. All the strides are set to 1. Furthermore, to mitigate the vanish of gradient and reuse previously computed features, densely connected blocks are applied in the local detail block to remedy feature loss, as shown in Fig. 2. In the densely connected block, all previously computed features are directly connected to the subsequent layers for feature reuse. In the second to the fourth layers, the spectral normalization is also applied to prevent the escalation of parameter magnitudes and avoid unusual gradients.

c) Merge block: The features obtained by the self-attention block and the local detail block are concatenated as the input of the merge block. The merge block consists of three convolutional layers. One difference is that the spectral normalization is only adopted in the first two layers but not in the third layer. The other difference is that the first two layers use the ReLU activation function and the last layer replaces it with the tanh activation function to obtain the final fused image.

2) Discriminator Architecture: Since the task of distinguishing between two images is easier than generating images for neural networks, the architecture of $D$ is less complicated than that of $G$. Specifically, there are five layers in the discriminator, as shown in Fig. 2(b). The first four layers are convolutional layers, followed by batch normalization or the ReLU function, or both of them. Among these four layers, the strides of the first three convolutional layers are set to 2 and spectral normalization is applied before these convolutional layers. To the final convolutional layer, the stride is set to 1. The output of the fourth layer is fed into a fully connected layer and activated by the tanh function to generate a scalar which estimates the probability that the input image is the ground truth rather than the generated fused image.

C. Implementations

1) Data: To validate the performance of the proposed MEF-GAN, we perform qualitative and quantitative experiments on the publicly available dataset¹ provided by [44] with multi-exposure sequences from 2 categories (indoor and outdoor) and the corresponding high-quality reference images (ground truth). The dataset contains image sequences from several exposure sequences from 2 categories (indoor and outdoor) and the corresponding high-quality reference images (ground truth). In this dataset, there are a series of photographs under different exposure settings which have been accurately aligned [45], [46]. We select the brightest image and the darkest one in each series as the over-exposed and the under-exposed images, respectively, as the fusion of the brightest and the darkest images is more challenging. 80 groups of over-exposed images, under-exposed images, and corresponding ground truths are selected and cropped into 20000+ patches as the training data. All patches are of size 144 × 144.

2) Training Details: To train the generator network and the discriminator network, during each iteration, some over-exposed patches $\{I^{o(1)}, \ldots, I^{o(B)}\}$, and some corresponding under-exposed patches $\{I^{u(1)}, \ldots, I^{u(B)}\}$ are sampled from the training dataset, where $B$ denotes the batch size. The corresponding ground truth patches are $\{I^{t(1)}, \ldots, I^{t(B)}\}$. Then fused patches can be generated by $G$: $\{f^{(1)}, \ldots, f^{(B)}\} = \{G(I^{o(1)}, I^{u(1)}), \ldots, G(I^{o(B)}, I^{u(B)})\}$.

In order to maintain the stability of the adversarial process and improve the results, the training details of the generator and the discriminator are set in different ways. Parameters in $D$, i.e., $\theta_D$ are updated by SGD Optimizer to minimize $\mathcal{L}_D$ in Eq. (10). Then parameters in $G$, i.e., $\theta_G$ are updated by Adam Optimizer to minimize $\mathcal{L}_G$ in Eq. (6). For $G$ and $D$, we perform different learning rate, as the two-timescale update rule (TTUR) [47] is confirmed to be effective. The learning rate of $G$ is set to 0.0006 with exponential decay and that of $D$ is set to two times that of $G$. The batch size $B$ is set to 19. The epoch is set to 2, $\lambda = 500$ and $\bar{c} = 0.6$. During the testing phase, only the trained generator is used to generate the results. And the test source images do not need to and cannot be cropped into patches to generate fused patches which are stitched together as the fused images (as explained in Section III-B.1-a)). That is, the input is the entire over-exposed and under-exposed images rather than image patches.

IV. EXPERIMENTAL RESULTS

To evaluate MEF-GAN, we firstly compare it with several state-of-the-art methods qualitatively. Then, with assistance, qualitative experiments are performed. For the sake of comprehensiveness, three types of metrics are employed for evaluation. Finally, comparative experiments are performed to validate the effectiveness of each block in the generator and the gradient loss. For generalization, the experiments about MEF-GAN under other conditions are also performed.

