Electronic Research Archive

https://www.aimspress.com/journal/era

ERA, 33(9): 5471–5495.
DOI: 10.3934/era.2025245
Received: 05 June 2025
Revised: 06 September 2025
Accepted: 09 September 2025
Published: 15 September 2025

Research article

Dual-channel fusion and dual-discriminator GAN for infrared and visible image fusion

Qianying Wang, Haiyan Xie*and Huimin Qu

School of Science, Dalian Maritime University, Dalian 116026, China

* **Correspondence:** Email: xiehy@dlmu.edu.cn.

Abstract: The fusion of infrared (IR) and visible (VIS) images aims to synthesize fused images with salient targets and enriched details. However, existing fusion methods face challenges in integrating modality-specific features. Accordingly, we proposed an image fusion method based on a dual-channel fusion strategy, termed DCGAN-Fuse. First, we created a dual-channel fusion strategy and constructed a dual-channel fusion module (DCFM) to integrate shared and complementary information across both modalities. Second, during the feature enhancement phase, we designed an attention-enhanced gradient retention module (AEGRM) to enhance edge feature extraction and enforce spatial consistency. Moreover, we used the multi-scale module (MSM) to capture fused features and avoid information loss from the two source images. We have improved the loss function by introducing the maximum intensity loss function for our proposed method. Experiments on public datasets demonstrated that our method generates fused images with highlighted infrared targets and enriched textures. Both subjective and objective assessments indicated that our DCGAN-Fuse is better than the other thirteen advanced algorithms.

Keywords: image fusion; infrared image; visible image; dual-channel fusion; dual-discriminator

1. Introduction

Visible (VIS) and infrared (IR) images exhibit different imaging principles, leading to their complementary information characteristics. VIS images contain rich scene details and texture information but are easily affected by adverse weather conditions. IR images contain high-contrast targets but lack comprehensive background details. The fusion of IR and VIS images effectively integrates their complementary advantages, and fused images showcase pronounced thermal targets and well-defined textures. Consequently, they have been widely used in image enhancement [1], target recognition [2], detection [3], tracking [4], agricultural automation [5], remote sensing [6], and other domains [7,8]. [7] applied them to accurate face recognition, and [8] extended to RGB-D image segmentation.

Current fusion methods can be broadly classified into two categories: traditional methods and deep learning-based methods. Traditional infrared and visible image fusion methods include multi-scale transformation-based methods [9], sparse representation-based methods [10], saliency-based methods [11], subspace-based methods [12], mixed-based methods [13], and others [14–16]. [14] transformed image fusion into an optimization problem. [15] used a contrast-based fusion approach in the discrete cosine transform (DCT) domain. [16] constructed four-tap size-limited filter banks to ensure lossless reconstruction during the multi-scale decomposition process. However, traditional fusion methods relying on manually designed feature extraction and fusion rules find it difficult to comprehensively consider all aspects and details in the fusion process.

In contrast to traditional methods, deep learning methods have better information expression ability [17]. Deep learning-based fusion techniques are primarily categorized by their network architectures, such as autoencoder (AE), convolutional neural network (CNN), generative adversarial network (GAN), and transformer. For CNN-based methods, Prabhakar et al. [18] designed a network with two convolutional layers and a feature reconstruction layer with three convolutional layers. However, shallow architectures struggle to capture high-level semantic features. [19] proposed a novel image fusion framework based on residual networks and zero-phase component analysis (ZCA) to generate feature weight maps, yet this method still inherits biases from manual rule design. Li et al. [20] used dense connections combined with the L_1 norm and softmax fusion rules to extract deep features and apply a fusion strategy. It is a typical AE-based fusion method.

Addressing the absence of ground truth in fusion tasks, GAN-based methods leverage adversarial training to refine output quality. Ma et al. [21] first utilized a GAN framework to preserve infrared thermal radiation and visible textures simultaneously. Nevertheless, a single discriminator often fails to balance multi-modal distributions. To address the imbalance issue, several dual-discriminator-based fusion frameworks have been proposed, such as [22, 23]. Transformer-based methods are gradually applied to the image fusion field, such as Swin Transformer [24] and DATFusion [25]. A convolution-guided transformer framework for visible and infrared image fusion was proposed by Li et al. [26]. In their approach, local features were computed via a convolutional feature extraction module, which was used to guide the transformer-based feature extraction module. Tang et al. [27] proposed a Y-shaped dynamic transformer. They employed Y-shaped branches for image feature extraction and designed a dynamic transformer module (DTRM) to capture local features and critical contextual information. Hu et al. [28] proposed a feature network with a dual-branch PoolFormer-CNN. In the PoolFormer blocks, basic spatial pooling is used to replace the attention module of the transformer for extracting low-frequency global information. However, these methods fail to consider the interactive information between different modalities. Tang et al. [29] took into account the information interaction in the fusion process of infrared and visible light, and proposed a cross-modal interactive transformer. They also designed cross-modal attention and transformer blocks to integrate features and establish multi-modal long-range relationships. These methods can better capture the global dependencies in images, thereby improving fusion performance.

Although these methods have achieved good fusion images, there are still some problems, such as inadequate retention, detail loss, and insufficient fusion.

Considering the above problems, this paper proposes a method based on the dual-channel fusion strategy and dual-discriminator for IR and VIS image fusion.

In summary, the main contributions of the proposed method are as follows.

- We present a novel generative adversarial network framework based on the dual-channel fusion strategy and dual discriminators. The dual-channel strategy ensures the full fusion of cross-modal information, while the dual discriminators collaboratively guide the generator to produce fused images with balanced brightness, prominent thermal targets, and rich background textures.
- We propose the dual-channel strategy, which reconstructs the working logic of the key link of fusion. During the feature fusion stage, we construct common features and differential features, and then combine them with edge features. By using this method, we effectively address the issue of incomplete information fusion and provide new insights for feature fusion.
- In different feature processing stages, we design three modules with positive synergistic effects, namely the attention-enhanced gradient retention module (AEGRM), dual-channel fusion module (DCFM), and multi-scale module (MSM). The synergy between these modules establishes a progressive processing mechanism of "first strengthening key features, then accurately fusing features, and finally extracting features at full scales". This mechanism enables the sufficient fusion and reconstruction of features from different modalities, thereby improving the model's fusion performance.
- Experiments on three mainstream datasets show that the proposed approach outperforms the other thirteen state-of-the-art methods in both quantitative and qualitative evaluations.

In Section 2, we discuss the Wasserstein generative adversarial network (WGAN). In Section 3, we describe the formulation of the problem and detail the proposed network architecture and loss function. In Section 4, we present experimental comparisons with state-of-the-art methods, including both qualitative and quantitative evaluations. Finally, the conclusions are given in Section 5.

2. Related work

In this section, we introduce the WGAN.

2.1. Wasserstein generative adversarial network (WGAN)

In 2017, Arjovsky et al. proposed WGAN [30]. Unlike the original GAN, WGAN replaces the Jensen-Shannon (JS) divergence with the Wasserstein distance to measure the similarity between two probability distributions. This replacement addresses the vanishing gradient issue in GAN training. To enforce the 1-Lipschitz continuity constraint on the network, WGAN employs weight clipping. The objective function of WGAN is formulated as follows:

$$\begin{aligned} \max_{D\sim 1-Lipschitz} \{ & E_{x\sim p_{data}}[D(x)] \\ & - E_{z\sim p_z}[D(G(z))]\}. \end{aligned} \tag{2.1}$$

However, in WGAN, the discriminator's weights are clipped to a fixed interval to enforce the 1-Lipschitz constraint, which easily leads to the gradient vanishing or gradient explosion in the model. Therefore, Petzka et al. [31] proposed a more moderate regularization term, instead of applying hard-weight clipping. The modified objective function incorporates a gradient penalty term, leading to more

stable training. The improved WGAN objective function for the discriminator is formulated as follows:

$$\begin{aligned}\max_{D} \mathbb{E}_{x\sim p_{\text{data}}}[D(x)] - \mathbb{E}_{z\sim p_z}[D(G(z))] \\ - \lambda \mathbb{E}_{\hat{x}\sim p_{\hat{x}}}\left[\left(\max\{0, \|\nabla_{\hat{x}} D(\hat{x})\|_2 - 1\}\right)^2\right],\end{aligned} \tag{2.2}$$

where the first two terms estimate the Wasserstein distance via the Kantorovich-Rubinstein duality, and the third term is the Lipschitz penalty proposed by Petzka et al. [31]. Here, $\hat{x}$ is sampled uniformly along straight lines between real x_{real} and generated x_{gen} data pairs,

$$\hat{x} = \epsilon x_{\text{real}} + (1-\epsilon)x_{\text{gen}}, \quad \epsilon \sim U[0,1], \tag{2.3}$$

where λ controls the penalty strength for violating the 1-Lipschitz constraint.

3. Our model

In this section, we comprehensively describe our model. First, we present the problem formulation of our method. Then, we elaborate on the detailed structure of the generator and discriminator. Finally, we define the loss functions for the generator and discriminator.

