Information Fusion 54 (2020) 85–98

Contents lists available at ScienceDirect

Information Fusion

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

Infrared and visible image fusion via detail preserving adversarial learning

Jiayi Ma[a], Pengwei Liang[a], Wei Yu[a], Chen Chen[b], Xiaojie Guo[c], Jia Wu[d], Junjun Jiang[e,f,*]

[a] Electronic Information School, Wuhan University, Wuhan, 430072, China
[b] Department of Electrical and Computer Engineering, University of North Carolina at Charlotte, NC, 28223, USA
[c] School of Computer Software, Tianjin University, Tianjin, 300350, China
[d] Department of Computing, Faculty of Science and Engineering, Macquarie University, Sydney, NSW 2109, Australia
[e] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, 150001, China
[f] Peng Cheng Laboratory, Shenzhen, 518055, China

ARTICLE INFO

Keywords:
Image fusion
Generative adversarial network
Infrared
Detail preserving
Convolution neural network

ABSTRACT

TargefTablets can be detected easily from the background of infrared images due to their significantly discriminative thermal radiations, while visible images contain textural details with high spatial resolution which are beneficial to the enhancement of target recognition. Therefore, fused images with abundant detail information and effective target areas are desirable. In this paper, we propose an end-to-end model for infrared and visible image fusion based on detail preserving adversarial learning. It is able to overcome the limitations of the manual and complicated design of activity-level measurement and fusion rules in traditional fusion methods. Considering the specific information of infrared and visible images, we design two loss functions including the detail loss and target edge-enhancement loss to improve the quality of detail information and sharpen the edge of infrared targets under the framework of generative adversarial network. Our approach enables the fused image to simultaneously retain the thermal radiation with sharpening infrared target boundaries in the infrared image and the abundant textural details in the visible image. Experiments conducted on publicly available datasets demonstrate the superiority of our strategy over the state-of-the-art methods in both objective metrics and visual impressions. In particular, our results look like enhanced infrared images with clearly highlighted and edge-sharpened targets as well as abundant detail information.

1. Introduction

Infrared images, which are captured by infrared sensors to record the thermal radiations emitted by different objects, are widely used in target detection and surface parametric inversion. Infrared images are minimally affected by illumination variations and disguises, and they can be easily captured at daytime and nighttime. However, infrared images usually lack texture, which seldom influences the heat emitted by objects. By contrast, visible images are captured and used to record the spectral information reflected by different objects, which contain discriminative characteristic information. Visible images also provide perceptual scene descriptions for the human eyes. Nevertheless, the targets in visible images may not be easily observed due to the influence of external environment, such as nighttime conditions, disguises, objects hidden in smoke, cluttered background, etc. Therefore, the purpose of fusion is to obtain a single complementary fused image that has rich detail information from the visible image and effective target areas from the infrared image [1–8].

Various infrared and visible image fusion methods have been proposed in recent years, and they can be divided into seven categories, namely, multi-scale transform [9], sparse representation [10], neural network [11], subspace [12], saliency [13], hybrid models [14], and deep learning [15]. In general, the current fusion methods involve three crucial challenges, i.e., image transform, activity-level measurement and fusion rule designing [16]. The three constraints have become increasingly complex, especially for designing fusion rules in a manual way which strongly limits the development of the fusion methods. Moreover, the existing methods typically select the same salient features of source images, such as edges and lines, to be integrated into the fused images, so that the fused images contain more detail information. However, the above approaches may not be suitable for infrared and visible image fusion. *In particular, infrared thermal radiation information is characterized by pixel intensities, while textural detail information in visible images is typically characterized by edges and gradients. These two scenarios differ and cannot be represented in the same manner.*

* Corresponding author.
E-mail addresses: jyma2010@gmail.com (J. Ma), erfect@whu.edu.cn (P. Liang), yuwei998@whu.edu.cn (W. Yu), chenchen870713@gmail.com (C. Chen), xguo@tju.edu.cn (X. Guo), jia.wu@mq.edu.au (J. Wu), junjun0595@163.com, jiangjunjun@hit.edu.cn (J. Jiang).

https://doi.org/10.1016/j.inffus.2019.07.005
Received 10 January 2019; Received in revised form 18 July 2019; Accepted 22 July 2019
Available online 22 July 2019

Fig. 1. Schematic illustration of image fusion. From left to right: infrared image, visible image, fusion results of a traditional method ADF [17], FusionGAN [15], and our method.

To address the abovementioned issues, in our previous work [15], we have proposed an end-to-end model, namely FusionGAN, to fuse the infrared and visible images based on a generative adversarial network (GAN). It is able to avoid the manual design of complicated activity level measurements and fusion rules, and the fusion results look like infrared images with clear highlighted targets and abundant details. However, there are two problems still remain. On the one hand, FusionGAN only relies on adversarial training to increase extra detail information, which is uncertain and unsteady, resulting in loss of a mount of detail information. On the other hand, the content loss in FusionGAN only attaches importance to the edge information in the visible image, ignoring that in the infrared image, and hence the edges of target in fusion results tend to be fuzzy. To overcome these two challenges, in this paper, we design two loss functions, i.e., the detail loss and target edge-enhancement loss, to improve the quality of detail information and sharpen the edge of infrared targets.

To illustrate the main idea of our method, we present a representative example in Fig. 1. The original infrared and visible images are shown in the left two images. The target, i.e. the bunker, is salient and easy to detect in the infrared image but difficult to distinguish in the visible image. However, the visible image has much more background information compared with the infrared one, which is beneficial to the accurate recognition of targets. Therefore, fusing these two types of information is desirable in achieving a comprehensive understanding of the captured scene. For qualitative comparison, in addition to FusionGAN, we also consider the anisotropic diffusion fusion (ADF) [17], a recently introduced infrared and visible image fusion method based on the traditional strategy. The fusion results of ADF, FusionGAN and our method are shown in the right three images. Clearly, ADF only preserves the textural information in the source images but the thermal radiation information is lost, which then leads to a low-contrasted target in the fused image. FusionGAN and our method both can retain thermal radiation distribution and textural details. Nevertheless, the target boundaries in our result are more clear than the result of FusionGAN, and the trees are of more details in our result.

This paper is an extension of our previous FusionGAN [15], and the primary new contributions can be summarized as follows. First, the fusion results of FusionGAN tend to be smooth and fuzzy, which is a common problem by optimizing the ℓ_2 norm. To address this issue, we propose the detail loss to constrain the fusion results and visible images to be more similar in semantic level, which can not only make the fusion results clearer, but also retain more useful detail information. Second, FusionGAN is designed to preserve the radiation information of infrared images, ignoring the detail information that can be reflected by textures (e.g., edges of salient objects) in infrared images. To solve this problem, we design the target edge-enhancement loss to further optimize the textures of targets, leading to a sharper representation of targets in the fusion results. The detail loss and target edge-enhancement loss keep the useful information in source images to a large extent compared with FusionGAN. Third, we deepen the generator and discriminator in the GAN framework. The deeper network has more powerful feature representation ability with stronger capacity to optimize our loss functions and get performance of the fusion results improved. Fourth, we provide qualitative and quantitative comparisons between our approach and nine state-of-the-art methods on two publicly available datasets. Unlike the competitors, our method can generate fused images with clearly highlighted and edge-sharpened targets as well as more textures.

The rest of the paper is organized as follows. Section 2 introduces the background and related work. Section 3 describes our method in detail. Section 4 validates the superiority of our model, especially the detail loss and target edge-enhancement loss, over other state-of-the-art methods on publicly available datasets. Section 5 presents the concluding remarks.

2. Related work

In this section, we briefly review the related work on infrared and visible image fusion and GANs. In addition, the perceptual loss for optimization is also discussed.

2.1. Infrared and visible image fusion

Numerous infrared and visible image fusion methods have been proposed due to the fast-growing demand and progress of image representation in recent years. According to the theoretical basis, the fusion methods can be divided into seven categories, as shown in Table 1. Multi-scale transform-based methods [18–20], the most actively used for fusion, assume that a source image can be decomposed into several levels. A final target fused image can be obtained by fusing its layers based on certain particular fusion rules. The most popular transforms used for decomposition and reconstruction are the wavelet [21], pyramid [22], curvelet [23], and their variants. The second category is the sparse representation-based methods [24,25]. It has been found that an image can be represented with a linear combination of sparse basis in an over-complete dictionary, which is the key factor in ensuring the good performance of this kind of method. The third category is the neural network-based methods [11,26], which have the advantages of strong adaptability, fault-tolerant capability, and anti-noise capacity, and they can imitate the perceptual behavior system of the human brain when dealing with neural information. The fourth category is the subspace-based methods [12,27], which aim to project high-dimensional input images into low-dimensional subspaces. Given that redundant information often exists in an image, low-dimensional subspaces can help capture the intrinsic structures of the original images. The fifth category is the saliency-based methods [28,29]. Human visual attention often captures objects or pixels that are more significant than their neighbors. For the methods in this category, the intensities of the regions with salient objects are highlighted, which then improve the visual quality of the fused image. The sixth category refers to the hybrid methods

Table 1
The category of infrared and visible image fusion methods.

Multi-scale transform	[18–20]
Sparse representation	[24,25]
Neural network	[11,26]
Subspace	[12,27]
Saliency	[28,29]
Hybrid	[14,30,31]
Deep learning	[4,33,34]

[14,30,31] that combine the advantages of the different methods and thus further improve the image fusion performance. In particular, Liu et al. [31] introduced an interesting fusion procedure guided by an integrated saliency map under the framework of joint sparse representation model. Recently, since much attention has been drawn to deep learning, some deep learning-based fusion methods have been developed [4,5,32–34]. However, current methods such as [4,5,33,34] typically only apply deep learning framework to some parts of the fusion process, e.g., extracting features or learning fusion strategies, while the overall fusion process is still in traditional frameworks and not end-to-end. In the field of exposure fusion, Prabhakar et al. [32] proposed an end-to-end model and has achieved promising fusion performance. However, the addressed problem is significant different with infrared and visible image fusion.

