A Multilevel Hybrid Transmission Network for Infrared and Visible Image Fusion

Qingqing Li, Guangliang Han, Peixun Liu, Hang Yang, Dianbing Chen, Xinglong Sun, Jiajia Wu, and Dongxu Liu

***Abstract*— Infrared and visible image fusion aims to generate an image with prominent target information and abundant texture details. Most existing methods generally rely on manually designing complex fusion rules to realize image fusion. Some deep learning fusion networks tend to ignore the correlation between different level features, which may cause loss of intensity information and texture details in the fused image. To overcome these drawbacks, we propose a multilevel hybrid transmission network for infrared and visible image fusion, which mainly contains the multilevel residual encoder module (MREM) and the hybrid transmission decoder module (HTDM). Considering the great difference between infrared and visible images, the MREM with two independent branches is designed to extract abundant features from source images. To avoid complicated fusion strategies, the concatenate convolution is applied to fuse features. Toward utilizing information from source images efficiently, the HTDM is constructed to integrate different level features. Experimental results and analyses on three public datasets demonstrate that our method not only can achieve high-quality image fusion, but also performs better than comparison methods in terms of qualitative and quantitative comparisons. In addition, the proposed method has good real-time performance in infrared and visible image fusion.**

***Index Terms*— Decoder module, encoder module, hybrid transmission, image fusion, multilevel.**

I. Introduction

IMAGE fusion is a significant technology in the field of computer vision, which can reduce the difficulty of image analysis and understanding by integrating images from different sensors into one image. Thus, it is widely applied in military, remote sensing, and surveillance [1]–[3].

With the development of sensors, infrared and visible image fusion has become a hot topic due to their strong information complementarity. Infrared sensors can capture thermal radiations emitted by objects so that it is able to distinguish targets from the background and is immune to variations of illumination and weather. However, infrared images have low spatial resolution and lack detailed textures. Visible images have the advantages of rich texture information and high resolution, but they are susceptible to weather and illumination [4]. Therefore, scholars focus on integrating infrared and visible images into a high-quality image with salient target information and rich texture details to provide sufficient information for computer vision tasks, such as object detection, recognition, and tracking [5].

Various image fusion methods have been developed in recent years, which can be divided into three dominant categories: multiscale transform-based methods, representation learning-based methods, and deep learning-based methods [6], [7].

The multiscale transform-based methods usually transform images into frequency domain to obtain different scale features. These features are combined through suitable rules to generate the fused image [8]. The dual-tree complex wavelet transform is a typical multiscale transform-based image fusion method, which can resolve the shift variance and the lack of directionality problems [9], [10]. For extracting plentiful direction information of source images, contourlet transform is proposed [11]. Non-subsampled contourlet transform is an improved form of contourlet transform, which has flexibility and shift invariance [12], [13]. In order to describe the structure information of the image preferably, shearlet transform is raised [14]. Non-subsampled shearlet transform is a modified method based on shearlet transform, which can achieve more precise directional decomposition [15]. The above image fusion methods based on multiscale transform need to transform images to the frequency domain, which may result in unrecoverable loss of information during the transformation process [16].

Different from multiscale transform-based methods, representation learning-based methods fuse images without transformation. Sparse representation is utilized to extract common and complementary information from source images to produce the high-quality fused image [17]. For solving the problem that sparse representation-based methods are sensitive to misregistration, Liu *et al.* [18] propose the convolutional sparse representation to fuse multimodal images. In recent years, image fusion methods based on the latent low-rank representation have attracted lots of attentions [19]–[21].

Manuscript received 9 April 2022; revised 22 May 2022; accepted 9 June 2022. Date of publication 30 June 2022; date of current version 21 July 2022. This work was supported in part by the Department of Science and Technology of Jilin Province under Grant 20210201132GX. The Associate Editor coordinating the review process was Mohamad Forouzanfar. *(Corresponding author: Guangliang Han.)*

Qingqing Li, Jiajia Wu, and Dongxu Liu are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China, and also with the University of Chinese Academy of Sciences, Beijing 100049, China (e-mail: liciomp@163.com; wujiajia17@mails.ucas.ac.cn; liudongxu18@mails.ucas.ac.cn).

Guangliang Han, Peixun Liu, Hang Yang, Dianbing Chen, and Xinglong Sun are with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China (e-mail: hangl@ciomp.ac.cn; liupx@ciomp.ac.cn; yanghang@ciomp.ac.cn; chendianbing1934@163.com; sunxinglong@ciomp.ac.cn).

Digital Object Identifier 10.1109/TIM.2022.3186048

These methods decompose input images into base and detail layers, which is conducive to design fusion strategies. Nevertheless, fusion methods based on representation learning are usually complex, which leads to the degradation of fusion performance and increase of running time [16]. In addition, both multiscale transform-based and representation learning-based methods require to manually design complicated activity level measurements and fusion rules [22].

In the past few years, deep learning-based methods have released huge potential in infrared and visible image fusion task [23]. The convolutional neural networks (CNNs) can extract multilevel deep features containing rich information, which is beneficial for image fusion. Li *et al.* [24] employ CNNs to effectively extract and integrate the features of infrared and visible images to finish image fusion. To get more useful features from source images, the dense block and nest connections are introduced to the fusion network [25]. Although these methods can generate satisfactory fusion images, they still rely on establishing traditional fusion rules (such as average, addition, l_1-norm, and attention-based) [16], [25], [26]. To avoid designing fusion strategies, the generative adversarial network is applied to image fusion. [27]. For example, Ma *et al.* [22] propose the GAN-based fusion method (FusionGAN) network to realize infrared and visible image fusion. Then, to improve the fusion performance of FusionGAN, Ma *et al.* [28] modify the loss function. However, these methods usually ignore the correlation between different level features and preserve insufficient detail information of source images [29].

To tackle these drawbacks above, we propose a multilevel hybrid transmission network for infrared and visible image fusion (MHTNet). The main contributions of this article are summarized as follows.

1) Considering the great difference between infrared and visible images, this article designs two independent branches containing a series of residual encoder blocks (REBs) to extract sufficient features of different levels from infrared and visible images, respectively.
2) Different from some fusion networks, the complex traditional fusion rule is replaced by the concatenate convolution (CC) to fuse infrared and visible features at the same level.
3) A hybrid transmission decoder module (HTDM) is proposed to improve the fusion performance by using features from encoder module adequately, which includes cross transmission and hybrid transmission. The former aims to make the information at different levels complement each other, and the latter is on the purpose of compensating for information loss in the decoding stage.
4) Extensive experiments are carried out on three public datasets. Experimental results demonstrate that the proposed network can comprehensively enhance the quality of infrared and visible fusion image and transcends comparison methods in terms of qualitative and quantitative analyses.

The rest of this article is organized as follows. In Section II, the related work on residual block is introduced. In Section III, the proposed infrared and visible image fusion network is described in detail. In Section IV, information about the experimental datasets and settings is given. In Section V, the self-comparison experiments, evaluation of experimental results, and the analyses of running time are discussed. In Section VI, conclusions are presented.

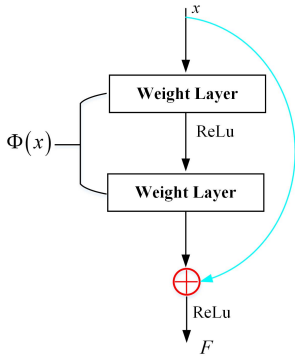


Fig. 1. Structural schematic of residual block.

II. Related Work

With the increase of network depth, the training accuracy will reach a saturation state and then deteriorate rapidly. He *et al.* [30] design a residual structure to address the above problem. The architecture of residual block is given in Fig. 1, and its mathematical representation is expressed as follows:

$$F(x) = \Phi(x) + x \tag{1}$$

where x and $F(x)$ represent the input and output of the residual block, respectively, and $\Phi(x)$ indicates the network operation which contains two weight layers.

As shown in Fig. 1, the multilayer information is utilized effectively through the "short-connection," which is beneficial for compensating the feature loss. Thus, residual block is widely used in image fusion. For example, Jian *et al.* [31] design a symmetric encoder–decoder infrared and visible image fusion network, which applies residual connection at the last layer of encoder to compensate for information loss. Li *et al.* [32] propose an infrared and visible image fusion method based on ResNet50 to fully utilize deep features. Mustafa *et al.* [33] present an end-to-end fusion network, which improves the quality of infrared and visible image fusion by residual attention module.

In view of the advantages of residual block and its good performance in the field of image fusion, the residual structure is also employed in our proposed fusion network.

III. Proposed Infrared and Visible Image Fusion Method

A. Overall Framework

This article proposes an MHTNet. The detailed structure and settings of MHTNet are provided in Fig. 2 and Table I. As shown in Fig. 2, MHTNet can be divided into four steps: initializing, feature encoding, feature decoding, and outputting the fused image.

