GANMcC: A Generative Adversarial Network with Multi-classification Constraints for Infrared and Visible Image Fusion

Jiayi Ma, Hao Zhang, Zhenfeng Shao, Pengwei Liang, and Han Xu

***Abstract*—Visible images contain rich texture information, while infrared images have significant contrast. It is advantageous to combine these two kinds of information into a single image so that it not only has good contrast, but also contains rich texture details. In general, previous fusion methods cannot achieve this goal well, where the fused results are inclined to either a visible or an infrared image. To address this challenge, a new fusion framework called generative adversarial network with multi-classification constraints (GANMcC) is proposed, which transforms image fusion into a multi-distribution simultaneous estimation problem to fuse infrared and visible images in a more reasonable way. We adopt a generative adversarial network with multi-classification to estimate the distributions of visible light and infrared domains at the same time, in which the game of multi-classification discrimination will make the fused result have these two distributions in a more balanced manner, so as to have significant contrast and rich texture details. In addition, we design a specific content loss to constrain the generator, which introduces the idea of main and auxiliary into the extraction of gradient and intensity information, which will enable the generator to extract more sufficient information from source images in a complementary manner. Extensive experiments demonstrate the advantages of our GANMcC over the state-of-the-art methods in terms of both qualitative effect and quantitative metric. Moreover, our method can achieve good fused results even the visible image is overexposed. Our code is publicly available at https://github.com/jiayi-ma/GANMcC.**

***Index Terms*—Image fusion, generative adversarial network, infrared, multi-classification, deep learning.**

I. Introduction

IMAGE fusion is to extract meaningful information from images captured by different sensors and then combine it to generate a single image, which contains richer information or more favorable for subsequent applications. Among them, infrared and visible image fusion is probably the most widely used [1]. Visible image is generated by the visible sensor capturing reflected light. It is characterized by rich texture detail information and conforms to the human eye observation law. Infrared sensors can perceive infrared band and convert the thermal radiation information to generate a gray-scale image. Infrared image is characterized by strong contrast and can effectively distinguish between background and target, even at night and in inclement weather. Infrared and visible image fusion combines these two characteristics together to generate images with significant contrast and rich texture details, which has good application prospects in the fields of military surveillance, object detection and vehicles night navigation [2], [3], [4]. Figure 1 shows an example to illustrate this point. The infrared image at night can effectively highlight vehicles and pedestrians on highways, while visible light images can retain traffic signs. The fused image can integrate these two advantages to be more conducive to the robot or vehicle understanding of the current scene.

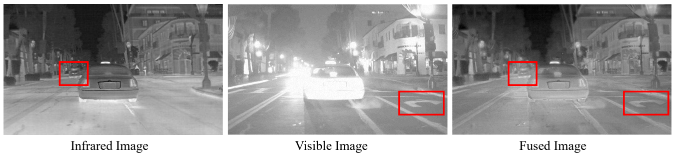


Fig. 1: An application example of infrared and visible image fusion in vehicle navigation.

The key to image fusion is the extraction and reconstruction of the most meaningful information [5], [6]. For infrared and visible image fusion, the most meaningful information is the significant contrast and rich texture, which is desirable to be preserved in the ideal result. In order to achieve this goal, researchers have proposed many image fusion methods, which can be divided into traditional methods and deep learning-based methods. Traditional methods measure the activity level of pixels or regions in the spatial domain or the transform domain, and realize image fusion according to specific fusion rules. Typical traditional methods are sparse representation-based methods [7], [8], multi-scale transform-based methods [9], [10], subspace-based methods [11], saliency-based method [12], [13] and hybrid methods [14], [15]. The deep learning-based methods [16], [17], [18], [19], [20] utilize the powerful nonlinear fitting ability of neural network to make the fused image have the desired distribution, and such methods can often produce more promising results.

Although the existing methods have achieved positive results under most conditions, there are several negativeness that should not be ignored. Firstly, the activity level measurement and fusion rules in traditional methods often require manual design, which become very complex because of the diversity

This work was supported by the Natural Science Fund of Hubei Province under Grant no. 2019CFA037, and the National Natural Science Foundation of China under Grant nos. 61773295 and 41890820. *(Corresponding author: Zhenfeng Shao.)*

J. Ma, H. Zhang, P. Liang and H. Xu are with the Electronic Information School, Wuhan University, Wuhan, 430072, China (e-mail: jyma2010@gmail.com, zhpersonalbox@gmail.com, erfect@whu.edu.cn, xu_han@whu.edu.cn).

Z. Shao is with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, 430079, China (e-mail: shaozhenfeng@whu.edu.cn).

of source images. In general, the rules of manual design are partial, limiting the fusion performance. Secondly, due to the lack of ground truth, the deep learning-based methods only realize image fusion by designing content loss function. The distribution learned in this way is not comprehensive enough. Third, even though the main information contained in different types of source images is different, there is still some secondary information contained in each other, which is desirable to be preserved in the final fused image. However, it has not been considered in previous methods. Fourth, it is difficult for most existing methods to achieve a good balance in maintaining infrared and visible inherent information. For example, the fused results of some methods tend to be visible images; although they contain rich texture details, they have no significant contrast and cannot clearly distinguish the targets from the background, such as [21], [22], [23]. Conversely, the results of some methods are closer to infrared images; they have better contrast information, but the texture is not rich enough, *e.g.*, more like sharp infrared images, such as [16], [24]. The unbalanced information in the fused result is harmful to subsequent upper-level tasks, such as the accuracy reduction in target detection.

The motivation of our method is mainly composed of two aspects. First of all, sufficient and effective information extraction is a prerequisite for good fusion. The previous methods believe that the expected contrast information only comes from the infrared image, and the desired texture information is only contained in the visible image. However, we find that infrared images also have some texture details, which in some cases are even very rich. Similarly, visible images also contain contrast information. We give two typical examples in Fig. 2. In the first example, the texture of some objects in the infrared image is even clearer than that in the visible image, such as the tree and grass highlighted in red boxes. In the second example, the visible image also contains some contrast information, such as the pavilion. This information should not be ignored. Second, aiming at the unbalanced fusion problem of existing methods, we think the key to ensuring that the fused image has both significant contrast and rich texture details is to ensure that the contrast and gradient information from source images is balanced, rather than biased. This is essentially a simultaneous estimation of the distribution of two different domains. Some specific network architectures can learn the distribution of observation data, such as SQAE [25], DBN [26], and LSTM [27], but they often need labels. Fortunately, the generative adversarial network (GAN) can better estimate the probability distribution of the target without supervision , while GAN with multi-classifier can further fit multiple distribution characteristics simultaneously. Therefore, it is suitable to solve this problem.

On the basis of the above observations, we design a generative adversarial network with multi-classification constraints (GANMcC), which can maintain contrast and texture details simultaneously. In particular, a specific content loss is designed to constrain the generator's extraction and processing of source image features, which addresses under-utilization of information. We not only construct the main infrared intensity loss between the fused image and infrared image, but also construct the auxiliary gradient loss, because the infrared image also contains texture details, and in some cases, they are even very rich. Similarly, we construct the main gradient loss and auxiliary intensity information loss between the fused image and visible image. This complementary loss also allows our method to generate good fused results when the visible images are overexposed (*e.g.*, the gradient changes greatly). In addition, we design a multi-classification generative adversarial network to address the challenge of unbalanced information fusion. In our model, the multi-modal image fusion is transformed into simultaneous estimation of multiple distributions. Concretely, we use a multi-classifier as discriminator, which can determine the probabilities that the input is an infrared image and is a visible image. For fused images, under the multi-classification constraints, the generator expects that the two probabilities are both high, that is, the discriminator considers it both an infrared image and a visible image; while the discriminator expects the two probabilities to be small at the same time, that is, the discriminator determines that the fused image is neither an infrared image nor a visible image. During this process, we constrain these two probabilities simultaneously to ensure the fused image to be true/false to the same degree in both categories. After continuous confrontation learning, the generator can simultaneously fit the probability distribution of the infrared image and the visible image, thus producing the result with both significant contrast and rich texture details. Through the cooperation of these two designs, our method can generate fused images with good visual effects.

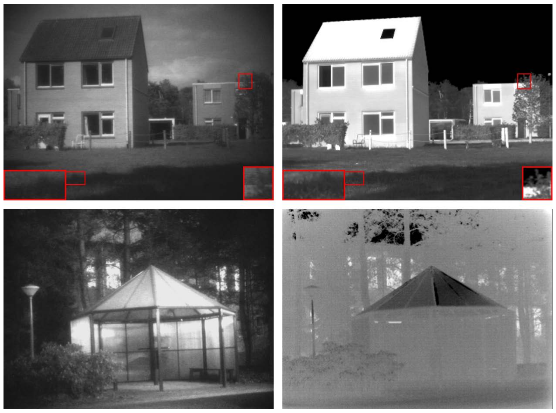


Fig. 2: Illustration of the existence of auxiliary information. The left column is visible images and the right column is infrared images. Clearly, infrared image may contain rich texture details (top row) while visible image may have high contrast (bottom row).

