Learning a Generative Model for Fusing Infrared and Visible Images via Conditional Generative Adversarial Network with Dual Discriminators

Han Xu[1] , **Pengwei Liang**[1] , **Wei Yu**[1] , **Junjun Jiang**[2] and **Jiayi Ma**[1*]

[1]Electronic Information School, Wuhan University, Wuhan 430072, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China
{xu_han, erfect, yuwei998}@whu.edu.cn, junjun0595@163.com, jyma2010@gmail.com

Abstract

In this paper, we propose a new end-to-end model, called dual-discriminator conditional generative adversarial network (DDcGAN), for fusing infrared and visible images of different resolutions. Unlike the pixel-level methods and existing deep learning-based methods, the fusion task is accomplished through the adversarial process between a generator and two discriminators, in addition to the specially designed content loss. The generator is trained to generate real-like fused images to fool discriminators. The two discriminators are trained to calculate the JS divergence between the probability distribution of downsampled fused images and infrared images, and the JS divergence between the probability distribution of gradients of fused images and gradients of visible images, respectively. Thus, the fused images can compensate for the features that are not constrained by the single content loss. Consequently, the prominence of thermal targets in the infrared image and the texture details in the visible image can be preserved or even enhanced in the fused image simultaneously. Moreover, by constraining and distinguishing between the downsampled fused image and the low-resolution infrared image, DDcGAN can be preferably applied to the fusion of different resolution images. Qualitative and quantitative experiments on publicly available datasets demonstrate the superiority of our method over the state-of-the-art.

1 Introduction

With the development of sensor technology, multi-modal images have been gaining in popularity in many fields, such as remote sensing, medical treatment, and target recognition [Li *et al.*, 2017; Zhang *et al.*, 2018]. Among different sensors, the combination of visible and infrared sensors has unique advantages. Visible images can represent texture details to the greatest content through the reflected light captured by visible sensors. Complementarily, the thermal radiation captured by infrared sensors can be represented in infrared images according to certain mapping relationships. Thus, the thermal targets can be highlighted by high contrast even in poor lighting conditions. Therefore, the fused images have the potential to present nearly all the inherent properties to improve visual understanding [Jin *et al.*, 2017], and play an important role in military and civilian applications [Ma *et al.*, 2019a].

The keystone of fusion is to extract the vital information in source images and merge it. For this purpose, researchers have proposed various feature extraction strategies and fusion rules, such as methods based on multi-scale transform [Zhou *et al.*, 2016], sparse representation [Zhu *et al.*, 2018], subspace, saliency [Ma *et al.*, 2017], hybrid [Paramanandham and Rajendiran, 2018], and other methods. Although these works have achieved promising performance, there are still some drawbacks. i) In traditional methods, the manually designed rules make methods more and more complex and complicated. ii) Drawbacks of deep learning-based methods are shown in Sec. 2.1. iii) As a whole, they focus on extracting and preserving features without considering the enhancement of vital features for more advantageous subsequent processing and applications. iv) Due to the limitations of hardware, infrared images always suffer from lower resolution. The method of downsampling visible images or upsampling infrared images will cause thermal radiation information blurring or texture detail loss. Thus, it remains a challenging task to fuse different resolution images.

To address the abovementioned challenges, we propose a method to learn a generative model via dual-discriminator conditional generative adversarial network (DDcGAN). The fusion task is accomplished through the adversarial process between a generator and two discriminators. The traditional GAN is adapted to GAN with dual-discriminators to preserve features in both types of source images. As for the discriminators, we apply the infrared image/gradient of the visible image as the real data, respectively. The downsampled fused image/gradient of the fused image should be indistinguishable with both types of real data, and hence the ground-truth fused image is not required. The entire network is an end-to-end model without the need to design fusion rules. Moreover, our model is suitable for the fusion of different-resolution images. Qualitative and quantitative results have revealed the advantages of our DDcGAN compared to other methods.