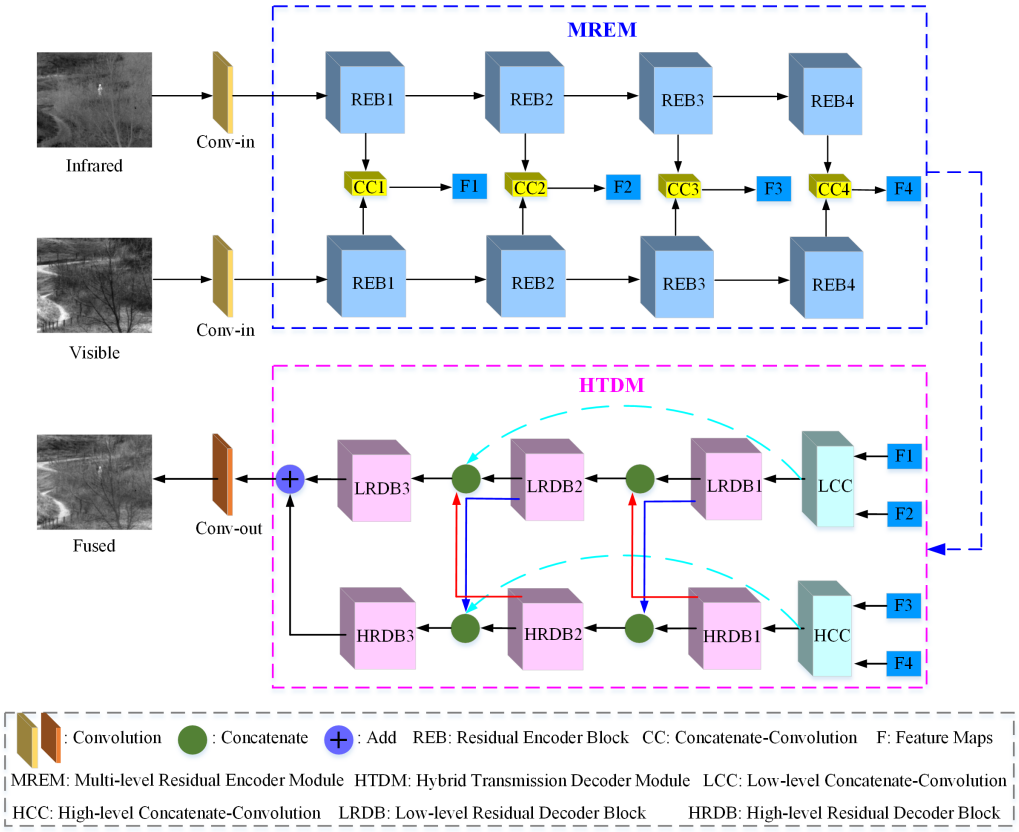


Fig. 2. Architecture of the proposed MHTNet.

TABLE I

DETAILED SETTINGS OF THE PROPOSED MHTNET, INCLUDING THE CONVOLUTIONAL KERNEL SIZE (SIZE), STRIDE, INPUT CHANNEL, AND OUTPUT CHANNEL OF EACH LAYER. NIN AND NC REPRESENT THE NUMBER OF CATEGORIES OF FEATURES THAT ARE ENTERED INTO CC AND RDB, RESPECTIVELY

Part	Layer	Size	Stride	Input channel	Output channel
	Conv-in	3	1	1	16
	Conv-out	3	1	16	1
MREM	REB1	-	-	16	16
	REB2	-	-	16	16
	REB3	-	-	16	16
	REB4	-	-	16	16
	CC1	-	-	32	16
	CC2	-	-	32	16
	CC3	-	-	32	16
	CC4	-	-	32	16
HTDM	LCC	-	-	32	16
	LRDB1	-	-	16	16
	LRDB2	-	-	32	16
	LRDB3	-	-	48	16
	HCC	-	-	32	16
	HRDB1	-	-	16	16
	HRDB2	-	-	32	16
	HRDB3	-	-	48	16
CC	conv	5	1	16×Nin	16
REB	conv	3	1	16	16
	conv	3	1	16	16
RDB	conv	3	1	16×Nc	16
	conv	3	1	16	16

1) Initializing: The infrared and visible images are, respectively, sent to the Conv-in layer to obtain the initialized feature maps.

2) Feature Encoding: The initialized feature maps of infrared and visible images are transferred to the multilevel residual encoder module (MREM) to extract different level features. Considering the great difference between infrared and visible images, MREM is consisted of two independent branches so as to obtain appropriate network model parameters and effective features. Each branch contains a series of REBs, which ensures the richness of feature information. In addition, in the encoding stage, in order to avoid the complicated fusion rules and decrease computation of the fusion network, infrared and visible features at the same level are fused by the CC.

3) Feature Decoding: Features from MREM are sent to the HTDM to communicate adequately. Different level features generally have certain differences; thus, in the decoding stage, features from MREM are first divided into low level and high level and encoded by two branches. The network structures of the two branches are identical, which ensures that the dimensions of features from the corresponding module of different branches are the same and provides convenience for the transmission of features between different levels. Moreover, HTDM aims to transfer features of different levels to the fused image effectively, which includes two specific transmission ways: cross transmission and skip transmission. The purpose of cross transmission is to realize information complementation of different level features, while the purpose of skip transmission is to compensate for information loss in decoding process.

4) Outputting the Fused Image: Features from HTDM are sent to the Conv-out layer to generate the fused image.

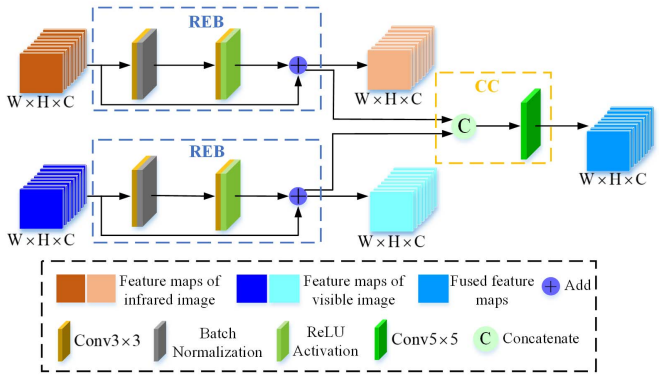


Fig. 3. Schematic of REB and CC.

In the above steps, MREM and HTDM are two pivotal components of MHTNet. These two parts are introduced in detail in the following.

B. Multilevel Residual Encoder Module

With the development of deep learning networks, a progressive structure named residual network is proposed to solve the degradation problem and increase training speed [30]. Residual network is easier to optimize and can provide satisfactory accuracy by increasing the network depth [32]. On account of its great advantages in training, residual network has been widely used in various visual tasks, such as hyperspectral classification, saliency detection, image segmentation, and target detection [34]–[37]. Furthermore, the residual network can adequately utilize multilevel information to extract features effectively and strengthen feature propagation ability [38]. Inspired by the above advantages of residual network, we raise an MREM to capture richer feature information of infrared and visible images.

As shown in Fig. 2, infrared and visible images are first initialized by Conv-in to derive two sets of feature maps with the size $256 \times 256 \times 16$; then, these feature maps are sent to MREM. MREM is consisted of two independent residual encoder branches and a series of CCs. Each branch of MREM deals with features from infrared and visible images, respectively, which contains four REBs. REBs obtain information at different levels from source images. In the process of encoding, the features of infrared and visible images at the same level are fused through CC. The schematic of REB and CC is given in Fig. 3. W, H, and C denote the width, height, and amount of channel of feature maps, respectively, where $W = H = 256$ and $C = 16$. Combining Figs. 2 and 3, it can be seen that feature maps produced by REB have two applications, one is be sent to the next REB and the other is to be sent to CC for fusion. REB and CC are introduced as follows.

1) *REB:* REB aims to extract feature maps with rich information through the residual connection. As shown in Fig. 3, REB consists of two 3×3 convolution layers, a batch normalization layer, a ReLU activation layer, and an add operation. Add operation is used to compensate for information loss in network computing, which can be written as follows:

$$\mathrm{FM}_{\mathrm{out}}(i) = \mathrm{FM}_{\mathrm{pre}}(i) + \mathrm{FM}_{\mathrm{cur}}(i), \quad i = 1, 2, 3, \ldots, C \tag{2}$$

where $\mathrm{FM}_{\mathrm{pre}}$, $\mathrm{FM}_{\mathrm{cur}}$, $\mathrm{FM}_{\mathrm{out}}$ are the previous layer feature maps of REB, the current layer feature maps generated by REB before adding operation, and the output feature maps of REB, respectively, and i represents the sequence number of feature map channels.

2) *CC:* The function of CC is to fuse infrared and visible encoding features at the same level. As shown in Fig. 3, CC is composed of a 5×5 convolution and a concatenate operation. The concatenate operation is used to arrange feature maps from the same level REB in a column, which can be described as follows:

$$\mathrm{FM}_{\mathrm{cat}} = [\mathrm{FM}_1; \mathrm{FM}_2; \ldots; \mathrm{FM}_n; \ldots; \mathrm{FM}_N], \quad n = 1, 2, 3, \ldots, N \tag{3}$$

in which, the size of FM_n is $W \times H \times C$, and the size of $\mathrm{FM}_{\mathrm{cat}}$ is $W \times H \times (N \times C)$. In this article, there are two modal images; thus, $N = 2$. FM_1 and FM_2 express the output of the same level REB on different encoder branches.