The contributions of our work include the following two aspects. First, we propose a new end-to-end GAN model with multi-classification constraints for infrared and visible image fusion, which can address the challenge of unbalanced fusion in existing methods. Our fused results not only retain the high contrast between thermal targets and background, but also contain rich texture details. Second, we propose a specific content loss for the generator. We construct two kinds of losses between the fused image and the two source images, namely intensity loss and gradient loss, and classify them as main loss and auxiliary loss. They are able to force the generator to get more information from the two source images (complementary to each other), so that the fused image contains more comprehensive and rich information. In addition, due to such a complementary loss function, when the visible image is overexposed, the corresponding information from the infrared image can make up for it, which makes our method be able to remove the highlight while maintaining the significant contrast.

The remainder of this paper is organized as follows. Section II introduces some work and techniques related to the proposed method. In Section III, we describe our GANMcC in detail, including the overall framework, loss functions and network architecture. In Section IV, our method is evaluated comprehensively, including qualitative and quantitative comparisons, complexity evaluation, ablation experiment and generalization verification. Discussion and conclusion are given in Section V.

II. Related Work

In this section, we review some of the work and techniques most relevant to our method, including deep learning-based image fusion methods and generative adversarial network (GAN).

A. Deep Learning-based Image Fusion

Deep learning has promoted tremendous progress in many fusion tasks,which relies on the powerful nonlinear fitting ability of neural network to estimate the expected distribution from the massive data. In contrast, the fusion rules of traditional algorithms typically require manual design, which cannot achieve robustness in various types of fusion tasks. Therefore, in recent years, deep learning-based fusion methods have become research hotspots.

In multi-focus image fusion, convolutional neural network (CNN) is used for the first time in [28] to learn the mapping from source images to decision maps, in which the network is trained to learn the detection of clear or fuzzy areas by artificially designing false labels. After realizing the difficulty of acquiring reference images, an unsupervised network [29] is proposed for generation of decision map, in which the post-processing is still required. The MFF-GAN proposed by Zhang *et al.* [30] realizes multi-focus image fusion with high detail preservation through a well-designed loss function, which can avoid the information loss near the boundary line that appears in the previous decision map-based method. In multi-exposure image fusion, a no-reference quality metric is used as the loss function to train an unsupervised network, which fuses a set of common low-level features extracted from each image to generate promising results [31]. In contrast, Xu *et al.* [32] used GAN to realize this goal, in which the self-attention mechanism is adopted to solve the negativeness caused by large luminance changes. In addition, deep learning has been applied to the remote sensing image fusion. PSGAN [33] uses GAN to fit the probability distribution of the high-resolution multi-spectral image, but it still needs to construct artificial ground truth for training. Analogously, Ma *et al.* [34] adopted GAN with dual discriminators to transform pansharpening into a multi-task learning problem, saying spectrum preservation and spatial preservation. The NDVI-Net designed by Zhang *et al.* [34] can generate the high-precision high-resolution normalized difference vegetation index (NDVI) by fusing low-resolution NDVI with the newly proposed high-resolution vegetation index. In medical image fusion, Liu *et al.* [35] introduced the neural network to implement measurement of activity information, then used the image pyramids to complete the fusion process. Similarly, Yin *et al.* [36] designed an interesting neural network named PA-PCNN to fuse the high-frequency bands of medical images, which performs well in four medical image fusion tasks, including computed tomography (CT) and magnetic resonance (MR), MR-T1 and MR-T2, MR and positron emission tomography, and MR and single-photon emission CT. Great progress also has been made in the field of infrared and visible image fusion. In particular, Liu *et al.* [37] proposed a method based on CNN for infrared and visible image fusion. This method can learn the weight map and solve the problem of activity level measurement and weight distribution. Li *et al.* [17] cleverly used the auto-encoder structure with dense blocks, in which two traditional fusion strategies are adopted to fuse features in the fusion layer. At present, there are also some deep learning-based methods that can uniformly realize the above-mentioned multiple image fusion tasks, and can produce promising results, such as U2Fusion [38] and PMGI [39].

These methods above either need to design artificial false ground truth to train the networks, or just use some metrics to construct content loss function. On the one hand, the ground truth does not exist in image fusion, and the so-called ground truth artificially constructed will set an upper limit for network learning. On the other hand, the loss defined only according to metrics will affect the fusion performance because of the rationality of metrics definition. To address this limitation, Ma *et al.* [16] innovatively introduced GAN into image fusion, which better guides the network to preserve the significant contrast and texture details without supervision through adversarial learning and a specific content loss. Subsequently, they introduced a detail loss and a target edge-enhancement loss based on FusionGAN to further enhance the texture details in the fused results [40]. In addition, they also introduced the dual-discriminator to better extract and reconstruct the information contained in source images [41]. However, these methods did not consider the secondary information contained in source images, nor the information balance in fused images. As a results, the fused images generated by them are more like sharped infrared images. In this work, a new fusion network is proposed, which uses multi-classification constraints and a specifically designed content loss to end-to-end achieve a good balance in maintaining infrared and visible inherent information.

B. GAN

The proposed method is based on adversarial learning of GAN, which realizes the infrared and visible image fusion through the guidance of content loss and the excitation of adversarial loss. Therefore, we introduce some GAN models related to our method, such as the original GAN and least squares GAN (LSGAN).

1) Original GAN: The original GAN was proposed by Goodfellow [42] in 2014, which can realize unsupervised distribution estimation through the mutual game between two modules.

Here we describe the adversarial process of GAN more formally. The two modules involved in the game are called generator G and discriminator D. The generator is dedicated to producing fake data that can fool the discriminator, while the discriminator is intended to distinguish the fake data produced by the generator from real data. Assuming that the training data input to the network are $X = \{x_1, x_2, \cdots, x_n\}$, which obey a specific distribution. The generator G estimates the distribution of X and tries its best to produce fake data $G(X)$ that subject to this specific distribution. Then the discriminator D need learning to distinguish between real training data X and fake data $G(X)$. In summary, the purpose of GAN is to gradually approach the distribution of fake data P_G to the distribution of real data P_{data}, which can be achieved by the following objective function:

$$\min_G \max_D E_{x\sim P_{\text{data}}}[\log D(x)] + E_{x\sim P_G}[\log(1 - D(G(x)))]. \tag{1}$$

As the adversarial relationship, the generator and discriminator promote each other in the continuous iterative training to continuously improve their forgery or discrimination ability. When the distance between these two distributions is small enough, the discriminator cannot distinguish between real data and fake data. Then the generator can be said to have successfully estimated the distribution of training data.

2) LSGAN: Subsequent research finds that the training process of original GAN is very unstable and the quality of generated images is not high. To improve this phenomenon, Mao *et al.* [43] proposed to use the least squares loss function to replace the cross entropy loss function to guide the optimization of GAN. The loss function is defined as follows:

$$\begin{aligned}\min_D V_{\text{LSGAN}}(D) = &\frac{1}{2}\mathbb{E}_{x\sim P_{\text{data}}}[(D(x) - a)^2] \\ &+ \frac{1}{2}\mathbb{E}_{x\sim P_G}[(D(G(x)) - b)^2],\end{aligned} \tag{2}$$

$$\min_G V_{\text{LSGAN}}(G) = \frac{1}{2}\mathbb{E}_{x\sim P_G}[(D(G(x)) - c)^2], \tag{3}$$

in which a and b are probability labels that guide the optimization of the discriminator. Specifically, a is the probability label corresponding to the real data, and b is the probability label corresponding to the fake data produced by the generator. In addition, c is the probability label that guides the optimization of the generator, that is to say, c is the label that the generator expects the discriminator to determine fake data. Clearly, b should be as close to 0 as possible. On the contrary, a and c should be as large as possible, approaching 1.

Compared with the above-mentioned existing methods, the proposed model mainly has two new technical contributions. First, a new and effective content loss function is designed. Different from the state-of-the-art FusionGAN [16], the proposed content loss function uses a concept of main and auxiliary information, which can extract more sufficient intensity and gradient information from source images. Second, we adopt a multi-classifier as the discriminator to simultaneously estimate distributions of two different domains, namely, visible light and infrared. Because the consistency of the probability distribution will make the fused result have the most significant characteristics of the target distribution, the generator can produce the fused result that has characteristics of both infrared and visible light, that is, significant contrast and rich texture details.

III. Method

This section introduces our proposed GANMcC in detail. First, we describe the overall framework of the proposed model. Second, the design of loss function is introduced. Finally, we give the detailed structure of the network.

A. Overall Framework

The image fusion can be described as the extraction and combination of the most meaningful information. Then the key to the problem is how to define the most meaningful information and how to combine them. The aim of infrared and visible image fusion is to produce the result that not only has significant contrast but also contains rich texture details. Therefore, the most meaningful information in infrared and visible image fusion can be defined as the contrast and texture details. For combining this information, a balanced way is needed to ensure that it is both prominent in the fused image. Our method is designed based on these two aspects. The overall framework is demonstrated in Fig. 3, which is an end-to-end model.