A typical example is shown in Figure 1. We compare the

*Corresponding author

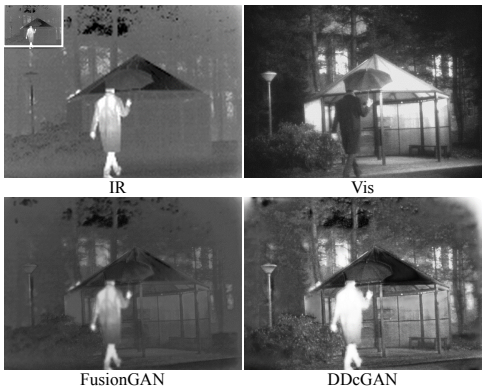


Figure 1: Schematic illustration of DDcGAN. The original low-resolution infrared image is enlarged and shown for better observation with the original infrared image in the upper left white box.

result of our method with FusionGAN [Ma *et al.*, 2019b]. With an additional discriminator, the fused image can highlight the thermal target to a greater extent. Moreover, with the discriminators employed to calculate the divergence between probability distributions rather than pixel-level differences, the generator is more likely to capture critical features and enhance them. In Figure 1, it represents as the contrast between thermal targets and the background. Compared with the thermal radiation information shown by defined mapping relationships in the infrared image, it is represented by higher contrast in our result for better target recognition. Meanwhile, more details in the visible image (*i.e.*, the lamp, the stool, and the bush) are kept in our result.

Contributions of our work include the following aspects:

- It has contributed in applying a deep learning framework for image fusion. On the one hand, it breaks through the limitation that most methods just apply deep learning framework in some sub parts. On the other hand, our work is not limited to applying deep learning to minimizing pixel-level losses. We solve it based on a min-max two-player game by the angle of probability distribution, in addition to the content loss.
- The architecture of dual-discriminator can avoid loss of information in one type of source image caused by introducing a discriminator on the other type of source image.
- As we solve it by angle of probability distribution, DDcGAN can not only extract, fuse, and reconstruction features, but also enhance vital features in source images, *i.e.*, the contrast between thermal targets and the background.
- In virtue of the downsampling operation before D_i and specially designed loss, our method demonstrates excellent performance for different-resolution image fusion.

2 Related Work

2.1 Deep Learning-based Fusion Methods

Since much attention has been drawn to deep learning recently, many deep learning-based fusion methods have been proposed. In some methods, deep learning is applied to extract image features. [Liu *et al.*, 2017] adopted CNN to generate a weight map while the overall process is based on pyramids. In [Li *et al.*, 2018], source images are decomposed into base parts and detail content, and deep learning is used in the detail content to extract features. In some methods, deep learning is also used for reconstruction. Prabhakar *et al*. [Prabhakar *et al.*, 2017] solved multi-exposure fusion by utilizing a novel CNN. Li *et al*. [Li and Wu, 2019] improved it while feature maps are still combined by manually designed rules. FusionGAN was proposed to fuse infrared and visible images using GAN [Ma *et al.*, 2019b]. The fused image generated is forced to have more details in the visible image by applying a discriminator to distinguish between them.

Although they have achieved promising performance, there are still some drawbacks. i) Existing methods perform neural network in feature extraction or reconstruction while the entire framework can not get rid of the limitations of traditional methods. ii) The stumbling block in utilizing deep learning for infrared and visible image fusion is the lack of ground-truth. Existing methods solve it by designing content loss functions. However, they may introduce new problems. For instance, the Euclidean distance suffers from blurred results. Therefore, it is difficult to design a comprehensive and adaptive loss function to specify a high-level goal. iii) Most artificially designed fusion rules lead to the extraction of same features while source images are manifestations of different phenomena. iv) Existing GAN-based methods force the fused image to obtain more details in visible images by introducing a discriminator. As the adversarial game proceeds, the prominence of thermal is gradually reduced. To address the problems, we apply GAN and adapt it with dual discriminators. We also adapt to different-resolution image fusion. In addition, for the stability of training process, we optimize the network architecture and training strategy.

2.2 Generative Adversarial Networks

GAN is designed to learn a probability distribution as an estimation of the real distribution $P_{data}(x)$. It solves the problem via an adversarial process by simultaneously training a generator G and a discriminator D [Goodfellow *et al.*, 2014]. G can generate samples by noise sampled from the latent space. The optimization formulation of G can be defined as:

$$G^* = \arg\min_G Div\left(P_G\left(x\right), P_{data}\left(x\right)\right), \tag{1}$$

where $Div(\cdot)$ denotes the divergence between two distributions. D can be used to calculate the divergence and the objective function can be formulated as:

$$D^* = \arg\max_D V\left(G, D\right), \tag{2}$$

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

$$V\left(G, D\right) = \mathbb{E}_{x\sim P_{data}}\left[\log D\left(x\right)\right] + \mathbb{E}_{x\sim P_G}\left[\log\left(1 - D\left(x\right)\right)\right]. \tag{3}$$

