Complex & Intelligent Systems (2022) 8:4753–4781
https://doi.org/10.1007/s40747-022-00722-9

ORIGINAL ARTICLE

MSAt-GAN: a generative adversarial network based on multi-scale and deep attention mechanism for infrared and visible light image fusion

Junwu Li[1] · Binhua Li[1,2] · Yaoxi Jiang[2] · Weiwei Cai[3]

Received: 27 October 2021 / Accepted: 5 March 2022 / Published online: 22 April 2022

Abstract

For the past few years, image fusion technology has made great progress, especially in infrared and visible light image infusion. However, the fusion methods, based on traditional or deep learning technology, have some disadvantages such as unobvious structure or texture detail loss. In this regard, a novel generative adversarial network named MSAt-GAN is proposed in this paper. It is based on multi-scale feature transfer and deep attention mechanism feature fusion, and used for infrared and visible image fusion. First, this paper employs three different receptive fields to extract the multi-scale and multi-level deep features of multi-modality images in three channels rather than artificially setting a single receptive field. In this way, the important features of the source image can be better obtained from different receptive fields and angles, and the extracted feature representation is also more flexible and diverse. Second, a multi-scale deep attention fusion mechanism is designed in this essay. It describes the important representation of multi-level receptive field extraction features through both spatial and channel attention and merges them according to the level of attention. Doing so can lay more emphasis on the attention feature map and extract significant features of multi-modality images, which eliminates noise to some extent. Third, the concatenate operation of the multi-level deep features in the encoder and the deep features in the decoder are cascaded to enhance the feature transmission while making better use of the previous features. Finally, this paper adopts a dual-discriminator generative adversarial network on the network structure, which can force the generated image to retain the intensity of the infrared image and the texture detail information of the visible image at the same time. Substantial qualitative and quantitative experimental analysis of infrared and visible image pairs on three public datasets show that compared with state-of-the-art fusion methods, the proposed MSAt-GAN network has comparable outstanding fusion performance in subjective perception and objective quantitative measurement.

Keywords Infrared and visible light image fusion · Multi-scale feature transfer · Deep attention mechanism · Multi-level receptive field · Generative adversarial network · MSAt-GAN

Introduction

Multi-modality image fusion is an important branch in the field of computer vision processing. Its purpose is to use appropriate image feature extraction methods and fusion strategies to fuse a series of source images obtained from different sensors and generate an image with salient features and complementary information [1]. Image fusion provides

✉ Binhua Li
lbh@kust.edu.cn

Junwu Li
lijunwu@stu.kust.edu.cn

Yaoxi Jiang
jiangyaoxi@kust.edu.cn

Weiwei Cai
vivitsai@ieee.org

[1] Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming 650500, China

[2] Key Laboratory of Applications of Computer Technologies of the Yunnan Province, Kunming University of Science and Technology, Kunming 650500, China

[3] School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China

rich information and highly reliable images for computer high-level vision tasks, and its advanced fusion algorithms are widely applied in many fields, such as visual tracking [2], video surveillance [3], target detection [4], person re-recognition [5], face recognition [6], semantic segmentation [7], and other fields [8, 9].

The fusion of infrared and visible images has its unique advantages. Infrared and visible images capture the same scene information through different sensors. Because of their different imaging methods, they can reflect multiple characteristics of the same scene in a comprehensive way. Infrared image sensors are sensitive to thermal radiation targets and can sense heat sources. It can also capture thermal targets under severe weather and extreme conditions with poor light. The heat source target is more prominent in the infrared image, but its spatial resolution is low, and there are problems with blurred details and textures. The visible image sensor captures light information through reflected light, and its image texture details are rich and the spatial resolution is high, which is suitable for human visual perception. However, in poor light or harsh environments, visible images often have poor visual quality. From the above analysis, it can be concluded that the visible image has rich texture details and the infrared image has a prominent target. Therefore, the fusion of infrared and visible images can fully harness and integrate complementary information to generate an image that not only conforms to human visual perception but also facilitates specific applications [10].

The focus of infrared and visible image fusion is to extract the typical features of the source image and design appropriate fusion rules. In recent years, infrared and visible image fusion methods have witnessed rapid development. Related scholars have made an overview and summary of the fusion algorithm [1, 11, 12]. Image fusion algorithms are roughly divided into two categories: traditional methods and deep learning-based methods. Based on different theories, traditional methods are divided into multi-scale transformation (MST) method [13–16], sparse representation (SR) method [17, 18], saliency representation method [19, 20], subspace method [21] and other methods [22, 23]. The above methods have good fusion effects. However, for these methods to obtain satisfactory fusion performance, it is necessary to design feature extraction algorithms and fusion strategies in an artificial manner. Moreover, the diversified feature extraction methods and complicated fusion rules make the fusion model increasingly complicated, limiting the practicability and real-time capability.

Deep learning has been widely used and developed in the field of computer vision and image processing with its excellent feature representation capability, especially in the fusion of infrared and visible images. According to the type of network structure, deep fusion methods are divided into: Convolutional Neural Network (CNN)-based methods [24–26] and Generative Adversarial Network (GAN)-based methods [27–30]. Literature [24] designed a Siamese Convolutional Neural Network (Siamese CNN), using image pyramids and local similarity strategies to integrate pixel activity information and adaptively adjust the fusion model. Literature [25] decomposes the source image into basic blocks and detail blocks and then combines the features extracted by the deep network with the basic part of the fusion to reconstruct the fused image. Literature [26] proposed a new network structure (DenseFuse) based on the dense block [31] and self-encoder to fuse infrared and visible images. In the training phase, the fusion layer is discarded to degenerate it into a self-encoder network; in the testing phase, the fusion strategy is used to fuse the depth features and reconstruct the image through the decoder.

Presetting Ground-Truth in the neural network will achieve better performance, so the CNN-based deep learning method is more suitable for supervised learning. However, in the field of image fusion, Ground-Truth is often difficult to obtain, especially there is no unified fusion standard for infrared and visible images, and the Ground-Truth does not exist. We should regard the image fusion task as an unsupervised learning problem. In addition, GAN has its unique advantages in solving unsupervised deep learning problems. In recent years, GAN research has shown a blowout type growth. Many GAN-based algorithms are also employed in image fusion. Literature [27] applied GAN to image fusion tasks for the first time and proposed an end-to-end fusion framework based on GAN (FusionGAN). The FusionGAN includes a generator and a discriminator. The generator is responsible for generating images that can preserve infrared light image intensity information and visible light image gradient information, and the discriminator is used to distinguish fusion image and visible image. Since there is one discriminator in the framework, the generated image contains only the gradient information of the visible image, thus inevitably ignoring the gradient information of the infrared image. To solve this problem, Literature [28] proposed a dual-discriminator conditional generation adversarial network (DDcGAN). In DDcGAN, the infrared discriminator forces the generated image to retain the infrared intensity, and the visible discriminator forces the generated image to have more visible texture information. In spite of that, the images generated by these two GANs cause texture detail loss and lack of integrity of neighboring pixels.

To fix the above problems, inspired by the deep multi-scale feature integration and attention mechanism [32–34], we propose a new generative adversarial network architecture based on multi-scale feature transfer and deep attention mechanism feature fusion, and name it MSAt-GAN. First, inspired the Inception V3: Using different sizes of convolution kernels to extract different sizes of receptive fields—the various kernel combination means different level feature fusion; the deeper

the network, the more abstract the extracted features, and the larger the receptive field involved in each feature. To better obtain the important features of the source image from different receptive fields and angles, we choose three classic convolution kernels of $3 \times 3, 5 \times 5$, and 7×7 as our three feature channels. The feature encoder adopts dense connection (DenseNet) during feature transfer, which makes full use of multi-scale features and strengthens the mapping relationship between features of different scales without changing the size of the source image. Second, to integrate the multi-level deep features of different receptive fields, we introduce a multi-scale deep attention mechanism in the encoder network to focus attention on the important features extracted from the multi-level receptive fields in both space and channel dimensions, and integrate them according to the level of attention. In addition, in the decoder network, the multi-level deep features fused in the encoder network are concatenated to compensate for the loss of previous features. Finally, we establish a generative adversarial network architecture with a generator and dual discriminators and an adversarial game is established in the generator and dual-discriminators to force the generated image to retain meaningful information from visible and infrared images.

To verify the efficiency of the MSAt-GAN fusion method, we compare and analyze three typical deep learning-based fusion methods on a set of examples of infrared and visible images. The experimental results are shown in Fig. 1. It can be seen from Fig. 1 that compared with the other three classic deep learning algorithms, the image fused by our method has the prominent target, clear texture details, high contrast, and the best visual perception. Especially in the woods on the upper side and the ripples of the lake in the middle, the outline is clear and the details are rich while the images fused by the other three algorithms have the defects of blurry contours and loss of texture details.

The main contributions of this paper are summarized as follows:

1. In the generator-encoder network, we introduce a multi-scale deep feature extraction module with multiple receptive fields. Three classic convolution kernels of $3 \times 3, 5 \times 5$, and 7×7 are used as convolution kernels for the three feature channels. Introducing 5×5 and 7×7 convolution kernels will increase computing time while expanding the receptive field. Therefore, a small convolution kernel is used instead of a large one. In this way, while maintaining the same receptive field, the parameters of the model are greatly reduced. Introducing a feature extractor with multiple receptive fields can better extract the features of the source image from all directions, not only from the depth and width, but also from the angle similar to the human visual system to comprehensively extract the deep features of the image, which greatly enhances the diversity and flexibility of the extracted features, and also provides a basis for the subsequent feature fusion module.
2. In the generator-encoder network, we propose a multi-scale deep attention fusion mechanism. It can calculate the important representations of attention in both spatial dimension and channel dimension of the multi-scale deep features obtained from different receptive fields, and fuse them according to the level of attention. This can better extract and fuse important features. The proposed multi-scale deep attention mechanism breaks through the limitations of the artificial design fusion strategy and significantly improves the fusion performance, and suppresses noise and undesirable artifacts to varying degrees.
3. In the generator-encoder network, we adopt dense connection (DenseNet) in all layers, which takes full advantage of multi-scale features and strengthens the mapping relationship between features of different scales. In addition, the multi-level deep features fused in the encoder and the deep features in the decoder are cascaded so that the feature transfer can be reinforced and the previous features can be better used.
4. In the network structure, we put forward an end-to-end dual-discrimination WGAN-LP generative adversarial network. The infrared discriminator distinguishes the generated image from the infrared source image, forcing the generated image to preserve more target background information from the infrared light source image. The visible discriminator distinguishes the generated image from the visible light source image, forcing the generated image to have more texture information of the visible light image. Therefore, the image generated by the generator retains more meaningful information from the two source images. It is well known that GAN lacks training instability and triggers mode collapse, so we use a new gradient penalty term (WGAN-LP) to strengthen the Lipschitz constraint to improve the training performance and stability of the MSAt-GAN model.
5. Existing infrared and visible training datasets are mostly generated by cropping the TNO standard database, or directly using the MS-COCO dataset. Yet, the TNO dataset itself does not contain many image pairs and the resolution is not high, and the generated sample image by cropping from the dataset is with a limited number, and lack of richness; the MS-COCO dataset is a high-resolution focus image, but it does not contain infrared light information. Using MS-COCO as a training set will lead to inaccurate infrared feature extraction. Therefore, we introduce a new RGB-NIR Scene Data Set and hereafter call it Nirscene. It contains lots of infrared and visible image pairs and rich scenes information, and its cropping images can greatly improve the richness of our training samples.

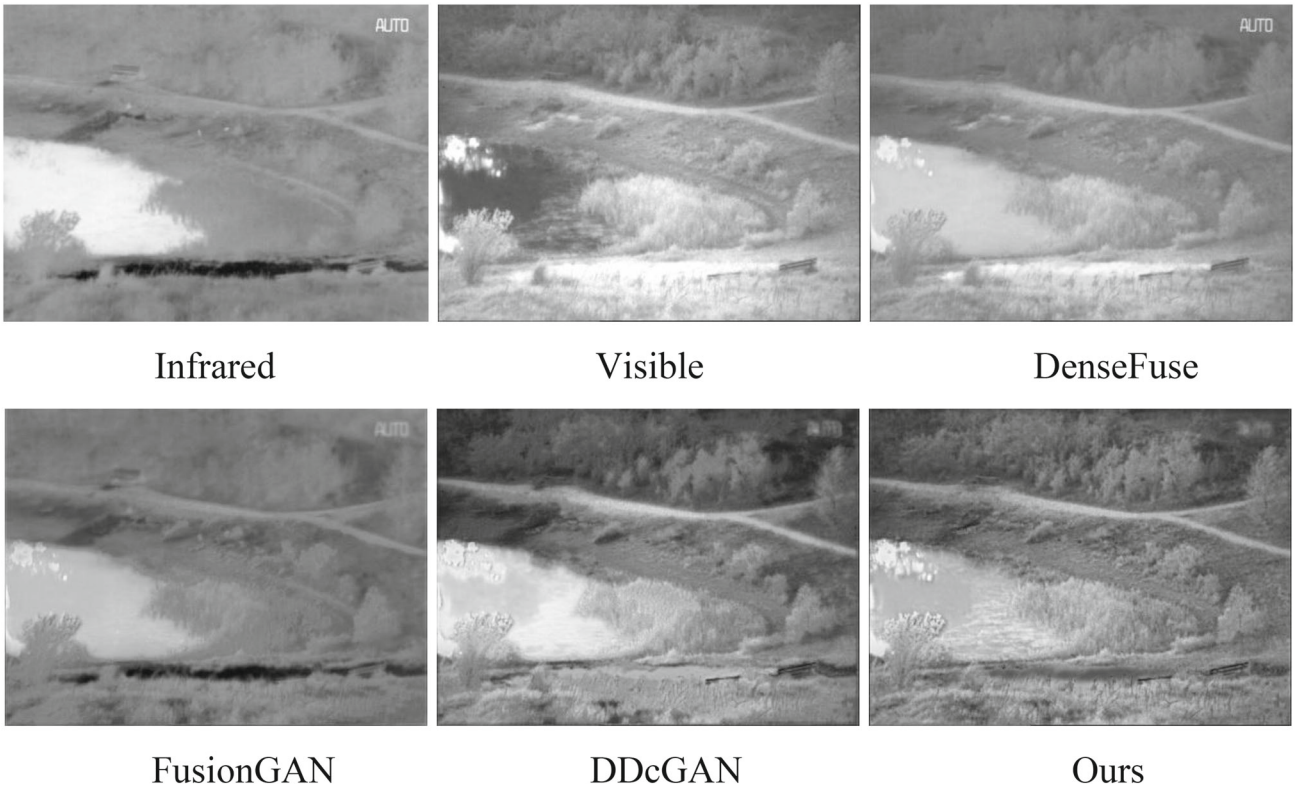


Fig. 1 Schematic illustration of MSAt-GAN. The first row shows the two source images and the fused results of DenseFuse [26], the second row presents the fused results of FusionGAN [27], DDcGAN [28] and proposed method MSAt-GAN

6. Extensive experiments and ablation studies have been conducted on two different public datasets, showing the necessity of a multi-scale deep feature extractor with multi-receptive fields and a multi-scale deep attention fusion mechanism. Experimental results prove that MSAt-GAN has excellent performance to other state-of-the-art methods in qualitative and quantitative comparison.

