Neurocomputing 527 (2023) 71–82

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

THFuse: An infrared and visible image fusion network using transformer and hybrid feature extractor

Jun Chen [a,b,c,*], Jianfeng Ding [a,b,c], Yang Yu [d,e], Wenping Gong [f]

[a] School of Automation, China University of Geosciences, Wuhan 430074, China
[b] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China
[c] Engineering Research Center of Intelligent Technology for Geo-Exploration, Ministry of Education, Wuhan 430074, China
[d] Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai 200083, China
[e] Key Laboratory of Infrared System Detecting and Imaging Technology, Chinese Academy of Sciences, Shanghai 200083, China
[f] Faculty of Engineering, China University of Geosciences, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 12 June 2022
Revised 25 September 2022
Accepted 8 January 2023
Available online 12 January 2023

Keywords:
Image fusion
Vision transformer
Infrared image
Visible image
Deep learning

ABSTRACT

Infrared and visible image fusion aims to integrate complementary information from different types of images into one image. The existing image fusion methods are primarily based on convolutional neural network (CNN), which ignores long-range dependencies of images, resulting in the fusion network unable to generate images with good complementarity. Inspired by the importance of global information, we introduced the transformer technique into the CNN-based fusion network as a way to improve the entire image-level perception in complex fusion scenarios. In this paper, we propose an end-to-end image fusion framework based on transformer and hybrid feature extractor, which enables the network to focus on both global and local information, using the characteristics of transformer to compensate for the shortcomings of CNN itself. In our network, the dual-branch CNN module is used to extract the shallow features of images, and then the vision transformer module is used to obtain the global channel and spatial relationship in the features. Finally, the fusion results are obtained through the image reconstruction module. We calculate the loss in the features of different depths according to the different kinds of original images by using the pre-trained VGG19 network. The experimental results show the effectiveness of adding the vision transformer module. Compared with other traditional and deep learning methods, our method achieves state-of-the-art qualitative and quantitative experiments performance.

1. Introduction

The image fusion technique can integrate multiple source images with complementary information into one fused image to generate a high-quality image to make up for the defects of a single camera sensor [1]. Due to the limitations of photographic hardware devices, the image information obtained from a single type of image sensor or a single shooting setup does not provide a complete description of the imaging scene. First, different types of sensors can acquire image information with different characteristics. For example, visible sensors can produce images with rich texture details, while infrared sensors can produce images with prominent targets. In the fusion of infrared and visible images, the salient information of the infrared images and the detail and texture information of the visible images are extracted, respectively. These two kinds of information are concentrated on one image to expand the image information, which is beneficial to later image processing [2]. For example, the infrared images, visible images, and image fusion results are shown in Fig. 1. Second, the same sensor with different shooting settings can only obtain limited information from the imaging scene. For example, different exposure settings can obtain images with different exposure levels in the same scene, while fixed focal length cameras only capture the objects within the depth-of-field [3]. With the improvement of the performance level of imaging equipment and image processing equipment, the image fusion technique has developed rapidly. It has a wide range of applications in many fields, such as security [4], military [5], medical diagnosis [6], recognition [7], and object tracking [8].

Image fusion tasks generally include infrared and visible image fusion, medical image fusion, multi-exposure image fusion and multi-focus image fusion. The first two types of image fusion belong to multi-modal image fusion, which integrates the valuable

* Corresponding author at: School of Automation, China University of Geosciences, Wuhan 430074, China.
E-mail address: chenjun71983@163.com (J. Chen).

https://doi.org/10.1016/j.neucom.2023.01.033
0925-2312/

Fig. 1. The examples of infrared and visible image fusion. From left to right: infrared images, visible images and fused images.

information in images collected by multiple sensors [9,10]. Subsequently, the latter two types of image fusion belong to digital photography fusion, which combines the meaningful information obtained via different settings to get a well-performing fused image [11]. Multi-modal image fusion is generally divided into feature extraction, feature fusion, and feature reconstruction [12]. Most of the existing image fusion research focuses on feature extraction and feature fusion to improve the performance of fusion networks [13]. Infrared and visible image fusion is an important topic in the field of image processing [14], with wide applications in the military and security fields, and this paper will focus on such problems.

Existing image fusion methods can be roughly divided into two categories, namely traditional methods [15] and deep learning-based methods [16]. The traditional image fusion methods mainly focus on the feature extraction of the original images. For example, the multi-scale transformation methods extract the multi-scale features from the original images and then fuse the extracted features through a carefully designed fusion strategy [17]. At last, the fusion image is reconstructed through an inverse multi-scale transformation. It can be seen that the fusion effect of such methods depends on the quality of the feature extraction. In addition to the above-mentioned multi-scale transformation methods, traditional image fusion methods include sparse representation-based, subspace-based, saliency-based, and total variation-based methods. Although traditional methods can obtain good fusion results after a long development period, there are many defects: (1) The improvement of the image fusion performance of traditional methods is limited by manual feature extraction and cannot cope with complex image fusion conditions [18]. (2) The traditional methods have to use the same strategy to extract features from different original images, and the generalization ability is relatively poor [19]. (3) The fusion strategy of traditional methods is relatively simple, resulting in ordinary fusion results [20].

Because of the shortcomings of the above traditional methods, researchers develop new fusion methods based on deep learning to solve some inherent problems of traditional image fusion methods. Fusion methods based on deep learning can be roughly divided into three categories: methods based on convolutional neural network (CNN) [21,22], methods based on autoencoder (AE) [23], and methods based on generative adversarial network (GAN) [24]. There are roughly two development routes for CNN methods. One is an end-to-end approach, where all tasks are done through the CNN, and the research focuses on how to design loss functions and network structures [25]. Another method is to use a pre-trained CNN as a feature fusion network and use traditional methods for feature extraction and feature reconstruction. The AE methods first train both encoder and decoder networks by reconstructing images from the training dataset. Using the encoder extracts the features of the original images, and then using the carefully designed fusion rules outputs the fusion features. Finally, the fused images are obtained through the decoder. The GAN methods continuously improve the performance of the fused images by establishing an adversarial game between the generator and the discriminator and completing the three necessary steps of image fusion in the above process [26].

Nowadays, the deep learning-based image fusion methods have driven significant progress in image fusion by their powerful feature extraction and generalization capabilities and have achieved performance far exceeding traditional methods [27,28]. However, those mentioned above deep learning-based image fusion methods use CNN for specific operations in the feature extraction stage [29]. Due to the small receptive field of CNN, it is difficult to model the long-range dependencies of images [30]. The limitation of the receptive field directly affects the quality of the fused images, so we need to model the global dependencies [31]. Therefore, we propose an end-to-end image fusion method combining CNN and vision transformer to solve the above problems. While retaining the advantages of CNN, the global dependencies of images are improved to obtain better-fused images. As shown in Fig. 12, the fusion results retain more salient features and detail information because of the addition of the transformer operation compared with the traditional CNN network, which is attributed to the increased receptive field and the full utilization of the global relationship of the image by adding the transformer.

The three main contributions of this paper are as follows:

- A hybrid feature extractor is proposed that combines dual-branch CNN and vision transformer to facilitate the simultaneous extraction of local and global information from images.
- The network structure of vision transformer is improved to make it more suitable for image fusion. In addition, the vision transformer is also used at the channel level of the features.
- A targeted perceptual loss function is designed. By calculating the loss of different depth features, the fusion images can retain more texture details and salient information.

2. Related work

This section mainly reviews the image fusion methods based on deep learning in recent years. In addition, we will briefly overview the development of vision transformer and its application in image fusion.

2.1. Deep learning-based image fusion algorithms

With CNN's powerful feature extraction ability and generalization ability, image fusion methods based on deep learning have been widely used in various image fusion fields, and they have shown excellent performance. DeepFuse was proposed by Prabhakar et al., who designed a CNN-based end-to-end fusion framework for multi-exposure image fusion without Ground Truth during the entire training and testing process [32]. Li et al. proposed DenseFuse [33]. Inspired by the DeepFuse above, they developed an infrared and visible image fusion method including an encoder, a decoder, and a fusion layer. The authors believe those dense connections can obtain more complementary information from the original images. In addition to feature extraction, the design of the fusion strategy is also crucial. Inspired by the DenseFuse network structure, Li et al. proposed NestFuse [34], which aims to enhance the salient features of the fused image while retaining more detail information about the original images. Besides, they designed a new spatial and channel self-attention model to extract more features. Based on the structure of NestFuse,

Li et al. proposed RFN-Nest, which aims to replace the ineffective handcrafted fusion strategies with a novel residual network structure [23].