According to (3), the concatenate operation changes the dimension of feature maps. To ensure the uniformity of the dimension, a convolution with the size 5×5 is added after the concatenate operation. It can adjust the size of feature maps to $W \times H \times C$ so as to provide convenience for subsequent decoding process.

C. Hybrid Transmission Decoder Module

In recent years, to satisfy higher task requirements, some scholars have improved network performance by integrating features of different layers. Luo *et al.* [39] design an LF3Net to increase the accuracy of salient detection, which utilizes the Stackelberg theory to make the low-level and high-level features to complement each other with a competitive way. Guo *et al.* [40] improve the dehazing capability of the network by fusing different level features. Ma *et al.* [41] present a hybrid network to improve the classification accuracy of hyperspectral images by mixing different level features. Inspired by above researches, this article designs an innovative HTDM to enhance the infrared and visible image fusion ability of the proposed network.

As shown in Fig. 4, for the purpose of obtaining parameters adapted to different level features, two independent decoder embranchments are adopted to deal with low-level and high-level features. In the decoding stage, F1 and F2 are considered as low-level features, and F3 and F4 are considered as high-level features. Each decoder embranchment mainly contains a CC and a series of residual decoder blocks (RDBs). The low-level concatenate-convolution (LCC) and high-level concatenate-convolution (HCC) are the same as CC. The structure of RDB is similar to REB in Fig. 3. It should be noted that hybrid transmission consists of two creative sections:

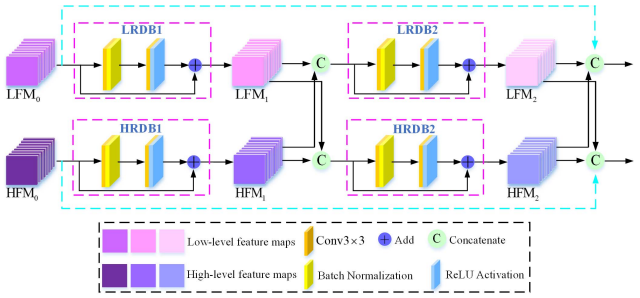


Fig. 4. Schematic of hybrid transmission.

cross transmission and skip transmission. Cross transmission aims to make the information of the high-level and low-level supplement each other, thus improving the information richness of the fused image. Skip transmission is on the purpose of remedying the information loss in the process of decoding. Cross transmission and skip transmission are explained in detail as follows.

1) *Cross Transmission:* As shown in Fig. 4, LFM_0 and HFM_0 are low-level and high-level feature maps obtained by LCC and HCC, respectively. LFM_0 is transmitted to the first low-level residual decoder block (LRDB1) to produce the LFM_1, and HFM_0 is transmitted to the first high-level residual decoder block (HRDB1) to produce the HFM_1. Then, LFM_1 is not only sent to the next step of the low-level decoder branch, but also sent to the high-level decoder branch. Similarly, HFM_1 is also inputted to the low-level decoder branch. This way of conveying information is named as cross transmission. Next, these feature maps are sent to the LRDB2 and HRDB2 to generate LFM_2 and HFM_2. LFM_2 and HFM_2 undergo the cross transmission.
2) *Skip Transmission*: Skip transmission is expressed by two cyan lines in Fig. 4. For the sake of compensating information loss in the decoder process, the feature information LFM_0 and HFM_0 is introduced into the LRDB3 and HRDB3, respectively.

In the low-level decoder branch, LFM_0, LFM_2, and HFM_2 are concatenated and sent to the LRDB3. In the high-level decoder branch, HFM_0, HFM_2 andLFM_2 are concatenated and sent to the HRDB3. As shown in Fig. 2, feature maps produced by LRDB3 and HRDB3 are combined together by an adding operation; then, the fused feature map is sent to a 3×3 convolution layer to output the final fused image.

D. Loss Function

A high-quality fusion image should contain strong intensity information and rich structure details simultaneously. Therefore, we construct the loss function from two aspects: the intensity loss and the structure loss, which is written by

$$L_{\text{total}} = L_{\text{intensity}} + \alpha L_{\text{structure}} \tag{4}$$

in which, L_{total}, $L_{\text{intensity}}$, and $L_{\text{structure}}$ represent the total loss, intensity loss, and structure loss, respectively. α is a balance parameter to control the tradeoffs between the intensity loss and structure loss. The intensity loss can be calculated as follows:

$$L_{\text{intensity}} = \sum_{x=1}^{X}\sum_{y=1}^{Y} \|I_{\text{f}}(x,y) - \text{mean}(I_{\text{ir}}(x,y), I_{\text{vis}}(x,y))\|_2^2 \tag{5}$$

$$\text{mean}(I_{\text{ir}}(x,y), I_{\text{vis}}(x,y)) = (I_{\text{ir}}(x,y) + I_{\text{vis}}(x,y))/2. \tag{6}$$

Here, I_{f} is the fused image, I_{ir} is the infrared image, and I_{vis} is the visible image. $\|\cdot\|_2^2$ denotes the L_2-norm. X and Y indicate the width and height of the image.

To fuse rich textural information, gradient information is generally utilized to calculate the structure loss [28]. Di Zenzo [42] proposes the structure tensor to express gradient information. According to [42], the gradient of infrared and visible images can be summarized by Jacobian matrix as follows:

$$J_{\text{I}}(x,y) = \begin{bmatrix} \nabla_x I_{\text{ir}}(x,y) & \nabla_y I_{\text{ir}}(x,y) \\ \nabla_x I_{\text{vis}}(x,y) & \nabla_y I_{\text{vis}}(x,y) \end{bmatrix} \tag{7}$$

where ∇_x and ∇_y indicate the derivative of the horizontal and vertical directions, respectively. Based on$J_{\text{I}}(x,y)$, the structure tensor is given by

$$S_{\text{I}}(x,y) = (J_{\text{I}}(x,y))^{\text{T}} \times J_{\text{I}}(x,y). \tag{8}$$

Combing (7) and (8), the structure tensor of input images can be described as follows:

$$S_{\text{I}}(x,y) = \begin{bmatrix} \tau_1 & \tau_2 \\ \tau_3 & \tau_4 \end{bmatrix} \tag{9}$$

$$\tau_1 = (\nabla_x I_{\text{ir}}(x,y))^2 + (\nabla_x I_{\text{vis}}(x,y))^2 \tag{10}$$

$$\tau_2 = \nabla_x I_{\text{ir}}(x,y) \times \nabla_y I_{\text{ir}}(x,y) + \nabla_x I_{\text{vis}}(x,y) \times \nabla_y I_{\text{vis}}(x,y) \tag{11}$$

$$\tau_3 = \nabla_y I_{\text{ir}}(x,y) \times \nabla_x I_{\text{ir}}(x,y) + \nabla_y I_{\text{vis}}(x,y) \times \nabla_x I_{\text{vis}}(x,y) \tag{12}$$

$$\tau_4 = (\nabla_y I_{\text{ir}}(x,y))^2 + (\nabla_y I_{\text{vis}}(x,y))^2. \tag{13}$$

Similarly, the gradient information $J_{\text{f}}(x,y)$ and structure tensor $S_{\text{f}}(x,y)$ of the fused image can be expressed as follows:

$$J_{\text{f}}(x,y) = \left[\nabla_x I_{\text{f}}(x,y) \nabla_y I_{\text{f}}(x,y)\right] \tag{14}$$

$$S_{\text{f}}(x,y) = \begin{bmatrix} (\nabla_x I_{\text{f}}(x,y))^2 \nabla_x I_{\text{f}}(x,y) \times \nabla_y I_{\text{f}}(x,y) \\ \nabla_y I_{\text{f}}(x,y) \times \nabla_x I_{\text{f}}(x,y) (\nabla_y I_{\text{f}}(x,y))^2 \end{bmatrix}. \tag{15}$$

On the basis of (9) and (15), the structure loss can be written by

$$L_{\text{structure}} = \sum_{x=1}^{X}\sum_{y=1}^{Y} \|S_{\text{f}} - S_{\text{I}}\|_{\text{F}}^2 \tag{16}$$

where $\|\cdot\|_{\text{F}}^2$ denotes the Frobenius norm.

To sum up, $L_{\text{intensity}}$ prefers the intensity content of the fused image and the input images to be the same, whereas $L_{\text{structure}}$ tends to make the structure tensors of the fused image and the input images to be identical. Consequently, the MHTNet can generate fused images with the strong intensity information and detailed structure textures of source images.

E. Training Setup

We select 36 pairs of infrared and visible images with various military and surveillance scenarios from TNO database [43] as the training data. In order to train a good model, each image of the training data is cropped with the stride 16, and each patch is the same size 120 × 120. Then, 15 712 pairs of infrared and visible patches are randomly chosen as training samples. The proposed MHTNet is trained on PyTorch 1.7 with 12-GB NVIDIA TITAN XP GPU. In the training of the MHTNet, the epoch is 150, the batch size is 16, and the learning rate is initialized as 10^{-3} and halved every 20 epochs until the end. In addition, the number of REB is 4, and the balance factor between intensity loss and structure loss is 0.01. Section IV-D will discuss the settings of REB number and α.

