DDcGAN: A Dual-Discriminator Conditional Generative Adversarial Network for Multi-Resolution Image Fusion

Jiayi Ma, Han Xu, Junjun Jiang, Xiaoguang Mei, and Xiao-Ping Zhang, *Senior Member, IEEE*

***Abstract*— In this paper, we proposed a new end-to-end model, termed as dual-discriminator conditional generative adversarial network (DDcGAN), for fusing infrared and visible images of different resolutions. Our method establishes an adversarial game between a generator and two discriminators. The generator aims to generate a real-like fused image based on a specifically designed content loss to fool the two discriminators, while the two discriminators aim to distinguish the structure differences between the fused image and two source images, respectively, in addition to the content loss. Consequently, the fused image is forced to simultaneously keep the thermal radiation in the infrared image and the texture details in the visible image. Moreover, to fuse source images of different resolutions, *e.g.*, a low-resolution infrared image and a high-resolution visible image, our DDcGAN constrains the downsampled fused image to have similar property with the infrared image. This can avoid causing thermal radiation information blurring or visible texture detail loss, which typically happens in traditional methods. In addition, we also apply our DDcGAN to fusing multi-modality medical images of different resolutions, *e.g.*, a low-resolution positron emission tomography image and a high-resolution magnetic resonance image. The qualitative and quantitative experiments on publicly available datasets demonstrate the superiority of our DDcGAN over the state-of-the-art, in terms of both visual effect and quantitative metrics. Our code is publicly available at https://github.com/jiayi-ma/DDcGAN.**

***Index Terms*— Image fusion, generative adversarial network, infrared image, medical image, different resolutions.**

I. Introduction

INFRARED and visible image fusion has been gaining in popularity in image signal processing due to its extensive applications in many fields such as computer vision, remote sensing, medical imaging, and military detection [1], [2]. Among these sensors, infrared and visible sensors are probably the most widely used type of sensors with wavelengths of 8-14 μm [3] and 300-530 nm [4] respectively. The uniqueness of the combining infrared and visible sensors lies in the fact that visible sensors capture reflected light to represent abundant texture details whereas infrared sensors map captured thermal radiation to gray images and can highlight thermal targets even in poor lighting conditions or under the circumstances of severe occlusion. Due to the strong complementarity between them, the fused result has the potential to present nearly all the inherent properties of the target to improve visual understanding [5]. Therefore, their fusion plays an important role in military and civilian applications [6], [7].

For multi-modality source images, the key of image fusion is to extract the most important feature information in source images taken from different imaging apparatus and merge it into a single fused image [8]. Therefore, the fused image can provide more complex and detailed scene representation while reducing redundant information. For this purpose, many fusion methods have been proposed in the past decades. According to corresponding schemes, these fusion methods can be divided into different categories, including multi-scale transform-based methods [9], [10], sparse representation-based methods [11], [12], neural network-based methods [13], subspace-based methods [14], saliency-based methods [15], hybrid methods [16], and other fusion methods [17], [18]. These methods are dedicated to design feature extraction and fusion rules in a manual way for better fusion performance. However, detailed and diverse feature extraction and fusion rule design make the fusion method more and more complex.

Since much attention has been drawn to deep learning recently, some deep learning-based fusion methods have been proposed. The detailed exposition of deep learning-based fusion methods will be discussed later in Sec. II-A. Although these works have achieved promising performance, there are still some drawbacks: (i) The deep learning framework is only applied in some part of the fusion process, *e.g.*, to extract features, while the overall fusion process is still in traditional frameworks [19], [20]. (ii) Faced with the lack of ground-truth, the solutions by merely designing loss functions are incomprehensive and inappropriate. (iii) The fusion rules designed in a manual way enforce the extraction of same features even if source images are multi-modality data. (iv) In existing fusion

Manuscript received December 10, 2018; revised September 30, 2019; accepted February 26, 2020. Date of current version March 12, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61773295, Grant 61903279, and Grant 61971165, and in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) under Grant RGPIN239031. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Jana Ehmann. *(Corresponding author: Xiaoguang Mei.)*

Jiayi Ma, Han Xu, and Xiaoguang Mei are with the Electronic Information School, Wuhan University, Wuhan 430072, China (e-mail: jyma2010@gmail.com; meixiaoguang@gmail.com; xu_han@whu.edu.cn).

Junjun Jiang is with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: junjun0595@163.com).

Xiao-Ping Zhang is with the Department of Electrical, Computer and Biomedical Engineering, Ryerson University, Toronto, ON M5B 2K3, Canada (e-mail: xzhang@ee.ryerson.ca).

Digital Object Identifier 10.1109/TIP.2020.2977573

methods based on traditional generative adversarial network (GAN) [21], [22], the fused image is trained to be similar to only one of the source images, leading to the loss of some information contained in the other source image.

Furthermore, due to the limitations of hardware and environments, the infrared images always suffer from low resolution and blurred details compared with corresponding visible images, and it is hard to improve the resolution of infrared images by upgrading hardware devices. For fusing multi-resolution infrared and visible images (*e.g.*, images of different resolutions), the strategy of downsampling visible images or upsampling infrared images before fusion will inevitably causes thermal radiation information blurring or visible texture detail loss. Therefore, it remains a challenging task to fuse multi-resolution infrared and visible images without loss of important information.

To address the above challenges, in this work, we propose a fusion method via dual-discriminator conditional generative adversarial network (DDcGAN). The problem is formulated as a particular adversarial process of two kinds of neural networks, *i.e.*, a generator and two discriminators, based on conditional GAN [23]. We adapt the architecture to dual-discriminators and the discriminators are pulling each other on the distribution of the generated data obtained by the generator, so that the fused image simultaneously keeps the most important feature information in infrared and visible images. We utilize source images as the real data and the fused image should be indistinguishable with both types of real images, and hence the ground-truth fused image is not required in our model. The entire network is an end-to-end model without the requirement of designing fusion rules. Moreover, our model can be generalized to fuse source images of different resolutions. In particular, we constrain the downsampled fused image to have similar properties with the infrared image, and utilize trainable deconvolution layers to learn a mapping between different resolutions. Last but not least, our proposed method can also be generalized to solve the medical image fusion problem, *e.g.*, positron emission tomography (PET) and magnetic resonance image (MRI) fusion, which can preserve the functional information and the anatomical information to a great extent in the fused image. Extensive results have revealed the advantages of our DDcGAN compared to other methods.

The major contributions of our work include the following four aspects. Firstly, our proposed method has contributed to applying a deep learning framework based on minmax two-player game to the overall fusion process of multi-modality images rather than just some sub parts of them. Secondly, the dual-discriminators architecture enables the generator to be more adequately trained to meet stricter requirements and avoid information loss caused by the introduction of discriminator on only one type of source images. Thirdly, in virtue of the utilization of trainable deconvolution layers and content constraints on downsampled fused images, our proposed method demonstrates better performances for multi-resolution source image fusion. Lastly, our method can also be extended to the fusion of medical images such as MRI and PET image fusion and achieves advantageous performances.

A preliminary version of this manuscript has appeared in [24]. The primary new contributions include the following five aspects. First, the generator network architecture is optimized, where we replace the U-net with the densely connected convolutional network. In virtue of the dense connections, the network architecture can strength the transmission of feature maps and make use of them more effectively. Without the loss caused by the large stride and the blur caused by the upsampling operations, the information in source images is preserved to a greater extent for clearer fusion performance. Second, the input of the discriminator D_v is no longer the gradients of image to be distinguished but the image itself. By expanding the probability space from the subspace of source images to the whole images, the fused images can have more similar properties with source images. When the network tries to minimize the divergence of different probability distributions in the subspace, it will introduce some additional noise into the source images. By expanding the probability space, the influence can be mitigated. Third, as for the input of the generator, *i.e.*, different-resolution source images, instead of upsampling the low-resolution source image with two upsamping layers, we employ a deconvolution layer to learn a mapping from low to high resolution. The difference is that the parameters in this layer are obtained during the training phase rather than pre-defined. And the high-resolution source image is fed into another deconvolution layer to generate same-resolution feature maps. Fourth, we add more detailed analysis experiments related to the generator and two discriminators to verify the effects of their subparts. Last, we apply the proposed method to fuse different-resolution multi-modal medical images, *i.e.*, low-resolution PET images and high-resolution MRI images, and compare our fused results with state-of-the-art methods qualitatively and quantitatively.

The remainder of this paper is organized as follows. Section II describes some related work, including an overview of existing deep learning-based fusion methods and a theoretical introduction of GANs. Section III provides the problem formulation, loss functions and network architecture design. In Section IV, our proposed method is generalized to fuse medical images. In Section V, we compare our method with several state-of-the-art methods on publicly available datasets by qualitative and quantitative comparisons both for infrared and visible image fusion and PET and MRI image fusion. The experiments of discriminator analysis are also conducted in this section. Conclusions are given in Section VI.

II. Related Work

In this section, we give a brief introduction of the existing deep learning-based image fusion methods. In addition, since our method is based on the GANs, we also provide a brief explanation of its basic theory and an improved network, namely conditional GAN.

A. Deep Learning-Based Fusion Methods

Since the study based on deep learning has become an active topic in the field of image fusion in the last three years [25], many deep learning-based fusion methods have been proposed and gradually formed a critical branch. In some methods, the deep learning framework is applied to extract image features in an end-to-end manner for reconstruction. Representatively, Liu *et al.* [19] applied the convolutional sparse representation (CSR) for image fusion, which is employed to extract multi-layer features and these features are used to generate the fused image. In [26], Liu *et al.* proposed a medical image fusion method based on convolutional neural networks (CNNs). The convolutional network is merely adopted to generate a weight map which integrates the pixel activity information and the overall fusion process is still conducted in a multi-scale mannar via image pyramids in a traditional way. In [20], Li *et al.* decomposed the source images into base parts and detail content. The deep learning framework is used to extract multi-layer features in the detail content while the base parts are fused by weighted-averaging. Then, the two parts are combined for reconstruction.

In other methods, the deep learning framework is used not only for feature extraction but also for reconstruction. For instance, based on a three-layer architecture for super-resolution, Masi *et al.* [27] proposed a convolutional neural network for projection, mapping, and reconstruction to solve pansharpening problem. Prabhakar *et al.* [28] proposed an unsupervised deep learning framework for multi-exposure fusion. They utilized a novel CNN architecture and designed a no-reference quality metric as the loss function. As weights are tied, the pre-fusion layers are forced to learn the same features and these features are added for fusion. On this basis, Li *et al.* [29] improved the architecture by introducing dense block. In the fusion layer, salient feature maps are combined by two manually designed fusion strategies (addition and ℓ_1-norm). Similarly, it utilizes no-reference metrics (the structural similarity index measure and the Euclidean distance) as the loss function for unsupervised learning. In our previous work [21], we proposed the FusionGAN to fuse infrared and visible images using a generative adversarial network. The fused image generated by the generator is forced to have more details existing in the visible image by applying the discriminator to distinguish differences between them. When fusing source images with different resolutions, the low-resolution infrared images are simply interpolated before fed into the generator.