Thus, Eq. (1) can be converted to:

$$G^* = \arg\min_G \max_D V\left(G, D\right). \tag{4}$$

The adversarial process makes up the two-player min-max game. Hence, the generated samples are getting more and

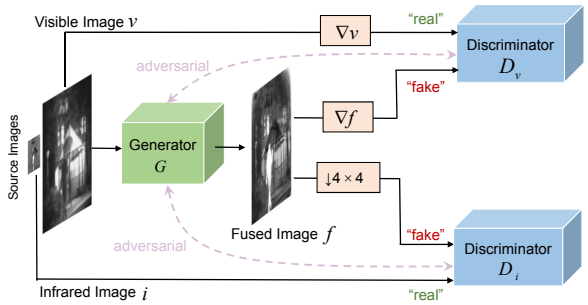


Figure 2: The entire procedure of DDcGAN for image fusion. ∇ denotes the gradient operator.

more indistinguishable from the real data. GAN can be extended to a conditional model if the generator and discriminator are conditioned on some extra information which is fed as additional input layer and this model is defined as conditional generative adversarial network (cGAN) [Mirza and Osindero, 2014].

3 Proposed Method

3.1 Problem Formulation

Because we are committed to solving the more challenging problem of different-resolution image fusion, without loss of generality, we assume that the ratio between the resolution of the visible image and that of the infrared image is set as 4. In other words, if the visible image is of size $m \times n$, the corresponding infrared image is of size $m/4 \times n/4$.

Given a visible image v and an infrared image i, the entire procedure of proposed DDcGAN is shown in Figure 2. The ultimate goal of our method is to learn a generator network G conditioned on v and i. Then the fused image $f = G(v, i)$ generated by G is encouraged to be realistic and informative enough to fool the discriminators. Simultaneously, we exploit two discriminator networks, D_v and D_i. Respectively, they generate a scalar that estimates the probability of the input from real data rather than G. The difference is that the real data of D_v and D_i is distinctive, even of different types. Specifically, D_v aims to distinguish the gradient of the generated image ∇f from the gradient of the visible image ∇v, while D_i is trained to discriminate between the original low-resolution infrared image i and down-sampled generated/fused image ψf, where ∇ is the gradient operator, and ψ is the downsampling operator.

A distinct change with traditional cGAN is that for the sake of the balance between the generator and discriminators, we don't feed ∇v and i as additional input layers to D_v and D_i. If so, the real data for D_v and D_i is the same with the extra input information. Thus, D_v and D_i are trained to discriminate whether two images are identical. As it is a simple enough task for neural networks and can be implemented through few layers of networks. However, for the generator, it will be a tough task to fool discriminators. Therefore, the adversarial relationship will fail to be established and the generator will tend to generate randomly. Consequently, the model will lose its original meaning.

Accordingly, the training target of G can be formulated as minimizing the following adversarial objective:

$$\begin{aligned}\min_{G}\max_{D_v,D_i} &\mathbb{E}\left[\log D_v\left(\nabla v\right)\right]+\mathbb{E}\left[\log\left(1-D_v\left(\nabla f\right)\right)\right]\\ &+\mathbb{E}\left[\log D_i\left(i\right)\right]+\mathbb{E}\left[\log\left(1-D_i\left(\psi f\right)\right)\right].\end{aligned} \tag{5}$$

Conversely, the goal of discriminators is to maximize Eq. (5).

Through the adversarial process of the generator and two discriminators, the divergence between two distributions, *i.e.*, $P_{\nabla F}$ and $P_{\nabla V}$, and the divergence between $P_{\psi F}$ and P_I will become smaller simultaneously. $P_{\nabla F}$ is the probability distribution of the gradients of the generated samples and $P_{\psi F}$ is that of the downsampled generated samples. $P_{\nabla V}$ is the probability distribution of the gradients of visible images and P_I is that of the infrared images.

3.2 Loss Function

Initially, the success of GANs was limited as they were known to be unstable to train and may result in artifacts and noisy or incomprehensible results [Zhang *et al.*, 2017]. A possible solution is to introduce a content loss to include a set of constraints into the networks. Thus, in this paper, the generator is not only trained to fool discriminators but also tasked to constraint similarity between the generated image and source images in content. Therefore, the loss function of the generator is composed by an adversarial loss $\mathcal{L}_G^{\text{adv}}$ and a content loss $\mathcal{L}_{\text{con}}$, with a weight λ controlling the trade-off:

$$\mathcal{L}_G = \mathcal{L}_G^{\text{adv}} + \lambda \mathcal{L}_{\text{con}}, \tag{6}$$