First, we observed that the texture structure mainly exists in visible images, while the contrast information mainly exists in infrared images. However, some structure information is also contained in the infrared image (even very rich in some cases). Similarly, the visible image also has a significant contrast, which can distinguish the target from the background. An intuitive example of this phenomenon can be seen in Fig. 2. In response to this phenomenon, we propose the idea of main and auxiliary information. On the one hand, we design a corresponding content loss to sufficiently extract such valuable information, in which the contrast information is represented by the intensity and the texture information is indicated by the gradient. On the other hand, we also design the structure of generator, in which we divide the input into a gradient path and a contrast path. For the gradient path, we concatenate two visible images and one infrared image along the channel dimension as input. Similarly, for the contrast path, we concatenate two infrared images and one visible image along the channel dimension as input. The detailed input and output of the generator can be found in Fig. 4. The input constructed in the form of difference ratio concatenation can drive the network to extract the contrast and gradient information unequally. Under this specific content loss and network design, the generator can obtain the main gradient

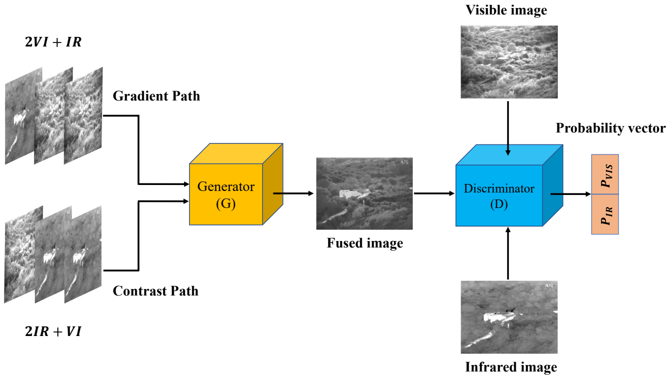


Fig. 3: Overall fusion framework of our GANMcC.

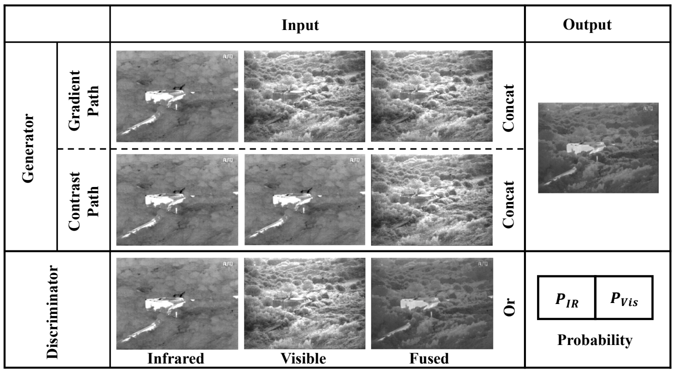


Fig. 4: Input and output of generator and discriminator.

and secondary contrast information from visible images, as well as the main contrast and secondary gradient information from infrared images. The above information can complement each other.

Second, balance of information fusion can be naturally achieved through the game idea of GAN. As the fusion task requires a game between two characteristics such as infrared and visible, we propose to use a multi-classifier in the discriminator. The overall framework of GANMcC is shown in Fig. 3. The output of discriminator is a 1×2 probability vector indicating the probability P_{vis} of the input image being a visible image, and the probability P_{ir} of the input being an infrared image. When the discriminator determines the fused image, the generator wants both probabilities to be large, that is to say, let the discriminator consider that the fusion image is both visible image and infrared image. In contrast, the discriminator is dedicated to precisely determine the fused image as pseudo data, that is, to make both probabilities small at the same time. In this way, an adversarial game is established between the generator and discriminator. When the discriminator determines that P_{vis} and P_{ir} of the fused image are both large, the fused image with balanced information is obtained. In more detail, we provide the input and output of the discriminator in Fig. 4.

Through the above designs, our method can generate good fused results, which not only have significant contrast but also contain rich texture details.

B. Loss Function

Loss $\mathcal{L}_G$ and loss $\mathcal{L}_D$ are used to guide the optimization of the generator and discriminator respectively, which are introduced below.

1) Loss Function of Generator: The loss function for guiding generator optimization consists of two parts, *i.e.*, the content loss $\mathcal{L}_{G_{\text{con}}}$ of the constraint information extraction, and the adversarial loss $\mathcal{L}_{G_{\text{adv}}}$ of the constraint information balance. We formalize it as:

$$\mathcal{L}_G = \gamma L_{G_{\text{con}}} + \mathcal{L}_{G_{\text{adv}}}, \tag{4}$$

where γ is the regularization parameter responsible for maintaining balance between two terms.

Our content loss follows the idea of main and auxiliary information. It depends on what information we want to extract from different types of source images and preserve in the fused image. For infrared image, its main feature is that it has significant contrast reflecting the thermal radiation information of the scene, and can highlight the target from the background. Therefore, the main information is its intensity distribution, and the main intensity loss is defined as:

$$\mathcal{L}_{\text{int}_{\text{main}}} = \|I_{\text{fused}} - I_{\text{ir}}\|_F^2, \tag{5}$$

where I_{fused} is the fused image, which can be formalized as $G(I_{\text{vis}}, I_{\text{ir}})$, I_{ir} is the infrared source image. As for the visible image, it contains rich texture details and conforms to the observation habit of human eyes. Therefore, the main information obtained from visible image is its gradient information, and the main gradient loss is defined as:

$$\mathcal{L}_{\text{grad}_{\text{main}}} = \|\nabla I_{\text{fused}} - \nabla I_{\text{vis}}\|_F^2, \tag{6}$$

where ∇ is the second-order gradient operator, and I_{vis} is the visible image.

As mentioned above, infrared images also have some texture details, and visible images also contain contrast information. Consequently, we propose the concept of auxiliary loss. That is, we construct an auxiliary gradient loss $\mathcal{L}_{\text{grad}_{\text{aux}}}$ between the fused and infrared images, and an auxiliary intensity loss $\mathcal{L}_{\text{int}_{\text{aux}}}$ between the fused and visible images as:

$$\mathcal{L}_{\text{grad}_{\text{aux}}} = \|\nabla I_{\text{fused}} - \nabla I_{\text{ir}}\|_F^2, \tag{7}$$

$$\mathcal{L}_{\text{int}_{\text{aux}}} = \|I_{\text{fused}} - I_{\text{vis}}\|_F^2. \tag{8}$$

In summary, the content loss consists of four parts, namely the main intensity loss, main gradient loss, auxiliary gradient loss and auxiliary intensity loss. It can be formulated as:

$$\begin{aligned}\mathcal{L}_{G_{\text{con}}} &= \mathcal{L}_{\text{int}_{\text{main}}} + \mathcal{L}_{\text{grad}_{\text{main}}} + \mathcal{L}_{\text{grad}_{\text{aux}}} + \mathcal{L}_{\text{int}_{\text{aux}}} \\ &= \beta_1\|I_{\text{fused}} - I_{\text{ir}}\|_F^2 + \beta_2\|\nabla I_{\text{fused}} - \nabla I_{\text{vis}}\|_F^2 \\ &\quad + \beta_3\|\nabla I_{\text{fused}} - \nabla I_{\text{ir}}\|_F^2 + \beta_4\|I_{\text{fused}} - I_{\text{vis}}\|_F^2,\end{aligned} \tag{9}$$

where $\beta_{(\cdot)}$ is a constant, which should be adjusted to realize primary and secondary relationships among these items. Besides, the gradient loss term is generally smaller than the intensity loss term, so $\beta_{(\cdot)}$ needs to be adjusted to make them equally important in the optimization process. Therefore, the setting rules of $\beta_{(\cdot)}$ can be summarized as:

$$\beta_1 > \beta_4, \;\; \beta_2 > \beta_3, \;\; \{\beta_2, \beta_3\} > \{\beta_1, \beta_4\}. \tag{10}$$

In order to achieve a balance between various kinds of information, we introduce the adversarial loss with discriminator into the loss function of generator, which can be defined as:

$$\mathcal{L}_{G_{\text{adv}}} = \big(D(I^n_{\text{fused}})[1] - d)^2 + (D(I^n_{\text{fused}})[2] - d)^2\big), \quad (11)$$

in which d is the probability label of the discriminator to determine the fused image. In our work, the discriminator is a multi-classifier that outputs a 1×2 probability vector. Therefore, $D(\cdot)[1]$ represents the first term of the vector, that is, the probability of the fused image being a visible image. Similarly, $D(\cdot)[2]$ represents the second term of the vector, that is, the probability of the fused image being an infrared image. It is worth noting that we use the same label d for both probabilities, and hence the discriminator has the same probability of determining that the fused image is an infrared image or a visible image. Here, because the generator expects that the discriminator cannot distinguish between the fused image and the real data, d is set to 1.

2) Loss Function of Discriminator: The discriminator is a multi-classifier whose loss function must constantly improve its discriminating ability, and can effectively identify what is an infrared image or a visible image. The discriminator's loss function $\mathcal{L}_D$ is composed of three parts, *i.e.*, the decision losses of visible image, infrared image and fused image. We denote these three losses as $\mathcal{L}_{D_{\text{vis}}}$, $\mathcal{L}_{D_{\text{ir}}}$, and $\mathcal{L}_{D_{\text{fused}}}$. That is:

$$\mathcal{L}_D = \mathcal{L}_{D_{\text{vis}}} + \mathcal{L}_{D_{\text{ir}}} + \mathcal{L}_{D_{\text{fused}}}. \quad (12)$$

Considering the 1×2 vectors output by the discriminator, we have $P_{\text{vis}} = D(x)[1]$ and $P_{\text{ir}} = D(x)[2]$. When the input is a visible image, it is expected that P_{vis} should be close to 1 and P_{ir} close to 0. The corresponding loss is defined as:

$$\mathcal{L}_{D_{\text{vis}}} = \frac{1}{2N}\sum_{i=1}^{N}((P_{\text{vis}}(I^n_{\text{vis}}) - a_1)^2 + (P_{\text{ir}}(I^n_{\text{vis}}) - a_2)^2), \quad (13)$$

where a_1 and a_2 are probability labels, a_1 is set as 1, and a_2 is set as 0. That is to say, when the visible image is input, the probability that the discriminator wants to judge that it is a visible image is large, and the probability of the infrared image is small.