Although most image fusion methods are aimed at a specific fusion task, the unified image fusion frameworks are gradually becoming a new research direction [35]. Such a unified framework can achieve better performance than training a framework of a specific task by jointly training on different fusion tasks. Zhang et al. proposed IFCNN, and they designed a supervised unified image fusion framework, which uses CNN to extract shallow information from the original images and fuses deep features through multiple fusion rules. And multiple fusion tasks share a trained model [36]. Zhang et al. proposed PMGI, an unsupervised unified image fusion framework. They transformed the image fusion task into preserving the ratio between texture information and intensity information of the original images. And based on the above ideas, a new loss function is proposed to constrain the fused images to contain more texture information and intensity information from the original images [16]. Xu et al. proposed U2Fusion and they designed an unsupervised unified image fusion network and proposed a novel loss function based on adaptive information preservation. The features are extracted by using pre-trained CNN in their method [37].

A generative adversarial network (GAN) is an algorithm that improves the quality of generated images by training a generator and a discriminator against each other. Goodfellow et al. first proposed Generative Adversarial Networks [38]. Alec et al. proposed DCGAN, the advantage of which is that its architectural state is relatively stable during training, which is longer than the training time of the original GAN. Mao et al. proposed LSGAN. Since the discriminators in other GAN methods are too powerful, they can quickly distinguish the true or false photos. LSGAN solves this problem by modifying the classification task of the discriminator into a regression task [39]. GAN-based image fusion research is gradually increasing. Ma et al. proposed FusionGAN, an unsupervised end-to-end fusion framework. They introduced GAN into the field of image fusion for the first time, providing a new idea for image fusion [40]. During training, the generator generates a fused image containing the visible image's detail information. It then forces the fused image to have more salient information from the infrared image through an adversarial game. Based on FusionGAN, they proposed DDcGAN to enhance the saliency of the fused images by introducing a dual discriminator structure and a new target-enhanced loss function [41].

The advantage of end-to-end image fusion methods is that no post-processing is required. Still, since the training datasets of most end-to-end methods are much smaller than sizeable natural image datasets (PASCAL VOC [42] and COCO [43]), this leads to overfitting of the trained model. Many methods will crop the image into small pieces to solve this problem. The disadvantage of this is that the global semantic information of the image is destroyed. The two-stage method can avoid the problem of too small datasets of the end-to-end method, but it also introduces new issues. Most two-stage methods do not consider the extraction of different features for different images through the autoencoder network when reconstructing images on the dataset. Even though novel fusion methods are designed in the second stage, this still limits the performance of the two-stage methods.

2.2. Vision transformer

In the field of image fusion, CNN and its variants are widely used due to its robust feature extraction ability and generalization ability. Not only that, its network structure is mature, and the corresponding computing equipment can accelerate its calculation. However, CNN has its own defects. CNN destroys the long-range dependencies of images, and the inherent small receptive field cannot effectively extract global information, which in turn affects the results of image fusion. Unfortunately, almost all existing image fusion frameworks use CNN networks as feature extractors without establishing long correlations in images [44]. The transformer method has long been a typical architecture in the field of natural language. Transformer has made many attempts on classic computer vision tasks, such as image classification, image segmentation, and object detection. The general framework of the vision transformer model is shown in Fig. 2, which is used for the image classification task [45]. Because transformer can achieve good results in modeling global features, we design a new feature extraction module by combining dual-branch CNN and vision transformer, which will effectively compensate for the inability of the CNN network to establish long-range dependencies in image fusion. It can make more effective use of local information and global information [46,47].

3. Method

This section will comprehensively introduce the infrared and visible image fusion network based on vision transformer and convolutional neural network. Firstly, we briefly introduce the network structure proposed in this paper. Then, we make a detailed analysis of the central part of the network. Finally, we present the loss function in detail.

3.1. Network architecture

The overall framework proposed in this paper is shown in Fig. 3, which is an end-to-end network. Our framework comprises three parts: a convolutional neural network module (CNN-Module), a vision transformer module (ViT-Module) and an image reconstruction module. The first two modules are called Hybrid Block, namely Hybrid Feature Extractor.

In the training phase, all input images will be resized to an appropriate size, the infrared image I_r and the visible image I_v are merged in the channel dimension and input to the Hybrid Block. The generated features from the Hybrid Block are sent to the image reconstruction module to obtain the fusion image I_f. Then, by calculating the loss function, the performance of the network model is continuously optimized until an ideal fusion framework is trained.

The loss function of our network is generally composed of two parts: per-pixel loss function, including structural similarity (SSIM) loss function, mean square error (MSE) loss function and total variation (TV) loss function; perceptual loss function, which uses VGG19 network to calculate the loss of the original images and the fused image separately in feature maps of different layers. The fused image can be generated directly using the trained module in the testing phase.

Inspired by Fu et al., the CNN-Module in this paper has two branches: the detail branch and the structure branch [48]. After the infrared image I_r and the visible image I_v are concatenated, they are sent to the initial convolutional layer as input, and then the output features are sent to the detail branch and the structure branch simultaneously. Finally, the output features from CNN-Module are obtained. The detail branch uses dense connections to extract the images' texture information and detail information. In contrast, the structure branch is used to extract the structural information of the image through the fast pooling method. The above two branches will output features, respectively, and then the features of the two branches will be concatenated and sent to the next module.

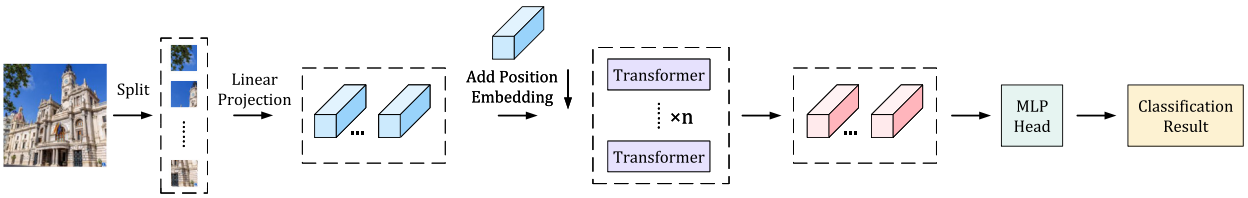

Fig. 2. Vision transformer model structure for classification tasks. They divide the input image into blocks of the same size, add the position embeddings for the classification task after linear projection, and then obtain the classification result through transformer and other operations.

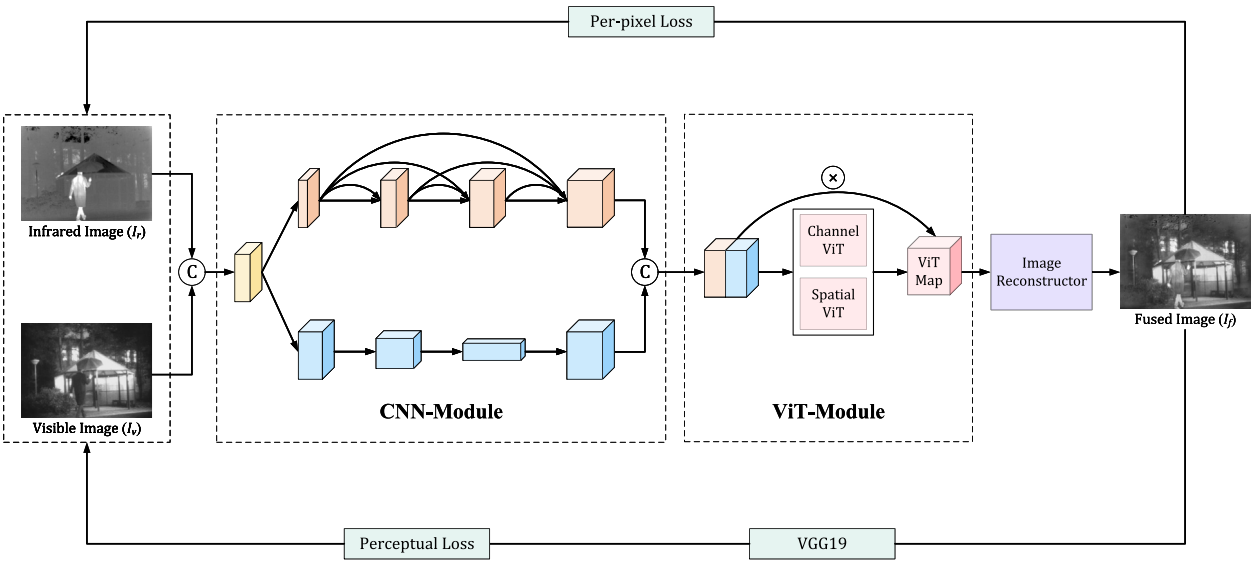

Fig. 3. The image fusion network based on vision transformer and convolutional neural network.

