Heliyon 10 (2024) e30798

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon

Research article

# Image fusion using Y-net-based extractor and global-local discriminator

Danqing Yang [a], Naibo Zhu [b,*], Xiaorui Wang [a,**], Shuang Li [b]

[a] School of Optoelectronic Engineering, Xidian University, Xi'an, 710071, China
[b] Research Institute of System Engineering, PLA Academy of Military Science, Beijing, 100091, China

ARTICLE INFO

*Keywords:*
Infrared and visible image fusion
Y-net
Multi-scale representation
Global-to-local detection
Contextual attention (CoA)
Generative adversarial network (GAN)

ABSTRACT

Although some deep learning-based image fusion approaches have realized promising results, how to extract information-rich features from different source images while preserving them in the fused image with less distortions remains challenging issue that needs to be addressed. Here, we propose a well worked-out GAN-based scheme with multi-scale feature extractor and global-local discriminator for infrared and visible image fusion. We use Y-Net as the backbone architecture to design the generator network, and introduce the residual dense block (RDblock) to yield more realistic fused images for infrared and visible images by learning discriminative multi-scale representations that are closer to the essence of different modal images. During feature reconstruction, the cross-modality shortcuts with contextual attention (CMSCA) are employed to selectively aggregate features at different scales and different levels to construct information-rich fused images with better visual effect. To ameliorate the information content of the fused image, we not only constrain the structure and contrast information using structural similarity index, but also evaluate the intensity and gradient similarities at both feature and image levels. Two global-local discriminators that combine global GAN with PatchGAN as a unified architecture help to dig for finer differences between the generated image and reference images, which force the generator to learn both the local radiation information and pervasive global details in two source images. It is worth mentioning that image fusion is achieved during confrontation without fusion rules. Lots of assessment tests demonstrate that the reported fusion scheme achieves superior performance against state-of-the-art works in meaningful information preservation.

## 1. Introduction

Practical applications such as video surveillance [1], vehicle night navigation [2] and fire rescue [3] often require a combination of infrared and visible sensors to adequately express scene information to enhance human and robotic visual understanding. Nevertheless, information redundancy is also a ubiquitous problem in multi-source data. Proverbially, infrared images captured by sensors that receive infrared wavelength information from object emission contain thermal radiation information characterized by significant intensities, which can avoid visual identification obstacles caused by illumination changes or camouflage. However, infrared images exhibit low-resolution and poor details, and are generally not convenient for human visual observation. In contrast, visible images are

* Corresponding author.
** Corresponding author.
*E-mail address:* 1922550996@qq.com (D. Yang).

https://doi.org/10.1016/j.heliyon.2024.e30798
Received 9 February 2024; Received in revised form 27 April 2024; Accepted 6 May 2024
Available online 11 May 2024

presented by receiving visible light band information reflected by objects, so these high-definition images display rich texture details of the object appearance represented by gradients, but they are susceptible to obstacles and environmental factors. Image fusion techniques can integrate vital information from different modalities to reduce redundancy and are widely used in multifarious imaging devices.

The taxing point of infrared and visible image fusion lies in extracting information-rich features at different scales from the different modalities and generating a fused image that merges them to improve image aesthetics and understanding without introducing any distortions/artifacts. Over the years, numerous solutions have been proposed to solve the above crucial aspect, including traditional approaches and data-driven methods based on deep learning (DL) [4,5].

According to different image processing methods, traditional approaches mainly include multi-scale transformation (MST), sparse representation (SR), saliency-based methods, subspace projecting (SP), hybrid models, and others. MST is one of the common practices in most traditional algorithms due to its flexibility and good visualization. Existing multi-scale techniques methods (including two-scale) can preserve the details of the source images and circumvent spectral degradation to some extent. For example, image pyramid transform [6], nonsubsampled contourlet transform [7], shearlet transform [8], guidance filter [9] and multiscale decomposition [10] are commonly-used MST techniques in image fusion task. Nonetheless, fusion algorithms based on MST typically have limited fusion performance for three reasons. First, this type of method, which blindly assembles transformations or representations of the source images with some dedicated rules instead of learning them from source images, is bearing heavy computation burden. Second, asymmetric feature information overlapping at multiple scales often results in halos and blurred edges (high-redundancy). Third, detail loss is inevitable in the multiscale transform, manual fusing, and inverse transform processes.

Recently, DL has become a mainstream method in image fusion tasks. On the one hand, DL methods can produce more filters for image feature extraction than that of traditional MST techniques. On the other hand, the parameters of filters in DL approaches can be learned adaptively to achieve various image fusion subtasks. Feature extraction and fusion at different scales based on DL have demonstrated their superiority in improving fusion performance. For example, off-the-shelf fusion models based on CNN backbonewhich show excellent fusion performance utilize multiple scales at the convolution kernels or feature levels to capture the meaningful information of the different modal images. UNIFusion [11] utilized Ghost module instead of the classical convolution layer to generate more feature maps. Fu [12] extracted the different level features via d dense connection operations. Zheng [13] achieved feature extraction at different scales and levels using HINBlock. Reference [14–17] employed convolution kernels of different sizes to extract common and unique features of source images. Reference [18–20] captured the multilevel features of the source images via residual learning. Moreover, modern GAN-based approaches [21–30] exploit multi-granularity convolution kernels of the same feature level, yielding different receptive fields and in turn improving fusion performance. For example, each network layer of the feature extractor in Refs. [21–24] utilized convolution kernels of different sizes to extract useful information from source images. Li [25,26] introduced multi-grained attention network to enable the fusion model to perceive the target region or detail information of the source image from multiple scales. TC-GAN [27] built the generator using convolutional layers with different scale filters to generate the combined texture maps in greater detail. Liu [28] proposed a contextualized dilated feature extraction module to obtain coarse-to-fine features. In order to balance the feature extraction capability and the number of parameters of the fusion model, reference [29,30] firstly utilized large-scale filters to expand the receptive field of the shallow network, and then used small-scale filters to further extract the deep features.

Although the above successful cases have witnessed a great improvement in the fusion effect, there are still some drawbacks that deserve to be emphasized. First of all, some fusion approaches do not fully extract the finer features and rich semantic information in the source images due to lack of downsampling operation. Fused results are often contaminated by blurring effects. In addition, during the reconstruction phase, CNN kernel-based multi-scale representation methods only utilize deep features and unreasonable fusion strategies to reconstruct the fused images, which fails to render the resulting images photorealistic. The third item is that some loss functions constrain the similarity between the generated image and reference images only at pixel domain, while neglecting the improvement in perceptual quality of the fusion image. Last but not least, the discriminator has limited ability to guide and facilitate the optimization of the generator during the adversarial process, as some GAN-based methods fail to simultaneously capture different scale features from the multimodal images in a global-to-local manner.

Based on the above weak points, we focus on improving the information richness of fused images from two aspects: network structure construction and loss function design. On the one hand, based on Y-Net, a generator network is constructed to directly downsample feature maps that have the same size as the inputs to obtain distinctive features at different scales, instead of using multiple scales at the convolution kernels or feature levels. Moreover, to preserve finer complementary features, we design two discriminator networks that combine global GAN with PatchGAN as a unified architecture to capture both local thermal radiation information and holistic features in the source images. On the other hand, since similarity constraints based on objective evaluation index have a comparatively limited ability to capture perception-related differences, we design a hybrid loss function based on both pixel and feature domains to achieve the retention of vital information in the source images and the improvement of perceptual quality of the fusion results.

In a nutshell, the main novelty of the work consists of the following four-fold.

(1) With Y-Net as the backbone, the generator capitalizes on residual dense block (RDblock) to extract shallow texture details and deep target structures at different scales from the source images.
(2) The cross-modality shortcut with contextual attention (CMSCA) is devised to strengthen the discriminative encoding features at different scales. By doing so, both shallow and deep enhancement features are used to maintain the saliency of the infrared targets and preserve rich details.

(3) We innovatively combine global GAN with PatchGAN to construct dual discriminator, so as to fully consider the information levels of the source images and enhance discriminative ability.
(4) A hybrid constraint is designed to guide the learning process of the proposed end-to-end fusion model from image and deep feature domains, respectively. As a result, the proposed method improves the information richness of the fusion images.

## 2. Technical backgrounds

### 2.1. Deep learning-based image fusion methods

Recently, it is a tendency to build performance-efficient deep neural networks for various image fusion tasks due to their strong nonlinear learning abilities. Learning-based fusion architectures, such as autoencoder (AE) [13,14,16,19], convolutional neural network (CNN) [15,18,20] and generative adversarial network (GAN) [21,22,24,27,29] have witnessed obvious improvements in fusion performance, but their single-scale frameworks can hardly capture the full-scale features of the real-world targets and fail to make the fused images photorealistic. More importantly, most methods directly capitalize on the features extracted in the last layer to reconstruct fused images, whereas earlier features do not. Consequently, some useful multi-layer information is lost in the deep cascaded network, resulting in unfriendly visual perception. In addition, some non-end-to-end methods [11,15–17,27,28] generate unsatisfied fusion results due to unreasonable fusion rules. To this end, in this work, we focus on developing more effective GAN frameworks that explicitly deal with the scale-space problems faced by visible and infrared image fusion task in an end-to-end fashion.

### 2.2. U-net framework

U-Net is originally proposed for image segmentation tasks [31]. With the powerful multi-scale representation advantage, more and more computer vision tasks are realized by using U-Net as the backbone network, such as image dehazing [32], salient object detection [33], facial emotion recognition [34], image denoising [35], image fusion [13,36–39]. U-Net architecture adopts a symmetric encoder-decoder manner that overcomes the disadvantages of local and global features loss in fully convolutional networks. In the contraction path (encoder), the features at different scales are extracted from the source images through the downsampling operations, and the resolution of the feature maps is gradually lessened. In the expand path (decoder), the image details are repaired by the up-sampling operations and the reconstructed image is restored to the input size. Furthermore, the skip connections in the U-Net architecture largely compensate for the information loss caused by the downsampling operation during fusion image restoration.

Although deep models based on U-Net have achieved remarkable performance in various application fields, there are inherent limitations in several aspects. In image fusion, dual convolution operation in U-Net has limited feature extraction capabilities, making it difficult to mine the intrinsic features of images with different modalities. Besides, existing U-Net often utilize simple skip connections to transfer features from convolutional blocks to their corresponding deconvolutional blocks in an elementwise max/concatenation way. They aggregate features of different scales without considering their discriminative contextual information that are crucial for infrared and visible image fusion task.