The structure of the paper is designed as follows. The related studies are shown in the next section, and more details of MSAt-GAN are presented in the third section. The experimental results and evaluation are given in the fourth section, and the conclusions are shown in the last section.

Related research

This section first introduces a typical dual-discriminator conditional generation adversarial network (DDcGAN) for infrared and visible image fusion, and then discusses the improved Wasserstein Generative Adversarial Network (WGAN-LP) for improving the stability of GAN training. Finally, discussion of the attention mechanism in deep learning is conducted.

DDcGAN [28]

DDcGAN establishes an adversarial game between a generator and two discriminators. The generator generates a fusion image according to specific loss functions to deceive the two discriminators. The two discriminators are used to distinguish the structural difference between the fusion image and the two source images. The loss function of DDcGAN is as follows:

Generator loss

$$\ell_G = \ell_G^{\text{adv}} + \lambda \ell_{\text{con}}, \tag{1}$$

$$l_G^{\text{adv}} = E[\log(1 - D_v(G(v, i)))] + E[\log(1 - D_i(G(v, i)))], \tag{2}$$

$$\ell_{\text{con}} = E[\|G(v, i) - i\|_F^2 + \eta \|G(v, i) - v\|_{TV}], \tag{3}$$

where ℓ_G^{adv} and ℓ_{con} represent adversarial loss and content loss, respectively; $\|\cdot\|_F$ is the Frobenius norm, $\|\cdot\|_{TV}$ is the TV norm, v and i represent the visible source image and the infrared source image, respectively, and λ and η represent balance coefficient. The information of thermal radiation and texture details are mainly characterized by the information

of pixel intensity and gradient changes, so Frobenius norm is adopted to constrain the fusion image and infrared image. TV norm can effectively solve the problem of non-deterministic polynomials. In the regularization term, TV norm is introduced to constrain the gradient of the fused image and visible image.

Discriminator loss

$$\ell_{Dv} = E[-\log D_v(v)] + E[-\log(1 - D_v(G(v, i)))], \quad (4)$$

$$\ell_{Di} = E[-\log D_i(v)] + E[-\log(1 - D_i(G(v, i)))], \quad (5)$$

where ℓ_{Dv} and ℓ_{Di} represent the loss functions of the visible discriminator and infrared discriminator, respectively.

WGAN-LP [35, 36]

On the basis of WGAN, WGAN-GP removes the weight trimming, and solves the problem of difficulty in training and slow convergence of the WGAN model. To make Critic meet the 1-Lipschitz constraint, a gradient penalty term is used in Critic's loss. WGAN-LP is an improvement of WGAN-GP, which strengthens the 1-Lipschitz constraint by squaring the penalty for large deviations, thus making the 1-Lipschitz constraint more reasonable. WGAN-LP is more stable than WGAN-GP model training. The loss functions of WGAN-LP is shown as follows:

$$\min_G \max_D L_W(D, G) = \mathrm{E}_{x\sim pr}[D(x)] + \mathrm{E}_{z\sim pz}[D(G(z))] + \mu \mathrm{E}_{\tilde{x}}[(\|\nabla_{\tilde{x}} D(\tilde{x})\|_2 - 1)], \quad (6)$$

$$\min_G \max_D L_W(D, G) = \mathrm{E}_{x\sim pr}[D(x)] + \mathrm{E}_{z\sim pz}[D(G(z))] + \mu \mathrm{E}_{\tilde{x}}[(\max\{0, \|\nabla D(\tilde{x})\|_2 - 1\})^2], \quad (7)$$

where Eq. (6) represents the loss of WGAN-GP, and Eq. (7) is the loss of WGAN-LP after improvement; the first two items are the Wasserstein distance estimation between the real image and the generated image, the last is the gradient penalty item, μ is the balance coefficient, and $\tilde{x}$ is the sample of the real image and the generated image.

Deep attention mechanism

Derived from the human visual attention mechanism, attention mechanism is much more a simulation of the attention behavior of humans in reading and listening and speaking by machines, which can be regarded as a bionic mechanism. It can focus on the important features of the object according to the importance of the object. In the past 2 years, the attention model has been widely applied in various types of deep learning tasks such as machine translation and text translation, image recognition and speech recognition in natural language processing [37–39], which is the most worthy of attention and in-depth research in deep learning. Due to its good algorithm performance, the model has also been widely applied and developed in computer vision, especially in image fusion tasks. Literature [40] deploys the deep attention mechanism for hyperspectral and multispectral image fusion, and its designed spatial attention network is used to extract tiny details and enhance the spatial structure of the image. Literature [41] designs a multi-resolution classification dual-branch attention fusion network (DBAF-Net) for remote sensing image fusion. The spatial resolution of panchromatic images (PAN) is higher than that of multispectral images (MS). In addition, based on different image types, the author sets up two attention models: spatial attention (SA) module and channel attention (CA) module. Through the dual-branch attention mechanism, unimportant information such as the image background is suppressed and the original feature information of the extracted image data is further enhanced. Literature [32] fuses multi-scale deep features, proposes a spatial/channel attention model fusion strategy and introduces it into infrared and visible image fusion. The spatial attention module is mainly based on the L1 operator; the channel attention module introduces and compares three kinds of global pooling operators: the average operator, the maximum operator and the nuclear norm operator. Experiments indicate that using the average pooling operator in the channel module has the best performance.

MSAt-GAN algorithm

In this section, we introduce in detail the proposed generative adversarial fusion network (MSAt-GAN) based on multi-scale feature transfer and deep attention mechanism feature fusion. First, we make a comprehensive statement for the fusion problem and illustrate the architecture of MSAt-GAN in this part. Then, the network architecture of the generator (encoder-feature generator-decoder) and the dual discriminator, and the deep attention fusion mechanism are introduced in detail. Finally, the loss functions of MSAt-GAN network are under discussion.

Network architecture

The proposed MSAt-GAN fusion architecture is shown in Fig. 2. The network architecture mainly includes three components: a generator (Generator) and two discriminators (Discriminator—Dir, Discriminator—Dvis); the generator is composed of an encoder (Encoder), a feature generator (Feature Generator) and a decoder (Decoder). The main purpose

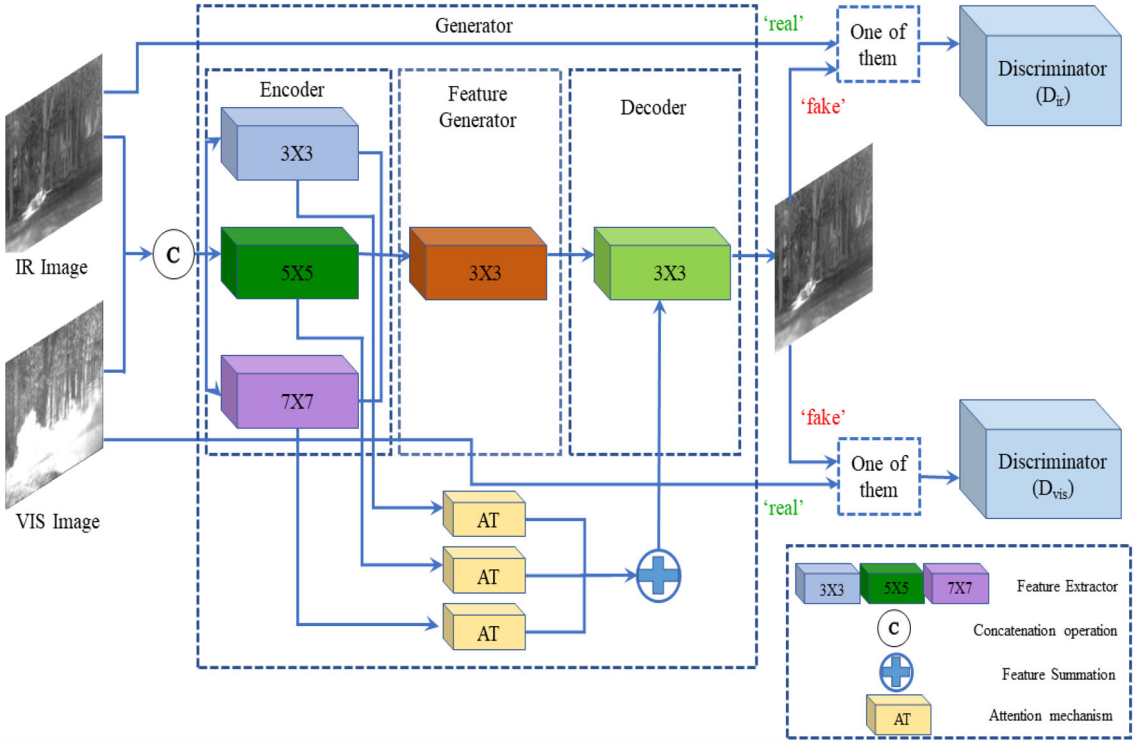


Fig. 2 Framework of the proposed MSAt-GAN

of the generator is to generate a fusion image with important information of two source images through network learning based on the inputted infrared and visible light source image pairs. Being two different modalities, the infrared and visible images show different representations of the same scene information and complement each other, so we concatenate the infrared and visible light image pairs and feed them to the encoder for encoding before feature extraction. The advantage of it is that the concatenated images have the properties of both infrared and visible light at the same time from the very beginning. Even if there is no discriminator to constrain the generator, the image generated will still have partial infrared intensity information and visible light gradient information, and this has been proved in Literature [27]. In the encoder, we introduce a multi-scale deep feature extraction module with multiple receptive fields. Three classic convolution kernels of 3, 5 and 7 are used in the feature channel to extract the multi-receptive field and multi-scale deep features of the source image. Through the feature extraction of multiple receptive field channels, the obtained source image information can be characterized in a more comprehensive manner. After that, the deep features extracted by the three receptive field feature extractors are concatenated and fed to the feature generator to further extract the multi-scale deep features of the source image. Finally, the output of the feature generator is used as the input of the decoder, and the fused image is generated through the decoder.

We design a multi-scale deep attention mechanism in the encoder network of the generator. For the multi-scale deep features obtained by the three different receptive field channels, the attention value of these features is calculated in the spatial dimension and the channel dimension, and the attention is focused on the important feature maps of each scale, and the redundancy information is eliminated. In the decoding process, the attention maps calculated by the encoder in different receptive fields are fused together according to the level of attention. In this way, the previous-level features extracted by the encoder can be fully utilized, and information loss is avoided while the feature transfer is enhanced. Therefore, the images generated by the MSAt-GAN fusion method can better capture the foreground information of the infrared image target and the detailed information of the visible light image scene.

What is more, we establish an adversarial game between a generator and two discriminators. The infrared discriminator (Dir) is designed to distinguish the generated image from the infrared source image, and constrains the image generated by the generator to preserve as much of the pixel intensity information in the infrared image as possible; the visible discriminator is used to distinguish the generated image from the visible source image, and constrains the generated image retains as much of the texture detail in the visible image as possible. In addition, we believe that when the two discriminators could not tell the generated image from the source

image, the generated image meets the fusion demand. Similar to most of the GAN fusion methods, the MSAt-GAN can be represented by generative and adversarial functions, as shown in Eq. (8). The aim of the generator is to meet minimize Eq. (9), and the goal of the discriminator is to meet maximize Eq. (9):

$$\begin{aligned} L_W(D_{ir},\, D_{\text{vis}},\, G) = {} & \mathrm{E}_{x\sim \text{pir}}[D_{ir}(x)] + \mathrm{E}_{g\sim pg}[-D_{ir}(G(g))] \\ & + \mathrm{E}_{x\sim \text{pvis}}[D_{vis}(x)] \\ & + \mathrm{E}_{g\sim pg}[-D_{\text{vis}}(G(g))] \\ & + \mu_1 \mathrm{E}_{\tilde{r}}[(\max\{0,\, \|\nabla D_{ir}(\tilde{r})\|_2 - 1\})^2 \\ & + \mu_2 \mathrm{E}_{\tilde{v}}[(\max\{0,\, \|\nabla D_{\text{vis}}(\tilde{v})\|_2 - 1\})^2], \end{aligned} \tag{8}$$

$$\min_{G} \max_{D_{ir}} \max_{D_{\text{vis}}} \{L_W(D_{ir},\, D_{\text{vis}},\, G)\}, \tag{9}$$

where D_{ir} and D_{vis} represent the infrared discriminator and visible discriminator, respectively; g is the fused image, pg is the distribution of the generated image, pir and pvis represent the real distribution of infrared and visible image, respectively, $\tilde{r}$ is the sample of generated image and infrared image, $\tilde{v}$ is the sample of generated image and visible light image, μ_1 and μ_2 denote the penalty quantity. Through the adversarial game between the generator (G) and the two discriminators (D_{ir} and D_{vis}), the Wasserstein distance between pg and the real distributions (pir and $pvis$) of the two source images will decrease at the same time, which means that the generated image is more similar to the source image.

Generator architecture

The generator consists of an encoder, a feature generator and a decoder, as shown in Fig. 3. In the test phase, only a pair of infrared and visible images needs to be input into the generator, and a complementary image that retains the infrared and visible light information as much as possible will be automatically generated.

Encoder architecture

A key of image fusion is to design a reasonable encoder to extract important information of the source image so as to represent the source image as much as possible. In this regard, the pros and cons of the encoded feature extractor exert a huge impact on the subsequent image fusion effect. In the encoding process, the existing deep learning methods tend to extract the multi-scale features of the source image by increasing the number of channels or the depth of the network as much as possible. Although it can expand the receptive field and extract deep features, the increase of the network model also means a sudden increase of parameters, which may probably lead to gradient information disappearance or explosion. Therefore, it can be concluded that the larger or deeper network model does not necessarily bring with a better performance, and this conclusion has been proved in ResNet [42]. To comprehensively extract the features of the source image and increase the receptive field of feature extraction, we design 3×3, 5×5 and 7×7 multi-receptive field feature extractors in the encoder as Fig. 3 shows, considering that the 3, 5, and 7 convolution kernels are the most commonly used and classic in existing deep learning. The increase of the receptive field will improve the accuracy of feature extraction but at the same time it will increase the computing time. Given the above situation, and taking an inspiration from the Inception V3, we replace the 5×5 convolution kernel with two 3×3 convolution kernels, and replace the 7×7 convolution kernel with three 3×3 convolution kernels. Such replacement can reduce parameters while keeping the receptive field unchanged. In the alternative convolution block, we add the PRelu loss function to increase the nonlinearity of the convolution. Our three receptive field feature extraction channels all consist of three convolutional blocks, and each of them contains three layers: a convolutional layer, a Batch Normalization (BN) layer and a PRelu layer. The channel dimensions of the three layers are all set to 32, and the step length is 1, and the padding operation is used in the convolution process to keep the feature resolution consistent before and after the convolution and avoid information loss. In the encoder, we adopt the DenseNet [31] connection to establish short direct connections in a feedforward manner between each layer and among all layers so that our multi-receptive field deep feature extraction can provide more accurate and complementary information.