Inspired by RFN-Nest [23], we not only consider using a regular spatial transformer but also want to make full use of the channel relationship between features, so we add the channel transformer. The ViT-Module in this paper calculates the vision transformer maps by performing spatial vision transformer and channel vision transformer operations on the mixed CNN features. Then, the vision transformer maps multiply the mixed CNN features, and we finally send the result to the image reconstructor to get the fusion image. The primary function of the ViT-Module is to establish long-range dependencies of images, which makes up for the defects of the CNN.

3.2. Hybrid block

As shown in Fig. 3, the Hybrid Block consists of two parts: CNN-Module and ViT-Module. These two modules complement each other and jointly extract features from the original images, which helps us obtain a global relationship from images.

3.2.1. CNN module

We believe that CNN has more image-specific inductive bias than vision transformer, so we can not wholly discard CNN in the feature extraction process. CNN-Module can preprocess images efficiently, and we use two CNN branches to extract useful information from images from different aspects.

The two input images will be resized to 256 × 256, respectively, and 16 channels of features are obtained through a convolution. The reason for resizing the images to this size is that this size does not affect the computational efficiency, but also satisfies the requirements of subsequent transformer operations in the training phase [49]. And two branches share the same set of feature maps, which can effectively improve the computational efficiency. Then, the features are copied into copies and sent to the detail branch and the structure branch, respectively. The specific parameters of the CNN-Module and image reconstructor are shown in Table 1.

- Detail branch

In this branch, we set up four convolutional layers with dense connection operations between them to extract deep features of images. The number of input features is 16, and the number of channels after convolution of each layer is 8, 16, 24 and 32. The kernel size of each convolutional layer is 3 × 3, and the stride is 1. To keep the size of the features fixed, we use the reflection mode to pad images. The detail branch will finally output a set of features of size 256 × 256 with 32 channels.

- Structure branch

In the structure branch, we perform three convolutional layers. After each convolution operation, the size of the features is half of the previous step to achieve the aim of downsampling. The kernel size of each convolutional layer is 3 × 3, and the stride is 2. The size of the input features is 256 × 256, and the number of channels is 16. The size of the features after each convolution operation is 128 × 128, 64 × 64 and 32 × 32, and the number of channels of the features is 32, 64 and 32, respectively. In order to ensure that the size and number of channels of this branch are consistent with the previous branch, a bilinear upsampling layer is added here to output features with 32 channels and the size of 256 × 256.

ReLU activation function is widely used in computer vision, especially in image recognition tasks, image classification tasks,

Table 1
The parameters of CNN-Module and image reconstruction module.

Module	Layer	Input Channel	Output Channel	Input Size	Output Size
Initial CNN	In_Conv	1	16	256 × 256	256 × 256
CNN Module	D1_Conv	16	8	256 × 256	256 × 256
	D2_Conv	8	8	256 × 256	256 × 256
	D3_Conv	16	8	256 × 256	256 × 256
	D4_Conv	24	8	256 × 256	256 × 256
	S1_Conv	16	32	256 × 256	128 × 128
	S2_Conv	32	64	128 × 128	64 × 64
	S3_Conv	64	32	64 × 64	32 × 32
	S4_Conv	32	32	32 × 32	256 × 256
Image Reconstructor	R1_Conv	64	32	256 × 256	256 × 256
	R2_Conv	32	16	256 × 256	256 × 256
	R3_Conv	16	8	256 × 256	256 × 256
	R4_Conv	8	1	256 × 256	256 × 256

etc. The ReLU function is equivalent to directly discarding the negative activation, which may be beneficial for tasks such as image classification, but it is not suitable for image fusion tasks. In order to better meet the needs of the image fusion, we set the activation function of the Hybrid Block as Leaky ReLU, which can retain the negative activation information.

3.2.2. ViT module

The mixed CNN features are input into ViT-Module to make up for the inherent defects of CNN-Module so that our network can make full use of the local and global information of the original image.

Most of the existing vision transformer architectures are used in the fields of image classification, target recognition, and image segmentation. This paper makes some improvements to the current vision transformer architectures to make them more suitable for the field of image fusion. The upgrades are as follows: we abandoned CLS because classification information is not required in the image fusion task. Inspired by the TGFuse [50], the features calculated by ViT-Module are not directly sent to the image reconstruction module but multiplied with the output of the features by CNN-Module. Transformer essentially calculates the interrelationship between images. If the processing result of the transformer is directly put into the reconstruction module for image reconstruction, the effect of the transformer will be weakened [46]. Therefore, we believe that the dot product operation in our paper can maximize the effect of the transformer. On the basis of spatial transformer, channel transformer is added. Our network structure pays attention to the global information on each feature and the channel relationship between each feature. The input image size is processed so that images other than the training image size can be input during testing phase. The transformer details of our network are shown in Fig. 4 (b).

The calculation process of spatial transformer and channel transformer is shown in Fig. 5 and Fig. 6. Among them, B represents the batch size in the training task, C represents the number of channels of the features, H represents the height of the features, and W represents the width of the features. In the spatial transformer, the image is first divided into blocks, and h and w represent the width and height of the image block, respectively. Then the blocks are pulled into the form of vectors and sent to the transformer for processing. Finally, we reshape these to get the same size as the original features. The difference between the channel transformer and the spatial transformer is how to stretch the vector. The channel transformer preserves the channel dimension and calculates the relationship between different channels.

3.3. Image reconstruction module

As shown in Fig. 4 (b), the image reconstructor receives the features calculated from the Hybrid Block. We reduce the number of channels through continuous convolution operations and finally reduce the number of channels to one to output a fusion image. We set four convolutional layers for the image reconstructor, the

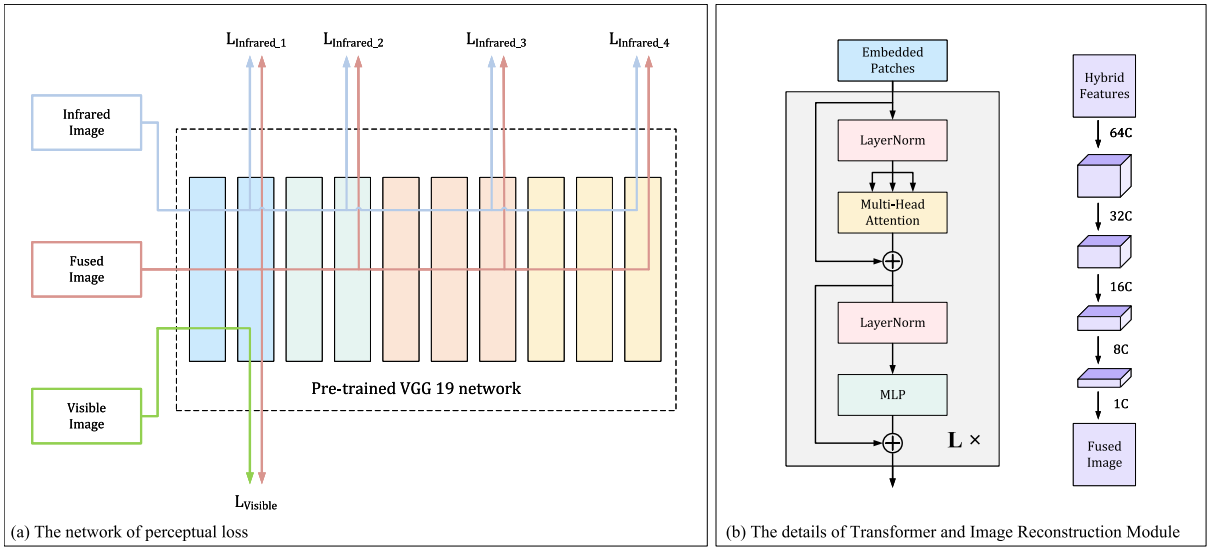

Fig. 4. (a) The network of perceptual loss in this paper. (b) The details of transformer and image reconstruction module.