Similarly, infrared loss term is defined as:

$$\mathcal{L}_{D_{\text{ir}}} = \frac{1}{2N}\sum_{i=1}^{N}((P_{\text{vis}}(I^n_{\text{ir}}) - b_1)^2 + (P_{\text{ir}}(I^n_{\text{ir}}) - b_2)^2), \quad (14)$$

where b_1 is set as 0, and b_2 is set as 1.

Finally, when the input image is a fused image, the loss function is formulated as:

$$\mathcal{L}_{D_{\text{fused}}} = \frac{1}{2N}\sum_{i=1}^{N}((P_{\text{vis}}(I^n_{\text{fused}}) - c)^2 + (P_{\text{ir}}(I^n_{\text{fused}}) - c)^2), \quad (15)$$

where c is the probability label of the discriminator to determine the fused image, which should be set as 0.

Again, we also use the same label c for both probabilities to achieve a balance. That is to say, in the view of the discriminator, the fused image is a pseudo-visible image and a pseudo-infrared image to the same degree.

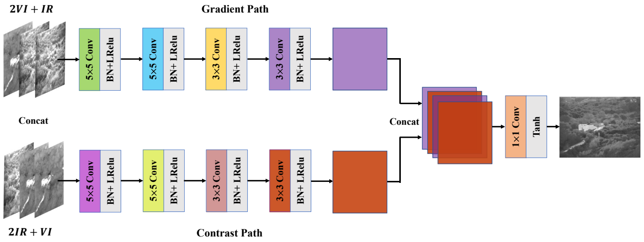


Fig. 5: Network architecture of the generator.

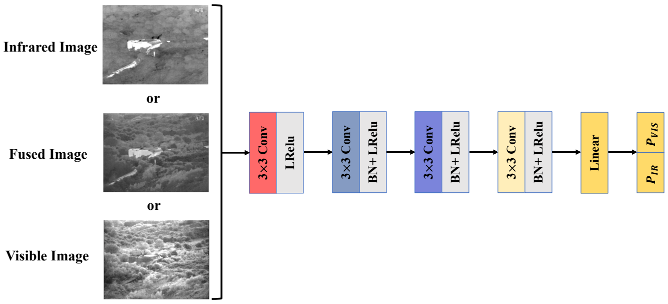


Fig. 6: Network architecture of the discriminator.

C. Network Architecture

1) Generator Architecture: As mentioned in Sec. III-A, we divide the generator into gradient path and contrast path for information extraction. Its structure is shown in Fig. 5.

For the gradient path, we hope that it is responsible for extracting texture information, that is, high-frequency features. In our view, texture information is mainly contained in visible images, and secondly in infrared images. Therefore, the primary and secondary concatenating strategy is used to construct the input. We use two visible images and one infrared image to concat along the channel as input. Similarly, for the contrast path, we expect it to be responsible for extraction of contrast information, which is mainly contained in infrared images, and secondary in visible images. So we use two infrared images and one visible image to concat along the channel.

In each path of information extraction, four convolutional layers are adopted for feature extraction. The 5×5 convolution kernel is used in the first two layers, and the 3×3 is adopted in the latter two layers, all with batch normalization and leaky ReLU activation function. Then, we fuse the features extracted from the two paths, and use the strategies of concating and convolution to achieve this purpose. In order to fully merge the information, we cross-concate the two feature maps along the channel. In the last layer, we use the kernel with size of 1×1, and tanh activation function. It is worth noting that we set the stride of all layers to 1, so the sizes of all feature maps do not changed.

2) Discriminator Architecture: The structure of the discriminator is demonstrated in Fig. 6. Our discriminator is essentially a multi-classifier, which can estimate the probability of each category of the input image. Its output is a probability vector of size 1×2. The proposed discriminator is composed of four convolution layers and one linear layer. The four convolutional layers use 3×3 convolution kernels and leaky

ReLU activation functions, and the latter three also use the batch normalization. We set the stride as 2 in all convolution layers. The last linear layer discriminates the input based on the features extracted by the first four convolutional layers, which outputs the classification probability.

IV. Experiments

In this section, we evaluate our proposed GANMcC on publicly available datasets. Seven popular methods are selected to compare with our method, including LPP [44], LP [45], CVT [46], DTCWT [47], GTF [24], CNN [37], and FusionGAN [16]. First, we provide the detailed experimental configuration. Second, we compare the proposed method with other popular methods qualitatively and quantitatively. In addition, we provide additional the ablation experiment and generalization experiment. Finally, we conduct experiments when visible images are overexposed.

A. Experimental Configurations

1) Datasets: We select the TNO and RoadScene datasets to evaluate our GANMcC and other comparative methods.

The image pairs in the TNO dataset mainly describe various military related scenes. In the TNO dataset, there are 60 infrared and visible image pairs, and 3 sequences involving 19, 32 and 23 image pairs, respectively.

The RoadScene dataset is a new image fusion dataset established in [38], in which 221 pairs are accurately aligned. The major scenes in the dataset is the road, including vehicles, pedestrians, traffic signs and other targets. It is worth noting that some images in this dataset are taken over exposure, and hence brings a new challenge for image fusion.

For testing, we use 16 and 30 image pairs on the TNO and RoadScene datasets, respectively. For training, we adopt overlapping cropping strategies to expand the dataset. Concretely, we crop the remaining images in the TNO and RoadScene datasets to $35,845$ and $69,029$ image patch pairs to train our network, respectively. In particular, we crop each image into multiple 120×120 image patches, and then fill it up to 132×132. Note that the input of the generator is 132×132 image patches, and the size of output is reduced to 120×120 after a series of operations. The 120×120 visible and infrared image patches, that are originally cropped and used as labels, are put into the discriminator along with the generated images.

2) Training Details: The generator and discriminator are trained iteratively, in which the ratio of training number is p. The batchsize is b, it takes m steps to traverse all the training data once, and the total number of training epoch is M. In practice, we empirically set $b = 32$, $p = 1/2$, $M = 10$, and m is set as the ratio between the whole number of patches and b. The initial learning rate is set as 0.0001, and we adopt the Adam as the optimizer to train the network. We summarize the entire training process in Algorithm 1. In addition, $\beta_{(\cdot)}$ of Eq. (9) in this work are set as follows according to the rules in Eq. (10): $\beta_1 = 1$, $\beta_2 = 5$, $\beta_3 = 4$ and $\beta_4 = 0.3$. The γ in Eq. (4) is adjusted until the generator and discriminator can form an effective confrontation. In our work, γ is set as 100. In order to make the training of GAN more stable, we adopt the soft label strategy. More specifically, we relax the labels a_1, b_2 and d that should be set to 1 to random numbers ranging from 0.7 to 1.2. In contrast, the labels a_2, b_1 and c that should be set to 0 are set to random numbers ranging from 0 to 0.3. All experimental work is carried out using GPU NVIDIA-RTX 2080Ti and CPU Intel i7-8750H.

Algorithm 1 Training procedure of GANMcC.

```
for $M$ epochs do
    for $m$ steps do
        for $p$ times do
            Select $b$ visible patches $\{I_{\text{vis}}^1, I_{\text{vis}}^2 \cdots I_{\text{vis}}^b\}$;
            Select $b$ infrared patches $\{I_{\text{ir}}^1, I_{\text{ir}}^2 \cdots I_{\text{ir}}^b\}$;
            Select $b$ fused patches $\{I_{\text{fused}}^1, I_{\text{fused}}^2 \cdots I_{\text{fused}}^b\}$;
            Update the parameters of the discriminator by AdamOptimizer: $\nabla_D(\mathcal{L}_D)$;
        end for
        Select $b$ visible patches $\{I_{\text{vis}}^1, I_{\text{vis}}^2 \cdots I_{\text{vis}}^b\}$;
        Select $b$ infrared patches $\{I_{\text{ir}}^1, I_{\text{ir}}^2 \cdots I_{\text{ir}}^b\}$;
        Update the parameters of the generator by AdamOptimizer: $\nabla_G(\mathcal{L}_G)$;
    end for
end for
```

3) Metrics: The quality evaluation of image fusion is a complex problem, so we not only carry out the qualitative assessment but also the quantitative evaluation.

Qualitative assessment starts from the human visual perception and judges the quality of results according to the goal of the task. For infrared and visible image fusion, the goal is to preserve significant contrast and rich texture at the same time. Conversely, quantitative evaluation relies on some existing statistical metrics to evaluate the quality of fused results from different aspects. In this work, six metrics are used to achieve this goal, which are structural similarity index measure (SSIM) [48], correlation coefficient (CC) [49], sum of the correlations of differences (SCD) [50], entropy (EN) [51], standard deviation (SD) [52], and mutual information (MI) [53].