Although the abovementioned works have achieved promising performance, there are still some drawbacks in existing deep learning-based fusion methods. (i) Existing methods typically perform neural network in feature extraction and reconstruction while fusion rules are still designed in a manual way. Thus, the entire method can not get rid of the limitations of traditional fusion methods. (ii) The major stumbling block in utilizing deep learning for infrared and visible image fusion is the lack of ground-truth fused image for supervised learning. Existing methods solve it by designing loss function to penalize differences between output and target in some aspects. However, these metrics will introduce new problems while penalizing certain aspects. For instance, the Euclidean distance suffers from relatively blurred results by averaging all plausible outputs [30]. Therefore, it remains to be difficult to design a comprehensive, appropriate and adaptive loss function to specify a high-level goal. (iii) Most artificially designed fusion rules lead to the extraction of same features for different types of source images, regardless of the fact that source images are manifestations of different phenomena and it is inappropriate for multi-source image fusion. (iv) Existing GAN-based fusion method merely applies GAN to force the fused image to obtain more details in visible images while the thermal radiation in infrared images is only obtained through the content loss. As the adversarial game proceeds, the fused image is more similar to the visible image and the prominence of thermal targets is gradually reduced.

To address the problems, we solve the fusion problem by applying GAN and adapt it with dual discriminators. On this basis, we introduce the deconvolution layers to adapt to the fusion of source images of different resolutions. In addition, for the stability of the training process, we optimize the network architecture and the training strategy.

B. Generative Adversarial Networks

Generative adversarial networks is one of the generative models. If samples are drawn from the real distribution $P_{\text{data}}(x)$, the generative model is designed to learn a probability distribution $P_{\text{model}}(x;\theta)$ parameterized by θ as an estimation of $P_{\text{data}}(x)$ from samples $\{x^1, x^2, \cdots, x^m\}$, where $P_{\text{model}}(x;\theta)$ are Gaussian mixture models. Likelihood of generating the samples is defined as follows:

$$L = \prod_{i=1}^{m} P_{\text{model}}\left(x^i;\theta\right). \tag{1}$$

Then we can perform maximum likelihood estimation [31]:

$$\theta^* = \arg\max_{\theta} \sum_{i=1}^{m} \log P_{\text{model}}\left(x^i;\theta\right). \tag{2}$$

It can be thought of as minimizing the Kullback-Liebler divergence between $P_{\text{data}}(x)$ and $P_{\text{model}}(x;\theta)$. However, if P_{model} is a much more complicated probability distribution, it will be quite difficult to calculate its likelihood function to perform maximum likelihood estimation. To deal with it, GANs estimate generative models via an adversarial process by simultaneously training two models: a generative model G and a discriminator model D [32].

The generator G is a network that can capture the data distribution and generate new samples. If we input the noise z sampled from the latent space, it generates a sample $x = G(z)$. In virtue of neural networks, the probability distribution $P_G(x)$ formed by generated samples has the ability to be much more complicated. The training objective of G is to make $P_G(x)$ and $P_{\text{data}}(x)$ as close as possible and the optimization formulation can be defined as:

$$G^* = \arg\min_{G} Div\left(P_G(x), P_{\text{data}}(x)\right), \tag{3}$$

where $Div(\cdot)$ denotes the divergence between two distributions. However, it is difficult to calculate the divergence because the formulations of P_G and P_{data} are unknown.

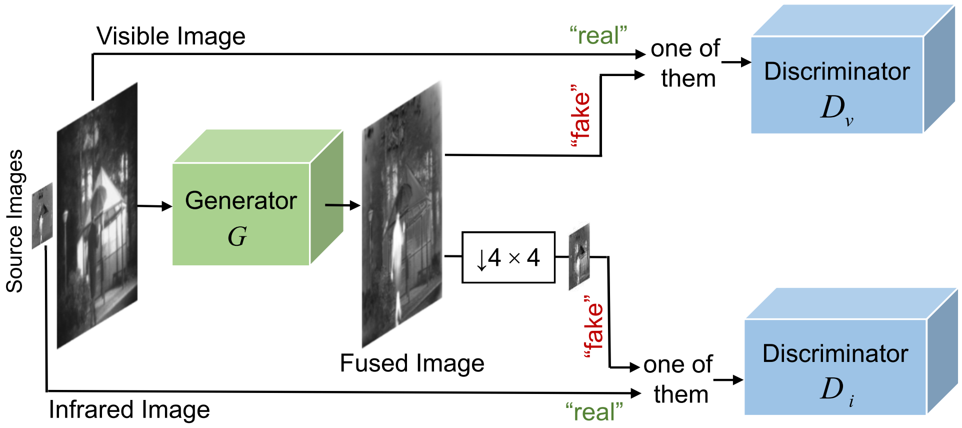


Fig. 1. The entire procedure of our DDcGAN for image fusion.

Ingeniously, the discriminator D can be used to solve this problem for it estimates the probability that a sample comes from the training data rather than G. The objective function for D can be formulated as:

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

where $V(G, D)$ is defined as follows:

$$V(G, D) = \mathbb{E}_{x\sim P_{\text{data}}}\left[\log D(x)\right] + \mathbb{E}_{x\sim P_G}\left[\log(1 - D(x))\right]. \tag{5}$$

A large objective value means that the Jensen-Shannon (JS) divergence of P_G and P_{data} is large and they are easy to discriminate. Thus, the optimization formulation of G can be converted to:

$$G^{*} = \arg\min_{G}\max_{D} V(G, D), \tag{6}$$

where the discriminator D is fixed when we are training G. The adversarial process of G and D makes up the two-player min-max game where G tries to fool D while D is trained to discriminate the generated data. Hence, the generated samples are getting more and more indistinguishable from the real data.

GANs can be extended to a conditional model if both the generator and discriminator are conditioned on some extra information which could be any kind of auxiliary information. We can perform the conditioning by feeding the extra information as additional input layer and this model is defined as conditional generative adversarial networks [23].

III. Proposed Method

In this section, with analysis of the characteristics of infrared and visible images, we provide our fusion formulation, the definition and design of loss functions. At the end of this section, the design of network architecture is shown concretely.

A. Problem Formulation

We formulate the fusion problem as a conditional GAN model by constructing a dual-discriminator conditional GAN. To fuse images of different resolutions, without loss of generality, we make an assumption that the resolution of the visible image υ is 4 × 4 times that of the infrared image i.

The entire procedure of our proposed DDcGAN is shown in Fig. 1. Given a visible image υ and an infrared image i, our ultimate goal is to learn a generator G conditioned on them and the generated image $G(\upsilon, i)$ is encouraged to be realistic and informative enough to fool the discriminators. Simultaneously, we exploit two adversarial discriminators D_υ and D_i, and they respectively generate a scalar that estimates the probability of the input from real data rather than G. Specifically, D_υ aims to distinguish the generated image from the visible image, while D_i is trained to discriminate between the original low-resolution infrared image and down-sampled generated/fused image. Average-pooling is employed here for downsampling due to its retention of low-frequency information compared with max-pooling and the thermal radiation information is mainly presented in this form. Put slightly differently, for the sake of the balance between the generator and discriminators, except for the input of discriminators, we do not feed the source images υ and i as additional/conditional information to D_υ and D_i. That is, the input layer of each discriminator is a single-channel layer containing the sampled data rather than a two-channel layer containing both the sampled data and the corresponding source image as the conditional information. Because when the condition and the sample to be discriminated are the same, the discrimination task is simplified to judge whether the input images are the same and it is a simple enough task for neural networks. When the generator is unable to fool the discriminator, the adversarial relationship will fail to be established and the generator will tend to generate randomly. Consequently, the model will lose its original meaning.

We denote the downsampling operator as ψ, which is implemented by two average pooling layers due to its retention of low frequency information. Both layers summarize a 3 × 3 neighborhood and use a stride of 2. Accordingly, the training target of G can be formulated as minimizing the following adversarial objective

$$\begin{aligned}\min_{G}\max_{D_\upsilon, D_i} \mathbb{E}\left[\log D_\upsilon(\upsilon)\right] + \mathbb{E}\left[\log(1 - D_\upsilon(G(\upsilon, i)))\right] \\ +\mathbb{E}\left[\log D_i(i)\right] + \mathbb{E}\left[\log(1 - D_i(\psi G(\upsilon, i)))\right].\end{aligned} \tag{7}$$

Conversely, the goal of discriminators is to maximize Eq. (7).

Through the adversarial process of the generator G and two discriminators (D_v and D_i), the divergence between P_G and two real distributions, *i.e.*, P_V and P_I, will become smaller simultaneously, where P_G is the probability distribution of the generated samples, P_V is the real distribution of the visible images and P_I is that of the infrared images.

B. Loss Function

Initially, the success of GANs was limited as they were known to be unstable to train and may result in artifacts and noisy or incomprehensible results [33]. A possible solution to solve the problem of artifacts and incomprehensible results is to introduce a content loss to include a set of constraints into the networks. Thus, in this paper, the generator is not only trained to fool discriminators but also tasked to constraint similarity between the generated image and source images in the content. Therefore, the loss function of the generator is composed by an adversarial loss $\mathcal{L}_G^{\text{adv}}$ and a content loss $\mathcal{L}_{\text{con}}$, with a weight λ controlling the trade-off:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda \mathcal{L}_{\text{con}}, \tag{8}$$

where $\mathcal{L}_G^{\text{adv}}$ comes from the discriminators and is defined as:

$$\mathcal{L}_G^{\text{adv}} = \mathbb{E}\left[\log\left(1 - D_v\left(G\left(v, i\right)\right)\right)\right] + \mathbb{E}\left[\log\left(1 - D_i\left(\psi G\left(v, i\right)\right)\right)\right]. \tag{9}$$

As the thermal radiation and texture details are mainly characterized by pixel intensities and gradient variation [17], respectively, we employ the Frobenius norm to constrain the downsampled fused image to have similar pixel intensities with the infrared image as the data fidelity term. By constraining the relationship of pixel intensities of downsampled fused image and the low-resolution infrared image, we can considerably prevent loss of texture information caused by compression or blur and inaccuracy due to forced upsampling. According to the aforementioned constraint, the thermal target remains prominent in the fused image. The TV norm [34] is applied in the regularization term to constrain the fused image to exhibit similar gradient variation with the visible image. Compared with the ℓ_0 norm, the TV norm is able to solve the non-deterministic polynomial-time hard problem effectively. With a weight η to tradeoff the differences of pixel intensities and gradient variation, we can obtain the content loss:

$$\mathcal{L}_{\text{con}} = \mathbb{E}\left[\|\psi G\left(v, i\right) - i\|_F^2 + \eta \|G\left(v, i\right) - v\|_{TV}\right]. \tag{10}$$

The discriminators in DDcGAN, *i.e.,* D_v and D_i, play a role of discriminating between source images and the generated fused image. The adversarial losses of discriminators can calculate the JS divergence between distributions and thus identify whether the intensity or texture information is unrealistic and thus encourage matching the realistic distribution. The adversarial losses are defined as follows:

$$\mathcal{L}_{D_v} = \mathbb{E}\left[-\log D_v\left(v\right)\right] + \mathbb{E}\left[-\log\left(1 - D_v\left(G\left(v, i\right)\right)\right)\right], \tag{11}$$

$$\mathcal{L}_{D_i} = \mathbb{E}\left[-\log D_i\left(i\right)\right] + \mathbb{E}\left[-\log\left(1 - D_i\left(\psi G\left(v, i\right)\right)\right)\right]. \tag{12}$$

C. Network Architecture

1) Generator Architecture: The generator consists of 2 deconvolution layers, an encoder network and a corresponding decoder network, as presented in Fig. 2. Since the infrared image has a lower resolution, we firstly employ a mapping before encoding. Rather than simple interpolation by the nearest, bilinear or bicubic method, we introduce a deconvolution layer [35] to learn a mapping from low to high resolution. Without defining an upsampling operator, this mapping is different from traditional upsampling and its parameters are obtained automatically by training. The output of the deconvolution layer is a high-resolution feature map rather than an upsampled infrared image. We also pass the visible image through an independent deconvolution layer which generates a feature map with the same resolution. Results obtained by deconvolution layers are concatenated and fed as the input of the encoder. The process of feature extraction and fusion are both performed in the encoder and fused feature maps are produced as the output. These maps are then fed to the decoder for reconstruction and the generated fused image is of the same resolution with the visible image.