As an improved version of U-Net, Y-Net with two encoders and one decoder can capture deep discriminative features from different source inputs. Therefore, we build the generator network based on the idea of Y-Net. Unlike previous architectures, the improvements to our generator lie in: 1) since simply stacking convolutional blocks to build deep networks is difficult to obtain good fusion performance, we introduce the dense-residual blocks (RDblocks) into Y-Net to enhance the abilities of feature extraction; 2) all the features extracted by the scaled-down layers are aggregated to their corresponding deconvolutional layers via cross-modality shortcuts with contextual attention (CMSCA), which can keep vital information in a fine-to-coarse manner for fusion image reconstruction and make the proposed model easier to be trained.

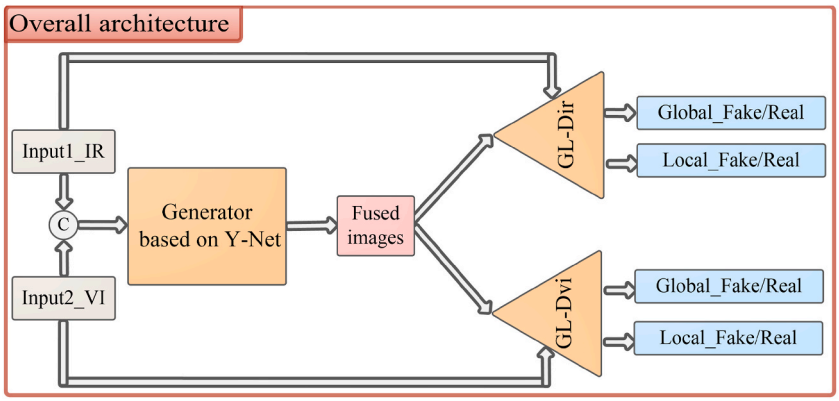


**Fig. 1.** The blueprint of the proposed method for infrared and visible image fusion.

## 3. Approach

### 3.1. Our motivation

Since no ground-truth image can serve as the optimization objective in infrared and visible image fusion task, the key point is to design deeper networks to fully mine the meaningful information within the source images and selectively retain it in the fusion result. Generally, the intensities of the dual source images at the same position often change significantly due to different imaging mechanisms. So, many deep methods pursue visually better fusion results in a multi-scale manner. Unfortunately, they have limited fusion performance for the following reasons. Above all, they all operate at the kernel level, such as convolution kernels of different sizes or dilated convolutions, to obtain multi-resolution features regardless of the sampling operations. It leads to the failure to capture high-level semantic features of the source images. Besides, existing GAN-based approaches classify features from the view of the overall image, but neglect the local information across different patches, resulting in the appearance of artifacts in their fusion results.

For multimodal image fusion problems, it is inefficient to achieve salient performance gains by simply stacking more convolution layers or building wider network layers. It is of great significance to design customized networks for fusion problems. Therefore, in this paper, the GAN architecture is used to characterize the features of the original images from the multi-resolution perspective. Fig. 1 shows the blueprint of the proposed method. Our entire model is composed of two functional modules, one generator based on Y-Net is to learn a powerful feature extractor that can generate realistic fusion images guided by both image and feature-level loss functions, and two discriminators aim to keep implicit details and enhanced radiative information of the source images from coarse to fine for visual performance. Lots of assessment tests demonstrate that our method not only extracts information-rich multi-scale features with low-redundancy from the two source images, but also ensures that they can be transferred to the fused images without loss of fidelity.

Simplistically, some symbols that appear frequently in this work are stated in advance as follows. **IR** represents the source infrared images, and **VI** represents the source visible images. **G** stands for generator. © denotes the concatenation operation. $\oplus$ denotes the element-wise add. $\otimes$ denotes the element-wise multiplication. **EN_ir** stands for the encoding path of the IR images, and **EN_vi** stands for the encoding path of the VI images. **RDblock** represents the residual dense block. **CMSCA** indicates the cross-modality skip-connection with the contextual attention, **CoA** indicates contextual attention, **GL-Dvi** stands for a discriminator that is used to capture more subtle differences between the generated and VI images from global to local level. Similarly, **GL-Dir** stands for a discriminator that can capture more subtle differences between the generated and IR images from global to local level.

### 3.2. Network architecture

#### 3.2.1. Generator architecture

Unlike other image fusion subtasks, the fusion results for IR and VI can retain salient features of the input images only through the deeper network to extract features containing more intrinsic information. In addition, for IR and VI with different modalities, there may be differences in pixel intensities in certain regions such that information from different image scales cannot be ignored. Hence, it is natural to think of employing Y-Net to achieve this fusion task. The Y-shaped structure fully takes into account the modal differences and completely addresses the problem that features at different scales have no interaction in the previous multi-resolution representation methods. Fig. 2 shows the architecture of the well-designed generator (G), and the architecture of the generator is shown in Table 1. Clearly, G builds a simple Y-Net structure for multi-scale representation of the original images, which mainly consists of three parts: encoder (EN_IR and EN_VI), cross-modality shortcuts with contextual attention (CMSCA) and decoder, three of which are elaborated below. In general, the preservation of shallow features improves the image quality, while deeper features help maintain the saliency of the thermal targets in the fused images. Therefore, the well-designed G not only extracts features at shallow and deep levels, but also uses them to restore the fused image.

(1) Scale-down path

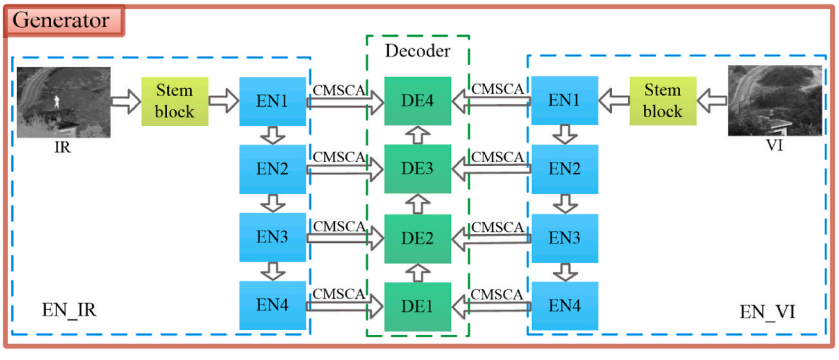


**Fig. 2.** The architecture of generator. The symbol EN represents the encoding module, DE represents the decoding module, Stem block represents the convolution operation for extracting coarse features from source images.

**Table 1**
The parameter settings for all layers of the generator. CB denotes convolution layer and activation function. IR_Pooling denotes average pooling operation, VI_Pooling denotes max pooling operation. RDblock denotes the residual dense block. Deconv denotes the deconvolution operation. ResBlock denotes the fusion image restore block.

| Subnetwork | Layer | Input Channel | Output Channel | Filter Size | Stride | Padding | Activation Function |
|---|---|---|---|---|---|---|---|
| Encoder_IR | CB1 | 1 | 16 | 5 | 1 | SAME | ReLU |
| | IR_Pooling1 | – | – | 2 | 2 | VALID | – |
| | CB2 | 16 | 32 | 3 | 1 | SAME | ReLU |
| | RDblock1 | 32 | 32 | – | – | SAME | ReLU |
| | IR_Pooling2 | – | – | 2 | 2 | VALID | – |
| | CB3 | 32 | 64 | 3 | 1 | SAME | ReLU |
| | RDblock2 | 64 | 64 | – | – | SAME | ReLU |
| | IR_Pooling3 | – | – | 2 | 2 | VALID | – |
| | CB4 | 64 | 128 | 3 | 1 | SAME | ReLU |
| | RDblock3 | 128 | 128 | – | – | SAME | ReLU |
| | IR_Pooling4 | – | – | 2 | 2 | VALID | – |
| | CB5 | 128 | 256 | 3 | 1 | SAME | ReLU |
| | RDblock4 | 256 | 256 | – | – | SAME | ReLU |
| Encoder_VI | CB1 | 1 | 16 | 5 | 1 | SAME | ReLU |
| | VI_Pooling1 | – | – | 2 | 2 | VALID | – |
| | CB2 | 16 | 32 | 3 | 1 | SAME | ReLU |
| | RDblock1 | 32 | 32 | – | – | SAME | ReLU |
| | VI_Pooling2 | – | – | 2 | 2 | VALID | – |
| | CB3 | 32 | 64 | 3 | 1 | SAME | ReLU |
| | RDblock2 | 64 | 64 | – | – | SAME | ReLU |
| | VI_Pooling3 | – | – | 2 | 2 | VALID | – |
| | CB4 | 64 | 128 | 3 | 1 | SAME | ReLU |
| | RDblock3 | 128 | 128 | – | – | SAME | ReLU |
| | VI_Pooling4 | – | – | 2 | 2 | VALID | – |
| | CB5 | 128 | 256 | 3 | 1 | SAME | ReLU |
| | RDblock4 | 256 | 256 | – | – | SAME | ReLU |
| Decoder | Deconv1 | 256 | 128 | 3 | 2 | SAME | – |
| | RDblock1 | 128 | 128 | – | – | SAME | ReLU |
| | Deconv2 | 128 | 64 | 3 | 2 | SAME | – |
| | RDblock2 | 64 | 64 | – | – | SAME | ReLU |
| | Deconv3 | 64 | 32 | 3 | 2 | SAME | – |
| | RDblock3 | 32 | 32 | – | – | SAME | ReLU |
| | Deconv4 | 32 | 16 | 3 | 2 | SAME | – |
| | RDblock4 | 16 | 16 | – | – | SAME | ReLU |
| | ResBlock | 16 | 1 | 1 | 1 | SAME | Tanh |

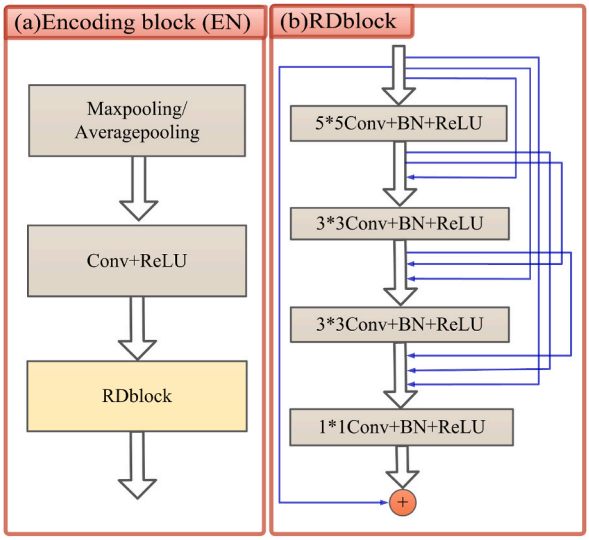


**Fig. 3.** Illustration of the encoding block in Fig. 2 a) the structure of encoding block; b) the structure of RDblock.