SSIM can measure the structural similarity between the fused image and the source images. We calculate the sum of SSIM between the fused image and the two types of source images as the final result. The larger the value of SSIM, the better the structure is maintained. CC measures the degree of linear correlation between the fused image and the source images, and the larger the value is, the more relevant the fused image is to the source images. SCD metric focuses on the difference between the fused image and the source image, and then measures the correlation between the difference and another source image. To a certain extent, SCD can also assess the pseudo-information, and the larger SCD value, the less pseudo information. EN is the most commonly used metric to represent the amount of information, and a large EN value indicates that the fused result contains a large amount of information. SD reflects the distribution of pixel values in the image. In general, a large SD value indicates high contrast and good visual perception. MI is a metric to assess the amount of

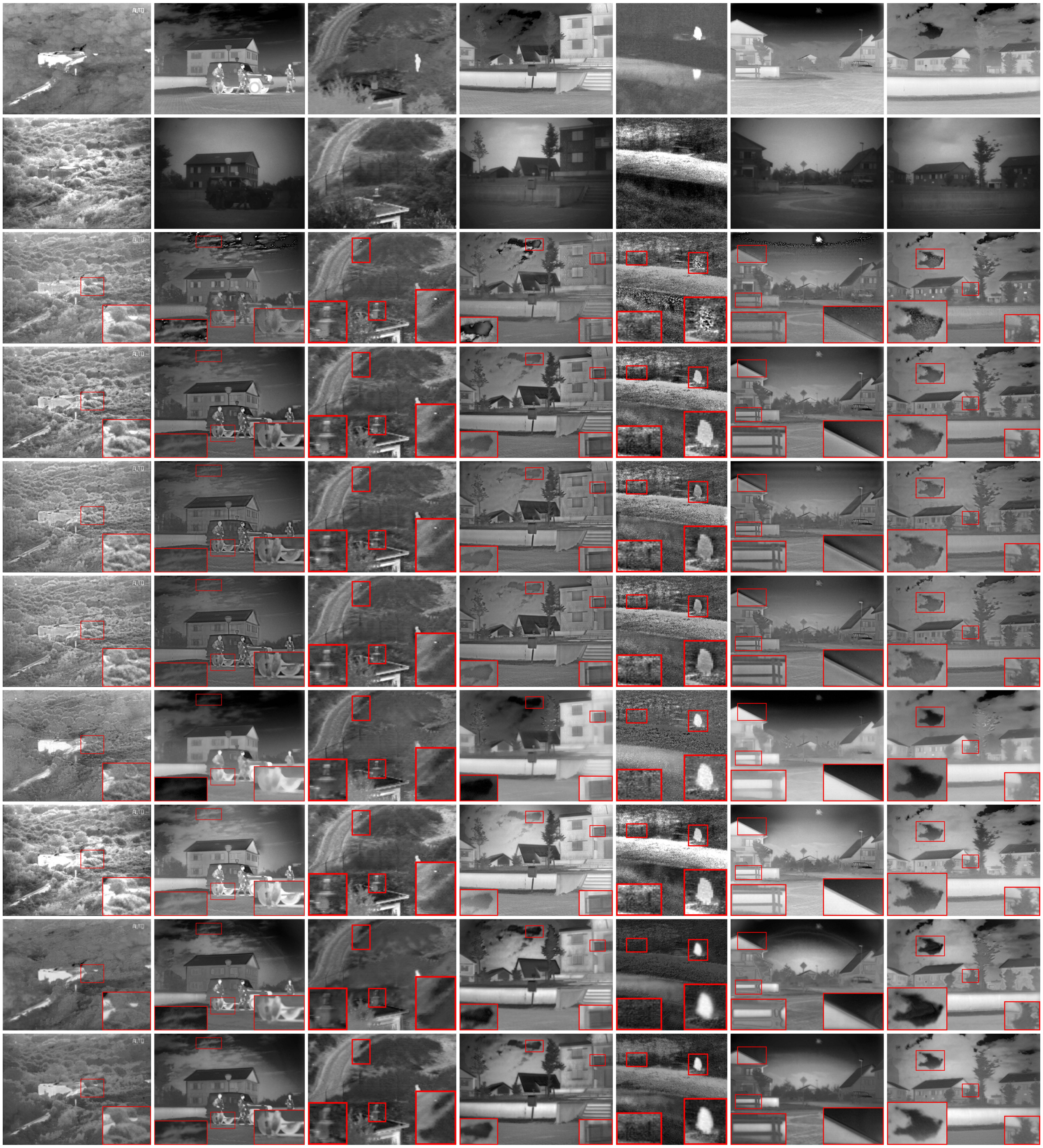


Fig. 7: Qualitative results on the TNO dataset. From top to bottom: infrared images, visible images, fused results of LPP [44], LP [45], CVT [46], DTCWT [47], GTF [24], CNN [37], FusionGAN [16], and our GANMcC.

information transmitted from the source images to the fused image. The larger the MI, the more information the fused image acquires from source images.

B. Results on The TNO Dataset

1) Qualitative Comparison: We provide seven typical qualitative results to demonstrate the characteristics of our GANMcC in Fig. 7. From the perspective of visual effect, our GANMcC has obvious advantages over the comparative methods. First, the proposed method maintains the major thermal radiation information of the infrared image, which can effectively distinguish the target from the background. This is important since most existing algorithms only have good texture details, but lose most of the thermal radiation information, which is harmful for target detection. In addition, while maintaining sufficient thermal radiation information, the results of our

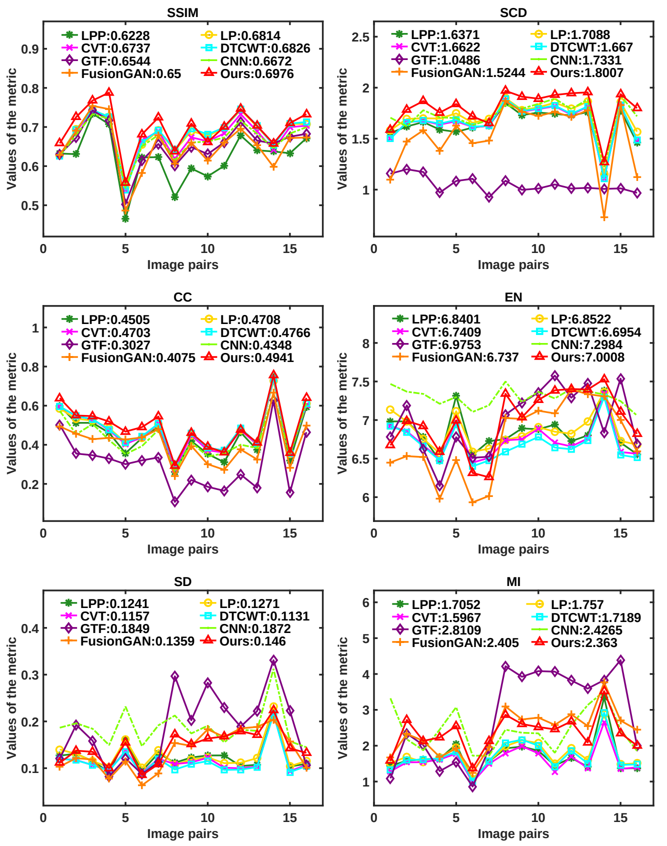


Fig. 8: Quantitative results of different methods on the TNO dataset. We select six metrics including SSIM, CC, SCD, EN, SD and MI.

method still have a lot of texture details.

According to the characteristics of the fused results, the comparative methods can be divided into two categories. The first is more inclined to visible image, such as LPP, LP, CVT and DTCWT. Their results have rich texture details, but the contrast between target and background is not that significant. On the contrary, the second category has good thermal radiation information, but the texture details are insufficient, such as GTF and FusionGAN. In addition, the fused results of CNN are more like an improvement on the first category of methods, which enhance the contrast while maintaining texture details. However, this contrast enhancement is limited. By contrast, our approach is more like an extension of GTF and FusionGAN. The fused results not only have significant contrast, but also rich texture details. For example, in the first column of Fig. 7, in our result the contrast between the building (*i.e.*, target) and the trees (*i.e.*, background) is strong, while the texture details of the trees are also rich.

2) Quantitative Comparison: Further quantitative comparison is performed on 16 image pairs of the TNO dataset, which is shown in Fig. 8. It can be seen that our GANMcC achieves the best values on SSIM, SCD and CC. In addition, our GANMcC achieves the second largest average on EN. As for SD, our method ranks second to GTF and CNN in average value, and follows behind GTF, CNN and FusionGAN on MI. From these results, we can conclude that our GANMcC is able to maintain the best structural information, has strong correlation with the source images, and contains minimal pseudo information. The results of our method also contain a large amount of information, second only to CNN, but in combination with SCD, CNN contains more fake information than our method. In addition, our method also maintains a significant contrast, which is only worse than GTF and CNN. However, the contrast of our results is more reasonable, which is more similar to the infrared image. Last but not least, although GTF, CNN and FusionGAN have larger metric values of MI, our method achieves a better balance between the information of two source images. Overall, our GANMcC performs best among all methods in objective evaluation.

C. Results on The RoadScene Dataset

1) Qualitative Comparison: We also conduct a qualitative comparison on the RoadScene dataset, and provide 6 typical results in Fig. 9. From subjective perception, only GTF, FusionGAN and our GANMcC have relatively significant contrast information. For example, the intensity distribution of the sky, sea water and trees is much closer to the infrared image. Nevertheless, compared with GTF and FusionGAN, our method has the advantage of retaining the light information that appears in visible images, such as the traffic signal light, the street lamp, and so on. In addition, our fused results also contain richer texture details. The fused results of the other five methods have good texture details but weak contrast. In general, infrared thermal radiation information is more important in some applications of road scenes, such as assisted driving at night and pedestrian and vehicle detection in automatic driving.