Feature generator architecture

The feature generator aims to further extract the deep features of the three concatenated receptive fields to improve the accuracy of the extracted features, which includes: a convolutional layer, a BN layer and a PRelu layer. The convolution kernel is set to 3×3, the channel dimension is 288, and the stride is 1.

Encoder architecture

The input of the decoder is equivalent to the output of the feature generator, and its purpose is to generate an image with rich information based on the multi-scale deep feature fusion of the extracted source image. The decoder contains 5 convolutional blocks. The first block contains a convolution layer and a PRelu layer. The second to fourth blocks are, respectively, composed of a convolutional layer, a BN layer, and a PRelu layer. The fifth block consists of a convolution layer and a Tanh activation function. The kernels of the first to fifth convolution blocks are 32, 32, 32, 16, and 1, in turn, and

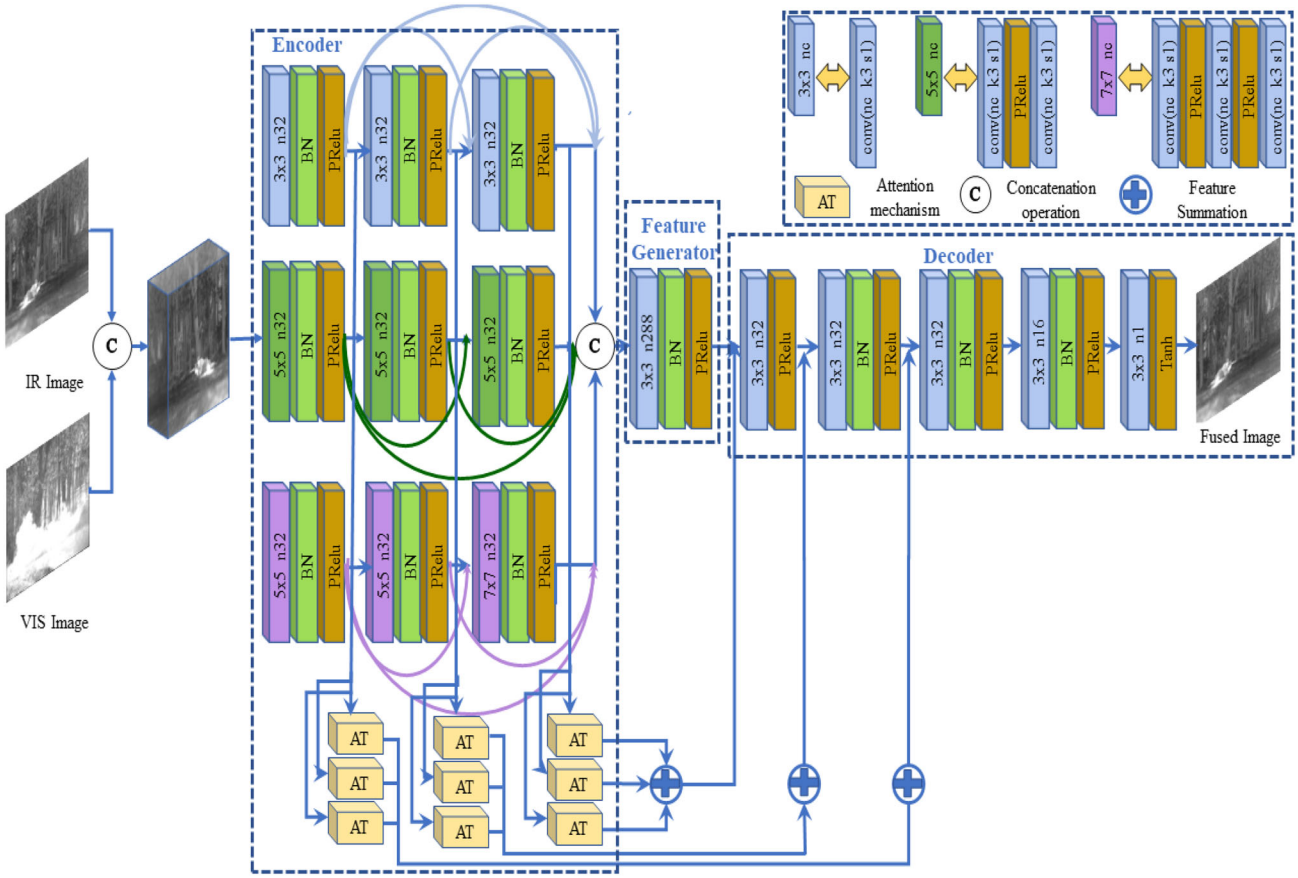


Fig. 3 Architecture of Generator. *k* denotes the kernel size, *n* denotes the number of filters, and s represents the stride

all of them are set to 3×3, with a step length of 1. We make skip connections between the first to third convolutional layers and the multi-scale deep attention map of the multiple receptive fields in the encoder. By connecting with the front layer, it can make compensation for the information loss of the feature map after multi-layer convolution and make full use of the important information that distinguishes between infrared and visible light extracted by the front layer. The decoder connection fully integrates the information of different receptive fields and ensures the structural consistency of the deep features.

All activation functions in the generator are PRelu [43] as it can adaptively learn and modify the parameters of the linear unit and improve the accuracy.

Deep attention fusion mechanism

The multi-attention mechanism can capture the salient areas in the visual scene information and is widely applied in computer vision tasks. Infrared and visible light images are of multi-modality, which are different representations of the same scene information. Therefore, we employ the mechanism to find the appropriate features of these two modalities, and put our attention on the focused modal region, that is, on the distinguishable parts of the same feature. For the feature map of the convolutional network, the channel dimension and the spatial dimension both contain rich attention information. By fusing the attention information from the above two dimensions, we can extract a more comprehensive and reliable feature information of the source image. Based on previous studies, we propose a multi-scale deep attention mechanism that focuses attention on the important features extracted by the multi-level receptive field feature in the two dimensions of spatial and channel and merge extracted features according to the level of attention. The whole mechanism module is shown in Fig. 4.

Fusion of channel attention

First, we input the feature maps of the three receptive field feature channels: $f_1(H \times W \times C)$, $f_2(H \times W \times C)$ and $f_3(H \times W \times C)$. Next, we employ the global average pooling on the three feature maps in the H and W dimensions, respectively, to obtain three $1 \times 1 \times C$ feature maps, and then make a two-layer full connection to the three feature maps, and add the activation function Relu. Sigmoid activation is

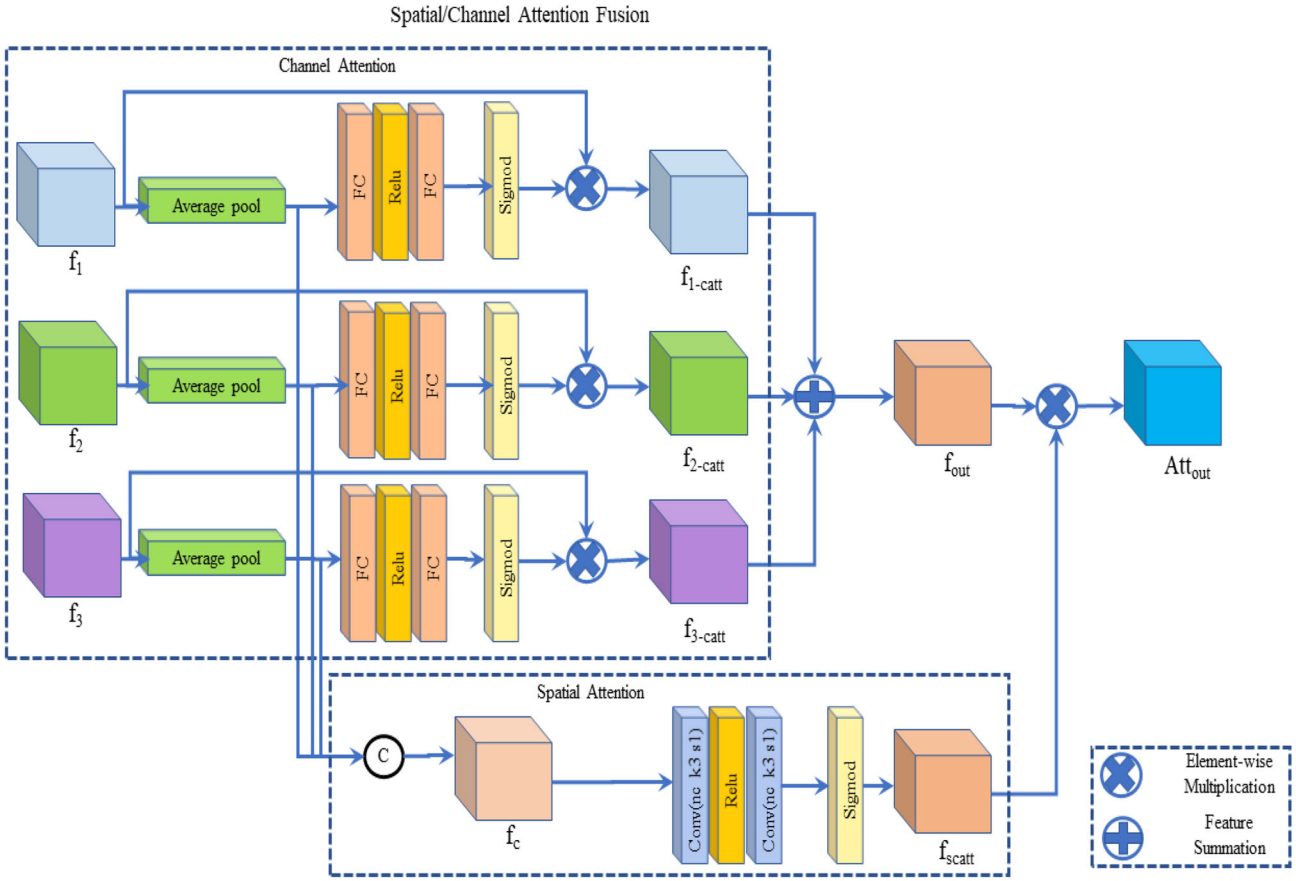


Fig. 4 Architecture of deep attention mechanism fusion

then performed on the generated feature maps to get a channel attention weight map. Finally, the three f weight maps and the three feature maps are multiplied and added to generate the final channel attention fusion feature, as shown in the following equation:

$$f_{\text{out}}^{c} = \sum_{i=1}^{3} \sigma(D_2(D_1(\text{avg}p(f_i^c)))) \times f_i^c, \tag{10}$$

where f_1^c, f_2^c and f_3^c, respectively, correspond to the channel features of the receptive field of the convolution kernel of 3, 5, and 7, D_1 and D_2 are two-layer full connections, respectively, avg$p$ is the global average pooling, and σ is the Sigmoid activation function.

Fusion of spatial attention

First, we input the feature maps of the three receptive field feature channels: $f_1(H \times W \times C)$, $f_2(H \times W \times C)$ and $f_3(H \times W \times C)$. Next, we employ the global average pooling on the three feature maps in the channel dimension to obtain three $H \times W \times 1$ feature maps, and then make concatenated connection and convolution among the three maps. Set the two-layer convolution kernel to 3×3, and use the Relu activation function. Then, the convolved feature map is activated by Sigmoid to generate a spatial attention weight map. Finally, we multiply the spatial attention weight map and the channel attention fusion feature (f_{out}^c) to obtain the final deep attention feature map Att_{out}^c, as shown in following equation:

$$Att_{\text{out}}^{c} = \sigma(\text{Conv}_2(\text{Conv}_1(C(\text{avg}p(f_i^c))))) \times f_{\text{out}}^c, \tag{11}$$

where f_i^c represents the channel feature of the receptive field of 3, 5 and 7 convolution kernel, Conv_1 and $Conv_2$, respectively, represent the two-layer fully convolutions, avg$p$ is the average pooling, σ is the Sigmoid activation function and C is concatenated operation.

Dual-discriminator architecture

The goal of the discriminator is to form an adversarial game with the generator to guide the generated image distribution as close to the actual distribution as possible. Since this article focus on image fusion of infrared and visible light images, the two forms with different modalities, we design two discriminators: infrared light discriminator (D_{ir}) and visible light

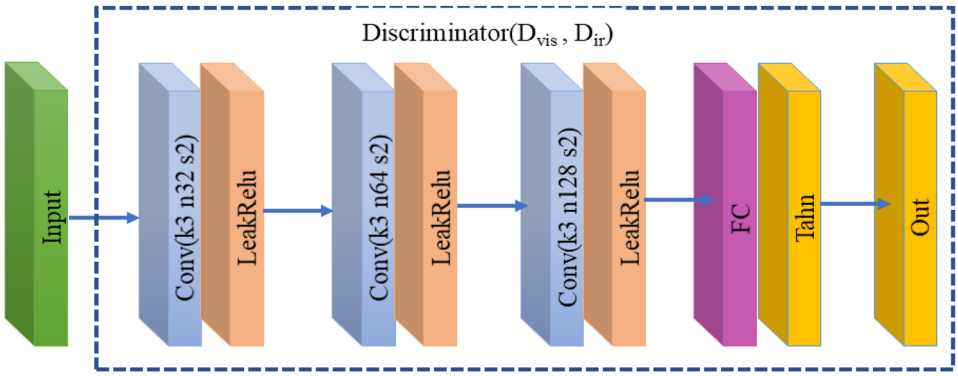


Fig. 5 Architecture of Discriminators. *k, n*, and *s* denote the kernel size, the number of filters, and stride, respectively

discriminator (D_{vis}) to distinguish the fused image from the source image of the two modalities. During the training process, the generator, infrared discriminator and visible light discriminator should maintain balance. Otherwise, the high efficiency of any one network will bring about the inefficiency of the other two networks, which would consequently lead to a worse performance of image fusion. The balance of our network training is maintained by designing a reasonable network structure and loss functions. The purpose of the discriminators is to distinguish whether the generated image is real or fake, which is a binary classification problem that is relatively simple for neural networks, so the structure of the discriminators is simpler than that of the generator. The two independent discriminators are designed as the same structure, as shown in Fig. 5.

There are four layers in the discriminators: the first three layers have three convolutional blocks, and each block contains a convolutional layer and a LeakRelu activation function; the fourth layer contains a fully connected layer and a Tanh activation function. The size of the convolution kernel is set to 3, and the stride is 2. The channel dimensions of the first three layers are 32, 64, and 128, respectively.

