IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 17, 2021, accepted May 28, 2021, date of publication June 3, 2021, date of current version July 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3086096

Infrared and Visible Image Fusion Method Based on ResNet in a Nonsubsampled Contourlet Transform Domain

CE GAO, DONGHAO QI, YANCHAO ZHANG, CONGCONG SONG, AND YI YU
Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China

Corresponding author: Ce Gao (50616636@qq.com)

This work was supported in part by the National Science Foundation of China under Grant 51675506, and in part by the National Science and Technology Major Project under Grant 2017ZX10304403.

ABSTRACT Although the traditional image fusion method can obtain rich image results, obvious artificial noise and artifacts are often present in the resulting image. Fusion algorithms based on neural networks can avoid the shortcomings of traditional methods, but they are more complex and less flexible. In this study, we proposed a fusion method using the deep residual neural network ResNet152, which can not only effectively suppress artificial noise but also preserve the edge details of the image and improve the efficiency of the neural network. The proposed method is characterized by a multiscale transformation of an infrared image and visible light image in the optimized nonsubsampled contourlet transformation domain, and the deep residual neural network ResNet152 is used to extract the deep features of the low-pass component to guide the fusion of the low-pass component. The bandpass component is fused by taking the modulus maximum. This method can fully retain the global features and structural information of the source image in the result image. Compared to existing fusion methods using public test image sets, the experimental results show that on a subjective level, the fusion method creates sharper depth edges and fewer noise artifacts than traditional fusion methods. From an objective perspective, the average value for the results of the evaluation function is greater than that of other fusion methods.

INDEX TERMS Image fusion, nonsubsampled contourlet transform, residual network, feature guidancet.

I. INTRODUCTION

In fields such as the military, navigation, stealth weapon detection, and medical imaging [1]–[4], a variety of different imaging bands are typically necessary to monitor the target scene to obtain a more comprehensive visual understanding. Surveillance methods commonly use both visible and infrared images. Using cameras at different wavebands to acquire images can provide rich and detailed scene information. Nevertheless, for specific observation scenarios, the imaging advantages of multiple image bands can be combined to show more detailed information.

Image fusion technology has been extensively studied in the past several decades. Initially, the multiscale transformation method based on Laplacian [5], [6] and contrast pyramids [7], [8] was proposed for image decomposition. Liu *et al.* [9] designed an image fusion method based on a steerable pyramid and expectation maximization. This method is superior to traditional fusion methods using steerable pyramids. The wavelet transform, which is also used in image decomposition, has coefficient incoherence, which differs from the pyramid transform, and has been widely used in the field of image fusion. Using the traditional wavelet transform, Chai *et al.* [10] proposed an image fusion model based on the quaternion wavelet transform. The traditional wavelet transform uses a set of filters to decompose the original image into a series of high-pass and low-pass subimages. However, this approach still has shortcomings such as oscillation, shift variance, and insufficient directivity, which cause artifacts to appear in the fused image.

Do and Vetterli [11] proposed a multidirectional multiresolution image transformation method called contour transformation. This method is impacted by the problem of displacement variance caused by upsampling and downsampling of the pyramid filter bank. To solve this problem, Da Cunha *et al.* [12] proposed a nonsubsampled contourlet

The associate editor coordinating the review of this manuscript and approving it for publication was Yongqiang Zhao.

transform (NSCT) model with full shift invariance. Due to its advantages in image decomposition, NSCT is used in image fusion research. Adu *et al.* [13] proposed a fusion method based on NSCT and gradient characteristics. The method can not only extract the target from infrared images but also preserve details in the visible image. However, because of rough feature extraction, more artificial noise is produced in the results.

A number of studies have focused on the high-frequency component of NSCT decomposition. For example, Jin *et al.* [35] used intersecting cortical models (ICMs) to extract the edge information of the image from the high-frequency component. Huang *et al.* [36] used a pulse coupled neural network (PCNN) to fuse different subbands of the high-frequency component. However, when deep learning technology was introduced into the field of image fusion based on multiscale decomposition, many studies focused on the low-frequency subbands [38] because the neural network can better extract the hierarchical features of the low-frequency image when processing low-frequency images containing contour information, which allows the resulting image to retain more information of the source image.

Neural networks have been applied in the field of infrared and visible image fusion due to their strong adaptability, fault tolerance, and antinoise ability [14], [15]. The PCNN [16]–[18] has been studied extensively to fuse infrared and visible light images. This network simulates the pulses of brain neurons and requires input excitation to generate pulses. Qu *et al.* [22] proposed a fusion method that uses spatial frequency (SF) to excite a PCNN. The decision map is then calculated using the pulse time of the PCNN and ultimately generates a fusion image. Liu *et al.* [29] and others proposed a fusion method based on Convolutional Sparse Representation (ConvSR). They extracted deep features from the source image and then used these features to generate a fusion image.

Ma *et al.* [37] proposed a fusion framework for generative confrontation networks with multiple classification constraints. This method innovatively transformed the fusion problem into a multiple distribution estimation problem, and it is currently one of the most advanced neural network fusion models. However, there is information loss in the fusion process using neural networks or fully connected networks, and the algorithms have poor performance in terms of complexity and robustness. Liu *et al.* [39] used the residual neural network ResNet50 to fuse the low-pass components of the nonsubsampled shearlet transform (NSST) domain and achieved good fusion results in multifocus images. A new fusion idea was thus introduced. However, the training of ResNet50 affects the flexibility of the algorithm.

A key factor for approaches to image fusion is the selection of rules for conversion and fusion. Combining the advantages of different methods to establish an enhanced image fusion model is a critical issue when fusing infrared and visible images. In this study, we proposed a new fusion method that uses the NSCT decomposition strategy to decompose infrared and visible images. This approach not only retains the direction information of the image but also increases the decomposition speed. The deep residual neural network ResNet152 was used to extract the features of the decomposed low-pass subbands to obtain a feature map. Therefore, the main information in the infrared and visible light images is retained in the fusion result. This method avoids the algorithm complexity caused by training the network and has good fusion efficiency. The fusion of the bandpass subband uses the method of taking the maximum value of the modulus so that the resulting image contains the most obvious detail information of the source image. Finally, NSCT inverse transformation was used to obtain the resulting image.

The structure of this paper is as follows: In Section 2, we focus on the NSCT and residual network (ResNet). Section 3 introduces the image fusion method based on ResNet in the NSCT. Section 4 provides the results of comparative tests. Section 5 presents our conclusion.

II. RELATE METHODS

A. NSCT

The NSCT is a multiscale decomposition method proposed by Do and Vetterli [11] that aims to overcome the displacement invariance and pseudo-Gibbs phenomenon (shift-invariant and pseudo-Gibbs phenomena). Contour transformation does not experience translation invariance because of the upsampling and downsampling operations in the Laplacian pyramid and the directional filter bank. To preserve the direction and multiscale properties of the transformation, the Laplacian pyramid is replaced with a nonsubsampled pyramid (NSP) structure in the NSCT to preserve the multiscale properties. Moreover, the directional filter bank is replaced with nonsubsampled directional filter banks (NSDFBs) that are used to preserve directionality. After decomposition, each subband image is similar in size to the original image.

The NSCT is an image decomposition strategy that employs multiple scales, multiple directions, and shift invariance [Figure 1a]. First, an NSP is used to perform multiscale image decomposition. Each NSP decomposition can produce a low-pass component and a bandpass component and iteratively decompose low-pass components to obtain the main information in the image. The two-dimensional division of the image during NSCT decomposition is shown in Figure 1b. If the NSCT decomposition level is x, then original image can be decomposed into 1 low-pass component and x bandpass components.

A nonsubsampled directional filter bank (NSDFB) is then used to decompose the bandpass components of each scale in different directions, generating directional subbands with the same size as the source image, which is beneficial for image fusion. In the application of image fusion [12], the NSCT method can effectively retain features of the original image while exhibiting outstanding decomposition performance.