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

$$\mathcal{L}_G^{\text{adv}} = \mathbb{E}\left[\log\left(1-D_v\left(\nabla f\right)\right)\right]+\mathbb{E}\left[\log\left(1-D_i\left(\psi f\right)\right)\right]. \tag{7}$$

On the one hand, as the thermal radiation information in the infrared image is characterized by pixel intensities [Ma *et al.*, 2016], we employ the Frobenius norm to constrain the downsampled fused image to have similar pixel intensities with the infrared image. The downsampling operation can considerably prevent loss of texture information caused by compression or blur caused by forced upsampling. On the other hand, texture details in the visible image are mainly characterized by gradient variation. Thus, the TV norm [Beck and Teboulle, 2009] is applied to constrain the fused image to exhibit similar gradient variation with the visible image. With a weight η to control the trade-off, we can obtain the content loss:

$$\mathcal{L}_{\text{con}} = \mathbb{E}\left[\|\psi f - i\|_F^2 + \eta\|f - v\|_{TV}\right], \tag{8}$$

where ψ denotes the downsampling operator, which is implemented by two average pooling layers due to its retention of low-frequency information.

The discriminators are trained to discriminate between the real data and the generated data. The adversarial losses of discriminators can calculate the JS divergence between distributions and thus identify whether the pixel intensities or texture information is realistic. The adversarial losses of the discriminators are defined as follows:

$$\mathcal{L}_{D_v} = \mathbb{E}\left[-\log D_v\left(\nabla v\right)\right]+\mathbb{E}\left[-\log\left(1-D_v\left(\nabla f\right)\right)\right], \tag{9}$$

$$\mathcal{L}_{D_i} = \mathbb{E}\left[-\log D_i\left(i\right)\right]+\mathbb{E}\left[-\log\left(1-D_i\left(\psi f\right)\right)\right]. \tag{10}$$

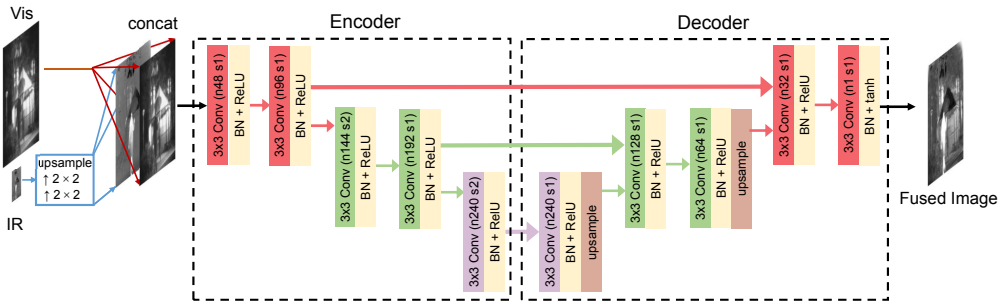


Figure 3: The overall architecture of our generator. Conv(np sq): convolutional layer which obtains p feature maps and the stride is set as q. The same color indicates that these feature maps have the same width and height.

3.3 Network Architecture

Generator Architecture

The generator network is the encoder-decoder network with 2 upsampling layers before the encoder, as presented in Figure 3. Since the infrared image has a lower resolution, we firstly introduce two upsampling layers by nearest neighbor interpolation to transform between two resolutions. The output of these 2 layers is an upsampled infrared image. The upsampled infrared image and the original visible image are concatenated and fed to the encoder. The process of feature extraction and fusion are both performed in the encoder and fused feature maps are produced. These maps are then fed to the decoder for reconstruction. The generated fused image is of the same resolution with the visible image.

The encoder consists of 5 convolutional layers. The number of output feature maps and the stride of each convolutional layer is shown in Figure 3. If the feature maps in red are of size $W \times H$, the feature maps in green and in purple are of size $W/2 \times H/2$, and $W/4 \times H/4$, respectively. Considering the loss caused by the stride set as 2 in the second and fourth layer in the encoder, *U-net* [Ronneberger *et al.*, 2015] is applied in the generator architecture. The feature maps obtained by the second and the fourth layers in the encoder are transferred to the corresponding layers in the decoder. These feature maps are concatenated with the feature maps obtained by the decoder itself for subsequent convolution and upsampling operations. The decoder is a 5-layer CNN and the setting of each layer is illustrated in Figure 3. The strides of all convolutional layers are set as 1. The feature maps obtained by the first and third convolutional layer are upsampled by nearest neighbor interpolation similarly. To avoid exploding/vanishing gradients and speed up training and convergence, batch normalization (BN) and ReLU activation function are applied.

