Infrared Physics & Technology 125 (2022) 104242

Contents lists available at ScienceDirect

Infrared Physics and Technology

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

Multi-scale unsupervised network for infrared and visible image fusion based on joint attention mechanism

Dongdong Xu *, Ning Zhang, Yuxi Zhang, Zheng Li, Zhikang Zhao, Yongcheng Wang

Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

ARTICLE INFO

Keywords:
Attention mechanism
Perceptual loss
Infrared and visible images
Image fusion
Deep learning

ABSTRACT

Infrared and visible image fusion can synthesize complementary features of salient objects and texture details which are important for all-weather detection and other tasks. Nowadays, the deep learning based unsupervised fusion solutions are preferred and have obtained good results since the reference images for fusion tasks are not available. In the existing methods, some prominent features are missing in the fused images and the visual vitality needs to be improved. From this thought, attention mechanism is introduced to the fusion network. Especially, channel dimension and spatial dimension attention are jointed to supplement each other for feature extraction. Multiple attention branches emphasize on multi-scale features to complete the encoding. Skip connections are added to learn residual information. The multi-layer perceptual loss, the structure similarity loss and the content loss together construct the strong constraints for training. Comparative experiments with subjective and objective evaluations on 4 traditional and 9 deep learning based methods demonstrate the advantages of the proposed model.

1. Introduction

With the rapid development of computer science, integrated circuit system and sensor technology, image acquisition is no longer limited to a single sensor. Multi-modal image fusion is able to combine the meaningful information in different source images to generate a synthesized image with rich features, so as to achieve all-round description and expression of same scene or object which is more beneficial for subsequent applications [1,2,3]. It can also meet the urgent needs of modern applications for comprehensive information. Infrared image gives expression to the thermal radiation intensity of objects. The penetrating force of infrared signal is quite strong. So, the image is not susceptible to weather conditions and environmental changes. But the spatial resolution is low which is unavoidable. However, the imaging of visible sensors depends on the reflectivity of objects. The images with high resolution can well retain the environmental details in the scene, and the information is rich. But the visible imaging is greatly affected by the light source and illumination conditions. There are some shortcomings, such as short detection range and poor environmental adaptability. From here we can see that the fusion of the two images is good for integrating prominent targets and rich environmental details [4], which is of great importance in all-weather tasks. Image fusion as an enhancement method is widely used in many fields such as military detection, medical diagnosis, public safety, and industrial production, etc. [5].

At present, correlative studies on infrared and visible image fusion are gradually mature. Different extraction strategies and transformations are promoted and combined in various methods. Generally, they can be roughly divided into two groups which are traditional methods and the deep learning (DL) based methods [6]. Related methods are introduced below. The traditional fusion methods are mainly in the transformation domain, multi-scale transformation (MST) [7,8,9] and sparse representation (SR) [10,11] are often adopted. In addition, spatial domain methods [12,13] and artificial neural network methods [14,15,16] as well as hybrid methods [17,18,19] also play important roles. There are mainly three steps when implementing above-mentioned methods. Image transformations are carried out at first. Then activity level measurement and corresponding fusion rules need to be designed. All these are operated manually and the realization process becomes very complicated when getting better results. As to DL-based fusion methods, the biggest advantage is the representation ability of features. The fusion process is realized by training multi-layer

* Corresponding author.
E-mail addresses: sdwhxdd@126.com, xudongdong@ciomp.ac.cn (D. Xu), neuq2013zn@163.com (N. Zhang), zhangyuxi18@mails.ucas.ac.cn (Y. Zhang), lizheng20@mails.ucas.ac.cn (Z. Li), zzkzhaozhikang@outlook.com (Z. Zhao), wangyc@ciomp.ac.cn (Y. Wang).

https://doi.org/10.1016/j.infrared.2022.104242
Received 21 October 2021; Received in revised form 24 May 2022; Accepted 5 June 2022
Available online 13 June 2022
1350-4495/© 2022 Elsevier B.V. All rights reserved.

convolutional neural networks. All the features can be extracted automatically and we do not have to perform scale transformation and activity level measurement manually. Liu et al. [20] firstly realized infrared and visible image fusion using the Siamese network. Also, MST was included. Then Li et al. [21] added the DenseNet [22] to the encoder so as to retain more features of middle layers. VGG-network [23], ResNet [24] and other trained modules were also used as feature extraction sub-networks for the reconstruction of fused image. The above-mentioned DL-based methods almost depend on convolutional neural networks (CNN). In the last few years, Ma et al. began to study on generative adversarial networks (GAN) to implement the fusion task, so that more detail information could be preserved through the adversarial process. FusionGAN [25] and its variant [26] are representative and fundamental models for this fusion task. Moreover, Li et al. [6] came up with the RCGAN model with shared weights. After that, DDCGAN [27] and D2WGAN [28] were designed with two discriminators to complement the information of the multi-modal images.

In fact, the imaging principles of infrared and visible images are quite different. One focuses on the highlighted thermal objects, the other on visible texture details in images. How to integrate these features is extremely important for fusion. Therefore, some researchers introduced the attention mechanism to stress the distinguishing features in DL-based fusion models. Li et al. [29,30] proposed two GAN-based models, called the MgAN-Fuse and AttentionFGAN, which could maintain the important parts of each source image better. Mustafa et al. [31] and Li et al. [32] also achieved competitive performance based on convolutional attention networks. However, aforementioned methods mainly pay attention to channel dimension. Spatial attention is largely ignored which is complementary to the channel attention. Meanwhile, the multi-scale features extracted and preserved by those deep neural networks are relatively deficient so that the features of source images cannot be reflected well.

To address the above-mentioned fusion problems, a CNN model with multi-scale attention mechanism is proposed. We introduce CBAM [33] with both channel attention and spatial attention to realize this multi-modal image fusion task. We learn from SKNet [34] and ResNeXt [35] to obtain multi-scale feature maps by designing various receptive fields and residual operations. Moreover, synthetical loss function is used to guide the process of parameter updating. Different levels of feature similarities are all well constrained. The main work of this article can be summarized to three points.

1) We build a multi-scale convolutional fusion model with joint channel attention and spatial attention, so that more salient features can be concerned and efficiently extracted. Different perceptive fields adopted in this multi-branch architecture help to get complementary feature maps which are important for image fusion tasks. Skip connections are added to learn residual information and assist the back propagation. The network can be trained and tested end-to-end.
2) The particular perceptual loss with the adjusted image is designed for dimension and feature matching. Also, it is associated with structural similarity (SSIM) loss and content loss to make a powerful feature constraint between original inputs and generated image during training. More prominent information and targets can be reserved. The loss can also accelerate the iterative convergence speed.
3) The deep fusion model is trained on the amplified dataset and tested on image pairs from three different datasets. Related ablation experiments are used for contrastive analysis. The visual and numerical results of the proposed network are worth affirming compared with traditional and other latest DL-based methods. Both subjective and objective evaluations prove the feasibility of this network. Other contents of the article are summarized below. Section II focuses on the CBAM and the DL-based models with attention mechanism. Section III mainly explains the designed network, attention module and loss function. The details of training and comparative experiments are shown in Section IV. Section V outlines brief conclusion and expectation at the end.

2. Related works

2.1. Attention mechanism

Attention mechanism can be regarded as a kind of allocation of resources. It can be understood that the resources originally distributed equally are redistributed according to the importance of the attentive objects. As to deep neural network, the weights for each layer are key resources that need to be noticed. Attention mechanism was first applied in natural language processing (NLP) [36] which greatly improved the capacity. Then, in computer vision (CV) field, some scholars have explored different strategies with attention mechanism to promote the deep neural networks. According to the different implementations, the attention mechanism can be classified into local-attention, soft-attention and hard-attention [37]. Among them, the soft-attention is the most widely used and generally consists of channel-wise and spatial-wise attention, as well as the joint module of the first two. The CBAM is the typical joint attention and can be embedded into any network branch easily. Therefore, the improved network will obtain 'what' and 'where' information simultaneously contained in different dimensions of the images. The content and location information are important for image process. Fig. 1 is the unique sketch map of CBAM.

In this paper, the CBAM is embedded as sub-network to every branch for feature extraction. Different perceptive fields are adopted in spatial attention module to get multi-scale feature maps for image reconstruction. All these help to preserve more details and targets for fusion tasks.

2.2. Attention networks used for image fusion task

Multi-layer networks embedded with attention modules now being used for fusing infrared and visible images. Mustafa [31] first put forward a dilated residual self-attention network to generate fusion images which included balanced details of source inputs. Li et al. [32] combined the DenseNet and attention unit [34] to produce fusion images, and got relatively accurate results. More typical DL-based models with attention mechanism were MgAN-Fuse and AttentionFGAN proposed by Li et al. They constructed the attention network by GAN and the final attention map was acquired through the mapping transformation in [38]. For channel attention, a series of global average pooling (GP), full connection (FC) and sigmoid function (SG) are applied to normalize the features. For spatial attention, they usually take two ways to get final map. One is to compute the maximum values across the channel dimension. Another way is to concatenate the extracted feature maps directly in channel wise without other calculation. Fig. 2 gives the schematic illustration of the attention module used.