The loss function of generator

The loss function is another method used to maintain a balance between the generator and the discriminators. The generator must not only form an adversarial game with the discriminators, but also keep the data distribution of the generated image consistent with the source image. Therefore, its loss function should restrict the similarity between the fused image and the source image in terms of structure and content, so as to deceive the discriminators. The total loss of MSAt-GAN includes two parts: discrimination loss and content loss. The definition is shown in the following equation:

$$L_{\text{tlos}} = L_G^{\text{adv}} + \mu L_{\text{contlos}}, \tag{12}$$

where μ is the balance coefficient to control the two losses.

The adversarial loss of the generator includes two parts: the adversarial loss of the infrared discriminator and the adversarial loss of the visible light discriminator. The definition is shown in the following equation:

$$L_G^{\text{adv}} = E_{g\sim pg}[-D_{ir}(g)] + E_{g\sim pg}[-D_{\text{vis}}(g)], \tag{13}$$

where g and pg denote the generated image and the data distribution of generated image.

Content loss: the texture detail information of visible light and the thermal target of infrared light can be, respectively, represented by gradient change information and pixel intensity information [28], so we also employ the Frobenius 2 norm constraint to generate the image to make it has similar pixel intensity information to the infrared source image. Introducing the Frobenius 2 norm to constrain the gradient information similarity between the generated image and the visible light image. The definition of content loss is shown in the following equation:

$$L_{\text{contlos}} = E[\|g - ir\|_F^2] + \lambda E[\|\nabla g - \nabla \text{vis}\|_F^2], \tag{14}$$

where F denotes the Frobenius norm, ∇g and ∇vis represent the gradient of the generated image and the visible light image, respectively, λ is the balance coefficient.

The loss function of dual discriminator

There are two losses in our discriminators: the discrimination loss of visible light and the discrimination loss of infrared light. The use of an independent dual-discriminator structure can make the fused image comprehensively preserve the important information of the two source images at the same time. For the stability of GAN training, we introduce the WGA-LP gradient penalty term in the discriminators. The loss function is shown in the following equations:

$$\begin{aligned} L_{D_{ir}} = {} & \text{E}_{ir\sim pir}[D_{ir}(ir)] + \text{E}_{g\sim pg}[-D_{ir}(G(g))] \\ & + \mu_1 \text{E}_{\tilde{r}}[(\max\{0, \|\nabla D_{ir}(\tilde{r})\|_2 - 1\})^2, \end{aligned} \tag{15}$$

$$L_{D_{\text{vis}}} = \mathrm{E}_{\text{vis}\sim p\text{vis}}[D_{vis}(\text{vis})] + \mathrm{E}_{g\sim pg}[-D_{\text{vis}}(G(g))] + \mu_2 \mathrm{E}_{\tilde{v}}[(\max\{0,\ \|\nabla D_{\text{vis}}(\tilde{v})\|_2 - 1\})^2], \quad (16)$$

where $\tilde{r}$ and $\tilde{v}$, respectively, represent the sample of generated image and the infrared and visible light image, and μ_1 and μ_2 are the gradient penalty coefficient.

Experimental results

In this section, we prove the effectiveness of the MSAt-GAN fusion method through experiments, and make a qualitative comparison on the two public infrared and visible datasets of TNO,[1] INO[2] and Nirscene.[3] For a more comprehensive analysis, we use eight quantitative indicators to evaluate the fusion results. Moreover, we carry out ablation research experiments to validate the proposed multi-scale deep feature extraction module with multi-receptive fields and multi-scale deep attention fusion mechanism.

Datasets and training details

Datasets

In fact, most of the existing infrared and visible light image training datasets are obtained by cropping the public TNO dataset to get enhanced training data. TNO image fusion dataset contains multi-band night images of military-related scenes. The type of infrared and visible light image information contained in the dataset is limited, and so is the number of images contained in each type. Using this dataset would consequently reduce the robustness of the model. There is also another dataset of MS-COCO, which contains 82,783 training sets and 1000 validation sets. However, the data itself is a multi-focus visible light image and does not contain infrared image information. Using this MS-COCO as a training dataset, infrared information will inevitably be lost during the training process. In this case, we adopt a new Nirscene dataset as our enhanced dataset, which contains 9 types of 477 image pairs captured by RGB and NIR. The scene categories in the dataset include: countryside, fields, forests, indoor scenes, mountains, old buildings, streets, cities and rivers. We crop these 477 infrared and visible light image pairs to enhance the training dataset. It is worth noting that the cropped image cannot be too small, as small infrared image blocks will usually cause invalid information; and the cropped image should not be too big, for it will bring a sharp increase of computational resources. Based on the current hardware configuration of our computer, these image pairs are cropped into 64,580 infrared and visible light image pairs for the sake of an enhanced training dataset.

Training details

Through substantial training experiments on MSAt-GAN, the model parameters are set as follows: μ=0.7, λ=1.3 and μ_1=μ_2=1. The exponential decay learning rate is used in network learning, the initial learning rate is set to 2×10^{-4}, the decay coefficient is set to 0.9 and the BatchSize is set to 24. Based on WGAN training techniques, our generator employs RMSProp optimizer, and both discriminators use SGD optimizer. To maintain the training balance between the generator and the discriminators, the discriminators usually have to train more than the generator. Here, we set the discriminators to train twice and then the generator to update once. In addition, we conduct training experiments on Nvidia RTX-3090 GPU-24 g memory, which adopts the TensorFlow framework. For the comparison algorithm running on the CPU, simulation software is the MTALB 2021a.

For the convenience of readers, all notation variables of the loss function of the MSAt-GAN model are presented in Table 1.

Image evaluation index and comparison algorithm

Generally, there are subjective and objective evaluation types of image fusion quality. The former is easily affected by human factors such as personal emotions and subjective vision, and the fusion effects of various algorithms have certain similarities and are hard to distinguish, while the latter can evaluate the pros and cons of the fused image in virtue of quantitative objective indicators. Therefore, we select eight objective indicators from five categories of performance index to make a quantitative compare of MSAt-GAN and other fusion methods, namely the information entropy (EN) of information theory, the multi-scale structural similarity index measurement (MS-SSIM) based on structural information [44], spatial frequency (SF) and standard deviation (SD) of image features, visual reality fidelity (VIF) and Qabf of human vision perception [45], the correlation coefficient (CC) and sum of difference correlation (SCD) based on source image and fused image [46].

EN measures the richness of information contained in the image: the higher the value, the richer the information and the better the quality of the fused image. MS-SSIM considers the similarity of two images in brightness, contrast, and structural information: the higher the value, the more similar the two images. SF reflects the change rate of image gray level, and the larger the spatial frequency, the clearer the image. SD reflects the dispersion of image pixel gray value relative to the

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

[2] https://www.ino.ca/en/technologies/video-analytics-dataset/

[3] https://ivrlwww.epfl.ch/supplementary_material/cvpr11/index.html

Table 1 Loss function variable notation table

Variables	Term	Variables	Term	Variables	Term
D_{ir}	Infrared discriminator	Dvis	Visible discriminator	g	Generated image
pg	Distribution of generated image	pir	Distribution of infrared image	$p$vis	Distribution of visible image
$\tilde{r}$	Sample of generated image and infrared image	$\tilde{v}$	Sample of generated image and visible image	μ_1	Infrared discriminator penalty quantity
μ_2	Visible discriminator penalty quantity	G	Generator	L_{tlos}	Generator loss
L_G^{adv}	Generator adversarial loss	L_{contlos}	generator content loss	μ	Generator balance coefficient
F	Frobenius norm	ir	Infrared image	vis	Visible image
∇g	Gradient of generated image	∇vis	gradient of visible image	λ	Content loss balance factor

mean value. The larger the SD value, the higher the contrast and the higher the fusion quality of the image. VIF shows the fidelity of the source image and the fusion image based on human vision: the higher value equals to the better the human visual performance. Qabf, being a pixel-level image fusion quality evaluation index, uses local metrics to estimate the performance degree of inputted salient information in the fused images. It reflects the quality of visual information obtained from fused images. The higher the Qabf value, the higher the subjective visual quality of the fused image. CC signifies the degree of linear correlation between the source image and the fused image and SCD surveys the amount of information transferred from the input image to the fused image. In short, the larger the above eight indexes' value, the better the fusion effect.

To verify the efficiency of the proposed MSAt-GAN, we select seven classic or state-of-the-art methods on three public infrared light datasets and make a comparison and analysis, including four traditional methods: transfer and total variation minimization fusion method (GTF) [47], cross bilateral filter fusion method (CBF) [48], guided filtering based fusion method (GFF) [49], and multiresolution singular value decomposition (MSVD) [50]. In addition, in view of that our method is based on GAN and the encoder–decoder, we choose three similar deep learning methods: DenseFuse [26], FusionGAN [27] and DDcGAN [28].

Experimental analysis on TNO dataset

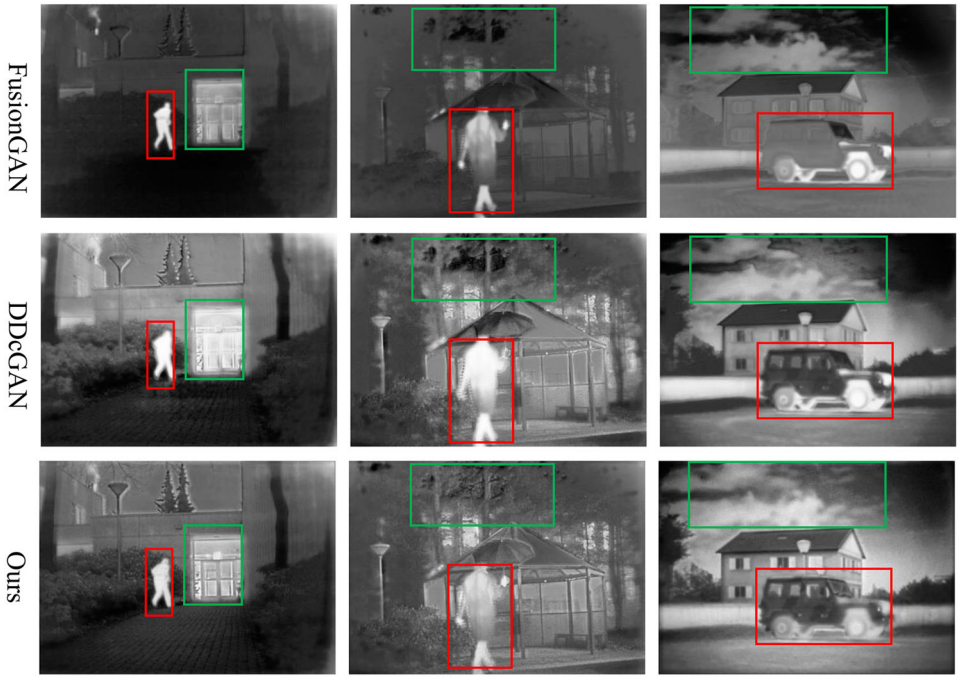
We select 25 infrared and visible light image pairs on the TNO dataset as experimental data, and compare and analyze the MSAt-GAN fusion method with 7 classic and state-of-the-art fusion algorithms to verify the efficiency of our proposed algorithm. All image pairs on the TNO dataset are well-matched, and the resolution of the two source images is the same.

Qualitative comparison: Six typical source images and fusion results on the TNO dataset are shown in Figs. 6 and 7. It can be seen from the Figs. 6 and 7 that our proposed method and the other seven fusion algorithms can achieve good fusion performance. However, compared with other methods, our method has three typical advantages. First, our fusion result has a higher contrast, which can better preserve the thermal target in the infrared source image. The prominent thermal targets in the fusion image can be more conducive to target detection and recognition, which are marked with a red box in Figs. 6 and 7. Second, our fused image has rich texture details, which can better retain the texture background information in the visible light source image. Rich texture detail information can better reflect contour information such as backgrounds, which are marked with green boxes in Figs. 6 and 7. Third, our fused images are clearer and conform to human vision, and thus have good visibility as there is no artifacts and noise.

In contrast, the other seven methods all have different defects. The CBF and GFF fused images have a lot of noise and artifacts, such as the sky in the first image in Fig. 6 and the trees in the second image. The images fused by GTF and MSVD are overall blurry and lack texture detail information, such as the woods in the second image and the ripple details in the third image in Fig. 7. Although the three methods of deep learning have better fusion quality than traditional methods, the images fused by DenseFuse, FusionGAN and DDcGAN all have some texture detail information loss, such as the background information of the door in the first image in Fig. 6, and the ripple in the third image in Fig. 7. It can be seen that the fused image of our method has prominent target background, rich texture detail information, and thus have a better overall subjective vision. This is because the multi-scale features of the multiple receptive fields can better extract a variety of features of the source image, which will produce a richer the texture detail information in the fusion image, and the fusion of the proposed multi-scale deep attention mechanism will make the fused image focus on the distinguishable parts of the source image, and highlight the prospect of the target.

Fig. 6 Qualitative comparison of our MSAt-GAN with 7 state-of-the-art methods on 3 typical infrared and visible image pairs on TNO dataset. From top to bottom: infrared image, visible image, fusion results of CBF [48], GFF [49], GTF [47], MSVD [50], DenseFuse [26], FusionGAN [27], DDcGAN [28] and our MSAt-GAN

Fig. 6 continued



Quantitative comparison: to further prove the effectiveness of the MSAt-GAN method, we use eight objective evaluation indicators to make a quantitative comparison between our method with seven state-of-the-art methods, as shown in Fig. 8. As can be seen from the figures, our fusion algorithm achieves the maximum value on EN, MS-SSIM, VIF and SCD indicators, and achieves the second largest value on SF, CC, SD and Qabf indicators. The difference between SF, CC SD and Qabf is 0.37, 0.83, 9.61 and 0.81 percentage points, respectively, from the maximum value, which is very small. Through analysis, we also see that the CBF algorithm has a large index value on SF, because the fusion image contains a lot of noise, and SF is the index that reflects the gray change rate of the image. In terms of subjective perception, this can be reflected in Figs. 6 and 7. The analysis of these eight indexes shows that our fusion algorithm can retain the foreground target information of the infrared source image and the rich edge and texture information of the visible light image to the maximum. Our fused image has the highest similarity with the two source images, according to the indicators of CC and MS-SSIM. It can also be seen through VIFF and Qabf that the image fused by our algorithm has the best visual effect. The subjective effect reflected by the objective index measurement of the fusion image is also verified in Figs. 6 and 7. In conclusion, through comprehensive analysis, it shows that the fusion quality of MSAt-GAN is better than any other algorithms.