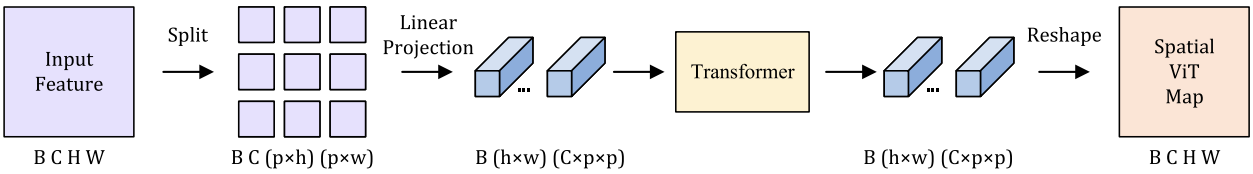

Fig. 5. The framework of spatial transformer in our network.

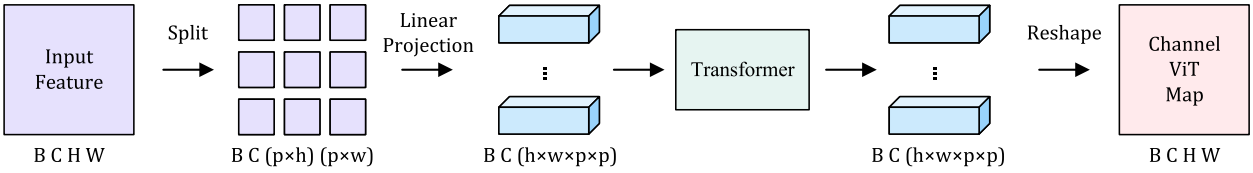

Fig. 6. The framework of channel transformer in our network.

number of input features is 64, and the number of feature channels after convolution of each layer is 64, 32, 16, 8, and 1. The kernel size of each convolutional layer is 3×3, and the stride is 1. To keep the size of the features fixed, we use the reflection mode to pad images.

3.4. Loss function

The loss function of our network is mainly composed of two parts, one is the per-pixel loss function, and the other is the perceptual loss function. λ is a hyperparameter to control the trade-off.

$$Loss = L_{Per-pixel} + \lambda L_{Perceptual} \tag{1}$$

3.4.1. Per-pixel loss

The per-pixel loss function mainly consists of three parts, which are defined as follows:

$$L_{Per-pixel} = L_{MSE} + \alpha L_{SSIM} + \beta L_{TV} \tag{2}$$

where L_{MSE} is the mean square error (MSE) loss function, L_{SSIM} represents the structural similarity (SSIM) loss function, L_{TV} denotes the total variance (TV) loss function, and α, β are two coefficients used to balance the loss function.

L_{MSE} is used for pixel-level reconstruction, which is defined as follows:

$$L_{MSE}(I_{out}, I_{in}) = \frac{1}{N}\sum_{n=1}^{N}(I_{out} - I_{in})^2 \tag{3}$$

where I_{out} is the output image by the image reconstruction module, and I_{in} is the original image of our network.

The SSIM loss function helps our network retain more structural information from the original images, which is defined as follows:

$$L_{SSIM} = 1 - SSIM(I_{out}, I_{in}) \tag{4}$$

The total variation loss is used to suppress the fused images' noise and retain the original images' gradient information. The formula is as follows:

$$R(p,q) = I_{out}(p,q) - I_{in}(p,q) \tag{5}$$

$$L_{TV} = \sum_{p,q}(||R(p,q+1) - R(p,q)||_2 + ||R(p+1,q) - R(p,q)||_2) \tag{6}$$

where $R(p,q)$ represents the difference between the output image and the original image, p is the horizontal coordinate of the image pixels, q is the vertical coordinate of the image pixels and $||\cdot||_2$ represents the Euclidean distance.

3.4.2. Perceptual loss

The general loss function calculates per-pixel loss between the input and output images. This loss cannot obtain the perceptual difference between the output and input images [51]. For example, two identical images that deviate from each other by only a few pixels, although perceptually similar, can be very different when measured by per-pixel loss. Moreover, it is groundless to design a fusion rule by assigning pixel-wise weight maps, which take the pixel-wise importance of feature maps into account. In this case, the roughness of the loss function limits the improvement of the fusion results. Therefore, even if there is a feature extraction way in the method that works well, it may not achieve its best performance due to the limitation of the loss function design [52]. The perceptual loss in this paper is calculated through the features of different depths of the pre-trained VGG19 network. For image fusion tasks, there is no so-called single correct output. We must make them semantically similar to calculate the loss between the output images and the original images. So, we generate high-quality images by combining perceptual loss and per-pixel loss.

The network of the perceptual loss function is shown in Fig. 4 (a). We divide the pre-trained VGG19 network into 4 levels according to the depth of the features. The deeper the features, the more semantic information contained in the extracted information. According to the characteristics of the image fusion task, when calculating the perceptual loss of the output images and the visible images, the calculation is performed on the relatively shallow features because the shallow features contain more structural information and detail information. In contrast, the perceptual loss of infrared images needs to be calculated on deeper features because we want to retain more salient features on infrared images [48].

Therefore, when calculating the perceptual loss between the output images by our network and the visible images and the infrared images, the levels of features used are 1st and 4th, respectively. The perceptual loss function between features is the Mean Absolute Error (MAE) loss function, and the formula is as follows:

$$L_{MAE}(\phi_{C_n}(\boldsymbol{I}), \phi_{C_m}(\boldsymbol{I})) = \frac{1}{N}\sum_{n=1}^{N}|\phi_{C_n}(\boldsymbol{I}) - \phi_{C_m}(\boldsymbol{I})| \tag{7}$$

where $\phi_{C_n}(\boldsymbol{I})$is the feature map by the convolutional layer before the n-th max-pooling layer in Fig. 4 (a).

4. Experiments and analysis

4.1. Setup

4.1.1. Datasets

The KAIST dataset contains outdoor scenes, and each pair of images is an infrared image and a visible image of the same scene. However, the number of images in the KAIST dataset [53] is too large, and there are many redundant image pairs. So we modify the KAIST dataset in the training phase, refer to the method proposed by Cao et al., and select 7601 pairs of infrared and visible images as training data [54]. In this paper, KAIST dataset is used to train the fusion network. In order to take into account the needs of the vision transformer in this paper, we uniformly resize the training images to 256 × 256 pixels and convert both images to grayscale images. Since we preprocess the image input to the Hybrid Block, the size of the test images we use in the testing phase can be inconsistent with the size of the training images. In addition, we select some images from the TNO dataset [55] and the RoadScene dataset [37] for testing.

4.1.2. Implementation Details

Our model is trained on NVIDIA GTX 3090 GPU, and our network is implemented with PyTorch. During the training phase, we set the learning rate to 1×10^{-4} and use the Adam optimizer. The training batch size is set to 44 and the training epoch is 50. The patch size of the spatial and channel vision transformer is set to 16.

4.1.3. Compared methods

We compare the method proposed in this paper with 9 traditional and deep learning methods in subjective evaluation and objective evaluation. These image fusion methods are Curvelet Transform (CVT) [56], DeepFuse [32], DenseFuse [33], FusionGAN [40], IFCNN [36], NestFuse [34], U2Fusion [37], PMGI [16], SDNet [57], and RFN-Nest [23].

4.2. Results analysis

Due to the particularity of the image fusion task, researchers cannot provide the unified Ground Truth, so evaluating the quality of an image fusion algorithm requires multiple investigations. In the field of image fusion, most people use two evaluation methods to evaluate fused images, namely subjective evaluation and objective evaluation. The subjective evaluation mainly starts from human visual perception and compares the fused image's brightness, sharpness, contrast, and overall visual effect. The salient features preservation of infrared images can also be considered in the infrared and visible image fusion task. Objective evaluation is to evaluate the fusion image through the objective evaluation metrics of the image. Now there are many kinds of objective evaluation metrics. We choose the more commonly used objective metrics for evaluation. In this section, our method will be compared with traditional methods and deep learning methods.

4.2.1. Subjective evaluation

In this section, we select some images from the TNO and RoadScene datasets, and then observe the experimental results of different methods intuitively. As shown in Fig. 7, the fusion results of our method are visually compared with the experimental results of other methods. From the comparison results, we can see that almost all methods have completed the image fusion task. However, from the perspective of preserving salient features, (c), (d), (e), (h) and (k) do not well maintain the salient information of infrared images, which is equivalent to discarding the salient information of infrared images. The essential information from the two original images is not well integrated. While our method retains the rich details of visible images, it still preserves the salient information of infrared images well, which is beneficial to subsequent image processing tasks.