In addition to the above methods, scholars [38] proposed the parallel fusion strategy based on spatial attention and channel attention models. They directly took L1 norm and global pooling with soft-max to perform channel and spatial dimension feature processing. Then all the feature maps extracted were added for subsequent integration.

We can see that although the channel and spatial dimensions are considered in those methods, there are still some improvements can be

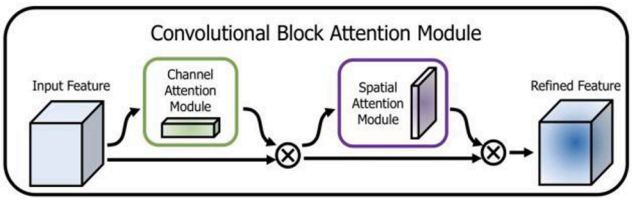


Fig. 1. The overview of CBAM [33]. The module has two sequential sub-modules: channel and spatial module.

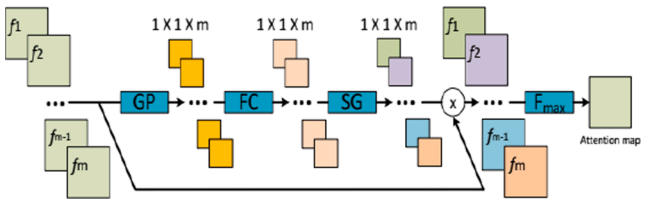


Fig. 2. The attention module of MgAN-Fuse [29]. f^m Denotes the m-th feature. F_{max} represents the selection of max value across the channel dimension of all the reweighted features. ⊗ is the element-wise multiplication.

made on the network architecture and the series–parallel mode of the two attention sub-module, to get more prominent feature maps.

3. Proposed method

The specific designs of the convolutional attention network will be introduced in detail. First, we will give a general description about the network. Fig. 3 shows the overview of the designed network architecture. Then, the channel attention, the spatial attention and the residual learning used are explained. Furthermore, we analyze the mixed loss functions. Especially, the perceptual loss is focused on.

3.1. Framework overview

We can see from Fig. 3 that the whole framework is an end-to-end model composed of convolutional layers. The proposed CNN model is easy to train compared with GAN which needs to take cooperative training of generator and discriminator into consideration. The input in the far left is the concatenation of the source images in channel dimension, since this 2-channel layer is considered to have the highest flexibility [39] compared with the Siamese form or its variants. The first convolution group consisted of Conv, BN and LReLU can get the basic features of the input. Then the attention modules are followed to transform and extract significant features. We take different kernels when carry out the spatial attention in each branch to pays attention to a certain scale of the basic features. Skip connection is used to learn the residual of every branch. The features extracted by the three attention branches are concatenated afterwards to construct new prominent maps in channel dimension. This will help to preserve multi-scale targets or textures in source images. What stated above is the encoding process of the network. The latter layers mainly complete the decoding operations. The dimensionality of the concatenated feature maps are reduced through two convolution groups so that the number of channels is consistent with the basic features. Skip connection is again introduced to guarantee the completeness of basic features and extracted features which is crucial for image fusion tasks. The next three layers accomplish the final integration. Especially, at the last layer of the network, the kernel size is changed to 1x1. There is no BN and the activation function is changed to Tanh. The convolution can be regarded as the point to point multiplication between the 32-channel feature vector and the previous layer. After the operation, the multi-channel feature maps are transformed to single layer fused image.

3.2. Attention modules

In each attention branch, the F_b are reformed and recombined to get the F_A with more salient features. Fig. 4 and Fig. 5 show the channel and spatial attention module respectively.

3.3. Channel attention

We can see from Fig. 4 that the channel attention mainly includes four steps: pooling, multi-layer perceptron, activation and multiplication. The goal of channel attention is to get a scalar which can reflect the degree of dependence on each channel. First, the average-pooling (AP) and max-pooling (MP) are introduced so that important clues of different objects can be gathered. Then a shared network is used to produce channel attention maps. The result of addition is activated by sigmoid function to obtain the final descriptor to calculate the F_C.

The transformation equation of F_C can be summarized as below:

$$F_C = F_b * \sigma(MLP(AP(F_b)) + MLP(MP(F_b))) \quad (1)$$

The channel attention can make feature adjustment channel by channel and improve the representation ability of the network. More useful and salient information of the source inputs will be extracted so that the pertinence and veracity of fusion are effectively enhanced.

3.4. Spatial attention

In the proposed method, the spatial attention is connected behind the channel attention and the F_C are given as the input. There are also four steps to realize the transformation. Other than channel attention, the AP and MP operations in spatial attention act on channel neurons and the concatenated feature maps are in the same size with F_C. Convolution is followed to make feature reforming. Then we take the sigmoid function to activate that spatial attention map which guides the tendency of F_S. Equation (2) shows the arithmetic process.

$$F_S = F_C * \sigma(Conv\ (AP(F_C), MP(F_C))) \quad (2)$$

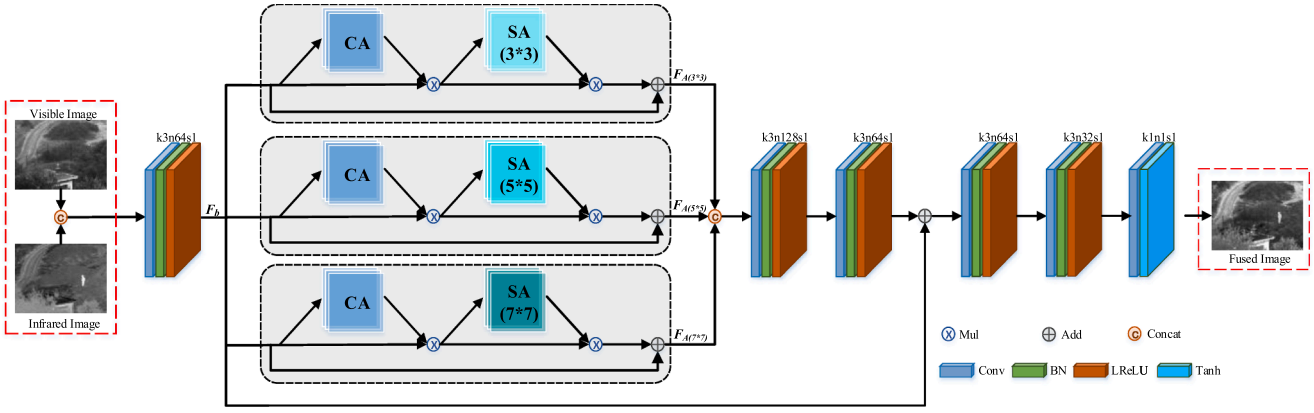


Fig. 3. Architecture of the proposed. The input is the channel concatenation of infrared and visible images. The features are extracted and transferred by convolution operations. F_b and F_A are the basic features and the features with attention. CA and SA represent the channel attention and the spatial attention. Different perceptive fields are marked in SA. ⊗, ⊕, © donate the element-wise multiplication, element-wise summation and concatenation respectively. Conv, BN and LReLU are the abbreviations of convolution, batch normalization and Leaky ReLU.

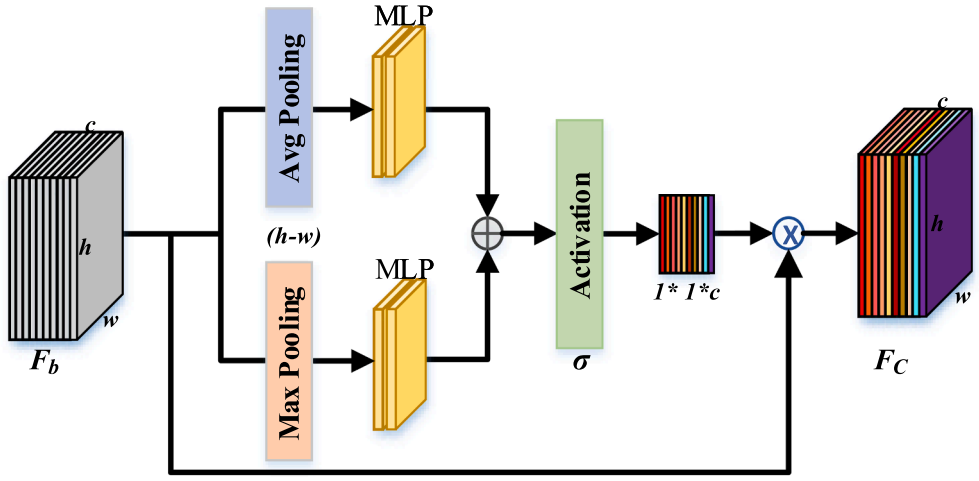


Fig. 4. The diagram of channel attention. F_b and F_C are the basic features and the features after channel attention. MLP represents the multi-layer perceptron. σ is the Sigmoid function. h, w, and c denote the height, width and the number of channels of the feature maps.