Liu *et al.* [38] compared two different decomposition strategies, NSST and NSCT, on the fusion of multifocus

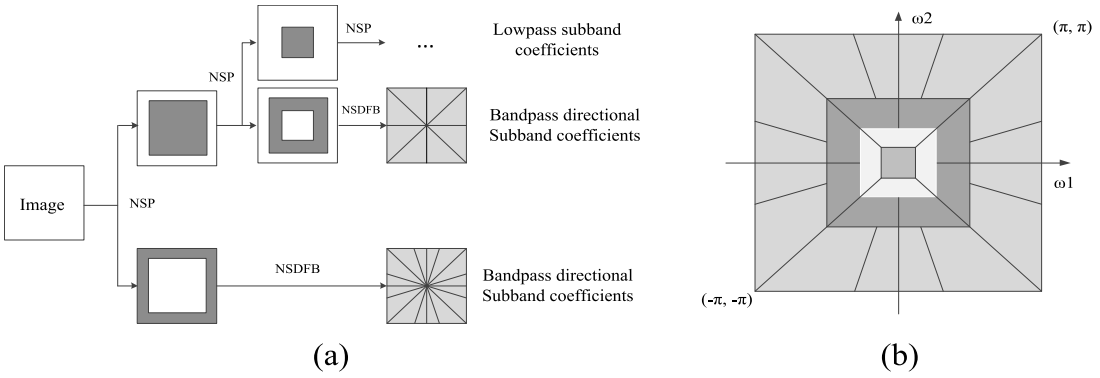


FIGURE 1. (a) NSCT decomposition framework; (b) NSCT frequency division.

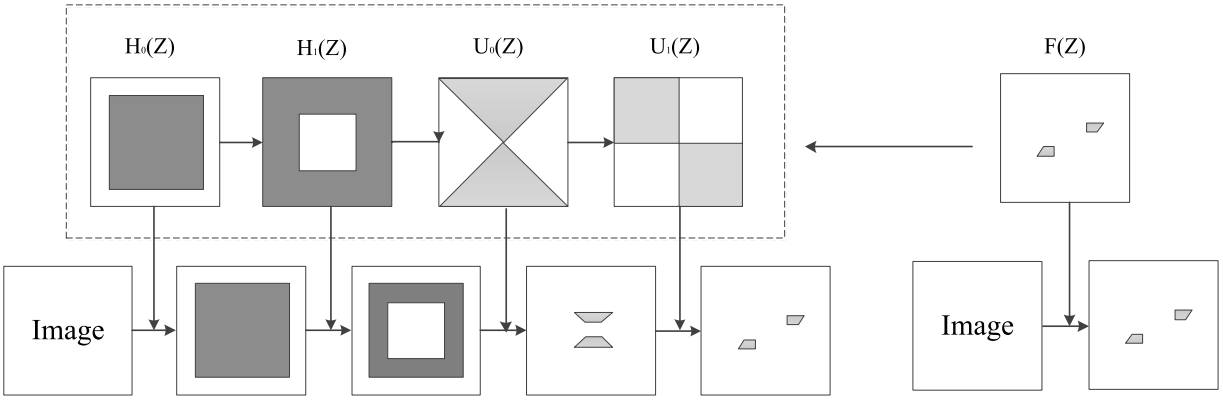


FIGURE 2. NSCT transformation process.

images, and the results showed that NSST is less time consuming. The reason for increased time during NSCT decomposition is that the source image is decomposed at multiple levels, and the image is sampled by different filters in the NSP and NSDFB stages. To reduce the time loss of the entire decomposition process, the multiple filtering processes of the image need to be integrated into a one-step filter. The essence of the NSCT decomposition method used in this study is to replace the binary tree structure in NSCT with a multichannel structure and use the reduced iterative process in the NSCT decomposition process to reduce time to a certain extent.

The NSCT transformation process is shown in Figure 2, where $H_0(Z)$ and $H_1(Z)$ represent the low-pass filter and high-pass filter in NSP, respectively, and $U_0(Z)$ and $U_1(Z)$ are the NSP Sector filters and quadrant filters. To satisfy the orthogonality and support between the filters and to maintain the size of the decomposed image, convolution calculation was carried out on the NSCT filter, and the convolution result was used as the replacement channel to replace the filter in the dashed box in Figure 2. The equation is as follows:

$$F(Z) = H_0(Z) \otimes H_1(Z) \otimes U_0(Z) \otimes U_1(Z) \tag{1}$$

When the decomposition level is x, the number of NSCT decompositions in this study is 2(x+1)−2. Compared with traditional NSCT decomposition (2(x+2)−4), the time is reduced by 50%. Therefore, the optimized NSCT decomposition is able to increase the efficiency.

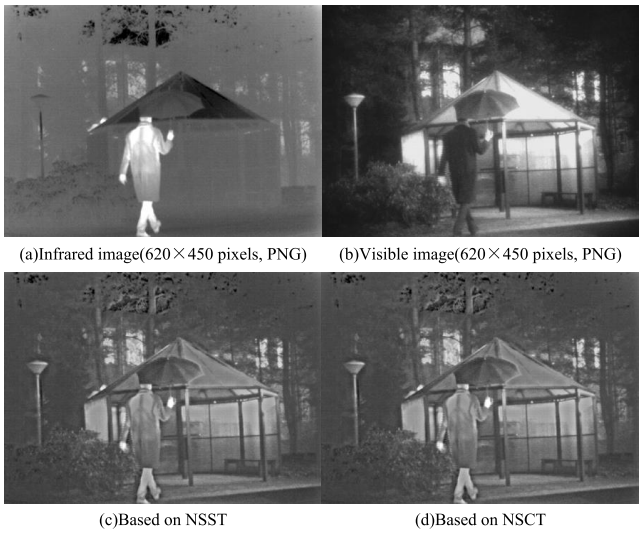


FIGURE 3. (a)-(b) Infrared and visible light images; (c)-(d) fusion results of NSST and NSCT-based method.

The proposed NSCT method was compared with the NSST method using the same parameters to verify its fusion performance. The image resolution used in the experiment was 640×450. Figure 3 shows the fusion results of the two decomposition strategies. The fusion results of the proposed method in this study have clear and detailed features. Table 1 shows the structural similarity $SSIM_a$ and the time consumption

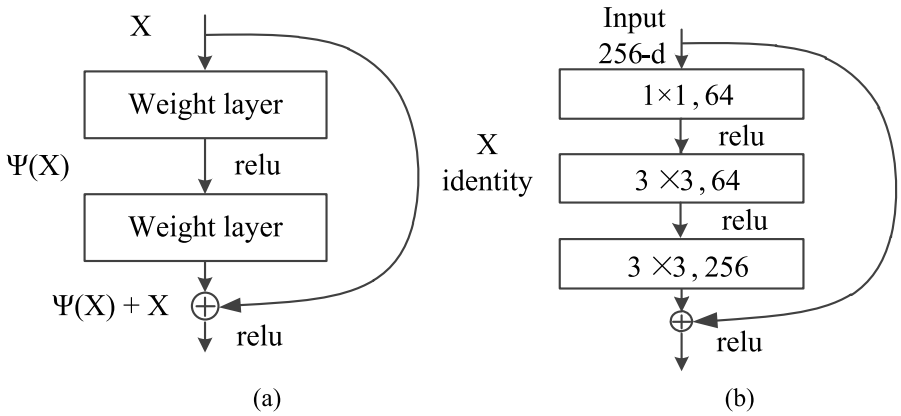


FIGURE 4. (a) Building block of residual architecture; (b) bottleneck block.

TABLE 1. Comparison of fusion results of two decomposition methods.

Method	$SSIM_a$	Time(s)
NSST-based method	0.63432	132.4345
NSCT-based method	0.68304	130.3453

of the algorithms. The results showed that the performance of the decomposition method in this study is more effective. Moreover, the time consumption of the algorithm proposed in this study was similar to that of the NSST method, indicating that the proposed method has the same efficiency as the NSST decomposition method. Therefore, collectively, the proposed NSCT-based fusion method has better performance.