3.1. Problem formulation

We propose a dual-discriminator generative adversarial network based on a dual-channel fusion strategy to comprehensively capture and fuse information from infrared and visible images. The whole network includes one generator G and two discriminators (D_{ir}, D_{vis}). Figures 1 and 2 show the specific training and testing processes. During training, the generator receives infrared images I_{ir} and visible images I_{vis} and generates fused images I_f through our designed network architecture. Each discriminator operates on either I_{ir} or I_{vis} to ensure the fused results retain comprehensive information from the source images through adversarial training with G. Specifically, within the generator, I_{ir} and I_{vis} are fed into residual blocks for feature extraction. The multi-level features extracted by each residual block undergo channel separation at different levels, while the remaining features continue to be further processed through deeper residual blocks for deeper feature extraction. Subsequently, the infrared and visible features are input into the DCFM. The dual-channel fusion strategy thoroughly fuses information from different modalities while avoiding detail loss. Within the fusion module, we design two branches to extract differential features and common features, respectively. These two types of features are then fused with spatial detail features refined through a gradient operator and an attention module. Finally, the fused features are processed through the MSM and the 1×1 convolutional layer, ultimately reconstructing I_f. In our network, we employ two structurally identical discriminators, D_{ir} and D_{vis}, to distinguish between I_f and source images of the corresponding modality, utilizing the Wasserstein loss as the adversarial loss for both the discriminators and the generator. Additionally, we construct the generator's loss function by combining adversarial loss L_{adv}, intensity loss L_{int}, perceptual loss L_{prep}, mean loss L_{mea}, and L_1 loss to constrain the generator in producing high-quality fused images. More specific details regarding the architectures of the generator, discriminators, and the formulation of loss functions can be found in Section 3.

In the testing phase, the trained generator generates fused images I'_f by processing input infrared images I'_{ir} and visible images I'_{vis}.

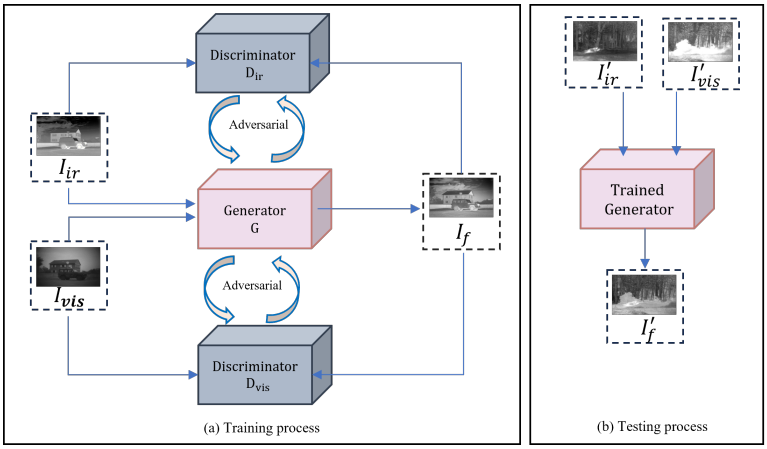

Figure 1. Overview of our method.

3.2. *Generator*

As shown in Figure 2, the infrared and visible images are fed as inputs to G, which outputs the final fused image. The generator consists of residual blocks (RBs), the attention-enhanced gradient retention module (AEGRM), the dual-channel fusion module (DCFM), the multi-scale module (MSM), and feature reconstruction layers. The generator comprises two parallel branches with residual blocks to extract dominant features from infrared and visible images, respectively. The DCFM is designed to fully integrate complementary and common features from both modalities. The AEGRM is employed to reinforce edge features. Subsequently, the MSM and the last two convolutional layers are utilized to extract fused features and reconstruct the fused image.

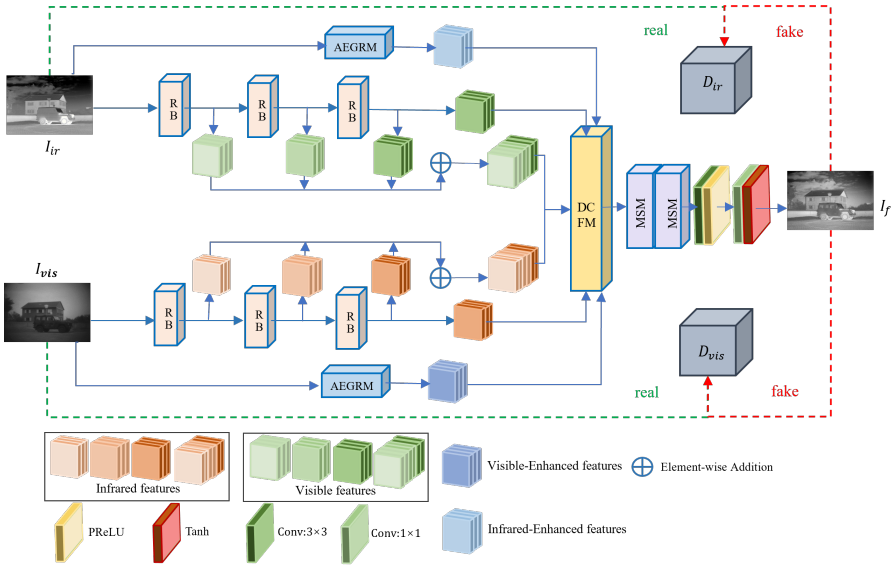

Figure 2. Framework of DCGAN-Fuse.

3.2.1. Design of different feature extraction stages

To comprehensively extract information from infrared and visible images, we create RBs, the MSM, and the AEGRM at distinct feature extraction stages. Specifically, RBs are adopted to extract primary features from infrared and visible images, the AEGRM enhances feature representation, and the MSM focuses on fused feature refinement.

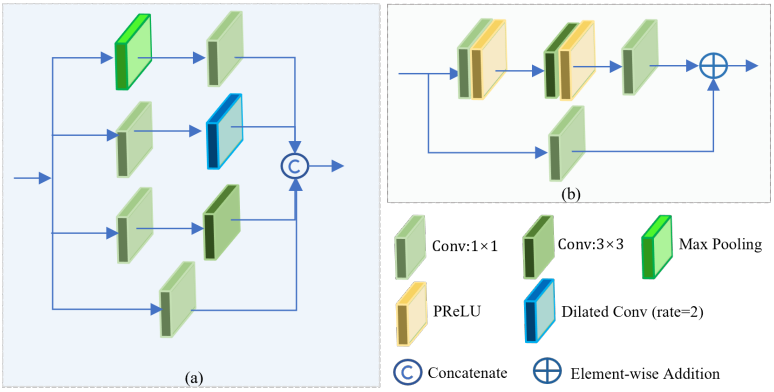

Figure 3. The structure of RBs and the MSM.

Part (b) in Figure 3 represents the structure of the RBs. To enhance feature extraction and alleviate gradient vanishing, we design these residual blocks so that each sequentially applies 1×1 and 3×3 convolutional layers, followed by element-wise addition for skip connections. PReLU is adopted to enhance non-linearity. The RB operation is formulated as:

$$F_{rb} = F_{1P}(F_{3P}(Conv(F_{input})_{1\times1}) + Conv(F_{input})_{1\times1}, \tag{3.1}$$

where F_{rb} is the output of the RB, F_{1P} indicates $Prelu(Conv(\cdot)_{1\times1})$, F_{3P} indicates $Prelu(Conv(\cdot)_{3\times3})$, and F_{input} represents the input of the RB. $Conv(\cdot)_{1\times1}$ and $Conv(\cdot)_{3\times3}$ denote the 1×1 convolution layer and 3×3 convolution layer, respectively.

Notably, for the feature channels after each RB are split, some features are retained for subsequent fusion, while the remaining features are fed into deeper RBs for deeper extraction.

For multi-scale feature fusion, we propose the MSM. As shown in (a) of Figure 3, the MSM employs a 1×1 convolutional layer, a 3×3 regular convolution, a 3×3 dilated convolution (dilation rate = 2), and a max-pooling layer in parallel branches to aggregate multi-scale contextual information. To reduce feature redundancy, a 1×1 convolutional layer is used before each layer, reducing channel dimensions and filtering less informative features. The MSM formulation is defined as:

$$\begin{aligned} F_{msm} = \text{Cat}(&Conv(Conv(F_{fuse})_{1\times1})_{3\times3}, \\ &DilatedConv(Conv(F_{fuse})_{1\times1})_{2\times2}, \\ &MaxPooling(Conv(F_{fuse})_{1\times1}), \\ &Conv(F_{fuse})_{1\times1}), \end{aligned} \tag{3.2}$$

where F_{msm} is the output of the MSM, $DilatedConv(\cdot)_{2\times2}$ indicates the 3×3 dilated convolution layer with a dilation rate of 2, and $MaxPooling(\cdot)$ represents the max pooling layer. $Cat(\cdot)$ denotes a channel connection operation. F_{fuse} indicates the finally fused features.