The encoder consists of 5 convolutional layers and each layer can obtain 48 feature maps by 3 × 3 filters. To mitigate the vanish of gradient, remedy feature loss and reuse previously computed features, *DenseNet* [36] is applied and short direct connections are built between each layer and all layers in a feed-forward fashion. The decoder is a 5-layer CNN and the setting of each layer is illustrated in Fig. 2. The strides of all convolutional layers are set as 1. To avoid exploding/vanishing gradients and speed up training, batch normalization is applied. ReLU activation function is used to speed up the convergence [37] and avoid gradient sparsity.

2) Discriminator Architecture: Discriminators are designed to play an adversarial role against the generator. In particular, D_v and D_i aim to distinguish the generated images from the visible and infrared images, respectively. However, these two types of source images are manifestations of different phenomena, thus have considerably different distributions. In other words, there are conflicts in the guidance of D_v on G and D_i on G. In our network, we should not only consider the adversarial relationship between the generator and discriminators but also take into account the balance of D_v and D_i. Otherwise, either the strength or weakness of one discriminator will finally lead to the inefficiency of the other as the training proceeds. In our work, the balance is achieved by the design of network architectures and training strategy (as discussed in Sec. V-A).

The discriminators D_v and D_i share the same architecture, which is set to be less complicated compared with the generator architecture, as shown in Fig. 3. The stride of all convolutional layers is set as 2. In the last layer, we employ the tanh activation function to generate a scalar that estimates the probability of the input image from source images rather than G.

IV. Application to Medical Image Fusion

In this section, we apply our proposed method to fuse medical images such as MRI and PET image fusion. We treat

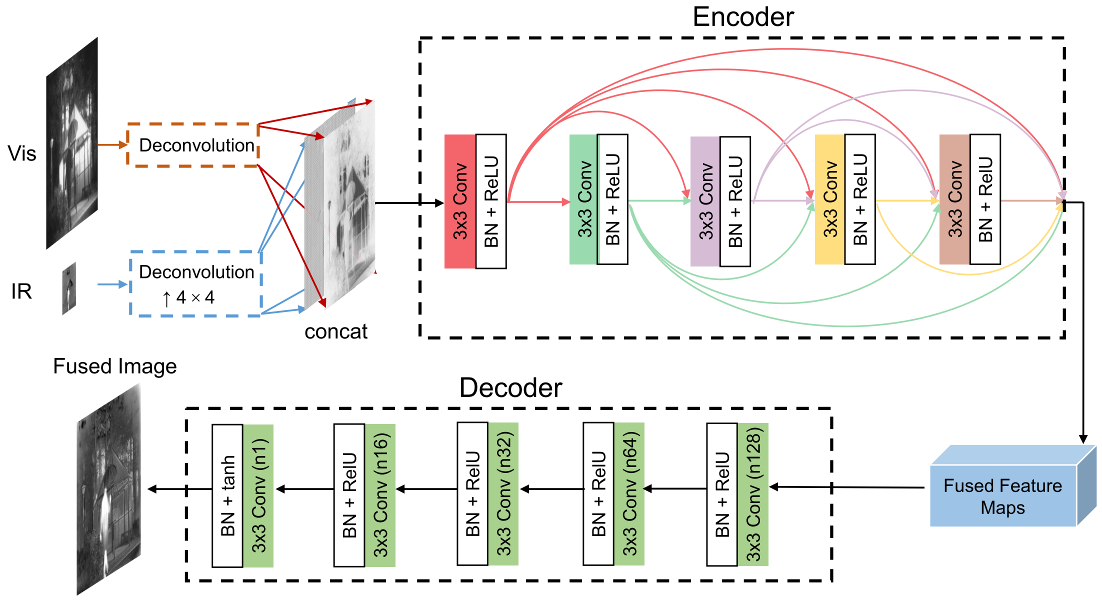


Fig. 2. The overall architecture of our generator, including layers of encoder and decoder. 3 × 3: filter size, Conv(nk): convolutional layer which obtains k feature maps, BN: batch normalization.

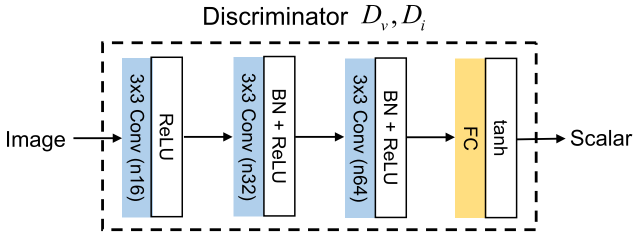


Fig. 3. The overall architecture of our discriminator. 3 × 3: filter size, Conv(nk): convolutional layer which obtains k feature maps, BN: batch normalization, FC: fully connected layer.

the PET images shown in pseu-do-color as color images, and DDcGAN is applied for fusing of high-resolution MRI image and low-resolution intensity component of PET image. In the following, we first introduce the background of medical image fusion, and then take the MRI and PET image fusion as an example and provide some implementation details.

A. Background

Multi-modal medical images have the advantage of offering diversified features to enhance robustness and accuracy and thus, the fusion of them provides a powerful tool for biomedical research and clinical applications, such as medical diagnostics, monitoring and treatment [38], [39]. These medical imaging can be divided into structural and functional systems [40]. Structure from motion methods [41] are typically used to obtain the structural information in natural image domain. While in medical imaging, X-ray, MRI and Computed Tomography are a typical structural system, which can provide structural and anatomical information with high resolution. The functional system can provide functional and metabolic information, such as PET and Single-Photon Emission Computed Tomography while these images are often accompanied by low resolution. The limited resolution restricts their clinical applications and encourages the fusion of functional and anatomical images.

According to the theories applied, existing medical fusion methods can be summarized into different categories, such as substitution methods [40], [42], arithmetic combination methods [43], and multi-resolution methods [44], [45]. In this paper, we take the MRI and PET image fusion as an example and apply our DDcGAN to solve this problem. MRI images are superior in capturing the details of soft tissue structures in organs such as brain, heart and lungs in high spatial resolution. The PET images are obtained by nuclear medicine imaging to provide functional and metabolic information, such as blood flow and flood activity. The captured images are usually rich in color but low in spatial resolution. Therefore, by fusing these two type medical images, the results will contain both spatial and spectral features in the source images for qualitative detection and quantitative determination.

The PET image in pseudo-color is traditionally treated as a color image and the color is the representation of the functional information, as shown in Fig. 4(a). In order to retain it, the color of the fused image should be as similar to that of the PET image as possible. For this purpose, de-correlated color models are used to separate the achromatic and chromatic information in the color into different channels. Then, the achromatic channel is substituted or fused with the MRI image [46]. In our work, we employ the intensity, hue and saturation (IHS) de-correlated color model and the intensity channel is the specific achromatic channel to be fused, as shown in Fig. 4(b). Because the other two channels are the representation of chromatic information, which ought to remain unchanged during the fusion process, the PET image is similar with the infrared image in using the intensity distribution to represent feature information. A slightly different

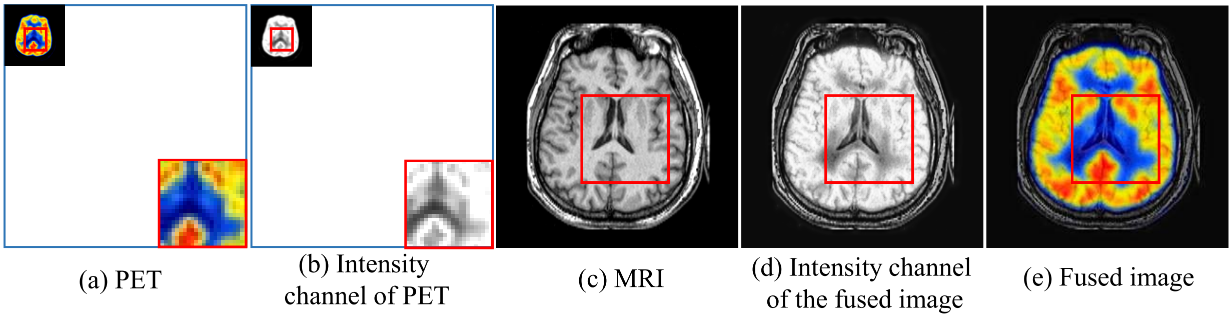

(a) PET
(b) Intensity channel of PET
(c) MRI
(d) Intensity channel of the fused image
(e) Fused image

Fig. 4. Schematic illustration of fusing the low-resolution PET image in RGB channels and the high-resolution MRI image in the gray channel to obtain the high-resolution fused image in RGB channels.

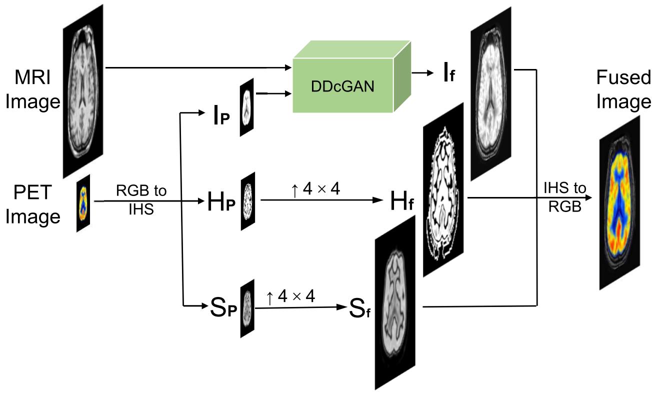


Fig. 5. The entire procedure of applying the proposed DDcGAN for MRI and PET image fusion.

point is that the PET image uses it to represent the functional information while in the infrared image, it is the reflection of thermal radiation. By contrast, the MRI image can provide detailed morphological information in the form of texture. It is mainly characterized by the gradients. Thus, like the visible image, the advantage of rich texture information in the MRI image can be applied to overcome the uncertainty of contouring the soft tissue structures on the PET image. From this point of view, the essence of fusing MRI and PET image has a great deal of similarity with that of fusing visible and infrared images. As shown in Fig. 4, the fused image is supposed to minimize both the spatial distortion caused by the spatial detail loss between the MRI (Fig. 4(c)) and the intensity channel (Fig. 4(d)) and the spectral distortion caused by color differences between the PET (Fig. 4(b)) and the fused intensity channel (Fig. 4(d)) simultaneously. Accompanied by the processed components of H and S channels, the final fused image is a three-channel image with abundant color and detail information, as shown in Fig. 4(e).