Experimental analysis on INO dataset

To make further verification of the robustness and scalability of the proposed MSAt-GAN method, we select a set of "Parking Snow" infrared and visible light video signals on the INO video dataset, and crop them to generate 2941 frames of infrared and visible light image sequences with a resolution of 448 × 324. In the generated sequences, 20 image pairs are selected as the test dataset, and qualitative and quantitative comparisons are made with 7 state-of-the-art image fusion algorithms. The comparative result is shown in Fig. 9 and Tables 2 and 3.

Figure 9 shows the fusion results of MSAt-GAN and 7 state-of-the-art fusion algorithms at frames 1051, 751 and 851 of the "Parking Snow" dataset. From the figures, we can see that the visible light image contains rich texture detail information, while the infrared image represents the foreground thermal target information. These eight algorithms can all achieve good fusion performance, but our MSAt-GAN has the best fusion quality as its infrared thermal target is prominent and detailed texture information is rich, which can be seen in the 10th row. For example, car marks on snowy roads are much clearer than that of other algorithms. The last row in Fig. 10 shows the target benchmark of the source image, which mainly reflects the thermal target information. Through comparison, it is found that our algorithm can highlight human thermal target information, and key points can be found in the benchmark map. The other seven algorithms have infrared target information loss, and the texture detail information are also relatively fuzzy. For example, there is

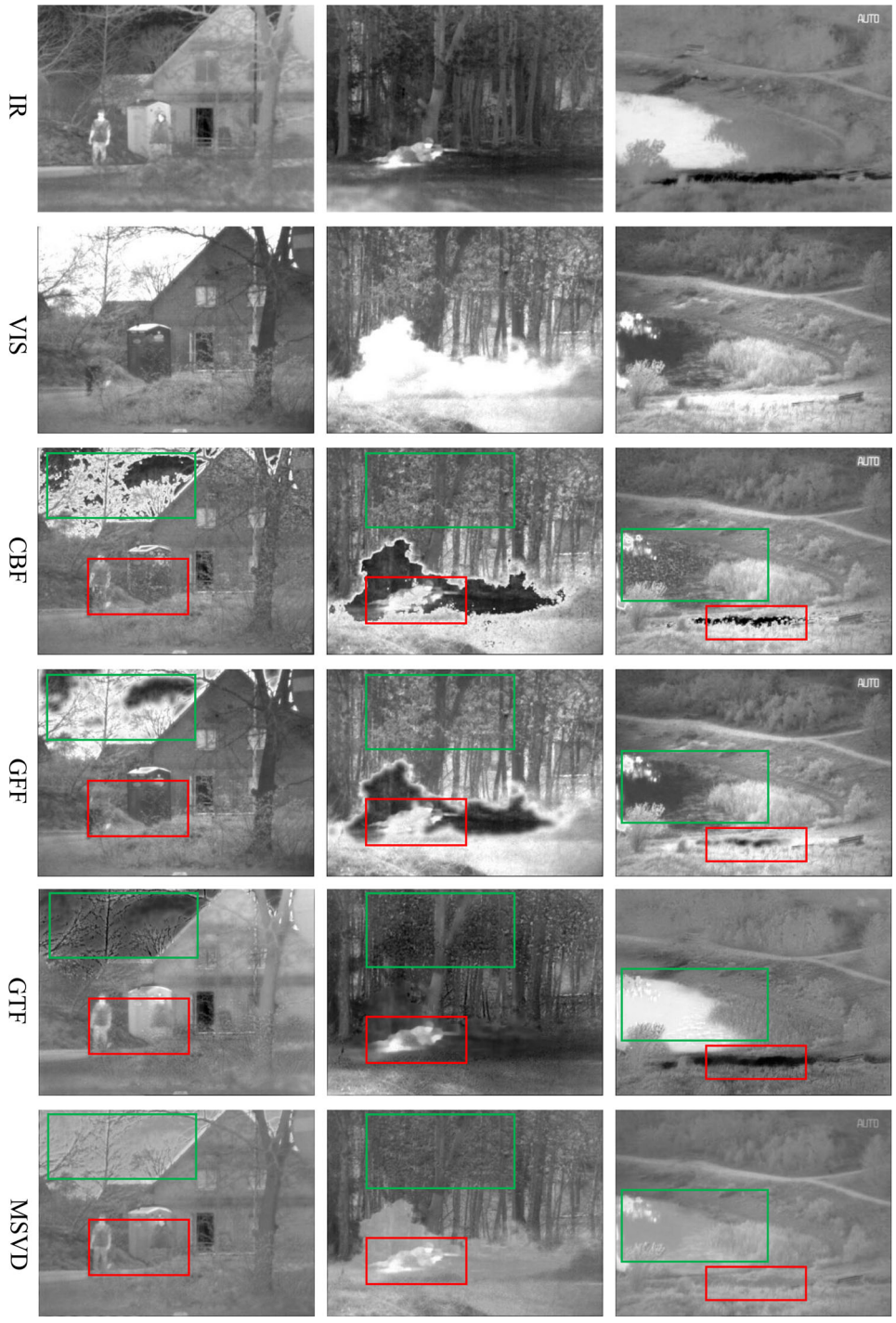
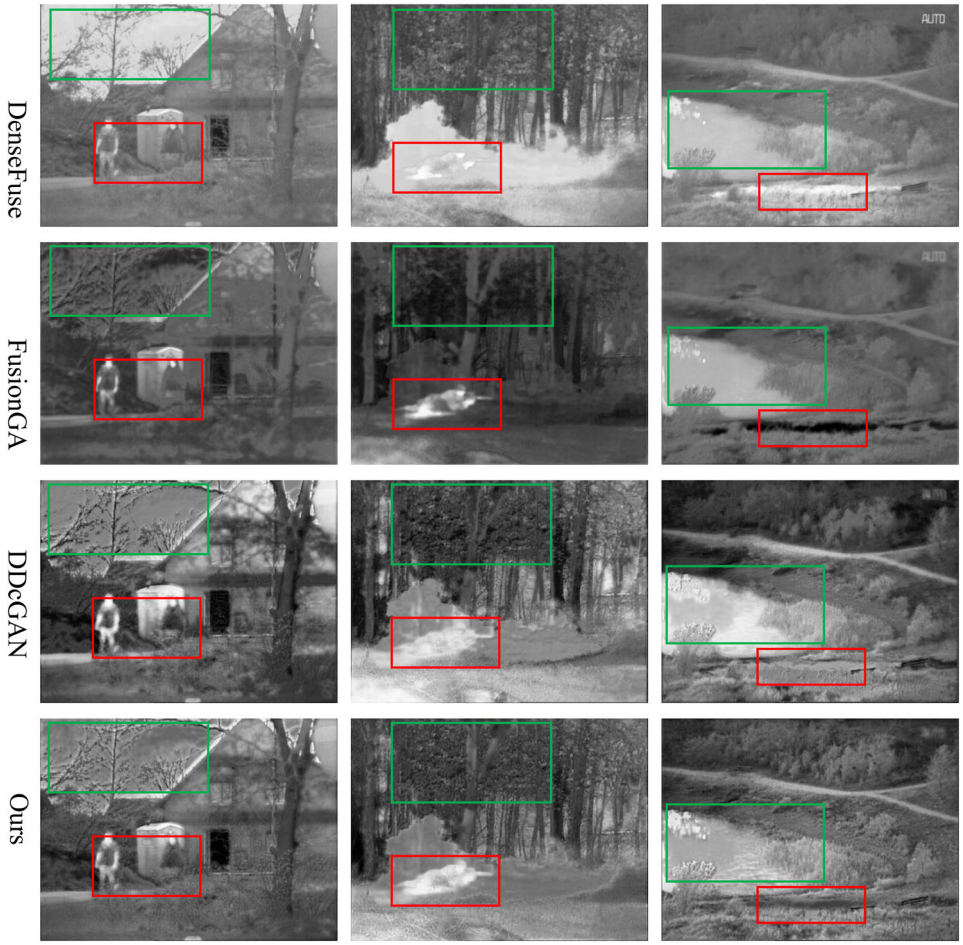


Fig. 7 Qualitative comparison of our MSAt-GAN with 7 state-of-the-art methods on 3 typical infrared and visible image pairs on TNO dataset. From top to bottom: infrared image, visible image, fusion results of CBF [48], GFF [49], GTF [47], MSVD [50], DenseFuse [26], FusionGAN [27], DDcGAN [28] and our MSAt-GAN

infrared target information loss and noise in CBF and GFF images. GTF, MSVD and FusionGAN have relatively low contrast. The image detail information of DenseFuse and DDcGAN are blur.

Table 2 is the evaluation of 20 pairs of infrared and visible image sequences selected on the "Parking Snow" dataset, according to MSAt-GAN and 7 state-of-the-art fusion algorithms. The index value is the mean value of 20 image pairs. Table 3 shows the standard deviations of the eight objective statistical indicators of eight respective algorithms corresponding to Table 2, which reflects the dispersion degree of each algorithm statistical indicator data. For the sake of analysis, the optimal values of the indicators are highlighted in red font, the second in green, and the third in blue in both Tables 1 and 2. For the sake of analysis, the optimal values of the indicators are highlighted in red font, the second

Fig. 7 continued



in green, and the third in blue in both Table 2 and Table 3. From Table 2, we can see that MSAt-GAN fusion algorithm ranks first in SF, CC, SCD and Qabf, second in MS-SSIM, and third in EN, VIF and SD. Specifically, our algorithm is only 0.43 percentage points less than MS-SSIM in GFF. The first-place algorithms are 0.0703, 0.0672 and 0.2341, respectively, higher than ours in EN, VIF and Qabf. In effect, the gap between our algorithm and that of the first place is not big. It can be seen from Table 3 that the standard deviations are very small, in terms of these eight algorithms on the eight respective metrics, indicating that the metrics of eight algorithms are relatively accurate with relatively small errors. Particularly, in terms of standard deviation, the MSAt-GAN algorithm proposed in this paper ranks first on SCD and MS-SSIM, second on CC, third on EN, SF and Qabf, and fourth on VIF and SD which is only 0.46 and 8.52 percentage points less than the first, respectively. It can also be seen that, compared with other advanced algorithms, our eight objective statistical indicators are more accurate and the model has better stability. From the overall objective evaluation index, it shows that the image fused by our proposed algorithm has high contrast, the highest similarity with the source image and good human visual perception effects. In general, MSAt-GAN also can achieve good fusion performance in INO video fusion.

Experimental analysis on Nirscene dataset

To further verify the advantages of our proposed MSAt-GAN method in the fusion of infrared and visible light images, 30 infrared and visible light image pairs on the "Nirscene" dataset are selected as test data. Then we compare and analyze the images among MSAt-GAN and seven advanced fusion algorithms to verify the robustness and efficiency of our proposed algorithm. The 30 selected infrared and visible image pairs are all registered and have the same resolutions of the 2 source images. The subjective effect comparison and objective experimental verification are shown in Figs. 10 and 11.

Figure 10 shows the subjective effect of fused images of MSAt-GAN and seven other fusion algorithms on three typical infrared and visible light image pairs in the "Nirscene" dataset, of which part of the infrared targets and visible light texture of the source images are highlighted in red and green

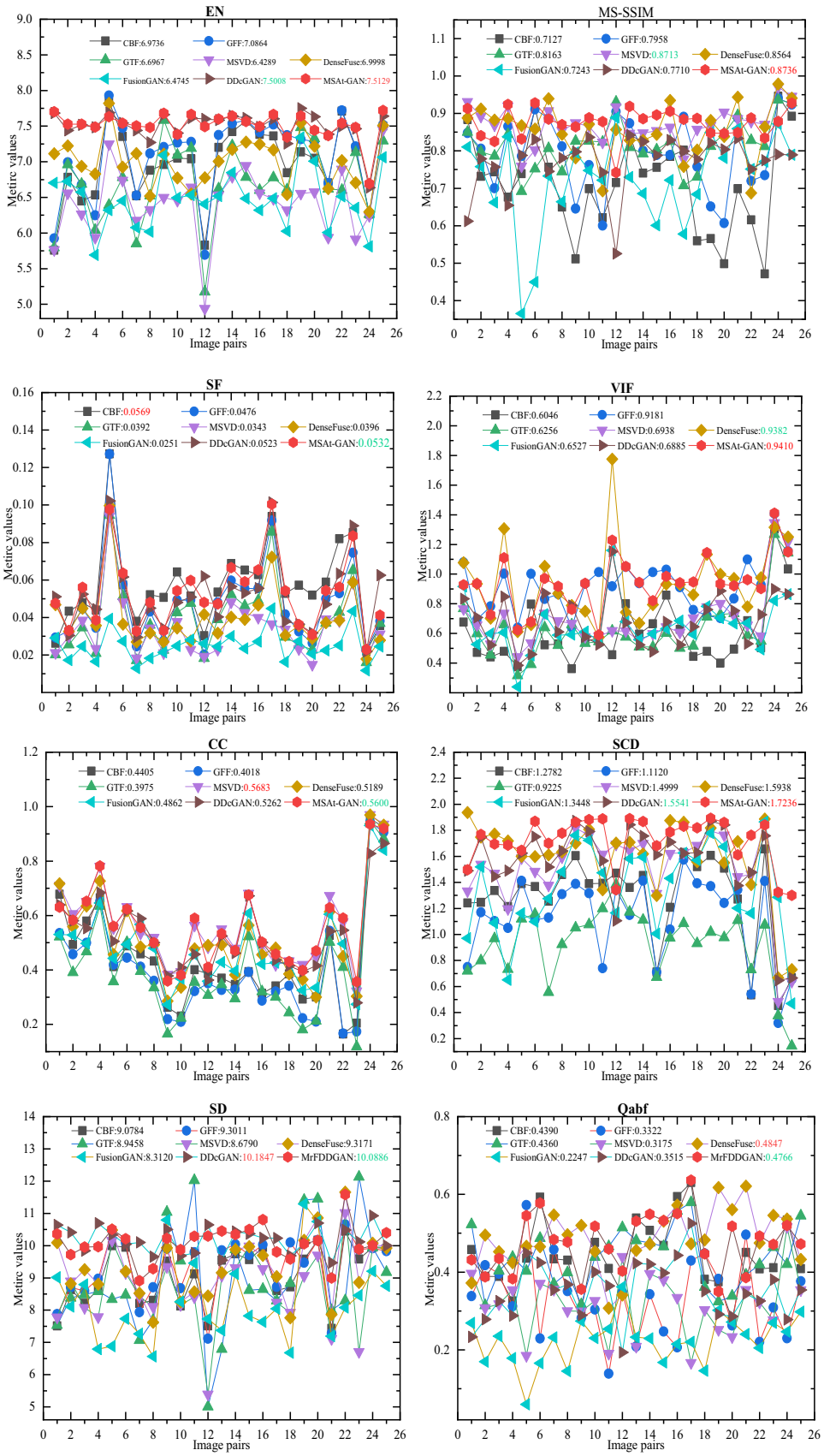


Fig. 8 Quantitative comparison between our MSAt-GAN and 7 state-of-the-art methods on 25 typical infrared and visible image pairs on the TNO dataset. Means of metrics for different methods are shown in the legends. The red font is the maximum value, and the green font is the second largest value