Discriminator Architecture

Discriminators are designed to play an adversarial role against the generator. In particular, D_v aims to distinguish gradients of generated images from gradients of visible images and D_i aims to distinguish the generated images from the infrared images, respectively. However, these two types of source images are manifestations of different phenomena, thus have considerably different distributions. In other words, there are conflicts in the guidance of D_v and D_i on G. In our network, we should not only consider the adversarial relationship between the generator and discriminators but also take into account the balance of D_v and D_i. Otherwise, either strength or weakness of one discriminator will finally lead to the inefficiency of the other as the training proceeds. In our work, the balance is achieved by the design of architectures and training strategy. D_v and D_i share the same architecture, as shown in Figure 4. The stride of all convolutional layers is set as 2. In the last layer, we employ the tanh activation function to generate a scalar that estimates the probability of the input image from source images rather than G.

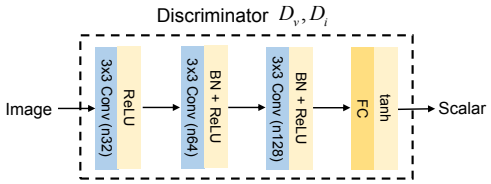


Figure 4: The overall architecture of our discriminator. 3×3: filter size, FC: fully connected layer.

4 Experimental Results

4.1 Dataset and Training Details

Dataset. To validate the effect of DDcGAN, we perform experiments on the publicly available *TNO Human Factors* dataset[1] for the infrared and visible image fusion. 36 infrared and visible images are selected and cropped to 27000+ patch pairs as the training dataset. As we assume that the resolution of visible images is 4×4 that of infrared images, all visible patches are of size 84×84 and all infrared patches are down-sampled to size 21×21. λ is set as 0.8 and η is set as 3. The learning rate is set as 0.002 with exponentially decaying. The batch size is 24 and the epoch is set as 1.

Training Details. The principle is to make the generator and discriminators form adversarial relationships. In order to overcome the unstableness and improve results, rather than taking turns training G, D_v and D_i once per batch in principle, we train D_v or D_i more times if it fails to discriminate the data generated from G and vice verse. During the testing phase, only the generator is used to generate fused images. Since there are no fully connected layers in our generator, the input can be the entire image rather than image patches.

[1] Available at: https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.

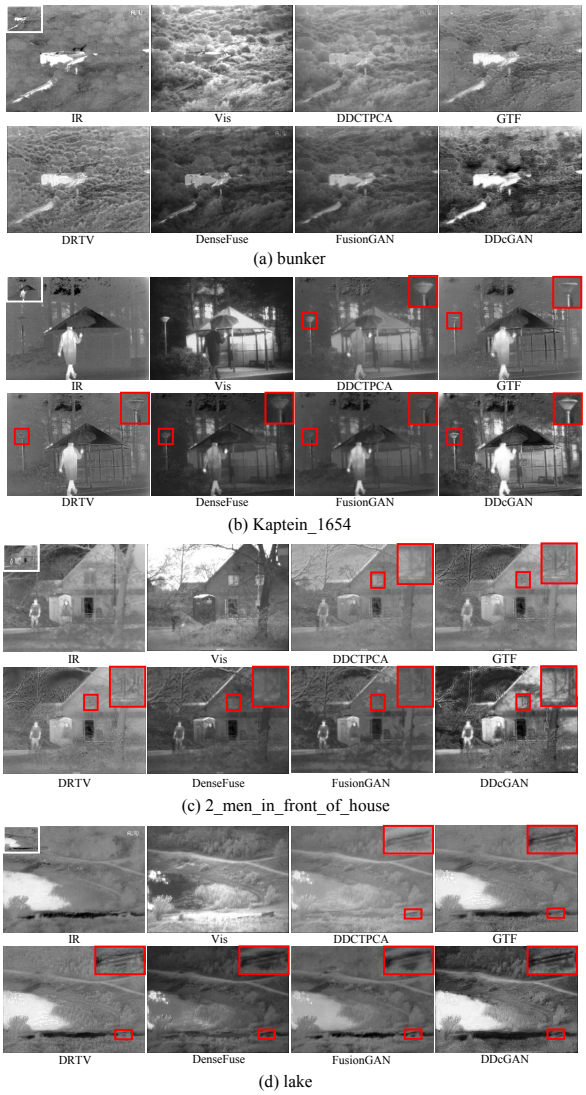


Figure 5: Qualitative comparison of our DDcGAN with 5 state-of-the-art methods on 4 typical infrared and visible image pairs.