IV. Experimental Datasets and Settings

A. Infrared and Visible Image Datasets

In this article, the MHTNet is tested on three public datasets: TNO [43], KAIST [44], and Bristol Eden Project Multisensor (BEPM) [45]. TNO dataset includes many pairs of registered infrared and visible images under different scenes, which can be freely used for research purpose. Images of TNO dataset have different sizes, such as 360 × 270, 505 × 510, and 768 × 576. KAIST dataset contains many registered infrared and visible image pairs under various regular traffic scenes. BEPM dataset is a registered infrared and visible image dataset with a man dressed in camouflage walking through thick foliage, which is given by the Rochester Institute of Technology (RIT), Rochester, NY, USA.

B. Comparison Methods

Several classical and advanced image fusion methods are selected to evaluate the proposed MHTNet, including dual-tree complex wavelet transform fusion method (DTCWT) [46], multiresolution singular value decomposition fusion method (MSVD) [47], gradient transfer and total variation minimization fusion method (GTF) [48], fourth-order partial differential equations fusion method (FPDE) [49], deep unsupervised fusion method (DeepFuse) [26], FusionGAN [22], dense block-based fusion method (DenseFuse) [25], deep image fusion (DIF) [50], and VIF-Net (VIF) [51]. Experiments of our method and other deep learning methods (including DeepFuse, FusionGAN, DenseFuse, DIF, and VIF) are implemented with a 12-GB NVIDIA TITAN XP GPU. The traditional methods (including DTCWT, MSVD, GTF, and FPDE) are conducted in MATLAB 2018a on a computer (Intel Core i7-9700F, 3.0-GHz CPU).

C. Evaluation Metrics

Six metrics are selected to comprehensively evaluate the fusion performance of the proposed MHTNet, which are the sum of the correlations of differences (SCD) [52], visual information fidelity fusion (VIFF) [53], entropy (EN) [54], standard deviation (SD) [55], mutual information (MI) [56], and image quality metric of Chen–Varshney metric (Qcv) [57].

These above metrics evaluate the quality of fused images from different perspectives. SCD measures the complementary information of the fused image, VIFF evaluates the visual information fidelity of the fused image, EN expresses the information richness of the fused image, SD indicates the spread of the information in the image, MI calculates the amount of information obtained from the source images, and Qcv is a comprehensive evaluation metric, which is associated with the edge, saliency, and similarity of images. Among these indicators, SCD, VIFF, EN, SD, and MI are positively correlated with fusion image quality, whereas Qcv is negatively correlated with the fusion image quality. Thus, in this article, Qcv is denoted as Qcv-.

D. Parameter Settings

In the proposed MHTNet, two vital parameters will affect the performance of image fusion. The one is the balance factor α of the loss function, and the other is the number of REB in MREM. In this section, we discuss the influences of these two parameters on image fusion.

1) *The Analyses on Different Values of the Balance Factor α:* In this article, the loss function consists of two terms: intensity loss and structure loss. α is a balance factor to control the tradeoffs between the two terms. The value of α will directly affect the image fusion ability of the proposed algorithm. In order to get an optimal α to obtain a good fusion model, we train the proposed algorithm under α with diverse orders of magnitude (including 0.1, 0.01, and 0.001) and test the image fusion performance on three public datasets mentioned in Section IV-A. As provided in Table II, when the value of α is 0.01, evaluation metric values are the best among these three conditions, which means MHTNet has the best performance. As a result, in this article, the balance factor α between $L_{\text{intensity}}$ and $L_{\text{structure}}$ is set to 0.01.
2) *The Analyses on the Number of REB:* As introduced in Section III, MREM contains several REBs. Different numbers of REB represent different depths of the proposed network and different levels of features. In Table II, K represents the number of REB. We analyze the effect of the number of REB by comparing the results from $K = 4$ with those from fewer REBs ($K = 2$) and more REBs ($K = 6$). Table II provides the average values of six evaluation metrics for the fused images obtained by the proposed MHTNet with different K. When $K = 4$, most of the evaluation metric values on three public datasets are the best, which means the fusion quality is the highest. Therefore, in this article, the number of REB is set to 4.

V. Experimental Results and Analyses

This section analyzes the self-comparison experiments of our algorithm and compares the proposed MHTNet with other fusion methods subjectively and objectively.

A. Self-Comparison Experimental Results and Analyses of the Proposed Fusion Network

This article proposes the MHTNet to realize infrared and visible image fusion. MHTNet mainly includes two parts:

TABLE II

AVERAGE VALUES OF SIX EVALUATION METRICS FOR THE FUSED IMAGES OBTAINED BY THE PROPOSED MHTNET WITH DIFFERENT α AND DIFFERENT K. α IS THE BALANCE FACTOR OF LOSS FUNCTION, AND K IS THE NUMBER OF REB. "−" INDICATES THAT THE SMALLER THE EVALUATION METRIC IS, THE BETTER THE FUSION PERFORMANCE (**Bold**: THE BEST)

Dataset	Metrics	α			K		
		0.1	0.01	0.001	2	4	6
TNO	SCD	1.52534	**1.54347**	1.52871	1.53173	**1.54347**	1.53090
	VIFF	0.49937	**0.50792**	0.49204	0.48916	**0.50792**	0.48839
	EN	6.85695	**6.86635**	6.83724	6.81720	**6.86635**	6.82410
	SD	74.11507	**74.15441**	72.94048	71.81896	**74.15441**	72.22933
	MI	13.71390	**13.73270**	13.67447	13.63440	**13.73270**	13.64820
	Qcv -	501.80465	**476.90174**	532.65575	489.93299	**476.90174**	477.28392
KAIST	SCD	1.36620	**1.38585**	1.38424	**1.39998**	1.38585	1.39822
	VIFF	0.62426	**0.64328**	0.62063	0.60589	**0.64328**	0.62822
	EN	6.52278	**6.53945**	6.51084	6.48068	**6.53945**	6.52223
	SD	71.12384	**71.89273**	70.41978	68.75801	**71.89273**	70.96237
	MI	13.04557	**13.07890**	13.02169	12.96137	**13.07890**	13.04445
	Qcv -	376.27192	**315.58684**	371.36541	371.67080	**315.58684**	330.72190
BEPM	SCD	1.67334	**1.68936**	1.67470	1.67734	**1.68936**	1.67713
	VIFF	0.40450	**0.41622**	0.40431	0.40385	**0.41622**	0.40373
	EN	6.85586	**6.88819**	6.85195	6.85036	**6.88819**	6.85310
	SD	63.64290	**64.88541**	63.47975 z	63.44695	**64.88541**	63.57550
	MI	13.71172	**13.77638**	13.70389	13.70073	**13.77638**	13.70621
	Qcv -	517.63420	**463.05821**	513.83797	493.57960	**463.05821**	488.92421

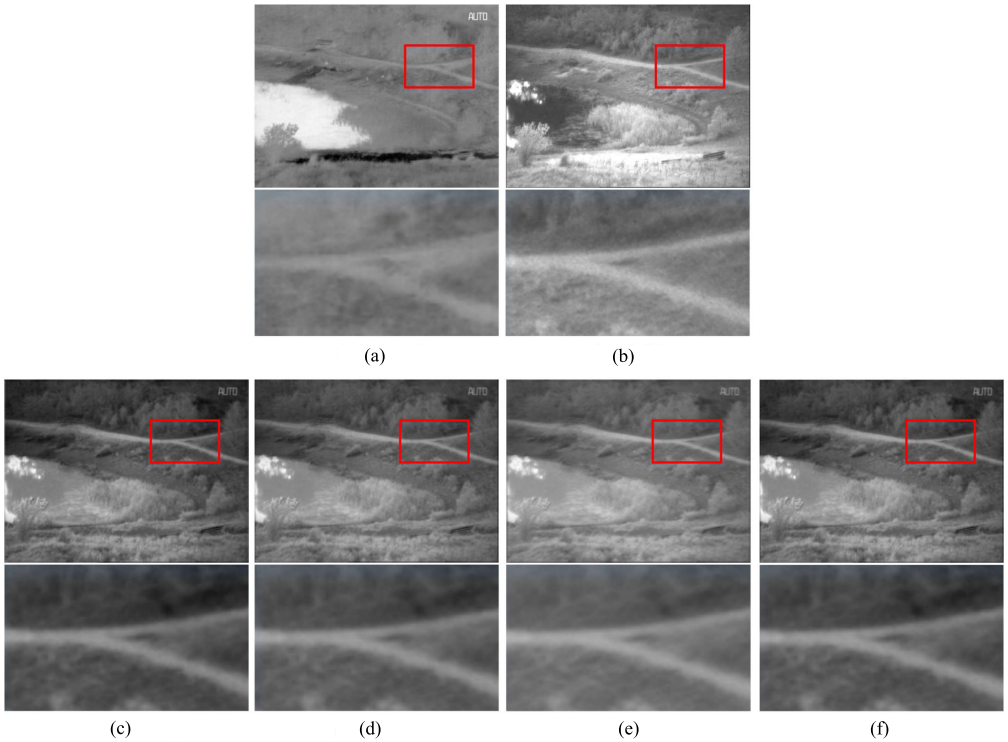


Fig. 5. Examples of self-comparison experimental results on "Lake" image of TNO dataset: (a) infrared; (b) visible; (c) CON1; (d) CON2; (e) CON3; and (f) CON4.