In general, the abovementioned methods aim at ensuring that the fused images contain abundant detail information. Therefore, they typically use the same image representations and select the same salient features of source images such as the textures to be fused. This may be problematic in infrared and visible image fusion, as the thermal radiation in infrared images is characterized by pixel intensity, and the high-contrast property will be lost in the result if only textures are considered during fusion. To address this issue, in our previous work [35], we proposed the gradient transfer fusion (GTF) method for image fusion to preserve the main intensity distribution in an infrared image and the gradient variation in a visible image. The result in [35] is highly similar to an infrared image with detailed appearance. The detail information in a visible image includes gradients, contrast, and saturability, *etc*. Thus, gradient variation cannot sufficiently preserve the useful detail information contained in visible images. To address this issue, we further proposed a GAN framework [15] to alleviate the problem with the loss function designed based on GTF. Nevertheless, only using adversarial training may still cause information loss due to its uncertainty and unstability, and the GTF loss ignores the edge information in infrared images, which will blur the targets in fusion results. In this paper, we introduce a new end-to-end model with two specifically designed loss functions based on detail preserving adversarial learning to overcome the abovementioned challenges.

2.2. Generative adversarial network

GAN was first proposed by Goodfellow et al. [36] to solve the problem of generating more realistic images. The main idea of GAN is to build a minmax two-player game between the learning of a generator and a discriminator. The generator takes noise as the input and attempts to transform this input noise into a more realistic image sample. Meanwhile, the discriminator takes generated samples or realistic samples as the input, the aim of which is to determine whether the input sample is derived from the generated sample or the realistic sample. The adversarial characteristic between the generator and the discriminator is continued until the generated samples cannot be distinguished by the discriminator. Subsequently, a relatively more realistic image sample can be produced by the generator. Although the original GAN can be used to generate digital images, such as those obtained from MNIST, noise and incomprehensible information still exist in the generated results. To improve the quality of the generated images, LAPGAN [37] is utilized with the Laplacian pyramid to generate a high-resolution image supervised by the low-resolution image; however, this approach is not suitable for images that contain wobbly objects. [38] and [39] succeeded in generating nature-type images, but they did not leverage the generators for supervised learning. [40] proposed the application of deeper CNNs to GANs and drafted a rule to design the CNN architecture of a generator and a discriminator for steady training. InfoGAN [41] can learn more interpretable representations. To solve the problem of weak GAN stability during the training process, the objective function of GANs was modified and WGAN [42] was proposed to relax the GAN training requirement, but the model was slow to converge as opposed to regular GANs. [43] resolved the problem by using the least square loss function for the discriminator.

The most widely used variant of GAN is the conditional GAN [44], which applies GANs in the conditional setting and forces the output to be conditioned on the input. Many studies based on conditional GANs, including image inpainting [45], image style transferring [46], image-to-image translation [47], product photo generation [48], *etc*., have been reported. The method presented in this paper is also mainly based on conditional GANs.

2.3. Perceptual loss for optimization

The pixel-wise loss function, such as mean square error (MSE), is widely used in image generation. However, this loss function typically renders the generated results over-smoothed, which then results in poor perceptual image quality. An increasing number of researchers in recent years have used perceptual loss to solve problems related to image style transferring and image super-resolutions. Perception loss is generally used to compare the high-level feature extracted from convolutional networks rather than by examining the pixel itself. [39] compared the feature extracted from neural networks, and the results showed that this loss can solve the ill-posed inverse problem caused by nonlinear representations. [49] and [50] replaced the low-level pixel-wise error measures with Euclidean distances between features extracted from a pre-trained VGG network. [51] adopted the perceptual loss and subsequently generated superior images. Inspired by the advantages of using perceptual loss, we introduce a detail item in our loss function to promote fusion performance. However, unlike the usual perceptual loss that is computed by a pre-trained VGG network, we use a discriminator as the feature extractor to compute the detail loss in our study.

3. Method

This section describes our proposed method. We first discuss the motivation of our method, and then present the network architectures and introduce the designed loss functions. Finally, we list some details in network training.

3.1. Motivation

Given a pair of infrared and visible image, our goal is to fuse both image types and construct a fused image that preserves both the saliency of targets in the infrared image and the abundant detail information in the visible image. Using CNN to generate fused image can overcome the difficulty of designing the activity level measurement and fusion rule in a manual way. However, two challenges exist with this approach. On the one hand, in the field of deep learning, training an excellent network requires a large number of labeled data. In other words, ground truth is essential for supervision during the CNN training procedure. However, a truly fused image does not exist in the image fusion problem. To address this issue, we convert the fusion problem into a regression problem, in which a loss function is required to guide the regression process. Given our fusion purpose, the objective function of GTF, which aims to preserve both thermal radiation information and visible textural details, is a good choice. On the other hand, the detail information in GTF is only represented as a gradient variation, which suggests that other important detail information, such as contrast and saturability, are abandoned. Nevertheless, such detail information usually cannot be characterized as a mathematical model.

Inspired by recent works about style transferring, GAN which builds a minimax game between a generator and a discriminator could be a better solver. We initially generate a fused image which looks like the result of GTF by solving the objective function in GTF using the generator. The result with the visible image is then sent to the discriminator to judge whether the image comes from source data or not. By building

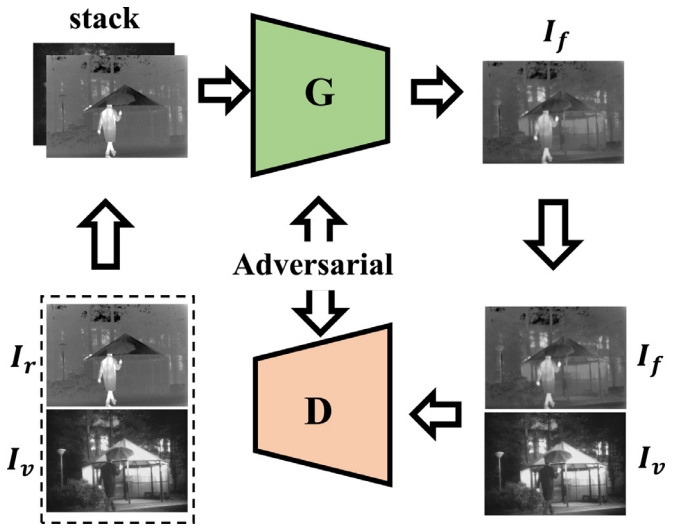

Fig. 2. The framework of our proposed method for infrared and visible image fusion.

the adversarial between generator and discriminator, when the discriminator cannot distinguish the fused image from the visible image, then we assume that our fused image contains sufficient detail information. By using this method, the detail information is represented and chosen automatically by a neural network rather than the manually designed rules. Moreover, our loss function contains an additional detail loss and a target edge-enhancement loss apart from adversarial loss. These items enable our model to be steady during the adversarial procedure, with greatly promising fusion performance.

The framework of our method is schematically illustrated in Fig. 2. During the training phase, we first stack infrared image I_r and visible image I_v in the channel dimension and then place the stacked image into the generator G similar to that in ResNet [52]. Guided by the loss function, we can then obtain the original fused image I_f from G. Subsequently, we input I_f with I_v into discriminator D whose architecture approximates that of the VGG-Net [53] to judge which of the samples is from the source data. The above training process is repeated until D cannot distinguish the fused image from the visible image. Finally, we obtain the G that has a strong ability to generate the fused image with a highlighted sharpening-edged targets and more abundant textures.

3.2. Network architecture

The proposed model is composed of a generator and a discriminator based on different network architectures, as shown in Fig. 3. Compared to our previous FusionGAN [15], we deepen the generator and discriminator which possess more powerful feature representation ability to improve the fusion performance. In particular, the generator is designed based on the ResNet [52]. In our generator network, the activation function of the residual block is a parametric rectified linear unit (RELU) [54] rather than the typical RELU. The parametric RELU is the same as leaky RELU [55] except that the slope is an adaptively learned parameter via back propagation. Furthermore, we use 1×1 convolution layer to replace the fully connected layer and build the fully convoluted network, which is not restricted by the size of an input image. In the fusion task, the aim is to extract valuable information from the source infrared and visible images. This approach therefore differs from the general GAN because our model does not contain deconvolution or pooling layers. Pooling layers will drop out some detail information, while deconvolution layers will insert extra information into the input, and both scenarios suggest inaccurate depiction of the real information of source images.