Given that features from images with different modalities encode discriminative information for feature multi-scale representation, it can help to mine more useful information in the original images. Thus, we use a two-stream network to encode IR and VI images, respectively. Fig. 3 (a) shows the structure of the encoding block. In the encoding paths, the stem block (i.e., 5 × 5 convolution layer) of the two encoders operates on a fine scale to extract features that have the same size as the source images, while later blocks transition (through pooling) to coarse scales to extract high-level semantic information. Both scale features are required, but occur at different positions in the EN_ir and EN_vi. Considering the computational complexity and time efficiency comprehensively, we consecutively downsample the two source images for four times (the resolution of the feature map is halved at each downsampling) to achieve feature extraction at different scales and levels. It is public knowledge that average pooling can preserve low-frequency information, while max pooling helps keep high-frequency information. Therefore, we utilize max pooling operation for downsampling in EN_vi and average pooling for downsampling in EN_ir. In addition, a convolution block (convolution layer and ReLU activation function) is adopted to guarantee that the number of channels is doubled with the increasement of the hierarchy. The problem of the original continuous convolutional encoding network is that the early texture details are lost. In addition, the batch normalization (BN) is abandoned, resulting in unstable training and low convergence efficiency. Residual connections and dense networks are the basic ideas of many deep frameworks to enhance the abilities of feature extraction by increasing the depth and width of the network. Therefore, the residual dense blocks (RDblocks) are introduced into the encoding blocks to fully extract more shallow details and deep semantic information, so as to obtain richer image representations. Fig. 3 (b) shows the structure of the RDblock. It is pretty easy to note that the dense network consists of a 5 × 5 convolution block (5*5Conv + BN + ReLU) that is used to increase the receptive field, and two 3× 3 convolution blocks (3*3Conv + BN + ReLU) that are adopted to extract deep features. Finally, a 1 × 1 convolution block (1*1Conv + BN + ReLU) is used to adjust the output of the dense network to satisfy the residual connection. These RDblocks make the proposed model easier to train and accelerate convergence.

(2) Scale-up path

During decoding, if the high-level features can be fully employed and transferred to the low-level features, it will help to produce satisfactory fusion results. On the one hand, the high-level features of both encoders provide rich semantic structure information that can be used to highlight the targets. On the other hand, the low-level features of both encoders represent rich texture details that help to improve the aesthetics of the fused images. Better fused images can be obtained by combining global semantic and detail information. However, if they are combined directly without considering their differences and global contextual information, the two features will lack interaction, making it difficult to restore the desired fusion results in the decoder. To this end, the decoder directly deconvolve the previously extracted high-level feature maps which combine different modal features with the output of the previous decoded block via cross-modality shortcuts with contextual attention (CMSCA) until the input resolution is restored. Fig. 4 (a) shows the structure of the decoding block, which enables feature integration across different scales and levels to reconstruct the fusion results. One can see that features from the two encoding blocks at the corresponding scale are added after enhancing by contextual attention (CoA) and then cascaded with the output of the previous decoding blocks to obtain the aggregated features. Subsequently, deconvolution operation is used to upsample the aggregated features. After four times of upsampling, the fused images are reconstructed through the restore block as shown in Fig. 4 (b).

(3) cross-modality shortcuts with contextual attention (CMSCA)

During feature extraction, details are inevitably lost after multiple downsampling operations of the encoders. Additionally, deconvolution can only recover the structural details of the source images from the encoded features. In other words, the outputs of the

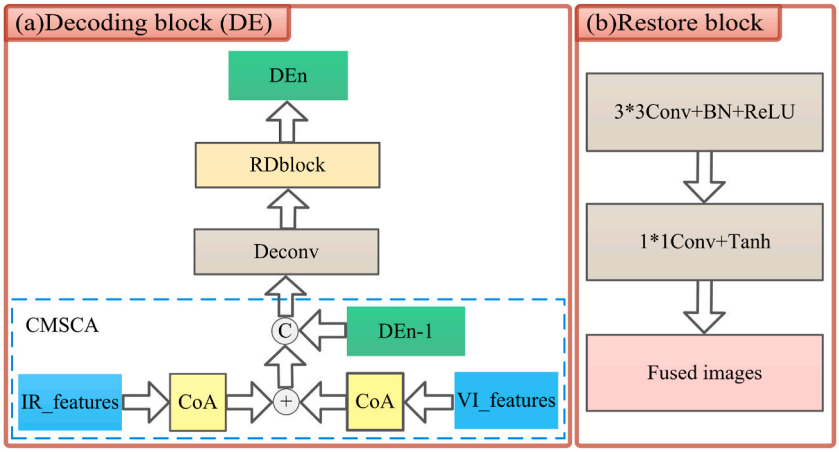


**Fig. 4.** Illustration of the decoding path in Fig. 2 a) the structure of decoding block; b) the structure of restore block.

decoder are the combined results of the source images, which leads to degraded fusion performance. More importantly, if the extracted multi-scale features are exploited in an incomplete or redundant way, the reconstruction of the information-rich fused image will be interfered. Therefore, we capitalize on skip connections to aggregate the same-scale features of both encoders into their corresponding upsampling blocks in an attentionally enhanced manner, thus reusing the earlier encoded feature maps and improving the reconstruction ability of the decoder. Reference [40], we employ contextual attention (CoA) that unifies both contextual information mining among features and self-attention learning over feature maps in a single architecture with favorable parameter budget to enhance the representative capacity of the discriminative features extracted from the both encoders. The structure of the CoA is shown in Fig. 5. CoA first captures the static contextual information among features via a $3 \times 3$ convolution operation. Then, two consecutive $1\times 1$ convolution operations are applied to conduct self-attention learning based on the input and contextualized features, yielding dynamic contextual information. On the other branch, a $1 \times 1$ convolution is used to capture global information of the input features, which is multiplied with the dynamic contextual information to obtain the global dynamic contextual information. Finally, the static and global dynamic contextual information are fused as the outputs.

CMSCA reinforces the feature interactions between the up- and down-sampled blocks in the corresponding phase by performing pixel/region adaptive selection and learning based on feature-level attention. Compared to manual rules, learnable feature aggregation strategies not only enable complex multi-resolution feature extraction networks have human-like attention perception, but also avoid the information loss caused by successive convolutions and sampling, so that the fused results show competitive brightness and contrast.

Summarily, G is constructed based on Y-Net that combines the superiorities of attention mechanism and residual dense network. Extensive ablation studies demonstrate that our designed fusion image generation network achieves a good balance in terms of computational load, training speed, and feature extraction.

#### 3.2.2. Discriminator architecture

Considering the types of dominant and secondary information contained in the source images, we construct two discriminators (GL-Dvi and GL-Dir) based on VGG16 to make the generated images more realistic with the gambling of the G and the GL-Dvi/GL-Dir. It is worth emphasizing that the newly designed single discriminator is able to mine both the holistic details and local radiative information within the source images and classify features by integrating global GAN and PatchGAN into a unified architecture. The branch of the global discriminator (i.e., global GAN) forces the fused image to learn the holistic distribution and feature of the source images, while the branch of the local discriminator (i.e., PatchGAN) focuses more on the degree of local information preservation. GL-Dvi and GL-Dir have the identical network structure but do not share training parameters. Fig. 6 shows the network structure of the designed discriminator, and the architecture of the global-local discriminator is shown in Table 2. The whole network consists of five convolution blocks, and each block is composed of two convolution operations, two batch normalization (BN) operations, two Leak ReLU functions, and one pooling operation. The latter part of the discriminator utilizes a split path with two separate confronted games to capture both holistic and local features in inputs. Concretely, the global path ends up with a fully connected layer to distinguish the whole fused image from the reference images, similar to global GAN. The local path, composed of a $1 \times 1$ convolution operation and a Tanh activation function, maps the input image to a matrix representing the probability of each true patch, similar to the PatchGAN. The final output of the local path is obtained by averaging over all probability values.

### 3.3. Loss functions

#### 3.3.1. Discriminator' loss functions

GL-Dvi and GL-Dir are trained to distinguish fused image from the reference images, and their loss functions are formulated based on least squares GAN [21] as follows:

$$L_{GP-Dvi}=\frac{1}{N}\sum_{i=1}^{N}\left(GP_{D_{vi}(VI_{local})}-a\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{vi}\left(VI_{global}\right)-a\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{vi}\left(F_{local}\right)-b\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{vi}\left(F_{global}\right)-b\right)^2 \quad (1)$$

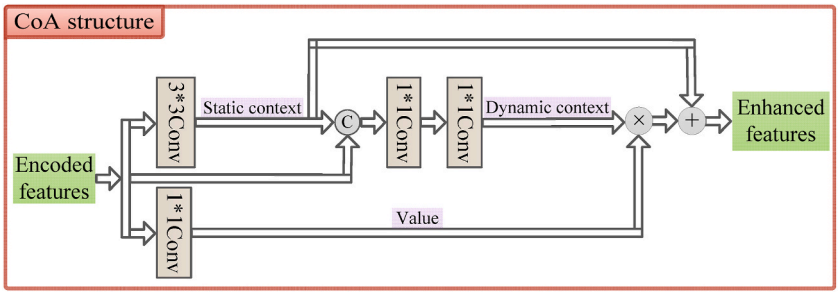


**Fig. 5.** Illustration of the CoA in Fig. 4.

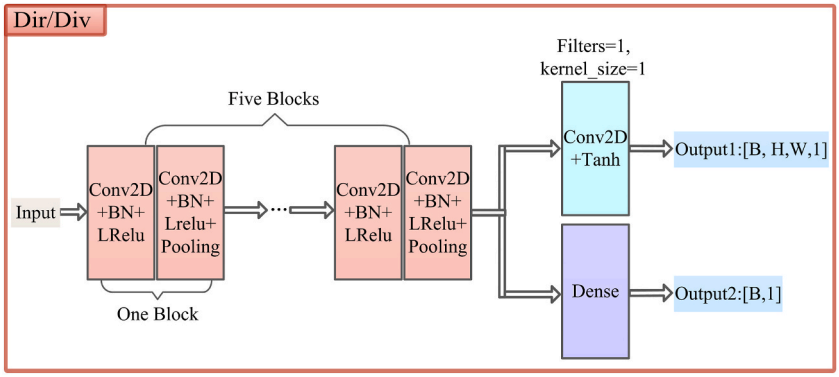


**Fig. 6.** The architecture of the global-local discriminator. Discriminator includes five convolution blocks, and each convolution block contains 2 convolution layers, 2 batch normalization layers, 2 leaky ReLU activation layers, and 1 pooling layer. At the end of the discriminator, a branch containing 1 convolution layer and Tanh activation layer outputs the local discriminant results. And the other branch outputs the global discriminant results via fully connected layer.