B. DEEP ResNet

Studies have shown [19], [21] that the deeper the neural network layers, the more image information is available, and the richer the features. However, experiments have shown that as the neural network deepens, the optimization effect declines, while the quality of the test data and the related accuracy rate decrease. These effects occur because the deepening of the network may cause the gradient to increase or decrease exponentially. At the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Zhang and Dana [20] and others proposed a new neural network structure, ResNet, to solve the problem of gradient descent. This network structure makes use of shortcut connections and residual representations, which can be more easily optimized than previous networks and can increase the depth to improve the accuracy.

ResNet is widely used in various fields of computer vision and has achieved good results in the field of multifocus image fusion [39]. However, the deep residual neural network has not been widely considered. Therefore, in this study, the deep residual neural network ResNet152 was applied to the fusion of infrared images and visible light images.

ResNet is composed of multiple subnetwork connections [Figure 4a]. X represents the input of the subnetwork, $\Psi(X)$ represents the network operation on the two weight layers, and "relu" represents the activation correction operation on the linear unit. The final mapping result can be calculated by $\Psi(X)+X$. The residual network used in this study is composed of bottleneck blocks as shown in Figure 4b. The number of channels was reduced through a 1×1 convolutional layer. The number of channels in the middle 3×3 convolutional layer was 1/4 of the original number of channels, and the number of channels output by the middle convolutional layer remains unchanged. The third 1×1 convolutional layer is used to restore the number of channels so that the input and output channels of the entire bottleneck block are the same. The first and last two 1×1 convolutional layers effectively reduce the number of convolution parameters and the amount of calculation.

The deep residual neural network is a chain structure composed of a large number of residual blocks or bottleneck blocks. There are details regarding the training process of the residual network ResNet50 in [39]. The ResNet152 used in this study was not simply an increase in the number of layers on the basis of ResNet50; there was additional processing in each convolutional block. Figure 5 shows the structure of ResNet152 and the processing of low-pass components of infrared and visible images. The first convolutional layer Conv1 of ResNet152 is a 7×7 convolutional layer. Conv2, Conv3, Conv4, and Conv5 are convolutional layers composed of bottleneck blocks shown in Figure 4b. There are 50 bottleneck blocks in total, which is 34 more than ResNet50. According to the principle of residual network, the increased convolution operation will not cause information loss and can extract deep information in the images. In this study, the deep feature extraction capability of ResNet152 was used to extract the main features of infrared and visible images, which were used to guide the fusion of low-pass subbands, thereby preserving the overall information and structural features of the source image to the greatest extent. The results of Conv5 were used in subsequent processing, which avoids the complicated

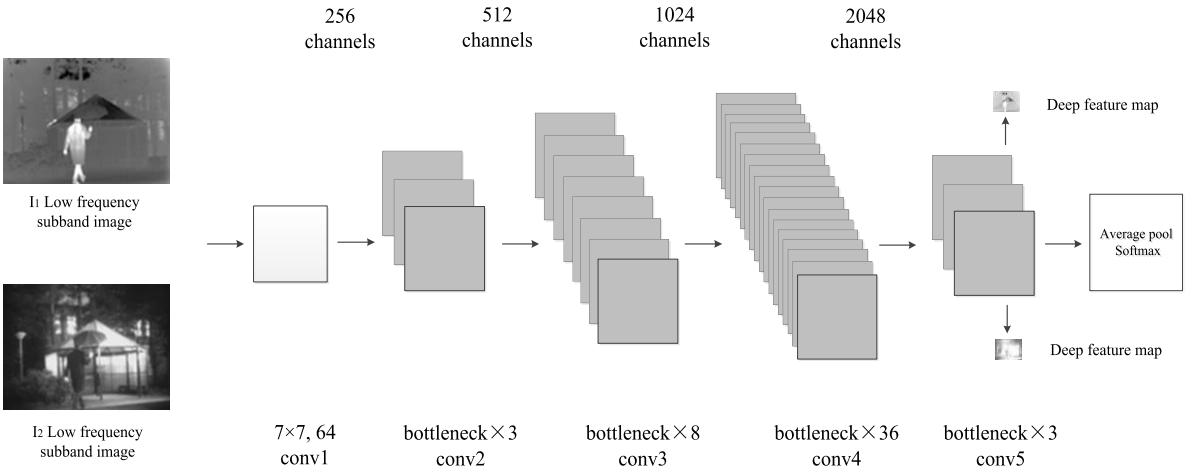


FIGURE 5. ResNet152 structure and extraction of deep feature map.

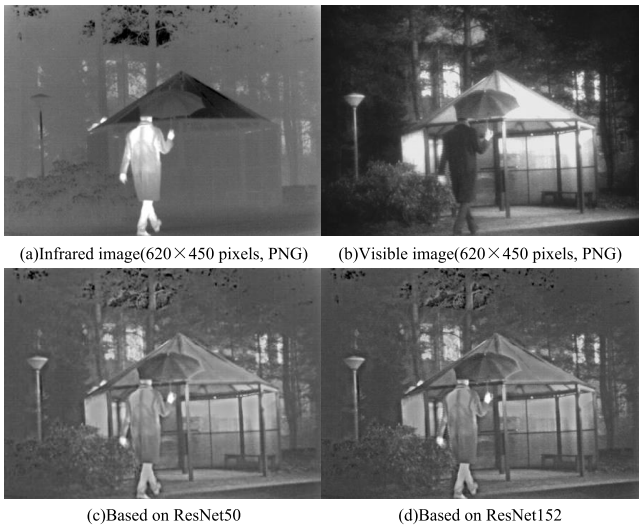


FIGURE 6. (a)-(b) Infrared and visible light images (c)-(d) fusion results of ResNet50 and ResNet152.

calculations associated with network training and enhances the flexibility of the algorithm.

Figure 6 is a comparison of the fusion results of ResNet50 and ResNet152 with the same NSCT decomposition parameters. The result of ResNet152 showed clear and detailed features. Table 2 shows the structural similarity $SSIM_a$ and SCD of the fusion results. The results show that the performance of ResNet152 is better when fusing low-pass components.

III. IMAGE FUSION METHOD BASED ON NSCT-ResNet

This paper assumes the use of K=2 preregistered source images. This method is also applicable to situations in which K>2 [34]. The source image is expressed as I_k, where k∈{1, 2}, in which 1 and 2 represent the infrared image and the visible light image, respectively. The NSCT multiscale decomposition method is used to decompose source images I_1 and I_2 into multiscale, multidirectional low-pass component groups (I_1^d, I_2^d) and bandpass component groups (I_1^b, I_2^b).

TABLE 2. Comparison of fusion results based on different residual networks.

Method	Structural similarity	SCD
ResNet50-based method	0.66234	1.57632
ResNet152-based method	0.68304	1.60938

The low-pass component group obtained after decomposition contains the main information in the source image, while the bandpass component reflecting the detailed information contains contour features from different directions and scales. Therefore, the fusion strategy of low-pass subband groups (I_1^d, I_2^d) is the key to fusing infrared and visible light images. Here, the ResNet is used to guide the method for feature fusion of the deepest layers, which involves fusing the main information of the two images. The bandpass component groups (I_1^b, I_2^b) mainly consist of detailed information from the image, and the weighted fusion is performed by taking a large modulus value, which can preserve the texture contour of the source image in the fusion image to the greatest extent. Finally, fusion image F is reconstructed by fusing fundamental part and contour part using the inverse NSCT. The framework of the fusion method proposed in this paper is shown in Figure 7.