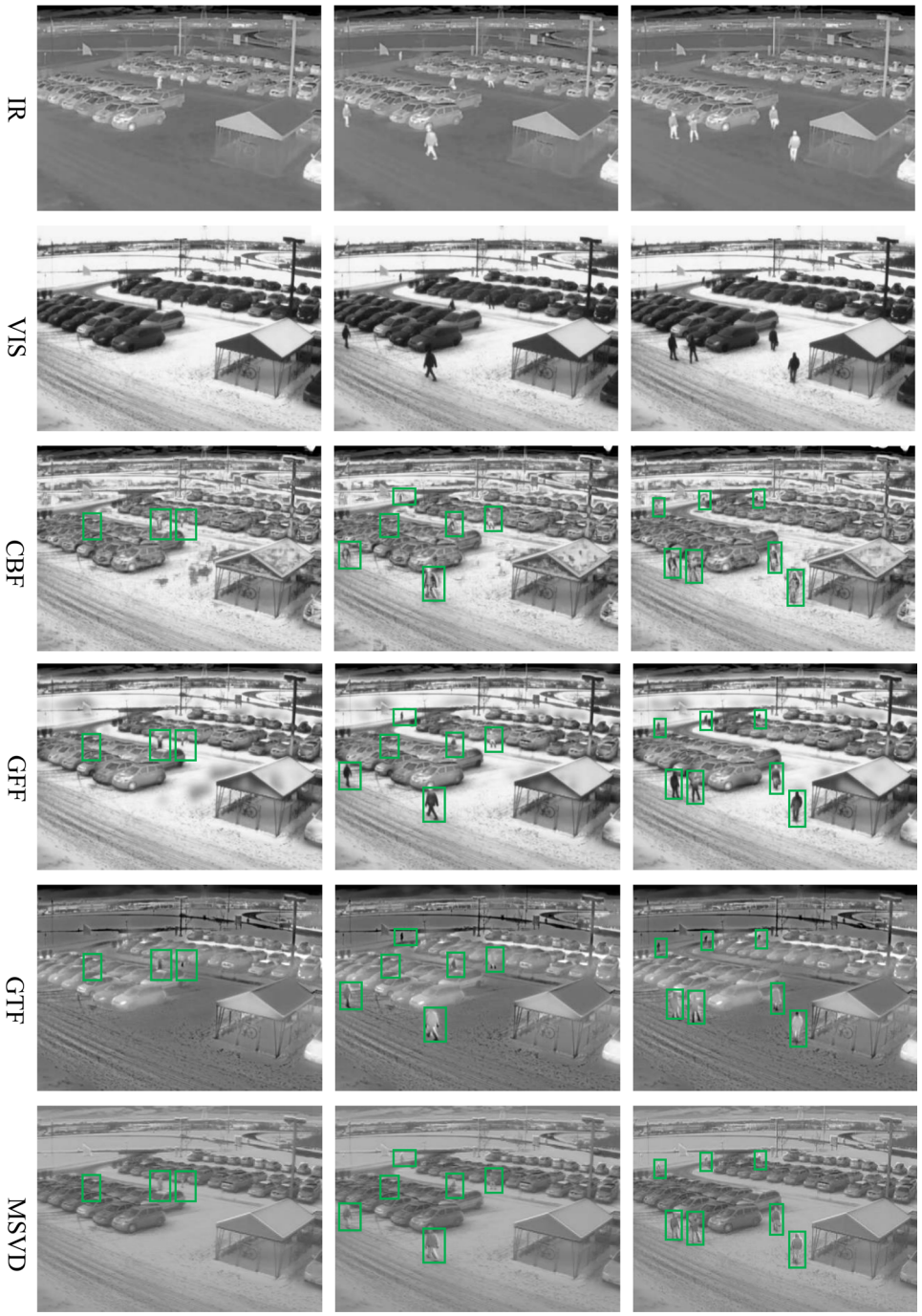


Fig. 9 Qualitative comparison of our MSAt-GAN with 7 state-of-the-art methods on 3 typical infrared and visible image pairs on the INO dataset. The first row shows the infrared source image at frames 1051, 751 and 851, the second rows shows the visible light source images at frames 1051, 751 and 851, the third to ninth row shows the results of CBF, GFF, GTF, MSVD, DenseFuse, FusionGAN, DDcGAN, the tenth row shows the results of the proposed method, and the last row is the ground truth of the moving target

frames, respectively. It can be seen from the Fig. 10 that these eight fusion algorithms can effectively fused infrared and visible image pairs. Yet, the fused images of CBF, GFF, GTF, MSVD, FusionGAN and DDcGAN algorithms is not good enough, for there are some defects of artifacts, noise or other unknown errors. For example, CBF, GFF, and GTF algorithms, there are a lot of noise and artifacts in the trees in the second fusion image and the sky in the third fusion image, which seriously affects the subjective visual effect; on the upper part of the third fused image of MSVD algorithm, there are many unknown vertical striped images. For deep learning and GAN-based algorithms—FusionGAN and DDcGAN, the effect of fused images is not ideal, neither. To be specific, the texture details of the grass are lost and the target of the vehicle is not prominent in the first fused image. Overall, DenseFuse and MSAt-GAN algorithms excel at fusion performance. Compared with the DenseFuse algorithm, the overall contrast of MSAt-GAN fusion image is higher, the

Fig. 9 continued

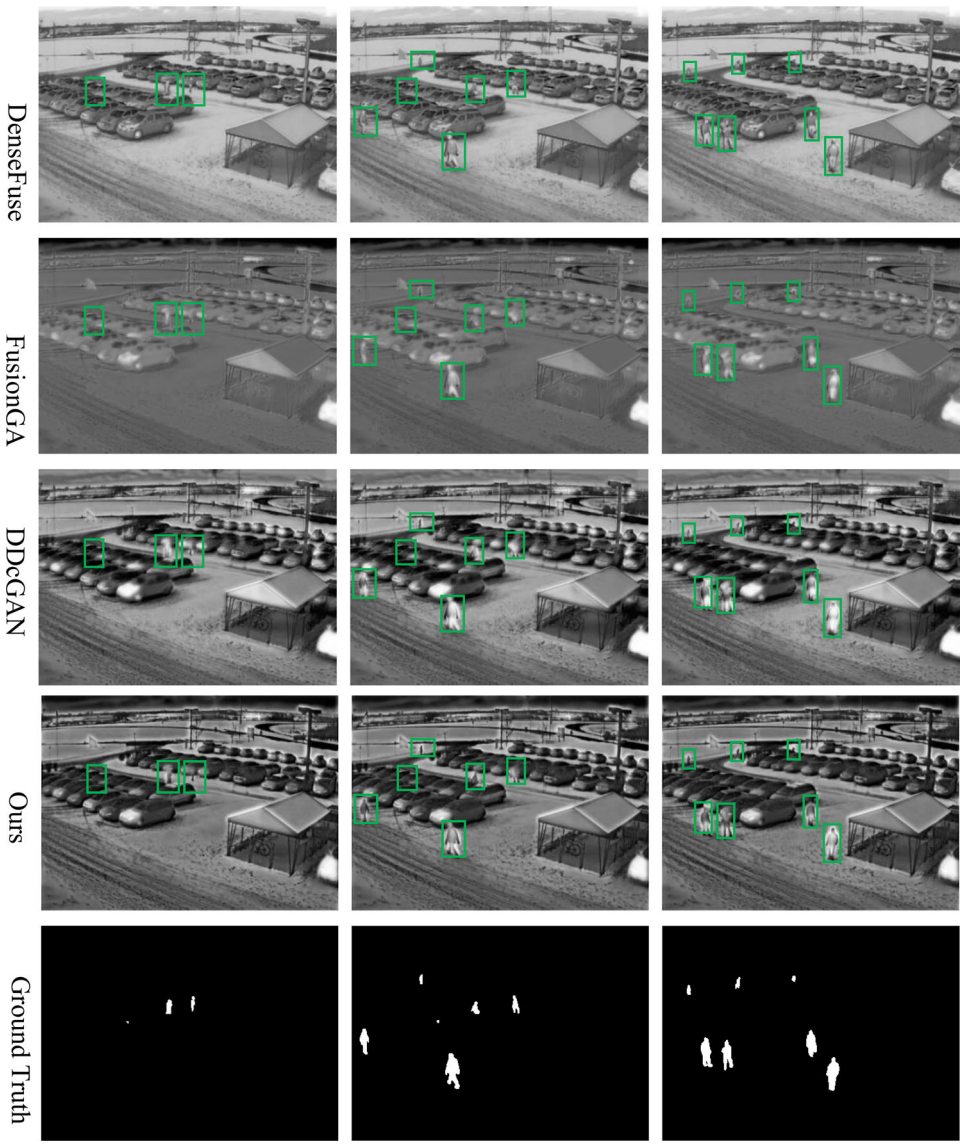


Table 2 The mean of the objective evaluation index of 8 advanced algorithms on the "ParkingSnow" image sequence pairs on the INO dataset

	EN	MS-SSIM	SF	VIF	CC	SCD	SD	Qabf
CBF	*7.5812*	0.7498	*0.0852*	0.4481	**0.3487**	**1.6441**	*11.1668*	0.5268
GFF	***7.6019***	***0.8765***	**0.0770**	*0.6728*	0.3480	1.6394	***11.3146***	**0.5434**
GTF	6.7166	0.5871	0.0597	0.3336	0.2013	1.1497	8.4455	0.4202
MSVD	6.6075	0.7814	0.0604	0.4658	*0.3518*	*1.7540*	10.7577	0.3647
Densefuse	7.2329	**0.8721**	0.0655	***0.6567***	0.2394	1.5584	10.9658	*0.5914*
FusionGAN	6.1160	0.4230	0.0348	0.3557	0.2756	1.4939	7.3390	0.1679
DDcGAN	7.4969	0.7967	0.0753	0.3900	0.3141	1.6338	10.6447	0.4396
Ours	**7.5316**	*0.8731*	***0.0872***	**0.6056**	***0.3732***	***1.7628***	**11.0805**	***0.5956***

The maximum values in the first three digits are highlighted in bolditalic, italic, and bold fonts

Table 3 The standard deviation of the objective evaluation index of 8 advanced algorithms on the "ParkingSnow" image sequence pairs on the INO dataset

	EN	MS-SSIM	SF	VIF	CC	SCD	SD	Qabf
CBF	***0.0141***	0.0118	0.0010	0.0101	0.0111	**0.011**3	*0.0929*	0.0048
GFF	0.0307	0.0076	**0.0008**	0.0113	0.0110	*0.0106*	**0.1195**	0.0050
GTF	*0.0203*	0.0102	*0.0006*	*0.0076*	0.0098	0.0160	***0.0745***	0.0070
MSVD	0.0722	**0.0067**	0.0067	0.0104	0.0088	0.0206	0.2029	0.0701
DenseFuse	0.0472	0.0078	0.0012	0.0131	*0.0057*	0.0190	0.2655	0.0066
FusionGAN	0.0632	0.0076	***0.0004***	***0.0051***	**0.0062**	0.0243	0.1746	*0.0031*
DDcGAN	**0.0234**	*0.0065*	***0.0004***	**0.0079**	***0.0040***	0.0277	0.1747	***0.0029***
Ours	**0.0234**	***0.0064***	**0.0008**	0.0097	*0.0057*	***0.0099***	0.1597	**0.0032**

The maximum values in the first three digits are highlighted in bolditalic, italic, and bold fonts

foreground target is more prominent, the background details are more abundant, and the visual effect is the best. For example, the MSAt-GAN algorithm has richer details in the grass on the left side of the first fused image, and the outline of the vehicle is clearer. Especially after careful observation, the first image fused by the DenseFuse algorithm has low contrast and overall blur. On the whole, the MSAt-GAN algorithm proposed in this paper produces a better subjective effect on the Nirscene dataset.

Figure 11 manifests the objective evaluation of MSAt-GAN and other 7 state-of-the-art fused algorithms on 30 infrared and visible image pairs selected on the "Nirscene" dataset, which is measured by 8 objective indicators. In addition, Fig. 11 shows the mean value of indicators for the abovementioned 30 image pairs, in which the maximum, the second and the third largest value of the indicators are highlighted in red, green, and blue fonts, respectively. From Fig. 11, we can see that the MSAt-GAN fusion algorithm ranks first in MS-SSIM, SF, and CC, second in EN, SCD and SD, and third in VIF and Qabf indicators. Specifically, there is only 0.2009, 0.0188 and 0.6291 less than the first in EN, SCD and SD, respectively, and 0.0467 and 0.0055 less than the first in VIF and Qabf, respectively. Through calculations, although the MSAt-GAN does not always rank first in the evaluation of each indicator, the gap between our algorithm and the first one is very small. The results of these eight objective indicators values are consistent with the subjective performance in Fig. 10. For example, in VIF and Qabf, the values of our algorithm are high, which is proved in Fig. 10 that the three typical fused images of MSAt-GAN excels in subjective effect. As SD reflects image contrast, the three images fused by MSAt-GAN in Fig. 10 show a higher contrast.

Through comprehensive analysis, it can be concluded that the images fused by the MSAt-GAN have good visual effects and the best overall fusion quality. The other seven state-of-the-art algorithms performed slightly worse on the "Nirscene" dataset, and the generalization ability of the model on this infrared and visible light dataset was not strong.

Comparative analysis of time consumption

The TNO and Nirscene dataset consists of images of different resolutions, while INO video dataset captures the same size of images. To make an objective evaluation of the algorithm time complexity, we test 25 infrared and visible light image pairs of TNO, 20 infrared and visible light image sequences of INO and 30 infrared and visible light image pairs of Nirscene. The traditional CBF, GFF, GTF, and MSVD algorithms are tested on the CPU, and the deep learning algorithms of DenseFuse, FusionGAN, DDcGAN and our proposed MSAt-GAN are tested on the GPU for time performance. The results are shown in Table 4. The maximum values of algorithm running time on CPU or GPU are marked with bold font and italics font for the convenience of further analysis.