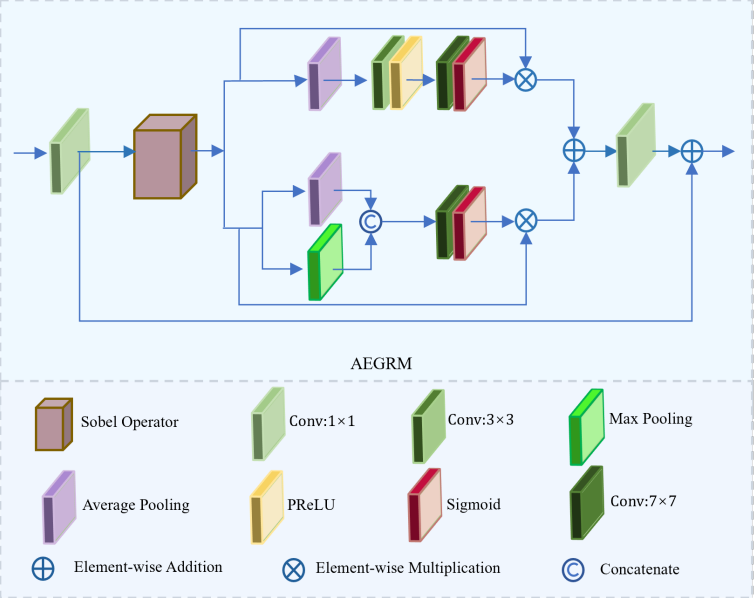

Figure 4. The structure of the AEGRM.

Finally, to enhance texture details and highlight salient targets, we propose an AEGRM based on the Sobel operator and attention mechanisms. This design enables gradient information to guide the attention mechanisms, directing their focus toward edge-sensitive regions. The Sobel operator is less affected by noise and more stable for calculating gradients in the horizontal and vertical directions of the image to preserve edge information. We choose the Sobel operator and explain some specific reasons in the ablation experiments section of Section 4. The structure of the AEGRM is shown in Figure 4. The AEGRM is composed of two convolutional layers, a Sobel operator, a channel attention block, and a spatial attention block. The first 1×1 convolutional layer extracts shallow features, and it can be expressed as:

$$F_{ir} = Conv(I_{ir})_{1\times1}, \tag{3.3}$$

$$F_{vis} = Conv(I_{vis})_{1\times1}, \tag{3.4}$$

where F_{ir} and F_{vis} represent the shallow feature map of the infrared and visible images, respectively. The second 1×1 convolutional layer reduces channel dimensionality discrepancies. The Sobel operator amplifies edge features by computing gradient maps along horizontal and vertical directions. The channel attention block reduces noise by adjusting key channels, while the spatial attention block removes background distractions by focusing on important regions. In addition, element-wise summation is applied to integrate gradient-enhanced edge features and attention-refined contextual features. The AEGRM formulation is defined as:

$$F_{ir_{enhanced}} = F_{ir} + (CA(Sobel(F_{ir})) + SA(Sobel(F_{ir}))), \tag{3.5}$$

$$F_{vis_{enhanced}} = F_{vis} + (CA(Sobel(F_{vis})) + SA(Sobel(F_{vis}))), \tag{3.6}$$

where $F_{ir_{enhanced}}$ and $F_{vis_{enhanced}}$ are the infrared and visible features enhanced by the AEGRM module, respectively. $Sobel(\cdot)$ represents the Sobel operator. $CA(\cdot)$ and $SA(\cdot)$ indicate channel and spatial

attention blocks, respectively.

3.2.2. Dual-channel fusion module (DCFM)

Since the extracted infrared (IR) and visible features contain both common and specific parts, we can express the features F_{ir} and F_{vis} as a combination of their common features and differential features.

$$F_{ir} = \frac{F_{ir} + F_{vis}}{2} + \frac{F_{ir} - F_{vis}}{2}, \tag{3.7}$$

$$F_{vis} = \frac{F_{vis} + F_{ir}}{2} + \frac{F_{vis} - F_{ir}}{2}. \tag{3.8}$$

These feature expressions provide theoretical support for the design of the DCFM. The dual-channel strategy we proposed reconstructs the operational logic of the key step of fusion. We use two channels to process the common features ($F_{ir} + F_{vis}$) and differential features $|F_{ir} - F_{vis}|$, respectively. This dual-channel processing method enables more information extraction. In one branch, we first sum the two modal features, then integrate edge features into the common features, and finally use convolution to extract the enhanced common features. In the other branch, we first compute the difference between features to diminish overlapping regions and emphasize distinct components, then incorporate edge features into the differential features, and finally use convolution to obtain the enhanced differential features. In this way, while capturing the common features and differential features, the network can also fuse these features with edge features. This approach not only addresses the limitation of insufficient information in single-channel fusion but also enhances the fused features, thereby facilitating subsequent feature processing.

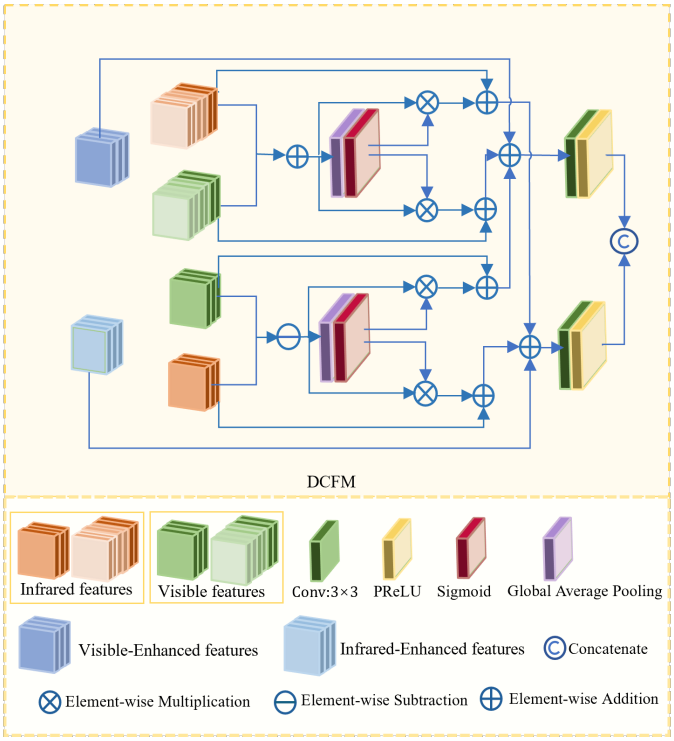

Figure 5. The structure of the DCFM.

The designed DCFM module processes features through parallel dual channels to achieve fusion: one channel extracts differential information from multimodal features, while the other captures the common part. As illustrated in Figure 5, the DCFM consists of pooling layers, convolutional layers, and activation functions. The convolutional and pooling operations effectively suppress noise interference while preventing feature redundancy. The DCFM employs two distinct computational strategies to extract complementary features ($F_{(\cdot)_{complementary}}$) and common features ($F_{common}$) from the infrared features ($F_{ir_{add}}$) and visible features ($F_{vis_{add}}$) extracted via RBs, respectively. Specifically, $F_{ir_{add}}$ and $F_{vis_{add}}$ are expressed as:

$$F_{ir_{add}} = F^3_{rb}(I_{ir}) + F^2_{rb}(I_{ir}) + F^1_{rb}(I_{ir}), \tag{3.9}$$

$$F_{vis_{add}} = F^3_{rb}(I_{vis}) + F^2_{rb}(I_{vis}) + F^1_{rb}(I_{vis}). \tag{3.10}$$

Complementary features and common features are defined as:

$$F_{common} = F_{ir_{add}} + F_{vis_{add}}, \tag{3.11}$$

$$F_{ir_{complementary}} = F^3_{rb}(I_{vis}) - F^3_{rb}(I_{ir}), \tag{3.12}$$

$$F_{vis_{complementary}} = F^3_{rb}(I_{ir}) - F^3_{rb}(I_{vis}), \tag{3.13}$$

where $F_{ir_{complementary}}$ and $F_{vis_{complementary}}$ represent the complementary features of infrared and visible, respectively. $F^{(i)}_{rb}$ denotes the output after i residual blocks.

Next, global average pooling and a sigmoid function are applied for channel weighting to obtain high-level features. Then, the AEGRM-enhanced features are combined with these features. High-level features of infrared images $F_{ir_{fuse}}$ and high-level features of visible images $F_{vis_{fuse}}$can be defined as:

$$\begin{aligned} F_{ir_{fuse}} = &(SG(F_{ir_{complementary}}) \times F_{ir_{complementary}} + F^3_{rb}(I_{ir})) \\ &+ (SG(F_{common}) \times F_{common} + F_{ir_{add}}) + F_{ir_{enhanced}} \end{aligned} \tag{3.14}$$

$$\begin{aligned} F_{vis_{fuse}} = &(SG(F_{vis_{complementary}}) \times F_{vis_{complementary}} + F^3_{rb}(I_{vis})) \\ &+ (SG(F_{common}) \times F_{common} + F_{vis_{add}}) + F_{vis_{enhanced}} \end{aligned} \tag{3.15}$$

where $Mul(\cdot)$ denotes element-wise multiplication, and $SG(\cdot)$ indicates the $Sigmoid(GlobalAveragePooling(\cdot))$ operation.

Finally, a convolutional layer integrates the enhanced and high-level features into effective fused features. The output of the DCFM is expressed as:

$$F_{fuse} = \mathrm{Cat}(F_{3P}(F_{ir_{fuse}}), F_{3P}(F_{vis_{fuse}})), \tag{3.16}$$

where F_{fuse} are the finally fused features.