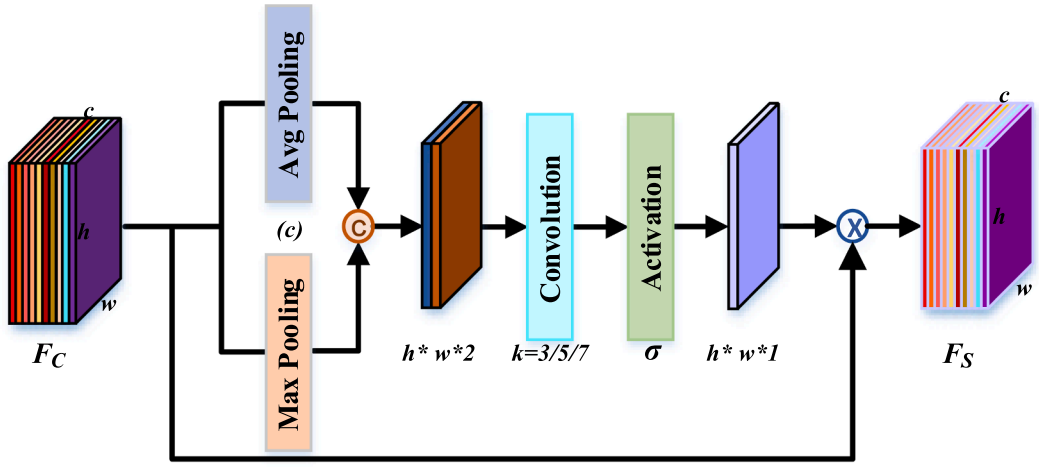


Fig. 5. The diagram of spatial attention. F_S are the features maps after spatial attention. h, w, and c denote the height, width and the number of channels of the feature maps. k denotes the kernel size of convolution.

The kernel size of the original CBAM is 7 × 7 when carrying out the convolution. In consideration of the particularity of the multi-modal image fusion, we creatively design different perceptive fields like 3 × 3 and 5 × 5 to complement the multi-scale spatial features in source inputs. The three branches with distinct kernels contribute to collect the highlighted information from source inputs to generated fusion image.

3.5. Residual learning

We all know that the fundamental features and deeply extracted features are all important for fusion tasks. In the proposed network, each attention branch conducts feature refining from F_b to F_A in a specific scale. Skip connection is added in every branch to realize residual learning. More shallow features are combined with noticed features when encoding. Also, the connection is built in the decoding process. So the final F_A of each branch is obtained by equation (3):

$$F_A = F_b + F_S \tag{3}$$

3.6. Loss function

The network structure proposed in this paper is a multi-layer CNN. We need to design forceful loss function so that the training process and fusion effect can be guaranteed. The loss function should have the ability to make full-scale constraint between input and output images. The similarities of low-level and high-level features need to be considered simultaneously. So, we come up with the mixed loss (L) composed of three parts which are SSIM loss (L_{SSIM}), perceptual loss (L_{Per}), and content loss (L_{Con}). The loss function with adjustable ratios is defined as equation (4).

$$L = \alpha L_{SSIM} + \beta_{Per} + \theta L_{Con} \tag{4}$$

3.7. SSIM loss

At present, the L_{SSIM} is most widely used in the training of unsupervised fusion networks. Mainly because the brightness, contrast and structure information are taken into account and it just similar to the observation mechanism of the human visual system. The whole calculation process effectively establishes the structure correlations between images. So, we take the L_{SSIM} as the fundamental loss to conduct the back propagation of the proposed network. The calculation equation is illustrated in (5):

$$L_{SSIM} = 1 - (w \cdot SSIM(I, F) + (1 - w) \cdot SSIM(V, F)) \tag{5}$$

I, V, and F indicate the infrared image, visible image and the fusion

image. w denotes the adjustable coefficient. $SSIM(\sim)$ executes the calculation of structural similarity which is consistent with [40].

3.8. Perceptual loss

Johnson et al. [40] first came up with the perceptual loss and the findings behaved well on image style transfer and super-resolution. Unlike normal losses, this loss acts on feature maps other than calculating the similarity between the original source inputs and outputs. There is an intermediate loss network used to extract the high-level features of the reference and the generated image on the same layer. Generally, if the two groups of output features are numerically similar, it can be considered that the reference and the generated image are closely related indirectly. As for image fusion, the loss will effectively ensure the similarity of the extracted information, and finally achieve the feature retention from source inputs to the fusion image.

SDPNet [41] and some methods took the networks they designed to calculate the loss. Since our purpose is high-level feature extraction and comparison, we tend to choose mature loss network with better performance for calculation. At present, the ready-made VGG networks [23] and residual networks are preferred for loss networks. The source infrared and visible images are scarce and the gray images contain limited information, so the shallow network is appropriate to calculate the loss and prevent the excessive extraction. So the VGG-16 is eventually selected. Moreover, another two issues have to be faced. Since we cannot get the ground-truth image for reference, how to simulate the inputs is important. The loss network is trained on colorful images, so we also need to make image concatenation to meet the requirements of three-channel form. Taking the two problems of channel correspondence into account at the same time, the definition of adjusted image is proposed to supplement the features of the source inputs. The image is targeted calculated by intensity and gradient weighting of infrared and visible images to adjust the lack of brightness, contrast and other information. Then, the infrared image, visible image and the adjusted image are organized as the reference. The three same fusion images are concatenated to simulate the generated image. The high-level features of the two simulated inputs are respectively extracted. Fig. 6 shows the diagram of the calculation process.

The perceptual loss can be formulated as equation (6).

$$L_{Per}(Y_1, Y_F) = \sum_{j=7,10,13} \frac{1}{C_j H_j W_j} \left\| \varphi_j(Y_I) - \varphi_j(Y_F) \right\|_2^2 \tag{6}$$

In the formula, φ means the VGG-16. Y_I represent the combined source images and the adjusted image. Y_F indicates the fused images with three channels. j is the serial number of convolution layer. $C_jH_jW_j$ is the total parameters in jth convolution layer. $\varphi_j(Y_I)$ and $\varphi_j(Y_F)$ denote the outputs of the jth layer and calculated with the L2 norm. The join of the L_{Per} greatly improve the characteristics of the fusion results on contrast and visual information fidelity.

3.9. Content loss

The content loss is mainly used in calculating the similarity of low-level features. For infrared and visible image fusion, not only the intensity, but also the gradient information is taken into account. The infrared radiation is strong and the visible textures are sufficient. So, the L_{Con} is organized as (7).

$$L_{Con} = \frac{1}{HW} \left(\|F - 1\|_F^2 + \|\nabla F - \nabla V\|_F^2 \right) \tag{7}$$

∇ indicates the gradient computation. HW means the size of the images. The Frobenius norm is applied to calculate the content loss.

4. Experimental results and analysis

4.1. Training and testing particulars

4.1.1. Training and test datasets

In reality, the infrared and visible image pairs accessible to the public are quite limited, so some researchers often make advance training on large datasets or make data augmentation. Since the fusion tasks are hugely dependent on the original information, data augmentation is adopted. 41 pairs of source images were collected from TNO [42] for training. Generally, the most direct and effective way for augmentation was image cropping. The minimum size of the source images and the computer capacity were together taken into consideration, so the cropping size 128 × 128 was adopted. Then the moving stride was set to 25 and the new dataset with 12,768 pairs of amplified image pairs was

