IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 6, 2019, accepted November 20, 2019, date of publication November 25, 2019,
date of current version December 19, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2955382

Multi-Source Medical Image Fusion Based on Wasserstein Generative Adversarial Networks

ZHIGUANG YANG[1], YOUPING CHEN[1], ZHULIANG LE[2], FAN FAN[2], AND ERTING PAN[2]
[1]School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Electronic Information School, Wuhan University, Wuhan 430072, China

Corresponding authors: Fan Fan (fanfan@whu.edu.cn) and Erting Pan (panerting@whu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61605146.

ABSTRACT In this paper, we propose the medical Wasserstein generative adversarial networks (MWGAN), an end-to-end model, for fusing magnetic resonance imaging (MRI) and positron emission tomography (PET) medical images. Our method establishes two adversarial games between a generator and two discriminators to generate a fused image with the details of soft tissue structures in organs from MRI images and the functional and metabolic information from PET images. Different information from source images can be effectively adjusted with a specifically designed loss function. In addition, we use WGAN instead of the traditional generative adversarial networks to make the training process more stable and allow our architecture to deal with source images of different resolutions. Qualitative and quantitative comparisons on publicly available datasets demonstrate the superiority of MWGAN over the state-of-the-art networks. Furthermore, our MWGAN is applied to the fusion of MRI and computed tomography images of different resolutions, achieving a satisfactory performance.

INDEX TERMS Medical image fusion, Wasserstein generative adversarial networks, end-to-end, different resolutions.

I. INTRODUCTION

Medical image fusion makes full use of multi-source images to obtain complementary information, which makes clinical diagnosis and treatment more accurate and perfect. The large amount of redundant information in medical images is a waste of storage space. Thus, the target of fusion is to preserve the vital information in them and merge it into a single image. Specifically, in the field of medical imaging, there are magnetic resonance imaging (MRI) images that capture details of organs' soft tissue structures (*e.g.*, texture detail information) and positron emission tomography (PET) images that provide functional and metabolic information (*e.g.*, pixel intensity information) [1]. Fusing these images ensures that the resulting image will have both soft tissue details and functional and metabolic information.

The key to fusing source images from different sensors is to extract the most important information of the source images into a single image. In the past decades, different schemes for image fusion have been developed, including subspace-based methods [2], hybrid methods [3], multiscale transform-based methods [4], saliency-based methods [5], sparse representation-based methods [6], and other fusion methods [7], [8]. However, these methods require manual design of complex activity level measurements and fusion rules and are thus limited by implementation difficulty and high computational costs [9]–[12].

Recently, due to the strong ability of extracting image features, the deep learning technique has been successfully applied to the problem of image fusion. In particular, Ma *et al.* [13], [14] proposed an end-to-end model based on generative adversarial networks (GAN) for infrared and visible image fusion, which can solve the defects of the traditional fusion methods to a certain extent and achieve promising fusion results. However, inadequacies still exist. On the one hand, in the current GAN-based fusion methods, balancing the training level of the generator and the discriminator is challenging, thereby posing difficulties in network training. On the other hand, the traditional GAN-based setting strategy of loss function causes ineffective adjustment of the ratio of information from source images, leading to a unbalance on preserving the various information in source images.

The associate editor coordinating the review of this manuscript and approving it for publication was Guobao Xiao.

In addition, due to the increasing diversity and complexity of scene requirements, demand for the fusion of medical images from different resolutions has been growing. However, low-resolution PET images have more blurred details than the corresponding high-resolution MRI images. Strategies to increase the resolution of PET images or reduce the resolution of MRI images result in loss of high frequency detail information because of the difficulty of upgrading hardware devices and algorithms. Therefore, studying the fusion of medical images from different resolutions without the loss of important information is of great significance.

To overcome the above difficulties, we proposed a new medical image fusion algorithm based on Wasserstein generative adversarial networks, termed as MWGAN. Unlike the two-player game in traditional GAN, our MWGAN is a three-player game which includes two types of networks such as a generator (G) and two discriminators (D_I and D_M). In particular, G constantly predicts the probability distribution of the real data in the training set to transform the fusion image generated by the two source images into an image that cannot be distinguished from the real one. The more similar the generated image is to the image in the training set, the better. D_I and D_M determine whether an image is real or not, *i.e.*, distinguishing the fused image from the original PET and MRI images, respectively. The goal is to distinguish the ''fake'' image generated by the generator G from the ''true'' image in the training set. MWGAN not only makes D_I and G play games, but also makes D_M and G play games. In the training process, the two networks are simultaneously enhanced through competition (*i.e.*, allowing the networks to compete with each other). Given the existence of D_I and D_M, G can better approximate the real data under the premise of prior knowledge and prior distributions and generate data that resembles the real data. D_I and D_M cannot distinguish the image generated by G and the real image. Hence, they reach a certain respective Nash equilibrium with G. In addition, we provide a new loss function setting strategy to effectively adjust the ratio of information from the source images. The proposed MWGAN is an end-to-end model with deep learning implemented throughout the entire fusion process. We use five bottlenecks as the encoders of G and a trainable deconvolution layer to effectively avoid the loss of information from source images. Finally, MWGAN can be extended to the fusion of MRI and computed tomography (CT) images. The experimental results verify the superiority of our MWGAN over the state-of-the-art methods.

The major contributions of this paper involve the following three aspects. First, we propose a more stable WGAN-based model with a new loss function setting strategy for fusion of MRI and PET medical images of different resolutions. To the best of our knowledge, it is the first time that the GANs are adopted for addressing the medical image fusion task. Second, the proposed MWGAN is an end-to-end model with deep learning implemented throughout the entire fusion process, exhibiting better fusion performance as indicated by qualitative and quantitative comparisons with state-of-the-art fusion methods. Third, we also apply the proposed MWGAN to MRI and CT image fusion, and has achieved excellent performance.

The remainder of this paper is organized as follows. Sec. II presents related fusion methods based on deep learning and the basic introduction of WGAN. Sec. III introduces the proposed MWGAN algorithm in detail. Sec. IV shows the fusion performance of our method on various types of MRI and PET image pairs compared with other state-of-the-art methods in terms of qualitative visual effect and quantitative metrics. The proposed method is applied to the fusion of MRI and CT image pairs from different resolutions, and the qualitative visual effect is also demonstrated in this section. Finally, we draw the conclusion in Sec. V.

II. RELATED WORKS

In this section, we present a brief introduction of existing image fusion methods based on deep learning. Furthermore, given that our fusion method is based on WGAN, we will discuss the traditional GAN and provide a basic explanation of WGAN.

A. FUSION METHODS BASED ON DEEP LEARNING

With the advent of deep learning, the outstanding performance of deep learning in feature learning and reconstruction has received extensive attention; thus, deep learning has been successfully applied in many image fusion fields. In the field of digital imaging, Tang *et al.* [15] proposed a multi-focus image fusion method that recognizes the focused and defocused pixels in source images from their neighborhood information. Prabhakar *et al.* [16] solved the problem of multiexposure fusion by proposing an unsupervised deep learning framework. In multimodal image fusion, Ma *et al.* [13], [14], [17] applied GAN to the fusion of infrared and visible images, and the generated image obtained more details from the visible image due to the existence of a discriminator. In the field of remote sensing image fusion, Masi *et al.* [18] proposed a new three-layer structure to solve the pan-sharping problem based on convolutional neural networks (CNNs), which effectively improved the fusion performance without adding complexity.