B. MRI and PET Image Fusion via DDcGAN

Uniformly, we assume that the resolution of the MRI image is 4 × 4 times that of the intensity component of the PET image and take it as an example. The entire fusion procedure is illustrated in Fig. 5. The multispectral input PET image with RGB channels are firstly transformed into IHS channels, as shown in Eq. (13), with the intensity channel displaying the brightness in a spectrum, the hue channel showing the property of the spectral wavelength, and the saturation channel demonstrating the purity of the spectrum:

$$\begin{pmatrix} I_{\text{PET}} \\ V1_{\text{PET}} \\ V2_{\text{PET}} \end{pmatrix} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \\ 1/\sqrt{6} & 1/\sqrt{6} & -2/\sqrt{6} \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix} \begin{pmatrix} R_{\text{PET}} \\ G_{\text{PET}} \\ B_{\text{PET}} \end{pmatrix}. \tag{13}$$

The components of H and S channels can be represented by variables V1 and V2 as follows:

$$H_{\text{PET}} = \tan^{-1}\left(\frac{V1_{\text{PET}}}{V2_{\text{PET}}}\right), \tag{14}$$

$$S_{\text{PET}} = \sqrt{V1_{\text{PET}}^2 + V2_{\text{PET}}^2}. \tag{15}$$

The fusion process is produced on the component of I channel of the PET image and the MRI image. Correspondingly, the input of the generator is the low-resolution I_{PET} and the high-resolution MRI image M. The output of the generator $I_{\text{fuse}} = G(M, I_{\text{PET}})$ is the new I channel of the fused image with high resolution. During the training procedure, the discriminator D_i is trained to discriminate differences between I_{fuse} and I_{PET}, while the probability of the input image from MRI images rather than G is obtained by the discriminator D_v. Therefore, the specific loss function of the generator can be expressed as follows:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda \mathcal{L}_{\text{con}}, \tag{16}$$

where the adversarial loss function $\mathcal{L}_G^{\text{adv}}$ is defined as:

$$\begin{aligned} \mathcal{L}_G^{\text{adv}} = \mathbb{E}\left[\log\left(1 - D_v\left(G\left(M, I_{\text{PET}}\right)\right)\right)\right] \\ + \mathbb{E}\left[\log\left(1 - D_i\left(\psi G\left(M, I_{\text{PET}}\right)\right)\right)\right]. \end{aligned} \tag{17}$$

And the content loss $\mathcal{L}_{\text{con}}$ is modified as:

$$\begin{aligned} \mathcal{L}_{\text{con}} = \mathbb{E}\Big[\|\psi G\left(M, I_{\text{PET}}\right) - I_{\text{PET}}\|_F^2 \\ + \eta \|G\left(M, I_{\text{PET}}\right) - M\|_{TV} \Big]. \end{aligned} \tag{18}$$

For the discriminators D_v and D_i, the adversarial losses are respectively defined as follows:

$$\begin{aligned} \mathcal{L}_{D_v} &= \mathbb{E}\left[-\log D_v\left(M\right)\right] \\ &\quad + \mathbb{E}\left[-\log\left(1 - D_v\left(G\left(M, I_{\text{PET}}\right)\right)\right)\right], \end{aligned} \tag{19}$$

$$\begin{aligned} \mathcal{L}_{D_i} &= \mathbb{E}\left[-\log D_i\left(I_{\text{PET}}\right)\right] \\ &\quad + \mathbb{E}\left[-\log\left(1 - D_i\left(\psi G\left(M, I_{\text{PET}}\right)\right)\right)\right]. \end{aligned} \tag{20}$$

To preserve the chromatic information in the PET image, the components of H and S channels of the PET image and the fused image should be as identical as possible. For these two channels, we directly employ the bicubic interpolation as the upsampling operation. The upsampled components are presented as H_{new} and S_{new} and their resolutions are both 4×4 time those of H_{PET} and S_{PET}. According to Eq. (14) and Eq. (15), the variables V1 and V2 can be updated by the components of H and S channels:

$$V1_{\text{new}} = S_{\text{new}} \sin H_{\text{new}}, \tag{21}$$

$$V2_{\text{new}} = S_{\text{new}} \cos H_{\text{new}}. \tag{22}$$

The inverse transform to obtain the final fused image in RGB channels from IHS channels can be represented as:

$$\begin{pmatrix} R_{\text{new}} \\ G_{\text{new}} \\ B_{\text{new}} \end{pmatrix} = \begin{bmatrix} 1/\sqrt{3} & 1/\sqrt{6} & 1/\sqrt{2} \\ 1/\sqrt{3} & 1/\sqrt{6} & -1/\sqrt{2} \\ 1/\sqrt{3} & -2/\sqrt{6} & 0 \end{bmatrix} \begin{pmatrix} I_{\text{fuse}} \\ V1_{\text{new}} \\ V2_{\text{new}} \end{pmatrix}. \tag{23}$$

V. Experimental Results

In this section, to validate the effectiveness of our DDcGAN, we firstly compare it with several state-of-the-art methods on publicly available datasets by qualitative comparisons both for infrared and visible image fusion and PET and MRI image fusion. For quantitative comparisons, we utilize six metrics to evaluate the fusion results. The experiments of discriminator analysis are also conducted.

A. Dataset and Training Details

1) Dataset: We validate the proposed DDcGAN on the publicly available *TNO Human Factors* dataset[1] for the infrared and visible image fusion. We select 36 infrared and visible image pairs from the dataset and crop them into 27, 264 patch pairs with 84×84 pixels. As we focus on fusing images of different resolutions while the source images in the dataset are of the same resolution, we downsample the infrared images to one quarter resolution. Therefore, all visible image patches are of size 84×84 and all infrared patches are of size 21×21. Parameters in our model are set as $\lambda = 0.5$ and $\eta = 1.2$. The entire network is trained with a learning rate of 2×10^{-3} with exponentially decaying to 0.75 of the original value after each epoch. The batch size is set as 24.

The application of our proposed DDcGAN to MRI and PET image fusion is validated on the publicly available Harvard medical school website.[2] The original PET and MRI images are all of size 256×256. For the purpose of validating the effectiveness of our method on fusing images of different resolutions, each channel of the PET images is downsampled to the size of 64×64. 83 PET and MRI pairs are downloaded and cropped into 9, 984 patch pairs as our training set. Similarly, all MRI patches are of size 84×84 and the intensity patches of all PET images are of size 21×21. The parameters, learning rate and the batch size are the same with those set in the infrared and visible fusion.

2) Training Details: During the training process, the principle is to make the generator and discriminators form an adversarial relationship with each other. In order to overcome some problems in training GAN and improve the training results, rather than taking turns training G, D_v and D_i once per batch in principle, we train D_v or D_i more times if it fails to discriminate the data from G and vice verse. The detailed training process is shown in Alg. 1. Except for $\mathcal{L}_{max}$, $\mathcal{L}_{min}$ and $\mathcal{L}_{G_{max}}$, a threshold for the number of iterations is additionally set. The reason is that the goal of updating the generator or discriminators more times is to keep the balance between them. However, there are still situations where these networks have been trained many times but still cannot achieve balance conditions. Especially for the generator, more training steps to minimize the adversarial loss may lead to higher content loss and higher $\mathcal{L}_G$, failing to achieve the balance condition. Thus, it can avoid the algorithm falling into an endless loop. Moreover, updating other networks timely will enable them to play a new role in guiding the current network, thus possibly avoiding the above-mentioned situation.

During the testing phase, we only use the trained generator to generate fused images. Since there are no fully connected layers in our generator, the input source images can be of any size with a predefined resolution ratio.

B. Results and Analysis on Infrared and Visible Image Fusion

To verify the effectiveness of our proposed DDcGAN, we compare it with seven state-of-the-art image fusion methods, including directional discrete cosine transform and principal component analysis (DDCTPCA) [14], hybrid multi-scale decomposition (HMSD) [47], fourth-order partial differential equations (FPDE) [48], gradient transfer fusion (GTF) [17], different resolution total variation (DRTV) [49], DenseFuse [29] and FusionGAN [21]. Due to some of the competitors require that source images share the same resolution, we upsample the low-resolution infrared images before performing these methods for fusion. While in DRTV and FusionGAN, as they can be applied to fuse images of different resolutions, the preprocessing of up-sampling low-resolution infrared images is unnecessary. The fused results of all methods are assessed both subjectively and objectively.

1) Qualitative Comparisons: We first report some intuitive results on six typical image pairs, as shown in Fig. 6. Compared with the existing fusion methods, our DDcGAN has three distinctive advantages. First, our results can maintain the high-contrast property of the infrared image, *e.g.*, the thermal targets are prominent in our fused images, as shown in the first and second examples, which is very important for the subsequent target detection task. Second, our results can preserve abundant texture details from the visible images, *e.g.*, the backgrounds contain more detail information in our fused images, as shown in the third to fifth examples, which is beneficial for accurate target recognition. Third, our results are clearer due to that it does not suffer from thermal radiation information blurring caused by upsampling of the low-resolution infrared images, as shown in the sixth example.