2) Quantitative Comparison: We further carry out quantitative experiments on 30 image pairs in the RoadScene dataset, and the results are shown in Fig. 10. Our GANMcC still achieves the best performance on SSIM, SCD and CC. As for EN and SD, our method ranks the second. For the metric MI, our method follows behind GTF and FusionGAN.

D. Complexity Evaluation

A complexity evaluation is performed to assess the cost of our GANMcC. We count the number of network parameters, which can describe the space complexity of the proposed method to a certain extent. During the training phase, the generator and discriminator are iteratively trained. At this time, the total number of parameters is the sum of parameters in the generator and discriminator, which is about 2.276 M. These parameters need about 125 minutes to be optimized. During the testing phase, only the generator is reserved to produce fused images, and the number of parameters used for testing is 1.867 M. Because the generator is the target network for image fusion, we compare the average running time of different methods in the test phase, which can indicate the time complexity of methods. The results are reported in Table I. Obviously, our GANMcC is competitive in running efficiency.

E. Ablation Experiment

In this work, we adopt two types of loss functions to guide the optimization of the network, saying content loss and

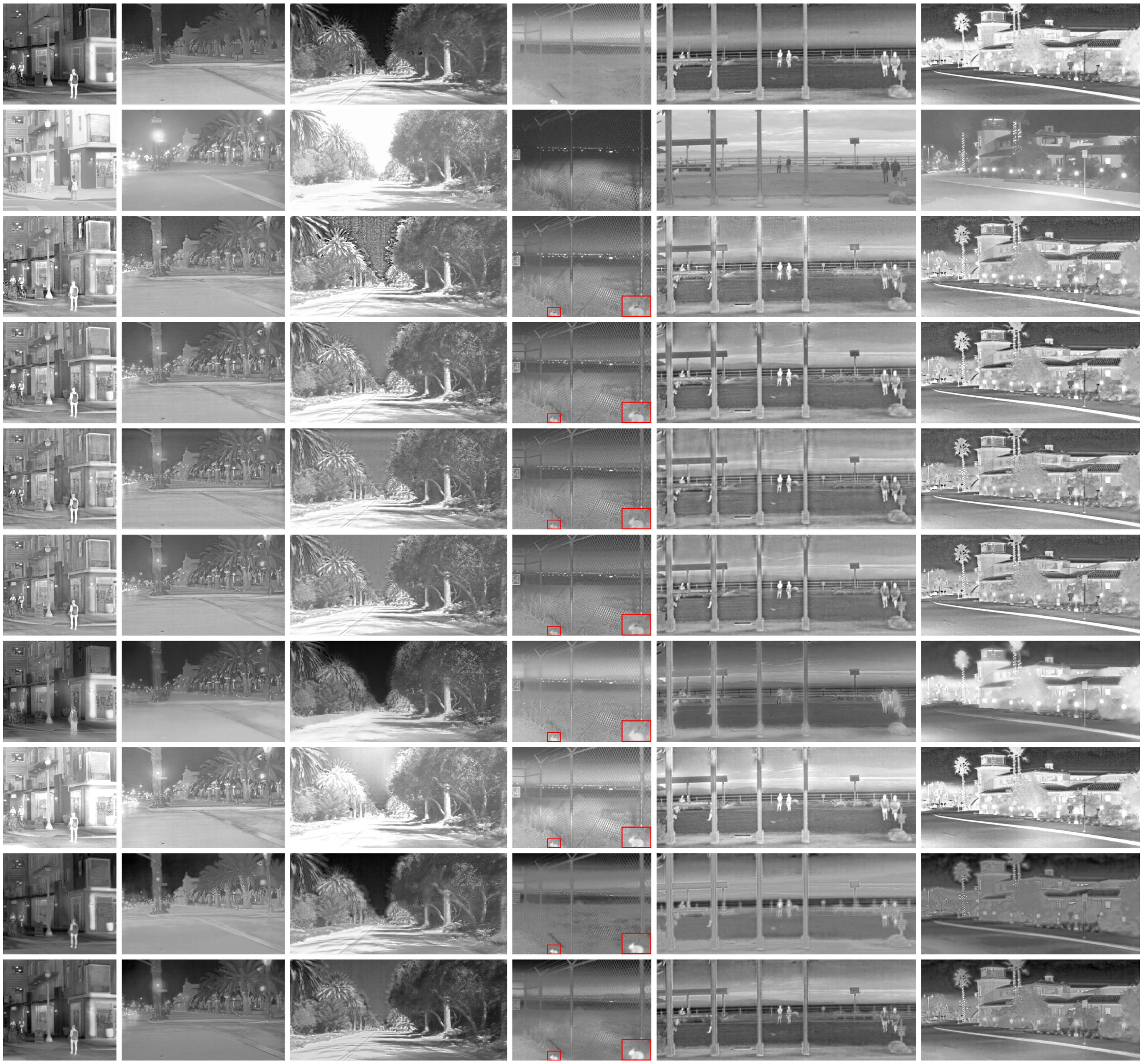


Fig. 9: Qualitative results on the RoadScene dataset. From top to bottom: infrared images, visible images, fused results of LPP [44], LP [45], CVT [46], DTCWT [47], GTF [24], CNN [37], FusionGAN [16], and our GANMcC.

adversarial loss. The content loss determines the type of information to be extracted, while the adversarial loss controls the balance of information in the fused result. More specifically, content loss divides the information extracted from the source images into intensity and gradient information, and the former corresponds to the contrast while the latter reflects the texture structure. Not only that, the main and auxiliary ideas in content loss make the information of source images complementary, which makes the extracted information more sufficient and reasonable. The role of adversarial loss is to simultaneously fit the two distributions in the form of the game, so that the extracted information can be fused in a more balanced manner. To verify the validity of the proposed loss functions, the related ablation experiments are conducted, including the content loss item and the adversarial loss item. Concretely, we first remove the discriminator and only use the content loss to guide the optimization of the generator. Then we remove the content loss and train the network only through the adversarial loss. Finally, we combine the adversarial loss and the content loss to jointly constrain the optimization of the network.

We show the results of ablation experiments in Fig. 11. It can be seen that when there is only the content loss, the texture details of the generated fused image are not rich enough. In particular, there is no light-dark transition in the fused result, so the texture structure is not vivid. More intuitively, it is undeniable that both contrast and texture information exists, but it is not balanced, and the texture information is clearly at a disadvantage. When only the adversarial loss exists, the fused result shows the game between two distributions of infrared and visible images. In other words, although the fused result has a certain of contrast and texture details, they do not match source images and are far from what we would expect. The

TABLE I: The mean and standard deviation of running time of different methods on two datasets (unit: second).

Dataset	LPP [44]	LP [45]	CVT [46]	DTCWT [47]	GTF [24]	CNN [37]	FusionGAN [16]	Ours
TNO	0.096±0.042	0.009±0.004	1.224±0.443	0.274±0.130	4.369±2.309	68.788±27.538	0.159±0.173	0.274±0.331
RoadScene	0.055±0.019	0.005±0.002	0.746±0.095	0.157±0.056	2.419±1.027	26.888±3.787	0.361±0.329	0.085±0.024

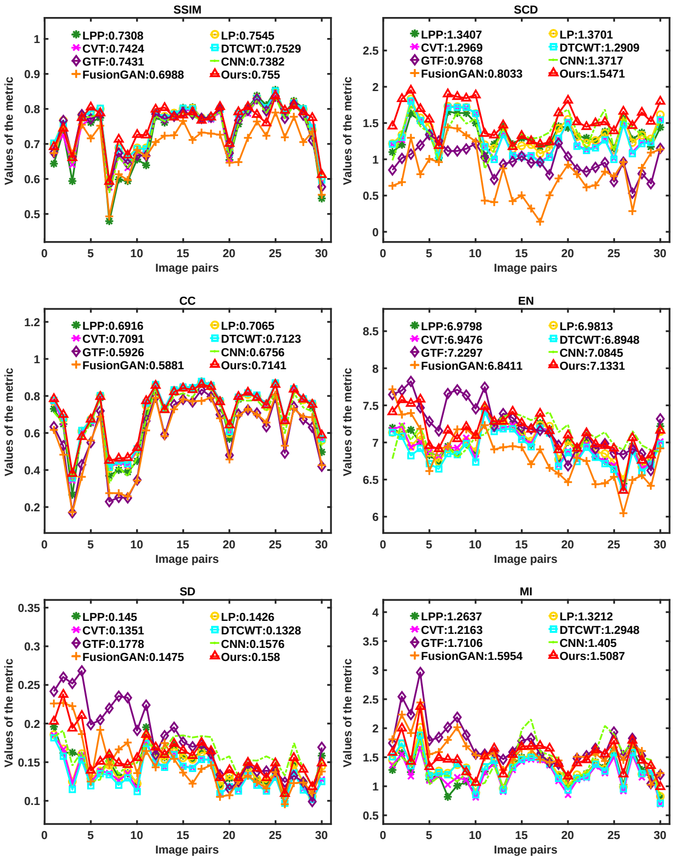
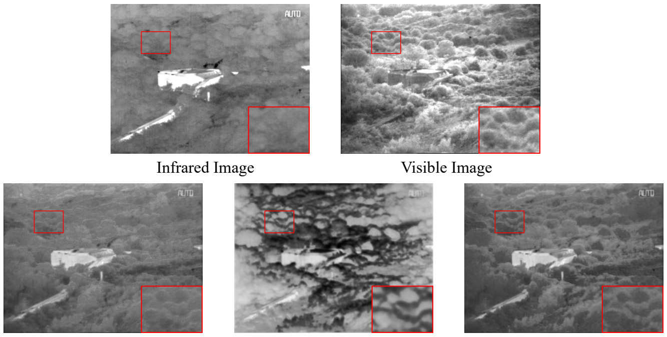


Fig. 10: Quantitative results on the RoadScene dataset. We select six metrics including SSIM, CC, SCD, EN, SD and MI.