**Table 2**
The parameter settings for all layers of the global-local discriminator. Conv denotes the convolution block (convolution layer + batch normalization + activation function).

| Layer | Input Channel | Output Channel | Filter Size | Stride | Padding | Activation Function |
|---|---|---|---|---|---|---|
| Conv1 | 1 | 16 | 3 | 1 | SAME | LeakyReLU |
| Conv2 | 16 | 16 | 3 | 2 | SAME | LeakyReLU |
| Conv3 | 16 | 32 | 3 | 1 | SAME | LeakyReLU |
| Conv4 | 32 | 32 | 3 | 2 | SAME | LeakyReLU |
| Conv5 | 32 | 64 | 3 | 1 | SAME | LeakyReLU |
| Conv6 | 64 | 64 | 3 | 2 | SAME | LeakyReLU |
| Conv7 | 64 | 128 | 3 | 1 | SAME | LeakyReLU |
| Conv8 | 128 | 128 | 3 | 2 | SAME | LeakyReLU |
| Conv9 | 128 | 256 | 3 | 1 | SAME | LeakyReLU |
| Conv10 | 256 | 256 | 3 | 2 | SAME | LeakyReLU |
| Branch1 | 256 | 1 | 1 | 1 | VALID | Sigmoid |
| Branch2 | 2 × 2 × 256 | 1 | – | – | – | – |

$$L_{GP-Dir}=\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{ir}(IR_{local})-a\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{ir}(IR_{global})-a\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{ir}(F_{local})-b\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{ir}(F_{global})-b\right)^2 \tag{2}$$

where $L_{GP-Dvi}$ stands for the loss function of the visible global-local discriminator, and $L_{GP-Dir}$ stands for the loss function of the visible global-local discriminator. $N$ represents the number of the training data, $GP\_D_{vi}(\bullet)$ and $GP\_D_{ir}(\bullet)$ denote the discriminant results, $VI_{local}$ and $IR_{local}$ indicate the patches of VI and IR images, $VI_{global}$ and $IR_{global}$ indicate the whole VI and IR images, $F_{local}$ indicates the patches of generated image, and $F_{global}$ indicates the whole generated image. The symbols a and b are soft labels.

### 3.3.2. *Generator' loss functions*

Previous fusion methods focus too much on the richness of the information, resulting in fused images that look like neutral results of IR and VI images, or distorted VI images, or sharpened IR images. In view of that, we design a hybrid loss function based on feature and pixel domains, which aims to improve both the information content and the perceptual quality of the fused images.

In order to keep more image content, the optimization of G considers both the guidance of deep feature and original image domains [18]. So, we define the similarities of the gradients and intensities at the image level as follows:

$$L_{content}^{pixel}=\frac{1}{N}\sum_{i=1}^{N}\left(Int(F)-Int(IR)\right)^2+\frac{1}{N}\sum_{i=1}^{N}\left(Grad(F)-Grad(VI)\right)^2 \tag{3}$$

where $Int(\bullet)$ represents the intensity operation, which is calculated using the mean filter. $Grad(\bullet)$ represents the gradient operation performed using the Sobel operator. F, IR, and VI indicate the generated, IR, and VI images, respectively. $L_{content}^{pixel}$ helps fused image preserves the intrinsic properties of images with different modalities.

Additionally, in the deep feature domain, the similarity constraint between the generated image and source images is formulated as follows:

$$L_{content}^{feature} = \lambda \bullet \frac{1}{N}\sum_{i=1}^{N}(\varphi(F) - \varphi(IR))^2 + \frac{1}{N}\sum_{i=1}^{N}(Grad(\varphi(F)) - Grad(\varphi(VI)))^2 \tag{4}$$

where $\varphi(\bullet)$ stands for the feature maps extracted from the 9th, 21st and 27th convolution layers of the GL-Dvi and GL-Dir. $L_{content}^{feature}$ can represent the intrinsic information by combining shallow and deep features. $\lambda$ is used to balance between the two terms and is fixed to 20 according to the experimental effects.

Image histogram [12] is also a common image comparison tool that reflects the statistical characteristics of image pixel values. In general, two images can be considered to be somehow identical if their histograms are extremely similar. The histogram similarity constraint for two images is expressed as follows:

$$L_{hist} = \frac{1}{255}\left(\|hist(F) - hist(I_{vi})\|_2^2 + \|hist(F) - hist(I_{ir})\|_2^2\right) \tag{5}$$

where $hist(\bullet)$ represents the histogram of the input images.

We also expect the generated image to share more structural similarities with the two source images. From this, the structural similarity constraint [29] is introduced and formulated as follows:

$$L_{SSIM} = 1 - \frac{SSIM(F, VI) + SSIM(F, IR)}{2} \tag{6}$$

where $SSIM(\bullet)$ stands for the structural similarity measure between the fused image and the two source images.

During adversarial process, G expects the GL-Dvi and GL-Dir to judge the generated image as the true data [21]. Hence, the adversarial loss can provide additional information complement and is defined as follows:

$$L_{adv} = \frac{1}{N}\sum_{i=1}^{N}\left(GP_{D_{ir}(F_{local})} - a\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(GP_{D_{ir}(F_{global})} - a\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{vi}(F_{local}) - a\right)^2 + \frac{1}{N}\sum_{i=1}^{N}\left(GP\_D_{vi}(F_{global}) - a\right)^2 \tag{7}$$

Summarily, the total loss of G is expressed as follows:

$$L_G = \alpha \bullet L_{content}^{pixel} + \beta \bullet L_{content}^{feature} + L_{SSIM} + \gamma \bullet L_{adv} + L_{hist} \tag{8}$$

where $\alpha$, $\beta$ and $\gamma$ are all hyperparameters, which are experimentally fixed as 18, 19, and 10, respectively.

## 4. Experimental verification

### 4.1. Experimental details

(1) Datasets

We selected 55 pairs of IR and VI images from the TNO dataset as the samples to train our model. However, this a little bit of data is insufficient to train a deep network model well. We adopt the measures of the non-overlapping cropping to cut each original pair of images into $88 \times 88$ patches with step size of 14 to extend the training dataset.

During testing, the TNO and RoadScene datasets are used to validate the fusion effect of the proposed method. It is worth to note that all image pairs used in this paper are aligned and grayed with high quality.

(2) Training details

During training stage, the batch size is 30, the training epoch is 20, and the learning rate is fixed at 1e-4. We initially train the two discriminators using Adam three times before alternating the training of G and the dual discriminator once per batch. The detailed training procedure is shown in Algorithm 1 below. At the test stage, only the generator remains. The complete pairs of test images are fed sequentially into the well-trained generator to produce satisfactory fusion results. All experiments, including ours, are programmed on TensorFlow and implemented on a computer configured with GPU NVIDIA GeForce RTX 3070 for fair comparison.

**Algorithm 1**: our model's training details

**Inputs**: infrared image and visible image
**Output**: fused image
**for** i in range maximum epoch **do**
  **for** t times **do**
    Select m visible and infrared image patches from training dataset;
    Select m fused image patches from generated set;
    Update global-local discriminator1 using the Adam according to Eq. (1);
    Update global-local discriminator2 using the Adam according to Eq. (2);

(*continued on next page*)

(*continued*)

| | |
|---|---|
| **7** | **End** |
| **8** | Select m visible and infrared from training dataset; |
| **9** | Update generator using the according to Eq. (8); |
| **10 end** | |

(3) Comparison methods

The proposed method adopts the idea of multi-scale representation in GAN architecture to fully extract the key information contained in the source images and selectively preserve it in the fusion images. Herein, nine representative fusion methods based on decomposition strategy are chosen as the baseline methods for comparison with our algorithm, including LP [41], RP [42], CVT [43], MSVD [44], DCHWT [45], MDLatLRR [46], Dualbranch [12], CUFD [14], GANFM [47]. U2Fusion [48] inspires us to assess the similarities between the resultant image and the two source images in deep feature domain. CSF [49] preserves valuable features by assessing the importance of features in a deep learning way, while our approach directly leverages relatively simple contextual attention to selectively aggregate important features for fused image reconstruction. It is therefore worth comparing the fusion effects of the two approaches. PMGI [50] is a unified image fusion method, which has been referred to by many works. FusionGAN [21] and GANMcC [24] are the typical GAN-based comparison methods.

(4) Objective assessment metrics

As we all know, it is a reasonable option to utilize a multi-metric evaluation system to comprehensively assess the fusion results. Given that the original intention of our method tends to improve both the information richness and visual perceptual of the resultant images, the following five metrics are selected to assess the fusion results. (1) Mutual information (MI) [51]: MI is used to assess the amount of information transferred from the two source images to the resultant image. The larger the MI, the more information the resultant image contains about the source images. (2) Visual information fidelity (VIF) [52]: VIF conforms to the human visual system and is used to evaluate the fidelity of information of the fused images. The higher the VIF is, the better performance the fusion method has. (3) Standard deviation (SD) [53]: SD reflects the contrast and distribution characteristics of the fused images. The larger the SD, the better the visual quality of the fused images. (4) Petrovic metric parameter (Nabf) [54]: Nabf represents the ratio of noise to