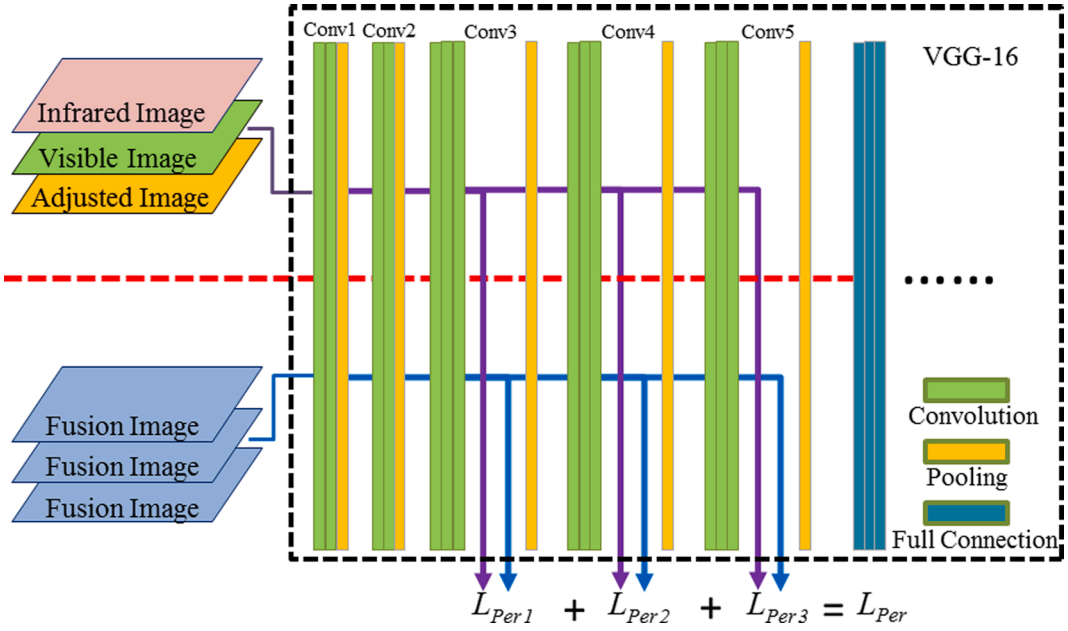


Fig. 6. The illustration of the proposed perceptual loss. L_{Per1}, L_{Per2}, and L_{Per3} respectively represents the perceptual loss calculated by the high-level features after the third, fourth and fifth convolution group.

established. The proposed network was trained with these small size pairs and the images were ergodic in every epoch. Notably, in order to show the generality of the proposed method, the test process was conducted on the TNO, RoadScene [1] and MFNet [3] datasets. It was an end-to-end module and the fusion results of the original source image pairs were obtained after each training epoch.

4.1.2. Detailed settings

Except above-mentioned cropping, other requirements and settings were introduced below. In order to enhance the training efficiency, the TITAN V GPU with Intel E5-2680 V3 processor was targeted. The way of mini-batch was used to guarantee the equilibrium of the input. The batch size and the learning rate were determined after multiple iterations, 32 and 10^{-5} were the final values. As to the optimizer, we chose the Adam which was a self-adaptive optimization algorithm with strong applicability and better convergence. The default parameters were used. In addition, the mixed loss function also involved some adjustable parameters existed in equation (4), (5). In the total loss, α, β, and θ were set as 160, 1, 1 after repeated experiments. In equation (5), w was 0.5 to balance the SSIM. During testing, in order to make the size of input and output consistent, the padding way of "SAME" was used when calculating the convolutions. But this would cause undesirable gray blocks around edges in fused images. To solve this problem, the source image pairs were padded in advance before put into the network. So, the obtained fused images with gray blocks were larger than original source images. Then we cut the gray blocks according to the padding value which was 6 in the proposed method. The value was closely related to the depth of the network.

4.2. Quality evaluation of the fused image

4 traditional fusion methods (CVT [43], DTCWT[44], LP [45], NSCT [46]) and 9 DL-based methods (CNN [20], Densefuse [21], Deepfuse [47], RCGAN [6], DDcGAN [27], U2Fusion [1], NestFuse [38], SeAFusion [3], CBAM [33]) were chosen together to evaluate the image quality more accurately. We tried our best to give an overall comparison on these methods. The traditional methods were implemented with the MATLAB toolbox and the DL-based models were carried out mainly referring to the program codes supplied by authors. For Deepfuse, we could not get the original codes and took the codes provided by [21]. As to CBAM, the original CBAM was embedded for feature extraction and the loss functions were consistent with the proposed method. Next, subjective and objective evaluation are synthesized to compare the fusion quality.

4.3. Subjective evaluation

As to subjective evaluation, the two source inputs and the images fused by these methods are arranged and marked for legible interpretation. Fig. 7, Fig. 8 and Fig. 9 are three typical groups of source images (a and b) and corresponding fused images (c to p) of *Marne_04*, *FLIR_04602* and *01023 N* from TNO, RoadScene and MFNet. In

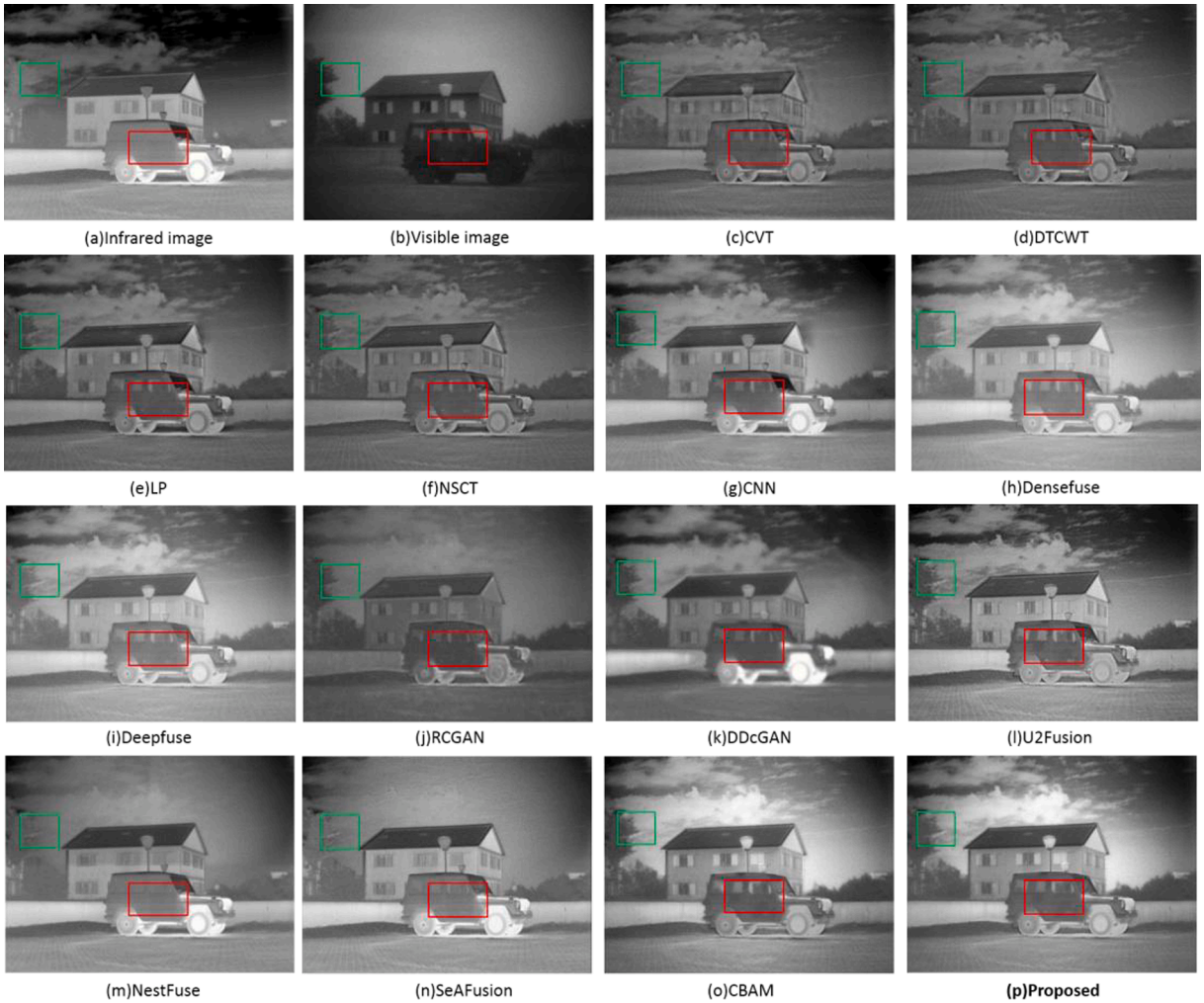


Fig. 7. Visual fused results of *Marne_04* image pair obtained by different fusion methods.