From the perspective of clarity, (c), (g) and (k) do not preserve the details of visible images well, and the fused images are blurred. The method (j) is to add too many gradient constraints, resulting in the excessive sharpening of the fusion image, and the fusion image is seriously distorted compared with the actual scene. Although the salient information of method (f) is obvious, doing so will discard essential details in the visible images, causing the details to be darkened. It is impossible to see the places outside the salient area. Although method (i) does not have the above problems, in the transparent glass scene in the picture, method (i) does not handle the exposed part well, and there is obvious overexposure compared with our method. We can observe the same result from Fig. 8, Fig. 9, Fig. 10, and Fig. 11.

In summary, our fused result is more like a visible image plus the most significant salient features in an infrared image. The advantage of this result is that it not only retains the richer details of the visible image and the fusion image is clearer but also enables the valuable information of the infrared image to be added to the fusion image naturally, which is more in line with the human visual perception in the actual scene.

4.2.2. Objective evaluation

In the field of image fusion research, there are many objective indicators. We selected 20 pairs of images from the RoadScene dataset for quantitative experiments, and measured the comparative results of 9 metrics to evaluate our method. These metrics are Entropy (EN) [58], SF (Spatial Frequency) [59], PSNR (Peak

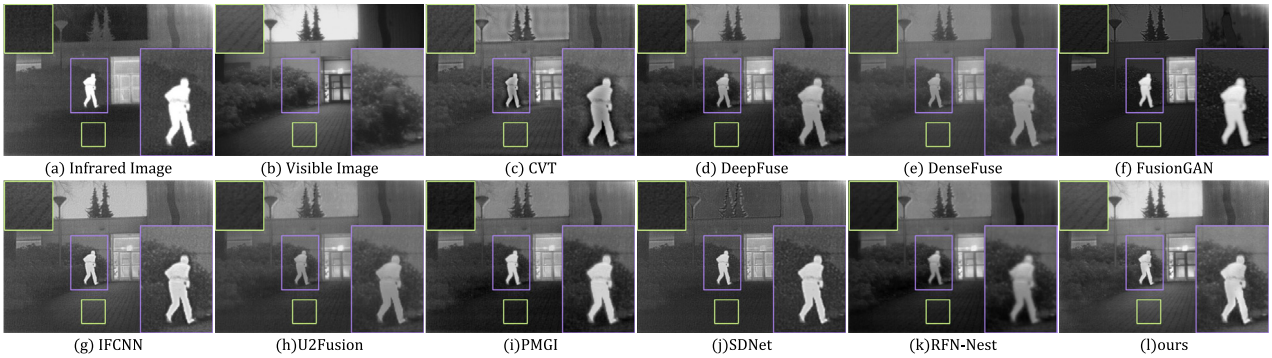

Fig. 7. Visualized results of our method compared with 9 state-of-the-art traditional and deep learning algorithms on *Kaptein_1123* scene of the TNO dataset. For a clear comparison, we select a salient region (*i.e.*, the purple box) in each image and zoom in it in the bottom right corner and highlight a texture area (*i.e.*, the green box) and zoom in it in the top left corner in each image.

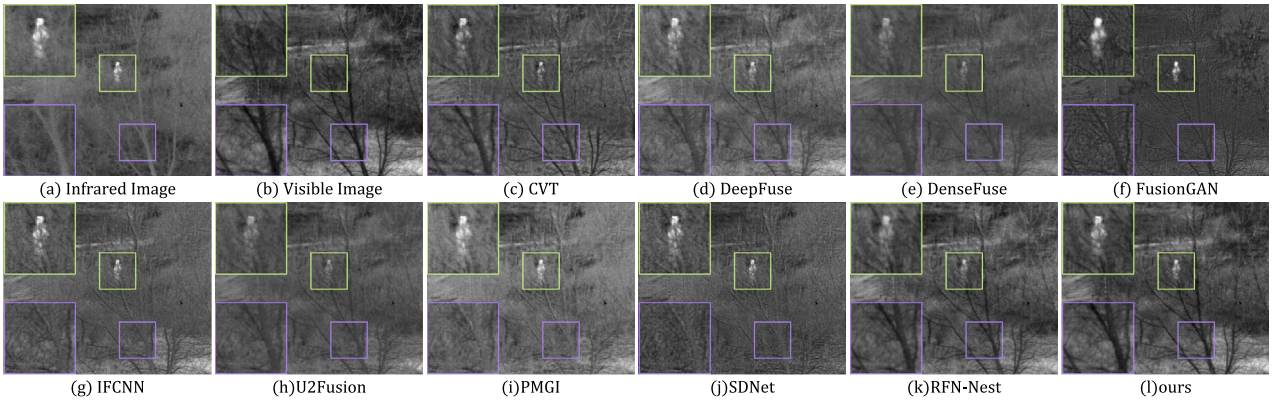

Fig. 8. Visualized results of our method compared with 9 state-of-the-art traditional and deep learning algorithms on *sandpath_18* scene of the TNO dataset. For a clear comparison, we select a salient region (*i.e.*, the green box) in each image and zoom in it in the top left corner and highlight a texture area (*i.e.*, the purple box) and zoom in it in the bottom left corner in each image.

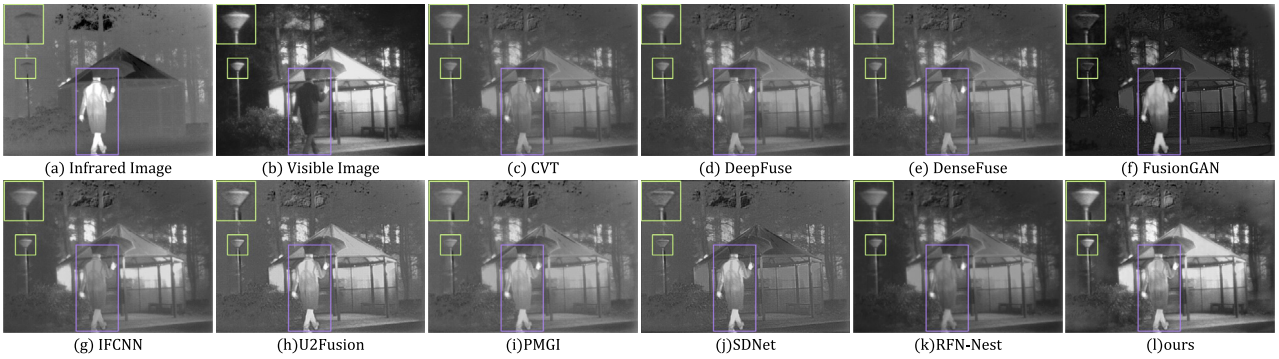

Fig. 9. Visualized results of our method compared with 9 state-of-the-art traditional and deep learning algorithms on *Kaptein_1654* scene of the TNO dataset. For a clear comparison, we select a salient region (*i.e.*, the purple box) in each image and highlight a texture area (*i.e.*, the green box) and zoom in it in the top left corner in each image.

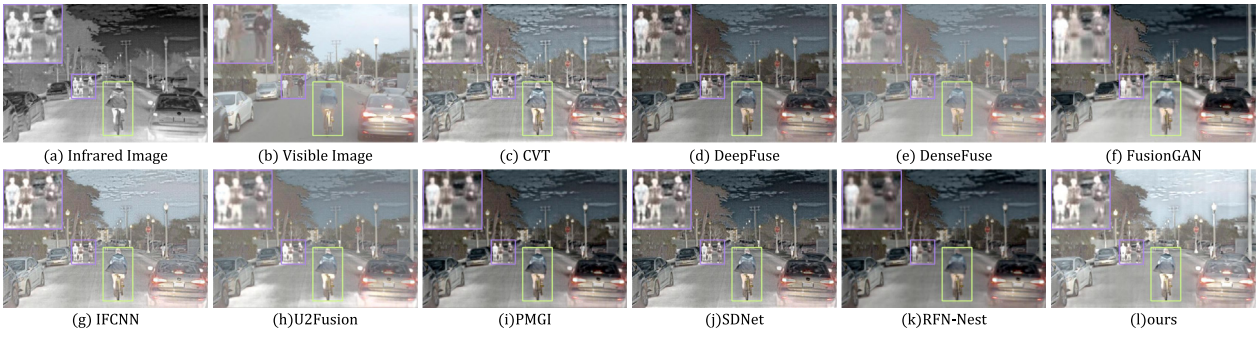

Fig. 10. Visualized results of our method compared with 9 state-of-the-art traditional and deep learning algorithms on *FLIR_06832* scene of the RoadScene dataset. For a clear comparison, we select a salient region (*i.e.*, the green box) and highlight an area (*i.e.*, the purple box) and zoom in it in the top left corner in each image.