The diversity of deep learning models also provides strong support for deep learning in medical image fusion. Liu *et al.* [19] proposed an image fusion framework based on convolutional sparse representation (CSR) to solve the problem of limited detail preservation and high sensitivity to misregistration, wherein each source image is decomposed into base and detail layers for fusing multi-focus and multi-modal images. In addition, Liu *et al.* [20] employed a Siamese convolutional network to generate a weight map that integrates the pixel activity information from two source images. To conform with human visual perception, multiscaling via image pyramids is applied in the fusion process.

The above methods have strong reference values due to their good performance. However, the following shortcomings are still unresolved: (i) The existing image fusion

methods based on GAN have unstable training and are prone to gradient vanishing/exploding; (ii) in GAN-based image fusion methods, the information of the source images is limited in the weight ratio of the fused image, making their adjustments difficult; and (iii) the limitations of traditional fusion methods still exist. For example, fusion rules are still designed in a manual way. Hence, deep learning cannot be implemented throughout the entire fusion process. The proposed MWGAN can solve the above problems properly. To overcome the first shortcoming, we introduce WGAN to solve the problem of instability in the training process of GAN. On this basis, we provide a new loss function setting strategy to make the weight ratio adjustment easier. Moreover, the proposed method does not require manual design of the fusion rules, and the whole process is free from the limitations of traditional methods.

B. WASSERSTEIN GENERATIVE ADVERSARIAL NETWORKS

GAN [21] was first proposed in 2014. This network has made many remarkable achievements in many fields, including image fusion. GAN is one of the current mainstream generative model. The application of GAN can avoid the problem of manually designing complex activity level measurements and fusion rules, thus can specify a high-level goal. However, unstable training is still an issue with GAN. Several works have attempted to solve this problem, but the results are not fully successful. For example, deep convolutional GAN [22] aimed to find the best set of network architecture settings on the basis of experimental enumeration of the generator and discriminator architectures. A temporary solution was formulated, but the problem was not completely solved. The birth of WGAN [23] has effectively solved the above problems.

The root causes of the original GAN problems can be reduced to two points: (i) unreasonable divergence measurement (Kullback Leibler [KL] and Jensen Shannon [JS] divergences) after the equivalent optimization and (ii) difficulty in overlapping the generated distribution of generator after random initialization with the real distribution. WGAN introduces the Wasserstein distance, which is defined as follows:

$$W(P_r, P_g) = \inf_{\gamma \in \prod(P_r, P_g)} \mathbb{E}_{(x,y)\sim\gamma}[\|x - y\|]. \tag{1}$$

Unlike the KL and JS divergences, the Wasserstein distance can still reflect distance even if the two distributions do not overlap. Furthermore, Wasserstein distance has superior smoothing characteristics compared with KL and JS divergences, which can solve the problem of gradient vanishing, that is, balancing the training level of the generator and discriminator. The Wasserstein distance is written in a solvable form (mathematical transformation and can be maximized using a neural network of D). Then, G is optimized under this approximate optimal D to reduce the Wasserstein distance, which effectively shortens the distance between the generated and the real distribution. The loss functions of G and D of WGAN are separately defined as follows:

$$L_G = -\mathbb{E}_{x\in P_g}[D_w(x)], \tag{2}$$

$$L_D = \mathbb{E}_{x\in P_g}[D_w(x)] - \mathbb{E}_{x\in P_r}[D_w(x)], \tag{3}$$

where w indicates the parameters with a limited value.

III. METHOD

In this section, we describe the application of WGAN on the fusion of PET and MRI images with different resolutions. We first explain the entire fusion procedure of MWGAN, and then we provide the problem formulation, the new setting strategy, and detailed definitions of our loss function. Finally, we discuss the network architectures of our MWGAN.

A. OVERVIEW OF THE FRAMEWORK

The PET image is a low-resolution pseudo-color image that represents the uptake of the radiotracer and provides important functional and metabolic information. The MRI image is a high-resolution grayscale image that provides texture details. Therefore, the fused image needs to be achieved in the de-correlated color model without any changes in the functional and metabolic information while retaining all the soft tissue structure details. The selection of the color model also greatly influences the fusion result. Therefore, to make the colors of the fused image as close as possible to those of the PET image, the de-correlated color model must separate the achromatic and chromatic information into different channels. The IHS color conversion model can well preserve the spectral information and spatial resolution of the source image, which is adopted in our work. Hue (H) and saturation (S) channels are chromatic information that does not need to be changed, whereas the intensity channel (I) is the specific achromatic channel participating in fusion with MRI image [24].

The objective of our work is to fuse the high-resolution MRI image and the low-resolution PET image. Initially, we uniformly set the resolution of the MRI image to be 4×4 of the PET image. Prior to the formal participation in MWGAN fusion, a multispectral PET image is converted from RGB channels to IHS channels, where I presents the brightness in the spectrum, H displays the spectral wavelength characteristics, and S demonstrates the spectral purity. The conversion process is expressed as follows:

$$\begin{pmatrix} I_{PET} \\ X1_{PET} \\ X2_{PET} \end{pmatrix} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} \begin{pmatrix} R_{PET} \\ G_{PET} \\ B_{PET} \end{pmatrix}, \tag{4}$$

where $X1_{PET}$ and $X2_{PET}$ are intermediate forms of the conversion process. The H and S channels are indicated as follows:

$$H_{PET} = \arctan\left(X1_{PET}/X2_{PET}\right), \tag{5}$$

$$S_{PET} = \sqrt{X1_{PET}^2 + X2_{PET}^2}. \tag{6}$$

Fusion occurs between the MRI image and the achromatic channel (I) of the PET image. The inputs to the MWGAN

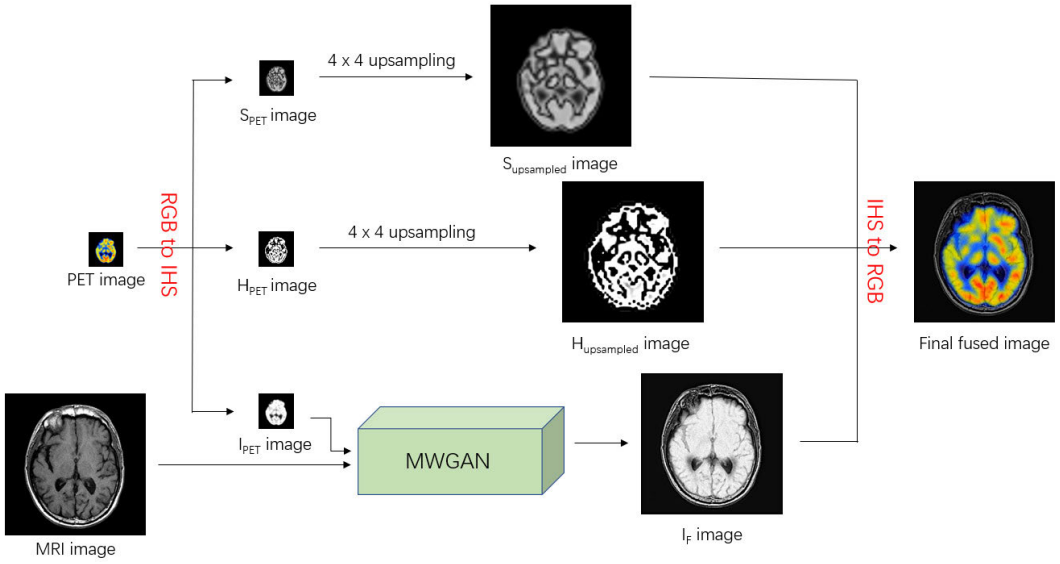

FIGURE 1. The whole procedure of the proposed fusion model.

