Information Fusion 48 (2019) 11–26

Contents lists available at ScienceDirect

Information Fusion

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

FusionGAN: A generative adversarial network for infrared and visible image fusion

Jiayi Ma[a], Wei Yu[a], Pengwei Liang[a], Chang Li[b], Junjun Jiang[c,*]

[a] Electronic Information School, Wuhan University, Wuhan 430072, China
[b] Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China
[c] School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Keywords:
Image fusion
Infrared image
Visible image
Generative adversarial network
Deep learning

ABSTRACT

Infrared images can distinguish targets from their backgrounds on the basis of difference in thermal radiation, which works well at all day/night time and under all weather conditions. By contrast, visible images can provide texture details with high spatial resolution and definition in a manner consistent with the human visual system. This paper proposes a novel method to fuse these two types of information using a generative adversarial network, termed as *FusionGAN*. Our method establishes an adversarial game between a generator and a discriminator, where the generator aims to generate a fused image with major infrared intensities together with additional visible gradients, and the discriminator aims to force the fused image to have more details existing in visible images. This enables that the final fused image simultaneously keeps the thermal radiation in an infrared image and the textures in a visible image. In addition, our FusionGAN is an end-to-end model, avoiding manually designing complicated activity level measurements and fusion rules as in traditional methods. Experiments on public datasets demonstrate the superiority of our strategy over state-of-the-arts, where our results look like sharpened infrared images with clear highlighted targets and abundant details. Moreover, we also generalize our FusionGAN to fuse images with different resolutions, say a low-resolution infrared image and a high-resolution visible image. Extensive results demonstrate that our strategy can generate clear and clean fused images which do not suffer from noise caused by upsampling of infrared information.

1. Introduction

Image fusion is an enhancement technique that aims to combine images obtained by different kinds of sensors to generate a robust or informative image that can facilitate subsequent processing or help in decision making [1,2]. Particularly, multi-sensor data such as thermal infrared and visible images has been used to enhance the performance in terms of human visual perception, object detection, and target recognition [3]. For example, infrared images capture thermal radiation, whereas visible images capture reflected light. These two types of images can provide scene information from different aspects with complementary properties, and they are also inherent in nearly all objects [4].

The image fusion problem has been developed with different schemes including multi-scale transform- [5–7], sparse representation- [8,9], neural network- [10,11], subspace-[12,13], and saliency-based [14,15] methods, hybrid models [16,17], and other methods [18,19]. Nevertheless, the major fusion framework involves three key components, including image transform, activity level measurement, and fusion rule designing [20]. Existing methods typically use the same transform or representation for different source images during the fusion process. However, it may not be appropriate for infrared and visible images, as the thermal radiation in infrared images and the appearance in visible images are manifestations of two different phenomena. In addition, the activity level measurement and fusion rule in most existing methods are designed in a manual way, and they have become more and more complex, having the limitations of implementation difficulty and computational cost [21].

To overcome the above mentioned issues, in this paper we propose an infrared and visible image fusion method from a novel perspective based on generative adversarial network (*FusionGAN*), which formulates the fusion as an adversarial game between keeping the infrared thermal radiation information and preserving the visible appearance texture information. More specifically, it can be seen as a minimax problem between a generator and a discriminator. The generator attempts to generate a fused image with major infrared intensities together with additional visible gradients, while the discriminator aims to force the

* Corresponding author.
E-mail addresses: jyma2010@gmail.com (J. Ma), yuwei998@whu.edu.cn (W. Yu), erfect@whu.edu.cn (P. Liang), lichang1214@gmail.com (C. Li), junjun0595@163.com (J. Jiang).

https://doi.org/10.1016/j.inffus.2018.09.004
Received 10 February 2018; Received in revised form 30 June 2018; Accepted 3 September 2018
Available online 5 September 2018

Fig. 1. Schematic illustration of image fusion. From left to right: the infrared image, the visible image, the fusion result of a classic method, and the fusion result of our proposed FusionGAN. Our result can simultaneously keep the thermal radiation distribution in the infrared image and the appearance textures in the visible image.

fused image to have more texture details. This enables our fused image to maintain the thermal radiation in an infrared image and the texture details in a visible image at the same time. In addition, the end-to-end property of generative adversarial networks (GANs) can avoid manually designing complicated activity level measurements and fusion rules.

To show the major superiority of our method, we give a representative example in Fig. 1. The left two images are the infrared and visible images to be fused, where the visible image contains detailed background and the infrared image highlights the target, *i.e.* the water. The third image is the fusion result by using a recent method [22]. Clearly, this traditional method is just able to keep more texture details in source images, and the property of high contrast between target and background in the infrared image cannot be preserved in the fused image. In fact, the key information in the infrared image (*i.e.*, the thermal radiation distribution) is totally lost in the fused image. The rightmost image in Fig. 1 is the fusion result by our FusionGAN. In contrast, our result preserves the thermal radiation distribution in the infrared image, and hence the target can be easily detected. Meanwhile, the details of the background (*i.e.*, the trees, road and water plants) in the visible image are also well retained.

The main contributions of this work lie in the following four folds. First, we propose a generative adversarial architecture and design a loss function specialized for infrared and visible image fusion. The feasibility and superiority of GANs used for image fusion are also discussed. To the best of our knowledge, it is the first time that the GANs are adopted for addressing the image fusion task. Second, the proposed FusionGAN is an end-to-end model, where the fused image can be generated automatically from input source images without manually designing the activity level measurement or fusion rule. Third, we conduct experiments on public infrared and visible image fusion datasets with qualitative and quantitative comparisons to state-of-the-art methods. Compared to previous methods, the proposed FusionGAN can obtain results looking like sharpened infrared images with clear highlighted targets and abundant textures. Last but not the least, we generalize the proposed FusionGAN to fuse source images with different resolutions such as low-resolution infrared images and high-resolution visible images. It can generate high-resolution resulting images which do not suffer from noise caused by upsampling of infrared information.

The rest of this paper is arranged as follows. Section 2 describes background material and related work on GAN. In Section 3, we present our FusionGAN algorithm for infrared and visible image fusion. Section 4 illustrates the fusion performance of our method on various types of infrared and visible image/video pairs with comparisons to other approaches. We discuss the explainability of our FusionGAN in Section 5, followed by some concluding remarks in Section 6.

2. Related work

In this section, we briefly introduce the background material and relevant works, including traditional infrared and visible image fusion methods, deep learning based fusion techniques, as well as GANs and their variants.

2.1. Infrared and visible image fusion

With the fast-growing demands of image representation methods, there are quantities of image fusion methods that have been proposed. They can be simply divided into seven categories including multi-scale transform-[5–7], sparse representation-[8,9], neural network-[10,11], subspace-[12,13], and saliency-based [14,15] methods, hybrid models [16,17], and other methods [18,19]. Next, we briefly discuss the main ideas of these methods.

Multi-scale transform-based methods are the most popular in image fusion, and multi-scale transform can decompose original images into components of different scales, where each component represents the sub-image at each scale and real-world objects typically comprise components at different scales [23]. In general, infrared and visible image fusion schemes based on multi-scale transforms comprise three steps [23]. First, each source image is decomposed into a series of multi-scale representations. Then, the multi-scale representations of the source image are fused according to a given fusion rule. Finally, the fused image is acquired using corresponding inverse multi-scale transforms on the fused representations. Sparse representation image fusion methods aim to learn an over-complete dictionary from a large number of high-quality natural images. Then, the source images can be sparsely represented by the learned dictionary, thereby potentially enhancing the representation of meaningful and stable images [24]. Meanwhile, sparse representation-based fusion methods divide source images into several overlapping patches using a sliding window strategy, thereby potentially reducing visual artifacts and improving robustness to misregistration [16]. Neural network-based methods imitate the perception behavior of the human brain to deal with neural information, the interactions among neurons characterize the transmission and processing of neuron information, and the neural network has the advantages of strong adaptability and fault tolerance and anti-noise capabilities, most neural network-based infrared and visible image fusion methods adopt the pulse-coupled neural network or its variants [10]. Subspace-based methods aim to project high-dimensional input images into low-dimensional spaces or subspaces. For most natural images, redundant information exists and low-dimensional subspaces can help capture the intrinsic structures of the original images. Thus, subspace-based methods, including principal component analysis, non-negative matrix factorization, and independent component analysis, have been successfully applied in infrared and visible image fusion [12]. Saliency-based methods are based on the fact that the attention is often captured by objects or pixels that are more significant than their neighbors, and saliency-based fusion methods can maintain the integrity of the salient object region and improve the visual quality of the fused image [14]. The above mentioned infrared and visible image fusion methods all have their advantages and disadvantages, and hybrid models combine their advantages to improve the image fusion performance [16]. Other infrared and visible image fusion methods can inspire new ideas and perspectives for image fusion, which are based on total variation [18], fuzzy theory [25], entropy [19] and so on.

2.2. Deep learning based image fusion

In recent years, deep learning has also been successfully applied to image fusion, due to its strong ability of extracting image features. In multi-focus image fusion, Liu et al. [26] trained a deep convolutional neural network (CNN) to jointly generate activity level measurement and fusion rule, and they also applied their model to fuse infrared and visible images [27]. In multi-modality image fusion, Zhong et al. [28] proposed a joint image fusion and super-resolution method based on CNN. Besides, Liu et al. [29] introduced the convolution sparse representation for image fusion, in which deconvolutional networks intend to build a hierarchy of layers, and each layer consists of an encoder and a decoder. In remote sensing image fusion, Masi et al. [30] proposed an effective three-layer architecture to solve the pansharpening problem, where the input is augmented by adding several maps of nonlinear radiometric indices to promote the fusion performance.