3.3. Discriminator

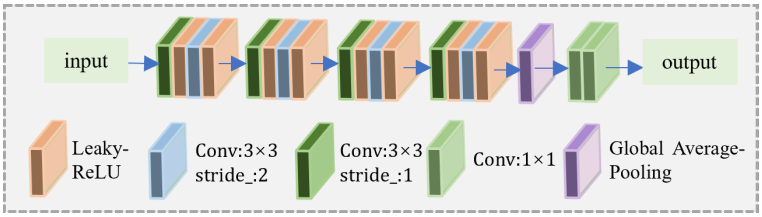

Figure 6. The structure of the discriminator.

Our network includes two discriminators, D_{ir} and D_{vis}, with identical structures, both of which function as classifiers. The input of D_{ir} is generated images and infrared images, and the input of D_{vis} is generated images and visible images. As shown in Figure 6, the front end of each discriminator comprises eight 3×3 convolutional layers. The stride alternates between 1 and 2 across the layers, with the corresponding number of convolutional kernels set to 32, 32, 64, 64, 96, 96, 128, and 128, respectively. At the back-end, a global average pooling layer compresses the spatial dimensions of the feature maps, followed by a 1×1 convolutional layer for classification. The discriminators exclude the batch normalization layer and use LeakyReLU for nonlinear activation.

3.4. Loss function

The overall loss function consists of the generator loss L_G and the discriminator loss L_D. The generator's loss is composed of four components: adversarial loss L_{adv}, L_1 loss, mean squared error loss L_{mse}, perceptual loss L_{percep}, and max intensity loss L_{int}. The selection of our loss functions is inspired by Zhang et al. [32]. Specifically, L_{adv} facilitates adversarial training between the generator and discriminator. Pixel-wise L_1 and L_{mse} losses are employed to refine image textures and semantic details. L_{percep} acts as a regularizer to enhance perceptual quality. L_{int} is introduced to control intensity variations and retain critical luminance information, designed for image fusion tasks.

$$L_G = L_{mse} + \alpha L_1 + \beta L_{percep} + \gamma L_{adv} + \delta L_{int}, \tag{3.17}$$

where α, β, γ, and δ are hyperparameters.

The training process of the network involves an adversarial game between the generator and the discriminator. The generator attempts to generate high-quality fused images to fool the discriminator, while the discriminator strives to distinguish generated images from source images. Therefore, the goal of the generator is to maximize the discriminator's score for the fused image I_f, which means minimizing the negative expectation of the discriminator's output. The adversarial loss of the generator is defined as:

$$L_{adv} = -E_{I_f \sim p_{I_f}}[D_{ir}(I_f)] - E_{I_f \sim p_{I_f}}[D_{vis}(I_f)], \tag{3.18}$$

where p_{I_f} denotes the distribution of the fused images I_f generated by the generator. The first term of the formula measures the ability of the fused image to be judged as a real infrared image by the infrared discriminator D_{ir}, and the second term measures the ability of the fused image to be judged as a source visible image by the visible discriminator D_{vis}. We employ L_1 and L_{mse} as pixel-level loss

functions:

$$L_1 = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left\|I_{ir} - I_f\right\|_1, \tag{3.19}$$

$$L_{mse} = \frac{1}{WH}\sum_{w=1}^{W}\sum_{h=1}^{H}\left\|I_{ir} - I_f\right\|_2^2, \tag{3.20}$$

where $\|\cdot\|_1$ and $\|\cdot\|_2$ represent the L_1-norm and L_2-norm, respectively, with W and H being the width and height of the images. To enhance visual quality, perceptual loss is introduced:

$$\begin{aligned} L_{percep} = \frac{1}{W_{5,4}H_{5,4}}\sum_{w=1}^{W_{5,4}}\sum_{h=1}^{H_{5,4}}&\left(\varphi_{5,4}(I_{vis})_{w,h}\right. \\ &\left.-\varphi_{5,4}(I_f)_{w,h}\right)^2, \end{aligned} \tag{3.21}$$

where $W_{5,4}$ and $H_{5,4}$ are the width and height of feature maps extracted from the 4th and 5th layers of a pre-trained VGG19 model.

Given that a high-quality fused image should retain the highlight regions of infrared images and the clear textures of visible images, we use the maximum intensity loss function to constrain the fused image to maintain an intensity distribution similar to that of the source images. Under the constraint of this loss function, the fused image can better present the contrast between targets and the background, making targets easier to identify. This loss can be defined as:

$$L_{int} = \frac{1}{WH}\left\|I_f - \max\left(I_{ir}, I_{vis}\right)\right\|_1. \tag{3.22}$$

The discriminator loss combines contributions from both infrared and visible light discriminators:

$$L_D = L_{D_{ir}} + L_{D_{vis}}, \tag{3.23}$$

where $L_{D_{ir}}$ and $L_{D_{vis}}$ are formulated as:

$$\begin{aligned} L_{D_{ir}} = &-E_{I_{ir}\sim p}[D_{ir}(I_{ir})] + E_{I_f\sim p_{I_f}}[D_{ir}(I_f)] \\ &+ \lambda E_{\hat{I}}\left[\left(\max\left\{0, \|\nabla_{\hat{I}}D_{ir}(\hat{I})\|_2 - 1\right\}\right)^2\right], \end{aligned} \tag{3.24}$$

$$\begin{aligned} L_{D_{vis}} = &-E_{I_{vis}\sim p}[D_{vis}(I_{vis})] + E_{I_f\sim p_{I_f}}[D_{vis}(I_f)] \\ &+ \lambda E_{\hat{I}}\left[\left(\max\left\{0, \|\nabla_{\hat{I}}D_{vis}(\hat{I})\|_2 - 1\right\}\right)^2\right]. \end{aligned} \tag{3.25}$$

The final term in Eqs (3.24) and (3.25) corresponds to the gradient penalty term, where λ is a regularization parameter and $\nabla(\cdot)$ denotes the image gradient.

4. Experimental analysis

In this section, we first introduce the experimental details. Next, ablation experiments are conducted to validate our proposed design. Finally, the experimental results are analyzed and compared to demonstrate the advantages of our algorithm.

4.1. Experimental settings

We use the TNO dataset [33] and the RoadScene dataset [34] in the experiment. The training set consists of 35 pairs of infrared and visible images from the TNO dataset. Due to the limited number of images in the TNO dataset, we employ a sliding window overlapping cropping method to crop source images sequentially from top to bottom and left to right. We set the sliding window size to 64×64 and the cropping step size to 24. Ultimately, we obtain 15,675 image pairs as the final training set. The learning rate is set to 0.001, the training iterations M is set to 30, and the batch size is 10. Additionally, the parameters involved in the loss function are as follows: $\alpha = 0.01, \beta = 0.06, \gamma = 0.001, \delta = 0.2$.

In our experiment, the quality of the fused images was measured by observing the details of the images and seven objective evaluation metrics. The seven metrics are: average gradient (AG) [35], quality assessment based on blur and noise factors (Qabf) [36], visual information fidelity for fusion (VIF) [35], entropy (EN) [35], edge intensity (EI) [37], and two information theory-based metrics, QMI [38] and QNCIE [39]. EN serves to quantify the amount of information from source images that is retained in the fused image. AG characterizes both the edge clarity and detail richness of the image by calculating its gradient distribution. SF describes the gradient distribution of the fused image from both horizontal and vertical directions, indicating the gray-scale change rate of the fused image within its spatial domain. VIF takes into account the crucial information in images related to visual perception, and it is an indicator proposed based on the characteristics of the human visual system to measure the quality of fused images. EI reflects the fused image's ability to preserve edges from the source images by measuring the gray-scale variation in the edge regions of the image. QMI assesses the interdependence between the input images and the resulting fused output. QNCIE evaluates the nonlinear correlation information entropy between the input images and the fused result. Higher values of these metrics indicate richer information content, finer texture details, and superior fusion performance in the fused image.

4.2. Ablation experiment

1) Effect of Different Modules: We conduct ablation studies to validate the impact of different design modules by comparing the eight model architectures. Structure 1(Base): Base architecture with only parallel RBs. Structure 2: Base + AEGRM. Structure 3: Base + DCFM. Structure 4: Base + MSM. Structure 5: Base + AEGRM + DCFM. Structure 6: Base + AEGRM + MSM. Structure 7: Base + DCFM + MSM. Structure 8(Ours): Our model.

Figure 7 shows the results of the ablation experiment. The results show that the images generated by the model with the AEGRM structure perform better in terms of brightness, contrast, and edge clarity, but suffer from blurred details of the landing gear. The DCFM and MSM structures can preserve more details from the source images (e.g., the connection between the helicopter's bottom and the landing gear). This indicates that the AEGRM, DCFM, and MSM modules all exhibit certain effectiveness in different stages of feature processing. Our model is an integration of these three modules. The fused images it generates can not only retain high-brightness infrared targets (e.g., aircraft fuselage) but also accurately capture detailed textures (e.g., aircraft contour).