module are a high-resolution MRI image and the corresponding low-resolution image I_{PET}, and the output is an image I_F with the same resolution as the input MRI image. For H_{PET} and S_{PET}, we use bicubic interpolation as the upsampling operation because the fused image needs to retain the chromatic information in the color of the PET image. The resolutions of upsampled $H_{upsampled}$ and $S_{upsampled}$ are both 4×4 higher than those of H_{PET} and S_{PET}. Finally, I_F, $H_{upsampled}$ and $S_{upsampled}$ are switched back to the RGB channel by using the conversion process below

$$X1_{RGB} = S_{upsampled} \sin(H_{upsampled}), \tag{7}$$

$$X2_{RGB} = S_{upsampled} \cos(H_{upsampled}). \tag{8}$$

The components of the R, G and B channels of the final fused image can be represented by variables $X1_{RGB}$, $X2_{RGB}$ and I_F, respectively, as follows:

$$\begin{pmatrix} R \\ G \\ B \end{pmatrix} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{6} & -1/\sqrt{2} \\ 1/\sqrt{3} & -2/\sqrt{6} & 0 \end{bmatrix} \begin{pmatrix} I_F \\ X1_{RGB} \\ X2_{RGB} \end{pmatrix}. \tag{9}$$

We summarize the whole fusion procedure in Fig. 1.

B. PROBLEM FORMULATION

The proposed MWGAN has two differences in network structure from the existing GAN. First, we use two Ds to constrain the fused image generated by G. Then, we use a more stable WGAN.

The training procedure of the proposed MWGAN is shown in Fig. 2. The high-resolution MRI image and low-resolution I_{PET} image are used as input to G, which in turn produces a fused image. Two aspects must be considered to obtain a fused image that has the texture detail information of the MRI image and the intensity information of the PET image. One is the design of G's loss function and the other is the existence of two Ds. We feed the fused image I_F and the source MRI image into D_M to distinguish I_F from the MRI image. Meanwhile, we feed the downsampled I_f through average pooling and the source I_{PET} image into D_I to distinguish I_f from the source I_{PET} image. We do not distinguish I_F from the upsampled I_{PET} image because the intensity information of the fused image needs to be closer to the intensity information of the input I_{PET} image rather than the upsampled I_{PET} image, which can ensure the authenticity of the information. We consider the fusion of I_{PET} and MRI as two adversarial games that occurs not only in G and D_M but also in G and D_I. During the training procedure, the two models are trained simultaneously by testing them against each other. As a result, D_M cannot distinguish the fused image I_F generated by G and the source MRI image, whereas D_I cannot distinguish between the downsampled fused image I_f and the source I_{PET} image.

C. LOSS FUNCTION

For GAN, the randomness and instability in the training process will inevitably result in uncontrollable or unexpected results for the fused image [25]. To obtain the expected fused image, we need to apply a series of constraints into the network. In the existing GAN, researchers are accustomed to separating the adversarial loss and the content loss [13]

$$L_G = L_G^{adv} + \lambda L^{con}. \tag{10}$$

However, it may not be suitable for image fusion, which owns two real datas. Moreover, the setting strategy poses difficulty in performing an effective adjustment to the overall information proportion of the MRI image and the I_{PET} image. The proposed MWGAN adjusts the loss function setting strategy, thereby optimizing the loss of MRI image and the loss of I_{PET} image separately to ensure that the overall information

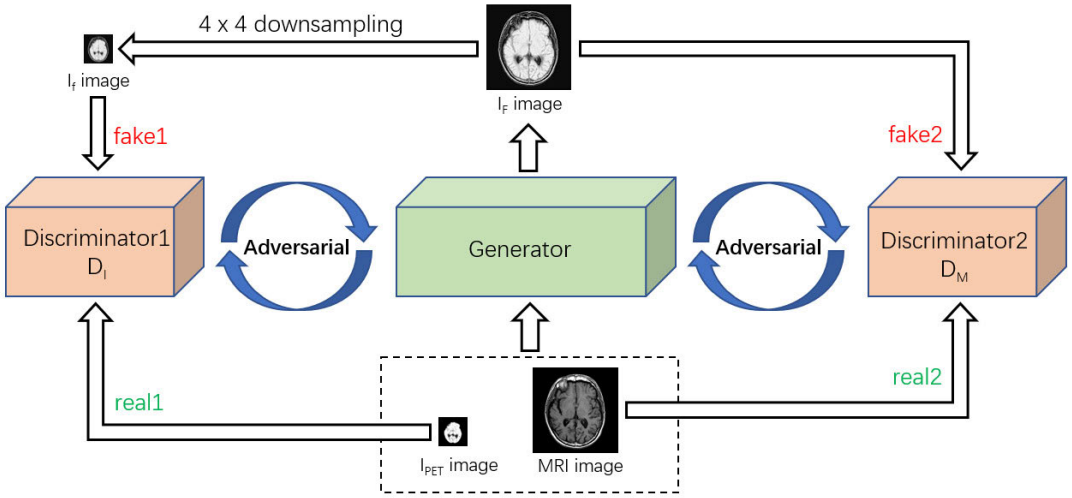

FIGURE 2. The training procedure of MWGAN.

ratio of MRI image and I_{PET} image is effectively adjusted while obtaining as much information of source images as possible. The new loss function setting strategy is defined as follows:

$$L_G = L_{I_{PET}} + \lambda L_{MRI}. \tag{11}$$

The loss function of G involves the losses of the I_{PET} and MRI images, whereas the weight λ controls the trade-off between the losses of the MRI and I_{PET} images. The first term $L_{I_{PET}}$ on the right hand denotes the loss of I_{PET} image, which is defined as follows:

$$L_{I_{PET}} = L^{adv}_{I_{PET}} + \alpha L^{con}_{I_{PET}}, \tag{12}$$

where the weight α is used to control the trade-off, $L^{adv}_{I_{PET}}$ conveys the adversarial loss between G and D_I, which is defined as follows:

$$L^{adv}_{I_{PET}} = \mathbb{E}[-D_I(\Gamma G(MRI, I_{PET}))], \tag{13}$$

where Γ denotes the downsampling operation. The intensity information of the I_{PET} image is represented by pixel intensities. $L^{con}_{I_{PET}}$ can constrain the pixel intensities of the downsampled fused image I_f so that it will have an intensity information that is as similar as possible to that of the source I_{PET} image. The $L^{con}_{I_{PET}}$ is expressed as follows:

$$L^{con}_{I_{PET}} = \mathbb{E}[\|\Gamma G(MRI, I_{PET}) - I_{PET}\|^2_F]. \tag{14}$$

The second term L_{MRI} in (11) represents the loss of MRI image, which is defined as follows:

$$L_{MRI} = L^{adv}_{MRI} + \beta L^{con}_{MRI}, \tag{15}$$

where the weight β is used to control the trade-off, L^{adv}_{MRI} indicates the adversarial loss between G and D_M, which is defined as follows:

$$L^{adv}_{MRI} = \mathbb{E}[-D_M(G(MRI, I_{PET}))]. \tag{16}$$

The texture details of the MRI image are mainly represented by gradient variation, and L^{con}_{MRI} can constrain the gradient variation of the fused image I_F so that it will have texture details that are as similar as possible to those of the source MRI image. The L^{con}_{MRI} is defined as follows:

$$L^{con}_{MRI} = \mathbb{E}[\|\triangle G(MRI, I_{PET}) - \triangle MRI\|]. \tag{17}$$

where $\triangle$ denotes Laplacian operator.