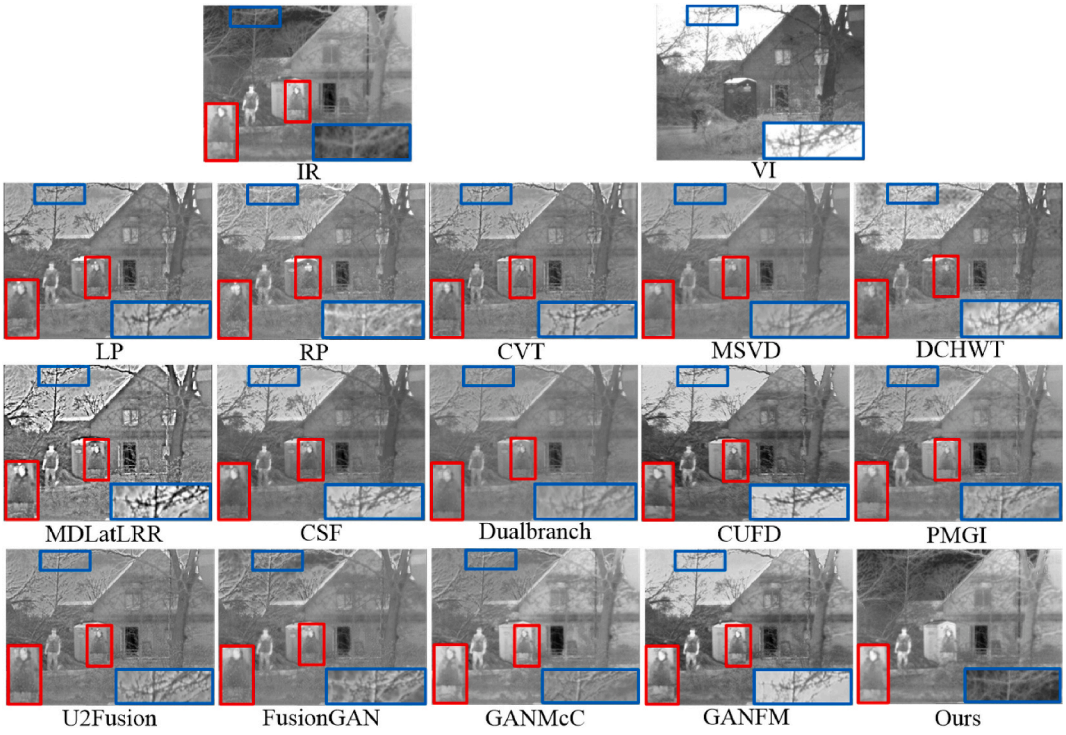


**Fig. 7.** Visual examples of different methods on image pairs in the TNO dataset. From top to bottom: infrared image, visible image, results of LP, RP, CVT, MSVD, DCHWT, MDLatLRR, CSF, Dualbranch, CUFD, PMGI, U2Fusion, FusionGAN, GANMcC, GANFM and our method. For clear comparison, we select two small regions (i.e., the red and blue boxes) in each image, and then zoom in it and put it in the bottom corner.

artifacts in the resultant results. The smaller the Nabf, the less noise and artifacts occur in the resultant images. (5) FMI_dct [55]: FMI_dct computes mutual information for the discrete cosine features. The larger the FMI_dct is, the better the fusion method performance has.

### 4.2. Comparison experiments

(1) Comparison results on the TNO dataset

Fig. 7 shows the fused results of *men_in_front_of_house* image in the TNO dataset. IR image provides thermal targets (such as two pedestrians) and clear structural features (such as protuberant branches and wall in the background), while texture details and better visual effect are reflected by VI image. Therefore, the combined results of the two source images will be biased towards the IR image and well fit the visual perception simultaneously. In each image, the thermal target is highlighted by a red rectangular region, while a blue box is used to frame out the background area. They were subsequently scaled for easier observation. Obviously, LP, RP, CVT, MSVD, DCHWT, CSF, and Dualbranch produce visually unfriendly blur effects and background noise/artifacts due to severe information loss during fusion. PMGI and U2Fusion methods, which incline to VI images, have lower brightness due to detail information loss and noise interference. MDLatLRR suffers from distortions and noise when restoring fused image, resulting in unnatural visual effects. FusionGAN overly inclined to IR image generate fused images with blurry visual perception because of the presence of noise/artifacts. GANMcC has higher contrast than other approaches, including ours, but the thermal targets are still somewhat blurred due to information loss. Although CUFD and GANFM perform best in maintaining the image contrast and thermal target saliency, rich levels of structural details are diluted by the VI images. Our method maintains both the saliency of the hot targets and the naturalness of the fusion image without losing the hierarchy of the background trees and introducing any unnecessary artifacts, while other comparison algorithms restore blurred targets, or introduce coarse edges along the wall, or make the branches present as though they are on a flat surface. Overall, our method not only maintains high brightness for objects that are skewed toward the IR images, but also conforms to human perceptual with less distortions. In other words, the proposed approach performs best than other competitors in preserving the targets saliency and realistic texture details.

The above intuitive analysis from the perspective of information richness of the fused images makes it difficult to say the best or worst fusion performance in a direct manner. Therefore, it is necessary to adopt some objective image evaluation metrics for further quantitative analysis. Table 3 shows the objective evaluation results of the above methods over the five metrics. Obviously, our metric ranks first in terms of MI, VIF, Nabf, and FMI_dct. The presence of multi-scale representation of source images allows the MI and FMI_dct metrics of the proposed method to surpass other competitors. In other words, our method extracts the maximum useful information (i.e., intensity feature of infrared image and gradient feature of visible image) from the different source images and transfers it to the resultant images. The largest VIF metric shows that our results suffer from less distortions due to the cross-scale feature aggregation submodule with contextual attention. This also means that the proposed method can enhance the visual perception features of visible image. The smallest Nabf indicates that our fused images suffer from less noise and artifacts. Our method assumes that intensity information exists in IR images while VI images convey gradient information. Therefore, a relatively simple content constraint is adopted to preserve the crucial information in the two source images and improve the image aesthetics rather than the maximum or complementary information constraints, which leads to a reduction in the contrast features of the resultant images. Thus, it is forgivable that the SD index is in the middle. In conclusion, our method can fully excavate meaningful information in source images and integrate it into the fusion images with the help of Y-Net-based generator and the image-feature domain-based loss function for cooperative guidance. As a result, the fusion results generated by our method improve the information content of the fused image.

(2) Comparison results on the RoadScene dataset

**Table 3**
The average of the five metrics among all algorithms on the TNO dataset (Bold: optimal).

| Method | MI | VIF | SD | Nabf | FMI_dct |
|---|---|---|---|---|---|
| LP | 1.4820 | 0.6871 | 8.3885 | 0.1165 | 0.2846 |
| RP | 1.3782 | 0.6181 | 8.3375 | 0.1812 | 0.2419 |
| CVT | 1.3905 | 0.5220 | 8.2301 | 0.1446 | 0.3755 |
| MSVD | 1.5213 | 0.5389 | 7.9573 | 0.1121 | 0.2311 |
| DCHWT | 1.4478 | 0.5111 | 7.9025 | 0.0619 | 0.3461 |
| MDLatLRR | 1.3199 | 0.5929 | 8.8746 | 0.5069 | 0.3553 |
| CSF | 1.6309 | 0.5952 | 8.3559 | 0.0729 | 0.2491 |
| Dualbranch | 1.6374 | 0.5490 | 7.9859 | 0.1921 | 0.2961 |
| CUFD | 3.0039 | 0.8156 | 9.3869 | 0.1650 | 0.1822 |
| PMGI | 2.0623 | 0.6769 | 9.3248 | 0.1021 | 0.3672 |
| U2Fusion | 1.5206 | 0.5945 | 8.8094 | 0.2700 | 0.3276 |
| FusionGAN | 1.7728 | 0.5643 | 9.0638 | 0.0675 | 0.3815 |
| GANMcC | 1.9488 | 0.6247 | **9.5547** | 0.0773 | 0.3363 |
| GANFM | 2.5939 | 0.8335 | 9.5035 | 0.2294 | 0.3275 |
| Ours | **3.4250** | **1.0317** | 8.6964 | **0.0576** | **0.3834** |

Fig. 8 shows the visualization results of the proposed approach compared to other alternatives on the RoadScene benchmark dataset. Pedestrians are bright thermal targets in the IR image, and signal poles, houses, trees are the background details in the VI image. Some distinct regions (pedestrians and signal poles) in the image are labeled with a red box and enlarged to clearly observe. On can see that only MDLatLRR and our method highlight the pedestrian targets, while the others reduce the brightness. In terms of texture detail recovery, almost all the fusion images produced by the comparison algorithms produce an unnatural visual perception due to the introduction of noise/artifacts (LP, RP, MDLatLRR, U2Fusion), or display similarly distorted VI images (CVT, MSVD, DCHWT, Dualbranch, FusionGAN), or look like neither VI nor IR images because of the pixel intensity changes so much (CSF, CUFD, PMGI, GANMcC, GANFM). On the contrary, our fusion image not only maintains the brightness of the thermal objects, but also restores most of the information with less distortions.

The calculation results of the five metrics of fusion results with different methods are shown in Table 4. One can notice that our method ranks first in MI, VIFF, and FMI_dct metrics and third in SD metric. The largest MI and FMI_dct demonstrate that the proposed method transfers the most amount of information from the two source images (i.e., intensity feature of infrared image and gradient feature of visible image) to the fusion image. That is, our fusion results contain the richest amount of information. The optimal VIF reflects that our results are more in line with human visual perception compared to other methods. Although the proposed method trails CUFD and GANFM by a narrow margin in the SD measurement, our method does not lose out to them in the overall effect. This is because the proposed method transfers and preserves more effective information from the two source images into the fusion image and the appearance is less distorted than that of other methods. For the Nabf metric, the proposed method ranks in the middle place, which is justifiable. The subtle differences in fusion performance between the TNO and RoadScene datasets can be attributed to their large discrepancy in scenarios and image contents.

### 4.3. Ablation study

In this section, we conduct ablation studies on six closely-related modules for image generation tasks to validate the superiority of the well-designed fusion model.

(1) Study on the number of DenseNet layers

To balance the computational burden and fusion performance, we first study the number of DenseNet layers in RDblock. The optimal DenseNet layers can be confirmed by observing the fusion effect against the layer number both in qualitative (the second row of Fig. 9) and quantitative results (Table 5). Obviously, the dense net with three convolution layers can preserves finer structural details of the bench in the scene.