It can be seen from Table 4 that among the deep learning methods, our MSAt-GAN takes the longest time on the INO dataset, and needs a shorter time than the DenseFuse but a longer time than FusionGAN and DDcGAN on the TNO dataset. The reasons are as follows. To begin with, our algorithm model is more complicated than that of FusionGAN and DDcGAN, both in terms of the width and depth of the network. Next, we introduce a multi-scale deep feature extraction module with multiple receptive fields to better extract deep features. With more receptive fields, it will necessarily increase the model width and depth of the deep neural network as well as model parameters. Finally, the deep multi-scale attention fusion module will also increase the running time of the algorithm. In addition, the reason for DenseFuse algorithm's high runtime complexity is that it adds a artificial fusion strategy to the fusion process, which increases its complexity. Among the traditional algorithms running on the CPU, CBF has the longest running time on the two datasets. This is because it adopts a series of complex decomposition and fusion strategies. In general, the operating efficiency of the algorithm is greatly improved due to the powerful matrix computing capabilities of the GPU. Set

Fig. 10 Qualitative comparison of our MSAt-GAN with 7 state-of-the-art methods on 3 typical infrared and visible image pairs on the Nirscene dataset. From top to bottom: infrared image, visible image, fusion results of CBF [48], GFF [49], GTF [47], MSVD [50], DenseFuse [26], FusionGAN [27], DDcGAN [28] and our MSAt-GAN

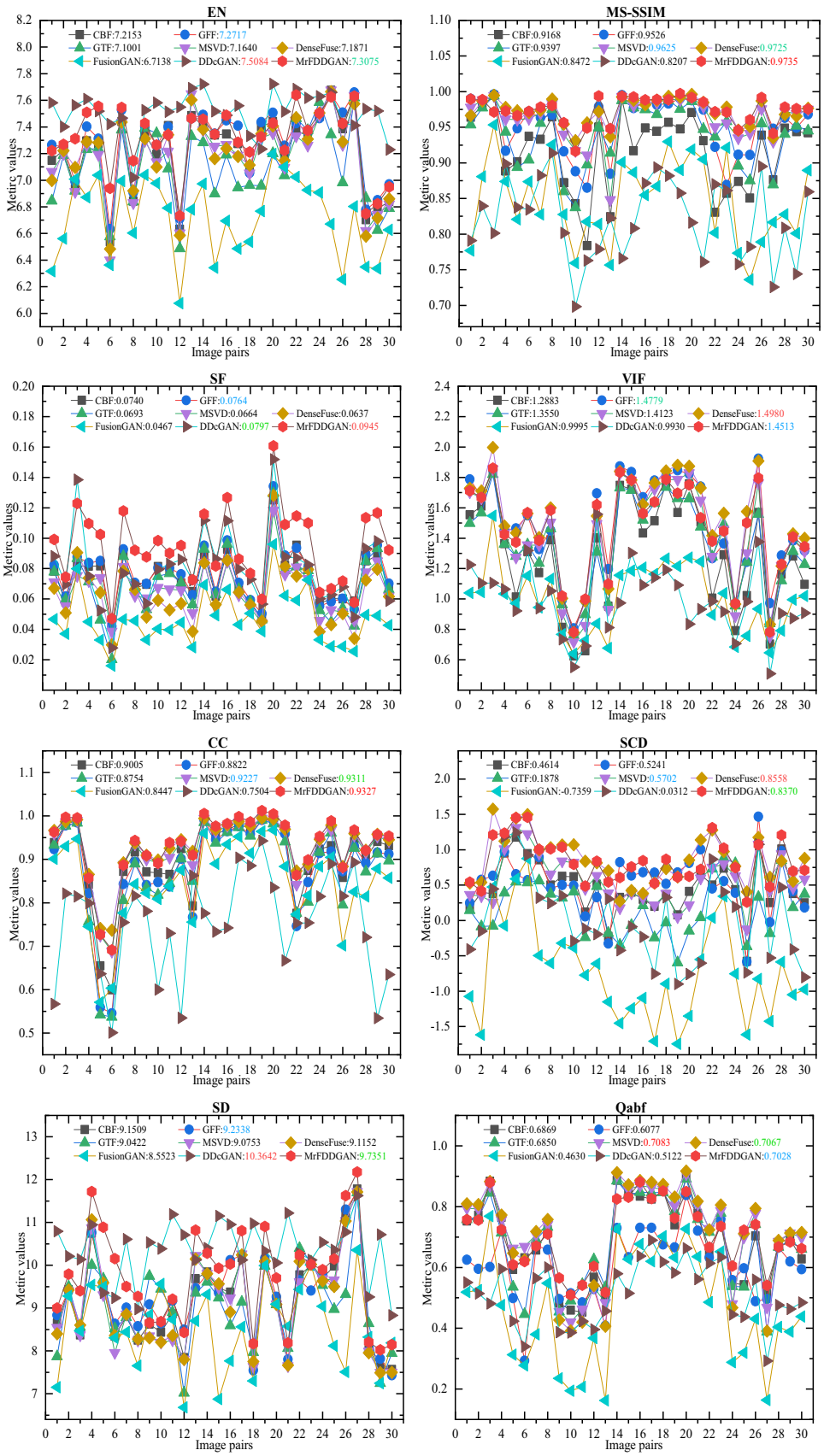


Fig. 11 Quantitative comparison between our MSAt-GAN and 7 state-of-the-art methods on 30 typical infrared and visible image pairs on the Nirscene dataset. Means of metrics for different methods are shown in the legends. The red font is the maximum value, green font is the second largest value, and blue font is the third largest value

Table 4 Run time of different methods (Mean/Std-On the left is the mean value of the running time of the algorithm, and on the right is the standard deviation of the running time of the algorithm)

	TNO dataset	INO dataset	Nirscene dataset
CBF	**14.5583/ 10.8104**	**6.4326/ 0.2280**	**18.5028/2.9578**
GFF	0.5040/0.8712	0.2358/ 1.0543	0.7186/0.1206
GTF	6.9854/ 6.8636	3.1623/ 0.1623	8.6335/1.9861
MSVD	0.5645/ 0.4038	0.2561/ 0.0088	0.6208/0.4002
DenseFuse	*2.4333/ 1.6542*	1.2711/ 1.0720	0.9569/1.0485
FusionGAN	0.8608/ 1.1193	0.1668/ 0.5186	1.0820/1.0695
DDcGAN	1.6407/ 1.4966	0.8844/ 1.301	1.9242/1.6306
Ours	2.3349/ 1.1397	*1.9919/ 1.1747*	*2.4191/1.9475*

hardware environment of GPU and CPU aside, the operating efficiency of MSAt-GAN algorithm is at an intermediate level among the eight fusion methods, and the processing time of each image is between 1 and 2 s, that is, it can basically meet the demands of real-time image fusion applications.

Model ablation study

Ablation study of deep attention mechanism

To prove the effectiveness of the multi-scale deep attention mechanism, we display several multi-level deep attention feature maps in the network, and verify our proposed mechanism through qualitative and quantitative experiments.

As infrared and visible light images are the images captured by two different modal sensors, they show different representations of feature information of the same scene. We introduce a deep attention mechanism in the feature extractor of the encoder that can focus the extracted features on key information, such as the target foreground information in the infrared image and the detailed background information in the visible light image. By introducing the attention mechanism into the model, it is easier to find the difference between the infrared light image and the visible light image, and eliminate the noise. Figure 12 shows the multi-scale attention feature maps of multi-receptive fields.

In Fig. 12, the first row from left to right is: infrared image, visible image and fused image. Starting from the left, the second row corresponds to: the first layer, the second layer and the third layer of attention feature map of the 3×3 receptive field in the encoder network. Likewise, the third row displays the first-layer, the second-layer, and the third-layer attention feature map with the receptive field of 5×5 in the encoder network; the fourth row exhibits the first layer, the second layer, and the third layer of attention feature map of the 7×7 receptive field in the encoder network. It can also be seen from Fig. 12 that by introducing a feature attention mechanism into the encoder network, the feature extractor can better focus attention on the target area information of infrared image and the detailed background information of visible light image, such as the infrared targets information of people and cars, and the detailed texture information of trees.

To verify the fusion effect of the deep attention mechanism, we train the network by comparing whether the attention mechanism is introduced into the encoder network. We select three pairs of infrared and visible light images from the TNO dataset, and make a comparison between MSAt-GAN images with fusion model and that without fusion model. The result is shown in Fig. 13. Figure 13 shows that the images fused by the deep attention mechanism we introduced have prominent infrared thermal target information and rich texture detail information. In Fig. 13, red and green boxes are used to mark the infrared thermal targets and visible texture features, respectively. It is because of the introduction of the attention mechanism that the important features of infrared images and visible images can be represented more accurately. On the contrary, the fused image without the attention mechanism has a poorer fusion effect. Figure 14 shows the quantitative comparison results of a model without the attention mechanism (Without_Att), a deep multi-scale feature extraction encoder model with multi-receptive field (convolution kernel receptive field as 3×3, 5×5 and 7×7) and our MSAt-GAN fusion method. It can be seen from Fig. 14 that our fusion method ranks first in the eight indexes of EN, MS-SSIM, SF, VIF, CC, SCD, SD and Qabf, which shows that our MSAt-GAN has achieved the best fusion performance on these six objective indicators, and it proves the necessity of introducing a deep attention mechanism into the encoder network.

Ablation study of deep multi-scale feature extraction module with multiple receptive fields

To extract the important features of the source image more comprehensively in the encoder network, we introduce a deep multi-scale feature extraction module with multiple receptive fields, and construct three feature extraction channels with the receptive fields of the convolution kernel as 3×3,

Fig. 12 Deep multi-scale feature attention map with multiple receptive fields

5×5 and 7×7. The feature extractor with multi-receptive fields can employ the depth and width of the deep network to expand the receptive field of the convolutional layer. The shallow convolutional layer can only extract local information of the feature, and the deeper the network, the richer the global information extracted. This has been proved in Fig. 12. We find that in the three feature extractors with convolution kernels of 3×3, 5×5 and 7×7, as the network deepens, the extracted global feature information becomes richer. For example, in the feature extraction channel in Fig. 12, there are many infrared target information extracted from the image in the first layer, but relatively few visible light textures. This is because the infrared target information in the source image is more prominent, and have a relative rich local information extracted from the lower layer. With the deepening of the second and third layers of the network, the extracted images not only retain the people and clouds in the infrared image, but also represent the tree in the visible light. The feature extraction process is a process from coarse to fine, from local to global. We can also see from Fig. 12 that in the three feature extractors of 3×3, 5×5 and 7×7, the feature information extracted by the same layer is not the same. This is because the feature extractors with convolution kernels of 5×5 and 7×7, has enlarged the receptive field, compared with that of 3×3 kernel on each layer. In addition, their extracted global information has more features than the smaller convolution kernel. For the multi-receptive field deep multi-scale feature extraction module, the purpose of such combination is to fully extract features on each convolutional layer of the neural network, thus ensuring the diversity and flexibility of feature extraction on each channel.

To verify the fusion effect of the deep multi-scale feature extraction module of multiple receptive fields, we set the feature extractors with convolution kernels of 3×3, 5×5 and 7×7 as independent training models, and keep the other modules of the network the same, and take them as the three verification models. We select three pairs of infrared and visible image pairs from the TNO dataset, and compare MSAt-GAN with the convolution kernel of 3×3, 5×5 and 7×7 fusion models in subjective perception. The results are shown in Fig. 13. It is found from Fig. 13 that the image fused by our method, the retained infrared thermal target information and the visible texture background information are clear. In Fig. 13, the infrared

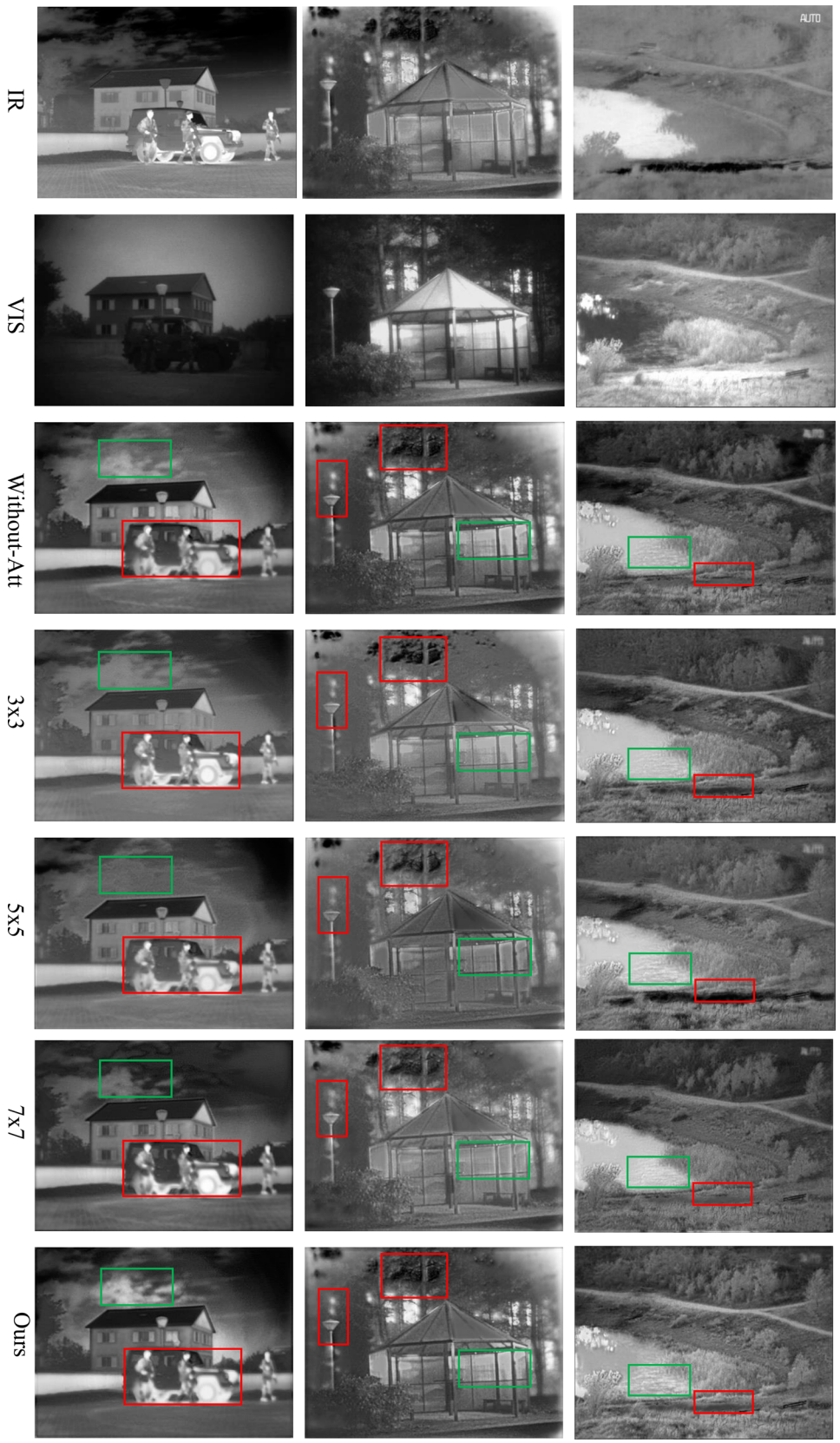


Fig. 13 Qualitative comparisons of ablation analysis on three image pairs from the TNO data set. The source images are shown in the first two rows, followed by the fused images of the model without attention mechanism (Without_Att), multi-scale deep feature extraction module with multiple receptive fields (3×3, 5×5, 7×7) and the fused images of our method