Signal-to-Nosie Ration) [60], VIF (Visual Information Fidelity) [61], MI (Mutual Information) [62], SCD (Sum of the Correlations of Differences) [63], Q_{abf} (Quality of Images) [64], MS-SSIM (Multiscale SSIM) [65], FMI_{pixel} (Feature Mutual Information with Pixel) [66].

The definitions and functions of the above metrics are as follows: EN measures the amount of information contained in the images, and the larger the calculated entropy value, the more details retained in the fused images. PSNR measures the difference between the fused images and the original images. SF measures the spatial frequency of the images, and the higher the spatial frequency, the higher the image quality. VIF is an image quality assessment metric based on natural scene statistics and the conception of image information extracted by the human visual system. MI is used to measure the similarity between two images. SCD measures the quality according to the amount of original image information contained in the fusion images. The larger the value, the more information the fusion images contain in the original images. Q_{abf} measures the gradient-based fusion performance. MS-SSIM measures the similarity of images from a multi-scale perspective. FMI_{pixel} is calculated for mutual information of features from pixel features. The above metrics are the larger the value, the better the fused image effect.

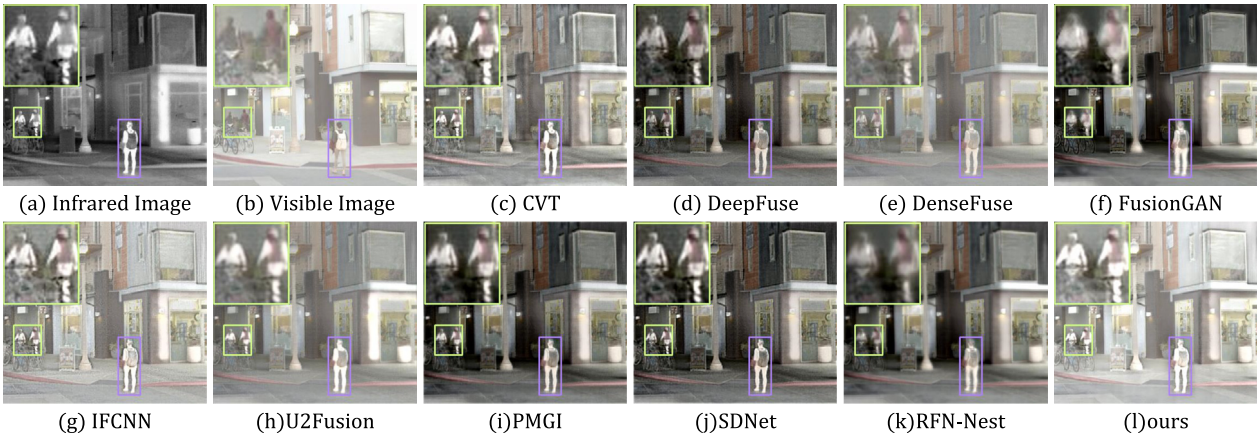

Fig. 11. Visualized results of our method compared with 9 state-of-the-art traditional and deep learning algorithms on *FLIR_08835* scene of the RoadScene dataset. For a clear comparison, we select a salient region (*i.e.*, the purple box) and highlight an area (*i.e.*, the green box) and zoom in it in the top left corner in each image.

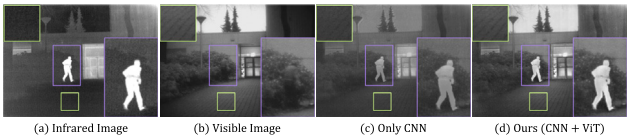

Fig. 12. Visualized results of our method compared with our network with only the ViT-Module removed on *Kaptein_1123* scene of the TNO dataset.

As shown in Table 2, our method performed best on six indicators, second best on two indicators, and third place on the remaining one indicator. Through subjective and objective evaluation, our method is proved to have obvious advantages in performance.

In order to evaluate the computational efficiency of different algorithms, we provide the average running times of different methods in Table 3 to demonstrate the efficiency advantage of our method in this paper. We can observe from the experimental results that our method is the fastest method on the three datasets. In addition, traditional methods usually take longer time to fuse images [57]. For example, curvelet transform is more time-consuming. In contrast, deep learning-based image fusion methods have a significant advantage in terms of running efficiency, mainly due to GPU acceleration and mature deep learning frameworks such as Pytorch [67]. Besides, our model does not have much complex structure and is a lightweight network [68]. The dual-branch design of CNN-Module also contributes to the reduced running time.

Table 3
Mean and standard deviation of the running times of all methods on the MFNet, RoadScene and TNO datasets (unit: second, RED indicates the best result and BLUE represents the second best result).

	MFNet	RoadScene	TNO
CVT	0.3092 ± 0.1735	0.0811 ± 0.0167	0.0461 ± 0.2964
DeepFuse	0.3268 ± 0.0554	0.0761 ± 0.0214	0.2665 ± 0.0241
DenseFuse	0.7775 ± 0.1535	0.5581 ± 0.1468	0.4149 ± 0.2143
FusionGAN	0.1872 ± 0.0119	0.3912 ± 0.0952	0.1996 ± 0.0945
IFCNN	0.0495 ± 0.0412	0.0574 ± 0.0014	0.0401 ± 0.0021
U2Fusion	0.3716 ± 0.1214	0.6886 ± 0.0987	0.3365 ± 0.3376
PMGI	0.3364 ± 0.0457	0.7464 ± 0.0845	0.3263 ± 0.2125
SDNet	0.0472 ± 0.1124	0.1652 ± 0.1204	0.0895 ± 0.0184
RFN-Nest	0.5288 ± 0.1741	0.1056 ± 0.0512	0.1192 ± 0.1927
Ours	0.0463 ± 0.1067	0.0489 ± 0.0102	0.0381 ± 0.0091

4.3. Ablation study

4.3.1. ViT module

In order to further verify the effectiveness of the ViT-Module in this paper, we compare the experimental results of the complete network in this paper with the network with only the ViT-Module removed.

According to the experimental results in Fig. 12, we can clearly see that the image fusion effect with the ViT-Module is much better than the network with only the CNN-Module. After adding the ViT-Module, more salient features in the infrared image are retained, making the target person appear brighter, with higher contrast and better retention of detail information. As shown in Fig. 13, the fusion result of (d) preserves the details in the purple rectangle better and maintains the salient features in the green rectangle more. Similarly, we can observe the same performance improvement in Fig. 14, Fig. 15, and Fig. 16.

From the above ablation experiments, adding the ViT-Module does help to improve the image fusion performance. The ViT-Module enhances long-range dependencies of images, allowing us to utilize local and global information fully. Furthermore, we

Table 2
Quantitative comparisons of representative methods. We select nine metrics to evaluate our method. RED indicates the best result. BLUE represents the second best result. BROWN indicates the third best result.

Method	EN	SF	PSNR	VIF	MI	SCD	Q_{abf}	MS-SSIM	FMI_{pixel}
CVT [56]	6.7779	0.0298	61.3181	0.6305	1.6692	1.5612	0.4552	0.8978	0.9389
DeepFuse [32]	6.9656	0.0281	63.4270	0.8142	2.2746	1.8251	0.4821	0.9053	0.9332
DenseFuse [33]	6.5276	0.0205	64.4970	0.6568	2.1339	1.5401	0.3104	0.8926	0.9321
FusionGAN [40]	6.1175	0.0241	61.2275	0.4714	1.9662	1.5018	0.2968	0.8044	0.9271
IFCNN [36]	6.9390	0.0310	63.3492	0.7747	2.4591	1.6900	0.4170	0.9272	0.9330
U2Fusion [37]	6.7580	0.0209	64.3618	0.6955	2.2164	1.6433	0.3517	0.9401	0.9306
PMGI [16]	6.9148	0.0303	63.4309	0.8388	2.5619	1.7881	0.4096	0.9159	0.9219
SDNet [57]	6.6325	0.0301	63.1688	0.6868	2.2468	1.4825	0.4051	0.8722	0.9192
RFN-Nest [23]	6.9988	0.0181	62.8070	0.7818	2.2184	1.8204	0.2614	0.9022	0.9357
Ours	6.9782	0.0311	64.2878	0.8605	2.6573	1.8272	0.4589	0.9543	0.9393

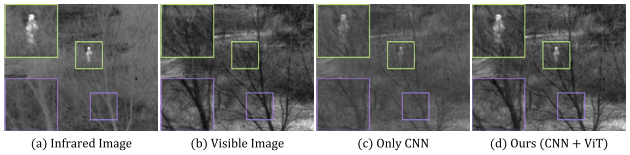

Fig. 13. Visualized results of our method compared with our network with only the ViT-Module removed on *sandpath_18* scene of the TNO dataset.