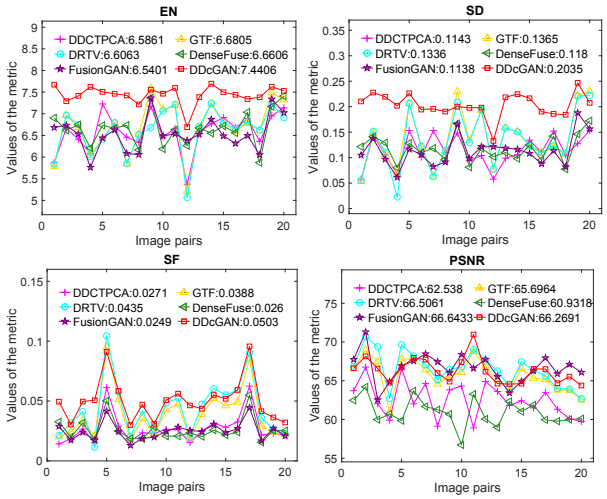


Figure 6: Quantitative comparison of our DDcGAN for infrared and visible image fusion with 5 state-of-the-art methods. Means of metrics are shown in legends.

4.2 Results and Analysis

We compare DDcGAN with 5 state-of-the-art methods, including three traditional methods, *i.e.*, DDCTPCA [Naidu, 2014], GTF [Ma *et al.*, 2016] and DRTV [Du *et al.*, 2018], and two deep learning-based methods, *i.e.*, DenseFuse [Li and Wu, 2019] and FusionGAN [Ma *et al.*, 2019b]. As some competitors require that source images have the same resolution, we upsample the low-resolution infrared images before performing these methods.

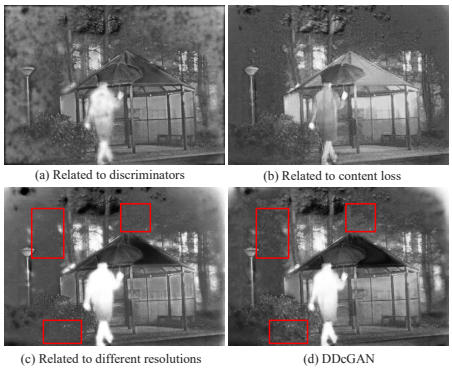
Qualitative Comparisons

Qualitative experiments are firstly performed. The intuitive results are shown in Figure 5 on four typical image pairs. Compared with existing methods, DDcGAN has three distinctive advantages. First, the thermal radiation information in the infrared image can be preserved and enhanced in our result. As shown in Figure 5(a), the contrast between the bunker and the background is higher than the original contrast in the infrared image, which is conducive to target detection. Second, our results can preserve texture details in visible images to a greater extent, as can be seen in Figure 5(b) and (c), which is beneficial for subsequent target recognition and the improvement of recognition accuracy. Third, our results are clearer due to that it does not suffer from thermal radiation information blurring, as shown in Figure 5(d).

As can be seen from Figure 5, DDCTPCA and DenseFuse cannot highlight the thermal targets well, while GTF, DRTV and FusionGAN cannot obtain abundant texture details. Besides, they all suffer from thermal radiation information blurring except DRTV and FusionGAN. Despite this, the results of DRTV inevitably suffer from staircase effects. In contrast, results of DDcGAN can obviously avoid staircase effects and details are more similar to those in the visible images. Compared with FusionGAN, due to the employment of the introduction of D_i, different network architecture, specially designed content loss and improved training strategy, our results can highlight thermal targets more obviously by higher contrast and meanwhile, contain more natural details which are more indistinguishable from the visible images. Generally, our DDcGAN works well and the fused images are more like super-resolved and contrast-enhanced infrared images which also contain more texture detail information in visible images.