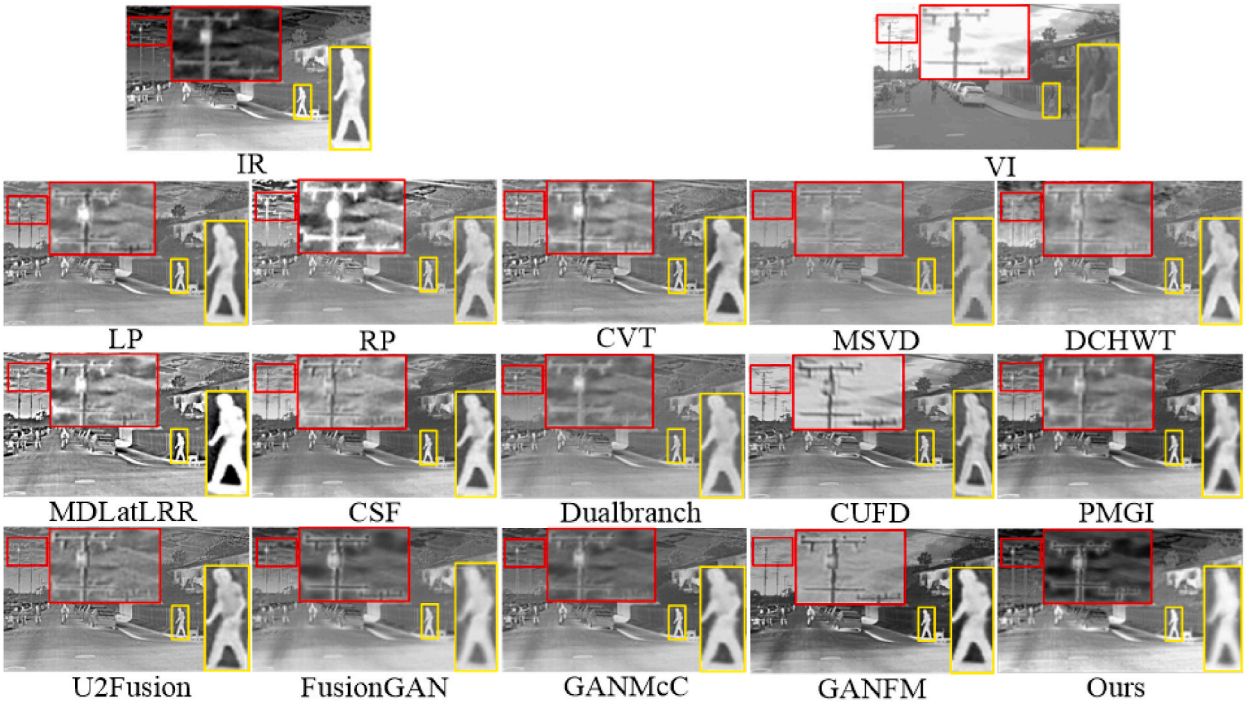


**Fig. 8.** Visual examples of different methods on image pairs in the RoadScene dataset. From top to bottom: infrared image, visible image, results of LP, RP, CVT, MSVD, DCHWT, MDLatLRR, CSF, Dualbranch, CUFD, PMGI, U2Fusion, FusionGAN, GANMcC, GANFM and our method. For clear comparison, we select two small regions (i.e., the red and yellow boxes) in each image, and then zoom in it.

**Table 4**
The average of the five metrics among all algorithms on the RoadScene dataset (Bold: optimal).

| Method | MI | VIF | SD | Nabf | FMI_dct |
|---|---|---|---|---|---|
| LP | 2.4781 | 0.8117 | 9.7681 | 0.1481 | 0.3203 |
| RP | 2.3815 | 0.7232 | 9.9558 | 0.2107 | 0.2671 |
| CVT | 2.2378 | 0.6104 | 9.8411 | 0.2013 | 0.3592 |
| MSVD | 2.6566 | 0.6604 | 9.8026 | **0.0289** | 0.2318 |
| DCHWT | 2.7426 | 0.6392 | 10.0968 | 0.1259 | 0.3366 |
| MDLatLRR | 1.9127 | 0.7125 | 10.0564 | 0.5539 | 0.3265 |
| CSF | 2.8683 | 0.7560 | 10.2582 | 0.1054 | 0.2551 |
| Dualbranch | 3.0517 | 0.6828 | 9.8994 | 0.0448 | 0.2577 |
| CUFD | 3.7967 | 0.8404 | 10.3422 | 0.1927 | 0.1995 |
| PMGI | 3.3732 | 0.8078 | 10.1556 | 0.0860 | 0.3573 |
| U2Fusion | 2.7289 | 0.7216 | 10.2790 | 0.2329 | 0.3177 |
| FusionGAN | 2.9472 | 0.6686 | 10.1548 | 0.0911 | 0.3486 |
| GANMcC | 3.1935 | 0.7241 | 10.2469 | 0.0699 | 0.3560 |
| GANFM | 3.6028 | 0.8803 | **10.6870** | 0.1786 | 0.3258 |
| Ours | **4.5508** | **1.0572** | 10.6142 | 0.1047 | **0.3669** |

(2) Study on the number of RDblocks

In general, the deeper the network, the easier it is to extract features that are close to the essence of the source images. Meanwhile, the training will be longer. Therefore, we also investigate the impact of the number of RDblocks on the fusion performance to determine the optimal number of RDblocks, so as to equilibrate the training efficiency and fusion performance. The third row of Fig. 9 shows visual examples against the RDblock number. Clearly, the use of two RDblocks for each scale feature extraction module leads to blurred fused image compared to our method that employs one RDblock. Worse still, fused image fails to be generated when three RDblocks are employed in each sampling module. The evaluation results in Table 5 further prove the above statement objectively.

(3) Study on the cross-modality shortcuts with contextual attention (CMSCA)

Since the skip connections designed in the previous fusion model integrate the features of different scales without considering their discriminative context information, we elaborate the cross-modality shortcuts with contextual attention (CMSCA) to transfer features from convolutional blocks to their corresponding deconvolution blocks. To demonstrate the effectiveness of CMSCA, we drop contextual attention (i.e., using only the concatenation aggregation rule), but all other settings remain unchanged. According to visual examples shown in the fourth row of Fig. 9, the concatenation operation slightly weakens the brightness of the IR targets and leads to a reduction in perceptual quality because of insufficient feature integration. Objective assessment results in Table 5 give further convincing results. This indicates that it is efficient to aggregate features at different scales via CMSCA.

(4) Study on the number of downsampling operations

It is a well-known phenomenon that downsampling IR images introduces noise, while downsampling VI images loses texture information. To strike the balanced tradeoff between computation/fusion performance, we further study the number of downsampling operations. Concretely, we downsample two source images 2, 3, 4, 5 times, so as to determine the optimal number of downsampling based on the qualitative and quantitative evaluation results of the fusion performance. Clearly, the best fusion performance can be achieved by downsampling two source images by a factor of 4.

(5) Study on the global-local discriminator (GL-Dir and GL-Dvi)

To demonstrate the effectiveness of applying global-local discriminators, we implement five sets of ablation studies, including 1) remove IR global-local discriminator; 2) remove VI global-local discriminator; 3) apply two global discriminators instead of two global-local discriminators; 4) apply two local discriminators instead of two global-local discriminators; 5) apply two global-local discriminators (ours). For the sake of simplicity, let's denote the above ablation experiments as follows: with-GL-Dir, with-GL-Dvi, with-Dgir and Dgvi, with-Dlir and Dlvi, and with-GL-Dir and GL-Dvi (ours). Evaluation of fusion performance from both subjective and objective aspects, there are two distinct phenomena: 1) The single adversarial game generates the fused image that overly inclines to IR or VI images. 2) The two-adversarial model established between two global/local discriminators and generator have a similar fusion effect in maintaining the brightness of IR targets and preserving the details of the two source images, which are slightly blurrier than our complete model. In contrast, the dual discriminator that combines global GNA and PatchGAN as a unified architecture can fully capture the local radiative information and global texture details in the source images.

(6) Study on the perceptual loss

In generative networks, regularization is often used at the pixel level to make the fused images have the same characteristics as the

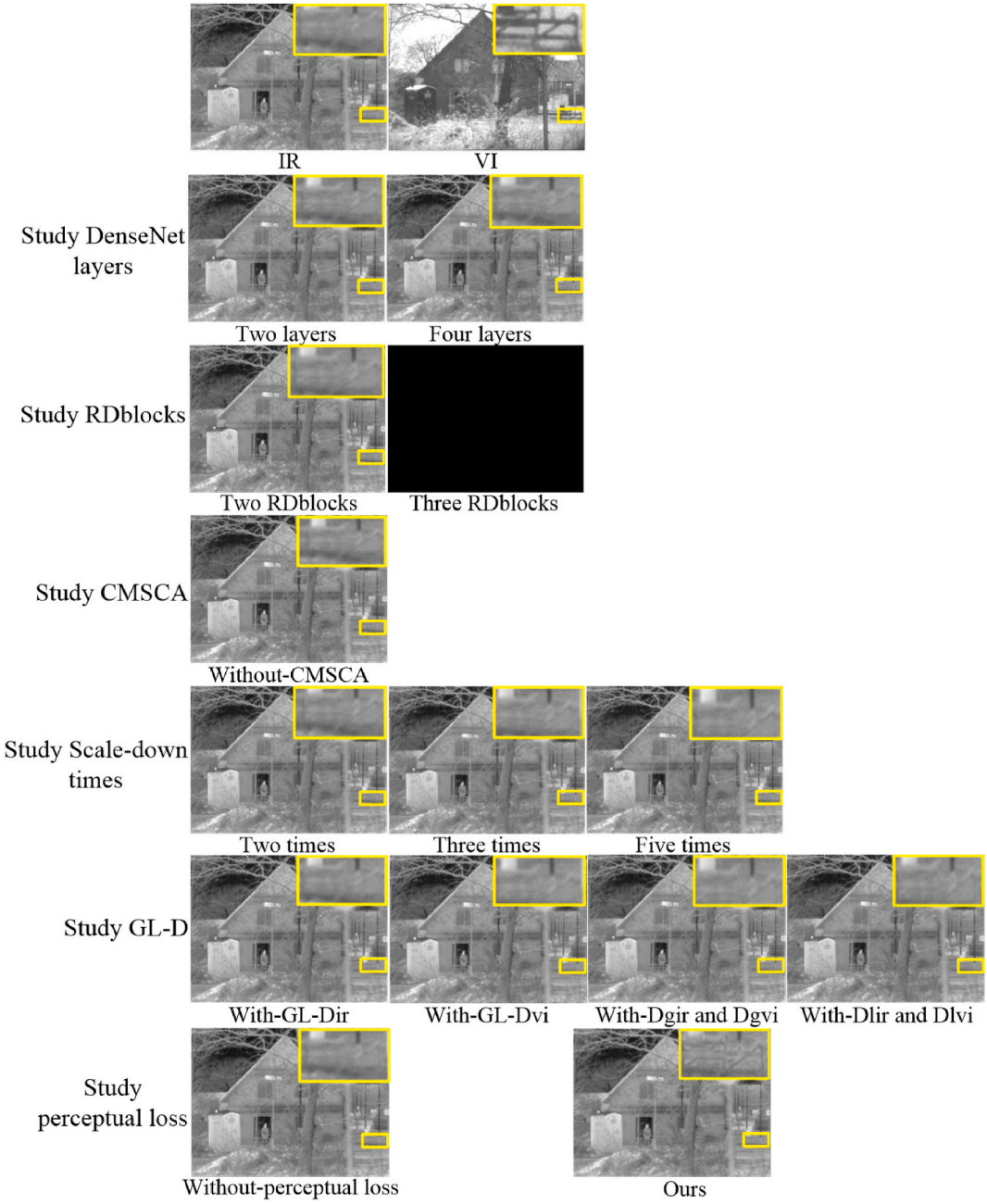


**Fig. 9.** Ablation examples on the TNO dataset. From top to bottom: DenseNet layers ablation study, RDblocks ablation study, CMSCA ablation study, downsampling times ablation study, global-local discriminator ablation study and perceptual loss ablation study.

source images. However, there are many limitations to the pixel-wise constraint. For example, given two identical images, if the pixels of one image are slightly shifted, the difference between the pixels of the two images becomes large, but the actual content remains the same. Therefore, perceptual loss is adopted to constrain the original images and the generated image in the deep feature space, so that the generated images can retain the high-level semantic information in the source images and improve the visual quality of the generated images. To verify the enhancement effect of the perceptual loss on the fusion performance, we remove it from the generator's loss for ablation studying and leave the others unchanged. In terms of subjective perception, fusion images degrade the visual effects regardless of the feature domain constraints. So, the ablation experiments demonstrate that the designed perceptual loss is effective in constraining the intensity and gradient between the fused image and source images in the feature domain through subjective and objective evaluations.

