Tian *et al*. *Intell Robot* 2024;4(4):524-43
DOI: 10.20517/ir.2024.30

Intelligence & Robotics

Research Article

Open Access

Infrared and visible image fusion based on multi-level detail enhancement and generative adversarial network

Xiangrui Tian, Xiaohan Xianyu, Zhimin Li, Tong Xu, Yinjun Jia

College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210000, Jiangsu, China.

Correspondence to: Dr. Xiangrui Tian, College of Automation Engineering, Nanjing University of Aeronautics and Astronautics, No. 29, General Avenue, Jiangning District, Nanjing 210000, Jiangsu, China. E-mail: xiangruitian@nuaa.edu.cn

How to cite this article: Tian X, Xianyu X, Li Z, Xu T, Jia Y. Infrared and visible image fusion based on multi-level detail enhancement and generative adversarial network. *Intell Robot* 2024;4(4):524-43. http://dx.doi.org/10.20517/ir.2024.30

Received: 10 Sep 2024 **First Decision:** 19 Nov 2024 **Revised:** 21 Dec 2024 **Accepted:** 23 Dec 2024 **Published:** 31 Dec 2024

Academic Editors: Simon Yang, Xin Jin **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

Abstract

Infrared and visible image fusion technology has a wide range of applications in many fields such as target detection and tracking. Existing image fusion methods often overlook the scale hierarchical structure information of features, with local and global features not being closely interconnected. Typically, improvements focus on the network structure and loss functions, while the intimate connection between the quality of the source images and the feature extraction network is often neglected. The aforementioned issues lead to artifacts and blurring of fused images; besides, the detailed edge information can not be well reflected. Therefore, a method of infrared and visible image fusion based on a generative adversarial network (GAN) with multi-level detail enhancement is proposed in this paper. Firstly, the edge information of the input source image is enriched by the multi-level detail enhancement method, which improves the image quality and makes it more conducive to the learning of feature extraction network. Secondly, the residual-dense and multi-scale modules are designed in the generator and the connection between local and global features is established to ensure the transmissibility and coherence of the feature information. Finally, by designing the loss function and dual discriminator constraints to constrain the fusion image, more structure and detail information are added in continuous confrontation. The experimental results show that the fused image contains more detailed texture information and prominent thermal radiation targets. It also outperforms other fusion methods in terms of average gradient (AG), spatial frequency (SF) and edge intensity (EI) metrics, with values surpassing the sub-optimal metrics of 65.41%, 65.09% and 55.22%, respectively.

Keywords: Image fusion, multi-level detail enhancement, generative adversarial networks, deep feature extraction

1. INTRODUCTION

With the continuous development of information communication and image sensing technologies, various types of image sensors are widely used across multiple industries. The image information obtained by a single sensor has certain limitations and can only reflect certain aspects of feature information in the scene which cannot meet application requirements. Therefore, image fusion technology has been vigorously developed. As a branch of image fusion, infrared and visible image fusion technology is widely used in many fields such as target recognition, target detection and tracking, intelligent monitoring, agricultural automation, remote sensing detection[1]. It makes full use of the different imaging characteristics of infrared and visible images, and adapts to various scenarios through information complementarity. Infrared images provide rich thermal radiation information, prominent targets, and are not affected by weather, but they have low resolution and unclear details. In contrast, visible images offer high resolution and rich color and texture information, but are susceptible to bad weather[2]. By combining the advantages of both, fused images with high quality and excellent visual effects can be obtained, which are favorable for subsequent image processing.

Traditional infrared and visible image fusion algorithms usually use multi-scale transformation[3,4] and sparse representation methods[5,6], which have made certain progress. Meng *et al.* used the non-subsampled contour wave multi-scale transformation method[7] to solve the problem of salient target prominence in the fused image. However, it relies too much on manual design, making it only suitable for complex and changeable fusion scenarios[8], which results in high time consumption and inefficiency. In recent years, with the development and progress of deep learning technology, increasing image fusion methods based on deep learning have been emerging. Convolutional neural networks (CNNs) were introduced to achieve multi-focus image fusion tasks[9], and relied on decision diagram classification to significantly improve the fusion effect. Li *et al.* proposed a fusion structure based on dense blocks and autoencoders (DenseFuse)[10]. Its dense network ensures that the significant deep features of the fused image will not be lost. However, this method still requires manual design of the fusion strategy and cannot achieve end-to-end infrared and visible image fusion. Subsequently, using a multi-scale approach based on DenseFuse, the fusion of feature maps output from multiple convolutional layers has achieved good results[11]. However, less complementary information between modalities has been extracted, and the fusion rules are still manually designed, without fusing the specific information of infrared and visible light. Adaptive fusion transformer (AFT) model revealed the latent relationship between the deep features of visible and infrared images, thereby achieving more accurate perception in fusion[12]. In 2019, Ma *et al.* applied generative adversarial networks (GANs)[13] to the task of infrared and visible image fusion, and realized end-to-end fusion of infrared and visible images without the need for manual design of fusion rules[14]. Subsequently, the proposed GAN-based multi-classification constrained fusion method effectively balanced the feature distribution of the source image[15], but the detailed texture and edge information of the image failed to be highlighted. A bi-discriminator conditional GAN modelnot only maintains source image information balance, but also enables multi-resolution image fusion[16], in addition to solving the problem of differentiating image gradients and intensities with a bi-discriminator structure[17]. Additionally, a kind of Unified Gradient and Intensity Discriminator GAN uses a dual discriminator to differentiate between gradient and intensity to ensure that the generated image contains the desired geometric structure and salient information[18]. In 2021, Li *et al.* employed a convolutional network to extract features from each source image, which were then amplified using a meta-amplification module with an adaptable factor based on practical requirements. Additionally, they designed residual compensation blocks that were applied iteratively within the framework to enhance the extraction of fine details from the images[19]. In 2023, Xu *et al.* proposed an improved fusion model for GANs, introducing densely connected modules in the generator and discriminator network structure to connect features between layers, improve network efficiency and enhance the network's ability to extract source image information[20]. Yi *et al.* introduced an enhanced infrared and visible light GAN image fusion model incorporating a Dropout layer, effectively addressing the generator's performance degradation caused by discriminator overfitting, without increasing memory consumption or training time[21]. However, existing GAN-based methods still face significant challenges due to

information bias between infrared and visible images, often leading to unnatural visual effects. To overcome this limitation, Yin *et al*. proposed a novel cross-scale pyramid attention GAN-based infrared and visible image fusion method (CSPA-GAN)[22]. The model employs a single generator and dual discriminators to better approximate the distribution of the fused image, ensuring more natural and visually appealing results.

The current fusion procedure of infrared and visible images based on deep learning mentioned is mainly achieved by modifying the image fusion framework, configuring the fusion network architecture, and formulating loss functions to restrict the fusion process.

At the same time, as the technology advances, the requirements in target recognition and detection continue to grow. Implementing target recognition technology based on deep learning methods can effectively solve the problems and limitations brought by traditional target recognition technology. Currently, deep learning-based target recognition methods can be divided into two categories: two-stage recognition algorithms based on candidate regions and single-stage recognition algorithms based on regression. Girshick *et al*. improved the AlexNet network and designed the regions with CNN features (R-CNN) algorithm[23] to address the scaling problem of candidate regions. He *et al*. used the spatial pyramid pooling (SPP)[24] method to design the SPPNet CNN model. Subsequently, the Fast R-CNN algorithm proposed on this basis effectively combined the advantages of the R-CNN algorithm and the SPPNet algorithm[25]. The two-stage recognition algorithm based on candidate regions has high recognition accuracy, but it does not fully utilize the global information of the image and involves redundant calculations. The two-stage classification-based recognition architecture greatly affects the recognition speed. Therefore, researchers are gradually conducting research on single-stage target recognition technology. Regression-based single-stage recognition algorithms include the You Only Look Once (YOLO)[26] series, single shot multibox detector (SSD)[27], feature pyramid network (FPN)[28], and RetinaNet[29]. Among them, the YOLO series algorithms have been widely studied and applied due to their superior performance. By integrating the information of infrared and visible light images, target recognition can be conducted based on fused images, which is conducive to achieving comprehensive and deeper analysis and research of targets.