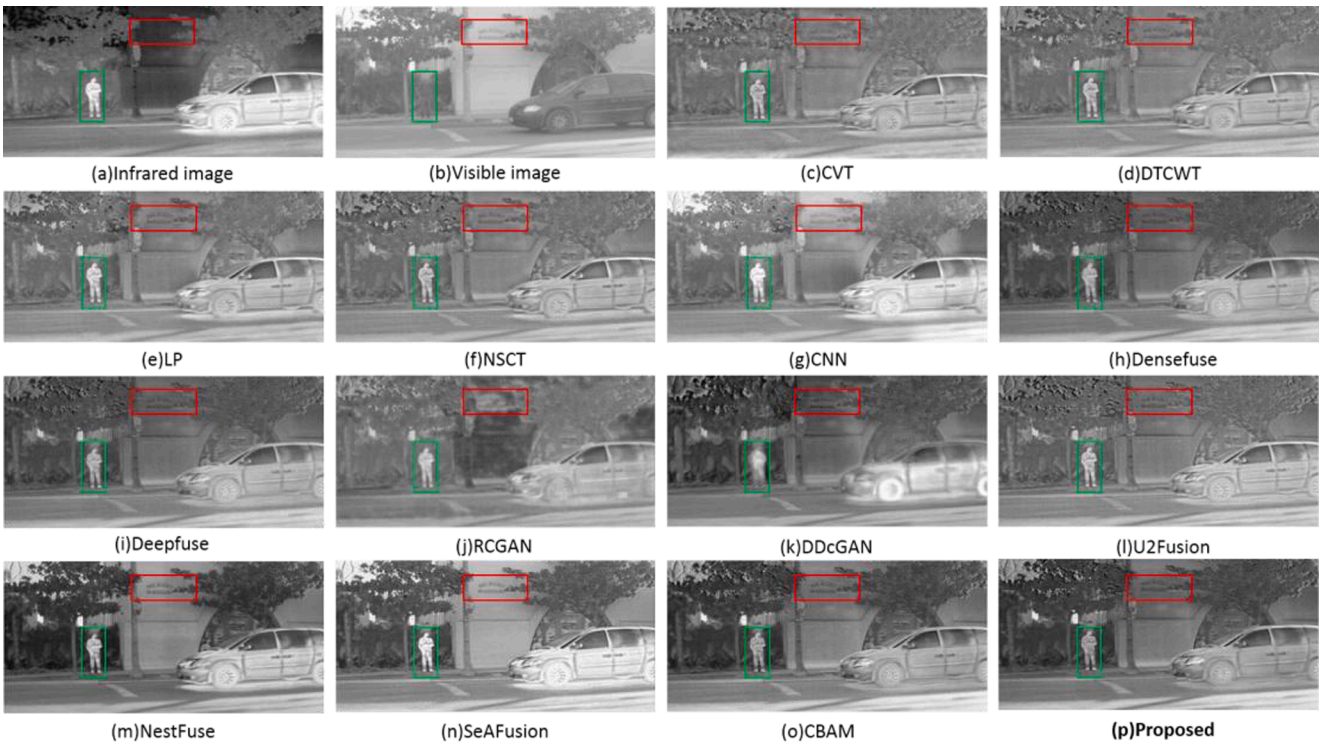


Fig. 8. Visual fused results of *FLIR_04602* image pair obtained by different fusion methods.

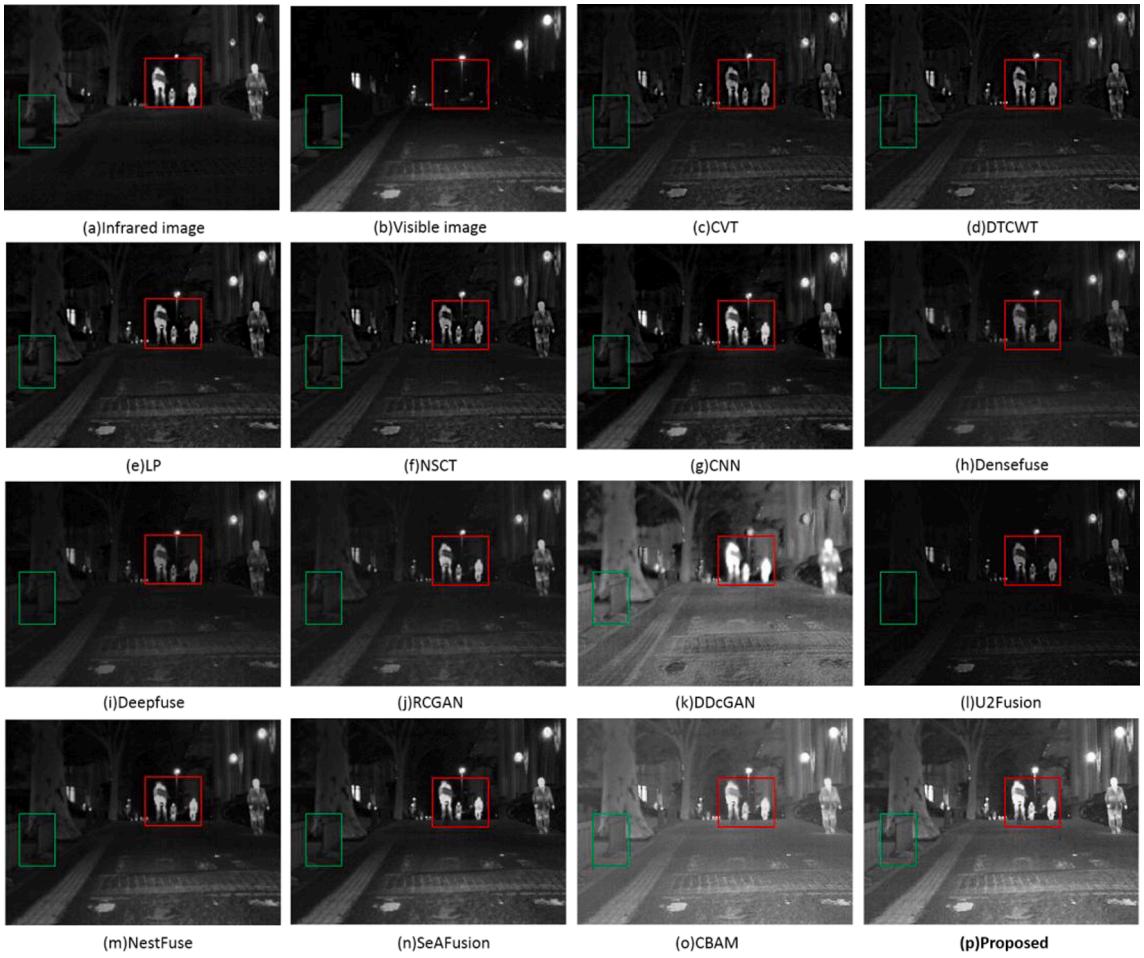


Fig. 9. Visual fused results of *01023 N* image pair obtained by different fusion methods.

Densefuse, there are two kinds of fusion strategies and the better fusion results are used for comparison. For purposeful observation, the red and green boxes on the images are advantageous to distinguish and recognize salient objects or features.

On the whole, the three pairs of the source images are all obviously distinguishable. The following four images acquired by traditional methods can only contain partial synthetical information of the source images. As a result, the brightness and the contrast are low leading to blurred visual effects. The fused images of CNN always have higher contrast and the targets can be easily recognized. Nevertheless, some undesirable and aggravated areas appear on these images, such as the roof in Fig. 7. From visual view, the Densefuse and Deepfuse methods both can retain visible details and infrared objects from the multi-modal inputs, and the fusion results are fine. But the contrast and the definition are slightly defective. Also, the fused images of RCGAN are not clear enough and respective specialties of the inputs are not highlighted. The fusion results of DDcGAN tend to reflect more infrared intensity and the objects are distinct. But the texture details in visible images are somehow neglected. The outputs of U2Fusion are relatively clear, but specific contents in few visible images are not reflected. The fusion results of NestFuse and SeAFusion contain comprehensive information, only the visual effect of several pairs like the one in Fig. 7 need to be improved.

With regard to CBAM and the proposed method, since the convolutional block attention module is introduced, the models can simultaneously lay more emphasis on obvious objects and details information of the source inputs. So the fusion results are always sufficient with prominent features. The images look clearer with less noises. Both in contrast and fidelity, the fused images behave quite well. Most notably, in our method, the multi-branch architecture is adopted and the different perceptive fields in spatial attention help a lot to get complementary feature maps. Skip connections are also added to learn residual information which are important for fusion tasks. If we enlarge the above three groups of pictures, corresponding images fused by our attention model have great advantages.

4.3.1. Evaluation metrics

For image fusion tasks, there are always no reference images. Different fusion purposes focus on different image characteristics and the standards for image quality evaluation are different. At the moment, most researchers adopt the way to use multiple objective indicators for multi-angle evaluation [48]. By drawing on their achievements, we have established a system for comprehensive evaluation by measuring the single generated image itself and the correlation between output and source inputs. The entropy (EN) [49] measures the amount of information of the fusion image. Spatial frequency (SF) [50] reflects the image definition. The standard deviation (SD) [51] helps to calculate the contrast. The following metrics focus on correlation between fusion images and the input images. The mean structural similarity (MSSIM) [52] can compare the overall structural information in detail. Correlation coefficient (CC) [53] and the sum of the correlation of differences (SCD) [54] can count the linear correlation degree between images from two angles. The last metric called visual information fidelity for fusion (VIFF) [55] is now widely applied. The metric is built up on image distortion and human visual distortion, so it can be used to measure the visual perception.