encoder and decoder. In the encoder part, we design the MREM with two independent branches to obtain practicable features. In the decoder part, we construct the HTDM to utilize the multilevel features adequately. HTDM contains two major structures: cross transmission and skip transmission. The purpose of cross transmission is to make the information

TABLE III

OBJECTIVE EVALUATION OF SELF-COMPARISON. "−" INDICATES THAT THE SMALLER THE EVALUATION METRIC IS, THE BETTER THE FUSION PERFORMANCE (**Bold**: THE BEST)

Dataset	Metrics	CON1	CON2	CON3	CON4
TNO	SCD	1.51931	1.52906	1.52907	**1.54347**
	VIFF	0.48814	0.49080	0.49244	**0.50792**
	EN	6.82826	6.83199	6.84346	**6.86635**
	SD	72.51015	72.71569	73.25597	**74.15441**
	MI	13.65653	13.66398	13.68693	**13.73270**
	Qcv -	530.21871	489.22161	507.09219	**476.90174**
KAIST	SCD	1.37033	**1.39371**	1.37255	1.38585
	VIFF	0.61627	0.62492	0.63076	**0.64328**
	EN	6.50585	6.51490	6.51522	**6.53945**
	SD	69.68374	70.80123	70.89069	**71.89273**
	MI	13.01170	13.02981	13.03044	**13.07890**
	Qcv -	370.45406	341.57127	363.24326	**315.58684**
BEMP	SCD	1.66114	1.67730	1.67326	**1.68936**
	VIFF	0.40425	0.40469	0.40448	**0.41622**
	EN	6.85014	6.85476	6.85417	**6.88819**
	SD	63.34231	63.61412	63.55872	**64.88541**
	MI	13.70027	13.70952	13.70835	**13.77638**
	Qcv -	503.78740	492.60854	513.15679	**463.05821**

between low-level and high-level complement each other, and the function of skip transmission is to compensate for information loss in the decoding process. To verify effectiveness and advantages of these novel designs in MHTNet, we conduct the self-comparison experiments under four conditions labeled CON1, CON2, CON3, and CON4, respectively.

CON1: using the MREM and the simple decoder module without cross transmission and skip transmission.

CON2: using the MREM and the decoder module with cross transmission.

CON3: using the MREM and the decoder module with skip transmission.

CON4: using the MREM and the decoder module with cross transmission and skip transmission. CON4 is the proposed MHTNet. The self-comparison experiments are carried out on three public datasets mentioned in Section IV-A. The subjective and objective experimental results are shown in Fig. 5 and Table III, respectively.

As shown in Fig. 5, in order to better analyze the quality of fused images under different conditions, the contents in red boxes are enlarged and exhibited at the bottom of experimental results. The fused image of CON1 contains the clear word "AUTO" of the infrared image and the lake wave light of the visible image, which means the proposed MREM can obtain features from source images effectively. Compared with CON1, the fused image produced by CON2 has more salient fork road and richer background information, which illustrates that the cross transmission can improve the image fusion quality by enhancing the communication between low-level and high-level features. The fused image of CON3 contains more intensity information than that of CON1 and CON2, which proves that skip transmission can further improve the image fusion quality through compensating the information loss during the decoding process. Compared with other three conditions, the fused image of CON4 has more salient path, more clear edges, richer background details, higher contrast, and visual quality, which demonstrates that the proposed network has better fusion performance by comprehensively utilizing the above innovative modules.

Table III provides the objective evaluation results of the three conditions. These values of CON1 on three public datasets are acceptable, which shows the effectiveness of MREM. Compared with CON1, the evaluation metric values of CON2 and CON3 are increased, which indicates that the cross transmission and skip transmission of decoder module are effective to enhance the fusion performance. Moreover, these evaluation metric values (excepting SCD in KAIST dataset) of CON4 are the best among four conditions, which proves that the proposed MHTNet can generate higher quality infrared and visible fusion images when combing MREM, cross transmission, and skip transmission.

B. Evaluation of the Experimental Results

To test the fusion performance of our proposed method, MHTNet and comparison methods are carried out on TNO, KAIST, and BEPM datasets. In the previous infrared and visible image fusion works, researchers generally choose about 20 pairs of images to test the fusion algorithms [24]. Thus, we select 20 infrared and visible image pairs of each dataset for testing. The subjective and objective analyses on three public datasets are presented as follows.

1) Experiments on TNO Dataset: Examples of the fused images on TNO dataset are provided in Fig. 6. Overall, the proposed method exhibits better fusion performance than its competitors. As shown in red boxes, fused images of DTCWT, MSVD, GTF, FPDE, DenseFuse, and VIF miss much salient information, e.g., the helicopter and the person. In addition, as shown in blue boxes, fused images produced by MSVD, FusionGAN, DenseFuse, DIF, and VIF have fuzzy edges and weak contrast. Especially in blue boxes of "movie_18," DTCWT, GTF, FusionGAN, and VIF can hardly integrate the texture details from the window of car into the fusion image. Compared with other methods, MHTNet can obtain fused images with more salient target information, more sharpened and clear textures, and higher visual quality.

Quantitative comparisons on TNO dataset are given in Table IV. Six metrics mentioned in Section IV-C are used for evaluation. In general, the evaluation metric values obtained by MHTNet on TNO dataset are all commendable. Specifically, the proposed method ranks best on EN, MI, SD, and Qcv-,

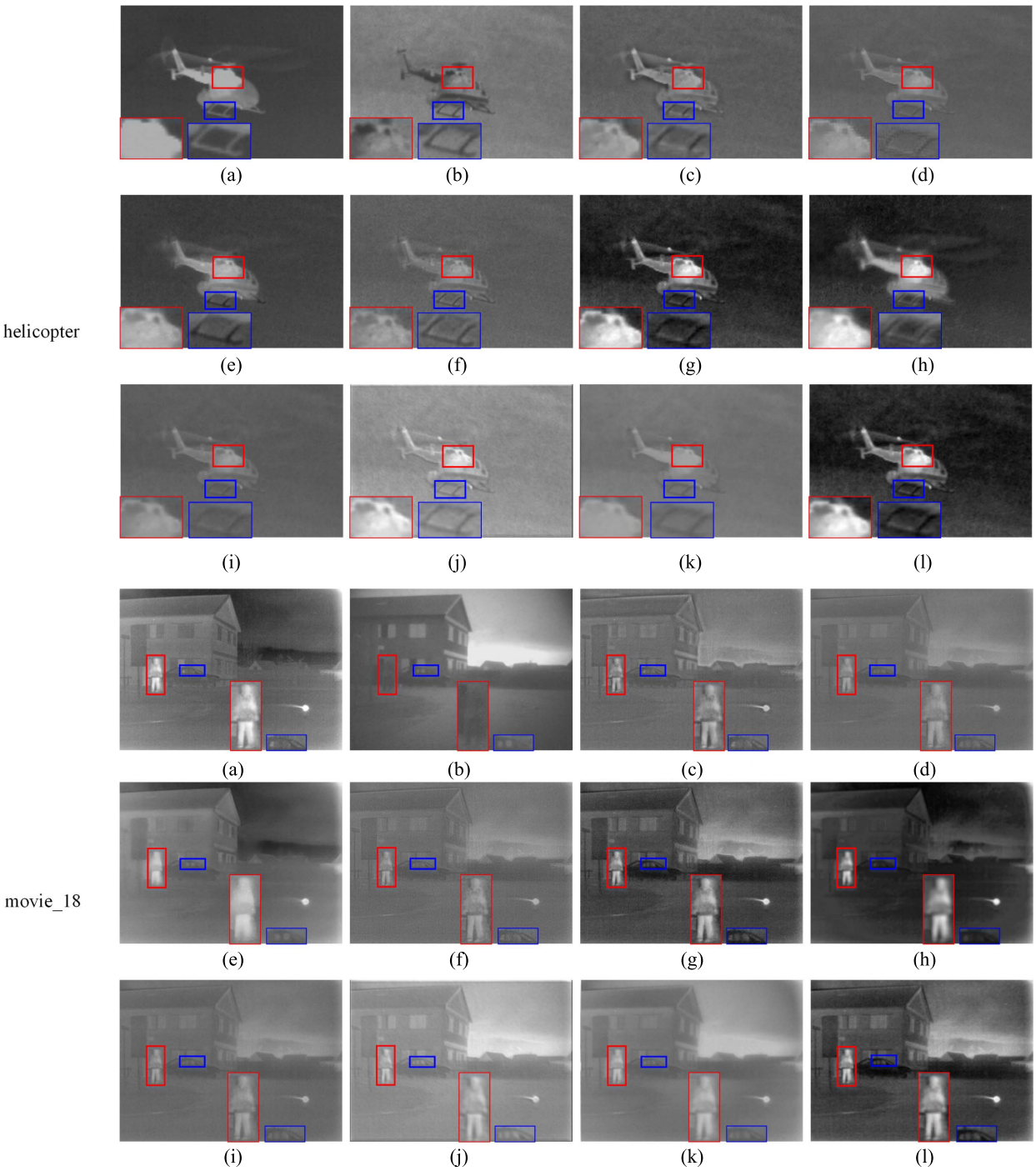


Fig. 6. Comparison experimental results on TNO dataset: (a) infrared; (b) visible; (c) DTCWT; (d) MSVD; (e) GTF; (f) FPDE; (g) DeepFuse; (h) FusionGAN; (i) DenseFuse; (j) DIF; (k) VIF; and (l) MHTNet.

which indicates that it can generate the fused image with sufficient intensity information, fine information distribution, clear edges, and high fidelity with source images. In addition, our method ranks second on VIFF, which means that MHTNet can produce fused images with high visual quality. These metric results prove that the proposed method has a better performance in infrared and visible image fusion field.