A. Comparison Methods

We compare MEF-GAN with five state-of-the-art methods, including GFF [12], DSIFT [11], the gradient-based method (GBM) [7], Deepfuse [3] and the method for fusing large-exposure-ratio images (FLER) [13]. GFF is a novel guided filtering-based method for creating a highly informative fused image. The code is at [48]. DSIFT is based on dense scale invariant feature transform and this is the first time the descriptor has been adopted as the activity level measurement. Its code is downloaded from [49]. In GBM, two different fusion strategies are applied for chrominance and luminance channels separately. The code is at [50]. Deepfuse is the landmark multi-exposure fusion method based on deep learning. The code is available at [51]. FLER is designed specifically for large-exposure-ratio image fusion and the code is available at [52].

B. Results and Analysis

1) Qualitative Comparisons: We perform qualitative comparisons on five typical image sequences and the results are shown in Figs. 4, 5, 6, 7 and 8. We analyze the results from two aspects, i.e., the whole image and details. From the perspective of the whole image, features lost due to inappropriate exposure settings are restored and enhanced in our results without local dark regions. Besides, our results are enriched with
XU et al.: MEF-GAN: MEF-GAN

Fig. 4. Qualitative comparison results on image sequence 1. Images in the leftmost column are source images and the corresponding ground truth. Fused results of five comparison methods (GFF [12], DSIFT [11], GBM [7], Deepfuse [3] and FLER [13]) and the fused result of our proposed MEF-GAN are shown on the bottom right corner. The bottom row are highlighted regions of different results corresponding to the red and green boxes of the ground truth image.

Fig. 5. Qualitative comparison results on image sequence 2.

abundant and bright colors, which are evident in Figs. 4, 5, 7 and 8. From the perspective of details, more texture details are preserved in our results, as shown in Figs. 4, 5, 6 and 7. In some local regions, our result can revise the imprecise or inappropriate representation in the ground truth, as shown in the highlighted parts in Figs. 4 and 5.

The ground truth is generated by thirteen representative algorithms and subjectively selecting the best quality one from them as the ground truth rather than taken directly by an optical lens. Thus, there may be imprecise or inappropriate representation in them.

Next, specific analysis and comparisons are presented. As can be seen from Figs. 4, 5, 6 and 7, the results of GFF and DSIFT suffer from obvious dark regions, especially in the regions with the strongest brightness in the over-exposed images. Furthermore, the tones of the results of GBM, Deepfuse and FLER are more grayish due to lower intensity and saturation compared with our results.

As can be seen in Fig. 4, the light of the pendant lamp and the colors of the wall in our result are of the highest intensity and saturation. The same phenomenon can be seen in the colors of the sky and the house in Fig. 5 and the colors of...
the highlighted regions shown in Fig. 8. Moreover, compared with the ground truths in Figs. 4 and 5, our results can locate the specific location of the bulb and the cloud respectively.

2) Quantitative Comparisons: In the multi-exposure image fusion community, MEF-SSIM [19] is a commonly used metric for quantitative evaluation. We apply this metric to evaluate the ground truth and fused results of five comparisons methods. The MEF-SSIM values are shown in Table I. As can be seen in the qualitative comparison results in Figs. 4, 5, 6, 7 and 8, the ground truths show the best qualitative effect in line with human visual perception compared with other fused results. However, as for the corresponding MEF-SSIM values in Table I, the results of this metric are not consistent with the visual effect. More concretely, the ground truths of best visual effect cannot reach the highest MEF-SSIM values or even have the lowest values. Even if the fused results of GFF and DSIFT suffer from obvious and large area of dark regions, their MEF-SSIM values are still higher than those of fused results obtained by GBM, Deepfuse and FLER. Although the results of other comparison methods suffer from low intensity and low saturation tones, their MEF-SSIM values are generally higher than those of the ground truths.

This phenomenon is also a reflection of a common problem not only in the multi-exposure image fusion community, but also in the field of multimodal image fusion. It is difficult to achieve unification between subjective visual effects and
objective evaluation results. The phenomenon that subjective effects are satisfactory but objective evaluation results are poor or the contrary phenomenon occurs from time to time. One reason is that the objective evaluation system is far from perfect. Metrics can only measure a certain aspect of characteristics or information, rather than perform a comprehensive analysis from a holistic point of view. In addition, results of objective evaluations are extremely sensitive and changeable and remained for future work. The other reason is that subjective evaluations are extremely sensitive and changeable and difficult to be expressed through metrics completely. Therefore, scholars generally adopt the evaluation method based on qualitative evaluations and supplemented by quantitative evaluations.