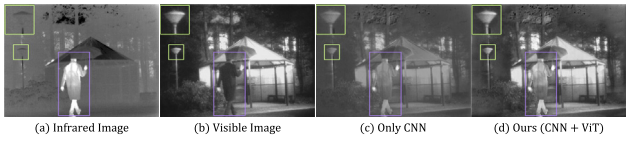

Fig. 14. Visualized results of our method compared with our network with only the ViT-Module removed on *Kaptein_1654* scene of the TNO dataset.

Fig. 15. Visualized results of our method compared with our network with only the ViT-Module removed on *FLIR_06832* scene of the RoadScene dataset.

Fig. 16. Visualized results of our method compared with our network with only the ViT-Module removed on *FLIR_08835* scene of the RoadScene dataset.

obtain valuable information about images from both spatial transformer and channel transformer perspectives. The above operations help us preserve more details of visible images and salient features of infrared images.

4.3.2. CNN Layers

The structural characteristics of ViT-Module determine the size of the features of CNN-Module. To facilitate subsequent calculations, we set the size of the features to 256 × 256. However, the number of feature channels of the CNN network is obtained through experiments. We prepared three schemes initially, namely 32 layers, 64 layers and 128 layers. The quantitative experimental results are shown in Table 4.

From the quantitative experimental results, it can be seen that the results of 64 channels are mostly better than the other two cases. This is because the number of channels of the features will also affect the fusion results. If the number of channels is too large, the extracted features may be redundant and computationally inefficient. If the number of channels is too small, the valuable information cannot be completely extracted from the images, and better fusion results cannot be obtained. Therefore, our structure finally selects the 64-channels CNN that performs better in most cases.

4.3.3. Transformer method

Based on the conventional spatial transformer, we add the channel transformer, as shown in Table 5. The combination strategy used in this method is better. So adding the channel transformer helps to improve the performance of fusion results.

5. Conclusion

This paper proposes an infrared and visible image fusion method based on vision transformer and convolutional neural network. Since our network is of the end-to-end type, post-processing of the fusion results is not required. The Hybrid Block integrates CNN-Module and ViT-Module, and the dual-branch CNN-Module has the more robust feature extraction capability. The addition of ViT-Module enables the network to take into account the local and global information of the images simultaneously, avoiding the problem of poor long-range dependencies of the traditional CNN network. In addition, we use the pre-trained VGG19 network to extract different features to calculate the loss and retain different types of image information in a targeted manner. Extensive experimental results show that the network structure with ViT-Module can achieve significantly better fusion results than CNN-Module alone. Compared with existing competitive methods, our method achieves desirable performance in both subjective and objective evaluations. The ultimate goal of image fusion is to combine with other computer vision tasks and make them better, so we will next try to improve the original results with image fusion driven by other computer vision tasks. In addition, although the focus of this paper is on infrared and visible image fusion, the network proposed in this paper can be used for other image fusion fields. We will try to apply this method to multi-exposure and medical image fusion in the future.

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.

Table 4
Quantitative results for different channels. RED indicates the best result.

	EN	SF	PSNR	VIF	MI	SCD	Q_{abf}	MS-SSIM	FMI_{pixel}
32 channels	6.9322	0.0278	64.4581	0.8464	2.5463	1.8159	0.4435	0.9379	0.9316
64 channels	6.9782	0.0311	64.2878	0.8605	2.6573	1.8272	0.4589	0.9543	0.9393
128 channels	6.9467	0.0302	65.5613	0.8141	2.5711	1.8061	0.4264	0.9532	0.9217

Table 5
Quantitative results for different transformer method. RED indicates the best result.

	EN	SF	PSNR	VIF	MI	SCD	Q_{abf}	MS-SSIM	FMI_{pixel}
spatial transformer	6.8457	0.0304	61.1558	0.8557	2.4524	1.8411	0.4414	0.9621	0.9233
channel transformer	6.2545	0.0289	62.1425	0.8425	2.1279	1.8257	0.4485	0.9485	0.9137
spatial + channel transformer	6.9782	0.0311	64.2878	0.8605	2.6573	1.8272	0.4589	0.9543	0.9393

Acknowledgment

This work was supported by the National Natural Science Foundation of China Nos. 62073304, 41977242 and 61973283.