The design of the discriminator is based on the VGG11 network [53]. Five convolution layers and five maxpooling layers are used in VGG11. By contrast, each convolution layer in our network is followed by a batch normalization layer, which has been proven to effectively accelerate network training. For the activation function, we replace the general RELU with parametric RELU to adjust the degree of leakiness during back propagation. Then, we add another convolution layer (1×1 filters) to reduce the dimension, which implies that the fully connected layers in VGG can be neglected. The discriminator is used to classify whether the image is a visible one or not, and thus, the large-scale fully connected network can be replaced with a simple convolutional layer. Therefore, both generator and discriminator networks can be regarded

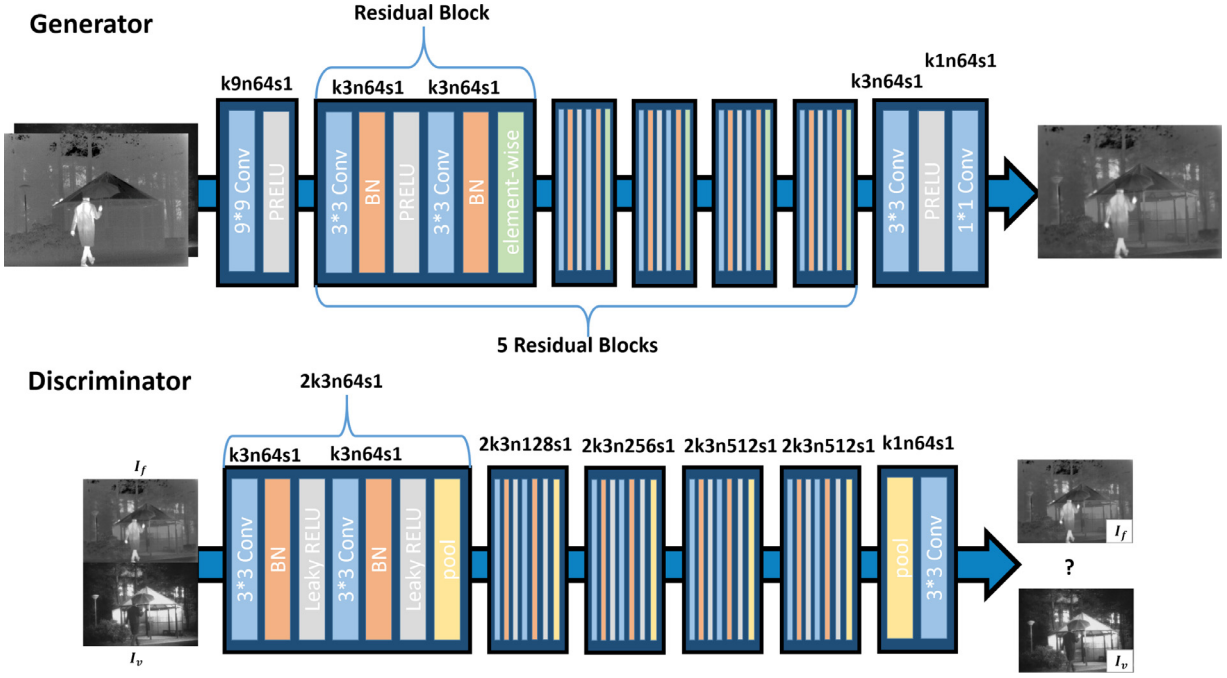

Fig. 3. Network architectures of generator and discriminator. The generator basic unit is a residual block proposed in ResNet. The network architecture of discriminator is similar to VGG11 network.

as fully convolutional networks that are robust for different sizes of input images.

3.3. Loss function

Generator loss consists of content loss, detail loss, target edge-enhancement loss, and adversarial loss, and they are expressed as follows:

$$Loss_{\text{total}} = \underbrace{L_{\text{image}} + \alpha L_{\text{gradient}}}_{\text{content loss}} + \beta L_{\text{detail}} + \delta L_{\text{tee}} + \gamma L_{\text{adversarial}}, \tag{1}$$

where the content loss constrains the fused image to one with similar pixel intensities as those of the infrared image and similar gradient variation as that of the visible image, which can be analogous to the objective function of GTF. The detail loss L_{detail} and adversarial loss $L_{\text{adversarial}}$ aim at adding more abundant detail information to the fused image. The target edge-enhancement loss L_{tee} is for sharpening the edges of highlighted targets in the fused image. We formulate the content loss as the sum of image loss L_{image} and gradient loss L_{gradient}. Then, we use the weight parameters of α, β, δ, γ to control the tradeoffs among different items in the generator loss.

3.3.1. Content loss

The pixel-wise image loss L_{image} is defined based on MSE as follows:

$$L_{\text{image}} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(I_{r_{(x,y)}} - I_{f_{(x,y)}}\right)^2, \tag{2}$$

where I_r is the original infrared image, I_f is the final output of the generator, and W and H denote the width and height of the image. The image loss renders the fused image consistent with the infrared image in terms of pixel intensity distribution. Note that we choose ℓ_2 norm due to that it is quadratic. Compared with ℓ_1 norm, ℓ_2 norm is derivable and easy to be optimized.

To fuse rich textural information, we design the gradient loss inspired by GTF as follows:

$$L_{\text{gradient}} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(D_{v_{(x,y)}} - D_{f_{(x,y)}}\right)^2, \tag{3}$$

where $D_{v_{(x,y)}}$ denotes the gradient of the visible image and $D_{f_{(x,y)}}$ denotes the gradient of the fused image. Gradient loss is defined as the MSE between $D_{v_{(x,y)}}$ and $D_{f_{(x,y)}}$.

3.3.2. Detail loss

We define the difference of the discriminator feature map between the fused image and the visible image as the detail loss as follows:

$$L_{\text{detail}} = \sum_{i=1}^{N}\sum_{j=1}^{M}\left(\phi_{v_{(i,j)}} - \phi_{f_{(i,j)}}\right)^2, \tag{4}$$

where ϕ depicts the feature map obtained by the convolution within the discriminator, ϕ_v and ϕ_f denote the feature representations of the visible and fused images, N and M denote the width and height of the result, which is the input image computed by conventional feature maps.

As for the other computer vision tasks, the perceptual loss produced by a pre-trained VGG-Net is usually used to improve performance. This approach is a good choice when using the VGG-Net to extract high-level features. However, the VGG-Net, which is pre-trained with the ImageNet dataset, does not contain infrared images. Moreover, the extraction of high-level features from the fused images (thermal radiation information and visible texture information) is uncertain in VGG-Net. Therefore, it will be problematic to mix up visible and fused images as VGG-Net input. Actually, the discriminator of our network is trained by fused and visible images. During the training process, the discriminator is able to extract relatively better features of fused and visible images, and this is the reason why the discriminator instead of VGG-Net is used to extract high-level features (we will validate this point in our experiments). Furthermore, the gradient loss will be decreased when the detail loss is optimized.

3.3.3. Target edge-enhancement loss

We formulate target edge-enhancement loss L_{tee} as follows:

$$L_{\text{tee}} = \frac{1}{WH}\sum_{x=1}^{W}\sum_{y=1}^{H}\left(I_{r_{(x,y)}} - I_{f_{(x,y)}}\right)^2 \cdot G(x,y). \tag{5}$$

In fact, this item is similar to L_{image}. In order to make target boundaries more sharpened, a weight map G is designed to pay more attention to the target boundary area and multiplied to L_{image}, where G is defined as follows:

$$G(x,y) = N_{k=3}\left(D_{r(x,y)}\right) + N_{k=5}\left(D_{r(x,y)}\right) + N_{k=7}\left(D_{r(x,y)}\right), \tag{6}$$

where N represents the Gaussian kernel, k corresponds to the kernel radius, and $D_{r(x,y)}$ denotes the gradient of the infrared image. Here we empirically use the combination $k = 3, 5, 7$ as our default configuration due to the satisfying visual effect. Obviously, there are three characteristics in our G map. First, the weights of most regions are 0, because these regions can be optimized well by L_{image}, and it does not need to optimize them again in L_{tee}. Second, the weights in the area of infrared target boundaries are large, which enables our model to focus on infrared target boundaries that may be ignored in visible image during training. Third, the parts close to the edge area can obtain small weights, which will achieve a smooth transition on both sides of an edge area.

3.3.4. Adversarial loss

The adversarial loss is adopted for our generator network with a discriminator to generate better fused images. The adversarial loss is defined based on the probabilities of the discriminator $\log D_{\theta_D}(G_{\theta_G}(I^{\text{mix}}))$ over all the training samples as follows:

$$L_{\text{adversarial}} = \sum_{n=1}^{N}\left(1 - \log D_{\theta_D}(G_{\theta_G}(I^{\text{mix}}))\right), \tag{7}$$

where I^{mix} is the stack of infrared and visible images, $\log D_{\theta_D}(G_{\theta_G}(I^{\text{mix}}))$ is the probability of the fusion image like a visible image, and N is the size of the batch.

3.4. Training detail

We train our proposed model on the TNO dataset[1] that contains 45 different scenes, and we select 45 infrared and visible image pairs for training. The image pairs has been aligned in advance, and image registration is required for unregistered image pairs [56,57]. We also adopt a random crop of 88×88 on the original infrared and visible image pairs in each iteration as the input during training. The input (i.e. the pixel intensity) is normalized to the range between -1 and 1. During the training process, we optimize the loss function using the Adam solver. For each iteration, the generator and the discriminator update their parameters. In the testing process, we place the whole stacked image into the generator and then obtain a fused image with the same size as input.

4. Experiments and evaluations

To evaluate the fusion performance of our proposed method, we conduct experiments on two publicly available datasets, such as TNO and INO[2], and compare them ten other fusion methods, namely, ADF [17], dual-tree complex wavelet transform (DTCWT) [58], fourth-order

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

partial differential equation (FPDE) [12], image fusion using multi-resolution singular value decomposition (IMSVD) [59], infrared and visible image fusion using deep learning framework (IVIFDLF) [33], two-scale image fusion based on visual saliency (TSIFVS) [13], wavelet [60], GTF [35], DenseFuse [34], and FusionGAN [15]. All these competitors are implemented based on publicly available codes, and we set their parameters by referring to their original reports. The experiments are conducted on a laptop with 3.3 GHz Intel Xeon CPU I5-4590, GPU GeForce GTX 1080TI, and 11 GB memory.

4.1. Training settings and fusion metrics

Our training parameters are set as follows: batch image size is 64, number of training iteration is 400, and discriminator training step is 2. Parameters α, β, δ and γ are set as follows: $\alpha = 100$, $\beta = 0.2$, $\delta = 5$ and $\gamma = 0.005$. The learning rate is set to 10^{-5}. All models are trained with the TNO dataset.