The existing deep learning based image fusion techniques typically rely on the CNN model, which has a critical prerequisite that the ground truth should be available in advance. For the multi-focus image fusion and pansharpening problems, the ground truth is well defined, for example, a clear image without blurred regions or a multispectral image with the same resolution as the corresponding panchromatic image. However, in the task of infrared and visible image fusion, defining a standard for fused images is unrealistic, and hence, establishing the ground truth is not considered. On this basis, rather than learning an end-to-end model which requires ground truth fused images, the existing techniques for infrared and visible image fusion just learn a deep model to determine the blurring degree of each patch in the source images, and then calculate a weight map accordingly to generate the final fused image [27]. In this paper, we formulate the fusion problem in the framework of GAN, which does not suffer from the aforementioned problem.

2.3. Generative adversarial networks and its variants

GAN is a popular framework for estimating generative models via an adversarial process, and deep convolutional GANs (DCGANs) successfully introduce a class of CNNs into GANs, while the least squares generative adversarial networks (LSGANs) overcome the vanishing gradients problem in regular GANs, which are more stable during the learning process. Next, we will briefly introduce the above mentioned three related techniques.

2.3.1. Generative adversarial networks

Goodfellow et al. [31] first proposed the concept of GAN, which has drawn substantial attention in the field of deep learning. The GAN is based on the minimax two-player game, which can provide a simple yet powerful way to estimate target distribution and generate new samples. The GAN framework consists of two adversarial models: a generative model G and a discriminative model D. The generative model G can capture the data distribution, and the discriminative model D can estimate the probability that a sample comes from the training data rather than G. More specifically, the GAN establishes an adversarial game between a discriminator and a generator, the generator takes the noise whose prior distribution is P_z as input and tries to generate different samples to fool the discriminator, and the discriminator aims to determine whether a sample is from the model distribution or the data distribution, finally the generator generates samples that are not distinguishable by the discriminator.

Mathematically, a generative model G aims to generate samples, whose distribution (P_G) tries to approximate the distribution (P_{data}) of real training data, G and D play the minimax two-player game as follows:

$$\min_G \max_D V_{\mathrm{GAN}}(G, D) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))]. \tag{1}$$

However, P_G cannot be represented explicitly, and D must be synchronized well with G during training. Thus, regular GANs are unstable, and it is hard to train a good model by regular GANs.

2.3.2. Deep convolutional GANs

The technique of Deep convolutional GANs (DCGANs) was first proposed by Radrord et al. [32]. DCGANs first successfully introduced CNNs, which can bridge the gap between CNNs for supervised learning and GANs for unsupervised learning. Since traditional GANs are unstable to train a good model, so the architecture of CNNs should be designed appropriately to make the traditional GANs more stable, and there are mainly five differences compared with traditional CNNs. First, the pooling layers are not used in both the generator and the discriminator. Instead, strided convolutions are applied in discriminator to learn its own spatial downsampling, and fractional-strided convolutions are used in generator to realize upsampling. Second, batchnormalization layers are introduced into both the generator and the discriminator. Since poor initialization always tends to create a lot of training problems, batchnormalization layers are able to solve these problems and avoid vanishing gradient in deeper models. Third, fully connected layers are removed in deeper models. Fourth, all activation layers in generator are rectified linear unit (ReLU) except the last activation layer, and the last layer is tanh activation. Last but not the least, all activation layers in discriminator are leaky ReLU activations. Thus, the training process becomes more steady, and the quality of generated results can be improved.

2.3.3. Least squares GANs

Despite the great success GANs have achieved, there still exits two critical issues to be solved. The first is how to improve the quality of generated images. In recent years, many works have been proposed to solve this problem, such as DCGANs. The second is how to improve the stability of training process. A lot of works have been proposed to deal with this problem by exploring the objective functions of GANs, such as Wasserstein GANs (WGANs) [33], which converge much slowly than regular GANs [34]. In addition, the regular GANs adopt the sigmoid cross entropy loss function for the discriminator, which may lead to the gradient-vanishing problem during the learning process. To overcome the above mentioned two problems, Mao et al. [34] proposed the least squares generative adversarial networks (LSGANs), which adopt the least squares loss function for the discriminator, and the objective functions for LSGANs are defined as follows

$$\begin{aligned}\min_D V_{\mathrm{LSGAN}}(D) &= \frac{1}{2}\mathbb{E}_{x\sim p_{data}(x)}[(D(x) - b)^2] + \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z)) - a)^2],\\ \min_G V_{\mathrm{LSGAN}}(G) &= \frac{1}{2}\mathbb{E}_{z\sim p_z(z)}[(D(G(z)) - c)^2],\end{aligned} \tag{2}$$

where the coding scheme is used for both the discriminator and generator, a and b denote the labels for fake data and real data, respectively, and c denotes the value that the generator wants the discriminator to believe for fake data. There are two methods to determine the values of a, b and c in Eq. (2). The first is to set the $b - c = 1$ and $b - a = 2$, thus minimizing Eq. (2) yields minimizing the Pearson χ^2 between $P_{data} + P_g$ and P_g. The second is to set $c = b$, which can make samples generated by generator as real as possible. The above mentioned two methods usually get similar performance.

In LSGANs, penalizing the samples that lie in a long way to the decision boundary make the samples generated by generator close to the decision boundary and generate more gradients. Thus, LSGANs have two advantages over regular GANs. On the one hand, LSGANs can generate higher quality images than regular GANs. On the other hand, LSGANs perform more stable than regular GANs during the training process.

3. Method

In this section, we describe the proposed FusionGAN for infrared and visible image fusion. We start by laying out the problem formulation

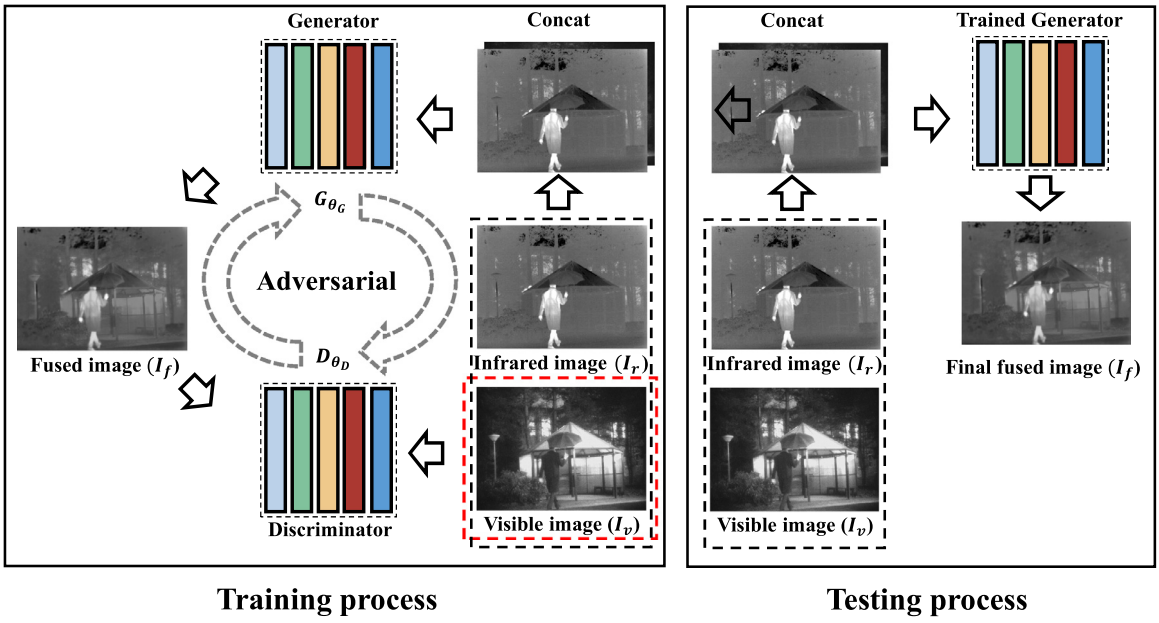

Fig. 2. Framework of the proposed FusionGAN for infrared and visible image fusion.

with GANs, and then discuss the network architectures of the generator and the discriminator. Finally, we provide some details for the network training.

3.1. Problem formulation

To keep the thermal radiation information of infrared image and the abundant texture information of visible image simultaneously, we propose a new fusion strategy from a novel perspective. We formulate the infrared and visible image fusion problem as an adversarial problem, as schematically illustrated in Fig. 2 (a). At the beginning, we concatenate the infrared image I_r and the visible image I_v in the channel dimension. Then, the concatenated image is fed into the generator G_{θ_G}, and the output of G_{θ_G} is a fused image I_f. Due to the loss function of the generator designed in this paper (explained later), without the discriminator D_{θ_D}, I_f tends to keep the thermal radiation information of infrared image I_r and preserve the gradient information of visible image I_v. After that, we input the fused image I_f and the visible image I_v into the discriminator D_{θ_D}, which aims to distinguish I_f from I_v. The proposed FusionGAN establishes an adversarial game between the generator G_{θ_G} and the discriminator D_{θ_D}, and I_f will gradually contain more and more detail information in visible image I_v. During the training phase, once the generator G_{θ_G} generates samples (*i.e.,* I_f) that cannot be distinguished by the discriminator D_{θ_D}, we can obtain the expected fused image I_f. The testing process is shown in Fig. 2 (b), we only input the concatenating image of I_f and I_v into the trained generator G_{θ_G}, and the output of G_{θ_G} is our final fused result.

Loss function. The loss function of our FusionGAN is consist of two parts, *i.e.*, the loss function of generator G_{θ_G} and the loss function of discriminator D_{θ_D}. In the following, we will introduce them separately. First, the loss function of generator G_{θ_G} consists of two terms:

$$\mathcal{L}_G = V_{\text{FusionGAN}}(G) + \lambda \mathcal{L}_{\text{content}}, \tag{3}$$

where $\mathcal{L}_G$ denotes the total loss, the first term $V_{\text{FusionGAN}}(G)$ on the right hand denotes the adversarial loss between generator G_{θ_G} and discriminator D_{θ_D}, which is defined as follows:

$$V_{\text{FusionGAN}}(G) = \frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_f^n) - c\right)^2, \tag{4}$$

where I_f^n denotes the fused image with $n \in \mathbb{N}_N$ and N denoting the number of fused images, and c is the value that generator wants discriminator to believe for fake data.