As can be seen from Fig. 6, DDCTPCA, HMSD, FPDE and DenseFuse cannot highlight the thermal targets well, while

[1] https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029
[2] http://www.med.harvard.edu/AANLIB/home.html

Algorithm 1 Training Process of DDcGAN

```
Parameter descriptions:
The numbers of steps to train $G$, $D_v$, and $D_i$ are denoted as $I_G$, $I_{D_v}$, and $I_{D_i}$ respectively.
$I_{max}$ is the max steps to train the networks. In our experiments, we used $I_{max} = 20$.
$\mathcal{L}_{max}$, $\mathcal{L}_{min}$ and $\mathcal{L}_{Gmax}$ are applied to determine a range to uncollapse training.
$\mathcal{L}_{max}$ and $\mathcal{L}_{min}$ are for adversarial losses of $G$, $D_v$, and $D_i$. $\mathcal{L}_{Gmax}$ is for the total loss of $G$.
We set $\mathcal{L}_{max} = 1.8$, $\mathcal{L}_{min} = 1.2$ and $\mathcal{L}_{G_{max}} = 0.8 \times \mathcal{L}_G$ in the first batch empirically in our experiments.

  Initialize $\theta_{Dv}$ and $\theta_{Di}$ for $D_v$ and $D_i$, and $\theta_G$ for $G$;
  In each training iteration:
  - Train discriminators $D_v$ and $D_i$:
      • Sample $m$ visible patches $\{v^1, \cdots, v^m\}$ and $m$ corresponding infrared patches $\{i^1, \cdots, i^m\}$;
      • Obtain generated data $\{G(v^1, i^1), \cdots, G(v^m, i^m)\}$;
      • Update discriminator parameters $\theta_{Dv}$ by SGDOptimizer to minimize $\mathcal{L}_{D_v}$ in Eq. (11); (step I)
      • Update discriminator parameters $\theta_{Di}$ by SGDOptimizer to minimize $\mathcal{L}_{D_i}$ in Eq. (12); (step II)
      • While $\mathcal{L}_{D_v} > \mathcal{L}_{max}$ and $I_{D_v} < I_{max}$, repeat step I.
        $I_{D_v} \leftarrow I_{D_v} + 1$;
      • While $\mathcal{L}_{D_i} > \mathcal{L}_{max}$ and $I_{D_i} < I_{max}$, repeat step II.
        $I_{D_i} \leftarrow I_{D_i} + 1$;
  - Train generator $G$:
      • Sample $m$ visible patches $\{v^1, \cdots, v^m\}$ and $m$ corresponding infrared patches $\{i^1, \cdots, i^m\}$;
      • Obtain generated data $\{G(v^1, i^1), \cdots, G(v^m, i^m)\}$;
      • Update generator parameters $\theta_G$ by RMSPropOptimizer to minimize $\mathcal{L}_G$ in Eq. (8); (step III)
      • While $(\mathcal{L}_{D_v} < \mathcal{L}_{min}$ or $\mathcal{L}_{D_i} < \mathcal{L}_{min})$ and $I_G < I_{max}$, update generator parameters $\theta_G$ by RMSPropOptimizer to
        minimize $\mathcal{L}_G^{\text{adv}}$ in Eq. (9).
        $I_G \leftarrow I_G + 1$;
      • While $\mathcal{L}_G > \mathcal{L}_{Gmax}$ and $I_G < I_{max}$, repeat step III.
        $I_G \leftarrow I_G + 1$;
```

GTF, DRTV and FusionGAN cannot obtain abundant texture details. Besides, they all suffer from thermal radiation information blurring except DRTV and FusionGAN. Although DRTV can prevent loss of texture information caused by upsampling when fusing source images of different resolutions, the results of DRTV inevitably suffer from staircase effects due to the application of first-order TV. In contrast, the results of DDcGAN can obviously avoid staircase effects and details in our results are more similar to those in the visible images. Compared with FusionGAN, due to the employment of the deconvolution layers, the introduction of the discriminator D_i, different network architecture and improved training strategy, our fused results can highlight thermal targets more obviously by higher contrast and meanwhile, contain more natural details which are more indistinguishable from the visible images. Excluding the effects of deconvolution layers, different network architecture and the training strategy, the influence of the additional discriminator will be analyzed later in Sec. V-B.3. Generally, our DDcGAN works well and the fused images are more like super-resolved infrared images which also contain abundant texture detail information in visible images.

2) Quantitative Comparisons: We further report quantitative comparisons of our DDcGAN and the competitors on the rest 15 image pairs in the dataset. Eight metrics such as entropy (EN) [50], mean gradient (MG), spatial frequency (SF), standard deviation (SD) [51], peak signal-to-noise ratio (PSNR), and correlation coefficient (CC), structural similarity index measure (SSIM) [52] and visual information fidelity (VIF) [53] are used for evaluation.

- Entropy (EN): This metric can measure the amount information contained in the fused image from the perspective of information theory and is defined as follows:

$$EN = -\sum_{l=0}^{L-1} p_l \log_2 p_l,$$

  where p_l denotes the normalized histogram of corresponding gray level in the fused image. And the number of all the gray levels is set as L. The larger entropy means that there is more information reserved in the image and the method achieves a better performance.
- Mean gradient (MG): MG is mathematically defined as:

$$MG = \frac{\sum_{i=2}^{M}\sum_{j=2}^{N}\sqrt{\left(\left(x_{i,j}-x_{i-1,j}\right)^2+\left(x_{i,j}-x_{i,j-1}\right)^2\right)/2}}{(M-1)(N-1)}.$$

  The larger MG is, the more gradient information the image contains and the better fusion performance the algorithm has.
- Spatial frequency (SF): SF is based on the gradient distribution to effectively reveal the details and texture

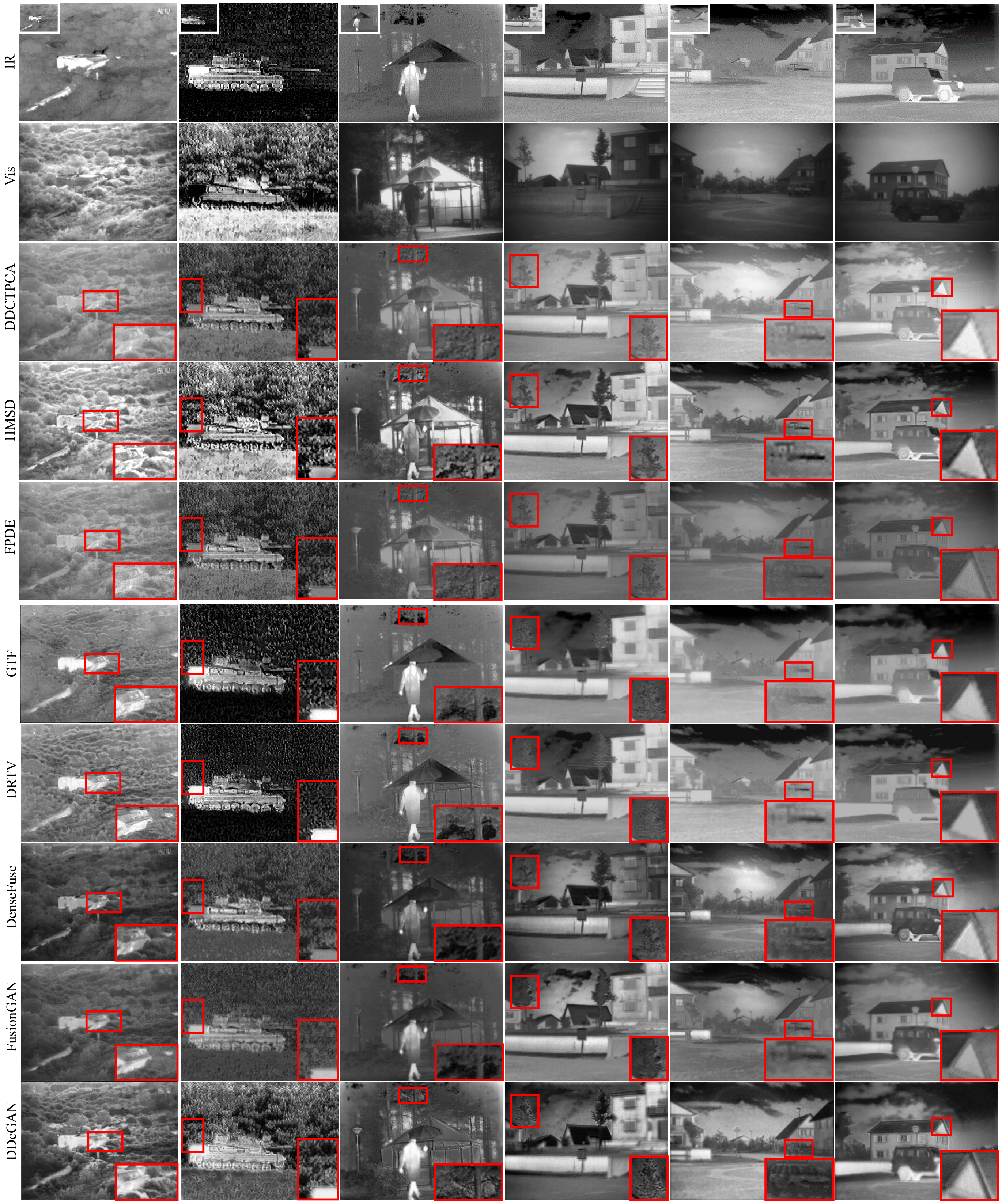


Fig. 6. Qualitative comparison of our DDcGAN with 7 state-of-the-art methods on 6 typical infrared and visible image pairs. From top to bottom: infrared image, visible image, fusion results of DDCTPCA [14], HMSD [47], FPDE [48], GTF [17], DRTV [49], DenseFuse [29], FusionGAN [21] and our DDcGAN. For more intuitive comparison, the infrared images are enlarged in the first row and the original low-resolution infrared images are shown in the white box in the top left corner.

of the image. It is defined by spatial row frequency (RF) and column frequency (CF):

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

where $RF = \sqrt{\sum_{i=1}^{M}\sum_{j=2}^{N}\left(x_{i,j} - x_{i,j-1}\right)^2}$ and $CF = \sqrt{\sum_{i=2}^{M}\sum_{j=1}^{N}\left(x_{i,j} - x_{i-1,j}\right)^2}$. The larger SF, the richer edges and texture details the image contains. And human perception is more sensitive to the image with larger SF.

- 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, the larger SD, the better visual effect the fused image achieves. Mathematically, SD is defined as:

$$SD = \sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}\left(x_{i,j} - \mu\right)^2},$$

where μ is the mean value of the image x.

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

$$PSNR = 10\log_{10}\frac{r^2}{MSE},$$

where r is the peak value of the fused image and is set as 256 in this paper. MSE is the mean square error that measures the dissimilarity between the fused image and source images and is defined as follows:

$$MSE = \omega_a MSE_{af} + \omega_b MSE_{bf},$$

where $MSE_{xf} = \frac{1}{MN}\sum_{i=0}^{M-1}\sum_{j=0}^{N-1}\left(x_{i,j} - f_{i,j}\right)^2$. A larger PSNR indicates the less distortion the fusion process produces and the fused image is more similar to the source images.

- Correlation coefficient (CC): The metric CC measures the degree of linear correlation between the source images and the fused image. It is mathematically defined as:

$$CC = \omega_a r_{af} + \omega_b r_{bf},$$

where $r_{xf} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j}-\mu_x)(f_{i,j}-\mu_f)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(x_{i,j}-\mu_x)^2\sum_{i=1}^{M}\sum_{j=1}^{N}(f_{i,j}-\mu_f)^2}}$, μ_x and μ_f denote the mean values of the source image x and the fused image f, respectively. A larger CC indicates that the fused image is more similar to the source images.