It is often difficult to judge the fusion performance by only subjective evaluation. Thus, quantitative fusion metrics are considered for objective evaluation. In this paper, we select six metrics, namely, entropy (EN) [61], standard deviation (SD) [62], correlation coefficient (CC) [63], spatial frequency (SF) [64], structural similarity index measure (SSIM) [65] and visual information fidelity (VIF) [66]. Their definitions are as follows. EN is based on information theory, which defines and measures the amount of information an image contains. SD is based on a statistical concept that reflects the distribution and contrast of an image. CC measures the degree of linear correlation of the fused image and the source images. SF metric is built based on horizontal and vertical gradients, which can measure the gradient distribution effectively and reflect the detail and texture of an image. SSIM measures the structural similarity between source images and fused image. VIF measures the information fidelity of fused image. For these six metrics, larger values indicate better performance.

4.2. Validation of detail loss

Detail loss plays an important role in our proposed method. By applying detail loss, our model gets steadier and the fusion performance gets better. Therefore, in this section, we focus on validating the detail loss without target edge-enhancement loss added in $Loss_{\text{total}}$. We design several experiments to demonstrate how to extract features from image for computing detail loss and confirm it is actually the detail loss that can improve the detail information in fused image.

Perceptual loss has been widely used in image style transferring. The existing methods typically consider a pre-trained VGG-Net as a feature extractor, and compare the feature map of the pool5 layer extracted from a generated image and the target image. Perceptual loss makes the generated image similar to the target image not only on pixel level but also on semantic level. In our proposed method, the function of detail loss is nearly the same as perceptual loss. But the pre-trained VGG-Net and the feature map of pool5 layer may not suitable for the task of infrared and visible image fusion because pre-trained VGG-Net is only trained on visible images, which almost cannot extract the high-level features of infrared information. In contrast, our discriminator is trained on the fused image and visible image, and hence infrared information may be extracted by the discriminator. To this end, it is more suitable to use the discriminator as feature extractor for detail loss computing.

To validate the abovementioned idea, in the following we conduct Experiments 1 to train two different models. The first one we call VGG-model where pre-trained VGG-Net is used as feature extractor. The second one we call D-model where discriminator is used as feature extractor. We compare the feature map of pool5 layer between the fused image and visible image. As the useful information for infrared and visible image fusion may not be contained in the feature map of pool5, we also conduct Experiment 2 to compare the feature maps of different layers, such as pool5, pool4, pool3, pool2 in two models. Finally, we conduct Experiments 3 to verify the role of detail loss in promoting the fusion performance.

4.2.1. Experiment 1

Fig. 4 illustrates some typical fusion results, where pre-trained VGG-Net and discriminator are respectively used as feature extractor. The first two rows present the original infrared and visible images of four scenes from the TNO dataset such as *smoke, men, bench* and *tree*. The remaining two rows correspond to the fusion results of VGG-model and D-model. From the results, we see that the fusion results of VGG-model

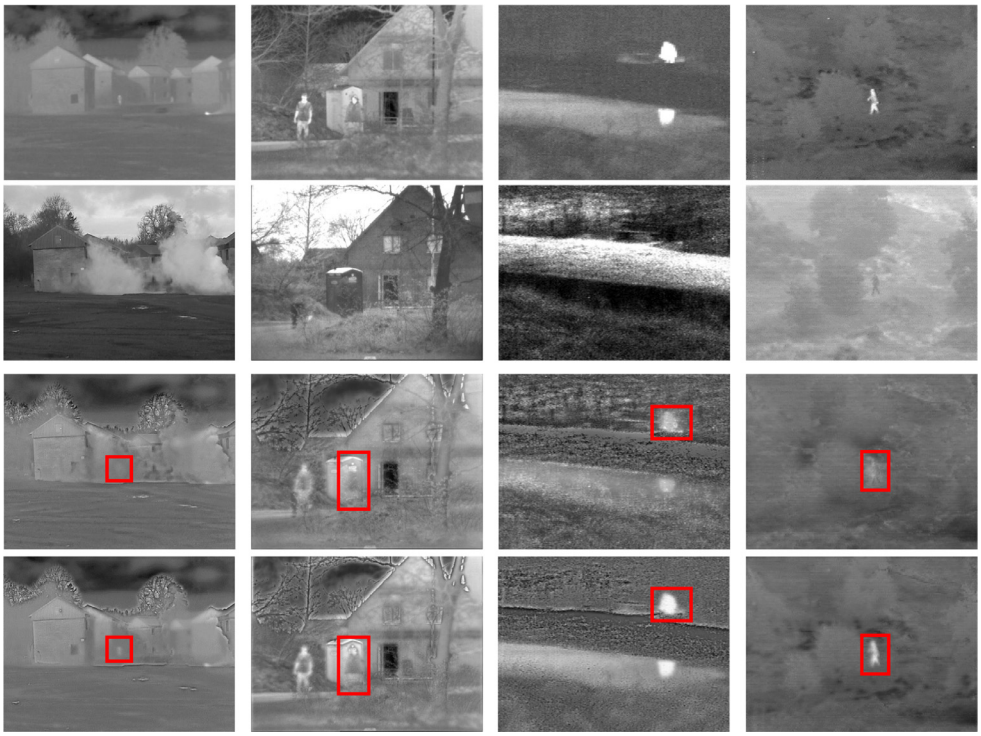

Fig. 4. Comparison of using pre-trained VGG-Net and discriminator as feature extractor on some typical infrared and visible image pairs. From left to right: *smoke, men, bench* and *tree*. From top to bottom: infrared image, visible image, fusion results of using VGG-Net as feature extractor, and fusion results of using Discriminator as feature extractor.

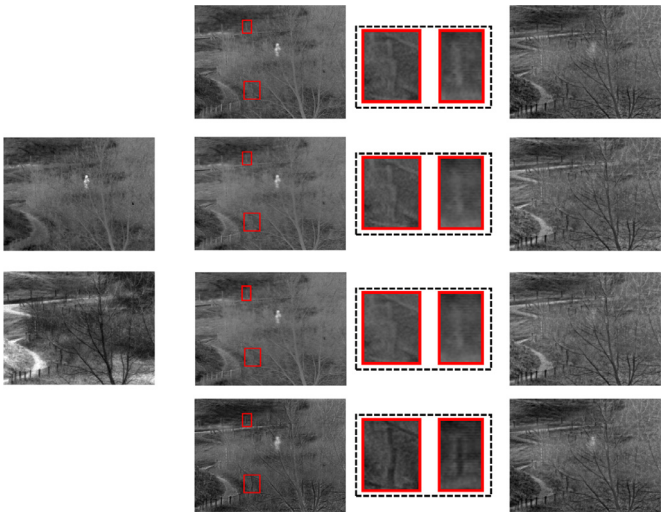

Fig. 5. Fusion results of D-model and VGG-model by using feature maps of different layers. Left: infrared and visible images. Middle: fusion results with highlighted regions (i.e., red boxes) by using feature maps of pool2 layer (top), pool3 layer (top middle), pool4 layer (middle bottom) and pool5 layer (bottom) in discriminator for computing detail loss. Right: fusion results by using feature maps of pool2 layer (top), pool3 layer (top middle), pool4 layer (middle bottom) and pool5 layer (bottom) in VGG-Net for computing detail loss. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Fig. 7. From left to right: *Kaptein_1654, sand path* and *bush*. From top to bottom: infrared image, visible image, fusion results of model without detail loss, and fusion results of model with detail loss.

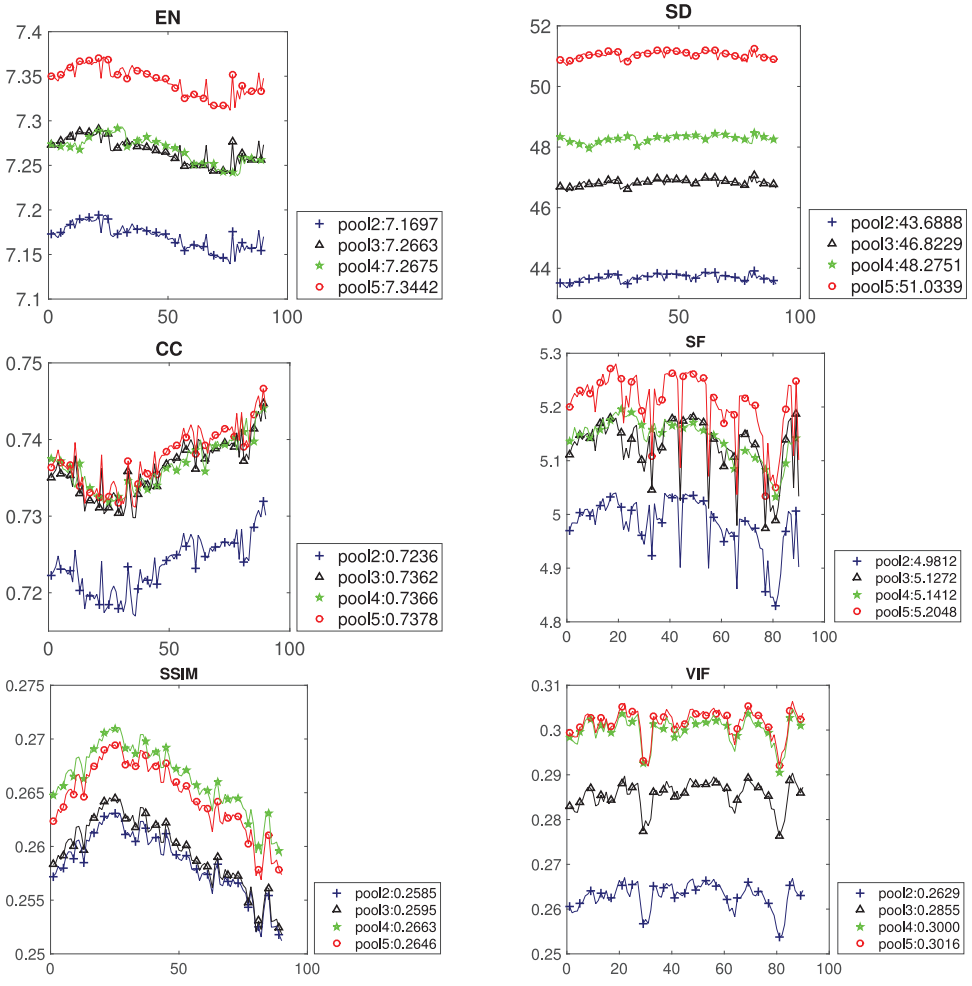

Fig. 6. The results of six fusion metrics on an infrared and visible image sequence pair from the INO dataset using feature map of pool2, pool3, pool4, and pool5 layers in discriminator for computing detail loss.