For assessing the proposed fusion framework intuitively, the bar charts of different fusion methods about six evaluation metrics on TNO dataset are shown in Fig. 7. Different color bars represent the average evaluation metrics values of different fusion methods. It can be seen that bars of MHTNet on metrics EN, MI, and SD are significantly higher than those of other methods, which proves that the proposed fusion method can retain sufficient information. Moreover, the Qcv- bar of MHTNet is the shortest among those of comparison methods, which illustrates that the proposed method performs the best comprehensive fusion ability.

2) Experiments on KAIST Dataset: As shown in red boxes of Fig. 8 on KAIST dataset, the light information in the fused images of MSVD, FPDE, FusionGAN, DenseFuse, and VIF is weaker than that of other methods. As presented

TABLE IV

AVERAGE EVALUATION METRIC VALUES OBTAINED BY DIFFERENT FUSION METHODS ON TNO, KAIST, AND BEPM DATASETS. "–" INDICATES THAT THE SMALLER THE EVALUATION METRIC IS, THE BETTER THE FUSION PERFORMANCE (RED: THE BEST AND BLUE: THE SECOND BEST)

Dataset	Metrics	DTCWT	MSVD	GTF	FPDE	DeepFuse	FusionGAN	DenseFuse	DIF	VIF	MHTNet
TNO	SCD	1.59085	1.58315	0.96545	1.58578	1.52973	1.01337	1.59213	1.52873	1.52869	1.54347
	VIFF	0.30440	0.24161	0.18818	0.23393	0.52923	0.18613	0.25512	0.22548	0.15345	0.50792
	EN	6.38778	6.18784	6.63534	6.23833	6.68435	6.36287	6.17403	6.34500	6.28288	6.86635
	SD	53.34962	48.16242	67.62603	48.29532	67.65344	54.35802	47.82040	55.37665	54.14508	74.15441
	MI	12.77557	12.37567	13.27069	12.47666	13.36869	12.72573	12.34807	12.69001	12.56577	13.73270
	Qcv -	535.83512	518.25340	1322.81681	503.81345	495.47642	1064.92911	485.83484	740.87967	481.93705	476.90174
KAIST	SCD	1.13576	1.23579	1.12716	1.25296	1.39019	1.19396	1.17195	1.05267	1.24213	1.38585
	VIFF	0.63166	0.38631	0.31328	0.43535	0.65289	0.23626	0.39088	0.54898	0.37206	0.64328
	EN	6.13092	5.98923	5.69266	6.04963	6.39691	5.63667	5.97557	6.11660	6.32094	6.53945
	SD	54.22110	48.42065	35.67014	49.18907	67.90102	39.68162	48.26599	60.94386	61.22929	71.89273
	MI	12.26184	11.97847	11.38532	12.09927	12.79382	11.27335	11.95113	12.23320	12.64188	13.07890
	Qcv -	407.03928	377.32373	1599.85104	366.16170	354.69058	1431.12742	374.71297	508.18508	836.87073	315.58684
BEPM	SCD	1.55656	1.55012	1.11964	1.55721	1.71017	1.00031	1.55760	1.42717	1.37309	1.68936
	VIFF	0.46236	0.33638	0.21934	0.29386	0.46620	0.08510	0.33656	0.31365	0.22909	0.41622
	EN	6.84276	6.53099	6.31508	6.52503	6.72591	5.66331	6.46943	6.79852	6.53485	6.88819
	SD	61.00927	49.42738	49.32724	48.76064	59.56136	38.58609	47.90606	60.26611	53.14512	64.88541
	MI	13.68552	13.06198	12.63016	13.05007	13.45183	11.32663	12.93887	13.59705	13.06970	13.77638
	Qcv -	516.98918	691.62846	1585.23442	624.53805	569.41968	2030.69450	584.57863	654.48057	601.13298	463.05821

Fig. 7. Bar charts of different fusion methods about six evaluation metrics on TNO dataset.

in blue boxes of Fig. 8 on KAIST dataset, fused images obtained by GTF, FPDE, FusionGAN, and DenseFuse have less intensity information of roadblocks. In addition, fused images of GTF and FusionGAN lack contour structures of the trees. In contrast, the proposed method can generate fused images with more intensity information, clearer structure, and higher visual quality.

Objective results on KAIST dataset are provided in Table IV. Metrics values (including EN, MI, SD, and Qcv-) of MHTNet are the best among these comparison methods. In addition, MHTNet ranks second on SCD and VIFF. These values prove the effectiveness and advantages of the proposed method in the field of infrared and visible image fusion.

The bar charts of different fusion methods about six evaluation metrics on KAIST dataset are presented in Fig. 9, which visually represents the superiority of the proposed method. Significantly, bars of MHTNet on SCD, VIFF, EN, SD, and MI are higher than those of most comparison methods, and bars of MHTNet on Qcv- are much lower than those of other methods. The above results reflect the excellent performance of the proposed method in infrared visible image fusion.

3) Experiments on BEPM Dataset: Qualitative results on BEPM dataset are shown in Fig. 8. As exhibited in red boxes on BEPM dataset, fused images of MSVD, GTF, FPDE, FusionGAN, DenseFuse, and VIF miss many detailed features of leaf texture in visible images. As exhibited in blue boxes on BEPM dataset, fused images produced by DTCWT, MSVD, FPDE, DenseFuse, DIF, and VIF have weaker human information of infrared images. By contrast, the fused image of the proposed method has stronger pixel contrast and clearer leaf texture structures.

As exhibited in Table IV, MHTNet obviously performs better than competitors on BEPM dataset. MHTNet ranks the best on EN, SD, MI, and Qcv-. The average values on SCD