References

[1] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fusion 76 (11) (2021) 323–336.
[2] H. Li, B. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, CVGIP Graph. Model. Image Process. 57 (3) (1995) 235–245.
[3] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, Y. Ma, Swinfusion: Cross-domain long-range learning for general image fusion via swin transformer, IEEE/CAA J. Autom. Sinica 9 (7) (2022) 1200–1217.
[4] T. Riley, M. Smith, Image fusion technology for security and surveillance applications, in: Optics and Photonics for Counterterrorism and Crime Fighting II, 2006, pp. 12–23.
[5] A.C. Muller, S. Narayanan, Cognitively-engineered multisensor image fusion for military applications, Inf. Fusion 10 (2) (2009) 137–149.
[6] G. Bhatnagar, Q.J. Wu, Z. Liu, Directive contrast based multimodal medical image fusion in nsct domain, IEEE Trans. Multim. 15 (5) (2013) 1014–1024.
[7] S. Singh, A. Gyaourova, G. Bebis, I. Pavlidis, Infrared and visible image fusion for face recognition, Biometric Technol. Human Identif. (2004) 585–596.
[8] Y. Zhu, C. Li, B. Luo, J. Tang, X. Wang, Dense feature aggregation and pruning for rgbt tracking, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 465–472.
[9] E. Daniel, J. Anitha, K. Kamaleshwaran, I. Rani, Optimum spectrum mask based medical image fusion using gray wolf optimization, Biomed. Signal Process. Control. 34 (4) (2017) 36–43.
[10] L. Tang, Y. Deng, Y. Ma, J. Huang, J. Ma, SuperFusion: A versatile image registration and fusion network with semantic awareness, IEEE/CAA J. Autom. Sinica 9 (12) (2022) 2121–2137.
[11] Q. Zhang, M.D. Levine, Robust multi-focus image fusion using multi-task sparse representation and spatial context, IEEE Trans. Image Process. 25 (5) (2016) 2045–2058.
[12] L. Tang, J. Yuan, H. Zhang, X. Jiang, J. Ma, Piafusion: A progressive infrared and visible image fusion network based on illumination aware, Inf. Fusion 83 (7) (2022) 79–92.
[13] M. Wu, Y. Ma, F. Fan, X. Mei, J. Huang, Infrared and visible image fusion via joint convolutional sparse representation, J. Opt. Soc. Am. A. 37 (7) (2020) 1105–1115.
[14] Y. Ma, J. Chen, C. Chen, F. Fan, J. Ma, Infrared and visible image fusion using total variation model, Neurocomputing 202 (8) (2016) 12–19.
[15] H. Li, X. Qi, W. Xie, Fast infrared and visible image fusion with structural decomposition, Knowl. Based Syst. 204 (9) (2020).
[16] H. Zhang, H. Xu, Y. Xiao, X. Guo, J. Ma, Rethinking the image fusion: A fast unified image fusion network based on proportional maintenance of gradient and intensity, in: Proceedings of the AAAI Conference on Artificial Intelligence, 2020, pp. 12797–12804.
[17] K.P. Upla, M.V. Joshi, P.P. Gajjar, An edge preserving multiresolution fusion: Use of contourlet transform and mrf prior, IEEE Trans. Geosci. Remote Sensing 53 (6) (2014) 3210–3220.
[18] J. Mou, W. Gao, Z. Song, Image fusion based on non-negative matrix factorization and infrared feature extraction, in: 2013 6th International Congress on Image and Signal Processing (CISP), 2013, pp. 1046–1050.
[19] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (7) (2017) 191–207.
[20] H. Li, X.-J. Wu, J. Kittler, Mdlatlrr: A novel decomposition method for infrared and visible image fusion, IEEE Trans. Image Process. 29 (2) (2020) 4733–4746.
[21] J. Ma, L. Tang, M. Xu, H. Zhang, G. Xiao, Stdfusionnet: An infrared and visible image fusion network based on salient target detection, IEEE Trans. Instrum. Meas. 70 (4) (2021) 1–13.
[22] L. Tang, J. Yuan, J. Ma, Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network, Inf. Fusion 82 (6) (2022) 28–42.
[23] H. Li, X.-J. Wu, J. Kittler, Rfn-nest: An end-to-end residual fusion network for infrared and visible images, Inf. Fusion 73 (9) (2021) 72–86.
[24] J. Ma, Y. Zhou, Infrared and visible image fusion via gradientlet filter, Comput. Vis. Image Underst. 197 (8) (2020).
[25] B. Ma, X. Yin, D. Wu, H. Shen, X. Ban, Y. Wang, End-to-end learning for simultaneously generating decision map and multi-focus image fusion result, Neurocomputing 470 (1) (2022) 204–216.
[26] A. Song, H. Duan, H. Pei, L. Ding, Triple-discriminator generative adversarial network for infrared and visible image fusion, Neurocomputing 483 (4) (2022) 183–194.
[27] H. Xu, X. Wang, J. Ma, Drf: Disentangled representation for visible and infrared image fusion, IEEE Trans. Instrum. Meas. 70 (2) (2021) 1–13.
[28] L. Liu, M. Chen, M. Xu, X. Li, Two-stream network for infrared and visible images fusion, Neurocomputing 460 (10) (2021) 50–58.
[29] C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, J. Zhang, Y. Sun, B. Zheng, Age-invariant face recognition by multi-feature fusionand decomposition with self-attention, ACM T. Multim. Comput. 18 (1s) (2022) 1–18.
[30] C. Yan, T. Teng, Y. Liu, Y. Zhang, H. Wang, X. Ji, Precise no-reference image quality evaluation based on distortion identification, ACM T. Multim. Comput. 17 (3s) (2021) 1–21.
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.
[32] K. Ram Prabhakar, V. Sai Srikar, R. Venkatesh Babu, Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 4714–4722.
[33] H. Li, X.-J. Wu, Densefuse: A fusion approach to infrared and visible images, IEEE Trans. Image Process. 28 (5) (2018) 2614–2623.
[34] H. Li, X.-J. Wu, T. Durrani, Nestfuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models, IEEE Trans. Instrum. Meas. 69 (12) (2020) 9645–9656.
[35] A. Fang, X. Zhao, J. Yang, B. Qin, Y. Zhang, A light-weight, efficient, and general cross-modal image fusion network, Neurocomputing 463 (11) (2021) 198–211.
[36] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, L. Zhang, Ifcnn: A general image fusion framework based on convolutional neural network, Inf. Fusion 54 (2) (2020) 99–118.
[37] H. Xu, J. Ma, J. Jiang, X. Guo, H. Ling, U2fusion: A unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2022) 502–518.
[38] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in neural information processing systems, 2014, pp. 2672–2680.
[39] X. Mao, Q. Li, H. Xie, R.Y. Lau, Z. Wang, S. Paul Smolley, Least squares generative adversarial networks, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 2794–2802.
[40] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (8) (2019) 11–26.
[41] J. Ma, H. Xu, J. Jiang, X. Mei, X.-P. Zhang, Ddcgan: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion, IEEE Trans. Image Process. 29 (3) (2020) 4980–4995.
[42] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The pascal visual object classes (voc) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft coco: Common objects in context, in: European conference on computer vision, 2014, pp. 740–755.
[44] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Q. Hou, J. Feng, Deepvit: Towards deeper vision transformer, arXiv preprint arXiv:2103.11886.
[45] C.-F.R. Chen, Q. Fan, R. Panda, Crossvit: Cross-attention multi-scale vision transformer for image classification, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 357–366.
[46] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, S.J. Oh, Rethinking spatial dimensions of vision transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 11936–11945.
[47] J.I. Olszewska, C. De Vleeschouwer, B. Macq, Multi-feature vector flow for active contour tracking, in: in: 2008 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2008, pp. 721–724.
[48] Y. Fu, X.-J. Wu, A dual-branch network for infrared and visible image fusion, in: 2020 25th International Conference on Pattern Recognition (ICPR), 2021, pp. 10675–10680.
[49] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: 9th International Conference on Learning Representations (ICLR), 2021, pp. 1–21.
[50] D. Rao, X.-J. Wu, T. Xu, Tgfuse: An infrared and visible image fusion approach based on transformer and generative adversarial network, arXiv preprint arXiv:2201.10147.
[51] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European conference on computer vision, 2016, pp. 694–711.
[52] H. Xu, H. Zhang, J. Ma, Classification saliency-based rule for visible and infrared image fusion, IEEE Trans. Comput. Imag. 7 (7) (2021) 824–836.
[53] S. Hwang, J. Park, N. Kim, Y. Choi, I. So Kweon, Multispectral pedestrian detection: Benchmark dataset and baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 1037–1045.
[54] Y. Cao, D. Guan, W. Huang, J. Yang, Y. Cao, Y. Qiao, Pedestrian detection with unsupervised multispectral feature learning using deep neural networks, Inf. Fusion 46 (3) (2019) 206–217.
[55] A. Toet, Tno image fusion dataset, URL (2014), https://doi.org/10.6084/m9.figshare.1008029.v1.
[56] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2) (2007) 143–156.
[57] H. Zhang, J. Ma, Sdnet: A versatile squeeze-and-decomposition network for real-time image fusion, Int. J. Comput. Vis. 129 (10) (2021) 2761–2785.
[58] J.W. Roberts, J.A. Van Aardt, F.B. Ahmed, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote. Sens. 2 (1) (2008).
[59] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, W. Wu, Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2011) 94–109.

[60] P. Jagalingam, A.V. Hegde, A review of quality metrics for fused image, Aquat. Procedia 4 (2) (2015) 133–142.
[61] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[62] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[63] V. Aslantas, E. Bendes, A new image quality metric for image fusion: The sum of the correlations of differences, AEU-Int. J. Electron. C. 69 (12) (2015) 1890–1896.
[64] B. Shreyamsha Kumar, Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform, Signal Image Video Process. 7 (6) (2013) 1125–1143.
[65] K. Ma, K. Zeng, Z. Wang, Perceptual quality assessment for multi-exposure image fusion, IEEE Trans. Image Process. 24 (11) (2015) 3345–3356.
[66] M. Haghighat, M.A. Razian, Fast-fmi: non-reference image fusion metric, in: 2014 IEEE 8th International Conference on Application of Information and Communication Technologies (AICT), 2014, pp. 1–3.
[67] Z.-R. Jin, L.-J. Deng, T.-J. Zhang, X.-X. Jin, Bam: Bilateral activation mechanism for image fusion, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4315–4323.
[68] G. Huang, Z. Liu, L. Van Der Maaten, K.Q. Weinberger, Densely connected convolutional networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 4700–4708.

Jun Chen received the B.S. degree in electronic and information engineering and the M.S. degree in communication and information system from the China University of Geosciences, Wuhan, China, in 2002 and 2004, respectively, and the Ph.D. degree in communication and information system from the Huazhong University of technology, Wuhan, China, in 2014. From 2004 to 2008, she was an Assistant Professor with the China University of Geosciences, where she is currently an Associate Professor with the School of Automation. Her research interests include computer vision and pattern recognition, geoscience and remote sensing.

Jianfeng Ding received the B.S. degree from the College of Mechanical Engineering, Chongqing University of Technology, Chongqing, China, in 2021. He is currently working toward the M.S. degree with the School of Automation, China University of Geosciences, Wuhan, China. His research interests include image fusion and machine learning.

Yang Yu received the B.S. degree in electronic and information engineering from University of Science and Technology of China, Hefei, China, in 2005, and the Ph.D. degree in circuits and systems from Shanghai Institute of Technical Physics, Chinese Academy of Sciences, Shanghai, in 2010. She is an Associate Research Fellow with Shanghai Institute of Technical Physics, Chinese Academy of Sciences. Her research interests include Infrared optical imaging technology, opto-electric imaging technology, etc.

Wenping Gong received the B.S. degree in civil engineering from Tongji University, Shanghai, China, in 2011, and the Ph.D. degree in civil engineering from Clemson University, Clemson, SC, USA, in 2014. He is a Professor with the Faculty of Engineering, China University of Geosciences, Wuhan, China. Dr. Gong is on the Editorial Board of Engineering Geology, Bulletin of Engineering Geology and the Environment, Marine Georesources and Geotechnology, and International Journal of Geotechnical Engineering.