**Table 5**
Objective ablation results on the TNO dataset (Bold: Our results).

| Ablation modules | Method | MI | VIF | SD | Nabf | FMI_dct |
|---|---|---|---|---|---|---|
| Study DenseNet layers | Two layers | 3.2506 | 1.0183 | 8.6590 | 0.0741 | 0.3735 |
| | Three layers (Ours) | – | – | – | – | – |
| | Four layers | 3.2863 | 1.0140 | 8.6873 | 0.0682 | 0.3777 |
| Study RDblocks | One RDblock (Ours) | – | – | – | – | – |
| | Two RDblocks | 3.2962 | 1.0230 | 8.6346 | 0.0709 | 0.3812 |
| | three RDblocks | – | – | – | – | – |
| Study CMSCA | Without-CMSCA | 3.2957 | 1.0136 | 8.6106 | 0.0679 | 0.3790 |
| | With-CMSCA(Ours) | – | – | – | – | – |
| Study Scale-down times | Two times | 3.2844 | 1.0231 | 8.6553 | 0.0723 | 0.3817 |
| | Three times | 3.2891 | 1.0220 | 8.6507 | 0.0724 | 0.3823 |
| | Four times (Ours) | – | – | – | – | – |
| | Five times | 3.2549 | 1.0176 | 8.6331 | 0.0746 | 0.3746 |
| Study GL-D | With-GL-Dir | 4.0616 | 1.0303 | 8.6352 | 0.0584 | 0.3635 |
| | With-GL-Dvi | 2.7684 | 0.9138 | 8.6957 | 0.1070 | 0.4104 |
| | With-Dgir and Dgvi | 3.2947 | 1.0136 | 8.6634 | 0.0685 | 0.3822 |
| | With-Dlir and Dlvi | 3.2874 | 1.0187 | 8.6777 | 0.0706 | 0.3832 |
| | With-GL-Dir and GL-Dvi (Ours) | – | – | – | – | – |
| Study perceptual loss | Without-perceptual loss | 3.2475 | 1.0168 | 8.6003 | 0.0755 | 0.3605 |
| | With-perceptual loss (Ours) | – | – | – | – | – |
| – | Ours | **3.4250** | **1.0317** | **8.6964** | **0.0576** | **0.3834** |

### *4.4. Comparison of running times and parameters*

Since traditional algorithms run on CPU, we only compare the complexity among various learning-based methods in Table 6. On the one hand, we evaluate the time complexity by computing the mean and standard deviation of the running times of different algorithms on the TNO dataset. On the other hand, we count the parameters of the different deep learning models to reveal the spatial complexity. It is worth mentioning that we only count the number of parameters in the generator network in Table 6, since only the generator is at work during image generation in the GAN-based methods. One can notice that PMGI realizes the minimum running time, while Dualbranch contains the smallest number of parameters. This is because PMGI and Dualbranch construct the simplest structures in the testing phase. Due to the introduction of RDblocks and contextual attention in our model, the proposed method is relatively time-consuming and has a large number of parameters.

## 5. Conclusion and discussion

In this work, we report a novel GAN-based solution for IR and VI image fusion that fully considers the multi-scale extraction and transfer of source image information and achieves realistic fusion results with less distortions. Specifically, the generator is designed based on Y-Net and RDblock is introduced to enhance the learning ability of discriminative multi-scale features of source images. Moreover, CMSCAs are employed to selectively aggregate features at different scales and different levels, which enhance the reconstruction capability of the generator. We design two discriminators that combine the structural advantages of global GAN and PatchGAN to study the distributions of the two source images from the global and local ranges, thus forcing the generator to produce fused images with richer information and less distortions. A hybrid loss function based on the intensity and gradient constraints in both feature and image domains is designed to guide the proposed model optimization. Upon lots of comparisons of our method with 14 other mainstream algorithms, our method far outperforms them in terms of source information extraction and transfer and the perceptual quality of the fusion images.

However, there is still a lot of potential that deserves further study or excavation. On the one hand, the loss function and network structure can be further improved to achieve the efficient fusion performance for IR and VI images in extreme environments such as illumination variations. The above test datasets usually contain image pairs that are captured under normal exposure settings. To validate the robustness of the proposed method to artifacts and illumination variations, 20 nighttime image pairs from the MSRS dataset are selected to conduct testing experiments. Fig. 10 provides a failure case to visually show the limitations of the proposed method. Obviously, all fusion methods except GANFM fail to eliminate the illumination degradation in nighttime images, but our method exhibits a more natural overall visual perception. Table 7 also gives credible evidence. As we can see, the five metrics of the proposed method is relatively low, compared with that of the TNO and RoadScene datasets. On the other hand, the improved solution can also be extended to other applications, such as medical image fusion and multiple exposure image fusion. Last but not least, the proposed method achieves numerous progresses in pre-registered multi-modality data. Recently, misaligned image fusion also a hot spot. In the future, we will continue to improve and optimize the proposed model to solve the misalignments and calibration discrepancies between the infrared and visible modalities in multi-modality image fusion task.

## Data availability statement

Since the data used in this work relate to our future work, the data associated with my study has not been deposited into a publicly

**Table 6**
Time and space complexity of different learning-based methods on the TNO dataset.

| Method | Running time/s | parameters/K |
|---|---|---|
| CSF | $5.06 \pm 2.19$ | 185.4 |
| Dualbranch | $1.06 \pm 0.09$ | 89.5 |
| CUFD | $21.42 \pm 1.78$ | 955 |
| PMGI | $0.093 \pm 0.340$ | 161 |
| U2Fusion | $0.469 \pm 0.749$ | 659 |
| FusionGAN | $0.14 \pm 0.62$ | 925.6 |
| GANMcC | $0.30 \pm 0.76$ | 1867 |
| GANFM | $0.96 \pm 1.67$ | 10210 |
| Ours | $0.35 \pm 1.12$ | 15214 |

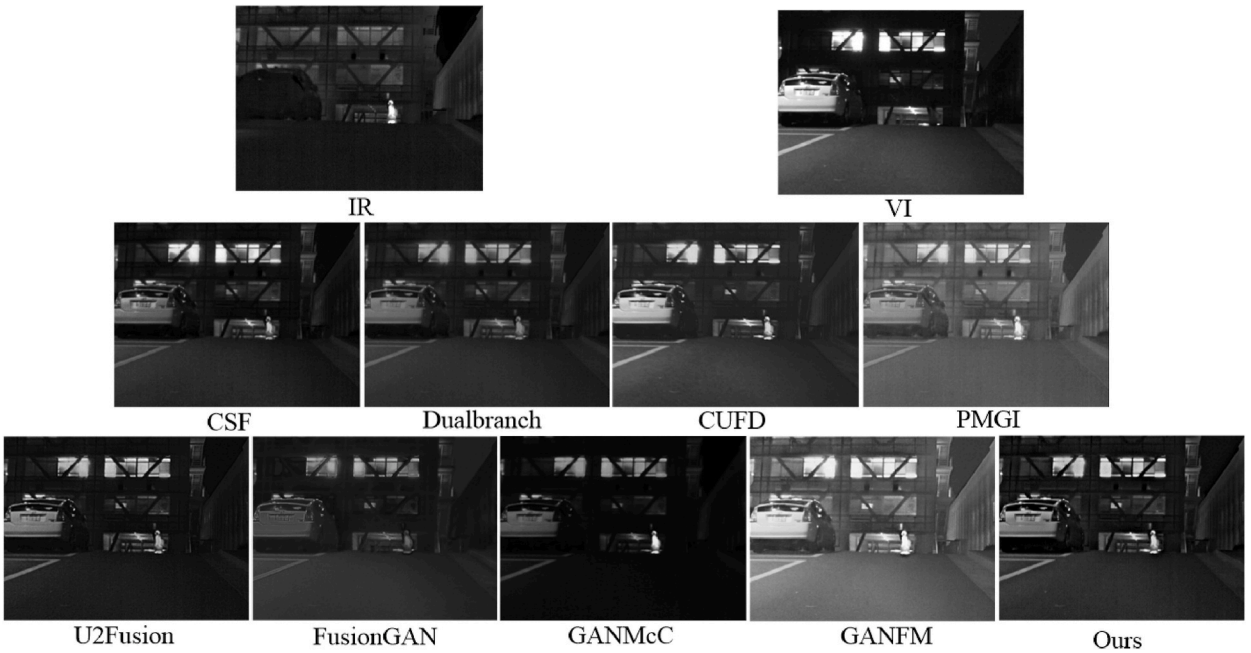


**Fig. 10.** A failure case. Fusion results of some methods on pairs of nighttime images. From top to bottom: infrared image, visible image, results of CSF, Dualbranch, CUFD, PMGI, U2Fusion, FusionGAN, GANMcC, GANFM and our method.

**Table 7**
The average of the five metrics among some algorithms on pairs of nighttime images (Bold: optimal).

| Method | MI | VIF | SD | Nabf | FMI_dct |
|---|---|---|---|---|---|
| CSF | 2.4110 | 0.7296 | 7.5557 | 0.0389 | 0.2339 |
| Dualbranch | 2.2848 | 0.6950 | 7.1422 | 0.0128 | 0.2545 |
| CUFD | 3.0436 | 0.6998 | 7.8705 | 0.0953 | 0.2062 |
| PMGI | 2.1371 | 0.7438 | 7.9849 | 0.1200 | **0.3368** |
| U2Fusion | 2.0144 | 0.6034 | 6.6207 | 0.0734 | 0.2608 |
| FusionGAN | 2.8405 | 0.4289 | 7.4037 | 0.0364 | 0.3213 |
| GANMcC | 2.1267 | 0.4193 | 5.7556 | **0.0099** | 0.2239 |
| GANFM | **3.2195** | **0.9421** | **8.9923** | 0.2512 | 0.2801 |
| Ours | 2.4523 | 0.7793 | 6.1525 | 0.0626 | 0.3283 |

available repository. However, data will be made available on request.