The second term $\mathcal{L}_{\text{content}}$ represents the content loss, and λ is used to strike a balance between $V_{\text{FusionGAN}}(G)$ and $\mathcal{L}_{\text{content}}$. Due to that the thermal radiation information of infrared image is characterized by its pixel intensities and the texture detail information of visible image can be partly characterized by its gradients [18], we enforce the fused image I_f to have similar intensities as I_r and similar gradients as I_v. Specifically, $\mathcal{L}_{\text{content}}$ is defined as follows:

$$\mathcal{L}_{\text{content}} = \frac{1}{HW}(\|I_f - I_r\|_F^2 + \xi\|\nabla I_f - \nabla I_v\|_F^2), \tag{5}$$

where H and W represents the height and width of the input images, respectively, $\|\cdot\|_F$ stands for the matrix Frobenius norm, and ∇ means the gradient operator. The first term of $\mathcal{L}_{\text{content}}$ aims to keep the thermal radiation information of infrared image I_r in the fused image I_f, the second term of $\mathcal{L}_{\text{content}}$ aims to preserve the gradient information contained in the visible image I_v, and ξ is a positive parameter controlling the trade-off between two terms.

Actually, without D_{θ_D}, we can also get a fused image, which can keep thermal radiation information in the infrared image and gradient information in the visible image. But that is often not enough, because the texture details in a visible image cannot be totally represented by using only gradient information (we will validate this issue in our experiments). Therefore, we establish an adversarial game between a generator G_{θ_G} and a discriminator D_{θ_D} to adjust the fused image I_f based on the visible image I_v. This can make I_f contain more texture details. Formally, the loss function of discriminator D_{θ_D} is defined as follows:

$$\mathcal{L}_D = \frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_v) - b\right)^2 + \frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_f) - a\right)^2, \tag{6}$$

where a and b denote the labels of fused image I_f and visible image I_v, respectively, $D_{\theta_D}(I_v)$ and $D_{\theta_D}(I_f)$ denote the classification results of the visible and fused images, respectively. The discriminator is designed to distinguish the fused images from the visible images based on the features extracted from them. We use the least square loss function, which obeys minimizing the Pearson χ^2 divergence. It makes the training process more steady and the loss function of discriminator converge quickly.

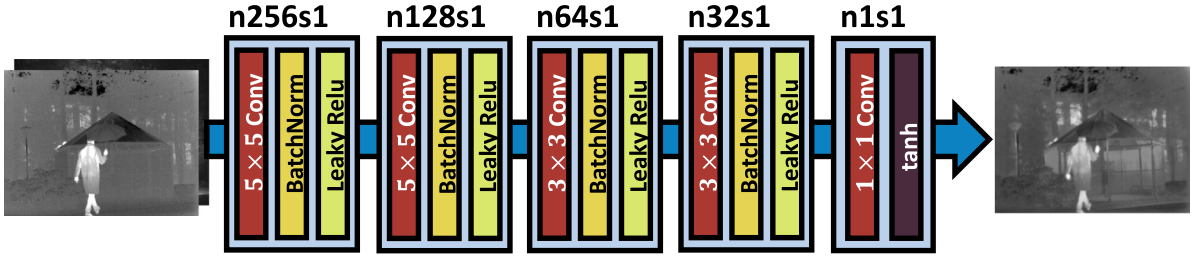

Fig. 3. Network architecture of generator G_{θ_G}. G_{θ_G} is a simple five-layer convolution neural network with 5 convolution layers, 4 batch normalization layers, 4 leaky ReLU activation layers, and 1 tanh activation layer.

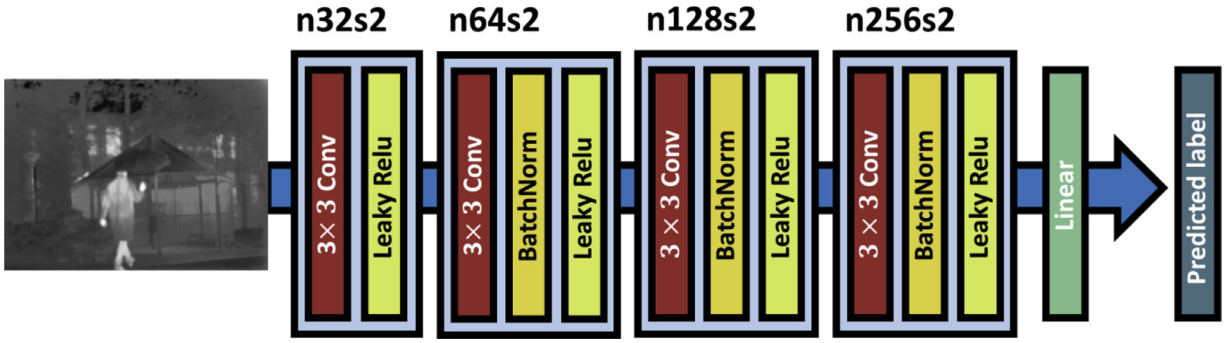

Fig. 4. Network architecture of discriminator D_{θ_D}. D_{θ_D} is a simple five-layer convolution neural network with 4 convolution layers to extract feature maps of input, 1 linear layer to do the classification, 4 batch normalization layers, and 4 leaky ReLU activation layers.

3.2. Network architecture

Our network architecture consists of two parts, *i.e.*, the generator and the discriminator. Architectures of them are designed based on the convolution neural network.

Network architecture of generator. Our network architecture of generator G_{θ_G} is presented in Fig. 3. As shown, G_{θ_G} is a simple five-layer convolution neural network, where 5×5 filters are used in the first and second layers, 3×3 filters in the third and fourth layers, and 1×1 filters in the last layer. The stride in each layer is set to 1, and there is no padding operation in convolution. Our generator's input is a concatenated image without noise. To improve the diversity of generated images, many works usually extract feature maps of input image by convolution layers, and then reconstruct an image to the same size of input image by transposed convolution layer. For infrared and visible image fusion, every down-sampling process will drop out some detail information in the source images, which is important for fusion [35]. Therefore, we only introduce the convolution layer without down-sampling. This can also keep the sizes of input and output the same, and hence, the transposed convolution layer is unnecessary in our network. In addition, to avoid the problem of vanishing gradient, we follow the rules of deep convolutional GAN [32] for batch normalization and activation function. To overcome the sensitivity to data initialization, we employ the batch normalization in the first four layers, and the batch normalization layer can make our model more stable and also can help the gradients to back propagate to every layer effectively. For activation function, we use leaky ReLU activation function in the first four layers, and the tanh activation function is used in the last layer.

Network architecture of discriminator. Our network architecture of discriminator D_{θ_D} is a simple five-layer convolution neural network, which is shown in Fig. 4. From the first layer to the fourth layer, we use 3×3 filters in convolution layers and set the stride to 2 without padding. This is different from the generator network. The underlying reason is that the discriminator is a classifier, and it first extracts feature maps from the input images and then classifies them. Therefore, it works the same way as pooling layer by setting the stride to 2. In order not to introduce noise in our model, we only perform padding operation on input images in the first layer, and no padding is performed in the rest three convolution layers. From the second layer to fourth layer, we use batch normalization layer. In addition, we use leaky ReLU activation function in the first four layers. The last layer is linear layer, which is mainly used for classification.

3.3. Training details

We select 45 pairs of infrared and visible images with different scenes from TNO database as our training data. However, it is insufficient to train a good model, so we crop each image by setting the stride to 14, and each patch is of the same size 120×120.[1] Thus, we can get 64,381 pairs of infrared and visible patches, and make them centralized to $[-1, 1]$. We select m pairs of infrared and visible patches from the training data, and then pad them to the size of 132×132, which are used as the input of the generator. The fused image patch output by the generator is of size 120×120. Next, we use m pairs of visible and fused image patches as the input of discriminator. We first train the discriminator k times, and the optimizer solver is Adam [36], then we train the generator until reaching the maximum number of training iterations. The procedure is summarized in Algorithm 1 . In the testing process, we crop the testing data without overlapping, and input them into the generator G_{θ_G} as a batch. Then we connect the results of generator according to the sequence of cropping, thus we can get the final fused image. The parameter setting will be discussed in the next section.

4. Experiments

In this section, we first briefly introduce the fusion metrics used in this paper and then demonstrate the efficacy of the proposed FusionGAN on public datasets, and compare it with eight state-of-the-art fusion

[1] Note that using smaller image patch size will make training process time-saving. However, for infrared images, small patch size usually leads to few valid information. For example, the image patch may be just a black or white patch without textures. Clearly, this will adversely affect the training process.

Algorithm 1: Training procedure of FusionGAN.

for number of training iterations **do**
  for *k steps* **do**
    Select m fusion patches $\{I_f^{(1)}, \cdots, I_f^{(m)}\}$ from G_{θ_G};
    Select m visible patches $\{I_v^{(1)}, \cdots, I_v^{(m)}\}$;
    Update discriminator by
    AdamOptimizer: $\nabla_{\theta_D}\left(\frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_v)-b\right)^2 + \frac{1}{N}\sum_{n=1}^{N}\left(D_{\theta_D}(I_f)-a\right)^2\right)$;
  end
  Select m infrared patches $\{I_r^{(1)}, \cdots, I_r^{(m)}\}$ and m visible patches $\{I_v^{(1)}, \cdots, I_v^{(m)}\}$ from training data;
  Update generator by AdamOptimizer: $\nabla_{\theta_G}\mathcal{L}_G$;
end

methods including adaptive sparse representation (ASR) [37], curvelet transform (CVT) [38], dual-tree complex wavelet transform (DTCWT) [39], fourth order partial differential equation (FPDE) [12], guided filtering based fusion (GFF) [22], ratio of low-pass pyramid (LPP) [3], two-scale image fusion based on visual saliency (TSIFVS) [40], and gradient transfer fusion (GTF) [18]. The implementation of all these eight methods are publicly available, and we set the parameters of competing methods according to the original papers. Subsequently, in order to verify the importance of adversarial training, we train two models based on whether to use adversarial loss and compare their fusion performances. Finally, we generalize our FusionGAN to fuse images with different resolutions such as low-resolution infrared images and high-resolution visible images, and we also compare it with the aforementioned eight state-of-the-art fusion methods. All the experiments are conducted on a desktop with 2.4 GHz Intel Xeon CPU E5-2673 v3, GeForce GTX 1080Ti, and 64 GB memory.