The existing fusion methods generally do not pay attention to the communication between network features, resulting in artifacts, blurring and other phenomena in the fused image, and focus on improving the network structure and loss function, ignoring the close connections between the source image quality and the feature extraction network. To solve the above problems, this paper proposed an infrared and visible image fusion algorithm based on a GAN under multi-level detail enhancement. Our contributions can be summarized as follows:

- The input source image is enhanced through a multi-level detail enhancement method to improve its quality, which helps the network learn more detailed edge features during training and image generation.
- Feature extraction modules such as residual-dense block and multi-scale residual block are added into the network structure to achieve refined feature extraction and ensure the coherence and transferability of image feature information.
- We designed a gradient and intensity information loss function based on the characteristics of the image. Together with the discriminator, the loss function constrains the network to generate fusion images that highlight both infrared targets and contain clear edge details.

The experimental results demonstrate that the proposed method achieves excellent fusion performance in subjective and objective evaluations. Specifically, the fused images outperform other methods significantly in terms of objective quality metrics, such as average gradient (AG), edge intensity (EI), and spatial frequency (SF). These images preserve more detailed textures and edge gradient information while ensuring that the infrared radiation targets are prominently highlighted, clear, and easily distinguishable. This demonstrates the superior fusion capabilities of the proposed approach.

2. RELATED THEORIES AND ANALYSIS

2.1. GANs

GANs are deep learning models proposed by Goodfellow *et al.* in 2014, evolving from the minimum binary zero-sum game in free game theory[13]. Unlike traditional deep neural networks, GANs consist of two components: generator (G) and discriminator (D). The generator receives random noise input and learns specific data feature distribution. The discriminator takes the generated results of the generator and real samples as input, determines whether the input data belongs to the real sample or the false sample, and feeds the judgment result back to the generator. These two components are trained alternately. In the constant confrontation game, they compete and drive each other to improve. During training, the parameters are continuously optimized until the generator and discriminator become indistinguishable, reaching a Nash equilibrium state.

In essence, the generator model belongs to the great likelihood estimation under the machine learning branch, which generates the original data distribution into the specified data by capturing the distribution of the sample data and adopting the parameter transformation calculation method in the great likelihood estimation. The generator model can be expressed as

$$\mathrm{x} = \mathrm{G}(\mathrm{z}; \theta^{(\mathrm{G})}) \tag{1}$$

where z is the random input noise, obeying a multivariate Gaussian distribution sampling. $\theta^{(G)}$ are the generator network parameters, and the network is passed through a series of nonlinear computations to obtain the output sample x. The essence of the discriminator is a binary classification model. It takes the generated sample x and real sample data as input, calculates the discriminant probability, and judges whether the sample is true or false. The discriminator can be expressed as

$$\mathrm{y} = \mathrm{D}(\mathrm{x}; \theta^{(\mathrm{D})}) \tag{2}$$

where $\theta^{(D)}$ is the network parameter of the discriminator and y is the output label, that is, the probability of discrimination.

The ultimate goal of GAN is to realize that the generated sample is infinitely close to the real data, so that the two are indistinguishable. The confrontation process between the two is expressed as

$$\min_{G} \max_{D} V(G, D) = E_{x \sim P_{\text{data}}} [\log D(x)] + E_{z \sim P_z} [\log (1 - G(z))] \tag{3}$$

where x is the input data, P_{data} is the real data distribution, z is the input noise, and P_z is the prior variable of the input noise. It can be seen from Equation (3) that the purpose of the generator is to minimize the objective function $V(G, D)$, making it difficult for the discriminator to distinguish real samples from the generated samples. The discriminator aims to improve the ability to distinguish between true and false samples, thereby maximizing the objective function $V(G, D)$. The two are trained alternately to optimize their own capabilities, and the final iteration is relatively stable.

2.2. Deep feature extraction

2.2.1. Residual dense block

As the depth of the network increases, the features of each convolutional layer have hierarchical structures with different receptive fields, and the traditional convolutional network structure cannot fully utilize the in-

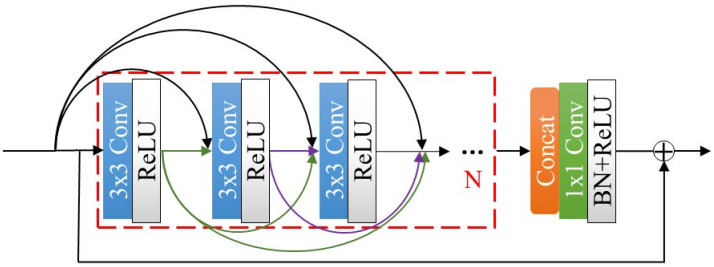

Figure 1. The overall structure of the dense residual block. 3 x 3: filter size. Conv: Convolutional layer; Concat: establish dense connections between the front and back layers; BN: batch normalization.

formation. In order to fully utilize the hierarchical features of all convolutional layers, Zhang *et al.* proposed a new feature extraction structure, named residual dense block (RDB)[30]. The structure extracts rich local features through convolutional layers densely connected, establishes dense connections between the front and the back layers, and reuses features in the channel dimension. Combined with residual connections, it enhances the information flow and ensures the transmission and coherence of feature information, effectively slowing down the phenomenon of gradient disappearance and enabling the network to achieve better performance with fewer parameters and calculations. The structure of the RDB module is shown in Figure 1.

The RDB module contains both dense and residual structures. Firstly, through the dense layer with N convolution kernel sizes of 3×3, the local feature learning can be realized effectively through the jump connection between Concat and the convolution layer in front. Then, the 1×1 convolution kernel is used for dimension reduction to prepare for the subsequent feature fusion. Finally, the global residual connection is used to improve the information flow and realize the fusion of local features and residual features.

2.2.2. Multi-scale residual blocks

In many visual tasks, it is very important to extract multi-scale features. Gao *et al.* proposed a novel multi-scale feature extraction method which is multi-scale residual block Res2Net in 2021[31]. Its structure is shown in Figure 2. As a variant of the residual network ResNet, Res2Net inserts more hierarchical residual connection structures into the residual unit. Unlike the typical multi-scale improvement of general architecture, this approach combines scales based on hierarchical structures. It performs multi-scale extensions with hierarchical and layered feature sets in a given block. This module divides the input into several equally sized groups. The first group is passed directly through, while the remaining groups are processed by the convolution kernel and combined with the next group. This process continues, and the processed features are finally concatenated.

As shown in Figure 2, after the first 1×1 convolution, it is divided into N subsets. Each subset can potentially receive all the feature information on its left after a 3×3 convolution operation, increase the receptive field, and obtain feature combinations with different quantities and sizes of receptive fields, so that the receptive field can represent multi-scale features at a finer granularity level. Finally, N groups of features are spliced through Concat, sent to 1×1 convolution, and then combined with the front residual connection to realize the fusion of feature information at various scales.

3. FUSION METHOD

3.1. Overall network framework

The overall framework of the fusion network is shown in Figure 3, which mainly includes three parts: image preprocessing, generator and discriminator.

First of all, the image preprocessing part mainly uses a multi-level detail enhancement method to enhance the

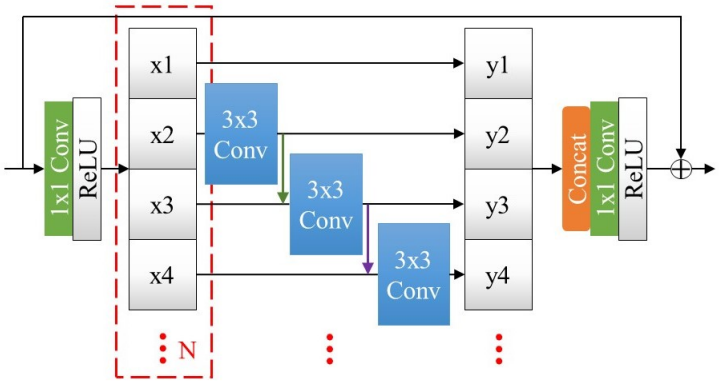

Figure 2. Schematic diagram of the structure of a multi-scale residual block. 3 x 3: filter size. Conv: Convolutional layer; Concat: establish dense connections between the front and back layers; BN: batch normalization. (xi expresses the N subsets into which it is divided after the first 1 x 1 convolution and yi expresses the N set of features obtained after xi undergoes a 3 x 3 convolution operation.)