The existence of Ds allows the generated fused image to be closer to the source images. The adversarial loss of D_M and D_I respectively judges the closeness of the texture details and the pixel intensities of the fused image to the source images by calculating the Wasserstein distance between the generated and the real distributions to maximize the ability of the fused image to obtain the texture details of the MRI image and the intensity information of the I_{PET} image, which is respectively defined as follows:

$$L_{D_M} = \mathbb{E}[D_M(G(MRI, I_{PET}))] - \mathbb{E}[D_M(MRI)], \tag{18}$$
$$L_{D_I} = \mathbb{E}[D_I(\Gamma G(MRI, I_{PET}))] - \mathbb{E}[D_I(I_{PET})]. \tag{19}$$

D. NETWORK ARCHITECTURE

From the training procedure of MWGAN, we can see that our network architecture consists of one G and two Ds. In this section, we will introduce the network architecture of the G and Ds separately.

1) GENERATOR

The network structure of G consists of two deconvolution layers, namely, one encoder and the corresponding decoder, as shown in Fig. 3. The MRI and I_{PET} images are source images that have different resolutions. Thus, we first convert the resolution of the I_{PET} image to that of the MRI image through deconvolution before officially entering the fusion process. To obtain the feature of the MRI image at the same time, the equivalent deconvolution processing for the MRI

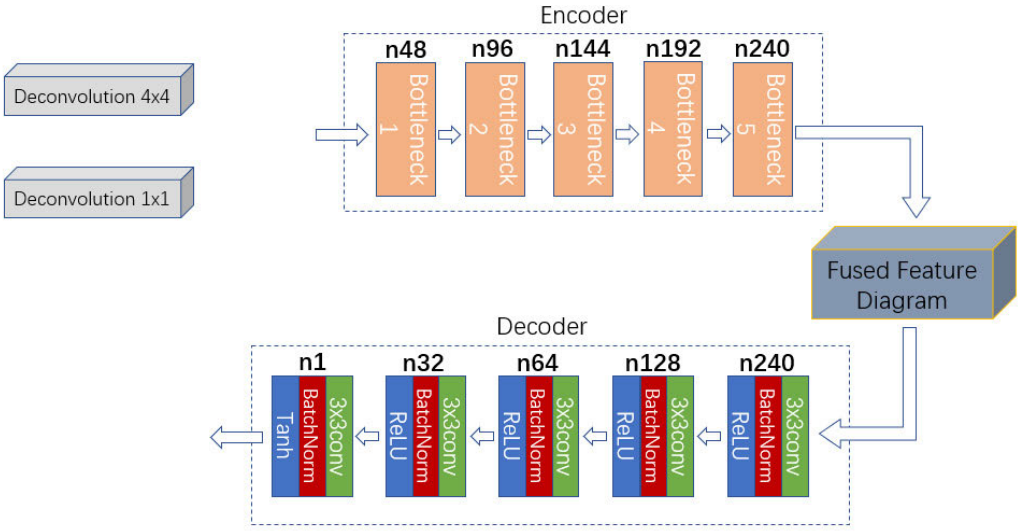

FIGURE 3. The network architecture of the generator.

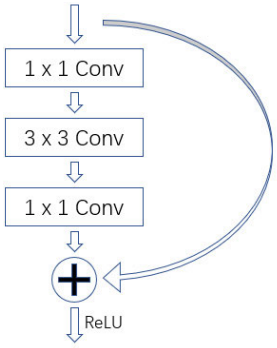

FIGURE 4. The network architecture of the bottleneck.

image is performed, which generates a feature diagram with the same resolution. The deconvolution method differs from traditional interpolation by the nearest, bilinear, or bicubic method in terms of obtaining the parameters, which are obtained automatically by training. We then concatenate the two feature diagram results from the two deconvolution layers and feed them into the encoder. The fused feature diagram is also the input of the following decoder for the construction of fused image.

The network structure of the encoder draws on Resnet v1's network structure [26]. Our encoder consists of five bottlenecks, as shown in Fig. 4. To avoid any loss in the source images' information, the stride of each convolutional layer is set to 1 to avoid downsampling. We define the rates of all bottlenecks as 1 to remove forged boundaries.

Five CNN layers are present in the decoder. Batch normalization and ReLU are applied to alleviate gradient exploding/vanishing and accelerate the training and avert gradient sparsity, respectively. The number of feature diagrams of the final generated image is reduced to one.

2) DISCRIMINATOR

The two Ds act as two judges in the MWGAN. Given the existence of D_M and D_I, the generated fused image can approach the source MRI and I_{PET} images. The same network structure is owned by the two Ds, which is illustrated in Fig. 5. The stride of all convolutional layers is set to 2 without padding, and the activation function is not employed in the last layer.

IV. EXPERIMENTAL RESULTS

In this section, we verify the performance of our MWGAN on publicly available datasets with comparison to the state-of-the-art fusion methods. We first introduce the dataset and training details, and then provide qualitative and quantitative results. We also conduct the clipped analysis of G's parameters and stability analysis. In addition, the fusion results on MRI and CT images are also showed.

A. DATASET

The proposed MWGAN is verified on the publicly available dataset on the Harvard Medical School website. We selected 83 pairs of MRI and PET images with the same resolution (256×256) from the dataset's website. Then, we extract I from the PET images, which leads to 83 pairs of MRI and I_{PET} images with a same resolution of 256×256. Next, these pairs are cropped into 9, 984 84×84 patch pairs to serve as our training set. Before training, we downsample the I_{PET} patches to fuse the source images of different resolutions. Note that the original image pairs have been aligned in advance, and image registration is required for unaligned image pairs [27]–[29].

B. TRAINING DETAILS

Our proposed MWGAN with two Ds is different from the traditional GAN. We need to make the G and two Ds form two adversarial relationships and ensure the balance between

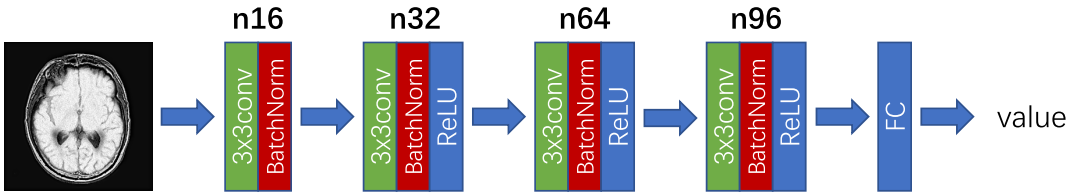

FIGURE 5. The network architecture of the discriminator.

the D_M and D_I so that one is not too strong and the other one is not too weak. The G, D_M, and D_I are not traditionally trained in turns (once per batch). If D is strong that it always distinguish the data from G, then we train G a few more times, and vice versa. We not only clip the absolute value of D's parameters to a fixed parameter w, but also clip the absolute value of G's parameters to obtain a better fused image, which is verified later in Section IV-C.3. The detailed training procedure is summarized in Algorithm 1. During the testing stage, we input the source images whose resolution must be of a certain proportion into the trained G to obtain the fused image I_F.