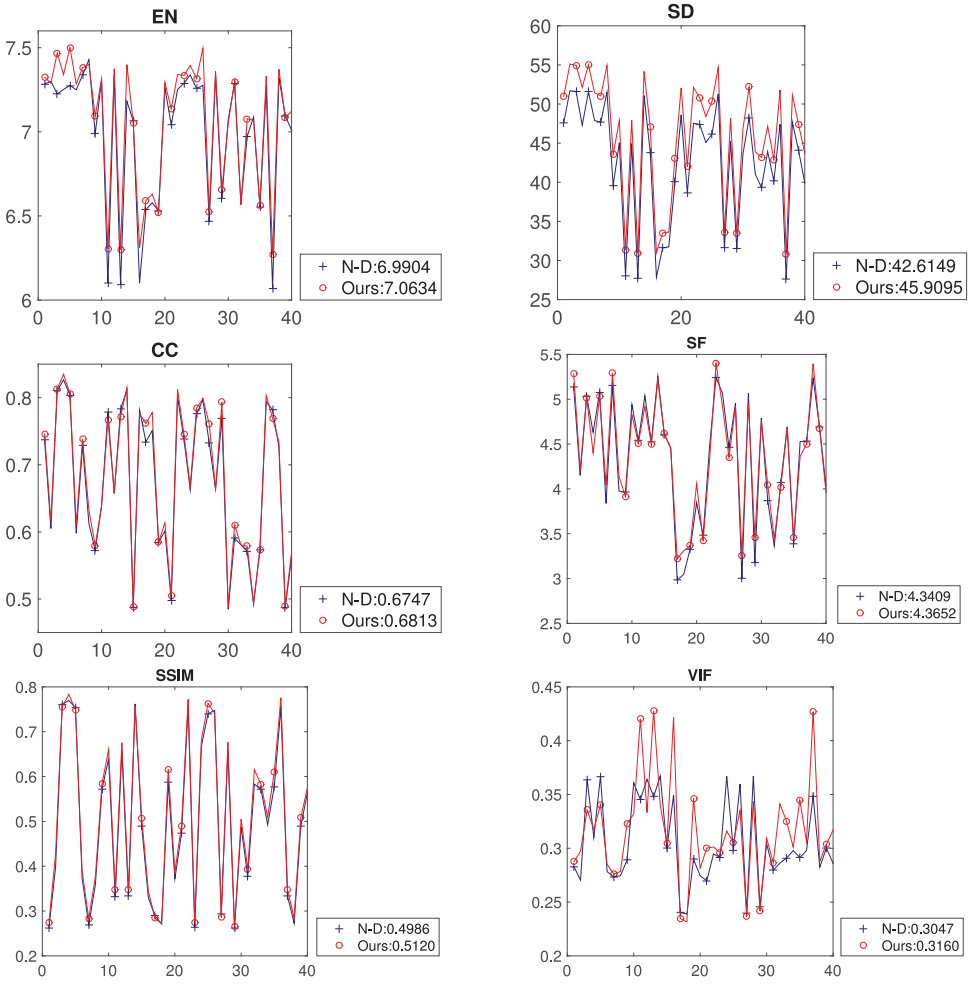

Fig. 8. The results of six fusion metrics on 40 infrared and visible image pairs from TNO dataset with detail loss (Ours) and without detail loss (N-D).

nearly cannot keep the high contrast property of target in the infrared image with only texture information in the visible image. For example, the people behind the smoke in the first example totally cannot be seen, while in rest three examples, fusion results can only preserve blurry outlines of the people which are not salient anymore. However, the results of D-model preserve the highlighted targets well and also contain abundant details in visible images, especially the sharpened branches in the first two examples. This demonstrates that the pre-trained VGG-Net has strong ability to extract high-level features in visible image but not in infrared image. Therefore, the detail loss in VGG-model makes fusion results concentrate on preserving more detail information rather than highlighting the targets. By contrast, D-model is more suitable to preserve both thermal radiation and texture detail information.

4.2.2. *Experiment 2*

We next test our D-model and VGG-model by using the feature maps of different layers, such as pool2, pool3, pool4 and pool5 layers, to compute the detail loss. An image pair named *sand path* is used for evaluation, as shown in Fig. 5. According to the results, for D-model, all four fused images nearly have the same characteristics that images look like sharpened infrared images with clear highlighted targets and abundant detail information. However, the fence along with the road is clearer in the result of pool5 layer. For VGG-model, we observe that no matter which layer of VGG-Net is used for computing detail loss, the fusion results cannot keep the high contrast property of target in the infrared image. This demonstrates that VGG-Net pretrained on visible image cannot extract high-level features of infrared information. To perform a

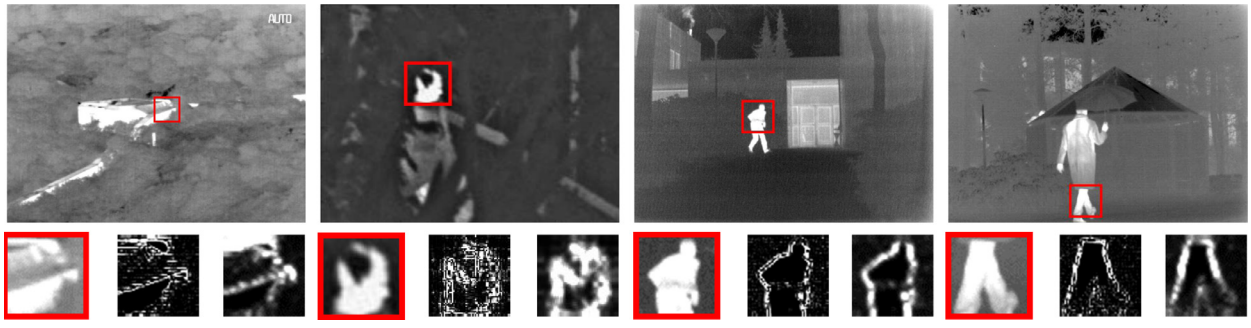

Fig. 9. Infrared images and their corresponding highlight target parts, edge map and *G* map. From left to right: *bunker, bush, Kaptein_1123* and *Kaptein_1654*.

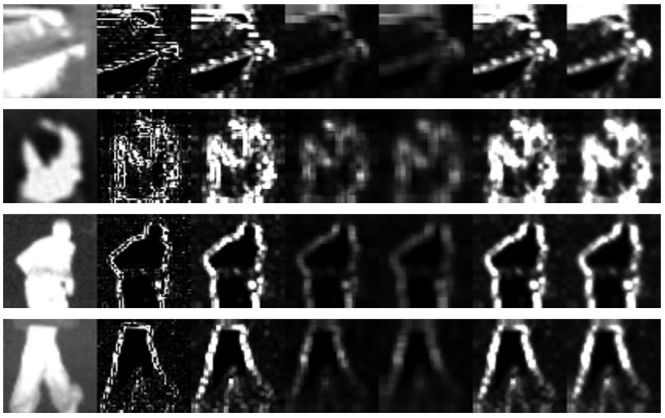

Fig. 10. Illustration of different kernel radius combinations. From left to right: infrared image patch from *bunker, bush, Kaptein_1123* and *Kaptein_1654*. From left to right: infrared image patch, edge map, G map at $N_{k=3}$, $N_{k=5}$, $N_{k=7}$, $N_{k=3}+N_{k=5}$ and $N_{k=3}+N_{k=5}+N_{k=7}$.

comprehensive assessment on selecting the best layer of D-model, we test these four candidates on an infrared and visible image sequence pair from the INO dataset and compute the six fusion metrics such as EN, SD, CC, SF, SSIM and VIF for comparison. The results are reported in Fig. 6. The pool5 layer clearly has the overall best performance for most image pairs. Consequently, we use discriminator as feature extractor with feature map of pool5 layer for computing detail loss.

4.2.3. Experiment 3

We further demonstrate the fusion results of our model with and without detail loss to verify the role of detail loss in promoting the fusion performance. Three different scenes from the TNO dataset such as *Kaptein_1654, sand path* and *bush* are used for evaluation, as shown in Fig. 7. As can be seen from the tent in *Kaptein_1654*, the fence in *sand path*, as well as the leaves in *bush*, the detail information is clearly more abundant in the results of model with detail loss, although both of them can well preserve the salient targets in the infrared images.

In addition, we quantitatively evaluate the six fusion metrics on 40 samples from the TNO dataset, and the results are reported in Fig. 8. From the results, we see that our model with detail loss consistently outperforms the one without detail loss in terms of all the six metrics on every image pair. Therefore, the detail loss does can enhance the visual effect of fused image and improve the quantitative fusion metrics.