Infrared Image

Visible Image

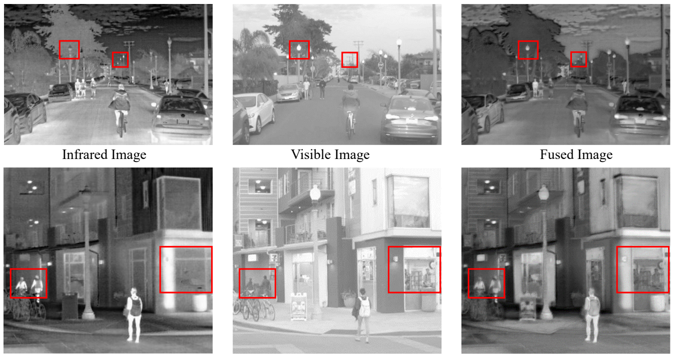
Result with only content loss

Result with only adversarial loss

Result with adversarial loss and content loss

Fig. 11: Ablation experiment.

reason for this phenomenon is that in an unsupervised GAN, the generator cannot extract the desired information without the guidance of content loss. When combining the content loss and adversarial loss, the generated result not only has obvious contrast and rich texture details, but also fits the characteristics of source image very well. For instance, the change of light-dark between textures is very similar to visible light, so it has a better visual experience. Therefore, it can be concluded that the designed content loss and the adversarial loss are complementary, and they work together to enable the generator to produce the desired fused result.



Infrared Image

Visible Image

Fused Image

Fig. 12: Visualization of generalization experiment.

F. Generalization Verification

In order to verify the generalization performance of our GANMcC, we implement the generalization experiment of the network. Specifically, we train the GANMcC on the TNO dataset and then test it on the RoadScene dataset. Fig. 12 shows the performance of the transferred model on the RoadScene dataset. From the results, we can see that our GANMcC has good generalization, and the transferred model still performs well on the RoadScene dataset. On the one hand, the fused results can maintain the significant contrast similar to the infrared image, such as pedestrians and street lights. On the other hand, the fused images contain rich texture details, which make the results have good visual experience.

G. Fused Results on Overexposed Images

Sometimes, the visible image is overexposed, and some details are blurred or even invisible, which is a new challenge for infrared and visible image fusion. This is more like a mixture of two fusion tasks, namely multi-exposure image fusion and infrared and visible image fusion. Among the comparative methods, only GTF, FusionGAN and our GANMcC can better maintain the thermal radiation information. These methods obtain gradient information from visible images to enhance texture details. However, if the visible image is overexposed, its local gradient changes greatly, which is a great challenge to achieve satisfactory results. Next, we compare the performance of these three methods when the visible image is overexposed.

We select six typical image pairs for evaluation, and their visible images all have a certain degree of overexposure. The fused results are shown in Fig. 13. First of all, the fused

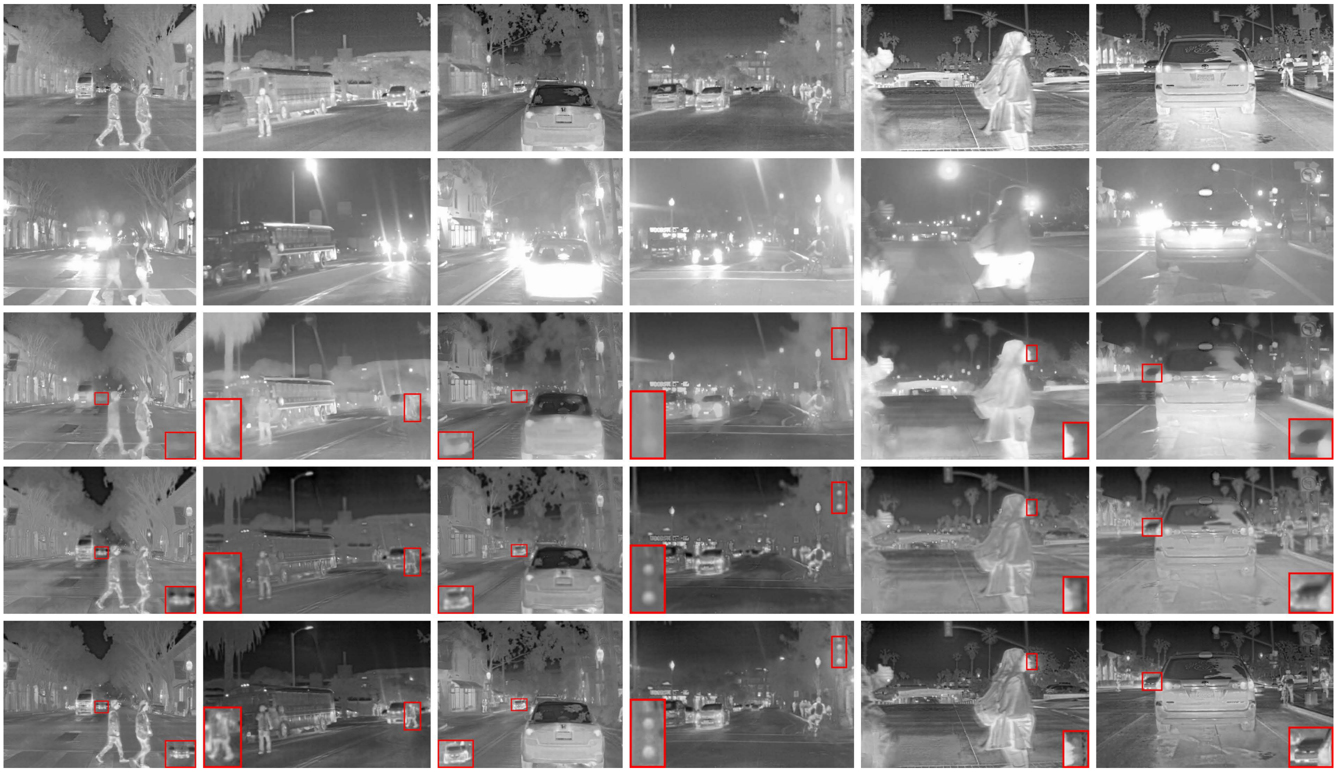


Fig. 13: Qualitative results on overexposed images. The first two rows are the infrared and visible images to be fused, and the rest three rows are the results of GTF, FusionGAN and our GANMcC.

results of three methods all have significant contrast, which is the prerequisite of comparison. In addition, our method demonstrates better fusion effects in the overexposed regions, such as the rear car, traffic lights and pedestrians. In contrast, GTF and FusionGAN perform poorly in these regions. In particular, the results of FusionGAN in these regions are blurred, while the results of GTF are not even visible.

The invisible phenomenon of the fused results of GTF is because the objective function constructs the gradient consistency between the fused image and the visible image, where the local gradient value of the overexposed in the visible image is large, which affects the normal fusion. Compared with GTF, FusionGAN uses the adversarial learning of GAN, so that the effect of overexposure can be reduced to some extent. However, the content loss of FusionGAN determines that it cannot further eliminate the effect of overexposure. Our method uses the main and auxiliary ideas to define a new content loss, and uses multiple classifiers as the discriminator to deal with the fusion challenge caused by overexposure.

H. Discussion of Limitation

The limitation of our method is that it is greatly affected by the shadows in some scenarios, which leads to unnatural shadow transitions of the fused result. We provide a typical example to illustrate this issue intuitively, as shown in Fig. 14. On closer inspection, there is a certain correspondence between these unnatural shadow transitions and the insignificant brightness changes in the visible image. When the discriminator determines that it is an important feature of the visible light image distribution, it will force the generator to strengthen such features, which leads to this phenomenon. A possible solution is to use PatchGAN [54] instead of LSGAN to estimate the visible image distribution from the probabilities of multiple patches, reducing the influence of certain specific areas on the discrimination process.

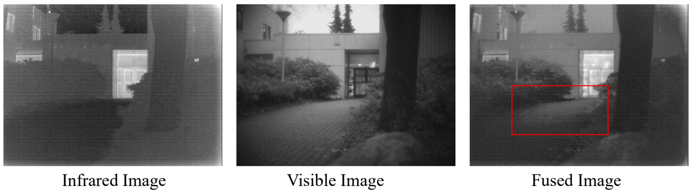


Fig. 14: A failure case. Unnatural shadow transitions appear in the fused image.

V. Discussion and Conclusion

In this paper, we propose a new end-to-end infrared and visible image fusion network, called GANMcC. Based on the requirement that image fusion should not only extract meaningful information, but also achieve a balance among various information, we design a generator with two paths and a discriminator that can realize multi-classification. In addition, we also design a new content loss function, involving the concepts of major loss and auxiliary loss, and use the same label to balance constraints in the judgment of fused images. Both qualitative and quantitative experiments show that our GANMcC has advantages over the state-of-the-art methods. In addition, our method can achieve good fused results when the visible image is overexposed.