C. FUSION RESULTS ON MRI AND PET IMAGES

To verify the effectiveness of our proposed MWGAN, we select five state-of-the-art fusion methods to compare with our MWGAN, namely, wavelet transform (DCHWT) [30], FPDE [2], GTF [7], FusionGAN [13], and Structure-Aware [31]. DCHWT, FPDE, GTF, and Structure-Aware are relatively classic, successful traditional fusion methods. FusionGAN is a new fusion method based on deep learning. The parameters of these five methods are the same as those in the original papers. These five fusion methods will be compared not only qualitatively but also quantitatively with our MWGAN. FusionGAN can be applied to the fusion of images with different resolutions. Hence, we will directly perform the fusion operation to obtain the results. However, the remaining fusion methods, whose source images demand the same resolution, require an upsampling operation on the low-resolution I_{PET} image before the fusion operation.

1) QUALITATIVE COMPARISONS

The qualitative comparisons with the five state-of-the-art fusion methods are illustrated in Fig. 6. We demonstrate four representative and intuitive generated images from four different axes of the brain hemisphere, including the cerebral hemisphere associated with disease (the two columns on the left) and the normal cerebral hemisphere (the two columns on the right). The generated images are evaluated from three aspects: (i) Retention degree of functional and metabolic information (pixel intensity information) in the PET image: The color intensities of DCHWT and Structure-Aware are obviously inferior to those of other fusion methods, resulting in the loss of functional and metabolic information in the PET image. By contrast, the fused images generated by FPDE, GTF, and MWGAN have brighter and stronger colors. The fused images of FPDE have a stiffer jagged outline compared with others. (ii) Retention degree of details of soft tissue structures (texture details information) in the MRI image: GTF and MWGAN shows less fuzziness than FPDE and FusionGAN, whereas FusionGAN and MWGAN have more detailed retention than GTF on the edges. DCHWT and Structure-Aware experience difficulty in recognizing some texture details due to the darker colors. (iii) Image naturalness: The images generated by the proposed MWGAN have the best performance in terms of naturalness.

2) QUANTITATIVE COMPARISONS

To obtain a more accurate evaluation of the experimental results, we introduced some fusion metrics to have an objective evaluation of the experimental results. To avoid the contingency of the experimental results, we introduce eight metrics, namely, entropy (EN) [32], mean gradient (MG), edge intensity (EI), peak signal to noise ratio (PSNR), correlation coefficient (CC) [33], spatial frequency (SF), standard deviation (SD) [34], and structural similarity index measure (SSIM), to evaluate the performances of different fusion methods. The eight metrics are defined as follows:

(1) EN: EN measures the amount of information contained in the fused image, which is a kind of metric based on information theory. It is mathematically defined as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l. \tag{20}$$

The number of all the gray levels is set as L, (L = 256 in our experiments), and p_l indicates the normalized histogram of corresponding gray level in the fused image. A large EN means that the fused image contains more information and shows better performance.

(2) MG: This metric can measure the amount of gradient information contained in the image. A large MG means that the method achieves better performance, which is defined as follows:

$$MG = \frac{\sum_{i=2}^{M}\sum_{j=2}^{N}\sqrt{X}}{(M-1)(N-1)}, \tag{21}$$

$$X = ((x_{i,j} - x_{i-1,j})^2 + (x_{i,j} - x_{i,j-1})^2)/2. \tag{22}$$

(3) EI: This metric reflects the gradient amplitude of the edge point, which is mathematically defined as follows:

$$EI = \frac{\sum_{i=2}^{M}\sum_{j=2}^{N}\sqrt{\nabla_x F(i,j)^2 + \nabla_y F(i,j)^2}}{(M-1)(N-1)}, \tag{23}$$

Algorithm 1 Training Procedure of MWGAN

Parameter descriptions:

N_G, N_M, N_I are the numbers of steps to train G, D_M, D_I.

$\mathcal{L}_{max}$, $\mathcal{L}_{min}$ and $\mathcal{L}_{Gmax}$ are applied to determining a range to uncollapse training.

$\mathcal{L}_{max}$ and $\mathcal{L}_{min}$ are for adversarial losses of G, D_M, and D_I.

$\mathcal{L}_{Gmax}$ is the total loss of G.

We set $L_{max} = 1.8$, $L_{min} = -1.8$ in the first batch empirically in our experiments.

Initialize θ_{D_M} and θ_{D_I} for D_M and D_I, and θ_G for G;
In each training iteration:
(1) **Train Discriminators** D_M **and** D_I:
- Sample m *MRI* patches $\{M^1, \cdots, M^m\}$ and m corresponding I_{PET} patches $\{I^1, \cdots, I^m\}$;
- Obtain generated data $\{G(M^1, I^1), \cdots, G(M^m, I^m)\}$;
- Update Discriminator parameters θ_{D_M} by RMSPropOptimizer to minimize $\mathcal{L}_{D_M}$ in (18); (**step I**)
- Update Discriminator parameters θ_{D_I} by RMSPropOptimizer to minimize $\mathcal{L}_{D_I}$ in (19); (**step II**)
- While $\mathcal{L}_{D_M} > \mathcal{L}_{max}$ *and* $N_M < 20$, repeat **step I**.
  $N_M \leftarrow N_M + 1$;
- While $\mathcal{L}_{D_I} > \mathcal{L}_{max}$ *and* $N_I < 30$, repeat **step II**.
  $N_I \leftarrow N_I + 1$;
(2) **Train Generator** G:
- Sample m *MRI* patches $\{M^1, \cdots, M^m\}$ and m corresponding I_{PET} patches $\{I^1, \cdots, I^m\}$;
- Obtain generated data $\{G(M^1, I^1), \cdots, G(M^m, I^m)\}$;
- Update Generator parameters θ_G by RMSPropOptimizer to minimize $\mathcal{L}_G$ in (11); (**step III**)
- While $\left(\mathcal{L}_{D_M} < \mathcal{L}_{min}$ *or* $\mathcal{L}_{D_I} < \mathcal{L}_{min}\right)$ *and* $N_G < 20$, repeat **step III**.
  $N_G \leftarrow N_G + 1$;
- While $\mathcal{L}_G > \mathcal{L}_{Gmax}$ *and* $N_G < 30$, repeat **step III**.
  $N_G \leftarrow N_G + 1$;

where ∇ is the Sobel matrix. A large EI means that the contrast is stronger and the algorithm has better fusion performance.

(4) PSNR: This metric can measure the distortion by the ratio of peak value power and noise power, which is defined as follows:

$$PSNR = 20log_{10}\frac{r}{\sqrt{MSE}}, \tag{24}$$

where r is set as 256 (peak value of the fused image), and the mean square error (MSE) reflects the dissimilarity between the source image and fused image, which is defined as follows:

$$MSE = (MSE_{AF} + MSE_{BF})/2, \tag{25}$$

$$MSE_{XF} = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(X_{i,j} - F_{i,j}\right)^2. \tag{26}$$

A large PSNR corresponds to less distortion. Also, the image is more similar to the source images, which results in better fusion performance.

(5) CC: CC is designed to measure the degree of linear correlation of the fused image and source images, which is defined as follows:

$$CC = \frac{1}{2}(r_{AF} + r_{BF}), \tag{27}$$

where

$$r_{XF} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j) - \overline{X})(F(i,j) - \overline{F})}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j) - \overline{X})^2 \sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j) - \overline{F})^2}}, \tag{28}$$

where $\overline{X}$ and $\overline{F}$ denote the mean values of the source image A or B and the fused image F, respectively. A large CC reflects that the fused image is more similar to the source images.