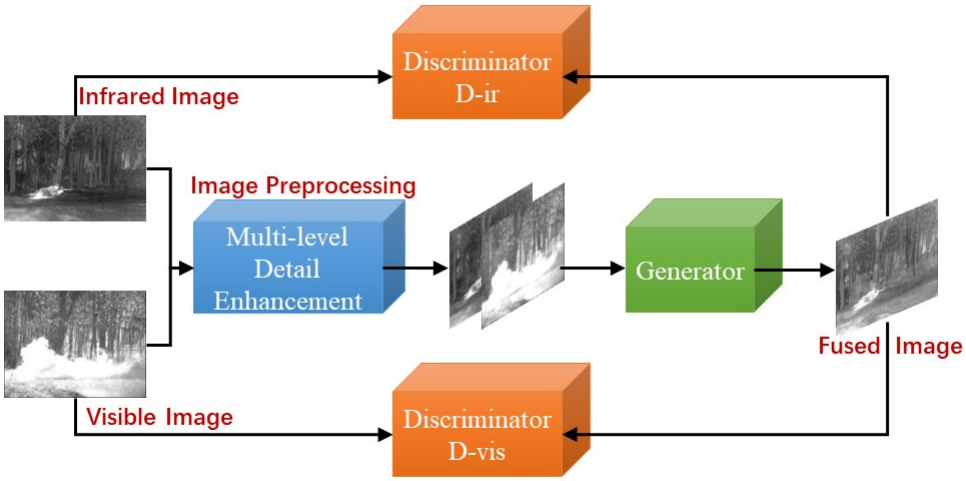

Figure 3. The overall framework of the fusion network.

edge details and other information of infrared and visible images, so that the input source image has more detailed information, which is beneficial to neural network processing. Secondly, Concat pairs of the registered infrared and visible images are used as the input of the generator for feature extraction and fusion image reconstruction. And the dual discriminator form of infrared discriminator D-ir and visible light discriminator D-vis is used to distinguish the infrared and visible images from the fusion image output by the generator, respectively, perceiving the feature differences. Ultimately, by engaging in an adversarial game with the generator and under the constraints imposed by the loss function, we can iteratively enrich and integrate image details to produce high-quality fused images that align with our expectations.

3.2. Multi-level detail enhancement module

In order to improve the quality of input source images and enhance the ability of neural networks to learn details, we can use the multi-level detail enhancement module to enhance the detail edge information of input source images. A method of multi-scale detail enhancement was proposed in literature [32], which decomposed the image into multiple scales through Gaussian filter kernel of different sizes, and fused detail information into the original image through a certain combination. Although this method could improve local details, it had limitations in improving global details and enhancing the balance between them. On this basis, this paper

proposed a multi-level detail enhancement method based on guided image filtering, which uses multiple levels to enhance image details without producing a gray saturation artifact phenomenon or excessive noise.

Guided filtering, as a nonlinear filtering technique, is widely used for edge preservation and image denoising. Unlike conventional denoising methods, which often struggle to distinguish between image edges and noise, guided filtering effectively preserves edges while suppressing noise. This is achieved by leveraging the structural information of a guidance image to adjust pixel weights within each region, thereby enabling image smoothing while retaining fine edge details. These characteristics make guided filtering highly effective in tasks such as image denoising, enhancement, and edge preservation.

In traditional guided filtering, the edge-preserving coefficient within any arbitrary window is fixed, which has a significant impact on the output image. A coefficient value that is too small may result in noticeable noise in the output image, whereas a value that is too large may lead to over-smoothing of the image. This study leverages the intrinsic properties of the edge-preserving coefficient in guided filtering by setting different filter sizes and edge-preserving coefficients to appropriate values. Through iterative and recursive filtering, this approach generates output images at various structural levels, enabling detailed and nuanced multi-level enhancement.

Firstly, the input image is processed through guided filtering and decomposed into different levels of background images, as given in

$$\begin{cases} B_1 = \text{GuildFilter}(I,\ I,\ \sigma,\ \varepsilon) \\ B_2 = \text{GuildFilter}(B_1,\ B_1,\ \sigma,\ \varepsilon) \\ B_3 = \text{GuildFilter}(B_2,\ B_2,\ \sigma,\ \varepsilon) \end{cases} \tag{4}$$

where I is the input image, and B_1, B_2 and B_3 are background images at different levels. σ and ε determine the filter size and edge retention coefficient, respectively. The values of (σ, ε) of the corresponding background images B_1, B_2 and B_3 are set to (3, 0.01), (9, 0.1) and (15, 0.5) respectively. The input image I, along with background images B_1 and B_2, is used as guide images, and each undergoes filtering to produce smooth background images B_1, B_2 and B_3.

Secondly, different layers of detail are extracted to prepare for subsequent enhancement; the detail layers are extracted by

$$\begin{cases} D_1 = I - B_1 \\ D_2 = I - B_2 \\ D_3 = I - B_3 \end{cases} \tag{5}$$

where I is the input image, and D_1, D_2 and D_3 represent high-, medium- and low-level details, respectively.

Finally, by using special weighting factors of set proportions, the details at levels D_1, D_2, and D_3, which correspond to fine details, moderate details, and large-scale edge structure details, are integrated into the original image through

$$I^* = (1 - w_1 \cdot sat\,(D_1)) \cdot D_1 + w_2 \cdot D_2 + w_3 \cdot D_3 + I \tag{6}$$

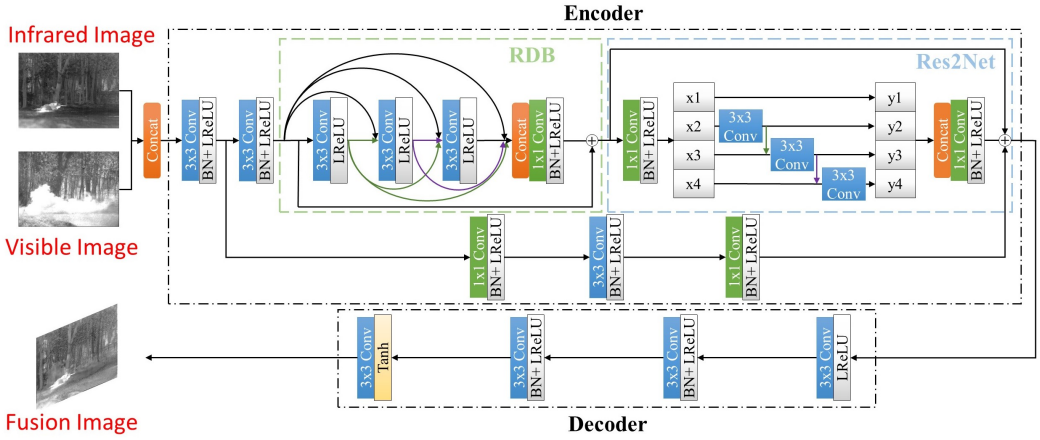

Figure 4. The overall architecture of the generator, including the number of layers of the encoder and decoder.

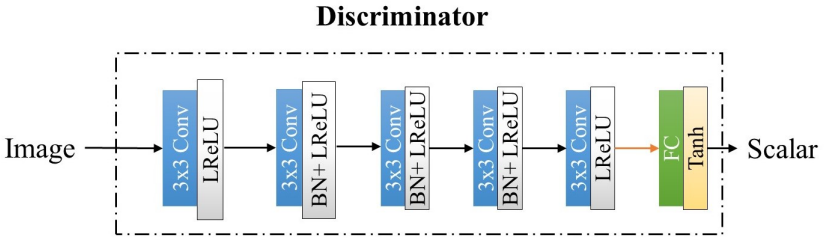

Figure 5. Discriminator structure diagram.

where I^* is the enhanced image, w_1, w_2 and w_3 are weight coefficients, with values of 0.4, 0.25 and 0.15 through many tests, and sat (·) represents the saturation function, defined as

$$sat(x) = \begin{cases} 1, & x > \Delta \\ kx, & |x| \le \Delta, k = \frac{1}{\Delta} \\ -1, & x < -\Delta \end{cases} \tag{7}$$

where Δ is the boundary threshold. As Δ approaches infinity, sat(·) is approximated as a symbolic function sgn(·). On the one hand, the sat(·) function is used to balance the positive and negative components of fine details D_1 to prevent artifacts caused by excessive grayscale saturation; on the other hand, it can also control the boundary threshold Δ to suppress a certain degree of noise caused by adding details, ensuring the balance of global and local details.