A. FUSION OF LOW-FREQUENCY SUBBAND COEFFICIENTS

For low-pass components and, this paper proposes a strategy of using deep residual neural networks to guide the image feature fusion process (Figure 8). ResNet152 [23] is used to extract feature maps for low-pass components in infrared and visible light images. Weight mapping is then obtained through feature mapping operations in formula (3) and formula (4). Finally, fundamental part F_d is reconstructed using weight mapping and low-pass components. In x-level decomposition, the ideal low-pass subband supporting frequency

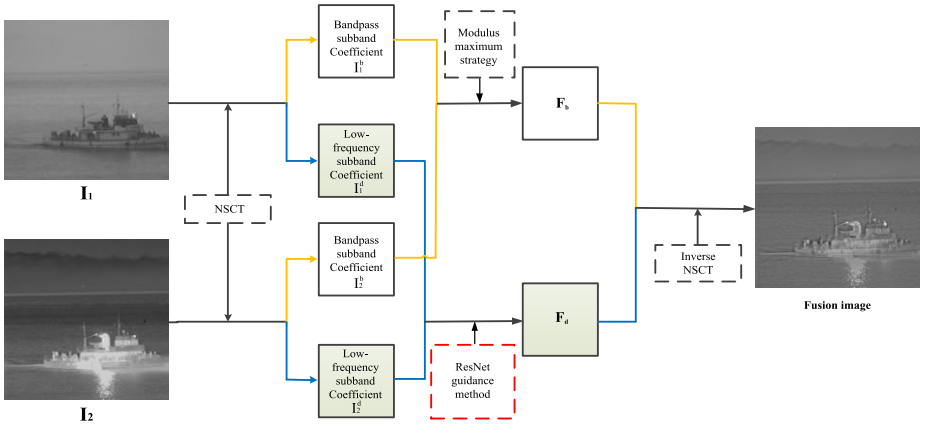


FIGURE 7. Proposed framework of image fusion.

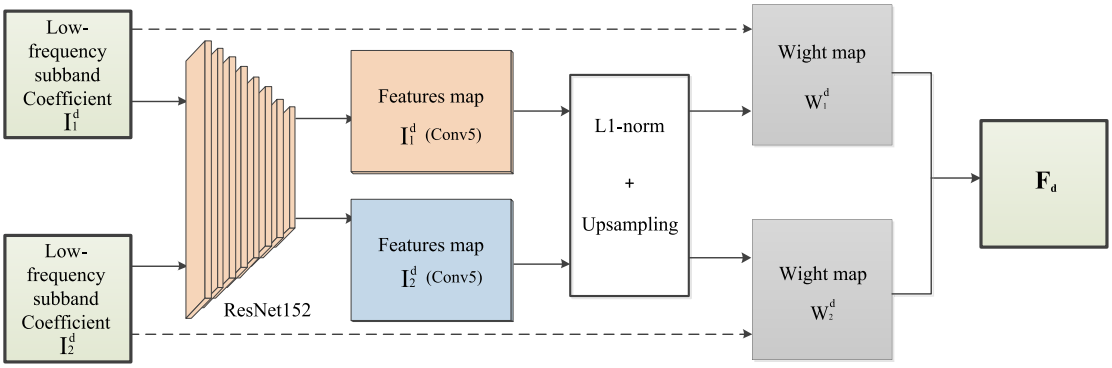


FIGURE 8. Framework of ResNet guidance method.

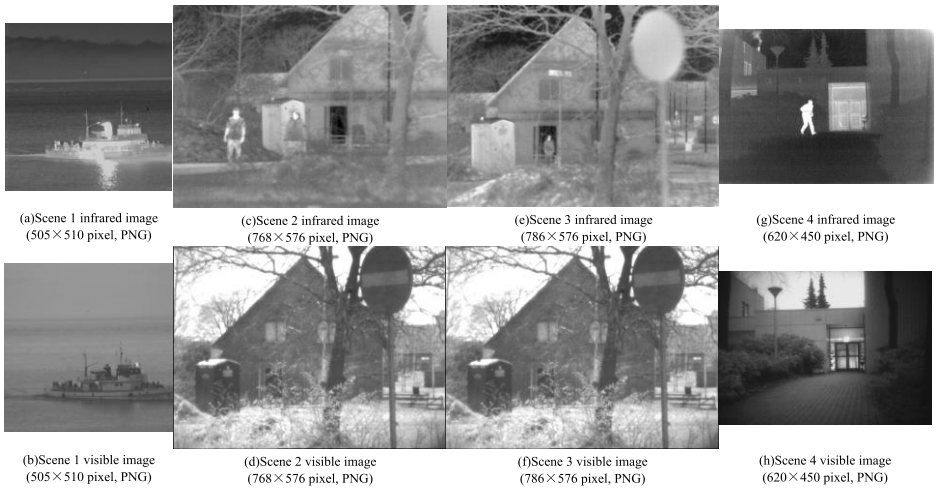


FIGURE 9. Four pairs of source images. First row is infrared images, and second row is visible images.

domain is $[-\pi/2^x, \pi/2^x]^2$, and that of the ideal high-pass subband is $[-\pi/2^{x-1}, \pi/2^{x-1}]^2/[-\pi/2^x, \pi/2^x]^2$. The range of the high-pass subband is the complement of the low-pass subband [40]. The low-pass subband contains the global features and content information of the source image, and the high-pass subband is the detailed information of the source image. To obtain more image features, neural networks were used to the fusion of low-pass subbands [35]. Compared to the fusion of high-pass subbands that contains image details and edges [41], the fusion of low-frequency component based on deep residual network can obtain a more complete image contour.

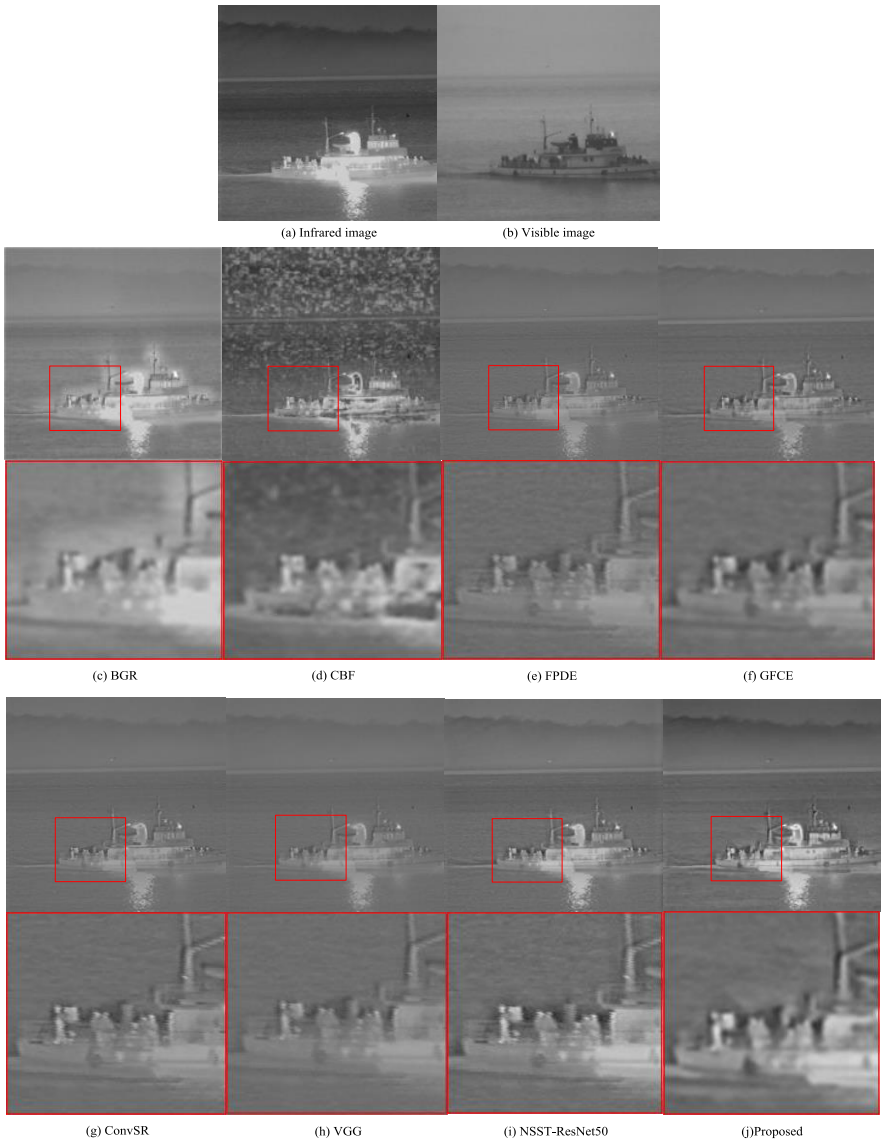


FIGURE 10. Results for "ship" images. (a) Infrared image; (b) visible image; (c) BGR result; (d) CBF result; (e) FPDE result; (f) GFCE result; (g) ConvSR result; (h) VGG result; (i) NSST-ResNet50 result; (j) the result for the proposed method.