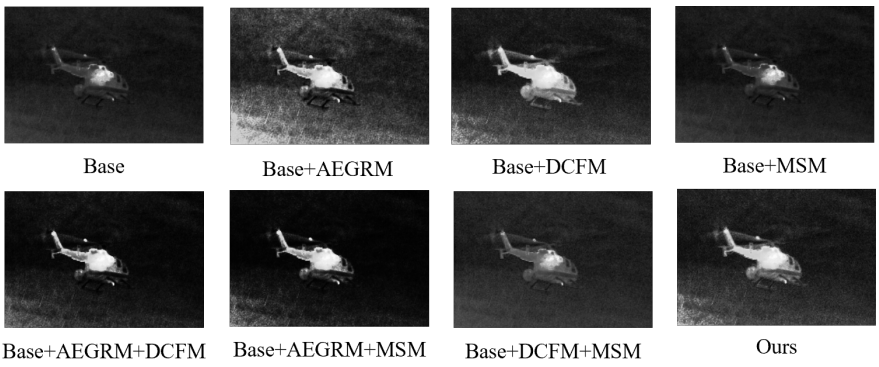

Figure 7. Examples of ablation experiments under different modules.

Table 1. Average results of ablation experiment with different modules. Red boldface indicates the best result.

Method	EN	AG	SF	VIF	EI	QMI	QNCIE
Base+AEGRM+DCFM	5.9681	4.2380	14.8593	0.62067	43.9178	0.3352	0.8050
Base+AEGRM+MSM	5.8555	4.1658	15.0557	0.6395	42.7898	0.3309	0.8048
Base+DCFM+MSM	6.7838	4.5790	12.1699	0.5965	46.9482	0.3106	0.8048
Base+AEGRM	5.8286	4.2303	**15.3826**	0.6329	44.0376	0.3324	0.8048
Base+DCFM	6.6526	4.2558	12.8807	0.6593	43.1078	0.3492	0.8050
Base+MSM	6.4124	3.3989	10.4859	0.4556	34.7184	**0.3578**	0.8050
Base	6.2143	3.2905	10.2338	0.3174	33.7626	0.3294	0.8048
Ours	**6.8231**	**5.1467**	14.3390	**0.7720**	**53.1825**	0.3297	**0.8051**

To further evaluate the effectiveness of the modules, objective metrics were used to analyze 14 pairs of fused images, with the results presented in Table 1. The model incorporating the AEGRM achieved the optimal and suboptimal values in the SF metric. However, it obtained a lower value in the EN metric. This indicates that the AEGRM structure enhances edge sharpness but suppresses a small amount of information. The models with the MSM and DCFM structures obtained the optimal and suboptimal values in the QMI metric, respectively, demonstrating that these two structures can preserve more detailed information. These results indicate that the DCFM structure has achieved positive effects in strengthening common information and capturing complementary information, and also demonstrate that the MSM can extract multi-scale information more comprehensively. Additionally, our model achieved the highest values in most metrics, which are consistent with the subjective evaluation. The experimental results demonstrate that integrating these functional modules into the base architecture enhances fusion performance. Moreover, the combination of these modules exhibits a positive synergistic effect. Specifically, there exists a specific mechanism between the AEGRM, DCFM, and MSM, and this mechanism enables the model to generate high-quality fused images. The AEGRM module utilizes attention and the Sobel operator to focus on key regions of the image and extract comprehensive edge features. This design provides enhanced feature inputs for the fusion step, while also effectively avoiding the loss of key edge information caused by subsequent convolution operations. The DCFM module further integrates the features output by the feature extraction stage and the edge-enhanced

features from the AEGRM module, addressing the limitation of insufficient information fusion in the single-channel mode. For the fused features of the DCFM module, the MSM module processes them by performing large-scale global fusion, medium-scale transition, and small-scale local refinement, resulting in the comprehensively fused multi-scale features. The synergistic interaction between different modules forms a progressive mechanism that enhances key features and fuses them precisely. The combination of these modules exhibits a positive synergistic effect and ultimately generates high-quality fused images.

2) Effect of Different Gradient Operators: Given that our research focuses more on image fusion rather than edge detection, we only conducted comparative experiments on commonly used edge detection operators.

Figure 8. Examples of ablation experiments under different gradient operators.

We randomly select three images and applied the Sobel, Canny, and Laplace operators to these visible images. The results are shown in Figure 8. To facilitate observation, important regions in the images are marked with red bounding boxes. The Canny operator can obtain clear single-pixel edge images by defining thresholds, but compared with images extracted by the Sobel operator, it exhibits partial edge loss in the vertical direction (e.g., the pillars in the first row of images and the street lamps in the third row of images). The result obtained by the Canny operator shows loss of background information (e.g., the wall surface in the first-row image and the body contours of the pedestrian in the third-row image). The results of the Laplace operator exhibit a large amount of noise and blurred edge lines (e.g., the edges of human targets, numbers, and letters in Figure 10 are all unclear). Compared with the Laplace operator, the results of the Sobel operator exhibit better visual effects. Additionally, in image fusion tasks, fused images need to retain as much information from the source images as possible. From this perspective, the Sobel operator retains more complete edges compared with the Canny operator. Therefore, we choose to use the Sobel operator in the AEGRM.

3) Effect of Different Feature Fusion Structures: We conducted an ablation study on the dual-channel design to verify its effectiveness. The fusion module comprises two components: a common-feature fusion channel (Structure A) and a differential-feature fusion channel (Structure B). Experi-

ments were performed using each structure independently, with results shown in Figure 9.

Figure 9. Examples of ablation experiments under different feature fusion structures.

Observation reveals that Structure A generates a fused image with partial infrared target representation, as evidenced by the unclear lower-body contours of the person. This suggests that Structure A fails to fully preserve critical thermal characteristics. The output of Structure B exhibits deficiencies in detail retention, particularly in the brightness contrast between the sky and the target person, which indicates limitations in feature extraction. These results demonstrate that neither fusion channel alone can effectively integrate infrared and visible features.

Table 2. Average results of the ablation experiment of different feature fusion structures. Red boldface indicates the best result.

Method	EN	AG	SF	VIF	EI	QMI	QNCIE
Structure A	6.83482	3.8323	10.4147	0.5258	38.8261	0.3390	0.8052
Structure B	6.9579	4.6733	12.6676	0.5908	45.1612	0.3360	0.8050
Ours	**7.0654**	**5.1366**	**14.9311**	**0.6828**	**49.3266**	**0.3872**	**0.8474**

The average results on the 16 TNO dataset pairs are shown in Table 2. Our method achieves the best scores in all metrics, consistent with subjective evaluations, proving that our DCFM better fuses infrared and visible features.

4) Effect of the Hyperparameter of Generator Loss

Since this paper refers to the research of Zhang et al. [32] in the selection of hyperparameters for L_{adv}, L_1, L_{mse}, and L_{percep}, only an experimental analysis of the hyperparameter δ for L_{int} is required. To determine the optimal value of δ, this section experiments by setting different values of δ. Figure 10 shows the results of a set of images from TNO under different δ values. The results show that as δ gradually increases, the overall brightness and contrast of the image also increase. However, a further increase in δ will lose infrared target information and reduce the brightness of the fused image. A relatively low δ value tends to weaken the texture features of the original image, leading to the loss of some key information in the fused image (such as the outline of the aircraft in Figure 10). When δ is set to 0.2, it works together with the hyperparameters of other loss functions to achieve a balance between pixel intensity and texture details of the image.

Figure 10. Examples of ablation experiments on different hyperparameters for the generator loss function.

To further quantitatively evaluate the fusion effect under different parameters, objective evaluation metrics are used to analyze the images. The results are shown in Table 3. It is shown that when $\delta = 0.2$, all evaluation metrics reach optimal values except for MI and Qabf. After comparative analysis with the other four groups of hyperparameters, the model under this parameter setting achieves better fusion performance, which is also consistent with the results of subjective evaluation. Therefore, this hyperparameter setting will be used in subsequent experiments.

Table 3. Average results of the ablation experiments of different hyperparameters. Red boldface indicates the best result.

Method	EN	AG	SF	VIF	EI	QMI	QNCIE
$\delta = 0.1$	6.6674	4.0542	11.5707	0.3584	40.7478	0.3678	0.8058
$\delta = 0.2$	**7.0058**	**5.6426**	**16.1410**	**0.5729**	**55.8437**	0.3694	0.8058
$\delta = 0.3$	6.7793	4.3861	12.1108	0.3600	44.0599	0.3838	**0.8064**
$\delta = 0.5$	6.9251	5.0647	15.0252	0.5625	51.7368	0.3558	0.8055
$\delta = 1$	6.8113	4.5230	13.3302	0.3855	45.3519	**0.4306**	0.8054

4.3. Comparison experiment

In this section, we evaluate our proposed method on publicly available datasets. We select thirteen commonly used methods for comparison, including dual-tree complex wavelet transform (DTCWT) [40], convolutional sparse representation (CSR) [41], ResNet and zero-phase component analysis (RZP) [19], DenseFuse [20], FusionGAN [21], DDcGAN [22], DATFusion [25], U2Fusion [33], generative adversarial networks with multi-class constraints (GANMcC) [42], PMGI [43], YDTR [27], ITFuse [29], and MMAE [44]. We perform qualitative and quantitative comparisons of our method with the approaches mentioned above.

We selected 9 pairs of images from the TNO dataset and 20 pairs of images from the RoadScene dataset to observe the differences in the fusion quality among different methods. Three pairs of representative images were chosen, and their fusion results under 14 methods are shown in Figure 11.