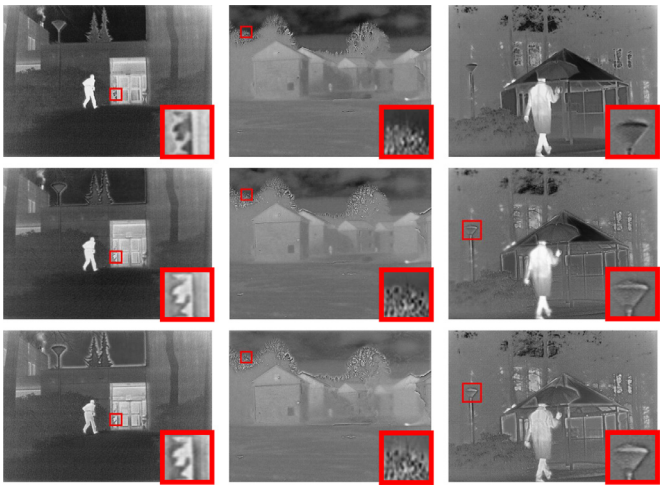

Fig. 12. Fusion results of three architectures on three scenes. From top to bottom: ShallowNet, Ours and DenseNet; from left to right: *Kaptein_1123, smoke* and *Kaptein_1654*.

4.3. Validation of target edge-enhancement loss

Next, we explain why we design G map to compute target edge-enhancement loss, and validate the function of target edge-enhancement loss based on the D-model.

In order to preserve the edges of infrared targets effectively, the most intuitive idea is to design a loss function like L_{gradient}, replacing $D_{v_{(x,y)}}$ with $D_{r_{(x,y)}}$. However, as can be seen in Fig. 9, the edge maps of infrared images are discrete and clutter because infrared images always contain lots of noise, which will influence the fusion performance. Therefore, we choose to adopt Gaussian kernel of different radiuses to filter the edge map, and then we can obtain a continuous and smooth map called G map, as shown in Fig. 9. The radiuses of kernel in our paper are empirically set to 3, 5 and 7. In addition, we also provide some qualitative results on different kernel radius combinations in Fig. 10. From the results, we see that the G map at $N_{k=3}+N_{k=5}+N_{k=7}$ in general could produce the best visual effect. Therefore, we set is as the default setting.

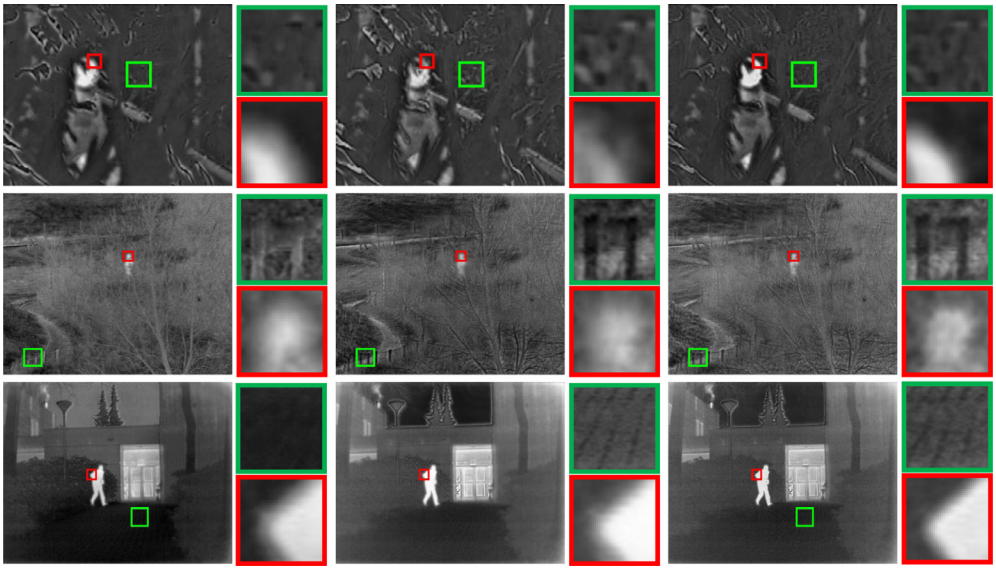

Fig. 11. Fusion results of FusionGAN (left), D-model (middle), and our approach (right), where red boxes highlight infrared target boundaries and green boxes highlight visible detail information. From top to bottom: *bush, sandpath* and *Kaptein_1123*. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In Fig. 11, we present some representative fusion results of FusionGAN, D-model (FusionGAN with detail loss) and our approach (FusionGAN with detail loss and target edge-enhancement loss). Whether results of FusionGAN or D-model, it is clear that the edges of infrared targets contain distinct burrs, such as the forehead edge in *bush* and elbow edge in *Kaptein_1123*. In contrast, our approach with target edge-enhancement loss can well address this problem, where the target boundaries of our results are well preserved and sharpened. In addition to the sharpened infrared target boundaries, we also find that the detail loss and target edge-enhancement loss can be optimized simultaneously without obvious conflicts. The evidence is that our fusion results also contain lots of detail information which is kept in D-model but not existed in FusionGAN, such as the leaves in *bush*, the fence in *sandpath* and streaks in *Kaptein_1123*. This demonstrates the effectiveness of our target edge-enhancement loss.

4.4. Influence of different architectures

In this section, we investigate the influence of different architectures in our framework. On the one hand, we investigate the influence of the depth of a network. Considering our 5-residual-block network is deep enough, we choose a shallower network named as ShallowNet, e.g. a

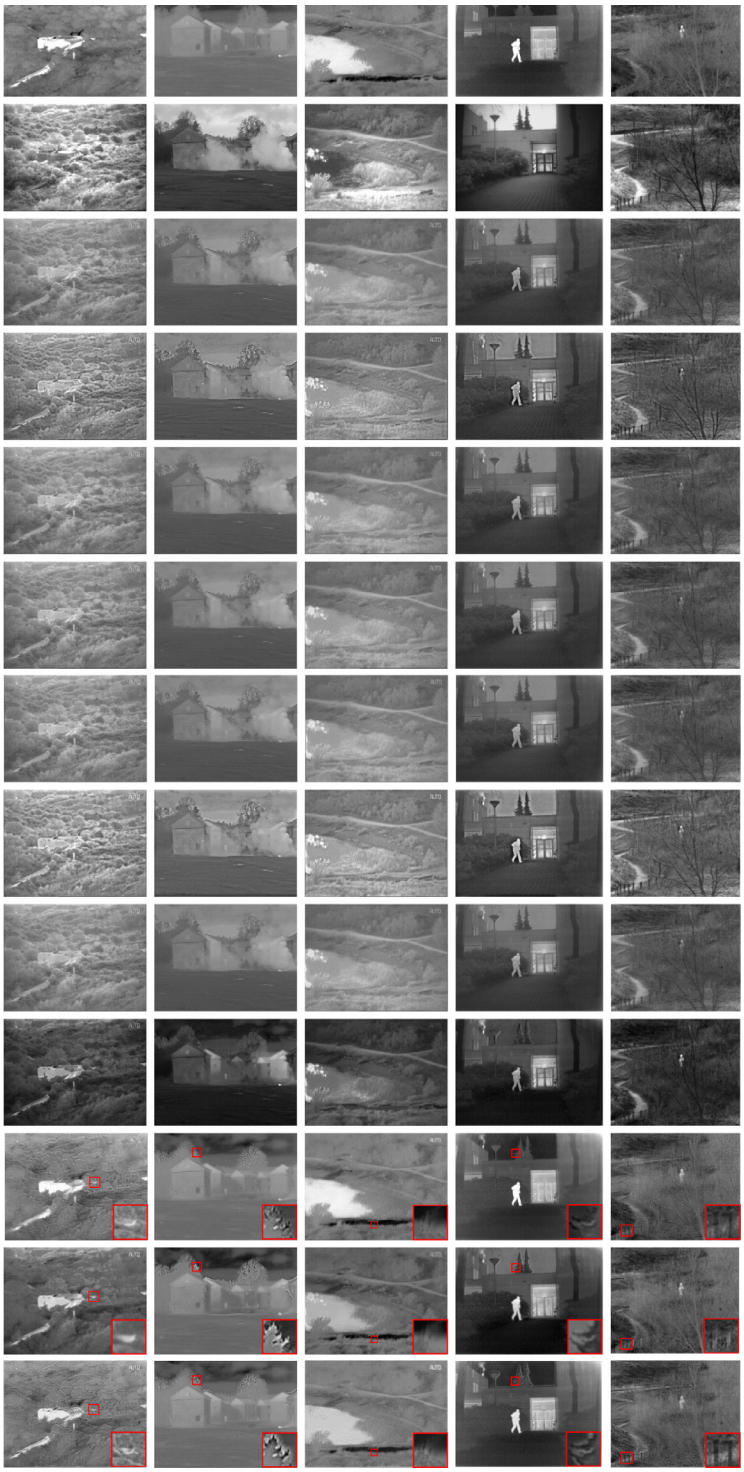

Fig. 13. Qualitative fusion results on five typical infrared and visible image pairs from the TNO dataset. From left to right: *bunker, smoke, lake, Kaptein_1123* and *sand path*. From top to bottom: infrared images, visible image, results of ADF [17], DTCWT [58], FPDE [12], IMSVD [59], IVIFDLF [33], TSIFVS [13], Wavelet [60], DenseFuse [34], GTF [35], FusionGAN [15] and our proposed method. To compare the preserved detail information of GTF, FusionGAN and our method in the last three rows, we have zoomed in some regions and put them at the lower right corner in each subplot.

4-residual-block network, for comparison. On the other hand, we investigate the influence of applying a different type of architecture named as DenseNet, e.g., using dense connections. In particular, we add dense connections to a 4-residual-blocks network.

Fig. 12 shows the fusion results of three different architectures on three different scenes. We can find that all the three architectures keep the radiation information well, but exhibit differences on detail information preservation. For example, the details in the red boxes in the results of ShallowNet are difficult to identify, but they are clear in the other two architectures. Moreover, compared with DenseNet, the targets in our fusion results are more salient, such as the people in all three scenes. Therefore, we conclude that both deeper networks and dense connections can improve the detail quality of fused image, and compared with dense connections, deeper architectures can preserve infrared information better.