In this paper, in order to quantitatively evaluate fused results comprehensively and avoid one-sidedness, we measure the following three types of metrics (The fused image is denoted as $f$ of size $M \times N$ and the ground truth is denoted as $g$):

a) Metrics reflecting the similarity between the ground truth and the fused image: As ground truths are available in the dataset, we evaluate the similarity between the fused image and the corresponding ground truth by the following metrics. Besides, as these metrics can merely measured on one channel, in order to avoid the loss of chrominance information caused by converting the image to a gray-scale image, we perform these metrics on the intensity (I) and saturation (S) channels respectively. The I channel and the S channel can be transformed by the original RGB channels according to the following formulations:

\[
I = \frac{1}{\sqrt{3}} \times R + \frac{1}{\sqrt{3}} \times G + \frac{1}{\sqrt{3}} \times B, \tag{11}
\]

\[
V_1 = \frac{1}{\sqrt{6}} \times R + \frac{1}{\sqrt{6}} \times G - \frac{2}{\sqrt{6}} \times B, \tag{12}
\]

\[
V_2 = \frac{1}{\sqrt{2}} \times R - \frac{1}{\sqrt{2}} \times G, \tag{13}
\]

\[
S = \sqrt{V_1^2 + V_2^2}, \tag{14}
\]

where $V_1$ and $V_2$ are intermediate variables.

- Peak signal-to-noise ratio (PSNR): PSNR is a metric reflecting the distortion by the ratio of peak value power and noise power:

\[
\text{PSNR} = 10 \log_{10} \frac{r^2}{\text{MSE}}, \tag{15}
\]

where $r$ is the peak value of the fused image and is set to 256 in this paper. MSE is the mean square error that measures the dissimilarity between the ground truth and the fused image and is defined as follows: where $\text{MSE} = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (g_{i,j} - f_{i,j})^2$. A larger PSNR indicates the less distortion the process produces and the fused image is more similar to the ground truth.

- Correlation coefficient (CC): CC measures the degree of linear correlation between the ground truth and the fused image. It is mathematically defined as follows:

\[
\text{CC} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (g_{i,j} - \mu_g)(f_{i,j} - \mu_f)}{\sqrt{\sum_{i=1}^{M} \sum_{j=1}^{N} (g_{i,j} - \mu_g)^2 \sum_{i=1}^{M} \sum_{j=1}^{N} (f_{i,j} - \mu_f)^2}}, \tag{16}
\]
where $\mu_g$ and $\mu_f$ denote the mean values of $g$ and $f$. A large $CC$ indicates that there is a strong linear correlation between the fused image and the ground truth.

b) Metric reflecting the similarity between the source images and the fused image:

- Mean structural similarity (MSSIM): MSSIM is calculated as the average of the individual SSIM values for each sliding window. SSIM is a metric used to model image loss and distortion. It is defined as follows:

$$SSIM_{X,F} = \frac{2\mu_X\mu_F + C_1}{\mu_X^2 + \mu_F^2 + C_1} \frac{2\sigma_X\sigma_F + C_2}{\sigma_X^2 + \sigma_F^2 + C_2} \frac{\mu_X \mu_F + C_3}{\sigma_X \sigma_F + C_3}, \quad (17)$$

where $x$ and $f$ are the image patches of the source image $X$ and the fused image $F$, respectively. $\sigma$ denotes the covariance or the standard deviation. $\mu$ denotes the mean values. $C_1$, $C_2$ and $C_3$ are the parameters for stability. Then, MSSIM can be evaluated by mean SSIM:

$$MSSIM(X,F) = \frac{1}{K} \sum_{j=1}^{K} SSIM(x_j, f_j), \quad (18)$$

where $K$ is the number of all the sliding windows. And the final MSSIM value can be calculated as:

$$MSSIM(A,B,F) = \frac{1}{2}MSSIM(A,F) + \frac{1}{2}MSSIM(B,F), \quad (19)$$

where $A$ and $B$ are the over-exposed and under-exposed images, respectively. Thus, MSSIM can be used to measure the mean structural similarity between the source images and the fused image.

c) Metric reflecting the nature of the fused image: Before measuring this metric, the fused images are firstly converted to gray-scale images.