The larger value means the better fusion result is suitable for all the seven metrics. They are encoded through MATLAB and the codes programmed by the original researchers are used for reference.

4.3.2. Objective evaluation

20 (10 of TNO, 5 of RoadScene, 5 of MFNet) infrared and visible pairs are selected to generated images for test. Average values of the indexes are calculated and the numerical results are listed in Table 1. The first four lines of the results are the quantitative values of traditional methods. The others are results of DL-based methods. The best values of different methods are in bold.

Among these comparative methods chosen, we can see that the numerical behaviors of the DL-based methods are generally better than traditional methods. This also indirectly proves the superiorities of deep neural networks. In these DL-based methods, there are no additional activity level measurements and the fusion rules are simplified. The whole transformation process is implemented by designing the network structure and restraining the loss. The features are extracted efficiently by repetitive training.

Next, numerical results of different methods are analyzed one by one. The proposed method behaves preferably when the average values are taken together. We get the best values on three metrics which are EN, CC and VIFF. The bigger EN means that the images are informative with details. The bigger CC indicates the correlation between the inputs and fusion image is close and more important features are transferred. Furthermore, the VIFF of this proposed model is very prominent. The fused images with higher fidelity are conductive to be observed by human visual system. The SF and SCD metrics of the proposed method are just inferior to the values of SeAFusion method. With regard to the SeAFusion method, gradient residual dense blocks are embedded to extract fine-grained detail information of feature maps and they creatively devise a semantic loss to adequately boost the semantic information of fused images. All these operations help to improves the image quality of SeAFusion method. The method also gets the biggest value on SD. The corresponding results of CNN, NestFuse and the proposed method are slightly smaller. As to MSSIM, the results of the CBAM and the proposed method are not desired. The Densefuse and Deepfuse methods achieve better values on it. In Densefuse, the densely connected structure is innovatively introduced which is useful to retain information of middle layers. Moreover, the encoding and decoding network are well adjusted by pre-training. For Deepfuse, the better performance is mainly attributed to its unique MEF SSIM. All these excellent designs of the above methods can be used for reference in our follow-up studies.

Table 1
The Numerical Values of Seven Metrics.

Methods	*EN*	*SF*	*SD*	*MSSIM*	*CC*	*SCD*	*VIFF*
CVT	6.7105	12.0074	30.3249	0.5506	0.5211	1.5998	0.3985
DTCWT	6.6530	11.9894	29.6035	0.5664	0.5269	1.6030	0.3929
LP	6.7876	12.3804	33.0202	0.5752	0.5229	1.6148	0.4894
NSCT	6.6762	12.1322	30.1654	0.5834	0.5314	1.6196	0.4441
CNN	7.1198	12.2810	45.9716	0.5698	0.4923	1.6215	0.5263
Densefuse	6.8092	9.4475	35.4260	**0.6049**	0.5507	1.6466	0.4950
Deepfuse	6.8257	9.4367	35.7491	0.6044	0.5506	1.6378	0.5019
RCGAN	6.5885	8.2742	29.7809	0.5634	0.4890	1.5175	0.3213
DDcGAN	7.2039	10.1917	41.1310	0.4470	0.4714	1.3316	0.3742
U2Fusion	6.7259	12.1825	35.0473	0.5755	0.5342	1.6349	0.5415
NestFuse	7.0422	11.2979	45.2945	0.5886	0.5096	1.5921	0.4435
SeAFusion	7.1253	**13.6405**	**46.9932**	0.5777	0.5142	**1.7011**	0.5274
CBAM	7.1201	12.3843	40.7179	0.5302	0.5564	1.6481	0.5934
Proposed	**7.2783**	13.5923	44.3380	0.5336	**0.5594**	1.6709	**0.6845**

Moreover, we will continue to optimize the composition and specific form of the loss function.

Through the numerical analyses, it can be summarized that the evaluation results are coincident whether on objective calculation or subjective judgments. For convenient comparison of the comprehensive performances of different methods on these metrics, we draw a radial map to reflect the numerical resultsas shown in Fig. 10.

4.4. Ablation study

For the sake of proving the originality of the designed method, ablation experiments are done to test the effects of the attention mechanism introduced, also check the influence of the perceptual loss.

4.4.1. Analysis of multi-scale attention module

The loss function is consistent with the proposed method when designing the two comparative experiments. The first network without attention module is consisted of eight layers of convolutional network. Five layers are used for encoding and three layers are used for decoding. The second network has the same structure with the proposed method whose perceptive fields are in the same size of 3×3 in spatial attention.

In Fig. 11, the images generated with attention module look clearer and more attractive. Whether the infrared objects or texture details are easy to identify and recognize. Especially, in the proposed method, different sizes of kernel are adopted in the three branches for convolution calculation in the spatial attention. This helps to acquire multi-scale salient features from the distinct inputs which are advantageous to construct better outputs. The multi-scale attention network achieves excellent effects on this multi-modal image fusion task. Hot maps are drawn to clearly show which area in the image is focused by the network.

4.4.2. Analysis of multi-layer perceptual loss

In this group of contrast experiments, the attention networks are consistent and we make differences on perceptual loss. One network is trained without perceptual loss while another one is trained with only L_{Per1}. SSIM loss and content loss are reserved.

In Fig. 12, the fused images gained by the attention network without perceptual loss are fuzzy. Meanwhile, a certain amount of information is missing. It is evident that the images obtained by the network constrained with perceptual loss are informative. The visual effect is also improved since less noise and useless information are brought. We can see that the attention network only with L_{Per1} is capable of achieving preferable fused images. In the proposed method, further promotion is made by combining the L_{Per1}, L_{Per2}, and L_{Per3} to construct a multi-layer perceptual loss. This makes the image quality be improved again. The targets and the outlines of the images become clearer. In fact, other combinations have also been tried, the selected mode has the best effect. Hot maps are also drawn at the end.

5. Conclusion

The paper focuses on the application of attention mechanism in infrared and visible image fusion. An unsupervised CNN model is built and mixed loss functions are constructed. As to the attention structure, channel attention and spatial attention are put into use together. Especially, different perceptive fields are adopted in spatial attention of each branch. This will help to obtain multi-scale targets or textures in source

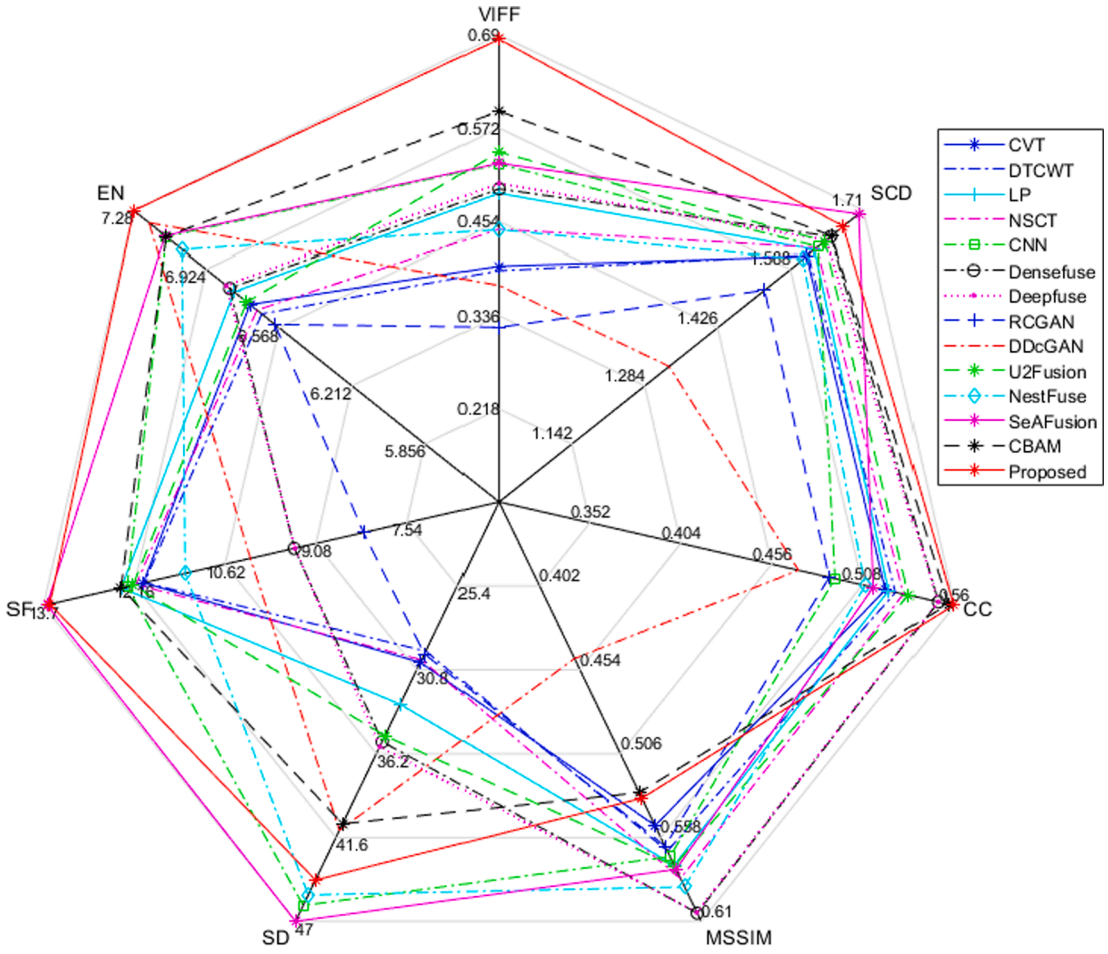


Fig. 10. The radial map of numerical results on different metrics.