4.5. Comparative experiments

In this section, we demonstrate the efficiency of the proposed method on publicly available datasets with comparison to other state-of-the-art fusion methods.

4.5.1. Results on TNO dataset

The TNO dataset contains multispectral (e.g. intensified visual, near-infrared, and longwave infrared or thermal) nighttime imageries of different military relevant scenarios that have been registered with different multiband camera systems. From the dataset, we choose 45 pairs of infrared and visible images as the training set and 12 pairs as the testing set. We select five typical pairs, such as *bunker, smoke, lake, Kaptein_1123*, and *sand path*, from the testing set for qualitative illustration, as shown in Fig. 13.

The first two rows in Fig. 13 present the original infrared and visible images. Our fusion results are shown in the last row, while the remaining ten rows correspond to the results of the competitors. All methods can fuse the information from the two source images to some extent. In this sense, it is difficult to judge which method is the best. However, the targets, such as bunker, window, lake, and human, of the other methods have low saliency in the fused images except those of GTF and FusionGAN, which suggests that the thermal radiation information in the infrared images is not well preserved. The observation can be attributed to the tendency of the methods to exploit the detail information in the source images, which then leads to difficulties in subsequent tasks, such as target detection and localization.

The results also show that our method, GTF and FusionGAN can highlight the targets in the fused images. However, the fusion results of our method contain much more detail information and sharpened edges of infrared targets. For example, in *Kaptein_1654*, the outline of the trees is much clearer and more sharpened in our result compared with GTF, while the streaks on the road is distinct in our result, but nearly cannot be observed in the result of FusionGAN. In *sand path*, the fence is perfectly fused in our result, but it is difficult to recognize in the results of GTF and FusionGAN. A similar phenomenon can also be observed in the other three examples. This finding demonstrates that our proposed method performs better than the other state-of-the-art methods when simultaneously preserving thermal radiation information, infrared target boundaries and texture detail information.

Furthermore, we perform a quantitative comparison of the eleven methods on all the 12 infrared and visible image pairs in the testing set.

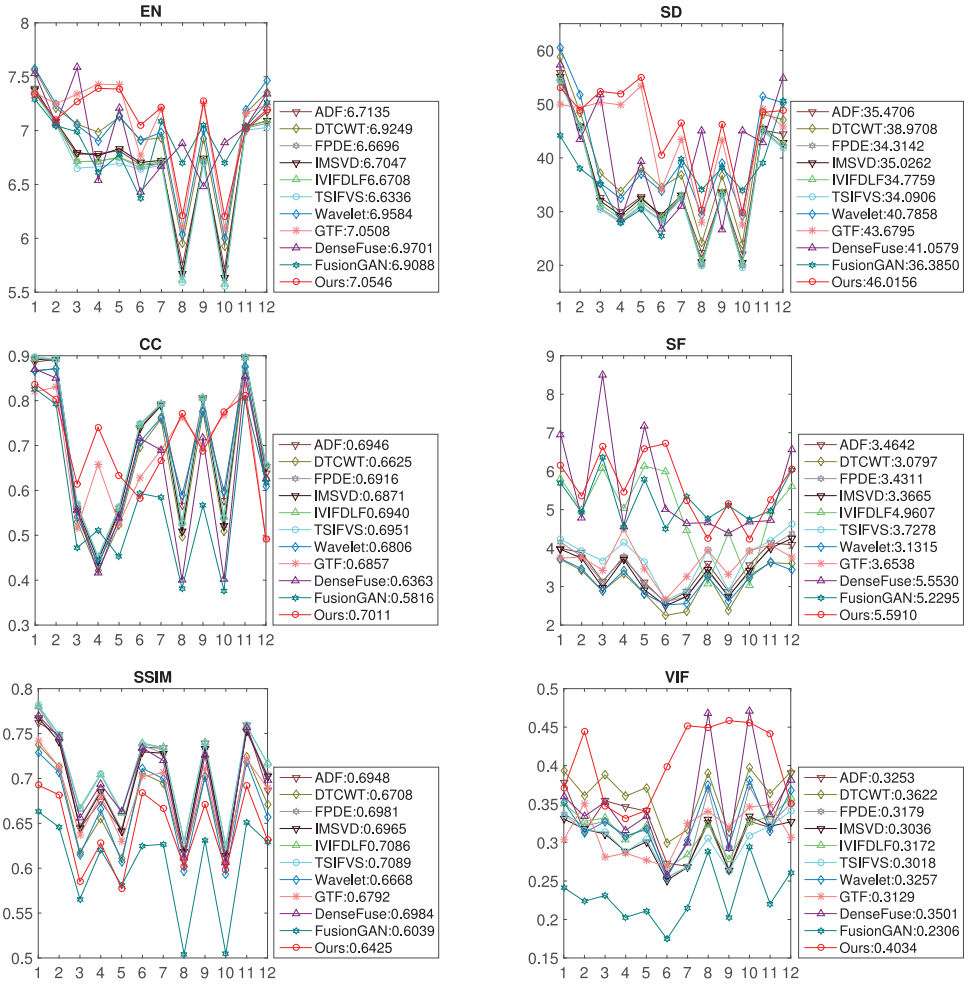

Fig. 14. Quantitative comparison of six fusion metrics on the TNO dataset. The ten state-of-the-art methods such as ADF [17], DTCWT [58], FPDE [12], IMSVD [59], IVIFDLF [33], TSIFVS [13], Wavelet [60] GTF [35], DenseFuse [34] and FusionGAN [15] are used for comparison.

Fig. 15. Schematic illustration of low SSIM. From left to right: infrared image, visible image, and our fusion result.

The results for the six metrics are shown in Fig. 14. Our method clearly obtains the best EN, SD, SF and VIF for most image pairs, and the average values for five evaluation metrics including CC are the largest relative to the other methods. The largest EN demonstrates that our fused image has much more abundant information than the other seven competitors. The largest SD suggests that our fused image has the best image contrast. The largest CC demonstrates that our fused image is strongly correlated with the two source images. The largest SF indicates that our fused image has richer edges and textures. The largest VIF means our fusion results are more consistent with the human visual system. Nevertheless, our method typically generates relatively low SSIM. This is due to that in order to preserve the radiation information and gradient information simultaneously, the pixel intensities of some areas in the fused image may be changed during training, and these areas will look neither like infrared image nor like visible image, leading to low structural similarity between source images and fused image. A typical example is illustrated in Fig. 15, where the stop line is white in visible image and cannot be visible in infrared image; however, in order to preserve the radiation information of ground and the edge texture of stop line, the area of stop line in the fused image gets black, which looks neither like infrared image nor like visible image. Similar phenomenon can be observed in the areas of truck and guide board. Therefore, the goal of simultaneously retaining the thermal radiation and abundant textural details will inevitably decrease the SSIM index.

4.5.2. Results on INO dataset

To verify its generalizability, we test our method on the INO dataset, which is trained on the TNO dataset. The INO dataset is provided by the National Optics Institute of Canada and contains several pairs of visible and infrared videos that represent different scenarios captured under different weather conditions. We captured 90 infrared and visible image pairs from the video named *trees and runner* for comparison.

The results of the quantitative comparison of the six fusion metrics are reported in Fig. 16. Our method has the best SD, CC, SF and VIF for all image pairs. Clearly, the average values of the evaluation metrics are the largest relative to those of the other ten methods. For the EN metric, our method is second to GTF by a narrow margin; limited by the content loss, we cannot get best SSIM, either. Moreover, we observe that the metrics of IVIFDLF [33] change greatly among different frames, especially for SSIM and VIF. This is due to that image reconstructions after downsampling operation in IVIFDLF will lead to misregistration between fusion results and source images, and this misregistration varies from frame to frame, which causes unsteady results.

We also present the run-time comparison of the eleven methods in Table 2. Our method has achieved comparable efficiency compared other ten methods.