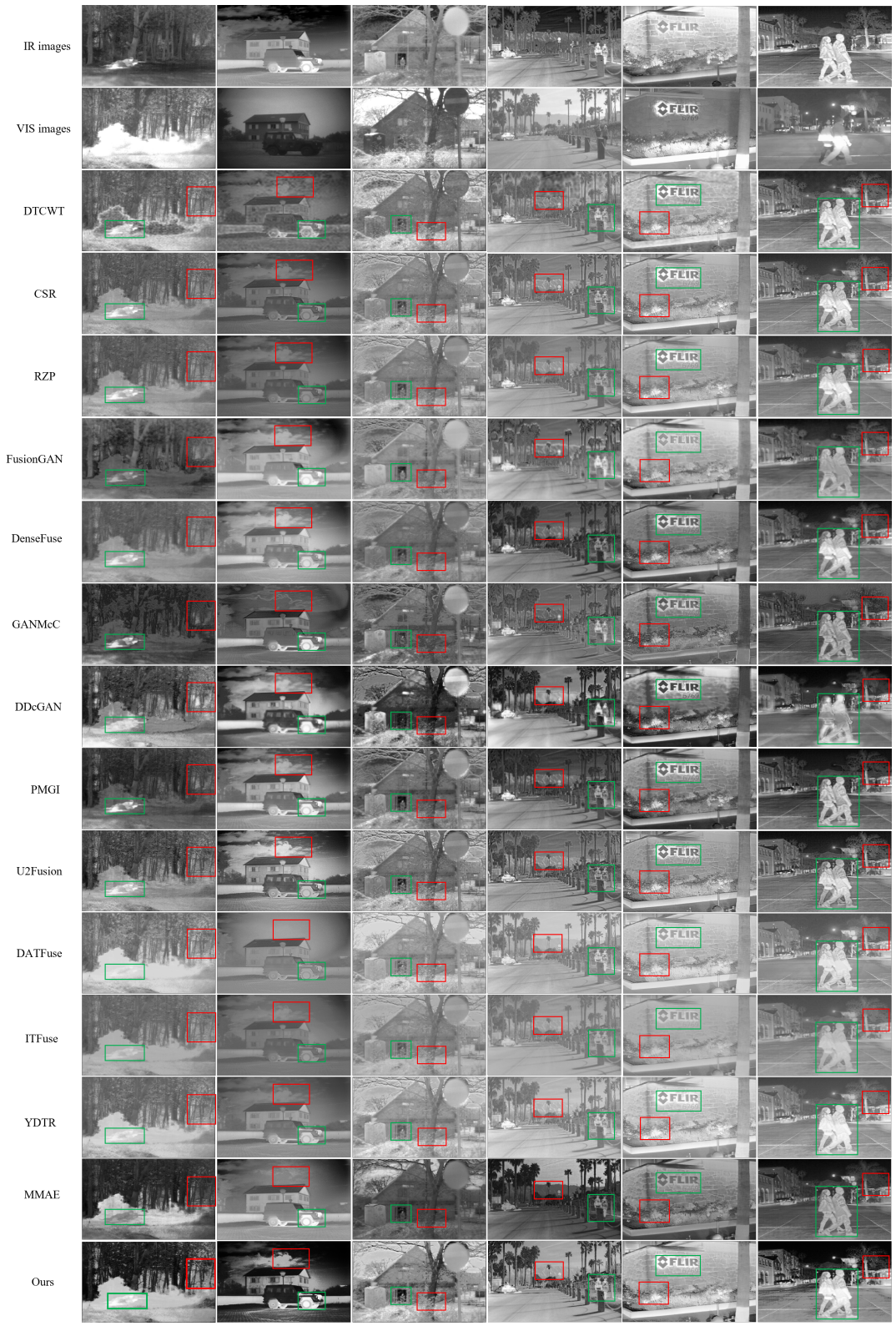

Figure 11. Visual quality comparison of our method with thirteen advanced methods on TNO and RoadScene datasets. Images in the first through third columns are from TNO, and images in the fourth through the sixth columns are from RoadScene.

Table 4. Average results of multiple image fusion methods under objective evaluation metrics based on the TNO and RoadScene datasets. Bold red, bold green, bold blue, and bold black values indicate the results ranking first, second, third, and fourth, respectively.

TNO							
Method	EN	AG	SF	VIF	EI	QMI	QNCIE
DTCWT	6.8638	**4.5546**	**11.8321**	0.2942	44.2002	0.2934	0.8039
CSR	6.6136	4.1093	10.9722	0.2796	40.2647	0.2700	0.8039
FusionGAN	6.6751	3.2469	8.1865	0.1612	31.7999	0.2896	0.8045
GANMcC	6.4384	4.2778	10.9053	0.1771	**41.4923**	0.2505	0.8038
RZP	6.5318	2.9013	7.1536	0.2647	28.5695	0.2874	0.8042
DenseFuse	**6.9805**	2.9528	6.0296	**0.4019**	31.1528	0.2653	0.8039
DDCGAN	**7.5364**	**5.7512**	**14.0039**	**0.5476**	**56.1588**	0.2273	0.8034
PMGI	6.8618	3.3842	7.9779	0.3092	34.5429	0.2742	0.8041
U2Fusion	**7.1033**	**5.6065**	**12.3800**	**0.5553**	**57.0102**	0.2740	0.8042
DATFuse	6.4533	3.6197	9.5288	0.1479	35.1481	**0.4523**	**0.8076**
ITFuse	6.7469	2.7364	6.2558	0.1117	28.8747	**0.5862**	**0.8145**
YDRT	6.5976	3.1514	8.3253	0.2322	31.0868	**0.3387**	**0.8052**
MMAE	6.7691	3.6637	10.6760	0.1726	35.9747	**0.5386**	**0.8098**
Ours	**7.1403**	**5.7945**	**16.4651**	**0.6905**	**56.1797**	0.3363	**0.8051**

RoadScene							
Method	EN	AG	SF	VIF	EI	QMI	QNCIE
DTCWT	7.1982	5.3118	13.0510	0.3539	55.6456	0.3153	0.8058
CSR	6.9219	5.2988	**14.1681**	0.3392	55.2819	0.3244	0.8057
FusionGAN	**7.3150**	3.8284	9.3142	0.2565	41.0627	0.3636	**0.8067**
GANMcC	6.9032	**5.4627**	**13.9610**	0.2526	**57.1035**	0.3278	0.8059
RZP	6.7783	3.6905	9.3521	0.3124	38.5024	0.3475	0.8059
DenseFuse	7.2745	4.0885	8.9189	**0.4490**	43.9134	0.3270	0.8056
DDCGAN	**7.6263**	**5.3351**	13.6549	0.4114	**56.5455**	0.3067	0.8051
PMGI	**7.3912**	4.8090	11.3704	**0.4534**	51.7305	**0.3923**	**0.8077**
U2Fusion	7.1284	**6.6583**	**16.3374**	**0.4539**	**70.0503**	0.3020	0.8052
DATFuse	6.7793	4.3406	12.3566	0.2194	44.2112	**0.5120**	**0.8087**
ITFuse	6.2760	2.2634	5.3219	0.1739	24.4513	0.3453	0.8055
YDRT	6.8612	4.2283	11.3760	0.3160	43.6292	**0.3642**	0.8060
MMAE	7.2216	4.8307	13.4527	0.2928	50.1285	**0.6059**	**0.8134**
Ours	**7.3252**	**6.4926**	**17.5133**	**0.5844**	**68.1374**	0.3632	0.8064

As observed from Figure 11, there are noise points in the fused image of DTCWT, and the noise introduced in the sky background is particularly obvious. DDCGAN, DenseFuse, and DATFuse weaken the infrared targets. CSR and RZP maintain relatively complete infrared targets, but the overall brightness is rather dim. In the results of FusionGAN and PMGI, some regions are not clear. GANMcC retains more information of the infrared image, but compared with the source visible images, the tex-

ture details are blurred. U2Fusion loses local detail information. However, YDTR and U2Fusion lose some local detail information. Specifically, U2Fusion misses part of the background information, such as the sky background in the third and fourth columns; in YDTR, the overall shape of the clouds in the second column appears blurry. In the results of ITFuse, more background information from the source visible images is preserved, but the edges of infrared targets are blurred (e.g., the outline of the person in the first column). The brightness of target objects in MMAE is relatively low, and some contrast information is lost (for instance, the human targets in the first column, the letters in the fifth column, and the person in the sixth column). Compared to these images, the images produced by our method highlight target areas and have clear texture details, and the fused images balance both clarity and contrast. When observing from the areas marked in green, the infrared targets in the results of DCGAN-Fuse are more prominent, and the outlines of people and objects are clearly visible. From the areas marked in red, our method can provide more complete details, such as the branches in the first and third columns, the clouds in the second column, the grass in the second column, and the texture structure of the wall in Figure 11.