ResNet152 is a pretrained network consisting of five convolutional blocks (conv1, conv2, conv3, conv4, conv5) and containing a total of 152 weight layers. Thus, the deep features output by the i ∈ {1, 2, 3, 4, 5}th convolution block can be expressed as follows:

$$I_k^{i,m} = \psi(I_k^d) \tag{2}$$

where represents a convolutional block of the ResNet, and m represents the number of channels in each feature layer, k∈{1, 2}.

Deep features $I_k^{i,m}$ are obtained, which initially undergo L1 regularization to obtain the initial weight mapping. The formula is as follows:

$$M_k^{i,*} = \frac{\sum_{p=x-\eta}^{x+\eta} \sum_{q=y-\eta}^{y+\eta} ||I_k^{i,m}(p,q)||_1}{\eta \times (2\eta+1)} \tag{3}$$

In the above formula, $\eta = 2$ is a matrix sparse operation with a step size of 5 × 5 on the feature map $I_k^{i,m}$ [24].

After two initial weight mappings $M_1^{i,*}$ and $M_2^{i,*}$ are obtained through the two low-pass components I_1^d and I_2^d, they are upsampled using bicubic interpolation and the initial weight mapping is adjusted to the source image size. The final weight mapping is calculated using the following formula:

$$W_k^i(x,y) = \frac{M_k^{i,*}(x,y)}{M_1^{i,*}(x,y) + M_2^{i,*}(x,y)} \tag{4}$$

The fusion result of low-pass components can be ultimately obtained using the following formula:

$$F_d(x,y) = \sum_{k=1}^{2} W_k^i(x,y) I_i^d(x,y) \tag{5}$$

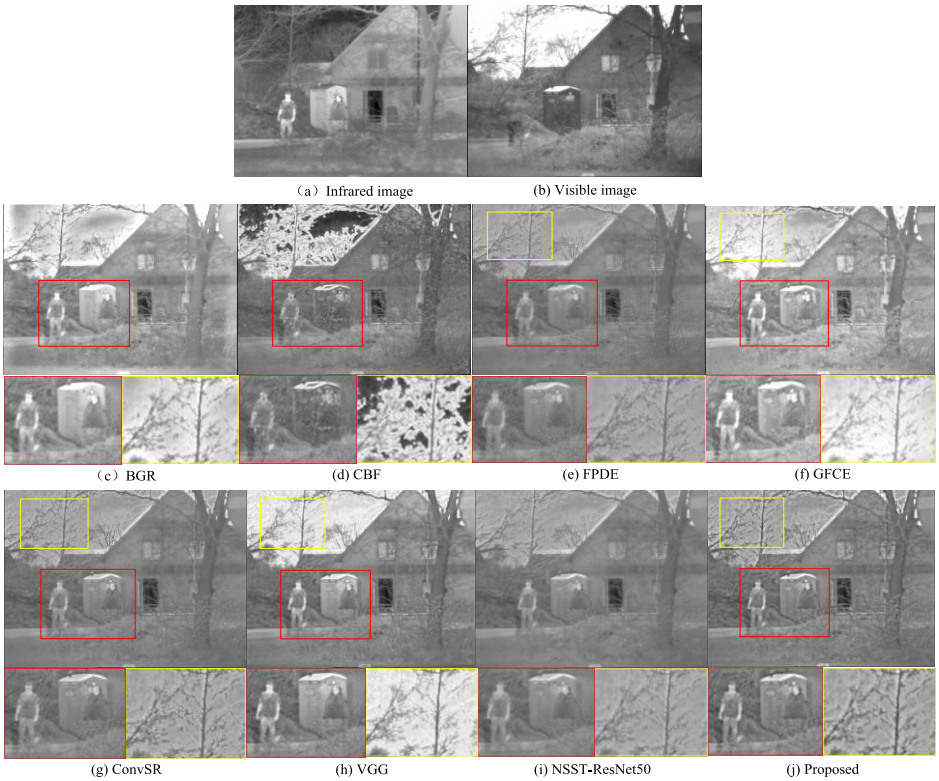


FIGURE 11. Results for "two person" images. (a) Infrared image; (b) visible image; (c) BGR result; (d) CBF result; (e) FPDE result; (f) GFCE result; (g) ConvSR result; (h)VGG result; (i) NSST-ResNet50 result; (j) the result for the proposed method.

B. FUSION OF BANDPASS DIRECTIONAL SUBBAND COEFFICIENTS

The part of the bandpass component separated from the source image contains contour and texture information. In this method, the maximum modulus method has been chosen to fuse the bandpass components. The maximum value of the bandpass component group is calculated using formula (6). Formula (7) is then used to fuse the bandpass components of the infrared and visible images to preserve as much edge and contour information as possible in the source image. The calculation equation is as follows:

$$W_{\max} = \max(|I_1^b|, |I_2^b|) \tag{6}$$

$$F_b(x, y) = W_{\max} I_1^b(x, y) + W_{\max} I_2^b(x, y) \tag{7}$$

In the formula, W_{max} is the result of taking the maximum value after taking the modulus of the bandpass component, and (x, y) represents the pixel positions of I_1^b, I_2^b, and F_b.

C. FUSION STEP

The process of this fusion method is as follows:

1) Preregistered source images I_1 and I_2 are subjected to multiscale decomposition through the NSCT to obtain a series of bandpass component images I_1^b and I_2^b and a set of low-pass component images I_1^d and I_2^d.

2) The feature layer of low-pass components is extracted using ResNet152, and a feature map is calculated. By guiding the fusion process for low-pass component images I_1^d and I_2^d, fusion result F_d for the low-pass components containing the main content of the image is obtained.

3) By taking the maximum modulus to fuse bandpass component images I_1^b and I_2^b, fusion result F_b for the bandpass component is obtained, which contains the edge information in the image.

4) Finally, F_b and F_d are transformed using the inverse NSCT to obtain the final fusion image.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

The purpose of this section is to use experiments to verify the proposed fusion strategy based on subjective and objective standards and to compare the results with previous research results.

A. EXPERIMENTAL SETTINGS

The 21 sets of infrared and visible light images used in the experiment are all preregistered images provided by Toet [25] and others. Four of sets of images were selected for subjective evaluation, as shown in Figure 9. The image resolution is 505 × 510, 768 × 576, and 620 × 450, respectively. All experimental codes were implemented in MATLAB R2017a. The neural network operation uses MATLAB toolbox MatConvNet, which provides a friendly and efficient environment for researchers in the field of deep learning and image processing.