3.3. Generator and discriminator

3.3.1. Generator structure

The generator structure is shown in Figure 4. The overall structure of the model can be divided into two parts: the encoder (Encoder) and the decoder (Decoder). The encoder mainly extracts the feature information of the image; firstly, two ordinary convolutional layers are used to extract the shallow feature, and then, the shallow feature is sent into the RDB and the multi-scale residual block Res2Net. The RDB module adopts three dense layers with a convolution kernel size of 3 × 3; the number of subsets N divided by the Res2Net module is selected as 4. It achieved the extraction of multi-scale features and the fusion of deep features, ensuring the richness of local features while also enhancing the circulation and transmission of receptive fields and feature information. Finally, the output of the first ordinary convolutional layer is short-circuited with the output

of the Res2Net module by using global residual connection. Through the application of three convolutional layers with kernel sizes of 1 × 1, 3 × 3, and 1 × 1, respectively, the receptive field is enlarged and the feature map is enhanced, facilitating the integration of local and global features. This enables the resultant image to preserve superficial structural features while incorporating deep-seated detailed features, thereby ensuring the continuity of feature information.

The decoder decodes and reconstructs the fused feature map, reduces the channel number of the feature map through four convolution layers, and finally outputs the reconstructed feature fusion image.

3.3.2. Discriminator structure

The task of the discriminator itself is to effectively distinguish between real and false data (fusion data). Given that infrared and visible light images have different structures and features, separate discriminators are designed for infrared and visible light data. The contrast of the fused image can be improved by adversarial training of the D_i discriminator, and the texture details can be enriched by learning the D_V discriminator. The adversarial network containing two discriminators enables the generator to produce fused images that not only contain rich detail information but also have significant contrast for better results.

The discriminator structure is shown in Figure 5, which consists of five convolutional layers with a convolution kernel size of 3 × 3 and one fully connected layer. These convolutional layers use the LeakyReLU activation function to adjust the problem of zero gradients caused by negative input. Convolutional layers 2, 3 and 4 use batch normalization to speed up network convergence. The fully connected layer FC employs the Tanh activation function and performs a linear transformation to produce a scalar, indicating the probability of the input being real data as opposed to fused data.

The dual discriminators adopt the same structure but the parameters are not shared. The discriminator and the generator play an adversarial game to guide the network to continuously supplement the visible detail gradient information and the infrared target intensity information into the fusion image. Figure 6 illustrates the specific process of image feature extraction and fusion in the generator. Firstly, the image feature extraction operation is carried out in the encoder, and the superimposed image is processed through two ordinary convolutional layers to achieve shallow feature extraction, and then passed to the multi-scale feature extraction module, which consists of RDB and ResNet modules. A global residual connection is used to fuse local and residual features. Finally, the decoder reduces the dimensionality of the fused image features.

3.4. Loss function design

The loss function of a fusion network includes two parts: generator loss function L_G and discriminator loss function L_D. L_G is composed of content loss L_{con} and counter loss L_{adv}, as give in

$$L_G = L_{adv} + \lambda \cdot L_{con} \tag{8}$$

where L_G is the total loss of the generator, L_{adv} is the adversarial loss, L_{con} is the content loss, λ is the balance parameter used to balance the two losses, and $\lambda = 0.6$ in this article. Among them, L_{con} is the main loss function, so that the fused image can effectively retain the effective feature information of the source image, which is obtained as:

$$L_{con} = \frac{1}{HW}\left[\|G_F - i\|_{MSE} + \alpha \cdot \|G_F - i\|_{TV} + \beta \cdot \|\nabla G_F - \nabla v\|_F^2\right] + \gamma \cdot (1 - MS\_SSIM\,(G_F, v)) \tag{9}$$

where HW is the product of the height and width of the input image, G_F is the fusion image output by the gen-

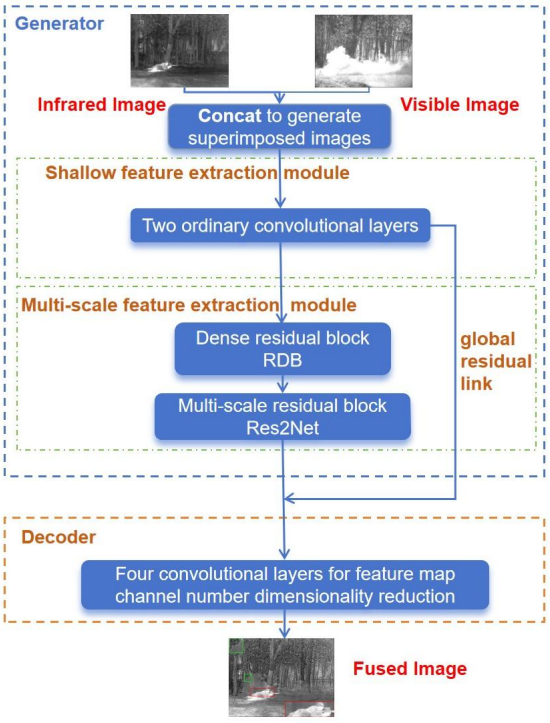

Figure 6. Flowchart of image feature extraction and fusion.

erator, v and i are the visible and infrared images, respectively, ∇ represents the gradient calculation, and MSE represents the mean square error loss, whose purpose is to constrain the difference in intensity distribution between fusion image G_F and infrared image i. TV is the Total Variation norm, which is used as a regularization loss to compensate for the image blur problem that may be caused by using MSE alone. F is the Frobenius norm, which enhances the learning of detailed texture features by constraining the gradient information of the fused image G_F and the visible image v. MS_SSIM is the multi-scale structural similarity, which constrains the similarity between the fused image G_F and the visible image v from multiple scale directions. α, β and γ are the parameters of the balance loss, which are respectively taken as 0.5, 0.5 and 100 in this paper.

In the content loss formulation, as given in Equation (9), the first and second terms $\frac{1}{HW}[\|G_F - i\|_{MSE} + \alpha \cdot \|G_F - i\|_{TV}]$ address the relationship between the fused and infrared images. The purpose of MSE loss is to constrain the intensity distribution differences between the fused and infrared images, enabling the network to effectively learn the infrared intensity distribution. However, relying solely on MSE loss may result in image blurring. To mitigate this, TV regularization is introduced as an additional term to preserve image sharpness and compensate for the limitations of MSE.

The third term $\frac{1}{HW} \cdot \beta \cdot \|\nabla G_F - \nabla v\|_F^2$ in Equation (9) enforces a constraint on the gradient information between the fused image and the visible light image using the F-paradigm. This term aims to enhance the model's ability to learn detailed texture features from the visible image, as gradient information encapsulates critical edge and detail characteristics.

Finally, the incorporation of the $MS - SSIM$ loss, based on human visual perception, enables a multi-scale evaluation of the similarity between the fused and visible light images. By considering structural features across multiple scales, $MS - SSIM$ enables the model to capture fine-grained details at various levels, ensuring that these details are effectively integrated into the fused image. This multi-faceted approach ensures a more comprehensive representation of the structural and textural attributes in the fused output.

L_{adv} is the loss function of information interaction between the discriminator and generator, aiming to promote the fusion image to retain more useful information, guide the generator's output, and reduce feature information loss in the counter. The process is expressed as

$$L_{adv} = -E_{G_F \sim P_{G_F}} \left[D_i \left(G_F \right) \right] - E_{G_F \sim P_{G_F}} \left[D_v \left(G_F \right) \right] \tag{10}$$

where P_{G_F} is the data distribution of the fused image, $D_i(G_F)$ and $D_v(G_F)$ are the discrimination probability values of the fused image G_F walk by the infrared and visible light discriminators, respectively.

L_D is to enable the discriminator to effectively identify the source and generated images, laying the foundation for the confrontation between the discriminator and the generator, which is obtained by

$$\begin{cases} L_{D_v} = -E_{v \sim P_v} \left[D_v(v) \right] + E_{G_F \sim P_{G_F}} \left[D_v \left(G_F \right) \right] \\ L_{D_i} = -E_{i \sim P_i} \left[D_i(i) \right] + E_{G_F \sim P_{G_F}} \left[D_i \left(G_F \right) \right] \end{cases} \tag{11}$$

where D_v and D_i are visible light and infrared discriminators, respectively.