(6) SF: This metric effectively reflects the texture details of an image by calculating the gradient distribution, which is composed of spatial row frequency (RF) and column frequency (CF). It is mathematically defined as follows:

$$SF = \sqrt{RF^2 + CF^2}, \tag{29}$$

$$RF = \sqrt{\sum_{i=1}^{M}\sum_{j=2}^{N}(x_{i,j} - x_{i,j-1})^2}, \tag{30}$$

$$CF = \sqrt{\sum_{i=2}^{M}\sum_{i=1}^{N}(x_{i,j} - x_{i-1,j})^2}. \tag{31}$$

Richer edges and texture details are represented by a large SF.

(7) SD: SD reflects the extent to which the values of individual pixels in the image deviate from the average value.

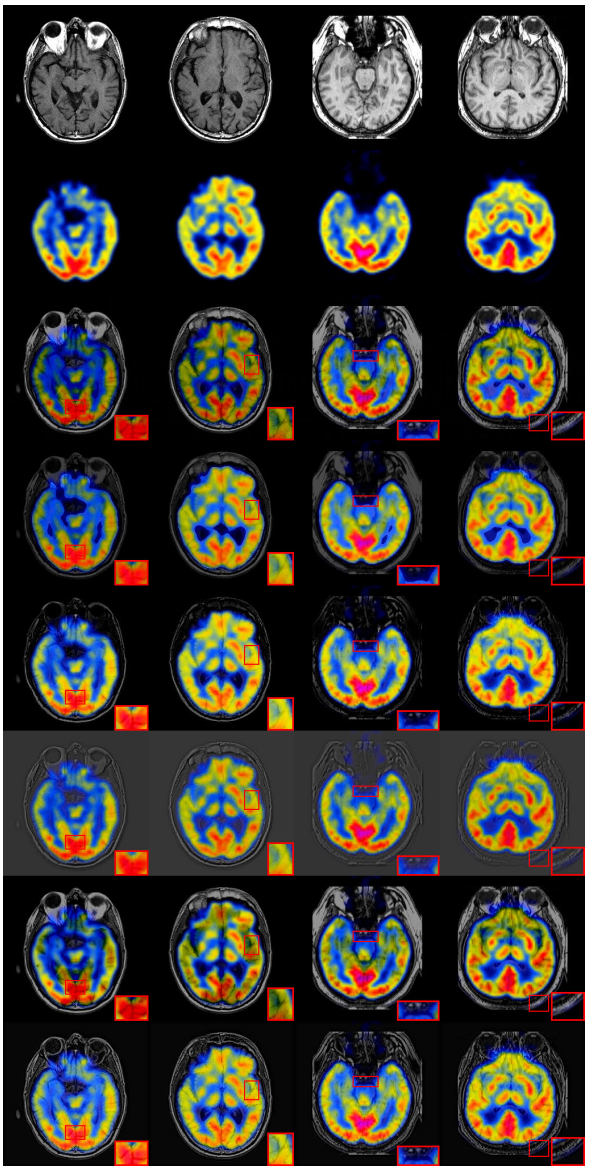

FIGURE 6. Qualitative comparison of our MWGAN with 5 state-of-the-art methods. From top to bottom: high-resolution MRI image, low-resolution PET image, fused images of DCHWT [30], FPDE [2], GTF [7], FusionGAN [13], Structure-Aware [31], and our MWGAN.

SD is mathematically defined as follows:

$$SD = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(F_{i,j} - \overline{F})^2}, \tag{32}$$

where F and $\overline{F}$ indicate the fused image with dimensions of $M \times N$ and the mean value of the fused image F, respectively. A large SD means that the attention of humans is more likely to be attracted by the high-contrast area, which leads to a better visual effect.

(8) SSIM: SSIM measures the structural similarity between the fused image and the source images, which is defined as follows:

$$SSIM = SSIM_{A,F} + SSIM_{B,F}, \tag{33}$$

where $SSIM_{X,F}$ is the structural similarity between fused image F and source image X. The $SSIM_{X,F}$ is mathematically defined as follows:

$$SSIM_{X,F} = \sum_{X,F} \frac{2\mu_x\mu_f + C_1}{\mu_x^2 + \mu_f^2 + C_1} \cdot \frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2 + \sigma_f^2 + C_2} \cdot \frac{\sigma_{x,f} + C_3}{\sigma_x\sigma_f + C_3}. \tag{34}$$

A large SSIM indicates that the algorithm achieves better performance in structural similarity, which means better fusion performance.

The quantitative comparisons of the eight fusion metrics on two sets of image pairs from the Harvard medical school website are illustrated in Fig. 7. The first 10 pairs of MRI and PET images in one set are from the cerebral hemisphere associated with disease, whereas the last 10 pairs of MRI and PET images in the other one set are from the normal cerebral hemisphere. In terms of the metrics of EN, MG, EI, PSNR, CC, and SF, our MWGAN achieved the largest

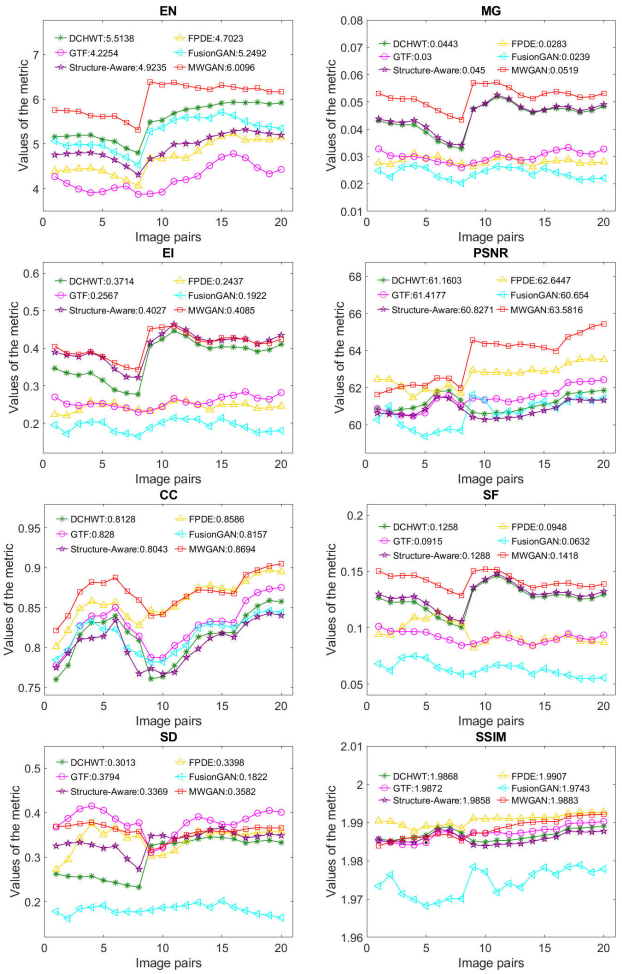

FIGURE 7. Quantitative comparisons of the eight metrics. Mean values of the metrics are presented in the legends.

TABLE 1. Average runtime comparison of different methods on the fusion of MRI and PET images (S-A: Structure-Aware, unit: second).

	DCHWT	FPDE	GTF	FusionGAN	S-A	MWGAN
Mean	0.80	0.14	0.23	0.09	0.03	0.46
STD	0.04	0.01	0.04	0.20	0.03	0.22

mean values, indicating that our MWGAN has the greatest amount of information of fused image, gradient information of fused image, contrast, similarity between the fused image and source images, linear correlation of the fused image and source images, and texture details. Although the mean values of the SD and SSIM are not the largest, they still achieved the second largest mean values among the six fusion methods; these mean values are 0.02123 and 0.0024, which are lower than those of GTF and FPDE, respectively. The results also demonstrate that our MWGAN performs better in terms of SD and SSIM.