4.1. Fusion metrics

Since it is difficult to get an accurate evaluation of the fusion performance only by subjective evaluation, we also need fusion metrics for objective evaluation. In recent years, many studies have proposed various fusion metrics, but it seems that none of them is definitely better than the others [41]. Therefore, it is necessary to choose multiple metrics to evaluate different fusion methods. We quantitatively evaluate the performances of different fusion methods using six metrics, *i.e.*, entropy (EN) [42], standard deviation (SD) [43], structural similarity index measure (SSIM) [44], correlation coefficient (CC) [45], spatial frequency (SF) [46], and visual information fidelity (VIF) [47]. The definitions of these six metrics are as follows.

EN is defined based on information theory, which measures the amount of information the fused image contains. Mathematically, EN is defined as follows:

$$\mathrm{EN} = -\sum_{l=0}^{L-1} p_l \log_2 p_l, \tag{7}$$

where L denotes the number of gray level, we set it to 256 in our experiments. p_l is the normalized histogram of corresponding gray level in the fused image. The larger entropy is, the more information fused image contains, and the better performance fusion method achieves.

SD is defined based on the statistical concept, which reflects the extent to which the values of individual pixels in the image from the average value. Mathematically, SD is defined as follows:

$$\mathrm{SD} = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-\mu)^2}, \tag{8}$$

where F is the fused image with dimensions of $M \times N$, and μ is the mean value of the fused image F. The regions with high contrast always attract the attention of human, and the fused image with higher contrast often leads to larger SD, which means that the fused image achieves better visual effect.

SSIM is used to model the image loss and distortion, which measures the structural similarity between source images and fused image. SSIM mainly consist of three components: loss of correlation, luminance distortion, and contrast distortion. The product of three components is the assessment result of the fused image, and SSIM is defined as follows:

$$\mathrm{SSIM}_{X,F} = \sum_{x,f}\frac{2\mu_x\mu_f + C_1}{\mu_x^2+\mu_f^2+C_1}\cdot\frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2+\sigma_f^2+C_2}\cdot\frac{\sigma_{xf}+C_3}{\sigma_x\sigma_f+C_3}, \tag{9}$$

$$\mathrm{SSIM} = \mathrm{SSIM}_{A,F} + \mathrm{SSIM}_{B,F}, \tag{10}$$

where $\mathrm{SSIM}_{X,F}$ denotes the structural similarity between source image X and fused image F, x and f represent the image patch of the source image and fused image in a local window of size $M \times N$, σ_x and σ_f denote the standard deviation, σ_{xf} is the standard covariance correlation of source and fused image, μ_x and μ_f denote the mean value of source image and fused image. C_1 and C_2 and C_3 are parameters to make the algorithm stable. And $\mathrm{SSIM}_{A,F}$ and $\mathrm{SSIM}_{B,F}$ denote the structural similarities between infrared/visible images and fused image. The larger value of SSIM indicates better performance.

CC measures the degree of linear correlation of the fused image and source images and is defined as follows:

$$CC = \frac{(r_{AF} + r_{BF})}{2}, \tag{11}$$

where $r_{XF} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\bar{X})(F(i,j)-\mu)}{\sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-\bar{X})^2(\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-\mu)^2)}}$ and $\bar{X}$ denotes the mean value of source image X. The larger the CC, the more similar the fused image is to the source images and the better the fusion performance.

SF is designed to measure the gradient distribution of an image, which is defined as follows:

$$SF = \sqrt{RF^2 + CF^2}, \tag{12}$$

where RF is spatial row frequency defined as $RF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i,j-1))^2}$, and CF is column frequency defined as $CF = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N}(F(i,j)-F(i-1,j))^2}$. The larger SF is, the richer edges and textures the fused image has.

VIF measures the information fidelity of the fused image, and it is computed by four steps: firstly, the source images and fused image are divided into different blocks; then, evaluate the visual information of each block with and without distortion; subsequently, evaluate the VIF for each sub-band; finally, calculate the overall metric based on VIF.

4.2. Experimental validation of fusion performance

4.2.1. Databases and training settings

In this experiment, we first focus on qualitative and quantitative comparisons on the fusion performance of different methods on the surveillance images from TNO Human Factors, which contain multispectral nighttime imagery of different military relevant scenarios, registered with different multiband camera systems.[2] We choose seven typical pairs for qualitative illustration, *i.e., Bunker, Bench, Sandpath, Kaptein_1123, Kaptein_1654, Marne_04*, and *Nato_camp*. The Nato_camp is an image sequence, which contains 32 image pairs, and this sequence is also used for quantitative comparison. In addition, we also test our method on the INO database,[3] which is provided by the National Optics Institute of Canada, and contains several pairs of visible and infrared

[2] Available at https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.
[3] Available at https://www.ino.ca/en/video-analytics-dataset/.

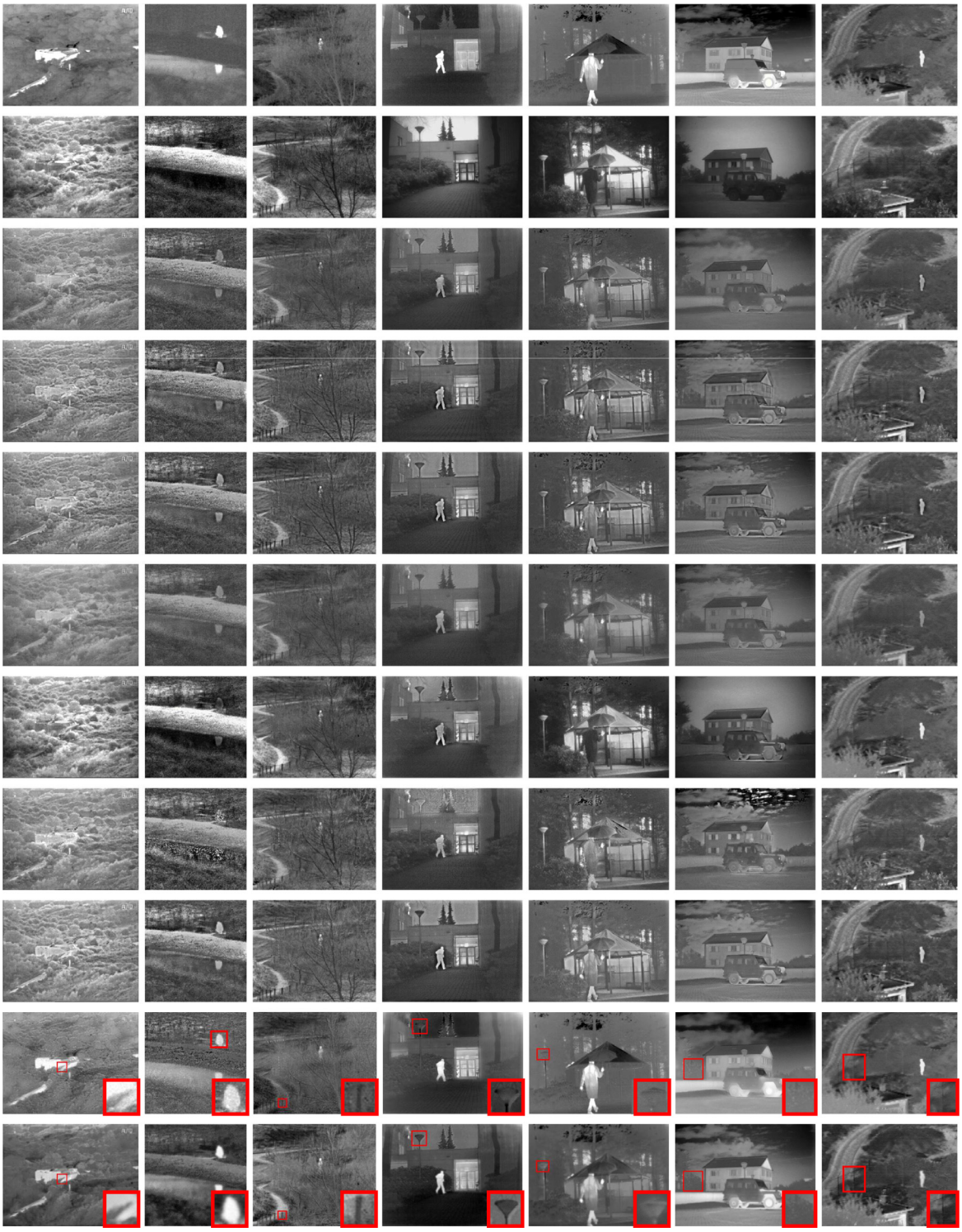

Fig. 5. Qualitative fusion results on seven typical infrared and visible image pairs from the TNO database. From left to right: *Bunker, Bench, Sandpath, Kaptein_1123, Kaptein_1654, Marne_04*, and *Nato_camp*. From top to bottom: infrared images, visible images, results of ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS, GTF and our FusionGAN. Note that in the last two rows, for clear comparison we select a small region (*i.e.*, the red box) in each fused image, and then zoom in it and put it in the bottom right corner. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

videos representing different scenarios captured under different weather conditions. Specifically, we capture 31 visible and infrared image pairs from the video named *Trees and runner* for both qualitative and quantitative comparisons.