## CRediT authorship contribution statement

**Danqing Yang:** Writing – original draft. **Naibo Zhu:** Resources. **Xiaorui Wang:** Writing – review & editing. **Shuang Li:** Methodology.

## Declaration of competing interest

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

## References


[1] J. Ma, Y. Ma, C. Li, Infrared and visible image fusion methods and applications: a survey, Inf. Fusion 45 (1) (2019) 153–178.
[2] X. Jin, Q. Jiang, S.W. Yao, A survey of infrared and visual image fusion methods, Infrared Phys. Technol. (2017) 478–501.
[3] H.M. Hu, J.W. Wu, B. Li, An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels, IEEE Trans. Multimed. 19 (12) (2017) 2706–2719.
[4] W.H. Ma, K. Wang, J.W. Li, Infrared and visible image fusion technology and application: a review, Sensors 23 (2) (2023) 599.
[5] Gaurav Choudhary, Dinesh Sethi, From conventional approach to machine learning and deep learning approach: an experimental and comprehensive review of image fusion techniques, Arch. Comput. Methods Eng. 30 (2) (2023) 1267–1304.
[6] D. He, Y. Meng, C. Wang, Contrast pyramid-based image fusion scheme for infrared image and visible image. Proceedings of the IEEE International Symposium on Geoscience and Remote Sensing, 2011, pp. 597–600.
[7] H. Li, Y. Chai, Z. Li, Multi-focus image fusion based on nonsubsampled contourlet transform and focused regions detection, Optik 124 (2013) 40–51.
[8] X. Liu, Y. Zhou, J. Wang, Image fusion based on shearlet transform and regional features, AEU-Int. J. Electron. Commun. 68 (2014) 471–477.
[9] L. Jian, X. Yang, Z. Zhou, Multi-scale image fusion through rolling guidance filter, Future Generat. Comput. Syst. 83 (2018) 310–325.
[10] D.P. Zou, B. Yang, Infrared and low-light visible image fusion based on hybrid multiscale decomposition and adaptive light adjustment, Opt Laser. Eng. 160 (2023) 107268.
[11] C.Y. Cheng, X.J. Wu, T.Y. Xu Unifusion, A lightweight unified image fusion network, IEEE Trans. Instrum. Meas. 70 (2021) 1–14.
[12] Fu Y., Wu X.J., A dual-branch network for infrared and visible image fusion, ICPR (2021),10675-10680.
[13] X. Zheng, Q.Y. Yang, P.B. Si, A multi-stage visible and infrared image fusion network based on attention mechanism, Sensors 22 (0) (2022) 3651.
[14] H. Xu, M.Q. Gong, X. Tian, CUFD: an encoder-decoder network for visible and infrared image fusion based on common and unique feature decomposition, Comput. Vis. Image Underst. 218 (2022) 103407.
[15] K. Ram Prabhakar, V. Sai Srikar, R. Venkatesh Babu, DeepFuse: A Deep Unsupervised Approach for Exposure Fusion with Extreme Exposure Image Pairs, CVPR, 2017, pp. 4724–4732.
[16] J.Y. Ma, L.F. Tang, M.L. Xu StdfusionNet, An infrared and visible image fusion network based on salient target detection, IEEE Trans. Instrum. Meas. 70 (2021) 1–13.
[17] J.W. Liu, J.Y. Liu, S.H. Zhou, Learning a coordinated network for detail-refinement multi-exposure image fusion, IEEE Trans. Circ. Syst. Video Technol. 33 (2) (2023) 713–727.
[18] Y.Z. Long, H.T. Jia, Y.D. Zhong Rxdnfuse, A aggregated residual dense network for infrared and visible image fusion, Inf. Fusion 69 (2021) 128–141.
[19] J.X. Wang, X.L. Xi, D.M. Li, Fusion GRAM: an infrared and visible image fusion framework based on gradient residual and attention mechanism, IEEE Trans. Instrum. Meas. 72 (2023) 1–12.
[20] J.W. Li, J.Y. Liu, S.H. Zhou, Infrared and visible image fusion based on residual dense network and gradient loss, Infrared Phys. Technol. 128 (2023) 104486.
[21] J.Y. Ma, W. Yu, P. Liang FusionGAN, A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (2019) 11–26.
[22] J.Y. Ma, P.W. Lang, W. Yu, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.
[23] Z.L. Le, J. Huang, H. Xu Uifgan, An unsupervised continual-learning generative adversarial network for unified image fusion, Inf. Fusion 88 (C) (2022) 305–318.
[24] J. Ma, H. Zhang, Z. Shao, GANMcC: a generative adversarial network with multiclassification constraints for infrared and visible image fusion, IEEE Trans. Instrum. Meas. 70 (2020) 1–14.
[25] J. Li, H. Huo, C. Li, Multi-grained attention network for infrared and visible image fusion, IEEE Trans. Instrum. Meas. 70 (2020) 1–12.
[26] J. Li, H. Huo, C. Li, Attention FGAN: infrared and visible image fusion using attention-based generative adversarial networks, IEEE Trans. Multimed. 23 (2020) 1383–1396.
[27] Y. Yang, J.X. Liu, S.Y. Huang, TC-GAN: infrared and visible image fusion via texture conditional generative adversarial network, IEEE Trans. Circ. Syst. Video Technol. 31 (12) (2021) 4771–4783.
[28] J.Y. Liu, X. Fan, J. Jiang, Learning a deep multi-scale feature ensemble and an edge-attention guidance for image fusion, IEEE Trans. Circ. Syst. Video Technol. 32 (1) (2022) 105–119.
[29] Y. Fu, X.J. Wu, Tariq Durrani, Image fusion based on generative adversarial network consistent with perception, Inf. Fusion 72 (0) (2021) 110–125.
[30] S. Yi, X. Liu, L. Li Cheng Wang, Infrared and visible image fusion based on blur suppression generative adversarial network, Chin. J. Electron. 32 (1) (2023) 177–188.
[31] Nabil Ibtehaz, M. Sohel Rahman, MultiResUNet: rethinking the U-Net architecture for multimodal biomedical image segmentation, Neural Network. 121 (2020) 74–87.
[32] H. Dong, J.S. Pan, L. Xiang, Multi-Scale Boosted Dehazing Network with Dense Feature Fusion, CVPR, 2020, pp. 2154–2164.
[33] S.P. Zhou, J.J. Wang, J.Y. Zhang, Hierarchical U-shape attention network for salient object detection, IEEE Trans. Image Process. 29 (2020) 8417–8428.
[34] X.H. Wang, J.Q. Gong, M. Hu, LAUN: improved StarGAN for facial emotion recognition, IEEE Access 8 (2020) 161509–161518.
[35] Y.C. Li, X.H. Zeng, Q. Dong, RED-MAM: a residual encoder-decoder network based on multi-attention fusion for ultrasound image denoising, Biomed. Signal Process Control 79 (2023) 104062.
[36] B. Xiao, B.C. Xu, X.L. Bi, Global-feature encoding U-Net (GEU-Net) for multi-focus image fusion, IEEE Trans. Image Process. 30 (0) (2021) 163–175.
[37] L.H. Jian, X.M. Yang, Z. Liu Sedrfuse, A symmetric encoder-decoder with residual block network for infrared and visible image fusion, IEEE Trans. Instrum. Meas. 70 (2021) 1–15.
[38] D. Han, L. Li, X.J. Guo, DPE-MEF: multi-exposure image fusion via deep perceptual enhancement, Inf. Fusion 79 (C) (2022) 248–262.
[39] Jun-Hyung Kim, Youngbae Hwang, Infrared and visible image fusion using a guiding network to leverage perceptual similarity, Comput. Vis. Image Und 227 (C) (2022) 103598.
[40] Y.H. Li, T. Yao, Y.W. Pan, Contextual transformer networks for visual recognition, IEEE Trans. Pattern Anal. Mach. Intell. 45 (2) (2023) 1489–1500.
[41] P. Burt, E. Adelson, The Laplacian pyramid as a compact image code, IEEE Trans. Commun. 31 (4) (1983) 532–540.
[42] A. Toet, Image fusion by a ratio of low-pass pyramid, Pattern Recogn. Lett. 9 (4) (1989) 245–253.
[43] F. Nencini, A. Garzelli, S. Baronti, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2) (2007) 143–156.
[44] V. Naidu, Image fusion technique using multi-resolution singular value decomposition, Defence Sci. J. 61 (5) (2011) 479–484.
[45] B.K. Shreyamsha Kumar, Multifocus and multi-spectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, Signal image video process 7 (6) (2013) 1125–1143.
[46] H. Li, X.J. Wu, J. Kittler, MDLatLRR: A Novel Decomposition Method for Infrared and Visible Image fusion, IEEE Trans. Image Process. 29 (2020) 4733–4746.
[47] H. Zhang, J.T. Yuan, X. Tian, GAN-FM: infrared and visible image fusion using GAN with full-scale skip connection and dual Markovian discriminators, IEEE Trans Comput Imag 21 (7) (2021) 1134–1147.
[48] H. Xu, J.Y. Ma, J. Jiang, U2fusion: a unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2020) 502–518.
[49] H. Xu, H. Zhang, J.Y. Ma, CSF: classification saliency-based rule for visible and infrared image fusion, IEEE Trans. Comput. Imaging 7 (2021) 824–836.

[50] H. Zhang, H. Xu, Y. Xiao, Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity, AAAI 34 (7) (2020) 12797–12804.
[51] G. Qu, D. Zhang, P. Yan. Information measure for performance of image fusion, Electron. Lett.38 (7)(2002)313-315.
[52] H.R. Sheikh, Image information and visual quality, IEEE Trans. Image Process 15 (2006) 430–444.
[53] Y.J. Rao, In-fibre Bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355.
[54] B.S. Kumar, Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, Signal, Image and Video Processing 7 (6) (2013) 1125–1143.
[55] M. Haghighat, M.A. Razian, Fast-FMI: non-reference image fusion metric, IEEE 8th Int. Conf. Appl. Inf. Commun. Technol. (AICT) (2014) 1–3.