4. EXPERIMENT AND ANALYSIS

4.1. Experimental settings

This paper selected 45 pairs of registered and corrected infrared and visible images under different scenes from the TNO dataset, which can be downloaded from the website https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029?file=1475454, and cropped them with a stride of 20, resulting in 37,164 pairs of infrared and visible images with a size of 108 × 108, which were normalized into the training set. During the training process, the Adam optimizer was used, the initial learning rate was set to 0.00016, the batch size was set to 36, and the iterations were 1,032 times, with a total of two rounds of training. Fusion algorithm model was built and trained under the Tensorflow framework, and the hardware platform is GeForce RTX 3090 GPU and Interi9-10900K CPU.

The test set selected 42 groups of infrared and visible images in the TNO data set, evaluated from both subjective and objective aspects. The fusion method in the paper is experimentally compared with six fusion methods: GAN with multiclassification constraints (GanMcC) method[15], DDCGAN method[16], U2Fusion method[33], NestFuse method[11], DIDFuse method[34], and FusionGAN method[35].

4.2. Comparative experiment and result analysis

The test set was used to conduct comparative experiments between the fusion method in the paper and other fusion methods, and three typical scenarios in the test set were selected for subjective evaluation analysis. Scenario 1 features dimly lit trees on both sides, making it difficult to distinguish the central figure from the background. Scenario 2 is affected by smoke interference, where soldiers are hard to identify, and the intricate details of interwoven background trees are challenging to discern. Scenario 3 focuses on identifying detailed features such as wheels and rooftops in low-light conditions. This analysis highlights the performance of the proposed method in addressing diverse and challenging visual fusion scenarios.

The experimental results of Scenario 1 are shown in Figure 7, in which the red box represents the images of infrared target extraction, the green box represents the visible light images, and the red and green arrows respectively indicate the enlarged details of the targets. Figure 7A and B represents the infrared and visible images, respectively. The infrared target in Figure 7C has insufficient thermal radiation information and low

Figure 7. Comparison of the results of the fusion experiment on seven fusion methods in scenario 1. The red box represents the magnified image of infrared target extraction, and the green box represents the magnified image of visible light details.

contrast. The images in Figure 7D and G are clearly distinguished from the background; the details of the trees are clear, but the overall tone is too dark, and the details of the grass only have a rough outline. As can be seen from Figure 7E and F, the fusion effect of the GanMcC method and the FusionGAN method is not good. The infrared character targets have a blur artifact phenomenon, and the grass of the visible light details enlarged by the green frame is also blurry. The details of the grass in Figure 7H have obvious outlines, but the edges of the details in the middle part are incoherent. Figure 7I shows the fusion algorithm of this paper. It not only has high-contrast infrared character targets, but also has the most prominent visual effect. The edges of the grass in visible light are clear, and the detailed textures are the most obvious and coherent, with strong recognition.

The experimental results of Scenario 2 are shown in Figure 8, where the infrared soldier target is effectively distinguished from the background, but details such as the outline of the soldier and the trees are blurred. The texture of the trees and the edge of the soldier appearance in Figure 8C and E compared with Figure 8F are clearer, but its infrared radiation information is not retained enough, and the contrast is low, which cannot highlight the soldier's infrared target. The smoke outlines in Figure 8D and G produced some artifacts, and the soldiers and the smoke could not be effectively distinguished. The smoke edges in Figure 8H are clear, but the soldiers are also disturbed by the smoke causing blurry contours of soldiers. The method presented in Figure 8I in this paper not only effectively retains the edge details of trees and smoke, but makes soldiers which are the infrared key targets clearly visible. The fused image has high contrast and good visual effect, which effectively

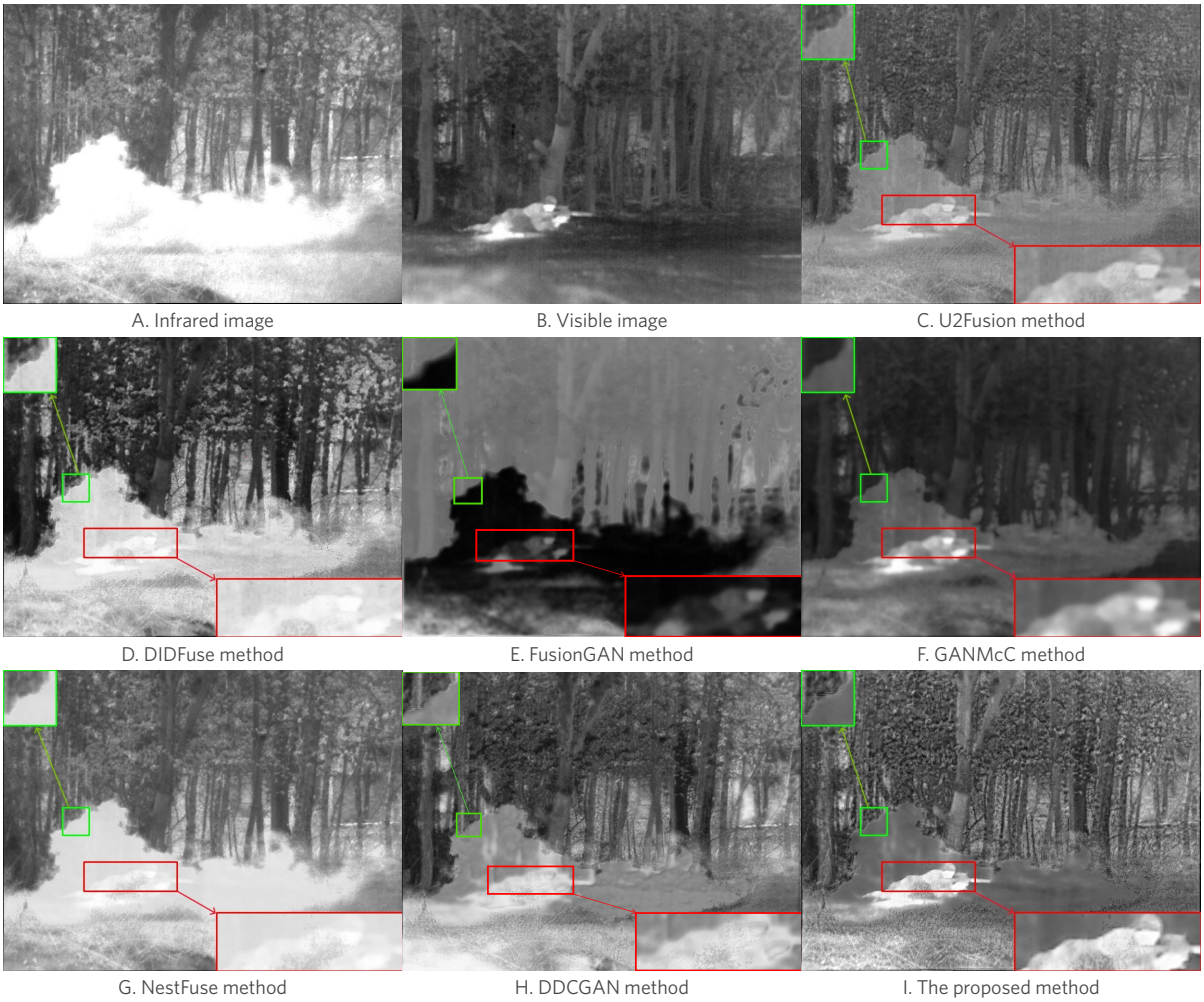

Figure 8. Comparison of the results of the fusion experiment on seven fusion methods in scenario 2. The red box represents the magnified image of infrared target extraction, and the green box represents the magnified image of visible light details.

inhibits the interference of smoke and highlights the target, laying a good foundation for subsequent visual tasks such as target recognition and target positioning.