Our FusionGAN is trained on the TNO database, from which we choose 45 infrared and visible images from different scenes, and our training parameters are set as follows: the size of batch images m is set to 32, the number of training iterations is set to 10, and the discriminator training step k is set to 2. λ is set to 100, ξ is set to 8, and the learning rate is set to 10^{-4}. The label a of fused image is a random number ranging from 0 to 0.3, the label b of visible image is a random number ranging from 0.7 to 1.2, and the label c is also a random number ranging from

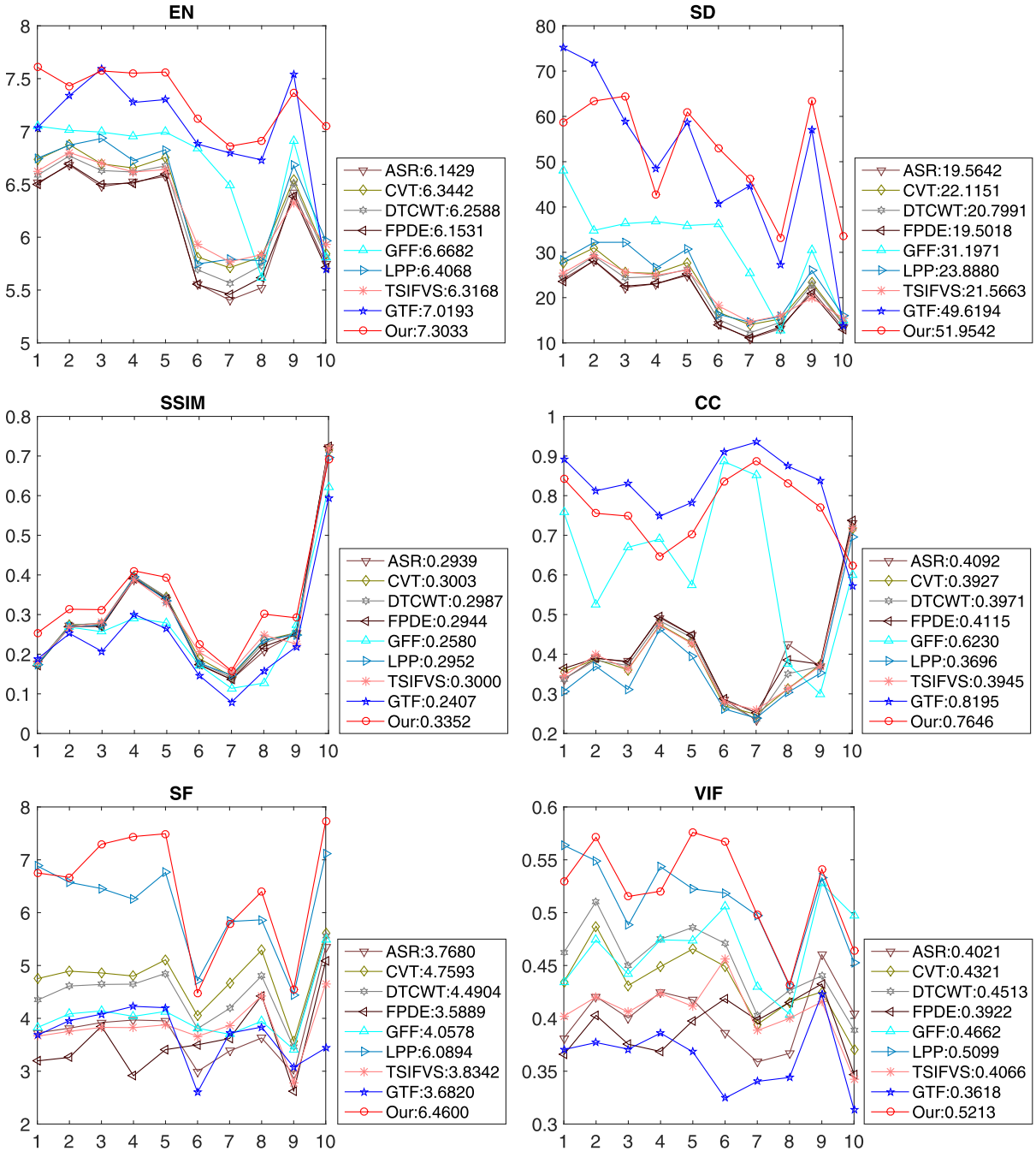

Fig. 6. Quantitative comparisons of the six metrics, *i.e.*, EN, SD, SSIM, CC, SF and VIF, on ten image pairs from the TNO database. The eight state-of-the-art methods such as ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS and GTF are used for comparison.

0.7 to 1.2. The labels *a, b* and *c* are not specific numbers, which are the so-called soft labels [34].

4.2.2. Results on TNO database

Qualitative comparisons. To give some intuitive results on the fusion performance, we select seven typical image pairs for qualitative evaluation, *Bunker, Bench, Sandpath, Kaptein_1123, Kaptein_1654, Marne_04*, and *Nato_camp*. The fusion results of the proposed FusionGAN and the other eight comparison methods are shown in Fig. 5. The first two rows in Fig. 5 present the original infrared images and visible images, the last row is the fusion results of our FusionGAN, and the rest eight rows correspond to the fusion results of eight comparison methods. From the results, we see that all methods can well fuse the information of visible image and infrared image to some extent. In this sense, we cannot judge which method is the best or worst. However, we also find that for the comparison methods except GTF, the targets (*e.g.*, the building, human, or the car) in the fused images are not that obvious, which means that the thermal radiation information in the infrared images is not well preserved. This can be attributed to that the comparison methods all focus on exploiting the detail information in the source images.

In contrast, GTF and our method can highlight the target regions in the fused images better than those in the visible images, which is beneficial for automatic target detection and localization. Both our

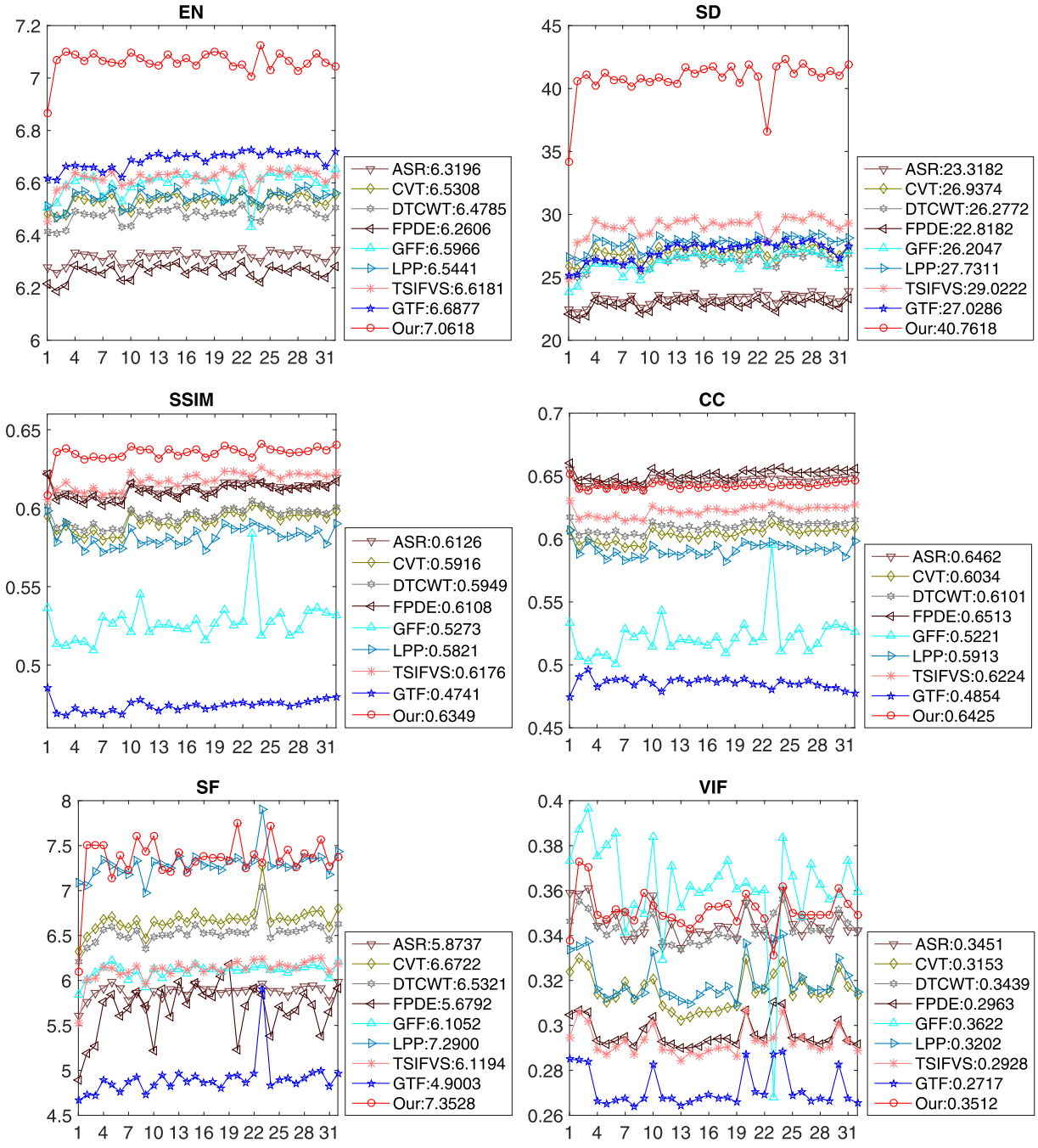

Fig. 7. Quantitative comparisons of the six metrics, *i.e.*, EN, SD, SSIM, CC, SF and VIF, on the *Nato_camp* sequence from the TNO database. The eight state-of-the-art methods such as ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS and GTF are used for comparison.

method and GTF can well preserve the thermal radiation. Nevertheless, compared with GTF, the fusion results of our FusionGAN obviously contain more abundant detail information, and our results are more appropriate for human visual perception. For example, in *Kaptein_1123*, the human is equally highlighted in two results, but the lamp on the street is much more clear in our result. In *Marne_04*, the tree is fused appropriately in our FusionGAN result; however, in the result of GTF, it is almost impossible to be recognized. Similar phenomenon can also be observed in the other five examples, as shown in the red boxes. This demonstrates that our FusionGAN has better performance than the other state-of-the-arts in terms of simultaneously preserving thermal radiation information and texture detail information.