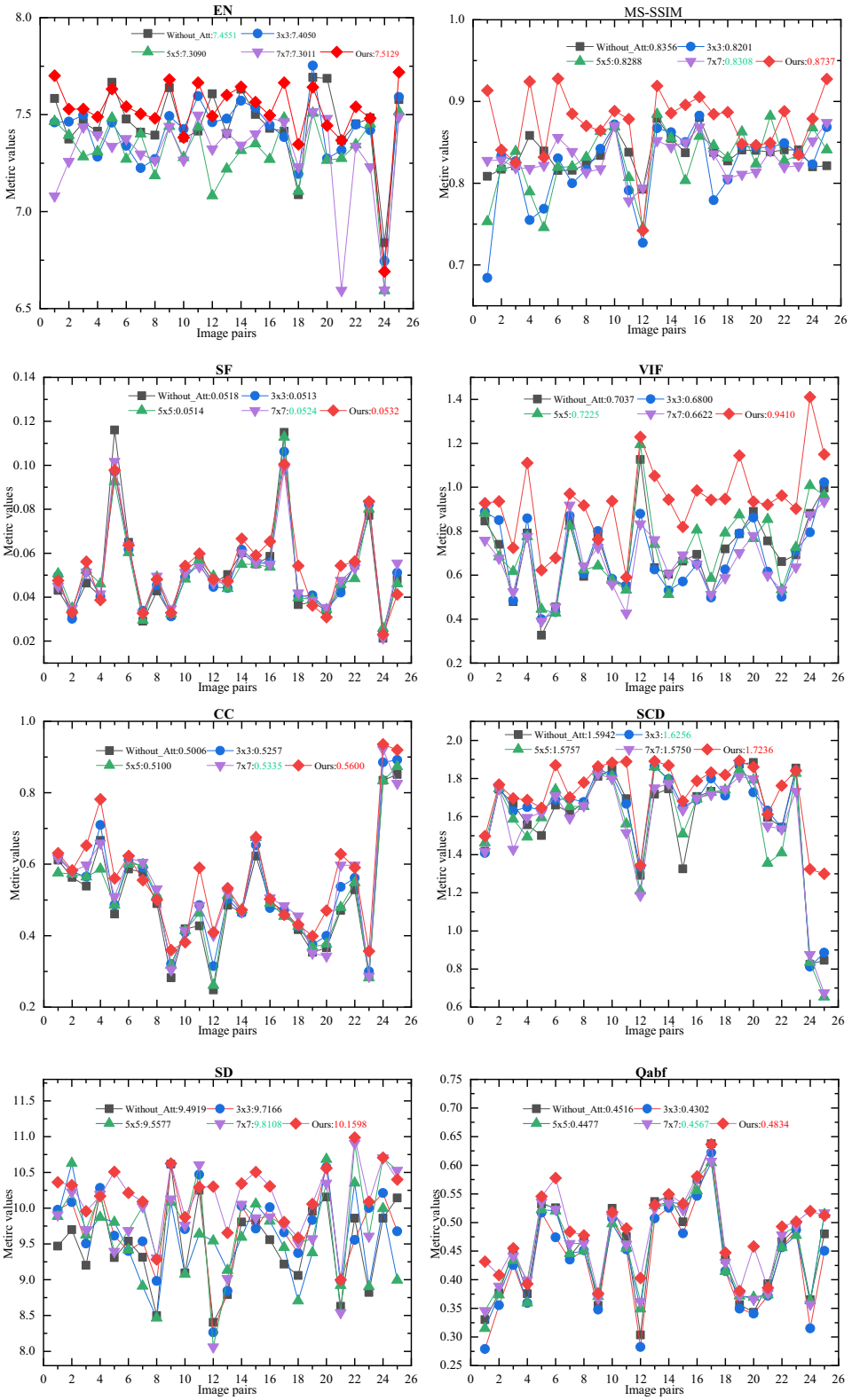


Fig. 14 Quantitative comparisons of ablation analysis on three image pairs from the TNO data set. The source images are shown in the first two rows, followed by the fused images of the model without attention mechanism (Without_Att), multi-scale deep feature extraction module with multiple receptive fields (3×3, 5×5, 7×7) and the fused images of our method

thermal target information and the visible texture feature are marked with a red box and a green box, respectively. To be specific, the infrared target information such as the outlines of cars and people in the first image, and street lights in the second image; the texture background information such as the railing of the house in the second image

and the ripple in the third image. Figure 14 demonstrates the quantitative comparison results of deep multi-scale feature extraction encoder models with multiple receptive fields (convolution kernel receptive field as 3×3, 5×5 and 7×7) and our MSAt-GAN fusion methods, based on the 25 pairs of infrared and visible image datasets selected on TNO. It can be seen from Fig. 14 that our fusion method is on the top in the eight indicators of EN, MS-SSIM, SF, VIF, CC and SCD, which indicates that our MSAt-GAN achieves the best fusion performance in the eight different indexes. From subjective and objective perspective, it is obvious that the introduction of deep multi-scale feature extraction modules with multi-receptive fields can enhance infrared target information and enrich texture information, which further proves the necessity of introducing a deep multi-scale feature extraction module with multiple receptive fields into the encoder network.

Conclusions

In the paper, we propose a generative adversarial network-based on multi-scale feature transmission and a deep attention mechanism for infrared and visible light image fusion. The deep multi-scale feature extraction module of the multi-receptive field we introduce in the generator-encoder network can effectively and comprehensively extract the global information and deep features of the source image. In addition, the deep attention mechanism introduced in the encoder network can make the model put more emphasis on the important features of the source image, such as the important part of the difference distinguishable between infrared image and visible image. What is more, we concatenate the fused multi-scale attention features to the decoder network of the generator, which can complete the feature transfer effectively, avoid detail loss in the decoding process and suppress the noise interference. The dense connection in all layers of the generator-encoder network can help make full use of multi-scale features and strengthen the mapping relationship between different scale features. This paper also introduces a new gradient penalty term to strengthen the Lipschitz constraint, which improves the training performance and stability of the MSAt-GAN model.

Through extensive qualitative and quantitative experiments and ablation research analysis on three public infrared and visible light image datasets, the advantages of MSAt-GAN proposed in this paper in infrared and visible light image fusion are proved. However, there are still limitations of our proposed algorithm: (1) the hyperparameters of the model loss function in most image fusion methods are determined by the empirical value and experimental research of other relevant literature, which may face the problem of model tuning. In addition, this is the same problem of this paper in terms of hyperparameters of the loss function of the MSAt-GAN model. (2) It should be noted that MSAt-GAN model only verifies the efficiency of infrared and visible light image fusion, but there is a certain correlation between each image fusion task, so its generality is insufficient.

Therefore, future studies should lay emphasis on finding a reasonable design of a certain indicator function to adaptively determine the proportional relationship between various hyperparameters. In future research, we will extend the MSAt-GAN model to other fusion tasks, such as multi-focus, multi-exposure, and medical image fusion, by designing a general loss function or adjusting only the hyperparameter values in the existing loss function.

Acknowledgements This work was supported by the National Natural Science Foundation of China under Grant 11673009. The authors would like to thanks Prof. K. Ji from Yunnan Observatory, Chinese Academy of Sciences, for their valuable comments and suggestions for this study.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Li S, Kang X, Fang L, Hu J, Yin H (2017) Pixel-level image fusion: a survey of the state of the art. Inf Fusion 33:100–112
2. Li C, Liang X, Lu Y, Zhao N, Tang J (2019) RGB-T object tracking: Benchmark and baseline. Pattern Recognit 96:106977
3. Kristan M et al (2019) The seventh visual object tracking vot2019 challenge results. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), IEEE, pp 1–36
4. Zhang X, Ye P, Leung H, Gong K, Xiao G (2020) Object fusion tracking based on visible and infrared images: a comprehensive review. Inf Fusion 63:166–187
5. Duan Z, Lan J, Xu Y, Ni B, Zhuang L, Yang X (2017) Pedestrian detection via bi-directional multi-scale analysis. In: Proceedings of the 25th ACM international conference on Multimedia. ACM, pp 1023–1031

6. Sun K, Zhang B, Chen Y et al (2021) The facial expression recognition method based on image fusion and CNN. Integr Ferroelectr 217(1):198–213
7. Xu J, Lu K, Wang H (2021) Attention fusion network for multi-spectral semantic segmentation. Pattern Recognit Lett 146:179–184
8. Kuanar S, Athitsos V, Mahapatra D, Rao K, Akhtar Z, Dasgupta D (2019) Low dose abdominal CT image reconstruction: an unsupervised learning based approach. In: 2019 IEEE International Conference on Image Processing (ICIP). IEEE, pp 1351–1355
9. Liu R, Mu P, Chen J, Fan X, Luo Z (2020) Investigating task-driven latent feasibility for nonconvex image modeling. IEEE Trans Image Process 29:7629–7640
10. Li J, Huo H, Li C et al (2020) Multi-grained attention network for infrared and visible image fusion. IEEE Trans Instrum Meas 70:1–12
11. Ma J, Ma Y, Li C (2019) Infrared and visible image fusion methods and applications: a survey. Inf Fusion 45:153–178
12. Liu Y, Chen X, Wang Z, Wang ZJ, Ward RK, Wang X (2018) Deep learning for pixel-level image fusion: recent advances and future prospects. Inf Fusion 42:158–173
13. Junwu L, Li B, Jiang Y (2020) An infrared and visible image fusion algorithm based on LSWT-NSST. IEEE Access 8:179857–179880
14. Vishwakarma A, Bhuyan MK (2018) Image fusion using adjustable non-subsampled shearlet transform. IEEE Trans Instrum Meas 68(9):3367–3378
15. Jin X, Jiang Q, Yao S et al (2018) Infrared and visual image fusion method based on discrete cosine transform and local spatial frequency in discrete stationary wavelet transform domain. Infrared Phys Technol 88:1–12
16. Chen J, Li X, Luo L et al (2020) Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. Inf Sci 508:64–78
17. Liu CH, Qi Y, Ding WR (2017) Infrared and visible image fusion method based on saliency detection in sparse domain. Infrared Phys Technol 83:94–102
18. Zhang Q, Liu Y, Blum RS, Han J, Tao D (2018) Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review. Inf Fusion 40:57–75
19. Ma J, Zhou Z, Wang B, Zong H (2017) Infrared and visible image fusion based on visual saliency map and weighted least square optimization. Infrared Phys Technol 82:8–17
20. Li J, Huo H, Sui C, Jiang C, Li C (2019) Poisson reconstruction-based fusion of infrared and visible images via saliency detection. IEEE Access 7:20676–20688
21. Rasti B, Ghamisi P (2020) Remote sensing image classification using subspace sensor fusion. Inf Fusion 64:121–130
22. Singh S, Anand RS (2019) Multimodal medical image sensor fusion model using sparse K-SVD dictionary learning in nonsubsampled shearlet domain. IEEE Trans Instrum Meas 69(2):593–607
23. Zhang H, Xu H, Xiao Y et al (2020) Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity. Proc AAAI Conf Artif Intell (AAAI) 34(07):12797–12804
24. Liu Y, Chen X, Cheng J et al (2018) Infrared and visible image fusion with convolutional neural networks. Int J Wavelets Multiresolut Inf Process 16(03):1850018
25. Li H, Wu X-J, Kittler J (2018) Infrared and visible image fusion using a deep learning framework. In: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, pp 2705–2710
26. Li H, Wu X-J (2018) DenseFuse: a fusion approach to infrared and visible images. IEEE Trans Image Process 28(5):2614–2623
27. Ma J, Yu W, Liang P et al (2019) FusionGAN: a generative adversarial network for infrared and visible image fusion. Inf Fusion 48:11–26
28. Ma J, Xu H, Jiang J et al (2020) DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Trans Image Process 29:4980–4995
29. Ma J, Zhang H, Shao Z et al (2020) GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion. IEEE Trans Instrum Meas 70:1–14
30. Li J, Huo H, Li C et al (2020) AttentionFGAN: infrared and visible image fusion using attention-based generative adversarial networks. IEEE Trans Multimed 23:1383–1396
31. Huang G, Liu Z, Van Der Maaten L, Weinberger KQ (2017) Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 4700–4708
32. Li H, Wu XJ, Durrani T (2020) NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans Instrum Meas 69(12):9645–9656
33. Liu J, Fan X, Jiang J et al (2021) Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion. In: IEEE Transactions on Circuits and Systems for Video Technology
34. Szegedy C, Vanhoucke V, Ioffe S et al (2016) Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 2818–2826.
35. Petzka H, Fischer A, Lukovnicov D (2017) On the Regularization of Wasserstein Gans, arXiv preprint 1709 (08894)
36. Li J, Huo H, Liu K et al (2020) Infrared and visible image fusion using dual discriminators generative adversarial networks with Wasserstein distance. Inf Sci 529:28–41
37. Vaswani A, Shazeer N, Parmar N et al (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
38. Zhong Z et al (2020) Squeeze-and-Attention Networks for Semantic Segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 13062–13071
39. Ulutan O, Iftekhar ASM, Manjunath BS (2020) VSGNet: Spatial Attention Network for Detecting Human Object Interactions Using Graph Convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp 13614–13623
40. Yang Q, Xu Y, Wu Z et al (2019) Hyperspectral and multispectral image fusion based on deep attention network. In: 2019 10th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS). IEEE, pp 1–5
41. Zhu H, Ma W, Li L et al (2020) A Dual-Branch Attention fusion deep network for multiresolution remote-Sensing image classification. Inform Fusion 58:116–131
42. He K, Zhang X, Ren S et al (2016) Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). IEEE, pp 770–778.
43. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In: Proceedings of the IEEE international conference on computer vision (ICCV). IEEE, pp 1026–1034
44. Wang Z, Li Q (2010) Information content weighting for perceptual image quality assessment. IEEE Trans Image Process 20(5):1185–1198
45. Han Y, Cai Y, Cao Y, Xu X (2013) A new image fusion performance metric based on visual information fidelity. Inf Fusion 14(2):127–135
46. Aslantas V, Bendes E (2015) A new image quality metric for image fusion: the sum of the correlations of differences. Aeu-Int J Electron Commun 69(12):1890–1896

47. Ma J, Chen C, Li C, Huang J (2016) Infrared and visible image fusion via gradient transfer and total variation minimization. Inf Fusion 31:100–109
48. Shreyamsha BK (2015) Image fusion based on pixel significance using cross bilateral filter. Signal Image Video Process 9(5):1193–1204
49. Li S, Kang X, Hu J (2013) Image fusion with guided filtering. IEEE Trans Image Process 22(7):2864–2875
50. Naidu VPS (2011) Image fusion technique using multi-resolution singular value decomposition. Def Sci J 61(5):479–484