The experimental results for Scene 3 are shown in Figure 9. From Figure 9E and F, it can be seen that the fusion effect of FusionGAN and GANMCc methods is not good, and the contours of tires and roofs are not clear. In Figure 9G, the roofs are clearer than in Figure 9E and F, but the tire contours are still fuzzy and some artifacts are generated, which have lost a lot of details. Infrared radiation information is not retained enough to highlight the tire target and cannot be effectively distinguished from the background [Figure 9C]. The tire can be distinguished from the background but the edge part is blurred, and the roof part lacks details [Figure 9H]. The contrast is high, and the contours of the tire and the roof are clear, but some details are lost [Figure 9D]. The method in this paper not only effectively preserves the edge details of the tire and the roof, but also has a high contrast and a better visual effect [Figure 9I], making it suitable for subsequent visual tasks such as target recognition, target location, and target positioning. These improvements lay a good foundation for further tasks.

Figure 9. Comparison of the results of the fusion experiment on seven fusion methods in scenario 3. The red box represents the magnified image of infrared target extraction, and the green box represents the magnified image of visible light details.

4.3. Index evaluation

Individual differences will affect the subjective evaluation results, so it is necessary to conduct objective quantitative data comparative analysis. We selected information entropy (EN), AG, SF, EI, standard deviation (SD) and visual information fidelity for fusion (VIFF) as objective evaluation indicators of fusion quality. Among them, the information EN represents the information richness of the image. AG and SF evaluate the clarity of the detailed texture of the image. EI reflects whether the edges of the image are obviously prominent. SD of high-contrast images is often higher, more eye-catching, and the visual effect is good. VIFF is used to evaluate image quality using the characteristics of the human vision system. The above six evaluation indicators are all positive indicators, and the larger the value, the better the fusion quality.

We conduct comparative experiments using the test set to obtain the objective evaluation index values of each fusion methods, with the average value used for quantitative analysis. The results are shown in Table 1. The horizontal coordinates of Figure 10 represent 42 image pairs of different scenes, which were selected for the comparative experiments. The values of each index were evaluated across all scenes, and a comprehensive analysis was subsequently conducted based on the average values presented in Table 1. In terms of information EN, the method in this paper yields the best result, because more details are added to the image by the fusion network, enriching its content. The three values of AG, SF and EI have all reached the highest, and have been greatly improved compared to other fusion methods, higher than the sub-optimal values of 65.41%, 65.09% and

Table 1. Comparison of the values of the objective evaluation indicators for each fusion algorithm

Methods	EN	AG	SF	EI	SD	VIFF
GanMcC	6.782	1.600	3.744	17.575	38.451	0.373
DDcGAN	6.941	2.646	5.864	28.506	36.518	0.502
U2Fusion	6.423	3.489	8.338	35.440	25.949	0.362
NestFuse	7.047	3.848	10.047	38.281	41.874	0.431
DIDFuse	6.958	4.267	11.315	42.636	46.620	0.505
FusionGAN	6.814	1.852	5.568	21.689	39.704	0.304
Proposed	7.193	7.058	18.680	66.181	40.855	0.623

EN: Entropy; AG: average gradient; SF: spatial frequency; EI: edge intensity; SD: standard deviation; VIFF: visual information fidelity for fusion.

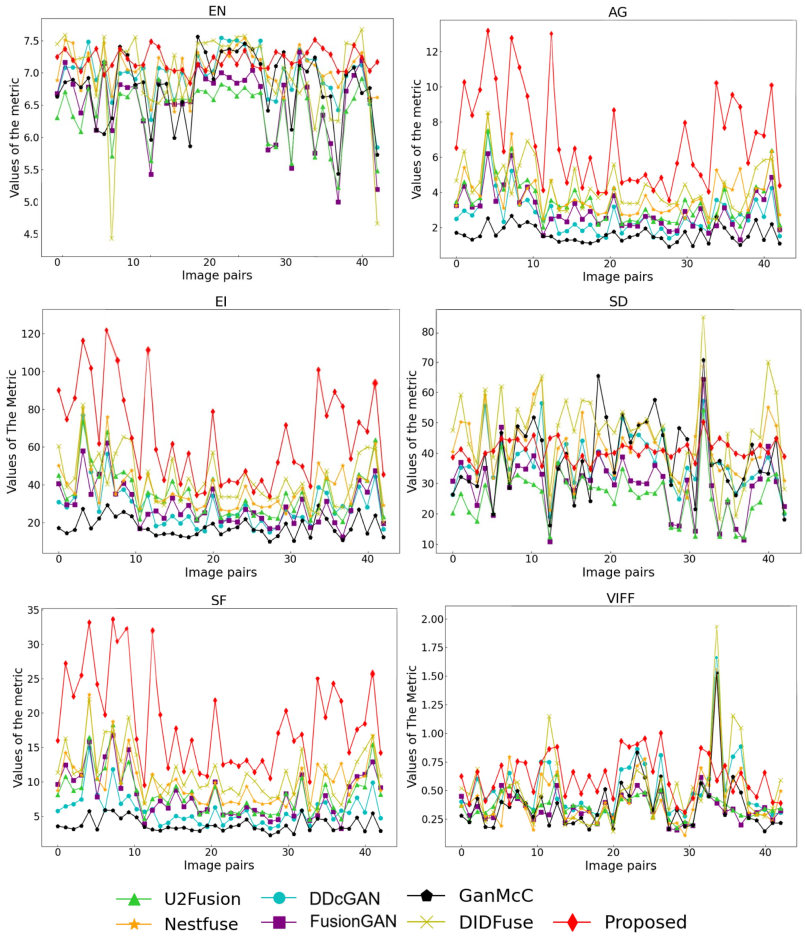

Figure 10. The proposed method is used for quantitative comparison of infrared and visible image fusion with the other six different fusion algorithms. The legend shows the mean values of the indicators for the different methods.

55.22%, respectively, indicating that the proposed method has obtained more detailed edge information, clearer texture details, and more prominent edge transitions. In terms of SD, the proposed method performs similarly to NestFuse, ranking second only to DIDFuse, indicating its ability to clearly distinguish the foreground from the background and achieve higher infrared target contrast. The method in this paper attains the optimal value in the VIFF index, aligning with the properties of the human visual system, and offering an excellent visual experience. Figure 8 shows the line chart of the objective evaluation indicators of each fusion algorithm. The trend observed in this chart, when combined with the results from Table 1, more intuitively and clearly

Table 2. Comparison of three sets of ablation experiments

Experiment	EN	AG	SF	EI	SD	VIFF
1	7.049	4.214	10.619	41.143	38.511	0.410
2	7.014	6.370	15.882	60.298	40.63	0.515
3	7.112	6.743	15.977	61.532	37.314	0.533
Proposed	7.193	7.058	18.680	66.181	40.855	0.623

Experiment 1: Fusion directly after image input. Experiment 2: Replace the RDB feature extraction module with the same number of layers instead of ordinary convolution. Experiment 3: Replace the Res2Net feature extraction module with the same number of layers instead of ordinary convolutions. EN: Entropy; AG: average gradient; SF: spatial frequency; EI: edge intensity; SD: standard deviation; VIFF: visual information fidelity for fusion.

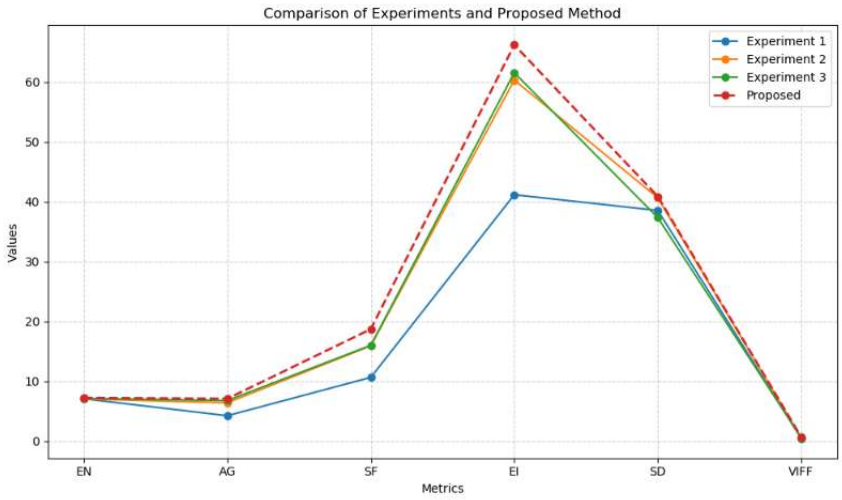

Figure 11. Schematic diagram of the results of the ablation experiment.

demonstrates the superiority of the proposed method, as evidenced by good fusion results in both overall performance and objective data.