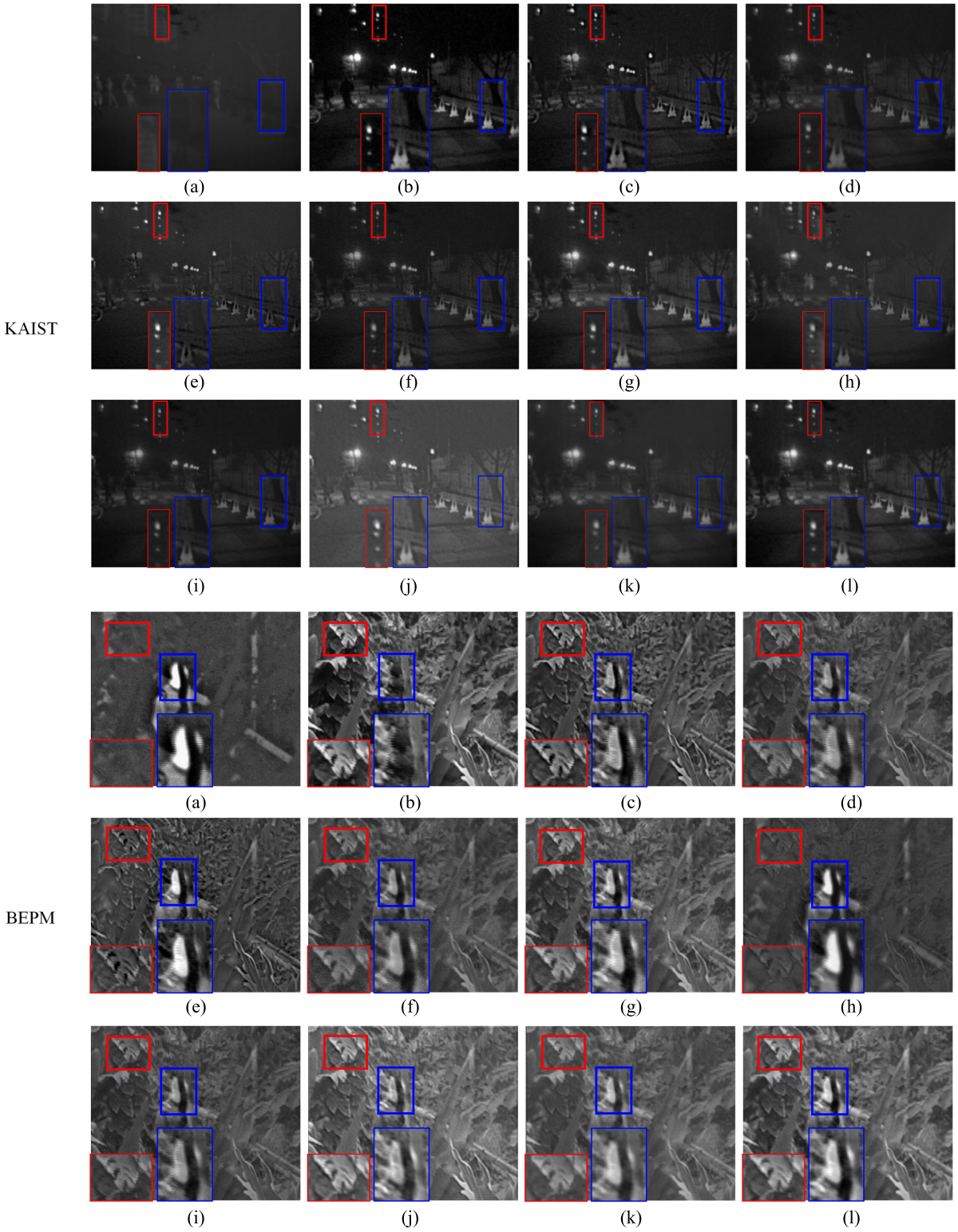


Fig. 8. Comparison experimental results on KAIST and BEPM datasets: (a) infrared; (b) visible; (c) DTCWT; (d) MSVD; (e) GTF; (f) FPDE; (g) DeepFuse; (h) FusionGAN; (i) DenseFuse; (j) DIF; (k) VIF; and (l) MHTNet.

and VIFF metrics of MHTNet are not the maximum, but they are still acceptable. These values mean that compared with other algorithms, the proposed method has more competitive ability in the field of infrared and visible image fusion.

To observe the advantages of the proposed method directly, Fig. 10 shows the bar charts of different fusion methods about six evaluation metrics on BEPM dataset. Obviously, bars (including SCD, VIFF, EN, SD, and MI) of MHTNet are higher than those of most comparison methods. The Qcv- bar of MHTNet is the lowest among these methods. As a result, the proposed method still shows better performance on BEPM dataset than competitors.

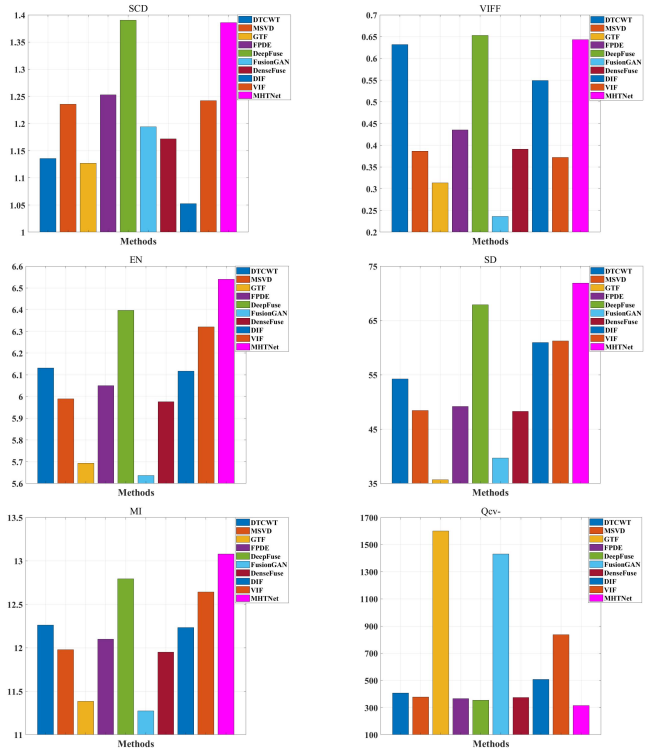


Fig. 9. Bar charts of different fusion methods about six evaluation metrics on KAIST dataset.

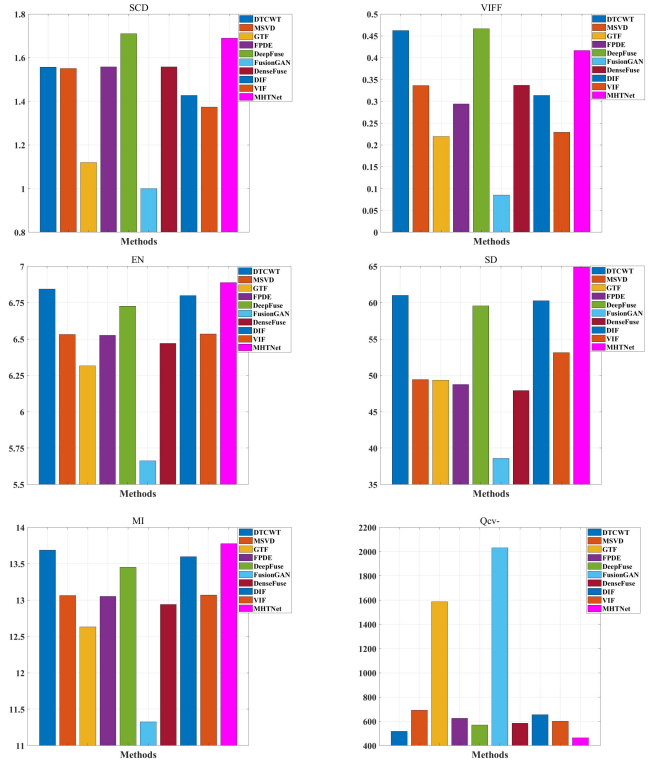


Fig. 10. Bar charts of different fusion methods about six evaluation metrics on BEPM dataset.

Above subjective and objective analyses of the experimental results on TNO, KAIST, and BEPM datasets demonstrate that the proposed method can generate fused images with more salient target information, more sharpened edges, richer details, and higher visual quality than comparison methods. In addition, as shown in Table IV, DTCWT shows good performance on BEPM dataset, and DeepFuse expresses well on TNO and KAIST datasets, whereas MHTNet can achieve satisfactory image fusion on TNO, KAIST, and BEPM datasets. Although the fused images of DeepFuse and MHTNet are visually similar, values of these evaluation metrics in Table IV demonstrate that MHTNet outperforms DeepFused in three datasets. As a result, MHTNet has better fusion performance than DeepFuse.

To sum up, the proposed method has not only better infrared and visible image fusion performance, but also better robustness to environmental changes than competitors.

4) Analyses of Time Complexity: The advantages of the proposed method in terms of fusion performance and robustness have been described above. Running time is also an important factor that must be considered in the image processing system. Therefore, we test the running time of the proposed and comparison methods.

The same fusion method has different computational complexities to process different size images. We compare the fusion time of different fusion algorithms under three image sizes 359 × 247, 505 × 510, and 768 × 576. As shown in Table V, the running time of MHTNet is significantly smaller than that of DTCWT, MSVD, GTF, FPDE, FusionGAN, DenseFuse, and VIF. Although the processing time of MHTNet is larger than that of DIF, it already has good real-time performance. Moreover, the fusion speed of MHTNet is obviously faster than DeepFuse, which demonstrates that MHTNet has more comprehensive fusion capability than DeepFuse. The average running time of MHTNet to fuse a pair of images is about 0.013630 s, that is, MHTNet can fuse 73 pairs of infrared and visible images per second. Experimental results of the time loss show that the proposed method has high computational efficiency and satisfies the requirement of real-time while maintaining good fusion performance.

TABLE V
TIME LOSS OF IMAGE FUSION WITH DIFFERENT IMAGE SIZES BY DIFFERENT METHODS

Methods	Image Size		
	359×247	505×510	768×576
DTCWT	0.073112	0.199686	0.340017
MSVD	0.081196	0.240129	0.406605
GTF	0.491226	3.663870	6.560461
FPDE	0.246746	1.258907	2.231045
DeepFuse	0.165018	0.331668	0.518540
FusionGAN	1.555519	4.095427	7.814400
DenseFuse	0.031200	0.078100	0.109300
DIF	0.007972	0.008251	0.009166
VIF	0.103360	0.150726	0.192517
MHTNet	0.011968	0.013963	0.014960

VI. Conclusion

In this article, an MHTNet has been developed. To extract features from source images effectively, an MREM with two independent branches has been designed. In order to avoid complex manual fusion strategies, the CC has been utilized to fuse feature maps at the same level. Moreover, to make full use of the information of different levels and improve the fusion performance, an HTDM has been presented, which includes two ingenious parts: cross transmission and skip transmission.

Extensive experiments have been conducted on TNO, KAIST, and BEPM datasets to verify the fusion performance of the proposed network. Experimental results convincingly demonstrate that the proposed network not only can produce fusion images with strong intensity information, sharpened edges, rich texture details, and high visual quality, but also performs better than comparison methods in terms of qualitative and quantitative aspects. In addition, the time loss of the proposed network has been tested, and values of the running time with different image sizes show that the proposed network has better real-time performance and faster computing speed than most comparison algorithms.