- Structural similarity index measure (SSIM): SSIM is the widely used metric which models the loss and distortion between two images according to their similarities in light, contrast and structure information. Mathematically, SSIM between images x and y can be defined as follows:

$$SSIM_{xy} = \sum_{x_i,y_i}\frac{2\mu_{x_i}\mu_{y_i}+c_1}{\mu_{x_i}^2+\mu_{y_i}^2+c_1}\cdot\frac{2\sigma_{x_i}\sigma_{y_i}+c_2}{\sigma_{x_i}^2+\sigma_{y_i}^2+c_2}\cdot\frac{\sigma_{x_iy_i}+c_3}{\sigma_{x_i}\sigma_{y_i}+c_3},$$

where μ denotes the mean value, σ is the standard deviation/covariance, c_1, c_2 and c_3 are the parameters to make the algorithm stable. Thus, SSIM between source images a, b and the fused image f can be defined as:

$$SSIM = \omega_a SSIM_{af} + \omega_b SSIM_{bf}.$$

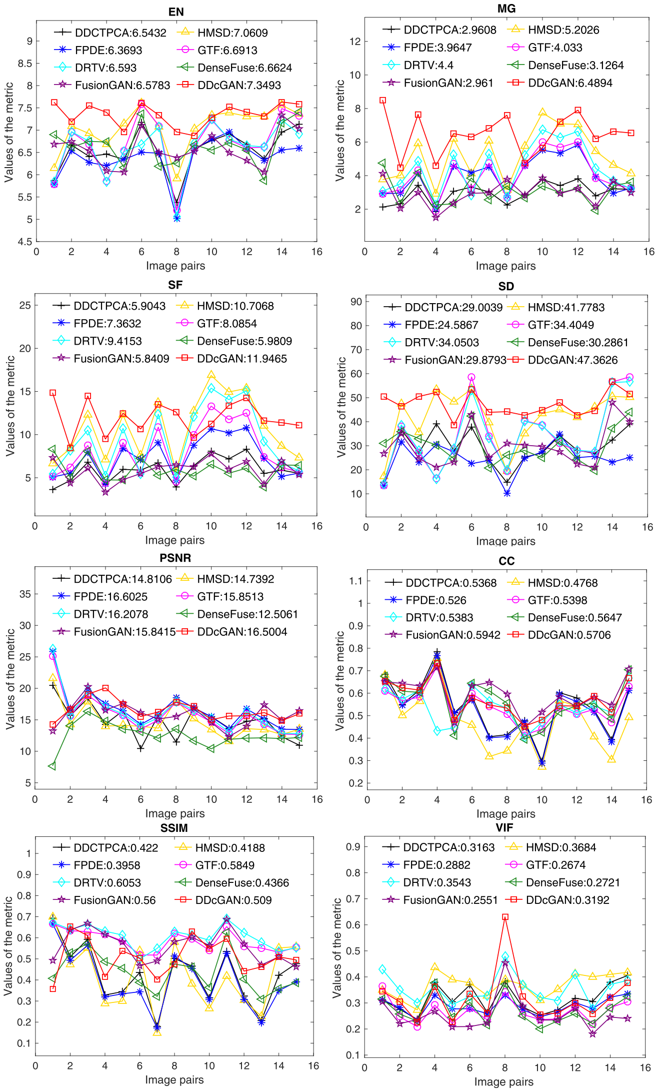


Fig. 7. Quantitative comparison of our DDcGAN for infrared and visible image fusion with 7 state-of-the-art methods. Means of metrics for different methods are shown in the legends.

- Visual information fidelity (VIF): The metric is consistent with human visual system and measures the information fidelity. It can be computed by four steps: (a) filter and divide the source images and the fused image into different blocks; (b) evaluate the visual information of each block; (c) calculate the VIF for each subband; (d) calculate the overall metric. A large VIF indicates that the fusion method has a good performance.

The results of quantitative comparisons are summarized in Fig. 7. As can be seen from the statistical results, our DDcGAN can generate the largest average values on the first 4 metrics: EN, MG, SF and SD. In particular, our DDcGAN achieves the best values of EN, MG, SF and SD on 13, 13, 10 and 8 image pairs, respectively. For the metric PSNR and CC, our DDcGAN can achieve comparable results with the average values being the second largest. These metrics only follow behind FPDE and FusionGAN by a narrow margin, respectively. As for VIF and SSIM, our result is the third and fourth largest respectively. These results demonstrate

TABLE I

AVERAGE RUNTIME COMPARISON OF DIFFERENT METHODS ON THE 15 TESTING IMAGE PAIRS (UNIT: SECOND). AS FOR THE RUNTIME OF DEEP LEARNING-BASED METHODS, THE FIRST VALUE IS TESTED ON CPU AND THE SECOND VALUE IS TESTED ON GPU

Methods	DDCTPCA [14]	HMSD [47]	FPDE [48]	GTF [17]	DRTV [49]	DenseFuse [29]	FusionGAN [29]	DDcGAN
Mean	106.83	1.70	1.93	7.27	5.99	5.43 / 0.54	14.12 / 0.58	18.96 / 1.36
STD	52.11	0.87	1.14	3.85	4.37	2.54 / 0.19	7.05 / 0.60	10.11 / 0.91

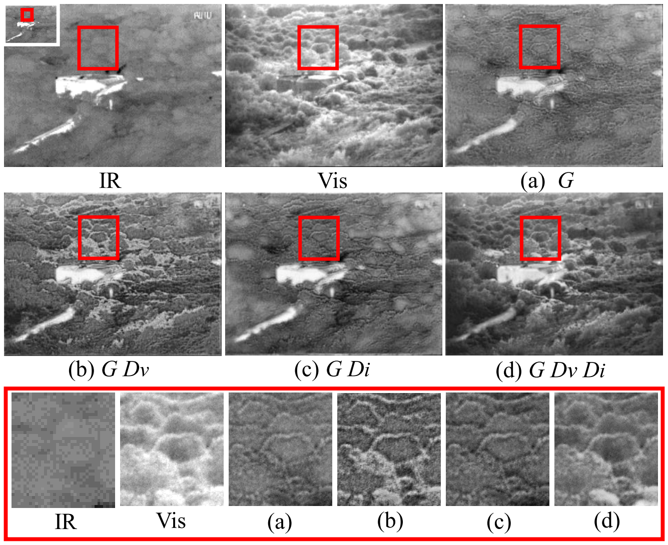


Fig. 8. Fused results on *bunker* when the discriminators in the entire networks change. We highlight a region and zoom in it as shown in the bottom red box. The infrared images are enlarged in the first plot, and the original low-resolution infrared images are shown in the white box in the top left corner.

that our method can reserve information to the greatest extent, especially the most gradient information, the richest edges and texture details, and the highest contrast, as shown in the first four metrics. In addition, the results of our methods can achieve considerable similarity with the source images.

The average runtime of different methods on the testing data is provided in Table I. All the methods are tested on a desktop with 3.4 GHz Intel Core i5 CPU. Since there are three deep learning-based methods (*i.e.*, DenseFuse, FusionGAN and DDcGAN), these methods are also tested on NVIDIA Geforce GTX Titan X. The reason why the runtime of DDcGAN is slower is that in the testing phase, the input of our model is the whole image. Thus, for each test image pair, our model is rebuilt according to their size and the parameters of the trained model are restored into the rebuilt model to avoid the possible seam effects caused by cropping tested images into patches and the distortion caused by resizing images. Another reason is that our model is deeper than other deep learning-based methods, resulting in more test runtime.

3) Discriminator Analysis: There are two discriminators presented in our proposed model, *i.e.*, D_v and D_i. In order to illustrate the effect of each discriminator, we perform four comparative experiments: (a) The entire networks merely consist of the generator G and the ultimate training objective is reduced to minimize $\mathcal{L}_{\text{con}}$ in Eq. (10). (b) D_i is not employed and the adversarial relationship exists only between G and D_v. (c) D_v is not embraced in the entire networks. Thus, the adversarial game is established between G and D_i. (d) The fused images are generated by the method proposed in this paper. All of G, D_v, and D_i play a part in the networks. All the comparative experiments are under the same experimental settings and the fused results are shown in Fig. 8.

In method (a), the training objective is to minimize the content loss $\mathcal{L}_{\text{con}}$, which is the first-order TV model in essence. This model performs well in preserving edges of the object in the piecewise constant image while it inevitably produces staircase effects [54], as can be seen in Fig. 8(a). With the introduction of D_v, the staircase effects have been alleviated in Fig. 8(b). However, the disadvantage is that the intensity distribution of the fused image is modified according to that of the visible image, leading to the reduction of the prominence of the thermal targets. The separate introduction of D_i increases the contrast between the thermal targets and the background, which is particularly evident in the prominence of the bunker between the results shown in Fig. 8(a) and Fig. 8(c). Nevertheless, the result of method (c) lacks in detail information compared with method (b).

With a comprehensive consideration of advantages and disadvantages of method (b) and (c), we propose a new structure based on conditional generative adversarial networks with dual discriminator: D_v and D_i. The use of D_i can correct the distinct differences of the intensity distribution between the result of method (b) and the infrared image. Meanwhile, more details and texture information can be added to the result of method (c) by introducing D_v. Worthy of note that since the discriminators increase from just D_v or D_i to both of them, the requirement and the training target of the generator become harsher. Under the condition that there exists a contradictory relationship between the discrimination tasks of D_v and D_i and according to the training strategy in Alg. 1, the training of G, D_v or D_i can be adjusted in case any of them loses its specific function, the generation ability of the generator can be further improved. On the promise that the thermal targets are still prominent, the results of method (d) include more details and these details look more similar to those in the visible images by effectively solving the problem of staircase effects compared with those shown in Figs. 8(b) and (c).

4) Generator Analysis: In the loss function of the generator G, there are two subitems, *i.e.*, the adversarial loss $\mathcal{L}_G^{\text{adv}}$ and the content loss $\mathcal{L}_{\text{con}}$. To verify the effect of each subitem, three comparative experiments are performed: (a) $\mathcal{L}_G = \lambda\mathcal{L}_{\text{con}}$. This comparative experiment is the same with method (a) in Sec. V-B.3. G is trained to minimize $\mathcal{L}_{\text{con}}$ in Eq. (10). (b) $\mathcal{L}_G = \mathcal{L}_G^{\text{adv}}$. The content loss is not introduced in $\mathcal{L}_G$. Then G is only trained to fool D_v and D_i. It should be noted that in this method, due to the lack of pixel-wise constraints, the introduction of the deconvolution

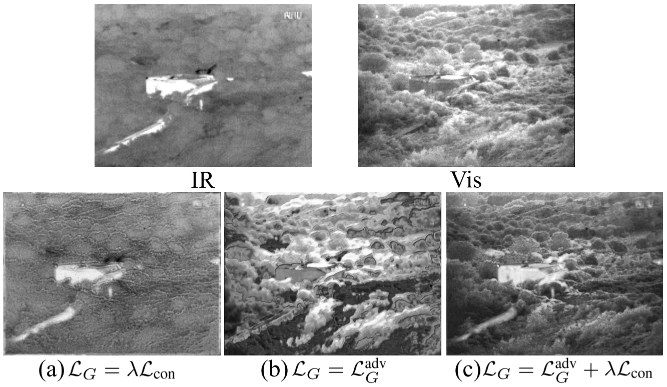

IR
Vis
(a) $\mathcal{L}_G = \lambda\mathcal{L}_{\text{con}}$
(b) $\mathcal{L}_G = \mathcal{L}_G^{\text{adv}}$
(c) $\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda\mathcal{L}_{\text{con}}$

Fig. 9. Fused results on *bunker* when the loss function of the generator $\mathcal{L}_G$ changes.

layers may cause the cavity effect. Thus, we replace these layers with two upsampling layers to avoid this influence. (c) $\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda\mathcal{L}_{\text{con}}$. It is the proposed method. With the same experimental settings, the fused results of these three methods are shown in Fig. 9.