We also compare the average runtime of different methods on the testing data. The result is shown in Table 1. FusionGAN and MWGAN are tested on GPU, and the others are tested on CPU. Our MWGAN achieves satisfying efficiency.

3) CLIPPED ANALYSIS OF G's PARAMETERS

In the process of the generating fused images, the existence of some discrete points will result in obvious black point-like noise, which can affect the visual effect of the generated image. The discrete points can be effectively controlled in a reasonable interval to reduce the impact of noise on visual effect by clipping the absolute value of G's parameters to a fixed parameter, as shown in Fig. 8. The black point-like noise in the fused images with clipping is effectively reduced, thereby exhibiting an enhanced visual effect.

4) STABILITY ANALYSIS

Since there is no activation function in the last layer of discriminator in our MWGAN, its value ranges $(-\infty, +\infty)$. As shown in the Fig. 9, the values of D_M fake (fused image) and D_M real (MRI image) are both controllable and close, which proves that our MWGAN is a stable model.

D. FUSION RESULTS ON MRI AND CT IMAGES

CT images are generally applicated in precise localization of dense structures such as bones and implants, which in line with the essence of our MWGAN fusion. To further verify the suitability of our MWGAN in the fusion of MRI and CT images with different resolutions, we acquired 20 pairs of 256×256 MRI and CT images from the Harvard Medical School website. We cropped these images into 2, 304 84×84 patch pairs to serve as our training set. Before training, we downsampled the CT patches to 21×21 to fuse the MRI and CT images with different resolutions. The training and testing procedures are the same as the fusion of MRI and PET images, in which the CT image is regarded as the I_{PET} image. Aside from DCHWT, Structure-Aware, and GTF, we added DDCTPCA [35] for multisensor image fusion and ASR [36] for the image fusion and denoising of multi-focus and multimodal images. The result of qualitative comparisons is shown in Fig. 10. From the result, we see that except for GTF and MWGAN, the fused images of other methods are far from retaining the pixel intensity information

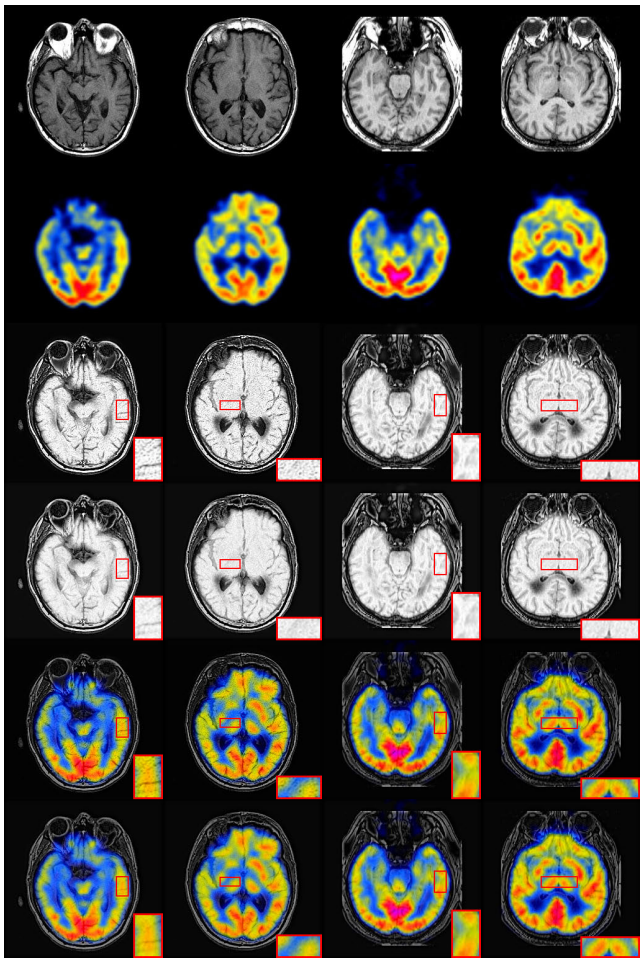

FIGURE 8. Qualitative comparison of the results with and without clipping. From top to bottom: high-resolution MRI image, low-resolution PET image, I_F images without clipping, I_F images with clipping, final fused images without clipping, final fused images with clipping.

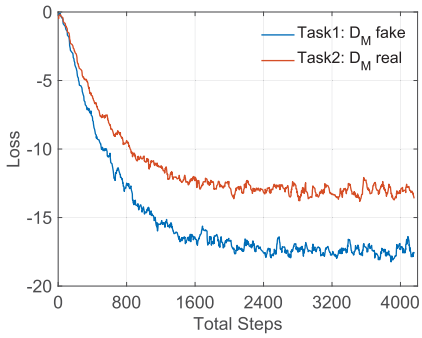

FIGURE 9. The loss of D_M fake and D_M real.

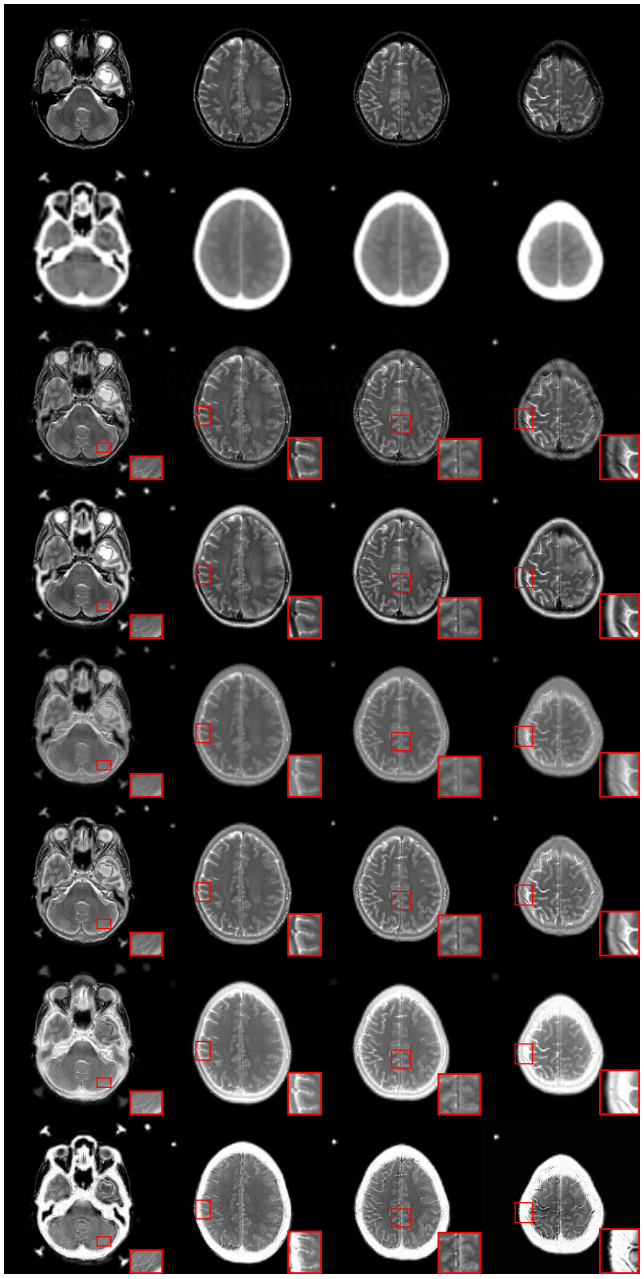

FIGURE 10. Qualitative comparisons of our MWGAN with 5 state-of-the-art methods. From top to bottom: high-resolution MRI image, low-resolution CT image, fused images of DCHWT, Structure-Aware, DDCTPCA, ASR, GTF and our MWGAN.

of the CT image. Our MWGAN maintained excellent performance in the texture details. Thus, our MWGAN can also be applied to the fusion of MRI and CT images with different resolutions.