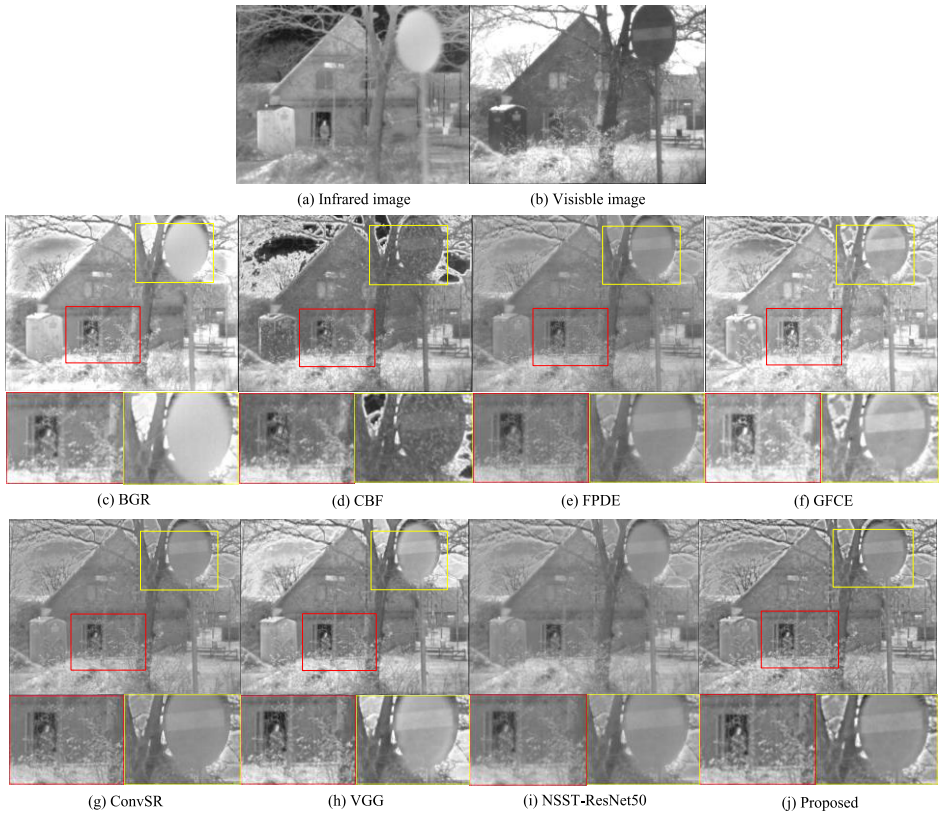


FIGURE 12. Results for "two person" images. (a) Infrared image; (b) visible image; (c) BGR result; (d) CBF result; (e) FPDE result; (f) GFCE result; (g) ConvSR result; (h) VGG result; (i) NSST-ResNet50 result; (j) the result for the proposed method.

To fully consider the decomposition speed and imaging quality [22], the initialized NSCT subband decomposition parameters were "9/7" and "pkva," representing the pyramid filter and directional filter, respectively. The decomposition level was 4, and the number of directions from coarse to fine scale was set to [0, 1, 3, 4], which represent the number of decompositions and the number of directions at each level. The low-pass component image is convolved through ResNet152, and the feature results from the last convolution layer of the fifth convolution block are extracted for the feature mapping calculation.

For comparison and analysis, several classic fusion methods along with methods based on convolutional neural networks were selected for the same experiment. The selected comparison methods include the cross bilateral filter (CBF) [2], infrared feature lifting and visual information preservation (BGR) [26], guided filter-based content enhancement (GFCE) [27], a method based on fourth-order partial differential equations (FPDE) [28], ConvSR [29], VGG [34], and NSST-ResNet50 [39].

B. SUBJECTIVE EVALUATION

The results obtained by comparing the fusion method proposed in this paper with five existing fusion methods are shown in Figures 10-13. Figure 10 is a scene of a ship on the sea with less detailed background information. Figures 11 and 12 have more detailed information with scenes of houses, trees, and people. In Figure 13, there is less content information in the visible image. Infrared images can effectively identify targets such as people obscured by trees or smoke, but they contain less background information. In visible images, target recognition is relatively poor, but there is rich background information and edge texture information.

Figure 10 indicates that the six fusion methods compared here were able to successfully fuse the visible and infrared images. However, the fusion result of CBF contains a substantial amount of noise (Figure 10d). The BGR fusion image has a low contrast and loses a substantial amount of background information (Figure 10c). The fusion result of GFCE is blurry around the main target, and artifacts emerge. The FPDE result has less texture information in the background (Figure 10e and 10f). The fusion result of ConvSR shows ripples on the edges of the people and boats in the image (Figure 10g). Figure 10h is the fusion result of a deep learning method where the deep network VGG19 was used to guide the fusion of low-pass components. The results showed that the overall image was dark, with poor visual effect. Figure 10i is the fusion result of NSST-ResNet50. The details of the characters were not as clear as the results of the method in this paper. Figure 10j is the fusion image obtained using the method in this article. It contains less artificial noise overall and retains more background information. Observing

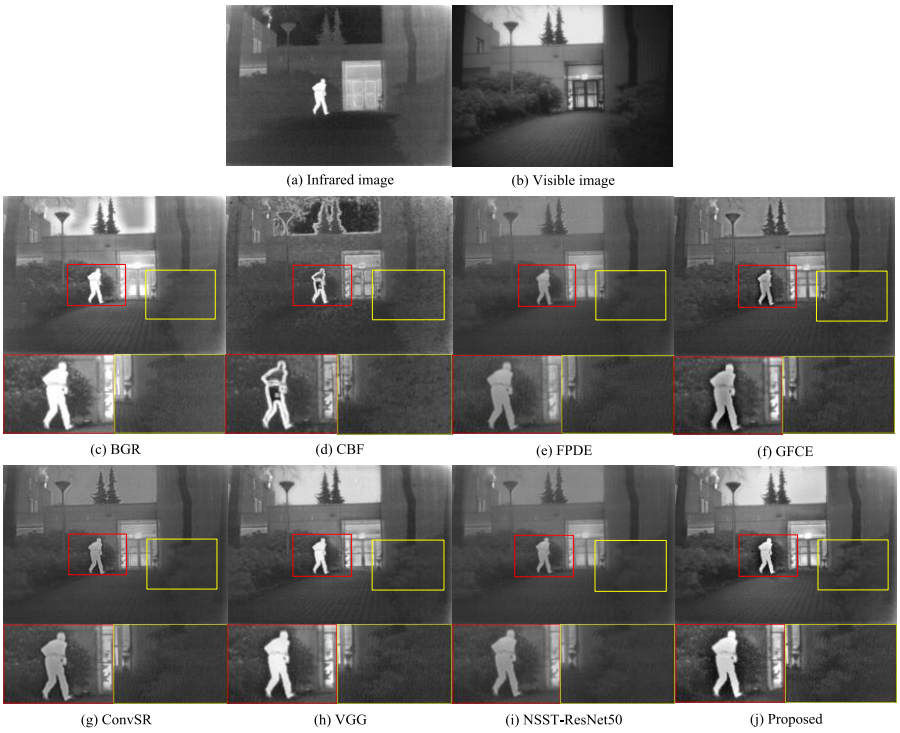


FIGURE 13. Results for "two person" images. (a) Infrared image; (b) visible image; (c) BGR result; (d) CBF result; (e) FPDE result; (f) GFCE result; (g) ConvSR result; (h) VGG result; (i) NSST-ResNet50 result; (j) the result for the proposed method.

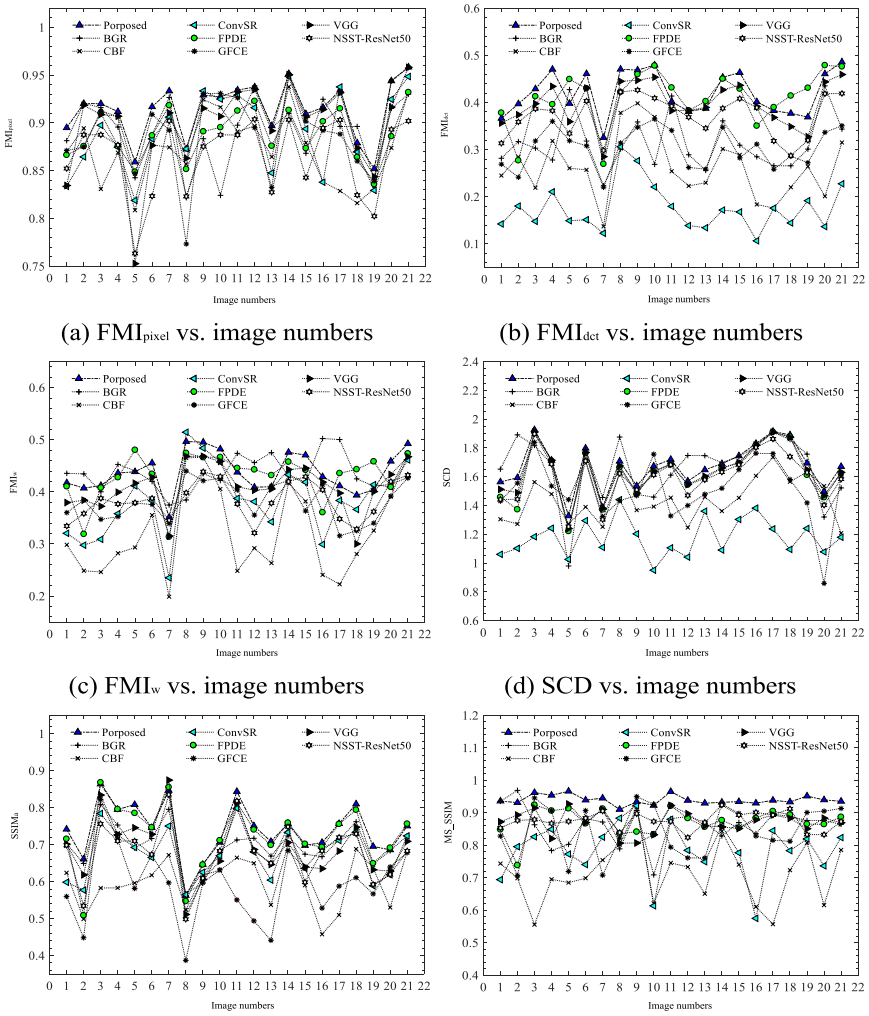
the area marked in the red window, the algorithm in this paper is shown to retain more detailed information, and the edge texture produced in the image is clearer than that obtained with the other algorithms.