Acknowledgment

The authors would like to thank the anonymous reviewers for their constructive comments.

References

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[2] S. G. Simone, A. Farina, F. C. Morabito, S. B. Serpico, and L. Bruzzone, "Image fusion techniques for remote sensing applications," *Inf. Fusion*, vol. 3, no. 1, pp. 3–15, 2002.

[3] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Computer Vision, Graphics and Image Processing*. Cham, Switzerland: Springer, 2006, pp. 528–539.

[4] G. Li, Y. Lin, and X. Qu, "An infrared and visible image fusion method based on multi-scale transformation and norm optimization," *Inf. Fusion*, vol. 71, pp. 109–129, Jul. 2021.

[5] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, vol. 63, pp. 166–187, Nov. 2020.

[6] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.

[7] G. He, J. Ji, D. Dong, J. Wang, and J. Fan, "Infrared and visible image fusion method by using hybrid representation learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 11, pp. 1796–1800, Nov. 2019.

[8] X. Jin *et al.*, "Infrared and visual image fusion method based on discrete cosine transform and local spatial frequency in discrete stationary wavelet transform domain," *Infr. Phys. Technol.*, vol. 88, pp. 1–12, Jan. 2018.

[9] J. Jinju, N. Santhi, K. Ramar, and B. S. Bama, "Spatial frequency discrete wavelet transform image fusion technique for remote sensing applications," *Eng. Sci. Technol., Int. J.*, vol. 22, no. 3, pp. 715–726, Jun. 2019.

[10] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, Nov. 2005.

[11] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.

[12] J. Adu, J. Gan, Y. Wang, and J. Huang, "Image fusion based on non-subsampled contourlet transform for infrared and visible light image," *Infr. Phys. Technol.*, vol. 61, pp. 94–100, Nov. 2013.

[13] Z. Wang, J. Xu, X. Jiang, and X. Yan, "Infrared and visible image fusion via hybrid decomposition of NSCT and morphological sequential toggle operator," *Optik*, vol. 201, Jan. 2020, Art. no. 163497.

[14] K. Guo and D. Labate, "Optimally sparse multidimensional representation using shearlets," *SIAM J. Math. Anal.*, vol. 39, no. 1, pp. 298–318, 2007.

[15] B. Zhang, X. Lu, H. Pei, and Y. Zhao, "A fusion algorithm for infrared and visible images based on saliency analysis and non-subsampled shearlet transform," *Infr. Phys. Technol.*, vol. 73, pp. 286–297, Nov. 2015.

[16] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.

[17] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infr. Phys. Technol.*, vol. 83, pp. 94–102, Jun. 2017.

[18] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.

[19] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," 2018, *arXiv:1804.08992*.

[20] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.

[21] T. Nie, L. Huang, H. Liu, X. Li, and B. He, "Multi-exposure fusion of gray images under low illumination based on low-rank decomposition," *Remote Sens.*, vol. 13, p. 204, Jun. 2021.

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

[23] H. Zhang, H. Xu, X. Tian, J. Jiang, and J. Ma, "Image fusion meets deep learning: A survey and perspective," *Inf. Fusion*, vol. 76, pp. 323–336, Dec. 2021.

[24] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 2705–2710.

[25] 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, Dec. 2018.

[26] 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, pp. 4714–4722.

[27] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.

[28] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, and J. Wu, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.

[29] H. Li, X.-J. Wu, and J. Kittler, "RFN-Nest: An end-to-end residual fusion network for infrared and visible images," *Inf. Fusion*, vol. 73, pp. 72–86, Sep. 2021.

[30] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[31] L. Jian, X. Yang, Z. Liu, G. Jeon, M. Gao, and D. Chisholm, "SEDR-Fuse: A symmetric encoder–decoder with residual block network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2020.

[32] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039.

[33] H. T. Mustafa, J. Yang, H. Mustafa, and M. Zareapoor, "Infrared and visible image fusion based on dilated residual attention network," *Optik*, vol. 224, Dec. 2020, Art. no. 165409.

[34] K. Li *et al.*, "Depthwise separable ResNet in the MAP framework for hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020.

[35] G. Song, K. Song, and Y. Yan, "EDRNet: Encoder–decoder residual network for salient object detection of strip steel surface defects," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9709–9719, Dec. 2020.

[36] L. Ding *et al.*, "MP-ResNet: Multipath residual network for the semantic segmentation of high-resolution PolSAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2021.

[37] J. Li, H. Zhang, R. Song, W. Xie, Y. Li, and Q. Du, "Structure-guided feature transform hybrid residual network for remote sensing object detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–13, 2021.

[38] B. Jiang *et al.*, "Deep dehazing network for remote sensing image with non-uniform haze," *Remote Sens.*, vol. 13, no. 21, p. 4443, Nov. 2021.

[39] H. Luo, G. Han, X. Wu, P. Liu, H. Yang, and X. Zhang, "LF3Net: Leader-follower feature fusing network for fast saliency detection," *Neurocomputing*, vol. 449, pp. 24–37, Aug. 2021.
[40] J. Guo, J. Yang, H. Yue, H. Tan, and K. Li, "RSDehazeNet: Dehazing network with channel refinement for multispectral remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2535–2549, Mar. 2020.
[41] W. Ma *et al.*, "A novel adaptive hybrid fusion network for multiresolution remote sensing images classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–17, 2021.
[42] S. Di Zenzo, "A note on the gradient of a multi-image," *Comput. Vis., Graph., Image Process.*, vol. 33, no. 1, pp. 116–125, 1986.
[43] T. Alexander, "TNO image fusion dataset," The Netherlands Org., The Netherlands, Tech. Rep. 26.04.2014, 2014.
[44] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.
[45] J. J. Lewis, S. G. Nikolov, A. Loza, E. F. Canga, and M. I. Smith, "The Eden Project multi-sensor data set," *Architecture Urbanism*, 2006.
[46] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel-and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, 2007.
[47] V. P. S. Naidu, "Image fusion technique using multi-resolution singular value decomposition," *Defence Sci. J.*, vol. 61, no. 5, p. 479, 2011.
[48] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[49] D. P. Bavirisetti, G. Xiao, and G. Liu, "Multi-sensor image fusion based on fourth order partial differential equations," in *Proc. 20th Int. Conf. Inf. Fusion (Fusion)*, Jul. 2017, pp. 1–9.
[50] H. Jung, Y. Kim, H. Jang, N. Ha, and K. Sohn, "Unsupervised deep image fusion with structure tensor representations," *IEEE Trans. Image Process.*, vol. 29, pp. 3845–3858, 2020.
[51] R. Hou *et al.*, "VIF-Net: An unsupervised framework for infrared and visible image fusion," *IEEE Trans. Comput. Imag.*, vol. 6, pp. 640–651, 2020.
[52] V. Aslantas and E. Bendes, "A new image quality metric for image fusion: The sum of the correlations of differences," *AEU-Int. J. Electron. Commun.*, vol. 69, no. 12, pp. 1890–1896, 2015.
[53] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, pp. 127–135, Apr. 2013.
[54] J. W. Roberts, F. B. Ahmed, and J. A. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, 2008, Art. no. 023522.
[55] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, p. 355, 1997.
[56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[57] H. Chen and P. K. Varshney, "A human perception inspired quality metric for image fusion based on regional information," *Inf. Fusion*, vol. 8, no. 2, pp. 193–207, Apr. 2007.

Qingqing Li received the B.E. degree from Hainan University, Haikou, China, in 2017. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

Her research interests include image registration, image fusion, and deep learning.

Guangliang Han received the M.S. and Ph.D. degrees from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China, in 2000 and 2003, respectively.

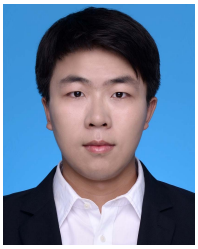
He is currently a Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. His current research interests are mainly focused on computer vision, image processing, and object tracking.

Peixun Liu received the Ph.D. degree from Jilin University, Changchun, China, in 2015.

He is currently an Associate Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun. His research interests include image processing, object detection, and robot automation.

Hang Yang received the B.S. and Ph.D. degrees from Jilin University, Changchun, China, in 2007 and 2012, respectively.

He is currently an Associate Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun. His research interests include image restoration and object tracking.



Dianbing Chen received the B.S. degree from Jilin University, Changchun, China, in 2012, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2017.

He is currently an Assistant Research Fellow with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun. His research interests include object tracking and sparse representation.

Xinglong Sun received the M.S. degree from the Beijing Institute of Technology, Beijing, China, in 2015. He is currently pursuing the Ph.D. degree with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

His current research interests are mainly focused on deep learning, object tracking, and image registration.

Jiajia Wu received the B.S. degree from Northeastern University, Qinhuangdao, China, in 2017. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China.

Her current research interests are mainly focused on RGBD saliency detection and deep learning.

Dongxu Liu received the B.E. degree from the Nanjing University of Information Science and Technology, Nanjing, China, in 2018. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences, Beijing, China, and the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China.

Her research interests include hyperspectral image classification and deep learning.