The average runtime of different methods on the testing data is also compared, and the result is shown in Table 2, where the MWGAN is tested on GPC and others are tested on CPU. Our MWGAN achieves comparable efficiency compared with other methods.

TABLE 2. Average runtime comparison of different methods on the fusion of MRI and CT images (S-A: Structure-Aware, unit: second).

	DCHWT	S-A	DDCTPCA	ASR	GTF	MWGAN
Mean	0.79	0.03	130.28	21.18	0.28	0.46
STD	0.02	0.01	1.15	0.88	0.04	0.22

V. CONCLUSION

In this paper, MWGAN, a novel WGAN-based algorithm that fuses MRI and PET medical images of different resolutions, is proposed. The proposed MWGAN is more stable than the algorithm based on traditional GAN and includes a new loss function setting strategy. Our proposed MWGAN is an end-to-end model with deep learning implemented throughout the entire fusion process. It can simultaneously retain the details of soft tissue structures in organs (texture detail information) from a high-resolution MRI image and the functional and metabolic information (pixel intensity information) from a low-resolution PET image without the loss of texture details and pixel intensity information. Quantitative comparisons with five state-of-the-art fusion methods on eight evaluation metrics confirm that MWGAN not only has better visual effects but can also retain approximately the largest amount of information of the source images. In addition, MWGAN achieves excellent performance in fusing MRI and CT images with different resolutions. In our future work, we will further apply our MWGAN to the fusion of more kinds image pairs and achieve better visual effects.

REFERENCES

[1] S. Daneshvar and H. Ghassemian, "MRI and PET image fusion by combining IHS and retina-inspired models," *Inf. Fusion*, vol. 11, no. 2, pp. 114–123, 2010.

[2] D. P. Bavirisetti, G. Xiao, and G. Liu, "Multi-sensor image fusion based on fourth order partial differential equations," in *Proc. 20th Int. Conf. Inf. Fusion*, Jul. 2017, pp. 1–9.

[3] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.

[4] J. Du, W. Li, B. Xiao, and Q. Nawaz, "Union Laplacian pyramid with multiple features for medical image fusion," *Neurocomputing*, vol. 194, pp. 326–339, Jun. 2016.

[5] X. Zhang, Y. Ma, F. Fan, Y. Zhang, and J. Huang, "Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 34, no. 8, pp. 1400–1410, 2017.

[6] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 12, pp. 3450–3459, Dec. 2012.

[7] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.

[8] H. Guo, Y. Ma, X. Mei, and J. Ma, "Infrared and visible image fusion based on total variation and augmented lagrangian," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 34, no. 11, pp. 1961–1968, 2017.

[9] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[10] J. Du, W. Li, K. Lu, and B. Xiao, "An overview of multi-modal medical image fusion," *Neurocomputing*, vol. 215, pp. 3–20, Nov. 2016.

[11] A. Dogra, B. Goyal, and S. Agrawal, "From multi-scale decomposition to non-multi-scale decomposition methods: A comprehensive survey of image fusion techniques and its applications," *IEEE Access*, vol. 5, pp. 16040–16067, 2017.

[12] Z. Zhu, M. Zheng, G. Qi, D. Wang, and Y. Xiang, "A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain," *IEEE Access*, vol. 7, pp. 20811–20824, 2019.
[13] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
[14] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 3954–3960.
[15] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Inf. Sci.*, vols. 433–434, pp. 125–141, Apr. 2018.
[16] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2017, pp. 4714–4722.
[17] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.
[18] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.
[19] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
[20] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. Int. Conf. Inf. Fusion*, Jul. 2017, pp. 1–7.
[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[22] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," 2015, *arXiv:1511.06434*. [Online]. Available: https://arxiv.org/abs/1511.06434
[23] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 214–223.
[24] P. Ganasala, V. Kumar, and A. D. Prasad, "Performance evaluation of color models in the fusion of functional and anatomical images," *J. Med. Syst.*, vol. 40, no. 5, p. 122, 2016.
[25] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," *IEEE Trans. Circuits Syst. Video Technol.*, to be published.
[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
[27] J. Ma, J. Jiang, C. Liu, and Y. Li, "Feature guided Gaussian mixture model with semi-supervised EM and local geometric constraint for retinal image registration," *Inf. Sci.*, vol. 417, pp. 128–142, Nov. 2017.
[28] J. Ma, J. Zhao, J. Jiang, H. Zhou, and X. Guo, "Locality preserving matching," *Int. J. Comput. Vis.*, vol. 127, no. 5, pp. 512–531, 2019.
[29] J. Ma, X. Jiang, J. Jiang, J. Zhao, and X. Guo, "LMR: Learning a two-class classifier for mismatch removal," *IEEE Trans. Image Process.*, vol. 28, no. 8, pp. 4045–4059, Aug. 2019.
[30] B. K. S. Kumar, "Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform," *Signal, Image Video Process.*, vol. 7, no. 6, pp. 1125–1143, 2013.
[31] W. Li, Y. Xie, H. Zhou, Y. Han, and K. Zhan, "Structure-aware image fusion," *Optik*, vol. 172, pp. 1–11, Nov. 2018.
[32] J. W. Roberts, F. B. Ahmed, and J. A. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, 2008, Art. no. 023522.
[33] M. M. Mukaka, "A guide to appropriate use of correlation coefficient in medical research," *Malawi Med. J.*, vol. 24, no. 3, pp. 69–71, 2012.
[34] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[35] V. P. S. Naidu, "Hybrid DDCT-PCA based multi sensor image fusion," *J. Opt.*, vol. 43, no. 1, pp. 48–61, 2014.
[36] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Process.*, vol. 9, no. 5, pp. 347–357, 2015.

ZHIGUANG YANG received the master's degree from the School of Mechanical Engineering, Wuhan Polytechnic University, Wuhan, China, in 2017. He is currently pursuing the Ph.D. degree with the School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan. His current research interests include computer vision and machine learning.

YOUPING CHEN received the B.S. and M.S. degrees in mechanical engineering from Shanghai Jiaotong University, in 1982 and 1984, respectively, and the Ph.D. degree in mechanical engineering from the Huazhong University of Science and Technology, in 1990. He is currently a Professor with the School of Mechanical Science and Engineering, Huazhong University of Science and Technology. His research interests include digital manufacturing, CNC system, intelligent control, and machine vision and online monitoring.

ZHULIANG LE received the B.E. degree from the School of Information Engineering, Wuhan University of Technology, Wuhan, China, in 2019. He is currently pursuing the master's degree with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

FAN FAN received the B.S. degree in communication engineering and the Ph.D. degree in electronic circuit and system from the Huazhong University of Science and Technology, Wuhan, China, in 2009 and 2015, respectively. He currently holds a postdoctoral position at the School of Remote Sensing and Information Engineering, Wuhan University, China. His current research interests include infrared thermal imaging, machine learning, and computer vision.

ERTING PAN received the B.S. degree in electrical engineering and its automation from Northeast Normal University, Changchun, China, in 2018. She is currently pursuing the M.S. degree with the Electronic Information School, Wuhan University, Wuhan. Her current research interests include hyperspectral imagery and deep learning.