4.4. Ablation experiment

In order to verify the function of the multi-level detail enhancement module, RDB and Res2Net modules in the algorithm network structure, the following three sets of ablation experiments were set up. Experiment 1: No multi-level detail enhancement was performed after the image input, and the source image was directly fused without preprocessing. Experiment 2: We removed the RDB feature extraction module, kept the other modules, and replaced them with ordinary convolutions of the same number of layers. Experiment 3: Instead of using the Res2Net feature extraction module, we replaced it with ordinary convolutions of the same number of layers. In the end, we used the test set to conduct objective evaluations by comparing the indicators of the three experimental methods, and the experimental results are shown in Table 2. Figure 11 visualizes the experimental results, and the specific values are presented in Table 2.

It can be seen from Experiment 1 that after using multi-level detail enhancement on the input image, the output fused image has higher values of various indicators, especially in the index of AG, SF and EI representing the edge of detail, proving that they make a lot of contributions to the retention of detail texture information. The enhanced image is also more informative. The contrast index SD is also enhanced to some extent, indicating that the quality of the fusion image has been greatly improved after the enhancement of multi-level details.

Table 3. Computational quantity comparison of three sets of experiments

Experiment	RDB-replaced model	Res2Net-replaced model	Proposed model
Params	1301252	1148740	1032932
GFlops	103.85	120.00	49.12

Params: The total number of weights and biases to be trained in the deep learning model. GFlops: The total amount of computation required by a model during forward propagation (inference) in units of 10^9 floating point operations. Params and GFlops are both important metrics for evaluating the computational effort of the model. RDB: Residual dense block.

Table 4. Comparison of convolutional layer configurations and params for RDB module between the proposed model in this paper and RDB-replaced model

Convolutional layer	Proposed model	RDB-replaced model
RDB_conv1	(3, 3, 64, 32, 18432)	(3, 3, 64, 96, 55296)
RDB_conv2	(3, 3, 96, 32, 27648)	(3, 3, 96, 128, 110592)
RDB_conv3	(3, 3, 64, 128, 36864)	(3, 3, 128, 160, 184320)
RDB_conv4	(1, 1, 64, 160, 10240)	(1, 1, 160, 64, 10240)
Sum of params	93184	360448

This table exhibits the network structure and the number of parameters for the four convolutional layers within the RDB module, denoted as (kernel height, kernel width, input channels, output channels, params). RDB: Residual dense block.

Table 5. Comparison of convolutional layer configurations and params for Res2Net module between the proposed model in this paper and Res2Net-replaced model

Convolutional layer	Proposed model	Res2Net-replaced model
Res2Net_conv1	(1, 1, 64, 128, 8192)	(1, 1, 64, 128, 8192)
Res2Net_conv2	(3, 3, 32, 32, 9216)	(3, 3, 128, 64, 73728)
Res2Net_conv3	(3, 3, 32, 32, 9216)	(3, 3, 64, 32, 18432)
Res2Net_conv4	(3, 3, 32, 32, 9216)	(3, 3, 32, 64, 18432)
Res2Net_conv5	(1, 1, 128, 64, 8192)	(1, 1, 64, 64, 4096)
Sum of params	44032	122880

This table exhibits the network structure and the number of parameters for the five convolutional layers within the Res2Net module, denoted as (kernel height, kernel width, input channels, output channels, params).

The comparison results of Experiments 2 and 3 with the original algorithm show that the RDB and Res2Net modules can effectively extract multi-scale features of the image, making the local and global features more closely connected, which is beneficial to the final fusion effect.

4.5. Computational quantity experiment

In order to analyze the impact of the RDB and Res2Net modules on the speed of inference, we set up three models for comparative analysis and calculate the variation in the number of parameters and GFlops, respectively: The Proposed Model in the paper: Include RDB and Res2Net modules. RDB-Replaced Model: The RDB feature extraction module is replaced by ordinary convolution corresponding to the same number of layers. Res2Net-Replaced Model: The Res2Net feature extraction module is replaced by ordinary convolution corresponding to the same number of layers. The results are presented in Table 3. The inclusion of RDB and Res2Net modules reduced both parameters and GFlops, improving inference speed compared to using standard convolutions.

The number of parameters in the model proposed was reduced by 20.6% and 10.08% relative to the two sets of comparison experiments, respectively. Tables 4 and 5 show the results of comparing the number of input channels to the convolutional layer, the number of output channels, and the number of output parameters per layer for the three sets of models, respectively. The RDB module extracts rich local features through densely connected convolutional layers and uses global residual connectivity to improve information flow; the channel count is optimized through concatenation and 1×1 convolutions, as shown in Table 4, thereby reducing the computational load of subsequent convolution compared to using ordinary convolutional layers. The core of the Res2Net module is the use of group convolution for multi-scale expansion and thus multi-scale feature extraction. The number of convolution kernels within each group is determined by that of input feature channels, which helps reduce the number of parameters. If a standard convolution layer was used instead of the Res2Net module, the absence of the group convolution structure would significantly increase the number of parameters. Specifically, after applying group convolutions, the number of parameters becomes 1/n of that of a standard convolution, where n is the number of groups. As shown in Table 5, in the proposed model, n is set to 4, resulting in a fourfold difference in the number of parameters.

Both the RDB and Res2Net modules significantly reduce computational complexity, with GFlops decreasing by 52.7% and 59.07%, respectively. The RDB module minimizes redundant computations through feature reuse and dense connections, reducing the number of channels processed in each convolution. The Res2Net module, by using group convolutions and multi-scale feature fusion, efficiently reuses feature information, reducing computation while enhancing model expressiveness without a large increase in parameters.

5. CONCLUSION

This paper proposes an infrared and visible light fusion method based on multi-level detail enhancement and GANs to solve the problems of blurred artifacts, unclear detail textures and unprominent target features in existing fusion methods. This method attaches great importance to the image preprocessing process, divides the image into multiple levels based on guided filtering, and adds the details of each level to the source image by combining the saturation function and weight allocation, so as to enhance the input infrared and visible images, making it more conducive to the training and learning of the feature extraction network. Secondly, we introduced modules such as global residual learning, dense residual blocks and multi-scale residual blocks as the feature extraction backbone of the generator, which enables the network to learn more abundant and comprehensive detail texture information. It enhanced the multi-scale feature extraction ability, realized the integration of local and global features, and ensured the transmission and circulation of feature information. Finally, we introduced the gradient calculation loss function and incorporated multi-scale structural similarity learning to enhance intensity information, co-constraining the fused image with the discriminator to preserve clearer gradient details and target strength characteristics.

The experimental results show that the fusion image obtained by this method achieves good fusion results both subjectively and objectively. The objective quality indicators such as image AG, edge strength and SF are much higher than those of other fusion methods. The fusion image contains more detailed texture and edge gradient information, and the infrared radiation targets in the image are prominent, clear and easy to identify, with excellent fusion performance.

This paper incorporates residual dense and multi-scale modules to introduce dense connections, enabling each layer to directly receive the outputs of all preceding layers. This design enhances the convergence speed and representational capacity of deep networks. However, the increased convolutional operations inevitably lead to a higher computational complexity. To address this, lightweight design principles and optimization strategies are adopted, effectively minimizing the impact on inference speed while preserving the advantages of the multi-scale modules. In the future, we will further consider practical applications in engineering, where

lightweight compression algorithm models are needed to meet hardware requirements. Meanwhile, we will consider guiding and improving the network models through specific image processing tasks after fusion such as recognition, detection, and tracking, so that they can be more effectively and reasonably applied.

DECLARATIONS

Authors' contributions

Made substantial contributions to conception and design of the study and performed data analysis and interpretation: Tian X, Xianyu X

Performed data acquisition and provided administrative, technical, and material support: Li Z, Jia Y, Xu T

Availability of data and materials

The TNO data set used in this article can be downloaded from the website https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029.

Financial support and sponsorship

This work was supported in part by the Natural Science Foundation of Jiangsu Province under Grant BK20210298 and the China Aeronautical Science Foundation (20240055052001).

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not appliable.

Consent for publication

Not applicable.