To comprehensively verify the effectiveness of our method, we performed objective evaluations on the aforementioned 9 and 20 image pairs, respectively. Table 4 shows the average results for each metric. Table 4 shows the excellent performance of our proposed DCGAN-Fuse on the TNO and RoadScene datasets. On the TNO dataset, our DCGAN-Fuse achieves the best values in the AG, SF, and VIF metrics, and reaches the second-best value in the EN and EI metrics. Additionally, our method achieves the fourth-best values in both the QMI and QNCIE metrics. On the RoadScene dataset, it achieves the best values in the two metrics of SF and VIF, and reaches the second-best values in the AG and EI metrics. Our method ranks third in the EN metric, and performs at a medium-to-high level in terms of the QMI and QNCIE metrics. The maximum values indicate that our DCGAN-Fuse can extract more information from the source images, retain the most edge information, and generate better visual images with higher contrast and clarity. Our method achieves competitive performance overall, but it is not excellent in some specific metrics (e.g., QMI and QNCIE). This is because during the feature extraction process of DCGAN-Fuse, we prioritized noise resistance rather than retaining all the details. Through qualitative and quantitative evaluations, DCGAN-Fuse demonstrates advantages. The fused results preserve richer information and achieve higher target contrast. Specifically, the fused images preserve the salient thermal targets from infrared images while maintaining the detailed texture information from visible images.

To verify the generalization ability of our model, we conducted comparative experiments on the MSRS dataset. We selected 20 pairs of images from the MSRS dataset. The qualitative and quantitative results are shown in Figure 12 and Table 5, respectively. The results demonstrate that our method outperforms many other state-of-the-art (SOTA) methods in terms of visual analysis and quantitative metrics on the MSRS dataset. Qualitatively, our method DCGAN-Fuse preserves important features from both inputs, such as edges and textures. Visually, it contains clear infrared targets and background textures. Quantitatively, our approach obtained the top value in the SF metric, while it performed as the second-best in the AG, EI, QMI, and QNCIE metrics. Although DDCGAN achieves the highest values in four metrics, its generated images are overall bright but introduce some noise (e.g., in the ground area shown in Figure 12). However, our method achieves a balance between visual performance and objective evaluation. Objective evaluations indicate that our method better preserves the structure, information, and visual effects of images, which is consistent with subjective evaluations.

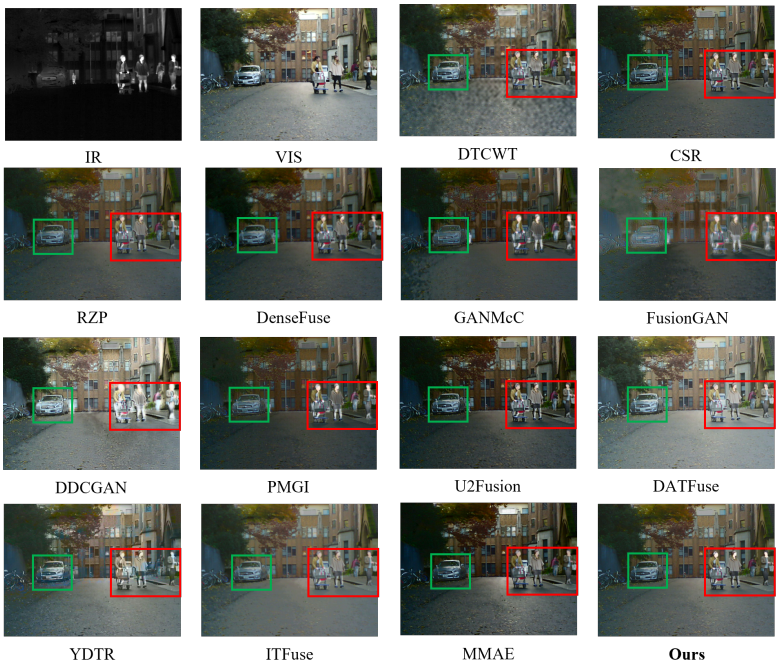

Figure 12. Visual quality comparison of our method with thirteen advanced methods on the MSRS dataset.

Table 5. Average results of multiple image fusion methods under objective evaluation metrics based on the MSRS dataset. Bold red, bold green, bold blue, and bold black values indicate the results ranking first, second, third, and fourth, respectively.

Method	EN	AG	SF	VIF	EI	QMI	QNCIE
DTCWT	**7.1293**	**5.0227**	**13.5260**	0.0286	**53.7291**	0.0848	0.8022
CSR	6.6212	4.3185	12.6689	0.0170	46.0619	0.0981	0.8022
FusionGAN	6.2493	3.4496	9.5148	0.0123	36.4104	0.0827	0.8021
GANMcC	6.1051	3.4882	9.4635	0.0127	36.9345	0.0846	0.8021
RZP	6.4653	2.8728	7.8224	0.0145	30.5586	0.1020	0.8022
DenseFuse	**6.9880**	3.2741	7.6755	0.5358	35.4984	**0.4024**	**0.8076**
DDCGAN	**7.3862**	**6.7744**	**17.6126**	**0.6468**	**71.8561**	0.2655	0.8040
PMGI	6.2991	3.7424	10.2049	0.3026	39.4935	0.3016	0.8042
U2Fusion	6.7759	4.7222	12.8874	**0.6111**	50.6271	**0.4022**	**0.8073**
DATFuse	**7.0791**	**5.1996**	**14.6030**	0.0262	**54.7332**	0.1073	0.8022
ITFuse	6.4964	2.5996	6.6466	0.0177	28.4026	0.1125	0.8022
YDRT	6.8496	4.2055	12.3540	0.0243	44.6545	0.1065	0.8022
MMAE	6.9142	4.7164	14.6878	**0.5380**	50.0868	**0.6074**	**0.8153**
Ours	6.9235	**5.8080**	**17.8607**	**0.5885**	**58.8944**	**0.5202**	**0.8124**

Extensive experiments have shown that our method can generate high-quality fused images. Our method shows certain superiority over other methods in both qualitative and quantitative analyses. It

can retain effective information from the original infrared and visible images, and generate fused images with highlighted infrared targets and rich texture information. To conduct a more comprehensive evaluation, we compare the parameter counts and floating point operations (FLOPs) of our method with those of other methods. The size of the source image used to calculate FLOPs is 120×120. Relevant results are reported in Table 6. The results show that our method has a medium-level parameter count and lower FLOPs. In future work, we will optimize the parameters to accelerate the fusion process and explore more advanced methods to adapt to more application scenarios and tasks, thereby meeting diverse practical needs.

Table 6. Parameters and FLOPs of eleven deep learning methods. Parameters are in millions; FLOPs are in GFLOPs. Bold red, bold green, and bold blue values indicate the results ranking first, second, and third, respectively.

	FusionGAN	GANMcC	RZP	DenseFuse	DDCGAN	PMGI	U2Fusion	DATFuse	ITFuse	YDRT	MMAE	ours
Parameters	1.32	2.17	8.57	**0.30**	1.10	**0.04**	0.66	**0.07**	0.08	0.87	0.84	1.02
FLOPs	28.54	126.48	**0.23**	2.54	31.55	1.19	81.19	**0.51**	2.37	9.00	21.03	**0.80**

5. Conclusions

In this work, we propose an infrared and visible image fusion method named DCGAN-Fuse, which is based on a dual-channel fusion strategy and dual discriminators. First, at different feature extraction stages, we design different modules to extract and enhance specific features. Second, we design the DCFM based on a dual-channel fusion strategy. This strategy improves the network's ability to fuse features of different modalities and supplements more detailed information for the fused features. Finally, the network is guided to optimize by the loss function. Experiments on the TNO, RoadScene, and MSRS show that the fused images obtained by our method contain highlighted thermal target areas and have rich background textures. The fusion effect is better than in the other thirteen comparison algorithms, providing a foundation for subsequent advanced visual tasks.

In future work, we will explore more advanced fusion mechanisms to improve the fusion effect and expand the application of this method to more scenarios and tasks.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Acknowledgments