Quantitative Comparisons

In addition to qualitative experiments, we further compare our DDcGAN with the above-mentioned competitors quantitatively on the rest 20 image pairs from the dataset. Four metrics, *i.e.*, entropy (EN), standard deviation (SD), spatial frequency (SF), and peak signal-to-noise ratio (PSNR) are used for evaluation. The results are summarized in Figure 6. As can be seen, our DDcGAN can generate the largest average values on the first 3 metrics: EN, SD and SF. In particular, DDcGAN achieves the best values of EN, SD and SF on 19, 18 and 15 image pairs, respectively. It is worth mentioning that SD is a metric reflecting contrast and distribution. The largest average value on SD is sufficient to prove that our results have



(a) Related to discriminators (b) Related to content loss

(c) Related to different resolutions (d) DDcGAN

Figure 7: Fused results of three comparative experiments and proposed DDcGAN on the image pair " Kaptein_1654".

the highest contrast between thermal targets and the background. For PSNR, DDcGAN can achieve comparable results with the average values being the third largest. By definition, PSNR is determined by the peak value and the mean square error (MSE) between the fused image and source images. Our method is designed with the aim of highlighting thermal targets by keeping thermal radiation information, leading to a large MSE between the fused image and the visible image. Besides, the higher contrast in the fused image also results in a comparatively large MSE between the fused and infrared images. These results demonstrate that our method can reserve information to the greatest extent, especially the most information, the highest contrast, the richest edges and texture details and considerable similarity with source images.

4.3 Comparative Experiments

In this part, we mainly perform three groups of comparative experiments, *i.e.*, comparative experiments related to discriminators, content loss, and different resolutions, respectively.

Experiment Related to Discriminators

To verify the effect of the conditional generative adversarial network with dual discriminators on fused results, we remove discriminators and the whole network merely contains the generator. Thus, the adversarial relationships no longer exist. The training goal is only to minimize the content loss $\mathcal{L}_{\text{con}}$ in Eq. (8). However, as the problem defined by the content loss is the first-order total variation model, the result is inevitably influenced by staircase effects [Lu *et al.*, 2016], as can be seen in Figure 7(a). While in the fused image of DDcGAN, the staircase effects are significantly weakened because the adversarial relationship requires that divergences between probability distributions should be as small as possible.

Experiment Related to Content Loss

To validate the benefit of the content loss $\mathcal{L}_{\text{con}}$, we compare the fused result of DDcGAN with the result by replacing the content loss with a commonly used loss, *i.e.*, MSE, for both the intensity differences between the fused image and two source images, which is defined as follows:

$$\mathcal{L}_{\text{con}} = \mathbb{E}\left[\|\psi f - i\|_F^2 + \eta\|f - v\|_F^2\right], \qquad (11)$$

where η is set as $1/16$. The fused result is shown in Figure 7(b). As can be seen, the fused result of modifying $\mathcal{L}_{\text{con}}$ as Eq. (11) cannot highlight thermal target, *i.e*, the pedestrian. Moreover, the edges are clearly serrated and the result is blurred due to the MSE losses.

Methods	DDCTPCA	GTF	DRTV	DenseFuse	FusionGAN	DDcGAN
Mean	107.15	7.19	5.85	0.51	0.56	0.78
STD	53.82	3.93	4.54	0.35	0.81	0.92

Table 1: Average runtime comparison of different methods on the 20 testing image pairs (unit: second).

Experiment Related to Different Resolutions

In order to deal with fusion of different resolution images, we constrain the intensity differences between the downsampled fused image and the original infrared image with the Frobenius norm in the content loss. Moreover, we adjust the input of the discriminator D_i as the downsampled fused image and the low-resolution infrared image. To verify the influence of the impact of these two operations on the fused results. We change the content loss $\mathcal{L}_{\text{con}}$ to the loss defined as:

$$\mathcal{L}_{\text{con}} = \mathbb{E}\left[\|f - \psi' i\|_F^2 + \eta\|f - v\|_{TV}\right], \qquad (12)$$

where ψ' denotes the inverse operator of the downsampling operator ψ, *i.e.*, the upsampling operator. Furthermore, D_i is employed to distinguish between the fused image f and the upsampled infrared image $\psi' i$. Results are shown in Figure 7(c). As can been seen from red boxes, the result of DDcGAN can prevent loss of texture information caused by compression or blur and inaccuracy caused by forced upsampling. Thus, it presents more texture details than Figure 7(c).

The average runtime of methods on the testing data is provided in Table 1. DDCTPCA, GTF, and DRTV are tested on on a desktop computer with 3.4 GHz Intel Core i5 CPU. The three deep learning methods, *i.e.*, DenseFuse, FusionGAN and DDcGAN are tested on NVIDIA Geforce GTX Titan X. Clearly, our DDcGAN can achieve comparable efficiency.