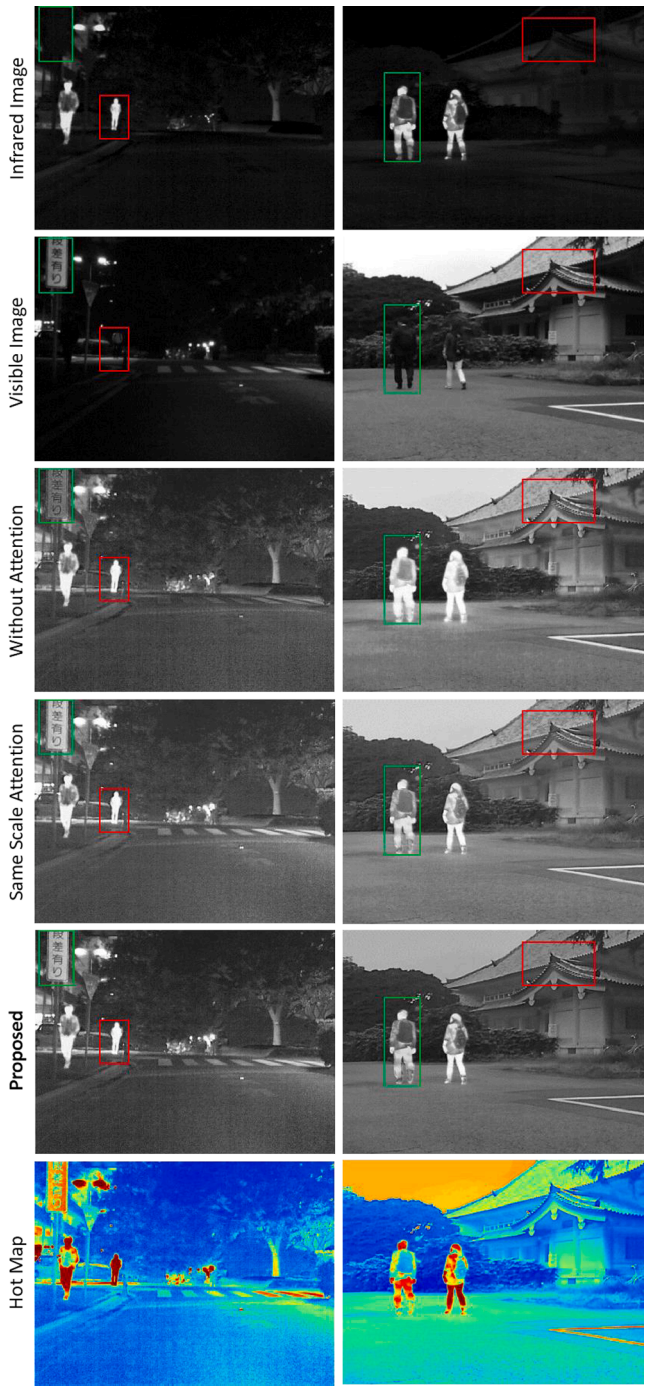


Fig. 11. The fused results of ablation experiments on attention module. From top to bottom: infrared image, visible image, fusion results of conventional convolutional neural network without attention module, attention network with same perceptive fields in spatial attention and the proposed method, the hot maps of the proposed method.

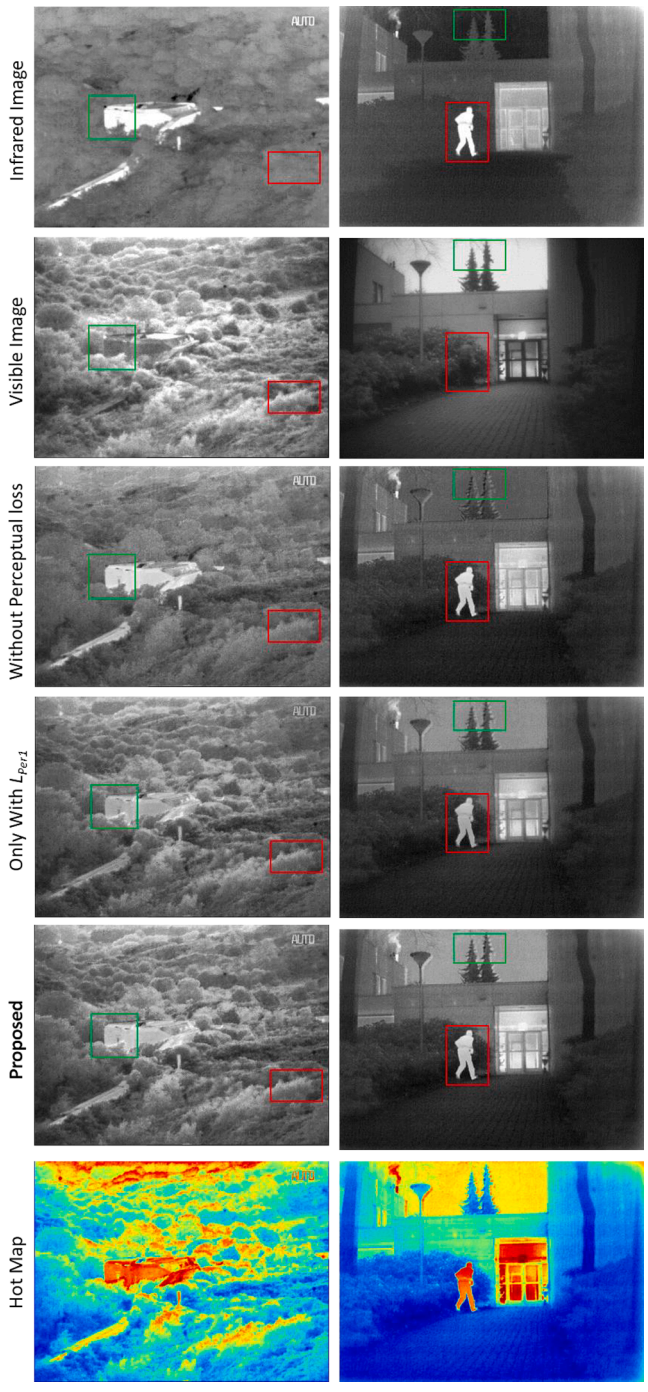


Fig. 12. The fused results of ablation experiments on perceptual loss. From top to bottom: infrared image, visible image, fusion results of attention network without perceptual loss, attention network with only L_{Per1} and the proposed method, the hot maps of the proposed method.

images. Moreover, infrared objects and visible details are extracted and preserved effectively. For better constraint of the back propagation, mixed loss functions are adopted. Especially, multi-layer perceptual loss is proposed so that the similarity of high-level features can be restrained. All these designs have got obvious quality improvements on fusion images. The image information is richer and the visual effect gets better. Both subjective observations and objective calculations show that the proposed method is feasible and has superiorities.

The framework and the loss of the proposed method have strong universality. With appropriate modifications, the model will be used for other multi-modal or multi-focus fusion tasks.

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] 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.

[2] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: A survey of the state of the art, Inf. Fusion 33 (Jan. 2017) 100–112.

[3] 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 (Jan. 2022) 28–42.

[4] H. Zhang, H. Xu, X. Tian, J. Jiang, J. Ma, Image fusion meets deep learning: A survey and perspective, Inf. Fus. 76 (2021) 323–336.

[5] H. Zhang, J. Ma, SDNet: A Versatile Squeeze-and-Decomposition Network for Real-Time Image Fusion, Int. J. Comput. Vis. 129 (2021) 2761–2785.

[6] Q. Li, L. Lu, Z. Li, W. Wu, X. Yang, Coupled gan with relativistic discriminators for infrared and visible images fusion, IEEE Sens. J. (2019).

[7] S. Li, B. Yang, J. Hu, Performance comparison of different multiresolution transforms for image fusion, Inf. Fusion 12 (2) (Apr. 2011) 74–84.