This work is supported by the Basic Scientific Research Project of the Department of Education of Liaoning Province under grant 984250013.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. L. Bai, W. Zhang, X. Pan, C. Zhao, Underwater image enhancement based on global and local equalization of histogram and dual-image multi-scale fusion, *IEEE Access*, **8** (2020), 128973–128990. https://doi.org/10.1109/ACCESS.2020.3009161
2. M. Rashid, M. A. Khan, M. Alhaisoni, S. H. Wang, S. R. Naqvi, A. Rehman, et al., A sustainable deep learning framework for object recognition using multi-layers deep features fusion and selection, *Sustainability*, **12** (2020), 5037. https://doi.org/10.3390/su12125037
3. X. Liang, J. Zhang, L. Zhuo, Y. Li, Q. Tian, Small object detection in unmanned aerial vehicle images using feature fusion scaling-based single shot detector with spatial context analysis, *IEEE Trans. Circuits Syst. Video Technol.*, **30** (2020), 1758–1770. https://doi.org/10.1109/TCSVT.2019.2905881
4. M. Jiang, Y. Zhao, J. Kong, Mutual learning and feature fusion siamese networks for visual object tracking, *IEEE Trans. Circuits Syst. Video Technol.*, **31** (2021), 3154–3167. https://doi.org/10.1109/TCSVT.2020.3037947
5. V. I. Adamchuk, R. A. Viscarra Rossel, K. A. Sudduth, P. S. Lammers, Sensor fusion for precision agriculture, *Sensor Fusion - Found. Appl.*, 2011. https://doi.org/10.5772/19983
6. W. Xiong, Z. Xiong, Y. Cui, L. Huang, R. Yang, An interpretable fusion siamese network for multi-modality remote sensing ship image retrieval, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 2696–2712. https://doi.org/10.1109/TCSVT.2022.3224068
7. M. A. N. U. Ghani, K. She, M. U. Saeed, N. Latif, Enhancing facial recognition accuracy through multi-scale feature fusion and spatial attention mechanisms, *Electronic Res. Arch.*, **32** (2024), 2267–2285. https://doi.org/10.3934/era.2024103
8. W. Zhang, M. Dai, B. Zhou, C. Wang, MCADFusion: a novel multi-scale convolutional attention decomposition method for enhanced infrared and visible light image fusion, *Electronic Res. Arch.*, **32** (2024), 5067–5089. https://doi.org/10.3934/era.2024233
9. Y. Liu, J. Jin, Q. Wang, Y. Shen, X. Dong, Region level based multi-focus image fusion using quaternion wavelet and normalized cut, *Signal Process.*, **97** (2014), 9–30. https://doi.org/10.1016/j.sigpro.2013.10.010
10. Y. Liu, X. Yang, R. Zhang, M. K. Albertini, T. Celik, G. Jeon, Entropy-based image fusion with joint sparse representation and rolling guidance filter, *Entropy*, **22** (2020), 118. https://doi.org/10.3390/e22010118
11. A. Wang, M. Wang, RGB-D salient object detection via minimum barrier distance transform and saliency fusion, *IEEE Signal Process. Lett.*, **24** (2017), 663–667. https://doi.org/10.1109/LSP.2017.2688136
12. W. Kong, Y. Lei, H. Zhao, Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization, *Infrared Phys. Technol.*, **67** (2014), 161–172. https://doi.org/10.1016/j.infrared.2014.07.019
13. R. Nie, J. Cao, D. Zhou, W. Qian, Multi-source information exchange encoding with PCNN for medical image fusion, *IEEE Trans. Circuits Syst. Video Technol.*, **31** (2021), 986–1000. https://doi.org/10.1109/TCSVT.2020.2998696

14. J. Zhao, G. Cui, X. Gong, Y. Zang, S. Tao, D. Wang, Fusion of visible and infrared images using global entropy and gradient constrained regularization, *Infrared Phys. Technol.*, **81** (2017), 201–209. https://doi.org/10.1016/j.infrared.2017.01.012
15. J. Tang, A contrast based image fusion technique in the DCT domain, *Digital Signal Process.*, **14** (2004), 218–226. https://doi.org/10.1016/j.dsp.2003.06.001
16. J. Tang, Q. Sun, Z. Wang, Y. Cao, Perfect-reconstruction four-tap size-limited filter banks for image fusion application, in *2007 International Conference on Mechatronics and Automation (ICMA)*, (2007), 255–260. https://doi.org/10.1109/ICMA.2007.4303550
17. W. Zhou, X. Lin, J. Lei, L. Yu, J. N. Hwang, MFFENet: multiscale feature fusion and enhancement network for RGB–thermal urban road scene parsing, *IEEE Trans. Multimedia*, **24** (2022), 2526–2538. https://doi.org/10.1109/TMM.2021.3086618
18. K. R. Prabhakar, V. S. Srikar, R. V. Babu, DeepFuse: a deep unsupervised approach for exposure fusion with extreme exposure image pairs, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 4724–4732. https://doi.org/10.1109/ICCV.2017.505
19. H. Li, X. J. Wu, T. S. Durrani, Infrared and visible image fusion with ResNet and zero-phase component analysis, *Infrared Phys. Technol.*, **102** (2019), 103039. https://doi.org/10.1016/j.infrared.2019.103039
20. H. Li, X. J. Wu, DenseFuse: a fusion approach to infrared and visible images, *IEEE Trans. Image Process.*, **28** (2019), 2614–2623. https://doi.org/10.1109/TIP.2018.2887342
21. J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: a generative adversarial network for infrared and visible image fusion, *Inf. Fusion*, **48** (2019), 11–26. https://doi.org/10.1016/j.inffus.2018.09.004
22. J. Ma, H. Xu, J. Jiang, X. Mei, X. Zhang, DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion, *IEEE Trans. Image Process.*, **29** (2020), 4980–4995. https://doi.org/10.1109/TIP.2020.2977573
23. H. Zhou, W. Wu, Y. Zhang, J. Ma, H. Ling, Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network, *IEEE Trans. Multimedia*, **25** (2023), 635–648. https://doi.org/10.1109/TMM.2021.3129609
24. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin Transformer: hierarchical vision Transformer using shifted windows, in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, (2021), 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986
25. W. Tang, F. He, Y. Liu, Y. Duan, T. Si, DATFuse: infrared and visible image fusion via dual attention transformer, *IEEE Trans. Circuits Syst. Video Technol.*, **33** (2023), 3159–3172. https://doi.org/10.1109/TCSVT.2023.3234340
26. J. Li, J. Zhu, C. Li, X. Chen, B. Yang, CGTF: Convolution-guided transformer for infrared and visible image fusion, *IEEE Trans. Instrum. Meas.*, **71** (2022), 1–14. https://doi.org/10.1109/TIM.2022.3175055
27. W. Tang, F. He, Y. Liu, YDTR: Infrared and visible image fusion via Y-shape dynamic transformer, *IEEE Trans. Instrum. Meas.*, **25** (2023), 5413–5428. https://doi.org/10.1109/TMM.2022.3192661

28. X. Hu, Y. Liu, F. Yang, PFCFuse: a Poolformer and CNN fusion network for infrared-visible image fusion, *IEEE Trans. Instrum. Meas.*, **71** (2024), 1–14. https://doi.org/10.1109/TIM.2024.3450061
29. W. Tang, F. He, Y. Liu, ITFuse: An interactive transformer for infrared and visible image fusion, *Pattern Recognit.*, **156** (2024), 110822. https://doi.org/10.1016/j.patcog.2024.110822
30. M. Arjovsky, L. Bottou, Towards principled methods for training generative adversarial networks, in *5th International Conference on Learning Representations (ICLR)*, (2017), 24–26. https://doi.org/10.48550/arXiv.1701.04862
31. H. Petzka, A. Fischer, D. Lukovnicov, On the regularization of wasserstein GANs, in *6th International Conference on Learning Representations (ICLR)*, (2018), 11–13. https://doi.org/10.48550/arXiv.1709.08894
32. A. Zhang, The research of single image super-resolution reconstruction based on improved generative adversarial network, *CNKI*, 2022. https://doi.org/10.26989/d.cnki.gdlhu.2022.001747
33. A. Toet, TNO Image Fusion Dataset, figshare, 2014. https://doi.org/10.6084/m9.figshare.1008029
34. H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2Fusion: a unified unsupervised image fusion network, *IEEE Trans. Pattern Anal. Mach. Intell.*, **44** (2022), 502–518. https://doi.org/10.1109/TPAMI.2020.3012548
35. Y. Yang, J. Li, Y. Wang, Review of image fusion quality evaluation methods, *J. Front. Comput. Sci. Technol.*, **12** (2018), 1021–1035. http://fcst.ceaj.org/EN/10.3778/j.issn.1673-9418.1710001
36. V. Petrovic, C. Xydeas, Objective image fusion performance characterisation, in *10th IEEE International Conference on Computer Vision (ICCV)*, (2005), 1866–1871. https://doi.org/10.1109/ICCV.2005.175
37. R. Balakrishnan, R. Priya, Hybrid multimodality medical image fusion technique for feature enhancement in medical diagnosis, *J. Med. Imaging Health Inf.*, **8** (2018), 52–60. https://www.researchgate.net/publication/326913363
38. M. Hossny, S. Nahavandi, D. Creighton, Comments on information measure for performance of image fusion, *Electron. Lett.*, **44** (2008), 1066–1067. https://doi.org/10.1049/el:20081754
39. Q. Wang, Y. Shen, J. Zhang, A nonlinear correlation measure for multivariable data set, *Physica D*, **200** (2005), 287–295. https://doi.org/10.1016/j.physd.2004.11.001
40. J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, *Inf. Fusion*, **8** (2007), 119–130. https://doi.org/10.1016/j.inffus.2005.09.006
41. Y. Liu, X. Chen, R. K. Ward, Z. J. Wang, Image fusion with convolutional sparse representation, *IEEE Signal Process. Lett.*, **23** (2016), 1882–1886. https://doi.org/10.1109/LSP.2016.2618776
42. J. Ma, H. Zhang, Z. Shao, P. Liang, H. Xu, GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion, *IEEE Trans. Instrum. Meas.*, **70** (2021), 1–14. https://doi.org/10.1109/TIM.2020.3038013

43. H. Zhang, H. Xu, Y. Xiao, X. Guo, 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*, **34** (2020), 12797–12804. https://doi.org/10.1609/aaai.v34i07.6975

44. X. Wang, L. Fang, J. Zhao, Z. Pan, H. Li, Y. Li, MMAE: A universal image fusion method via mask attention mechanism, *Pattern Recognit.*, **158** (2025), 111041. https://doi.org/10.1016/j.patcog.2024.111041

AIMS Press