- Standard deviation (SD): SD is a metric reflecting contrast and distribution. Attention of human is more likely to be attracted by the area with high contrast. Thus, a fused image with high contrast often results in a large SD, which means that the fused image achieves a good visual effect. Mathematically, SD is defined as:

$$SD = \sqrt{\frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} (f_{i,j} - \mu_f)^2}, \quad (20)$$

where $\mu_f$ is the mean value of the image $f$.

We test these six metrics on the rest 60 image sequences and the comparison results are shown in Fig. 9. As shown in Fig. 9, our results can achieve the highest average values on all the six metrics. The highest average values of PSNR$_R$, PSNR$_S$, $CC_1$, and $CC_3$ represent that our method allow the fused images to achieve the least distortion, and the strongest similarity and linear correlation with the ground truth images. Moreover, the optimal mean value of our MEF-GAN on MSSIM shows that our results can achieve the highest structural similarity with the source images besides the similarity between the fused result and the ground truth. In addition, the largest average values of our results on the metric SD illustrates that our results are of the highest contrast, and the best visual effect.

3) Efficiency Comparison: We compare the efficiency of the proposed model with competitors on the 60 test image pairs. The runtime is shown in Table II. Among them, GFF, DSIFT, GBM and FLER are traditional methods which are tested on a laptop with 1.9 GHz Intel Core i3-4030 CPU. Deepfuse and MEF-GAN are deep learning-based methods and are tested on NVIDIA GeForce GTX Titan X. Compared with traditional methods, MEF-GAN can achieve comparable efficiency. Compared with Deepfuse, it takes more time for our model to generate the result. The reason is that the generator architecture of our model is more complex with three blocks.
TABLE II
RUNTIME OF DIFFERENT METHODS ON THE 60 TEST IMAGE PAIRS (UNIT: SECOND)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>6.593</td>
<td>10.935</td>
<td>7.694</td>
<td><strong>0.269</strong></td>
<td>66.144</td>
<td>3.545</td>
</tr>
<tr>
<td>STD</td>
<td>3.115</td>
<td>5.086</td>
<td>3.649</td>
<td><strong>0.144</strong></td>
<td>37.072</td>
<td>2.641</td>
</tr>
</tbody>
</table>

Among them, the self-attention block has the largest amount of computation and is the most time consuming.

C. Ablation Study

1) Ablation Study of Blocks: There are three blocks in the generator, i.e., the self-attention block, the local detail block, and the merge block. Among them, the self-attention block and the local detail block complement each other.

To verify the effect of each block, we perform two comparative experiments here. First, to verify the effect of the local detail block, the detail block is not employed. And the generator consists of the self-attention block and the merge block. The fused result is shown in Fig. 10(a). It can be observed that the whole image has consistency and continuity. However, it is inevitably affected by blur and inaccuracy. Then, to verify the effect of self-attention block, the self-attention block is removed. The fused result is shown in Fig. 10(b). Compared with the result in Fig. 10(a), more texture details are presented in the result. Nevertheless, it comes with another problem. As shown in the highlighted regions shown in the bottom rows of Fig. 10, there is obvious distortion in the luminance or the chrominance in some local regions compared with the ground truth. Through the combination of the self-attention block and the local detail block, the generator can combine the advantages of both blocks to make up for their shortcomings. This is another contribution of the proposed MEF-GAN. More concretely, the fused result of MEF-GAN has a clear representation of local details while taking into account the attention-driven and long-range dependency to avoid local distortion, confusing results or artifacts.

The quantitative comparison is shown in Table III. The method of removing the local detail block shows the worst performance on the first five metrics. And the method of removing the self-attention block ranks second. By combining them, MEF-GAN shows the best results on all the six metrics. It shows that the combination of the local detail block and the self-attention block can achieve higher similarity with the ground truth and source images, and exhibit higher contrast.