Quantitative comparisons. We further give quantitative comparisons of the nine methods on two sets of image pairs from the TNO database. The first set contains ten pairs of infrared and visible images including the seven pairs in Fig. 5 and three additional pairs. The second set is an infrared and visible image sequence pair, *e.g.*, the *Nato_camp* sequence. The results of six metrics on the two datasets are shown in Figs. 6 and 7. On the first dataset, our FusionGAN is able to obtain the largest average values on the five evaluation metrics, *i.e.* EN, SD, SSIM, SF and VIF, and our FusionGAN only follows GTF by a narrow margin on the metric of CC. While on the second dataset, our FusionGAN clearly has the best EN, SD, SSIM and SF on most image pairs, and the average values of these evaluation metrics are also the largest compared to the

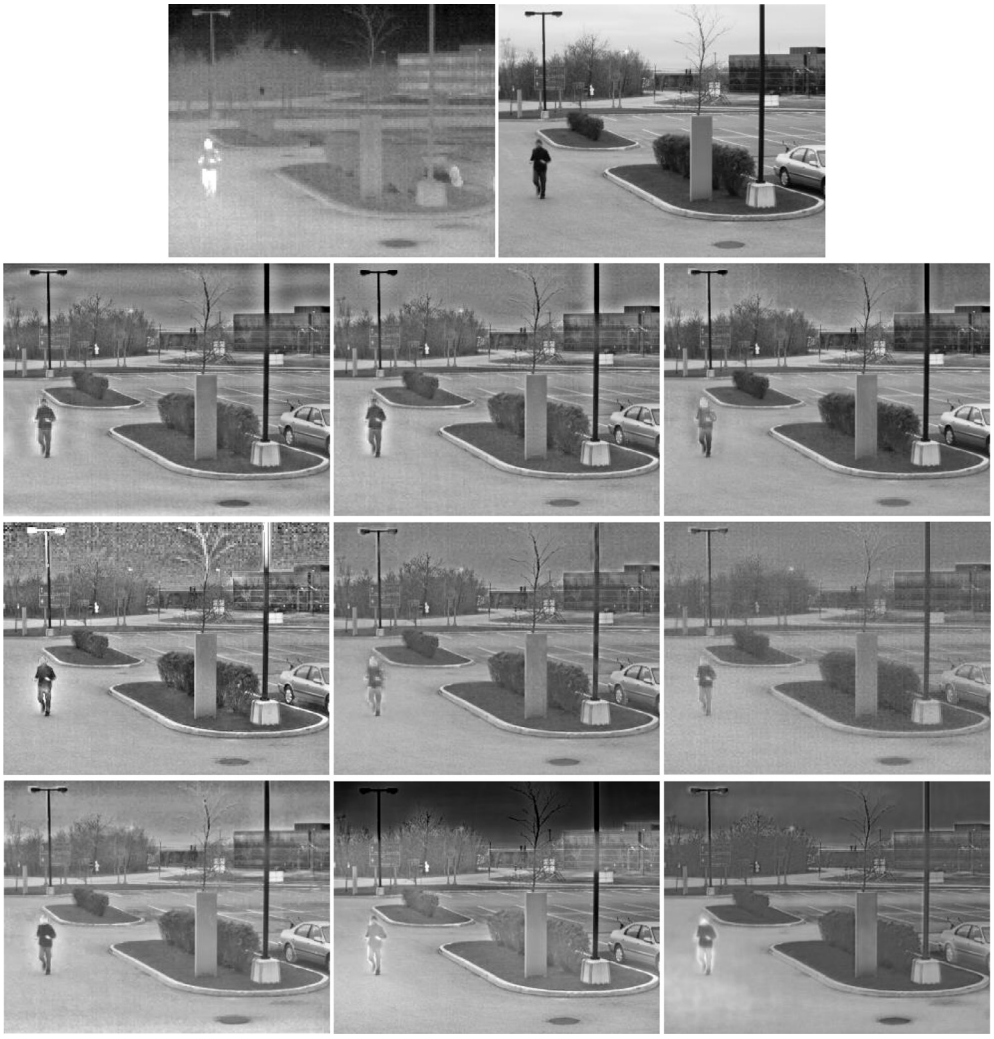

Fig. 8. Qualitative fusion results on Frame 27 of the selected *Trees and runner* from INO Dataset. The first row is the infrared image and visible image. The rest three rows (from left to right, top to bottom) are the fusion results of ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS, GTF and our FusionGAN.

rest eight methods. For the metric of CC, our FusionGAN follows behind FPDE and ASR by a narrow margin; for the metric of VIF, our FusionGAN also only follows behind GFF.

The largest EN demonstrates that our fused image has more abundant information than the other eight comparison methods. The largest SD demonstrates that our fused image has the largest image contrast. The largest SSIM demonstrates that our fused image is more similar to the infrared and visible image on the structures. While the largest SF demonstrates that our fused image contains rich edges and textures. Although CC and VIF of our method are not the best, the comparable results still demonstrate that our fused image has great correlation with two source images and is also more consistent with the human visual system. We also provide the run time comparison of nine methods in Table 1. From the results, we see that our FusionGAN can achieve comparable efficiency compared with the other eight methods.

Table 1
Run time comparison of nine methods on the two datasets from the TNO database and one dataset from the INO database. Our method is performed on GPU while all the other methods are performed on CPU. Each value denotes the mean of run times of a certain method on a dataset (unit: second).

Method	TNO1	TNO2	INO
ASR	2.62×10^{2}	9.13×10^{1}	8.97×10^{1}
CVT	1.46	6.53×10^{-1}	6.62×10^{-1}
DTCWT	3.32×10^{-1}	1.30×10^{-1}	1.28×10^{-1}
FPDE	5.02×10^{-1}	9.22×10^{-2}	1.09×10^{-1}
GFF	3.18×10^{-1}	8.35×10^{-2}	1.04×10^{-1}
LPP	9.60×10^{-2}	3.72×10^{-2}	4.28×10^{-2}
TSIFVS	3.08×10^{-2}	1.16×10^{-2}	3.60×10^{-2}
GTF	4.82	1.00	1.51
FusionGAN	2.54×10^{-1}	7.16×10^{-2}	3.50×10^{-2}

4.2.3. Results on INO database

We further test our method and other eight compared methods on the INO Database, in which we select 31 visible and infrared image pairs every other 18 frames from the video named *Trees and runner* for both qualitative and quantitative comparisons. Fig. 8 shows the fused results of the 27th frame in the selected 31 image pairs. We see that the visible image involves rich textures, whereas the infrared image contains poor textures, as can be seen from the car and tree on the ground. However, the human in the infrared image is more highlighted than that in the visible image. Considering the fusion results, the texture information can be well retained by all the nine methods. However, only our FusionGan can keep the thermal radiation distribution in the infrared image, *e.g.,* the pixel intensities of the human area. Fig. 9 shows the quantitative comparison results of four metrics, where our FusionGAN again has the

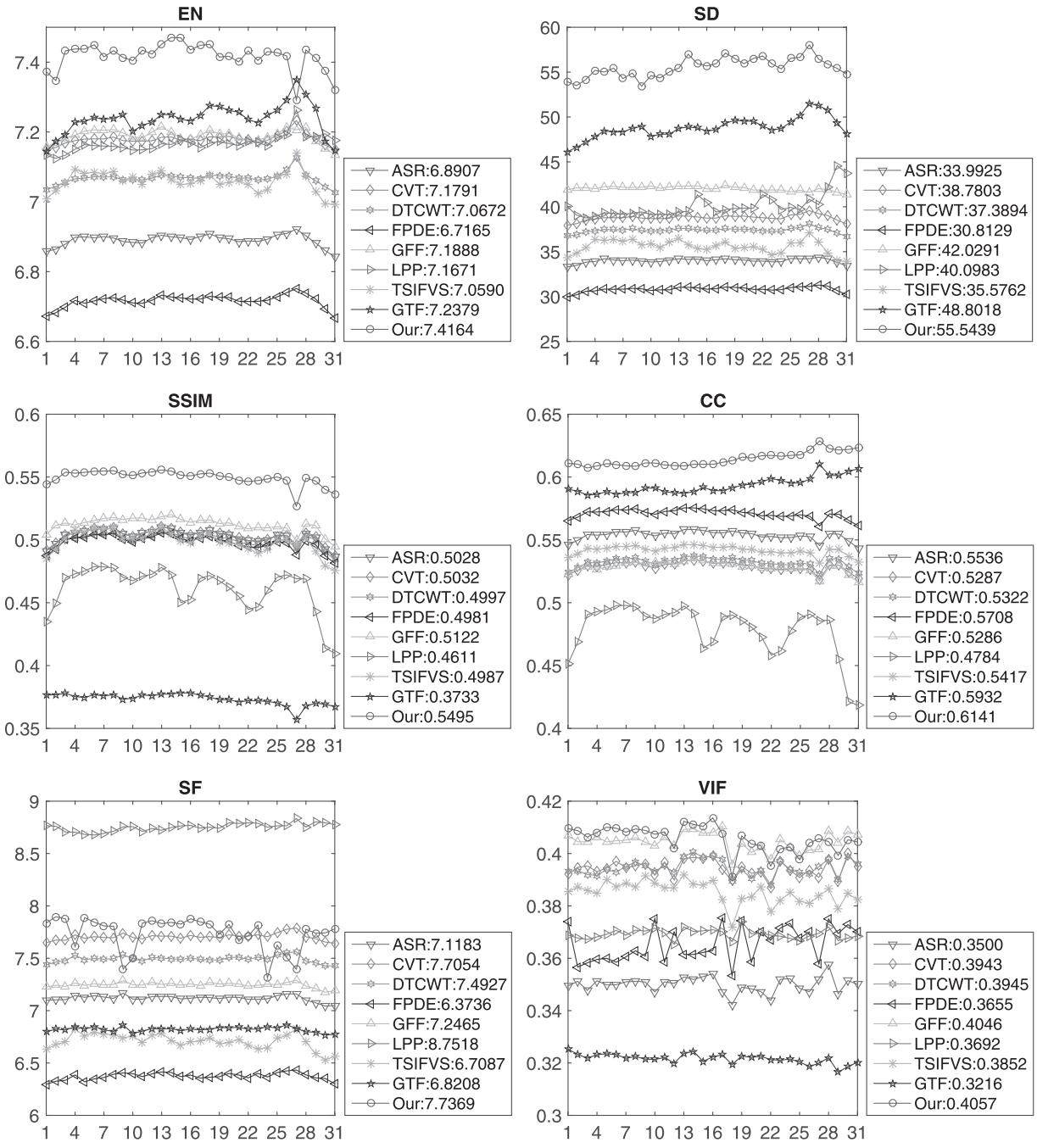

Fig. 9. Quantitative comparisons of the six metrics, *i.e.*, EN, SD, SSIM, CC, SF and VIF, on the *Trees and runner* sequence from the INO database. The eight state-of-the-art methods such as ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS and GTF are used for comparison.

best EN, SD, SSIM, CC and VIF on most image pairs, and the average values of these evaluation metrics are the largest compared to the eight compared methods. For the metric of SF, our FusionGAN only follows behind LPP. In addition, we also provide the run time comparison of different fusion methods in Table 1, and our FusionGAN can achieve comparable efficiency compared with the other eight methods.