5 Conclusion

In this paper, we propose a new deep learning and GAN-based infrared and visible image fusion method by constructing a dual-discriminator conditional GAN, named DDcGAN. It does not require the ground-truth fused images for training, and can fuse images of different resolutions without introducing thermal radiation information blurring or visible texture detail loss. Extensive comparisons on four metrics with other five state-of-the-art fusion algorithms demonstrate that our DDcGAN can not only identify the most valuable information, but also can keep the largest or approximately the largest amount of information in the source images. In our future work, we will apply our DDcGAN to multi-modal medical images of different resolutions, *e.g.*, low-resolution positron emission tomography (PET) images and high-resolution magnetic resonance imaging (MRI) images.

Acknowledgments

This paper was supported by the National Natural Science Foundation of China under Grant no. 61773295.

References

[Beck and Teboulle, 2009] Amir Beck and Marc Teboulle. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Transactions on Image Processing*, 18(11):2419–2434, 2009.

[Du *et al.*, 2018] Qinglei Du, Han Xu, Yong Ma, Jun Huang, and Fan Fan. Fusing infrared and visible images of different resolutions via total variation model. *Sensors*, 18(11):3827, 2018.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[Jin *et al.*, 2017] Xin Jin, Qian Jiang, Shaowen Yao, Dongming Zhou, Rencan Nie, Jinjin Hai, and Kangjian He. A survey of infrared and visual image fusion methods. *Infrared Physics & Technology*, 85:478–501, 2017.

[Li and Wu, 2019] Hui Li and Xiao-Jun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019.

[Li *et al.*, 2017] Shutao Li, Xudong Kang, Leyuan Fang, Jianwen Hu, and Haitao Yin. Pixel-level image fusion: A survey of the state of the art. *Information Fusion*, 33:100–112, 2017.

[Li *et al.*, 2018] Hui Li, Xiao-Jun Wu, and Josef Kittler. Infrared and visible image fusion using a deep learning framework. *arXiv preprint arXiv:1804.06992*, 2018.

[Liu *et al.*, 2017] Yu Liu, Xun Chen, Juan Cheng, and Hu Peng. A medical image fusion method based on convolutional neural networks. In *Proceedings of the International Conference on Information Fusion*, pages 1–7. IEEE, 2017.

[Lu *et al.*, 2016] Wenqi Lu, Jinming Duan, Zhaowen Qiu, Zhenkuan Pan, Ryan Wen Liu, and Li Bai. Implementation of high-order variational models made easy for image processing. *Mathematical Methods in the Applied Sciences*, 39(14):4208–4233, 2016.

[Ma *et al.*, 2016] Jiayi Ma, Chen Chen, Chang Li, and Jun Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016.

[Ma *et al.*, 2017] Jinlei Ma, Zhiqiang Zhou, Bo Wang, and Hua Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82:8–17, 2017.

[Ma *et al.*, 2019a] Jiayi Ma, Yong Ma, and Chang Li. Infrared and visible image fusion methods and applications: a survey. *Information Fusion*, 45:153–178, 2019.

[Ma *et al.*, 2019b] Jiayi Ma, Wei Yu, Pengwei Liang, Chang Li, and Junjun Jiang. Fusiongan: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Naidu, 2014] VPS Naidu. Hybrid ddct-pca based multi sensor image fusion. *Journal of Optics*, 43(1):48–61, 2014.

[Paramanandham and Rajendiran, 2018] Nirmala Paramanandham and Kishore Rajendiran. Multi sensor image fusion for surveillance applications using hybrid image fusion algorithm. *Multimedia Tools and Applications*, pages 1–32, 2018.

[Prabhakar *et al.*, 2017] K Ram Prabhakar, V Sai Srikar, and R Venkatesh Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4724–4732, 2017.

[Ronneberger *et al.*, 2015] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 234–241. Springer, 2015.

[Zhang *et al.*, 2017] He Zhang, Vishwanath Sindagi, and Vishal M Patel. Image de-raining using a conditional generative adversarial network. *arXiv preprint arXiv:1701.05957*, 2017.

[Zhang *et al.*, 2018] Qiang Zhang, Yi Liu, Rick S Blum, Jungong Han, and Dacheng Tao. Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review. *Information Fusion*, 40:57–75, 2018.

[Zhou *et al.*, 2016] Zhiqiang Zhou, Bo Wang, Sun Li, and Mingjie Dong. Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters. *Information Fusion*, 30:15–26, 2016.

[Zhu *et al.*, 2018] Zhiqin Zhu, Hongpeng Yin, Yi Chai, Yanxia Li, and Guanqiu Qi. A novel multi-modality image fusion method based on image decomposition and sparse representation. *Information Sciences*, 432:516–529, 2018.