Figure 11 and Figure 12 are fusion comparisons in similar scenarios. From the results, the image background obtained by the CBF method is blurred and has serious artifacts. The results obtained by the BGR and GFCE methods have less edge texture information, and artifacts are present in the background. Although there is no artifact problem in the FPDE results, the main target of the obtained image is blurred, and the edges are smoother than those in the results from the method proposed in this paper. The fusion result of the VGG method is the background is too bright, which is the result of inheriting too much infrared image content, and there are artifacts at the edges of the object. The NSST-ResNet50 method is similar to the method in this paper, but the fusion result is slightly blurred. This difference is because the method in this paper uses more feature information of the source image to guide the fusion process. Comparing the window areas marked in red and yellow in Figure 11, the edges of the people in Figure 11j are clearer, and the texture of the tree branches is clearer. In making a comparison with Figure 12, the contrast is even more apparent. The method used in this paper obtains even more prominent target information, including richer roads and shrubs in the background information and clearer edge texture information.

The visible image in Figure 13 has less information, and the figures in the picture cannot be clearly identified (Figure 13b). The BGR method has artifacts on the edges (Figure 13c), and there is obvious noise in the red marked box, while there is relatively little background information such as trees. A substantial amount of noise appears in the image in Figure 13d. Moreover, the feature information is chaotic, and there is a significant artifact issue. The remaining algorithms have ideal fusion effects. However, after comparing the details in the marked boxes, the method used in this paper is found to have a clearer texture for the tree trunks in the background, and the edges of the people are more distinct. When evaluated subjectively, the fusion method proposed in this paper retains more thermal radiation intensity information in the infrared images and more detailed texture information in the visible light images. The results from the proposed method also contain less artificial noise and fewer artifacts, indicating that they are more natural.

C. OBJECTIVE EVALUATION

Six typical image quality evaluation functions were selected to evaluate the performance of image fusion. The selected functions include the sum of the correlation of differences (SCD) [30], pixel mutual information (FMI_{pixel}) [31], discrete cosine mutual information (FMI_{dct}) [32], wavelet feature mutual information (FMI_w) [32], modified without reference image structural similarity ($SSIM_a$) [32], and a new nonreference image fusion performance metric (MS_SSIM) [33]. The value represents the ability of the fusion method to preserve the structural information of the source image. The increase in the function value of the six image quality evaluations



(a) FMI_{pixel} vs. image numbers

(b) FMI_{dct} vs. image numbers

(c) FMI_{w} vs. image numbers

(d) SCD vs. image numbers

(e) $SSIM_{a}$ vs. image numbers

(f) MS_SSIM vs. image numbers

FIGURE 14. **Quantitative comparisons of six metrics for all fused images. (a) FM_{Ipixel} vs. image numbers. (b) FM_{Idct} vs. image numbers. (c) FM_{Iw} vs. image numbers. (d) SCD vs. image numbers. (e) SSI_{Ma} vs. image numbers. (f) MS_SSIM vs. image numbers.**

represents the improvement in the performance of the fusion method, thus indicating an optimized fusion effect.

Figure 14 shows a comparison of the evaluation functions for the six comparison algorithms with 21 sets of fused images. The comparison results indicate that the evaluation function value of the algorithm proposed in this paper is more effective than that of the other comparison algorithms for most of the fusion results. To eliminate calculation errors between different image scenes and algorithms, the comparison idea proposed by Li [34] is used to compare the average value of the evaluation quality function values for the 21 images. The calculation method is as follows:

$$Value_{favg} = \frac{\sum_{n=1\sim21} Value_f}{21} \tag{8}$$

In the above formula, $Value_{favg}$ represents the average value of the evaluation functions FMI_{pixel}, FMI_{dct}, FMI_{w}, SCD, $SSIM_{a}$, and MS_SSIM under the different algorithms for the 21 evaluation images.

In Table 3, the maximum value from the average values of all evaluation functions is shown in bold. Based on the evaluation function results in the table, the fusion method proposed in this paper produced results that are better than those of the other fusion methods. The greatest evaluation function value is calculated among the five comparison algorithms using formula (9). Moreover, formula (10) is used to calculate the relative amount of improvement in each evaluation function value for the algorithm in this paper:

$$Max_{value} = \max(Value_{favg}) \tag{9}$$

$$\mathrm{Re} = \frac{Value_{Porposed} - \mathrm{Max}_{value}}{Max_{value}} \tag{10}$$

The calculation results show that the evaluation function value FMI_{pixel} of the algorithm in this paper increased by 2.74%, FMI_{dct} increased by 33.7%, FMI_{w} increased

TABLE 3. Average values of FMIpixel, FMIdct, FMIw, SCD, SSIMa, and MS_SSIM for 21 fused image.

Method	BGR	CBF	FPDE	GFCE	ConvSR	VGG	NSST-ResNet50	Propose
FMI_{pixel}	0.89115	0.87203	0.89017	0.8922	0.88693	0.89752	0.86498	**0.91428**
FMI_{dct}	0.31445	0.26309	0.17529	0.30339	0.20833	0.39723	0.37098	**0.42054**
FMI_{w}	0.43609	0.3235	0.38415	0.38062	0.38333	0.41019	0.38361	**0.42733**
SCD	1.65488	1.38962	1.14903	1.51735	1.14225	1.63624	1.59757	**1.67138**
$SSIM_a$	0.70037	0.59956	0.70786	0.59187	0.70893	0.69188	0.67781	**0.74004**
MS_SSIM	0.85713	0.7139	0.78837	0.8446	0.78544	0.87144	0.87517	**0.93947**

by 0.28%, SCD increased by 1.01%, $SSIM_a$ increased by 4.38%, and MS_SSIM increased by 9.61%. From an objective evaluation, the method in this paper has a better fusion performance.

V. CONCLUSION

Based on an optimized multiscale NSCT, this paper proposed a new method for the fusion of infrared and visible images. Because the deep residual neural network ResNet152 has powerful feature extraction capabilities, it can be used to guide the fusion process of low-pass components. This method has the advantages of multiscale decomposition in the NSCT. In addition, the feature guidance and modulus maximum methods can also be used to retain rich contour texture in the source image, which provides clearer edge information in the resulting image. In comparison with several classic fusion algorithms, the proposed fusion strategy was evaluated through subjective and objective evaluation methods. The experimental results show that the fusion image of this algorithm has a clearer visual expression and fewer noise artifacts.

In terms of objective evaluation, the results of several typical image quality evaluation functions were compared. The average evaluation function of the fusion image of the algorithm in this paper is greater than that of the other comparison algorithms. The FMI_{dct} value of the evaluation function is 33.7% higher than the next best comparison algorithm, indicating that the fusion image obtained by this algorithm has better image quality.

ACKNOWLEDGMENT

The authors would like to thank the editor for their support and the anonymous reviewers for their valuable suggestions.