4.3. Experimental validation of adversarial loss

To verify the importance of adversarial training in our FusionGAN, here we train two models on the TNO database based on whether to use adversarial loss, with all training settings the same as the first experiment. In Fig. 10, we schematically illustrate the fusion results of two models on four typical pairs including *Bunker, Sandpath, Kaptein_1123* and *Marne_04*. The first two rows present the infrared images and visible images, the third row is the fusion results of FusionGAN trained without adversarial loss, and the last row is the fusion results of our proposed method, *i.e.,* using the adversarial loss. From the results, we can clearly see that the fusion results of FusionGAN by using adversarial loss contain much more detail information and are also more consistent with the human visual system. For example, the textures of vegetation in *Bunker*, stumps in *Sandpath*, trees in *Kaptein_1123*, as well as car window in *Marne_04* are all fused appropriately in the last row. However, without using adversarial loss in the third row, the contrast of the

Fig. 10. Qualitative fusion results on four typical infrared and visible image pairs from the TNO database. From left to right: *Bunker, Sandpath, Kaptein_1123* and *Marne_04*. From top to bottom: infrared images, visible images, results of FusionGAN without adversarial loss and with adversarial loss. For clear comparison we select a small region (*i.e.*, the red box) in each fused image, and then zoom in it and put it in the bottom right corner. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

resultant images is low and all the above mentioned details are difficult to distinguish.

In fact, texture details cannot be totally represented by only gradient information. Other properties such as image contrast, saturation and illumination changes also play important roles in reflecting the details in visible images. However, it is difficult to mathematically characterize these properties in an objective function to be optimized. Nevertheless, the results in this section demonstrate that the adversarial training can help to exploit more texture details in a visible image and transfer them to the fused image.

4.4. Application to fusing images with different resolutions

In this experiment, we further apply our FusionGAN to fuse infrared and visible images with different spatial resolutions, say low-resolution infrared images and high-resolution visible images. To this end, we make the following three modifications. First, we downsample all infrared images to $\frac{1}{c^2}$ (*i.e.*, $\frac{1}{c} \times \frac{1}{c}$, here we set $c = 4$) of the original image scale as the new low-resolution source infrared images, while the visible images keep the original scale. The ten image pairs from the TNO database as described in Fig. 6 are used as testing data in this experiment. Second, due to the infrared and visible images are of different spatial resolutions, we cannot directly concatenate them for training or testing, as shown in Fig. 2. Therefore, we interpolate the low-resolution infrared images to the same resolution as the corresponding visible images before concatenating and putting them into the generator. Third, since the fused image and infrared image are also of different resolutions, we redesign the content loss $\mathcal{L}_{\text{content}}$ in Eq. (5) as follows:

$$\mathcal{L}_{\text{content}} = \frac{1}{HW}(\|\phi(I_f) - I_r\|_F^2 + \xi\|\nabla I_f - \nabla I_v\|_F^2), \tag{13}$$

where ϕ is a downsampling operation, which downsamples the fused image to the resolution of infrared image. All other training settings are the same as the first experiment. Note that in Eq. (13) we choose to downsample the fused image rather than upsampling the infrared image. This is because that upsampling infrared image will inevitably introduce additional noise which is likely to be transferred to the fused image, leading to an unsatisfying result.

For all the eight comparison methods, they have a prerequisite that the source images should share the same resolution. Therefore, we have to first eliminate resolution differences through downsampling visible images or upsampling infrared images. Clearly, downsampling visible images will cause texture information loss and upsampling infrared images will blur thermal radiation information. Nevertheless, to avoid loss of information, we choose to upsample the infrared images before fusion for all the comparison methods.

We select five typical image pairs for qualitative evaluation, including *Bunker, Sandpath, Kaptein_1123, Kaptein_1654* and *Marne_04*. Fig. 11 shows the fused image generated by our FusionGAN and other eight comparison methods. The first two rows in Fig. 11 present the original low-resolution infrared images and high-resolution visible images, the last row is the fusion results of our FusionGAN, and the rest eight rows correspond to the fusion results of eight comparison methods. From all these results, we can get the same conclusion as the first experiments, *e.g.,* thermal radiation in the infrared images is not well

Fig. 11. Qualitative fusion results on five typical infrared and visible image pairs from the TNO database. From left to right: *Bunker, Sandpath, Kaptein_1123, Kaptein_1654*, and *Marne_04*. From top to bottom: low-resolution infrared images, high-resolution visible images, results of ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS, GTF and our FusionGAN. Note that in the last two rows, for clear comparison we select a small region (*i.e.*, the red box) in each fused image, and then zoom in it and put it in the bottom right corner. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

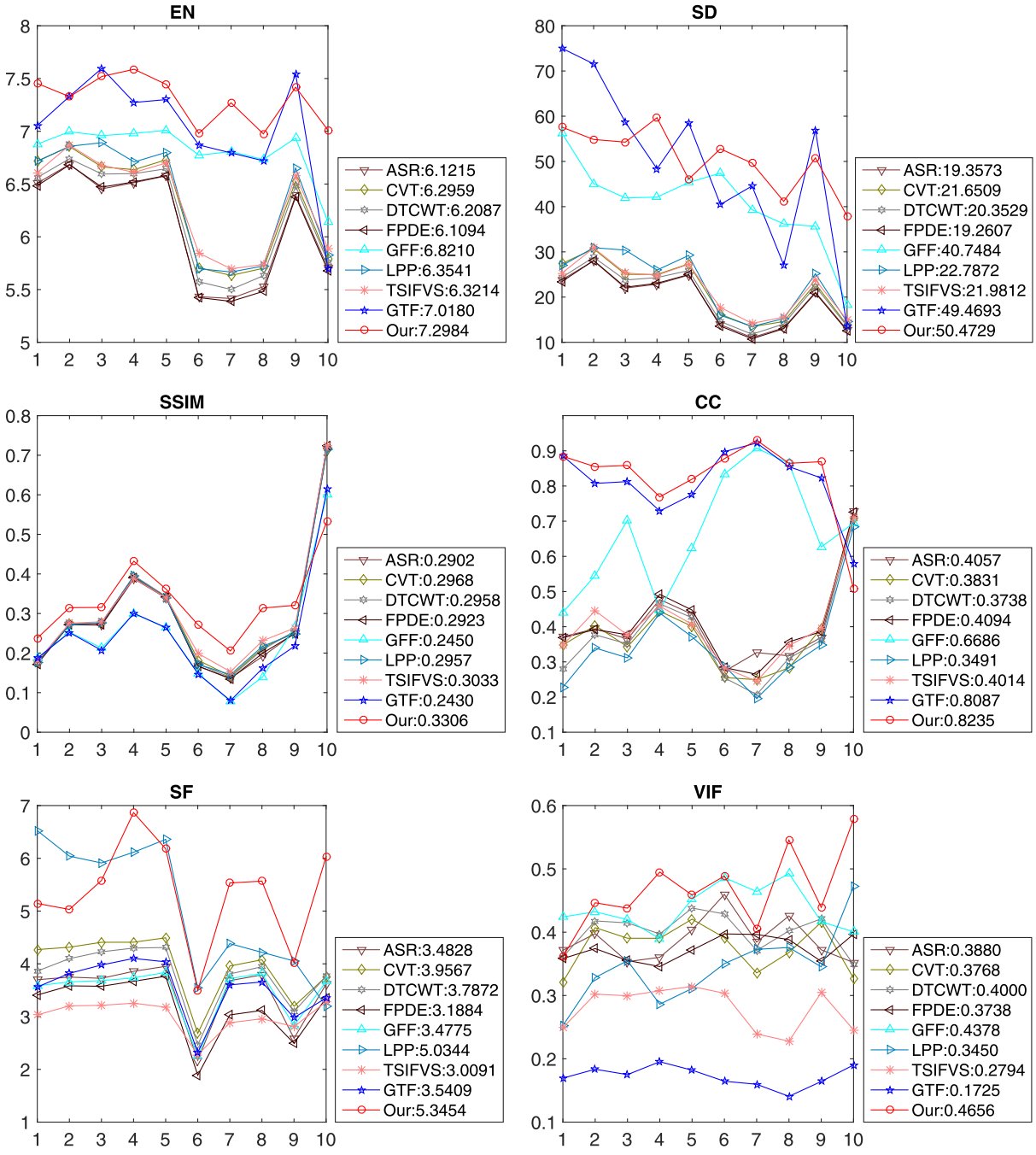

Fig. 12. Quantitative comparisons of the six metrics, *i.e.*, EN, SD, SSIM, CC, SF and VIF on ten image pairs from the TNO database. The eight state-of-the-art methods such as ASR, CVT, DTCWT, FPDE, GFF, LPP, TSIFVS and GTF are used for comparison.

preserved in the comparison methods except GTF, and compared with GTF, our results contain more abundant detail information and are more appropriate for human visual perception. In addition, the fused images generated by our FusionGAN are clearer and cleaner compared to the other eight methods. This is due to that our FusionGAN does not need to downsample or upsample source images in the content loss (*i.e.*, (13)), and hence does not suffer from noise caused by upsampling of infrared information.