On the one hand, without the adversarial loss, the fused result fails to exhibit more and clearer texture details in the visible image, as shown in Fig. 9(a). On the other hand, without the content loss, the generator is incapable of knowing which type of information should be retained from source images. Without the pixel-wise constraints, what the generator can do is to make the probability distribution of generated images close to that of source images. In this case, the fused image may have high contrast and texture details. However, the highlighted regions may not be the thermal targets in the infrared image and texture details may be different from the visible image, as shown in Fig. 9(b). Thus, when DDcGAN is trained without the content loss, it will generate artifacts and incomprehensible results. By combining these two subitems, DDcGAN can solve this problem and generate a high-quality fused image, as shown in Fig. 9(c).

C. Results on MRI and PET Image Fusion

According to corresponding schemes, we compare our method with six other fusion methods separately based on principal component analysis method such as DDCTPCA [14], sparse representation method such as adaptive sparse representation (ASR) [56], wavelet transform method such as discrete cosine harmonic wavelet transform (DCHWT) [55], saliency method such as Structure-Aware [57] and deep learning-based methods such as FusionGAN [21] and RCGAN [58]. Among these methods, PCA is a classic theory applied for the fusion of PET and MRI images. Based on PCA and taken as a representation of comparison methods for infrared and visible image fusion utilized in Sec. V-B, DDCTPCA is employed here for comparison once more. ASR can be applied for multi-modal image fusion and perform fusion and denoising simultaneously. DCHWT takes into account the fusion of multi-spectral image fusion. Structure-Aware is a method expressly proposed for multi-modal medical image fusion. FusionGAN and RCGAN are methods based on GAN and also representations of infrared and visible image fusion methods.

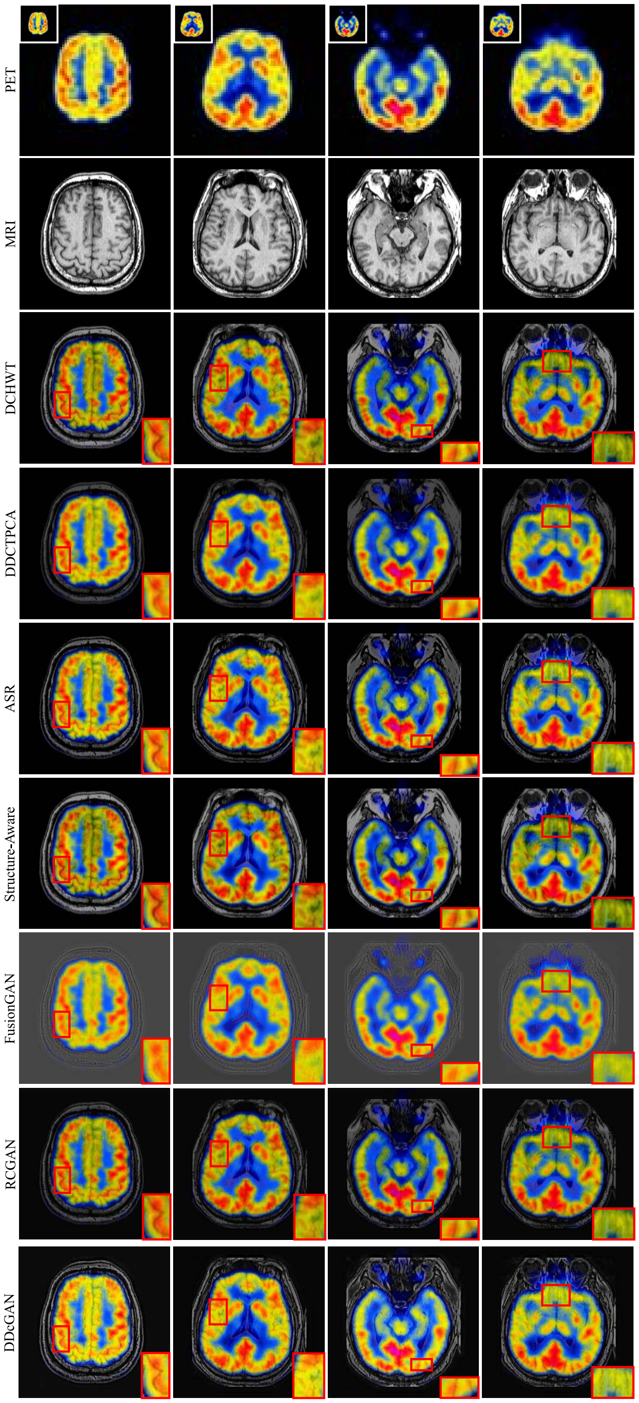


Fig. 10. Qualitative comparison of our DDcGAN with 6 state-of-the-art methods on 4 typical MRI and PET image pairs of different (from left to right: #99, #81, #60 and #70) transaxial sections of the brain-hemispheric. From top to bottom: PET image, MRI image, fusion results of DCHWT [55], DDCTPCA [14], ASR [56], Structure-Aware [57], FusionGAN [21] and RCGAN [58] and our DDcGAN. For more intuitive comparison, the PET images are enlarged in the first row and the original low-resolution PET images are shown in the white box in the top left corner.

In the remainder of this section, qualitative and quantitative experiments are conducted to demonstrate the effectiveness of our method on PET and MRI image fusion.

1) Qualitative Comparison: Four typical and intuitive results on four different transaxial sections of the brain-hemispheric are exhibited in Fig. 10. By comparison, DCHWT, Structure-Aware and RCGAN significantly reduce

TABLE II

AVERAGE RUNTIME COMPARISON OF DIFFERENT METHODS ON THE 20 TESTING IMAGE PAIRS (UNIT: SECOND). AS FOR THE RUNTIME OF DEEP LEARNING-BASED METHODS, THE FIRST VALUE IS TESTED ON CPU AND THE SECOND VALUE IS TESTED ON GPU

Methods	DCHWT [55]	DDCTPCA [14]	ASR [56]	Structure-Aware [57]	FuisonGAN [29]	RCGAN [58]	DDcGAN
Mean	0.91	24.07	29.04	0.04	3.24 / 0.05	3.01 / 0.09	5.25 / 0.52
STD	0.05	0.42	1.05	0.01	0.13 / 0.08	0.03 / 0.01	0.13 / 0.21

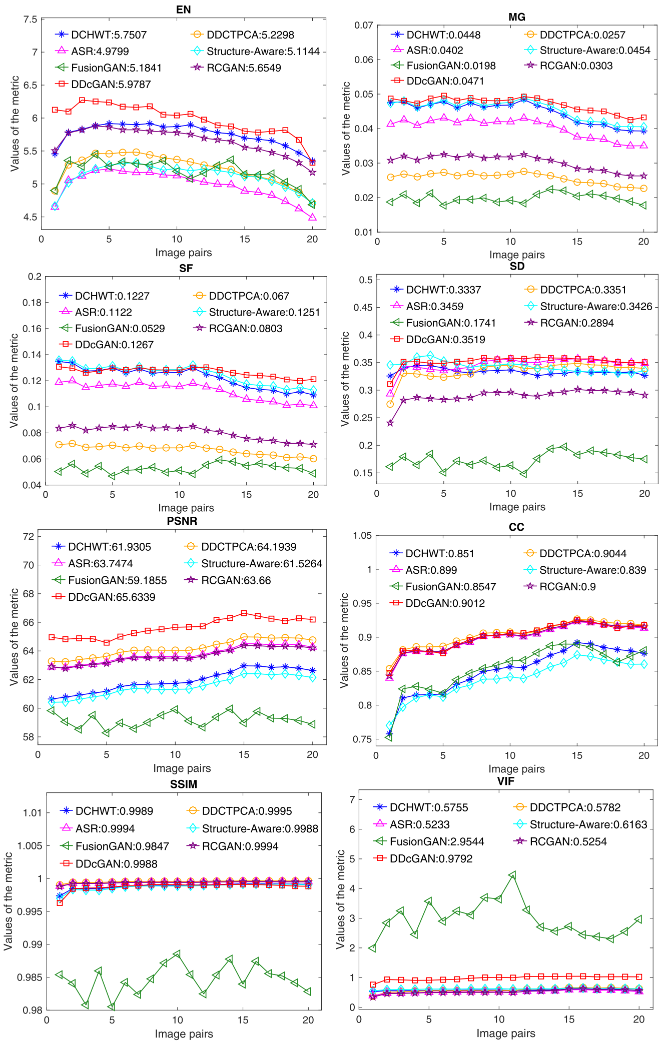


Fig. 11. Quantitative comparison of our DDcGAN for PET and MRI image fusion with 5 state-of-the-art methods. Means of metrics for different methods are shown in the legends.

the intensity of colors in the PET image, leading to the loss of functional information. By contrast, the results generated by DDCTPCA, ASR, FusionGAN and DDcGAN exhibit brighter and stronger colors. And among these four methods, the colors of our results are the closest to those of the original PET images. Furthermore, as a result of the unsampling of low-resolution PET image, the results of six comparison methods suffer from functional information blurring, presented as blurred color information, as shown in the first and second groups of results, and blurred details, which can be seen in the third group of results. In terms of the texture information retained from the MRI image, the results of DDCTPCA and FusionGAN show the most obvious fuzziness. Moreover, due to the fact that ASR performs fusion and denoising simultaneously, the impurities in the MRI image are eliminated in the fused image. However, some image details are blurred in the meantime. Compared with DCHWT, Structure-Aware and RCGAN, the details in our results avoid blurring and the difficulty of recognition due to darker colors, which can be seen in the fourth group.

2) Quantitative Comparison: Experiments of eight performance metrics are performed here and the results of quantitative comparisons on 20 test image pairs are shown in Fig. 11. The 20 test image pairs are of different transaxial sections of the brain-hemispheric. As for the first five metrics: EN, MG, SF, SD and PSNR, our proposed method can achieve the largest mean values with 19, 19, 10, 14 and 20 of all the 20 test pairs performing the best values, respectively. As for the metrics CC and VIF, our method also shows comparable results, generating the second largest average values and its average values merely follow behind that of DDCTPCA and that of FusionGAN respectively. As for SSIM, our method generates the fifth largest average value, the reason is that our method is designed to preserve the gradient variations in the MRI image regardless of the pixel intensity, leading to a small SSIM value between the fused intensity channel and the MRI image. Thus, it can be concluded from the statistical results that for PET and MRI image fusion, our method can also obtain relatively satisfactory results by reserving the texture information, *i.e.*, morphological information, and color information, *i.e.*, functional and metabolic information, to a great extent at the same time.

The average runtime of the 6 methods on the 20 testing image pairs is also reported in Table II.

VI. CONCLUSION

In this paper, we proposed a new deep learning-based infrared and visible image fusion method by constructing a dual-discriminator conditional GAN, named DDcGAN. It does not require the ground-truth fused images for training, and can fuse images of different resolutions without introducing thermal radiation information blurring or visible texture detail loss. Extensive comparisons on six metrics with other seven state-of-the-art fusion algorithms demonstrate that our DDcGAN can not only identify the most valuable information, but also can keep the largest or approximately the largest amount of information in the source images. Moreover, our proposed DDcGAN is applied to the fusion of PET and MRI images, and it can also achieve an advanced performance compared with five state-of-the-art algorithms.