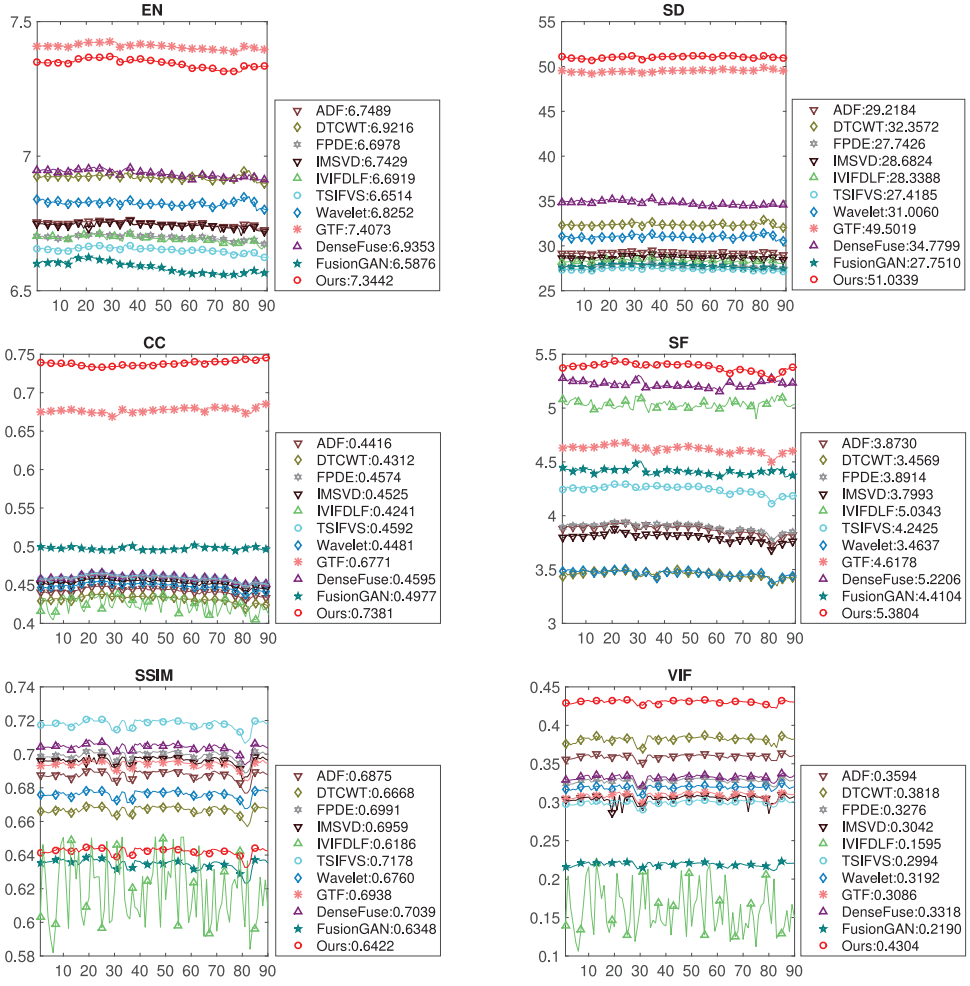

Fig. 16. Quantitative comparison of six fusion metrics on the INO dataset. The seven state-of-the-art methods such as ADF [17], DTCWT [58], FPDE [12], IMSVD [59], IVIFDLF [33], TSIFVS [13], Wavelet [60] GTF [35], DenseFuse [34] and FusionGAN [15] are used for comparison.

Table 2
Run time comparison of eleven methods on the TNO and INO datasets. The IVIFDLF, DenseFuse, FusionGAN and Our methods are performed on GPU while all the others are performed on CPU. Each value denotes the mean of run time of a certain method on a dataset (unit: second).

Method	TNO	INO
ADF [17]	6.47×10^{-1}	1.82×10^{-1}
DTCWT [58]	3.32×10^{-1}	1.30×10^{-1}
FPDE [12]	5.02×10^{-1}	9.22×10^{-2}
IMSVD [59]	5.19×10^{-1}	1.58×10^{-1}
IVIFDLF [33]	7.52	3.40
TSIFVS [13]	3.08×10^{-2}	1.16×10^{-2}
Wavelet [60]	2.23×10^{-1}	1.04×10^{-1}
DenseFuse [34]	5.26×10^{-1}	4.25×10^{-1}
GTF [35]	4.82	1.00
FusionGAN [15]	4.61×10^{-2}	4.50×10^{-2}
Ours	2.54×10^{-1}	7.16×10^{-2}

5. Conclusion

We propose a novel infrared and visible image fusion method based on GAN that can simultaneously retain the thermal radiation information in infrared images and the rich textural details in visible images. The proposed method is an end-to-end model, and can avoid the manual and complicated design of activity-level measurement and fusion rules of traditional fusion strategies. In particular, we design two loss functions i.e. detail loss and target edge-enhancement loss to improve the fusion performance. The detail loss is introduced to better exploit the textural details in the source images, while the target edge-enhancement loss aims to sharpen edges of infrared targets. Benefit from these two loss functions, our results can simultaneously well preserve thermal radiation information, infrared target boundaries and texture detail information. We demonstrate the effectiveness of using detail loss and target edge-enhancement loss in our experiments. The qualitative and quantitative comparisons reveal the superiority of our strategy over the state-of-the-art methods. Moreover, our method not only produces comparatively better visual effects, but also generally retains the largest or approximately the largest amount of information in the source images.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant nos. 61773295 and 61772512.

References

[1] G. Cui, H. Feng, Z. Xu, Q. Li, Y. Chen, Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition, Optics Commun. 341 (2015) 199–209.
[2] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (2019) 153–178.
[3] J. Han, H. Chen, N. Liu, C. Yan, X. Li, Cnns-based rgb-d saliency detection via cross-view transfer and multiview fusion, IEEE Trans. Cybernetics 48 (11) (2018) 3171–3183.
[4] 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.
[5] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.
[6] B. Ni, Y. Pei, P. Moulin, S. Yan, Multilevel depth and image fusion for human activity detection, IEEE Trans. Cybernetics 43 (5) (2013) 1383–1394.
[7] 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, Inf. Fusion 42 (2018) 158–173.
[8] J. Xiao, R. Stolkin, Y. Gao, A. Leonardis, Robust fusion of color and depth data for rgb-d target tracking using adaptive range-invariant depth models and spatio-temporal consistency constraints, IEEE Trans. Cybern. 48 (8) (2018) 2485–2499.
[9] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[10] B. Yang, S. Li, Visual attention guided image fusion with sparse representation, Optik-Int. J. Light Electron Optics 125 (17) (2014) 4881–4888.
[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] D.P. Bavirisetti, R. Dhuli, Two-scale image fusion of visible and infrared images using saliency detection, Infrared Phys. Technol. 76 (2016) 52–64.
[14] 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.
[15] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: a generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[16] 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.
[17] D.P. Bavirisetti, R. Dhuli, Fusion of infrared and visible sensor images based on anisotropic diffusion and karhunen-loeve transform, IEEE Sensors J. 16 (1) (2016) 203–209.
[18] S. Li, B. Yang, J. Hu, Performance comparison of different multi-resolution transforms for image fusion, Inf. Fusion 12 (2) (2011) 74–84.
[19] G. Pajares, J.M. De La Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (2004) 1855–1872.
[20] 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.
[21] Y. Liu, J. Jin, Q. Wang, Y. Shen, X. Dong, Region level based multi-focus image fusion using quaternion wavelet and normalized cut, Signal Process. 97 (2014) 9–30.
[22] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recognit. Lett. 9 (4) (1989) 245–253.
[23] M. Choi, R.Y. Kim, M.-R. Nam, H.O. Kim, Fusion of multispectral and panchromatic satellite images using the curvelet transform, IEEE Geosci.Remote Sens. Lett. 2 (2) (2005) 136–140.
[24] 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.
[25] 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.
[26] 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.
[27] 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. Technol. 67 (2014) 161–172.
[28] 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.
[29] 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.
[30] 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.
[31] C.H. Liu, Y. Qi, W.R. Ding, Infrared and visible image fusion method based on saliency detection in sparse domain, Infrared Phys. Technol. 83 (2017) 94–102.
[32] K.R. Prabhakar, V.S. Srikar, R.V. Babu, Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: IEEE International Conference on Computer Vision, 2017, pp. 4724–4732.
[33] H. Li, X.-J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in: International Conference on Pattern Recognition, 2018, pp. 2705–2710.
[34] H. Li, X.-J. Wu, Densefuse: a fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2019) 2614–2623.
[35] 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.
[36] 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.
[37] E.L. Denton, S. Chintala, R. Fergus, et al., Deep generative image models using a Laplacian pyramid of adversarial networks, in: Advances in Neural Information Processing Systems, 2015, pp. 1486–1494.
[38] K. Gregor, I. Danihelka, A. Graves, D.J. Rezende, D. Wierstra, Draw: a recurrent neural network for image generation, in: International Conference on Machine Learning, 2015, pp. 1462–1471.
[39] A. Dosovitskiy, T. Brox, Generating images with perceptual similarity metrics based on deep networks, in: Advances in Neural Information Processing Systems, 2016, pp. 658–666.
[40] A. Radford, L. Metz, S. Chintala, Unsupervised representation learning with deep convolutional generative adversarial networks (2015) arXiv:1511.06434.
[41] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, P. Abbeel, Infogan: Interpretable representation learning by information maximizing generative adversarial nets, in: Advances in Neural Information Processing Systems, 2016, pp. 2172–2180.
[42] M. Arjovsky, S. Chintala, L. Bottou, Wasserstein gan (2017) arXiv:1701.07875.
[43] 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.
[44] M. Mirza, S. Osindero, Conditional generative adversarial nets (2014) arXiv:1411.1784.

[45] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: Feature learning by inpainting, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[46] C. Li, M. Wand, Precomputed real-time texture synthesis with markovian generative adversarial networks, in: European Conference on Computer Vision, 2016, pp. 702–716.
[47] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.
[48] D. Yoo, N. Kim, S. Park, A.S. Paek, I.S. Kweon, Pixel-level domain transfer, in: European Conference on Computer Vision, 2016, pp. 517–532.
[49] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, 2016, pp. 694–711.
[50] J. Bruna, P. Sprechmann, Y. LeCun, Super-resolution with deep convolutional sufficient statistics (2015) arXiv:1511.05666.
[51] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[52] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[53] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014) arXiv:1409.1556.
[54] K. He, X. Zhang, S. Ren, J. Sun, Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1026–1034.
[55] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, ICML Workshop on Deep Learning for Audio, Speech and Language Processing, 2013.
[56] J. Ma, J. Zhao, J. Jiang, H. Zhou, X. Guo, Locality preserving matching, Int. J. Comput. Vision 127 (5) (2019) 512–531.
[57] J. Ma, X. Jiang, J. Jiang, J. Zhao, X. Guo, Lmr: Learning a two-class classifier for mismatch removal, IEEE Trans. Image Process 28 (8) (2019) 4045–4059, doi:10.1109/TIP.2019.2906490.
[58] 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.
[59] V. Naidu, Image fusion technique using multi-resolution singular value decomposition, Defence Sci. J. 61 (5) (2011) 479.
[60] H. Li, B. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, Graph. Models Image Process. 57 (3) (1995) 235–245.
[61] 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.
[62] Y.-J. Rao, In-fibre bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355.
[63] M. Deshmukh, U. Bhosale, Image fusion and image quality assessment of fused images, Int. J. Image Process. 4 (5) (2010) 484–508.
[64] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.
[65] Z. Wang, A.C. Bovik, A universal image quality index, IEEE Signal Process. Letters 9 (3) (2002) 81–84.
[66] 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.