We further give quantitative comparisons of the nine methods on the whole ten image pairs, and the results of six metrics are reported in Fig. 12. There are three metrics such as SSIM, CC and VIF that rely on the source images. More specifically, calculating these metrics requires the source images and the fused image to share the same resolution. It may be not appropriate to require the fused image to preserve as much information as possible in the upsampled infrared image, as it involves additional noise. Instead, it would be better to downsample the visible and fused images and require the downsampled fused image to preserve as much information as possible in the original infrared image and downsampled visible image. Therefore, in this paper we downsample the visible and fused images to the same resolution as the corresponding infrared image before calculating the metrics of SSIM, CC and VIF. From the results, we see that our FusionGAN has the best average values on all six metrics. This demonstrates the significant advantage of our FusionGAN over

the existing fusion methods for fusing source images with different resolutions.

In conclusion, our FusionGAN can keep the thermal radiation information in infrared images and the rich texture details in visible images simultaneously, no matter fusing source images in the same spatial resolution or different resolutions. Compared with existing state-of-the-art fusion methods, our results look like sharpened infrared images with clear highlighted targets and abundant detail information. In addition, our FusionGAN has comparable efficiency compared with the state-of-the-arts.

5. Discussion

The deep learning based techniques usually have a common problem that they are regarded as black-box models and even if we understand the underlying mathematical principles of such models they lack an explicit declarative knowledge representation, hence have difficulty in generating the underlying explanatory structures [48]. In this section, we briefly discuss the explainability of our FusionGAN.

The essence of traditional GAN is to train a generator to capture the data distribution, so that the data generated by the generator has the same distribution as the original data. In this procedure, the similarity of data distribution is measured by a discriminator. More specifically, a discriminator is trained to distinguish the generated data from the original data. When the discriminator cannot successfully distinguish these two types of data, we consider that the generated data has the same distribution as that of the original data. The essence of our FusionGAN is to generate a fused image which selectively preserves the information in source images (*i.e.,* the infrared and visible images), and the amount of information to be preserved is controlled by the parameters λ and ξ. In particular, the content loss aims to preserve the radiation information in the infrared image and the gradient information in the visible image, while the adversarial loss aims to preserve other important properties characterizing the detail information in visible image, such as image contrast, saturation and illumination changes. Therefore, during the adversarial process, the generator continually fits the distribution of detail information in the fused image to that in the visible image while simultaneously preserves the infrared radiation information. When the discriminator cannot distinguish the fused image from the visible image, the distribution of detail information in the fused image is then considered to be the same as that in the visible image, and hence the fused image visually possesses more textural details.

6. Conclusion

In this paper, we propose a novel infrared and visible image fusion method based on generative adversarial network. It can simultaneously keep the thermal radiation information in infrared images and the texture detail information in visible images. The proposed FusionGAN is an end-to-end model, which can avoid designing complicated activity level measurement and fusion rule manually as in traditional fusion strategies. Experiments on public datasets demonstrate that our fusion results look like sharpened infrared images with clear highlighted targets and abundant detail information, which is beneficial for target detection and recognition systems based on image fusion. The quantitative comparisons with eight state-of-the-arts on four evaluation metrics reveal that our FusionGAN can not only produce better visual effects, but also can keep the largest or approximately the largest amount of information in the source images.

Our FusionGAN is a general framework for dealing with fusion tasks aiming at fusing pixel intensities in one source image together with texture details in another source image. We have also generalized our FusionGAN to fuse source images with different resolutions. In our future work, we will further apply our FusionGAN to solve the well-known pansharpening problem from the remote sensing community, the goal of which is to fuse a low-resolution multispectral image and a high-resolution panchromatic image to generate a multispectral image with high spatial resolution [49].

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant nos. 61773295 and 61503288, and the Beijing Advanced Innovation Center for Intelligent Robots and Systems under Grant no. 2016IRS15.

References

[1] A. Dogra, B. Goyal, S. Agrawal, From multi-scale decomposition to non-multi-scale decomposition methods: a comprehensive survey of image fusion techniques and its applications, IEEE Access 5 (2017) 16040–16067.
[2] Y. Ma, J. Chen, C. Chen, F. Fan, J. Ma, Infrared and visible image fusion using total variation model, Neurocomputing 202 (2016) 12–19.
[3] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recognit. Lett. 9 (4) (1989) 245–253.
[4] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Phys. Technol. 85 (2017) 478–501.
[5] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, Inf. Fusion 12 (2) (2011) 74–84.
[6] G. Pajares, J.M. De La Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (2004) 1855–1872.
[7] Z. Zhang, R.S. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proc. IEEE 87 (8) (1999) 1315–1326.
[8] J. Wang, J. Peng, X. Feng, G. He, J. Fan, Fusion method for infrared and visible images by using non-negative sparse representation, Infrared Phys. Technol. 67 (2014) 477–489.
[9] S. Li, H. Yin, L. Fang, Group-sparse representation with dictionary learning for medical image denoising and fusion, IEEE Trans. Biomed. Eng. 59 (12) (2012) 3450–3459.
[10] T. Xiang, L. Yan, R. Gao, A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain, Infrared Phys. Technol. 69 (2015) 53–61.
[11] W. Kong, L. Zhang, Y. Lei, Novel fusion method for visible light and infrared images based on nsst–sf–pcnn, Infrared Phys. Technol. 65 (2014) 103–112.
[12] D.P. Bavirisetti, G. Xiao, G. Liu, Multi-sensor image fusion based on fourth order partial differential equations, in: International Conference on Information Fusion, 2017, pp. 1–9.
[13] W. Kong, Y. Lei, H. Zhao, Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization, Infrared Phys. Technology 67 (2014) 161–172.
[14] X. Zhang, Y. Ma, F. Fan, Y. Zhang, J. Huang, Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition, JOSA A 34 (8) (2017) 1400–1410.
[15] J. Zhao, Y. Chen, H. Feng, Z. Xu, Q. Li, Infrared image enhancement through saliency feature analysis based on multi-scale decomposition, Infrared Phys. Technol. 62 (2014) 86–93.
[16] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (2015) 147–164.
[17] J. Ma, Z. Zhou, B. Wang, H. Zong, Infrared and visible image fusion based on visual saliency map and weighted least square optimization, Infrared Phys. Technol. 82 (2017) 8–17.
[18] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.
[19] J. Zhao, G. Cui, X. Gong, Y. Zang, S. Tao, D. Wang, Fusion of visible and infrared images using global entropy and gradient constrained regularization, Infrared Phys. Technol. 81 (2017) 201–209.
[20] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: a survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[21] Y. Liu, X. Chen, Z. Wang, Z.J. Wang, R.K. Ward, X. Wang, Deep learning for pixel-level image fusion: recent advances and future prospects, Information Fusion 42 (2018) 158–173.
[22] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[23] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Inf. Fusion 4 (4) (2003) 259–280.
[24] Q. Zhang, Y. Liu, R.S. Blum, J. Han, D. Tao, Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review, Inf. Fusion 40 (2018) 57–75.
[25] S. Rajkumar, P.C. Mouli, Infrared and visible image fusion using entropy and neuro-fuzzy concepts, in: Proceedings of the Annual Convention of Computer Society of India, 2014, pp. 93–100.
[26] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.
[27] Y. Liu, X. Chen, J. Cheng, H. Peng, Z. Wang, Infrared and visible image fusion with convolutional neural networks, Int. J. Wavelets Multiresolution Inf. Process. 16 (3) (2018) 1850018.

[28] J. Zhong, B. Yang, Y. Li, F. Zhong, Z. Chen, Image fusion and super-resolution with convolutional neural network, in: Chinese Conference on Pattern Recognition, 2016, pp. 78–88.
[29] Y. Liu, X. Chen, R.K. Ward, Z.J. Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.
[30] G. Masi, D. Cozzolino, L. Verdoliva, G. Scarpa, Pansharpening by convolutional neural networks, Remote Sens. 8 (7) (2016) 594.
[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[32] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv:1511.06434v1 (2015).
[33] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan, arXiv:1701.07875v1 (2017).
[34] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S.P. Smolley, Least squares generative adversarial networks, in: IEEE International Conference on Computer Vision, 2017, pp. 2813–2821.
[35] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, arXiv:1511.07122v1 (2015).
[36] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv:1412.6980v1 (2014).
[37] B. Yang, S. Li, Visual attention guided image fusion with sparse representation, Optik-Int. J. Light Electron Opt. 125 (17) (2014) 4881–4888.
[38] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2) (2007) 143–156.
[39] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel-and region-based image fusion with complex wavelets, Inf. fusion 8 (2) (2007) 119–130.
[40] D.P. Bavirisetti, R. Dhuli, Two-scale image fusion of visible and infrared images using saliency detection, Infrared Phys. Technol. 76 (2016) 52–64.
[41] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (2019) 153–178.
[42] J.W. Roberts, J. Van Aardt, F. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens. 2 (1) (2008) 023522.
[43] Y.-J. Rao, In-fibre bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355.
[44] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Lett. 9 (3) (2002) 81–84.
[45] M. Deshmukh, U. Bhosale, Image fusion and image quality assessment of fused images, Int. J. Image Process. 4 (5) (2010) 484–508.
[46] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[47] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. fusion 14 (2) (2013) 127–135.
[48] A. Holzinger, C. Biemann, C.S. Pattichis, D.B. Kell, What do we need to build explainable ai systems for the medical domain?, arXiv:1712.09923v1 (2017).
[49] C. Chen, Y. Li, W. Liu, J. Huang, Sirf: simultaneous satellite image registration and fusion in a unified framework, IEEE Trans. Image Process. 24 (11) (2015) 4213–4224.