REFERENCES

[1] M. Eslami and A. Mohammadzadeh, "Developing a spectral-based strategy for urban object detection from airborne hyperspectral TIR and visible data," *IEEE J. Sel. Topics Appl. Earth Observat., Remote Sens.*, vol. 9, no. 5, pp. 1808–1816, May 2016.

[2] C. Lopez-Molina, J. Montero, H. Bustince, and B. De Baets, "Self-adapting weighted operators for multiscale gradient fusion," *Inf. Fusion*, vol. 44, pp. 136–146, Nov. 2018.

[3] N. Yamamoto, T. Saito, S. Ogawa, and I. Ishimaru, "Middle infrared (wavelength range: 8 μm-14 μm) 2-dimensional spectroscopy (total weight with electrical controller: 1.7 kg, total cost: Less than 10,000 USD) so-called hyperspectral camera for unmanned air vehicles like drones," *Proc. SPIE*, vol. 9840, May 2016, Art. no. 984028.

[4] J. Tian *et al.*, "Carbon quantum dots/hydrogenated TiO_2 nanobelt heterostructures and their broad spectrum photocatalytic properties under UV, visible, and near-infrared irradiation," *Nano Energy*, vol. 11, pp. 419–427, Jan. 2015.

[5] X. Jin *et al.*, "A survey of infrared and visual image fusion methods," *Infr. Phys. Technol.*, vol. 85, pp. 478–501, Sep. 2017.

[6] 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.

[7] 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.

[8] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[9] H.-M. Hu, J. Wu, B. Li, Q. Guo, and J. Zheng, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2706–2719, Dec. 2017.

[10] K. He, D. Zhou, X. Zhang, R. Nie, Q. Wang, and X. Jin, "Infrared and visible image fusion based on target extraction in the nonsubsampled contourlet transform domain," *J. Appl. Remote Sens.*, vol. 11, no. 1, 2017, Art. no. 015011.

[11] Y. Bin, Y. Chao, and H. Guoyu, "Efficient image fusion with approximate sparse representation," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 14, no. 4, Jul. 2016, Art. no. 1650024.

[12] Q. Zhang, Y. Liu, R. S. Blum, J. Han, and D. Tao, "Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review," *Inf. Fusion*, vol. 40, pp. 57–75, Mar. 2018.

[13] T. Xiang, L. Yan, and R. Gao, "A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking PCNN in NSCT domain," *Infr. Phys. Technol.*, vol. 69, pp. 53–61, Mar. 2015.

[14] V. P. S. Naidu, "Hybrid DDCT-PCA based multi sensor image fusion," *J. Opt.*, vol. 43, no. 1, pp. 48–61, Nov. 2013.

[15] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infr. Phys. Technol.*, vol. 82, pp. 8–17, May 2017.

[16] M. Yin, P. Duan, W. Liu, and X. Liang, "A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation," *Neurocomputing*, vol. 226, pp. 182–191, Feb. 2017.

[17] 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.

[18] Y. Ma, J. Chen, C. Chen, F. Fan, and J. Ma, "Infrared and visible image fusion using total variation model," *Neurocomputing*, vol. 202, pp. 12–19, Aug. 2016.

[19] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[20] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," 2018, *arXiv:1804.06992*. [Online]. Available: http://arxiv.org/abs/1804.06992

[21] 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.

[22] J. Ma *et al.*, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.

[23] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: http://arxiv.org/abs/1411.1784

[24] 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. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3954–3960.

[25] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[26] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–7.

[27] G. Masi, D. Cozzolino, L. Verdoliva, and G. Scarpa, "Pansharpening by convolutional neural networks," *Remote Sens.*, vol. 8, no. 7, p. 594, Jul. 2016.

[28] 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. (ICCV)*, Oct. 2017, pp. 4724–4732.

[29] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[30] X. Gao, K. Zhang, D. Tao, and X. Li, "Image super-resolution with sparse neighbor embedding," *IEEE Trans. Image Process.*, vol. 21, no. 7, pp. 3194–3205, Jul. 2012.

[31] I. Goodfellow, "NIPS 2016 tutorial: Generative adversarial networks," 2017, *arXiv:1701.00160*. [Online]. Available: http://arxiv.org/abs/1701.00160

[32] I. Goodfellow *et al.*, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[33] H. Zhang, V. Sindagi, and V. M. Patel, "Image de-raining using a conditional generative adversarial network," 2017, *arXiv:1701.05957*. [Online]. Available: http://arxiv.org/abs/1701.05957

[34] A. Beck and M. Teboulle, "Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems," *IEEE Trans. Image Process.*, vol. 18, no. 11, pp. 2419–2434, Nov. 2009.

[35] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2018–2025.

[36] G. Huang, Z. Liu, L. V. D. Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[37] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[38] R. Matthews and M. Choi, "Clinical utility of positron emission tomography magnetic resonance imaging (PET-MRI) in gastrointestinal cancers," *Diagnostics*, vol. 6, no. 3, p. 35, Sep. 2016.

[39] F. Kogan, A. P. Fan, and G. E. Gold, "Potential of PET-MRI for imaging of non-oncologic musculoskeletal disease," *Quant. Imag. Med. Surg.*, vol. 6, no. 6, pp. 756–771, Dec. 2016.

[40] 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, Apr. 2010.

[41] Y. Gao and A. L. Yuille, "Estimation of 3D category-specific object structure: Symmetry, manhattan and/or multiple images," *Int. J. Comput. Vis.*, vol. 127, no. 10, pp. 1501–1526, Aug. 2019.

[42] C. He, Q. Liu, H. Li, and H. Wang, "Multimodal medical image fusion based on IHS and PCA," *Procedia Eng.*, vol. 7, pp. 280–285, Jan. 2010.

[43] R. A. Mandhare, P. Upadhyay, and S. Gupta, "Pixel-level image fusion using brovey transforme and wavelet transform," *Int. J. Adv. Res. Electr., Electron. Instrum. Eng.*, vol. 2, no. 6, pp. 2690–2695, 2013.

[44] V. Bhateja, H. Patel, A. Krishn, A. Sahu, and A. Lay-Ekuakille, "Multimodal medical image sensor fusion framework using cascade of wavelet and contourlet transform domains," *IEEE Sensors J.*, vol. 15, no. 12, pp. 6783–6790, Dec. 2015.

[45] R. Singh and A. Khare, "Fusion of multimodal medical images using daubechies complex wavelet transform–a multiresolution approach," *Inf. Fusion*, vol. 19, pp. 49–60, 2014.

[46] 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, Apr. 2016.

[47] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, Jul. 2016.

[48] 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 (Fusion)*, Jul. 2017, pp. 1–9.

[49] Q. Du, H. Xu, Y. Ma, J. Huang, and F. Fan, "Fusing infrared and visible images of different resolutions via total variation model," *Sensors*, vol. 18, no. 11, p. 3827, Nov. 2018.

[50] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.

[51] 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.

[52] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[53] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013.

[54] W. Lu, J. Duan, Z. Qiu, Z. Pan, R. W. Liu, and L. Bai, "Implementation of high-order variational models made easy for image processing," *Math. Methods Appl. Sci.*, vol. 39, no. 14, pp. 4208–4233, Mar. 2016.

[55] B. K. Shreyamsha 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, Aug. 2012.

[56] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Process.*, vol. 9, no. 5, pp. 347–357, May 2015.

[57] W. Li, Y. Xie, H. Zhou, Y. Han, and K. Zhan, "Structure-aware image fusion," *Optik*, vol. 172, pp. 1–11, Nov. 2018.

[58] Q. Li *et al.*, "Coupled GAN with relativistic discriminators for infrared and visible images fusion," *IEEE Sensors J.*, to be published.

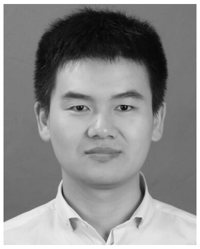
Jiayi Ma received the B.S. degree in information and computing science and the Ph.D. degree in control science and engineering from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively.

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 held a postdoctoral position at the Electronic Information School, Wuhan University, from August 2014 to November 2015, and received an accelerated promotion to Associate Professor and Full Professor in December 2015 and December 2018, respectively. He has authored or coauthored more than 120 refereed journal and conference papers, including the IEEE TPAMI/TIP/TSP/TNNLS/TIE/TGRS/TCYB/TMM/TCSVT, IJCV, CVPR, ICCV, IJCAI, AAAI, ICRA, IROS, and ACM MM. 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 the *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. She is currently pursuing the Ph.D. degree with the Multi-spectral Vision Processing Lab, Electronic Information School, Wuhan University, Wuhan.

Her current research interests include computer vision and pattern recognition.



Junjun Jiang received the B.S. degree from the Department of Mathematics, Huaqiao University, Quanzhou, China, in 2009, and the Ph.D. degree from the School of Computer, Wuhan University, Wuhan, China, in 2014.

From 2015 to 2018, he was an Associate Professor with the China University of Geosciences, Wuhan. Since 2016, he has been a Project Researcher with the National Institute of Informatics, Tokyo, Japan. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China. His research interests include image processing and computer vision.

Dr. Jiang won the Finalist of the World's FIRST 10K Best Paper Award at ICME 2017, and the Best Student Paper Runner-up Award at MMM 2015. He received the 2016 China Computer Federation (CCF) Outstanding Doctoral Dissertation Award and 2015 ACM Wuhan Doctoral Dissertation Award.

Xiaoguang Mei received the B.S. degree in communication engineering from the Huazhong University of Science and Technology (HUST), Wuhan, China, in 2007, the M.S. degree in communications and information systems from Central China Normal University, Wuhan, in 2011, and the Ph.D. degree in circuits and systems from HUST in 2016.

From 2010 to 2012, he was a Software Engineer with the 722 Research Institute, China Shipbuilding Industry Corporation, Wuhan. From 2016 to 2019, he was a Postdoctoral Fellow of the Electronic Information School, Wuhan University (WHU), Wuhan, where he is currently an Associate Professor. His current research interests include hyperspectral imagery, machine learning, and pattern recognition.

Xiao-Ping Zhang (Senior Member, IEEE) received the B.S. and Ph.D. degrees from Tsinghua University, in 1992 and 1996, respectively, both in electronic engineering, 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 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, respectively. He is a frequent consultant for biotech companies and investment firms. He is the Co-Founder and CEO of 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 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, 2021. He is the General Co-Chair of the 2017 GlobalSIP Symposium on Signal and Information Processing for Finance and Business, and the General Co-Chair of 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, 2015. He is the Publicity Chair of the International Conference on Multimedia and Expo 2006, and the Program Chair for International Conference on Intelligent Computing, in 2005 and 2010, respectively. 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 for 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 has served as a Guest Editor for the *Multimedia Tools and Applications* and the *International Journal of Semantic Computing*. He is awarded as an IEEE Distinguished Lecturer for the term from January 2020 to December 2021 by the IEEE Signal Processing Society. He is a registered Professional Engineer in Ontario, Canada.