[8] G. Pajares, J. Manuel de la Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (Sep. 2004) 1855–1872.

[9] Z. Zhang, R.S. Blum, A categorization of multiscale-decomposition based image fusion schemes with a performance study for a digital camera application, Proc. IEEE 87 (8) (Aug. 1999) 1315–1326.

[10] J. Wang, J. Peng, X. Feng, G. He, J. Fan, Fusion method for infrared and visible images by using non-negative sparse representation, Infr. Phys. Technol. 67 (Nov. 2014) 477–489.

[11] S. Li, H. Yin, L. Fang, Group-sparse representation with dictionary learning for medical image denoising and fusion, IEEE Trans. Biomed. Eng. 59 (12) (Dec. 2012) 3450–3459.

[12] T.M. Tu, S.C. Su, H.C. Shyu, P.S. Huang, A new look at ihs-like image fusion methods, Inf. Fusion 2 (3) (2001) 177–186.

[13] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, T. Wittman, An adaptive ihs pan-sharpening method, IEEE Geosci. Remote Sens. Lett. 7 (4) (2010) 746–750.

[14] R. Eckhorn, H.J. Reitbock, M. Arndt, P. Dicke, A neural network for feature linking via synchronous activity: Results from cat visual cortex and from simulations, Can. J. Microbiol. 46 (8) (1989) 759–763.

[15] Z. Wang, C. Gong, A multi-faceted adaptive image fusion algorithm using a multi-wavelet-based matching measure in the PCNN domain, Appl. Soft Comput. 61 (Dec. 2017) 1113–1124.

[16] Y. Lin, S. Le, Z. Xin, and Y. Huang, "Infrared and visible image fusion algorithm based on contourlet transform and PCNN," in *Proc. SPIE, Infr Mater., Devices, Appl.*, vol. 6835, 2008.

[17] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (Jul. 2015) 147–164.

[18] J. Ma, Z. Zhou, B. Wang, H. Zong, Infrared and visible image fusion based on visual saliency map and weighted least square optimization, Infr. Phys. Technol. 82 (May 2017) 8–17.

[19] X. Huang, G. Qi, H. Wei, Y. Chai, and J. Sim, "A novel infrared and visible image Inf. Fusion method based on phase congruency and image entropy," *Entropy*, vol. 21, no. 12, p. 1135, Nov. 2019.

[20] Y.u. Liu, X. Chen, J. Cheng, H.u. Peng, Z. Wang, Infrared and visible image fusion with convolutional neural networks, Int. J. Wavelets Multiresolut Inf. Process. 16 (03) (2018) 1850018.

[21] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May. 2019.

[22] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.

[23] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*. [Online]. Available: http://arxiv.org/abs/1409.1556.

[24] He, K.; Zhang, X.; Ren, S.; Jian, S. Deep Residual Learning for Image Recognition. *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, 27–30 Jun 2016; pp. 770–778.

[25] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, FusionGAN: A generative adversarial network for infrared and visible image fusion, Inf. Fusion 48 (Aug. 2019) 11–26.

[26] J. Ma, P. Liang, Y. Wei, C. Chen, X. Guo, J. Wu, J. Jiang, Infrared and visible image fusion via detail preserving adversarial learning, Inf. Fusion 54 (2020) 85–98.

[27] J. Ma, H. Xu, J. Jiang, X. Mei, X.P. Zhang, DDcGAN: A dual-discriminator conditional generative adversarial network for multiresolution image fusion, IEEE Trans. Image Process. 29 (2020) 4980–4995.

[28] J. Li, H. Huo, K. Liu, C. Li, Infrared and visible image fusion using dual discriminators generative adversarial networks with Wasserstein distance, Inf. Sci. 529 (2020) 28–41.

[29] J. Li, H. Huo, C. Li, R. Wang, C. Sui, Z. Liu, Multi-grained attention network for infrared and visible image fusion, IEEE Trans. Instrum. Meas. 70 (2021) 1–12.

[30] J. Li, H. Huo, C. Li, R. Wang, Q.i. Feng, Attentionfgan: Infrared and visible image fusion using attention-based generative adversarial networks, IEEE Trans. Multimedia 23 (2021) 1383–1396.

[31] H.T. Mustafa, J. Yang, H. Mustafa, M. Zareapoor, Infrared and visible image fusion based on dilated residual attention network, Optik- Int. J. Light Electron. Opt. 224 (9) (2020), 165409.

[32] Y. Li, J. Wang, Z. Miao, J. Wang, Unsupervised densely attention network for infrared and visible image fusion, Multimed. Tools Appl. 79 (45-46) (2020) 34685–34696.

[33] S. Woo, J. Park, J.Y. Lee, I.S. Kweon, CBAM: Convolutional block attention module, in: in *European Conference on Computer Vision (ECCV)*, 2018, pp. 3–19.

[34] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[35] S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, "Aggregated residual transformations for deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR))*, 2017.

[36] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. "Neural machine translation by jointly learning to align and translate," arXiv preprint arXiv: 1409.0473, 2014.

[37] S. Chaudhari, G. Polatkan, R. Ramanath, and V. Mithal, "An attentive survey of attention models," 2019.

[38] 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) (Dec. 2020) 9645–9656.

[39] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016.

[41] H. Xu, J. Ma, Z. Shao, H. Zhang, J. Jiang, X. Guo, SDPNet: A Deep Network for Pan-Sharpening With Enhanced Information Representation, IEEE Trans. Geosci. Remote. Sens. 59 (5) (May 2021) 4120–4134.

[42] T. Alexander, "TNO image fusion dataset," 2014.

[43] F. Nencini, A. Garzelli, S. Baronti, L. Alparone, Remote sensing image fusion using the curvelet transform, Inf. Fusion 8 (2) (2007) 143–156.

[44] J.J. Lewis, R. O'Callaghan, S.G. Nikolov, D.R. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inf. Fusion 8 (2) (2007) 119–130.

[45] P.J. Burt, E.H. Adelson, The laplacian pyramid as a compact image code, Readings Comput. Vis. 31 (4) (1987) 671–679.

[46] Q. Zhang, B. L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *SIGNAL PROCESSING -AMSTERDAM-*, 2009.

[47] K. R. Prabhakar, V. S. Srikar, and R. V. Babu, "DeepFuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, p. 3.

[48] G. Schwan and N. Scherer-Negenborn, "An approach to select the appropriate image fusion algorithm for night vision systems," *Proc. SPIE,* vol. 9649, Oct. 2015, Art. no. 964908.

[49] J. Van Aardt, Assessment of image fusion procedures using entropy, image quality, and multispectral classification, J. Appl. Remote Sens 2 (1) (2008) 023522.

[50] A.M. Eskicioglu, P.S. Fisher, Image quality measures and their performance, IEEE Trans. Commun. 43 (12) (1995) 2959–2965.

[51] Y.-J. Rao, In-fibre Bragg grating sensors, Meas. Sci. Technol. 8 (4) (1997) 355–375.

[52] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: From error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (Apr. 2004) 600–612.

[53] M. Deshmukh, U. Bhosale, Image fusion and image quality assessment of fused images, Int. J. Image Process. 4 (5) (2010) 484–508.

[54] V. Aslantas, E. Bendes, A new image quality metric for image fusion: The sum of the correlations of differences, AEU-Int. J. Electron. Commun. 69 (12) (2015) 1890–1896.

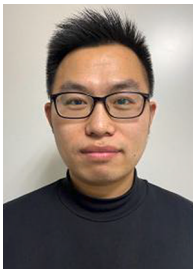
[55] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (Apr. 2013) 127–135.

DONGDONG XU received his bachelor's degree from Shandong University in 2013 and master's degree from Harbin Institute of Technology in 2015. He received his Ph. D. degree from Chinese Academy of Sciences in 2020. Now, he is a research assistant at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include deep learning, image fusion and embedded system.

NING ZHANG received her bachelor's degree from Northeastern University, Qinhuangdao, China, in 2017. She is currently a PhD student at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. Her research interests cover image processing, deep learning and remote sensing image super-resolution.

YUXI ZHANG received his bachelor's degree from Harbin Institute of Technology, Weihai, China, in 2018. He is a PhD student at Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include image processing, deep learning, and object detection on remote sensing images.

ZHENG LI received his bachelor's degree from Changchun University of Science and Technology, Changchun, China, in 2020. He is pursuing the master degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include image processing, remote sensing image object detection and deep learning.



ZHIKANG ZHAO received his bachelor's degree from Ocean University of China, Qingdao, China, in 2019. He is pursuing the master degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include image processing, deep learning and remote sensing image super-resolution.

YONGCHENG WANG received his bachelor's degree from Jilin University in 2003 and Ph. D. degree from Chinese Academy of Sciences in 2010. He is a researcher of Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. His research interests include artificial intelligence, image engineering, and embedded system of space payload.