REFERENCES

[1] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, and K. He, "A survey of infrared and visual image fusion methods," *Infr. Phys. Technol.*, vol. 85, pp. 478–501, Sep. 2017.

[2] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, Jul. 2015.

[3] D. P. Bavirisetti and R. Dhuli, "Multi sensor image fusion using saliency map detection," *Int. Rev. Comput. Softw.*, vol. 10, no. 7, p. 757, Jul. 2015.

[4] Q. Guihong, Z. Dali, and Y. Pingfan, "Medical image fusion by wavelet transform modulus maxima," *Opt. Exp.*, vol. 9, no. 4, p. 184, Aug. 2001.

[5] D. M. Bulanon, T. F. Burks, and V. Alchanatis, "Image fusion of visible and thermal images for fruit detection," *Biosyst. Eng.*, vol. 103, no. 1, pp. 12–22, May 2009.

[6] X. Yu, J. Ren, Q. Chen, and X. Sui, "A false color image fusion method based on multi-resolution color transfer in normalization YCC space," *Optik*, vol. 125, no. 20, pp. 6010–6016, Oct. 2014.

[7] 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.

[8] H. Jin and Y. Wang, "A fusion method for visible and infrared images based on contrast pyramid with teaching learning based optimization," *Infr. Phys. Technol.*, vol. 64, pp. 134–142, May 2014.

[9] Z. Jing, "Image fusion based on an expectation maximization algorithm," *Opt. Eng.*, vol. 44, no. 7, Jul. 2005, Art. no. 077001.

[10] P. Chai, X. Luo, and Z. Zhang, "Image fusion using quaternion wavelet transform and multiple features," *IEEE Access*, vol. 5, pp. 6724–6734, 2017.

[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] A. L. Da Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. Image Process.*, vol. 15, no. 10, pp. 3089–3101, Oct. 2006.

[13] J. Adu, M. Wang, Z. Wu, and J. Hu, "Infrared image and visible light image fusion based on nonsubsampled contourlet transform and the gradient of uniformity," *Int. J. Adv. Comput. Technol.*, vol. 4, no. 5, pp. 114–121, Mar. 2012.

[14] T. Xiang, L. Yan, and R. Gao, "A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking PCNN in NSCT domain," *Infr. Phys. Technol.*, vol. 69, pp. 53–61, Mar. 2015.

[15] W. Kong, L. Zhang, and Y. Lei, "Novel fusion method for visible light and infrared images based on NSST–SF–PCNN," *Infr. Phys. Technol.*, vol. 65, pp. 103–112, Jul. 2014.

[16] R. Eckhorn, H. J. Reitbock, M. Arndt, and P. Dicke, "A neural network for feature linking via synchronous activity: Results from cat visual cortex and from simulations," *Can. J. Microbiol.*, vol. 46, no. 8, pp. 759–763, 1989.

[17] Y. Chen and Z. Qin, "PCNN-based image fusion in compressed domain," *Math. Problems Eng.*, vol. 2015, pp. 1–9, Jan. 2015.

[18] W. W. Kong, Y. J. Lei, Y. Lei, and S. Lu, "Image fusion technique based on non-subsampled contourlet transform and adaptive unit-fast-linking pulse-coupled neural network," *IET Image Process.*, vol. 5, no. 2, pp. 113–121, Mar. 2011.

[19] J. Zhong, B. Yang, Y. Li, F. Zhong, and Z. Chen, "Image fusion and super-resolution with convolutional neural network," in *Proc. Chin. Conf. Pattern Recognit.*, Nov. 2016, pp. 78–88.

[20] H. Zhang and K. Dana, "Multi-style generative network for real-time transfer," 2017, *arXiv:1703.06953*. [Online]. Available: https://arxiv.org/abs/1703.06953

[21] X. Tao, H. Gao, R. Liao, J. Wang, and J. Jia, "Detail-revealing deep video super-resolution," *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 4472–4480.
[22] X.-B. Qu, J.-W. Yan, H.-Z. Xiao, and Z.-Q. Zhu, "Image fusion algorithm based on spatial frequency-motivated pulse coupled neural networks in nonsubsampled contourlet transform domain," *Acta Automatica Sinica*, vol. 34, no. 12, pp. 1508–1514, Dec. 2008.
[23] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, and A. C. Berg, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, 2015.
[24] 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.
[25] A. Toet. (2014). *TNO Image Fusion Dataset*. Figshare. [Online]. Available: https://figshare.com/articles/TN Image Fusion Dataset/1008029.
[26] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infr. Phys. Technol.*, vol. 83, pp. 227–237, Jun. 2017.
[27] Z. Zhou, M. Dong, X. Xie, and Z. Gao, "Fusion of infrared and visible images for night-vision context enhancement," *Appl. Opt.*, vol. 55, no. 23, pp. 6480–6490, Aug. 2016.
[28] 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*, Jul. 2017, pp. 10–13.
[29] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
[30] 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, Dec. 2015.
[31] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE 8th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2014, pp. 1–3.
[32] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Trans. Image Process.*, vol. 24, no. 11, pp. 3345–3356, Nov. 2015.
[33] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[34] 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.
[35] X. Jin, G. Chen, J. Hou, Q. Jiang, D. Zhou, and S. Yao, "Multimodal sensor medical image fusion based on nonsubsampled shearlet transform and S-PCNNs in HSV space," *Signal Process.*, vol. 153, pp. 379–395, Dec. 2018.
[36] X. Huang, G. Qi, H. Wei, Y. Chai, and J. Sim, "A novel infrared and visible image information fusion method based on phase congruency and image entropy," *Entropy*, vol. 21, no. 12, p. 1135, Nov. 2019.
[37] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–14, 2021.
[38] S. Liu, J. Wang, Y. Lu, H. Li, J. Zhao, and Z. Zhu, "Multi-focus image fusion based on adaptive dual-channel spiking cortical model in non-subsampled shearlet domain," *IEEE Access*, vol. 7, pp. 56367–56388, 2019.
[39] S. Liu, J. Wang, Y. Lu, S. Hu, X. Ma, and Y. Wu, "Multi-focus image fusion based on residual network in non-subsampled shearlet domain," *IEEE Access*, vol. 7, pp. 152043–152063, 2019.
[40] Y. Li, Y. Sun, X. Huang, G. Qi, M. Zheng, and Z. Zhu, "An image fusion method based on sparse representation and sum modified-Laplacian in NSCT domain," *Entropy*, vol. 20, no. 7, p. 522, Jul. 2018.
[41] B. Li, H. Peng, and J. Wang, "A novel fusion method based on dynamic threshold neural P systems and nonsubsampled contourlet transform for multi-modality medical images," *Signal Process.*, vol. 178, Jan. 2021, Art. no. 107793.

CE GAO was born in 1981. He received the bachelor's degree in computer science and technology from Tianjin University, in 2004, and the master's degree in optical engineering from the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, in 2011. He is currently pursuing the M.S. degree. He is also an Associate Research Fellow. His main research interests include computer applications, image processing, and optoelectronic measurement and control technology.

DONGHAO QI was born in 1994. He received the master's degree in mechanical and electronic engineering from Xidian University, in 2020. He is currently a Research Intern. His main research interest includes technology related to image processing.

YANCHAO ZHANG was born in 1985. She received the bachelor's and master's degrees in engineering from the School of Precision Instrument and Opto-electronics Engineering, Tianjin University, in 2007 and 2009, respectively, and the Ph.D. degree in optical engineering from the University of Chinese Academy of Sciences, in 2015. She is currently an Associate Research Fellow. Her main research interests include image processing and photoelectronic measurement and control technology.

CONGCONG SONG was born in Changchun, Jilin, China, in 1994. She received the bachelor's degree in communication engineering from the Changchun University of Science and Technology, in 2017, and the master's degree in electronics and communication engineering from Jilin University, in 2020. She currently works with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. Her current main research interests include image fusion and deep learning.

YI YU was born in 1979. He received the Ph.D. degree in optical engineering from the Changchun Institute of Technology, Changchun, China, in 2016. He is currently a Professor with the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun. His research interests include optoelectronic measurement, image/video processing, computer vision, and machine learning.