In the future, we will focus on the research of variational auto-encoders (VAE) technology for infrared and visible image fusion. VAE can sample different values from the distribution of hidden layers, thus reconstructing diverse fused results. In other words, the distribution of various attributes can be obtained through hidden layer encoding, and then the fused images with different contrast and texture richness can be reconstructed based on the different values sampled. Further, we will apply infrared and visible image fusion to a wider range of tasks, such as object detection, scene understanding and so on.

References

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

[2] S. Das and Y. Zhang, "Color night vision for navigation and surveillance," *Transportation Research Record*, vol. 1708, no. 1, pp. 40–46, 2000.

[3] H. Li, X.-J. Wu, and T. Durrani, "Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Transactions on Instrumentation and Measurement*, 2020.

[4] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Transactions on Instrumentation and Measurement*, 2020.

[5] Y. Yang, Y. Que, S. Huang, and P. Lin, "Multiple visual features measurement with gradient domain guided filtering for multisensor image fusion," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 4, pp. 691–703, 2017.

[6] H. Wang, S. Li, L. Song, L. Cui, and P. Wang, "An enhanced intelligent diagnosis method based on multi-sensor image fusion via improved deep learning network," *IEEE Transactions on Instrumentation and measurement*, vol. 69, no. 6, pp. 2648–2657, 2019.

[7] S. Li, H. Yin, and L. Fang, "Group-sparse representation with dictionary learning for medical image denoising and fusion," *IEEE Transactions on Biomedical Engineering*, vol. 59, no. 12, pp. 3450–3459, 2012.

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

[9] S. Li, B. Yang, and J. Hu, "Performance comparison of different multi-resolution transforms for image fusion," *Information Fusion*, vol. 12, no. 2, pp. 74–84, 2011.

[10] G. Pajares and J. M. De La Cruz, "A wavelet-based image fusion tutorial," *Pattern Recognition*, vol. 37, no. 9, pp. 1855–1872, 2004.

[11] W. Kong, Y. Lei, and H. Zhao, "Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization," *Infrared Physics & Technology*, vol. 67, pp. 161–172, 2014.

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

[13] J. Zhao, Y. Chen, H. Feng, Z. Xu, and Q. Li, "Infrared image enhancement through saliency feature analysis based on multi-scale decomposition," *Infrared Physics & Technology*, vol. 62, pp. 86–93, 2014.

[14] J. Ma, Z. Zhou, B. Wang, and H. Zong, "Infrared and visible image fusion based on visual saliency map and weighted least square optimization," *Infrared Physics & Technology*, vol. 82, pp. 8–17, 2017.

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

[16] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "Fusiongan: A generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[17] H. Li and X.-J. Wu, "Densefuse: A fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[18] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Processing Letters*, vol. 23, no. 12, pp. 1882–1886, 2016.

[19] T. Xiang, L. Yan, and R. Gao, "A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain," *Infrared Physics & Technology*, vol. 69, pp. 53–61, 2015.

[20] W. Kong, L. Zhang, and Y. Lei, "Novel fusion method for visible light and infrared images based on nsst–sf–pcnn," *Infrared Physics & Technology*, vol. 65, pp. 103–112, 2014.

[21] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

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

[23] X. Kong, L. Liu, Y. Qian, and Y. Wang, "Infrared and visible image fusion using structure-transferring fusion method," *Infrared Physics & Technology*, 2019.

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

[25] X. Yuan, J. Zhou, B. Huang, Y. Wang, C. Yang, and W. Gui, "Hierarchical quality-relevant feature representation for soft sensor modeling: a novel deep learning strategy," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 6, pp. 3721–3730, 2019.

[26] Y. Wang, Z. Pan, X. Yuan, C. Yang, and W. Gui, "A novel deep learning based fault diagnosis approach for chemical process with extended deep belief network," *ISA transactions*, vol. 96, pp. 457–467, 2020.

[27] X. Yuan, L. Li, Y. Shardt, Y. Wang, and C. Yang, "Deep learning with spatiotemporal attention-based lstm for industrial soft sensor model development," *IEEE Transactions on Industrial Electronics*, 2020.

[28] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[29] B. Ma, X. Ban, H. Huang, and Y. Zhu, "Sesf-fuse: An unsupervised deep model for multi-focus image fusion," *arXiv preprint arXiv:1908.01703*, 2019.

[30] H. Zhang, Z. Le, Z. Shao, H. Xu, and J. Ma, "Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion," *Information Fusion*, 2020.

[31] 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*, 2017, pp. 4714–4722.

[32] H. Xu, J. Ma, and X.-P. Zhang, "Mef-gan: Multi-exposure image fusion via generative adversarial networks," *IEEE Transactions on Image Processing*, vol. 29, pp. 7203–7216, 2020.

[33] X. Liu, Y. Wang, and Q. Liu, "Psgan: a generative adversarial network for remote sensing image pan-sharpening," in *Proceedings of the IEEE International Conference on Image Processing*, 2018, pp. 873–877.

[34] J. Ma, W. Yu, C. Chen, P. Liang, X. Guo, and J. Jiang, "Pan-gan: An unsupervised pan-sharpening method for remote sensing image fusion," *Information Fusion*, vol. 62, pp. 110–120, 2020.

[35] Y. Liu, X. Chen, J. Cheng, and H. Peng, "A medical image fusion method based on convolutional neural networks," in *Proceedings of the International Conference on Information Fusion*, 2017, pp. 1–7.

[36] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 1, pp. 49–64, 2018.

[37] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 16, no. 03, p. 1850018, 2018.

[38] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2fusion: A unified unsupervised image fusion network," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, to be published, doi: 10.1109/TPAMI.2020.3012548.

[39] H. Zhang, H. Xu, Y. Xiao, X. Guo, and J. Ma, "Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity." in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, pp. 12 797–12 804.

[40] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang, "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85–98, 2020.

[41] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4980–4995, 2020.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIM.2020.3038013, IEEE Transactions on Instrumentation and Measurement

[42] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Advances in Neural Information Processing Systems*, 2014, pp. 2672–2680.

[43] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, “Least squares generative adversarial networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 2794–2802.

[44] A. Toet, “Image fusion by a ratio of low-pass pyramid,” *Pattern Recognition Letters*, vol. 9, no. 4, pp. 245–253, 1989.

[45] P. Burt and E. Adelson, “The laplacian pyramid as a compact image code,” *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.

[46] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, “Remote sensing image fusion using the curvelet transform,” *Information Fusion*, vol. 8, no. 2, pp. 143–156, 2007.

[47] J. J. Lewis, R. J. O’Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, “Pixel-and region-based image fusion with complex wavelets,” *Information Fusion*, vol. 8, no. 2, pp. 119–130, 2007.

[48] Z. Wang and A. C. Bovik, “A universal image quality index,” *IEEE Signal Processing Letters*, vol. 9, no. 3, pp. 81–84, 2002.

[49] M. Deshmukh and U. Bhosale, “Image fusion and image quality assessment of fused images,” *International Journal of Image Processing*, vol. 4, no. 5, p. 484, 2010.

[50] V. Aslantas and E. Bendes, “A new image quality metric for image fusion: the sum of the correlations of differences,” *Aeu-International Journal of Electronics and Communications*, vol. 69, no. 12, pp. 1890–1896, 2015.

[51] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, “Assessment of image fusion procedures using entropy, image quality, and multispectral classification,” *Journal of Applied Remote Sensing*, vol. 2, no. 1, p. 023522, 2008.

[52] Y.-J. Rao, “In-fibre bragg grating sensors,” *Measurement Science and Technology*, vol. 8, no. 4, p. 355, 1997.

[53] G. Qu, D. Zhang, and P. Yan, “Information measure for performance of image fusion,” *Electronics Letters*, vol. 38, no. 7, pp. 313–315, 2002.

[54] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

Jiayi Ma received the B.S. degree in Information and Computing Science and the Ph.D. degree in Control Science and Engineering, both from the Huazhong University of Science and Technology, Wuhan, China, in 2008 and 2014, respectively. From 2012 to 2013, he was an Exchange Student with the Department of Statistics, University of California at Los Angeles, Los Angeles, CA, USA.

He is currently a Professor with the Electronic Information School, Wuhan University, Wuhan, China. He has authored or co-authored over 140 refereed journal and conference papers, including IEEE TPAMI/TIP/TSP, IJCV, CVPR, ICCV, ECCV, *etc*. He has been identified in the 2019 Highly Cited Researchers list from the Web of Science Group. He is an Area Editor of *Information Fusion*, Associate Editor of *Neurocomputing*, and a Guest Editor of *Remote Sensing*. His current research interests include the areas of computer vision, machine learning, and pattern recognition.

Hao Zhang received the B.E. degree from the School of Mechanical Engineering and Electronic Information, China University of Geosciences, Wuhan, China, in 2019. He is currently a Master student with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Zhenfeng Shao received the Ph.D. degree in aerial photogrammetry from Wuhan University, Wuhan, China, in 2004.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include remote sensing and data mining.

Pengwei Liang received the B.E. degree from the School of Information Engineering, Wuhan University of Technology, Wuhan, China, in 2017. He is currently a Master student with the Electronic Information School, Wuhan University. His research interests include computer vision, machine learning, and pattern recognition.

Han Xu received the B.S. degree from the Electronic Information School, Wuhan University, Wuhan, China, in 2018. She is currently a Ph.D. student in the Multi-spectral Vision Processing Lab, Electronic Information School, Wuhan University, Wuhan. Her current research interests include computer vision and pattern recognition.