2) Ablation Study of the Gradient Loss: In the content loss, to compensate for the deficiency of the MSE loss and retain the content with high frequency, we apply the gradient loss shown in Eq. (8). To verify the effect, we remove it and retrain the model. The other settings are the same as those of MEF-GAN. The qualitative results are shown in Fig. 11. By comparing the results obtained by removing the gradient loss and those of MEF-GAN, we can conclude that the introduction of this loss can retain more high-frequency textures, edges and details in the results. The quantitative results are also shown in Table III.
TABLE III
QUANTITATIVE COMPARISON OF ABLATION STUDY ON 60 IMAGE SEQUENCES WITH MEAN AND STANDARD DEVIATION SHOWN BELOW

<table>
<thead>
<tr>
<th>Method</th>
<th>PSNR₁</th>
<th>PSNR₅</th>
<th>CC₁</th>
<th>CC₅</th>
<th>MSSIM</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>Removing local detail block</td>
<td>66.1056±1.8900</td>
<td>66.6321±1.9713</td>
<td>0.8781±0.0715</td>
<td>0.9068±0.0673</td>
<td>0.9817±0.0048</td>
<td>0.2102±0.0552</td>
</tr>
<tr>
<td>Removing self-attention block</td>
<td>68.2941±2.5801</td>
<td>68.3017±2.5588</td>
<td>0.9403±0.0428</td>
<td>0.9546±0.0382</td>
<td>0.9820±0.0049</td>
<td>0.2031±0.0518</td>
</tr>
<tr>
<td>Removing the gradient loss</td>
<td>68.2020±2.8043</td>
<td>68.3523±2.5850</td>
<td>0.9386±0.0517</td>
<td>0.9529±0.0393</td>
<td>0.9815±0.0049</td>
<td>0.2395±0.0518</td>
</tr>
<tr>
<td>MEF-GAN</td>
<td>68.4200±2.6872</td>
<td>68.3721±2.4810</td>
<td>0.9420±0.0435</td>
<td>0.9548±0.0376</td>
<td>0.9820±0.0048</td>
<td>0.2103±0.0539</td>
</tr>
</tbody>
</table>

Fig. 13. MEF-GAN to fuse images of different exposure ratios of the same sequence ((1), (2) and (3) in the image sequence are under-exposed images; (4) and (5) are over-exposed images).

As for PSNR₁, PSNR₅, CC₁, CC₅ and MSSIM, MEF-GAN can achieve the better mean values. As for the metric SD, our MEF-GAN follows behind the comparative method. It is because SD focuses on the contrast of the fused image rather than details and edges. Also because of the adversarial relationship between the discriminator and the generator, there can be some changes in the contrast of the fused results.

D. MEF-GAN Under Other Conditions

In this paper, we propose a GAN-based model, MEF-GAN, which can directly fuse two source images (an under-exposed image and an over-exposed image) with large exposure ratio. To verify the generality of the proposed model, we perform our MEF-GAN under other conditions: MEF-GAN with multiple inputs and inputs with arbitrary exposure ratio.

1) Multiple Inputs: In the source image sequence, if more than two images are available, these images can be fused in turn. As shown in Fig. 12, we can firstly fuse two of them and then fuse the intermediate result and another source image. In this way, the proposed method can be employed to produce the input with any number of inputs.

2) Inputs With Different Exposure Ratios: Even though we select the brightest and darkest images as the inputs in the training dataset, the exposure ratio of the brightest and darkest images in each image sequence is constantly changeable. Moreover, fusing the images with large exposure ratio is more challenging as there is less information in the darkest and brightest images. The generator is trained to extract and fuse the deep features contained in them. As for the images with moderate exposure ratio, the features can also be extracted by the trained generator. Thus, the restriction of the exposure ratio on MEF-GAN is limited. To verify the effectiveness of MEF-GAN on source images with different exposure ratios, we select five images of the same scene which are taken with different exposure settings. The results of any two images in this sequence are shown in Fig. 13. Since our model is trained on images with large exposure ratio, the fused result of fusing (1) and (5) is the most similar with the ground truth. Although other results are less similar to the ground truth, MEF-GAN can also generate comparable results on the untrained images with comparatively moderate exposure ratio. Thus, MEF-GAN are applicable to fuse inputs with different exposure ratios.