REFERENCES

1. Li L, Wang H, Li C. A review of deep learning fusion methods for infrared and visible images. *Infrared Laser Eng* 2022;51:20220125. DOI
2. Shen Y, Huang C, Huang F, Li J, Zhu M, Wang S. Research progress of infrared and visible image fusion technology. *Infrared Laser Eng* 2021;50:20200467. DOI
3. Liu B, Dong D, Chen J. Image fusion method based on directional contrast pyramid. *J Quantum Electron* 2017;34:405-13. Available from: https://m.researching.cn/articles/OJb55f12fb8e8c310f. [Last accessed on 26 Dec 2024]
4. Meng F, Song M, Guo B, Shi R, Shan D. Image fusion based on object region detection and non-subsampled contourlet transform. *Comput Electr Eng* 2017;62:375–83. DOI
5. Zhang Y, Qiu Q, Liu H, Ma X, Shao J. Brain image fusion based on multi-scale decomposition and improved sparse representation. *J Shaanxi Univ Technology* 2023;38:39-47. Available from: https://kns.cnki.net/kcms2/article/abstract?v=8XtZWovJaIRKW_m-UDySgTjWyqyco1C29tm9qtlAQukS1yBmvKDlfsLyujV75oXhuqr7_fIir8qYF-i4Vh6zcRFxkf38gN_JP301fxZmMDCamAZzIfynKMzcrepn3ta_QzURcktRLBXYwBhm5QweFEojPKfTZQ3aUF62LXfeTAwfYBi0SoRJzwD8WebuubMP&uniplatform=NZKPT&language=CHS. [Last accessed on 26 Dec 2024]
6. Yang P, Gao L, Zi L. Image fusion of convolutional sparsity and detail saliency map analysis. *J Image Graph* 2021;26:2433–49. Available from: https://kns.cnki.net/kcms2/article/abstract?v=8XtZWovJaIQhF4EB97rzeF9qazTDbDP00WW97CVhjFMlUYqfPZElERlDygQxUOVyCEdhfJJfqK-SpxKnGhI8gRrOD41-g36P17UI3EDaxNoeNi_NkrjEJ4YYJFVx-S54oABS3i1gJJJ4sLwLa2QTElcwweP6dI7weqH_sBywZEIcq39PaojJ9iBeJoq1HEu9S4wxRERmaeNvPvURIk72CerLzBqE0KI3vlAQZ5RHbC-1Y6hWdXw16Q==&uniplatform=NZKPT&language=CHS. [Last accessed on 26 Dec 2024]
7. Chen H, Deng L, Zhu L, Dong M. ECFuse: edge-consistent and correlation-driven fusion framework for infrared and visible image fusion. *Sensors* 2023;23:8071. DOI
8. Min L, Cao S, Zhao H, Liu P. Infrared and visible image fusion using improved generative adversarial networks. *Infrared Laser Eng* 2022;51:20210291. DOI

9. Liu Y, Chen X, Peng H, Wang Z. Multi-focus image fusion with a deep convolutional neural network. *Inform Fusion* 2017;36:191–207. DOI
10. Li H, Wu XJ. DenseFuse: a fusion approach to infrared and visible images. *IEEE Trans Image Process* 2018;28:2614-23. DOI
11. Li H, Wu XJ, Durrani T. NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. *IEEE Trans Instrum Meas* 2020;69:9645-56. DOI
12. Chang Z, Feng Z, Yang S, Gao Q. AFT: adaptive fusion transformer for visible and infrared images. *IEEE Trans Image Process* 2023;32:2077-92. DOI
13. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Commun ACM* 2020;63:139-44. DOI
14. Ma J, Yu W, Liang P, Li C, Jiang J. FusionGAN: a generative adversarial network for infrared and visible image fusion. *Inform Fusion* 2018;48:11–26. DOI
15. Ma J, Zhang H, Shao Z, Liang P, Xu H. GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion. *IEEE Trans Instrum Meas* 2020;70:1–14. DOI
16. Ma J, Xu H, Jiang J, Mei X, Zhang XP. DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. *IEEE Trans Image Process* 2020;29:4980–95. DOI
17. Zhou H, Hou J, Zhang Y, Ma J, Ling H. Unified gradient-and intensity-discriminator generative adversarial network for image fusion. *Inform Fusion* 2022;88:184–201. DOI
18. Rao D, Xu T, Wu XJ. TGFuse: an infrared and visible image fusion approach based on transformer and generative adversarial network. *IEEE Trans Image Process* 2023. DOI
19. Li H, Cen Y, Liu Y, Chen X, Yu Z. Different input resolutions and arbitrary output resolution: a meta learning-based deep framework for infrared and visible image fusion. *IEEE Trans Image Process* 2021;30:4070–83. DOI
20. Xu X, Shen Y, Han S. Dense-FG: a fusion GAN model by using densely connected blocks to fuse infrared and visible images. *Appl Sci* 2023;13:4684. DOI
21. Yi Y, Li Y, Du J, Wang S. An infrared and visible image fusion method based on improved GAN with dropout layer. In: The Proceedings of the 18th Annual Conference of China Electrotechnical Society. Springer; 2024. p. 1–8. DOI
22. Yin H, Xiao J, Chen H. CSPA-GAN: a cross-scale pyramid attention GAN for infrared and visible image fusion. *IEEE Trans Instrum Meas* 2023;72:1–11. DOI
23. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2014. pp. 580–7. Available from: https://openaccess.thecvf.com/content_cvpr_2014/html/Girshick_Rich_Feature_Hierarchies_2014_CVPR_paper.html. [Last accessed on 26 Dec 2024]
24. He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Pattern Anal Mach. Intell* 2015;37:1904–16. DOI
25. Girshick R. Fast R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision; 2015. pp. 1440-48. Available from: https://openaccess.thecvf.com/content_iccv_2015/papers/Girshick_Fast_R-CNN_ICCV_2015_paper.pdf. [Last accessed on 26 Dec 2024]
26. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016. pp. 779–88. Available from: https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Redmon_You_Only_Look_CVPR_2016_paper.pdf. [Last accessed on 26 Dec 2024]
27. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Computer Vision - ECCV 2016: 14th European Conference; 2016 Oct 11-14; Amsterdam, the Netherlands. Springer; 2016. pp. 21–37. DOI
28. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2017. pp. 2117–25. Available from: https://openaccess.thecvf.com/content_cvpr_2017/papers/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.pdf. [Last accessed on 26 Dec 2024]
29. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision; 2017. pp. 2980–8. Available from: https://openaccess.thecvf.com/content_ICCV_2017/papers/Lin_Focal_Loss_for_ICCV_2017_paper.pdf. [Last accessed on 26 Dec 2024]
30. Zhang Y, Tian Y, Kong Y, Zhong B, Fu Y. Residual dense network for image super-resolution. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018. pp. 2472–81. Available from: https://openaccess.thecvf.com/content_cvpr_2018/papers/Zhang_Residual_Dense_Network_CVPR_2018_paper.pdf. [Last accessed on 26 Dec 2024]
31. Gao SH, Cheng MM, Zhao K, Zhang XY, Yang MH, Torr P. Res2net: a new multi-scale backbone architecture. *IEEE Trans Pattern Anal Mach Intell* 2019;43:652–62. DOI
32. Kim Y, Koh YJ, Lee C, Kim S, Kim CS. Dark image enhancement based onpairwise target contrast and multi-scale detail boosting. In: 2015 IEEE international conference on image processing (ICIP); 2015 Sep 27-30; Quebec City, Canada. IEEE; 2015. pp. 1404–8. DOI
33. Xu H, Ma J, Jiang J, Guo X, Ling H. U2Fusion: a unified unsupervised image fusion network. *IEEE Trans Pattern Anal Mach Intell* 2020;44:502–18. DOI
34. Zhao Z, Xu S, Zhang C, Liu J, Li P, Zhang J. DIDFuse: deep image decomposition for infrared and visible image fusion. *arXiv* 2020. arXiv:2003.09210. Available from: https://doi.org/10.48550/arXiv.2003.09210. [Last accessed on 26 Dec 2024]
35. Fu Y, Wu XJ, Durrani T. Image fusion based on generative adversarial network consistent with perception. *Inform Fusion* 2021;72:110–25. DOI