V. CONCLUSION

We propose a novel GAN-based multi-exposure image fusion method, termed as MEF-GAN. Rather than applying a network to minimizing the value of a metric which constrains the similarity of some information between the fused result and the source images, we construct a min-max two-player game. In our method, given the over-exposed and under-exposed images, a generator network is trained to generate the fused image. Meanwhile, a discriminator network is trained to form an adversarial relationship with the generator by distinguishing the fused image from the corresponding ground truth, in addition to the content loss. Furthermore, the self-attention mechanism introduced in the generator also improves the performance of the fused image by allowing for attention-driven and long-range dependency. Compared with other five state-of-the-art fusion methods, our method can achieve advanced performance both qualitatively and quantitatively.
REFERENCES


Jiayi Ma (Member, IEEE) received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. From 2012 to 2013, he was an Exchange Student with the Department of Statistics, University of California at Los Angeles, Los Angeles, CA, USA. He was a Postdoctoral Researcher with Electronic Information School, Wuhan University, from August 2014 to November 2015, and received an accelerated promotion to an Associate Professor and a Full Professor in December 2015 and December 2018, respectively. He has authored or coauthored more than 130 refereed journal and conference papers, including the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE (TPAMI)/TRANSACTIONS ON IMAGE PROCESSING (TIP)/TRANSACTIONS ON SIGNAL PROCESSING (TSP), IJCV, CVPR, and ICCV. His research interests include computer vision, machine learning, and pattern recognition. Dr. Ma has been identified in the 2019 Highly Cited Researchers list from the Web of Science Group. He was a recipient of the Natural Science Award of Hubei Province (first class), the Chinese Association for Artificial Intelligence (CAAI) Excellent Doctoral Dissertation Award (a total of eight winners in China), and the Chinese Association of Automation (CAA) Excellent Doctoral Dissertation Award (a total of ten winners in China). He is an Editorial Board Member of Information Fusion and Neurocomputing and a Guest Editor of Remote Sensing.

Han Xu received the B.S. degree from Electronic Information School, Wuhan University, Wuhan, China, in 2018, where she is currently pursuing the master’s degree with Multi-spectral Vision Processing Laboratory. Her current research interests include computer vision and pattern recognition.

Xiao-Ping Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from Tsinghua University, in 1992 and 1996, respectively, and the M.B.A. degree (Hons.) in finance, economics, and entrepreneurship from The University of Chicago Booth School of Business, Chicago, IL, USA.

Since Fall 2000, he has been with the Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Toronto, ON, Canada, where he is currently a Professor and the Director of the Communication and Signal Processing Applications Laboratory. He has served as the Program Director of Graduate Studies. He is also cross-appointed to the Finance Department at the Ted Rogers School of Management, Ryerson University. He was a Visiting Scientist with the Research Laboratory of Electronics, Massachusetts Institute of Technology, Cambridge, MA, USA, in 2015 and 2017. He is also a Frequent Consultant for biotech companies and investment firms. He is also the Co-Founder and the CEO for EidoSearch, an Ontario-based company offering a content-based search and analysis engine for financial big data. His research interests include image and multimedia content analysis, machine learning, statistical signal processing, sensor networks and electronic systems, and applications in big data, finance, and marketing.

Dr. Zhang is a registered Professional Engineer in Ontario, Canada, and a member of the Beta Gamma Sigma Honor Society. He is an elected member of the ICME steering committee. He received the 2020 Sarwan Sahota Ryerson Distinguished Scholar Award, the Ryerson University Highest Honor for Scholarly, Research, and Creative Achievements. He is the General Co-Chair of the IEEE International Conference on Acoustics, Speech, and Signal Processing, in 2021. He is also the General Co-Chair of 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business and the 2019 GlobalSIP Symposium on Signal, Information Processing and AI for Finance and Business. He is the General Chair of the IEEE International Workshop on Multimedia Signal Processing, in 2015. He is the Publicity Chair of the International Conference on Multimedia and Expo 2006 and the Program Chair of International Conference on Intelligent Computing in 2005 and 2010. He has served as a Guest Editor for Multimedia Tools and Applications and the International Journal of Semantic Computing. He was a Tutorial Speaker at the 2011 ACM International Conference on Multimedia, the 2013 IEEE International Symposium on Circuits and Systems, the 2013 IEEE International Conference on Image Processing, the 2014 IEEE International Conference on Acoustics, Speech, and Signal Processing, the 2017 International Joint Conference on Neural Networks, and the 2019 IEEE International Symposium on Circuits and Systems. He is a Senior Area Editor of the IEEE TRANSACTIONS ON SIGNAL PROCESSING and the IEEE TRANSACTIONS ON IMAGE PROCESSING. He was an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON MULTIMEDIA, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, the IEEE TRANSACTIONS ON SIGNAL PROCESSING, and the IEEE SIGNAL PROCESSING LETTERS. He is awarded as the IEEE Distinguished Lecturer for the term from January 2020 to December 2021 by the IEEE Signal Processing Society.