Hindawi
Journal of Sensors
Volume 2022, Article ID 9669142, 24 pages
https://doi.org/10.1155/2022/9669142

Research Article

Infrared and Visible Image Fusion in a Multilevel Low-Rank Decomposition Framework Based on Guided Filtering and Feature Extraction*

Chao Fang,[1] **Xin Feng**,[1,2] **Haifeng Gong**,[2] **and Xicheng Lou**[1]

[1]*School of Mechanical Engineering, Key Laboratory of Manufacturing Equipment Mechanism Design and Control of Chongqing, Chongqing Technology and Business University, Chongqing 400067, China*
[2]*Engineering Research Centre for Waste Oil Recovery Technology and Equipment of Ministry of Education, Chongqing Technology and Business University, Chongqing 400067, China*

Correspondence should be addressed to Xin Feng; 149495263@qq.com

Received 3 December 2021; Revised 9 March 2022; Accepted 12 April 2022; Published 30 April 2022

Academic Editor: Mu Zhou

A novel infrared and visible image fusion method in a multilevel low-rank decomposition framework based on guided filtering and feature extraction is proposed to address the lack of edge information and blurred details in fused images. Based on multilevel low-rank decomposition, the fusion strategy of base part and detail contents has been improved. Firstly, the source infrared and visible images are decomposed to the base part coefficients and n-level detail content coefficients by multilevel low-rank decomposition. Secondly, the base part coefficients are learned by the VGG-19 network to get the weight map, and then, the improved weight map is obtained by guided filtering, and the coefficients of the base part are fused to acquire the fused base part coefficients. The n-level detail content coefficients are fused using the rule of dynamic level measurement with maximum value and then reconstructed to obtain the final fused detail content coefficients. Finally, the fused base part and detailed content information are superimposed to get the final fusion result. The results show that the fusion algorithm can effectively preserve the edge and detail features of the source image. Compared with other state-of-the-art fusion methods, the proposed method performs better in objective assessment and visual quality. The average value of evaluation metrics *EN* and *MI* have been improved by 0.5337 and 1.0673 on the six pair images.

1. Introduction

Image fusion is an enhancement technology. Image fusion is aimed at combining different images to generate a steady and informative image, which can facilitate subsequent processing and help in decision making. Recently, many fusion methods have been proposed to fuse the features in infrared and visible images into a single image [1]. The visible images usually have high spatial resolution and large detail contrast but are easily affected by harsh environments and climatic conditions. Infrared images depict the temperature or radiation of an object, which is not easily affected by the environment and climatic conditions. However, infrared images contain a few shortcomings, such as inconspicuous texture details and poor resolution. So we can make full use of different modalities to convey complementary information. It applies in a lot of applications, such as surveillance [2], object detection, and target recognition [3–5]. The methods of multiscale transforms [6, 7] and representation learning [8] are generally used in image fusion field.

The traditional multiscale transformation method decomposes the source images into base parts and detail content at distinct dimensions. The base part mainly represents the contours and edges of the source image, and the detail content contains more detailed texture information. The base part and detail content are fused according to predefined rules in the transform domain. Then, the final fused image is obtained through inverse multiscale transformation [9]. There are some typical algorithms, such as the discrete wavelet transform (DWT) [10], contourlet transform [11,

12], shearlet transforms [13], and multilevel decomposition latent low-rank representation(MDLatLRR) [14]. These decomposition methods can be consistent with human visual characteristics but are easy to introduce artifacts. Hence, many other approaches have attracted great attention, such as sparse representation and low-rank representation.

In the sparse domain, the sparse representation(SR) [15] and dictionary learning [16] are widely used in image fusion. For instance, Li et al. [17] proposed a novel multimodal fusion method via three-layer decomposition and SR. Also, there are many methods that combine SR and other approaches for image fusion, such as low-rank representation (LRR) [18]. Zhu et al. [19] proposed a novel multimodality image fusion method based on image decomposition and sparse representation in which the texture components can be preserved well by a sparse representation based method. In [20], Liu et al. proposed a fusion method based on convolutional sparse representation (CSR) in which the detail of the source image can be retained well by multilayer features that can learn more about it in [21]. Besides, the joint sparse representation (JSR) [22] and cosparse representation [23] are also used in sparse domain. Although SR-based methods can improve image fusion performance, these methods are too time-consuming in dictionary learning operations [24]. These issues have prompted a growing study in deep learning to replace dictionary learning in SR.

In deep learning-based fusion methods, deep features of the source images can be extracted to reconstruct the fused images. For example, the VGG-19 [25], ResNet-50 [26], and DenseFuse [27] network architecture are commonly used in deep learning-based methods. Ma et al. [28, 29] proposed a multimodal image fusion based on adversarial networks, which improves the performance of image fusion to a large extent. Although the deep learning-based methods have performed well in image fusion, these methods still have some drawbacks, such as the deeper the network; the choice of parameters can be more complex.

To preserve many of the edge and detail features of the source image, we proposed a multilevel low-rank decomposition framework based on guided filtering and feature extraction algorithm for infrared and visible image fusion. This solution uses the MDLatLRR method to decompose the original images to extract the detail content coefficients. The fused detail content coefficients can be obtained by dynamic level measurement with maximum value. After superimposing these detail content coefficients, the edge and structure information of the original images will be well retained and improve the display of the object. Then, the VGG-19 network is used to extract the significant area, structure, and object characteristics of the base part coefficients, and the weight maps can be produced according to the base part's activity level. In order to better preserve the edge information of the base part, the improved weight map is obtained by guided filtering. The improved weight map and the base part coefficients then make the Hadamard product to acquire the fused base part coefficients. Finally, the fused base part and detailed content information are superimposed to get the final fusion result. After the above fusion scheme process, the experimental results show that the proposed method significantly outperforms the comparison methods in image information retention. The significant contributions of this paper are summarized as follows:

(1) We introduce MDLatLRR to decompose the source images and determine the optimal number of decomposition layers for infrared and visible image fusion

(2) Base part fusion: to obtain more features information, we use the VGG-19 network and guide filtering to fuse the base part. Firstly, the base part coefficients are learned by the VGG network to get the weight map. The weight map obtained in this way can well adapt the base part coefficients of the source image with a block distribution of pixel information. And then, the improved weight map is obtained by guided filtering, which can effectively preserve edge information and reduce noise in the weight map. Finally, the fused base part coefficients are acquired by multiplying the improved weight map and the base part coefficients

(3) Detailed content fusion: it is well known that the larger the detail content coefficient is, the more information it contains. The n-level detail content coefficients are fused using dynamic level measurement with maximum value and then reconstructed to obtain the final fused detail content coefficients, which can preserve more sufficient detail content information from the source images

(4) We first conducted ablative experiments on the number of decomposition layers of MDLattLRR and the number of layers of VGG-19 and finally selected a five-layer VGG-19 network to sufficiently extract features

The remainder of this thesis is as follows. Section 2 introduces multilevel decomposition latent low-rank representation to decompose the source images. Section 3 presents the fusion method of this paper. The base part is fused by the VGG-19 network and guide filtering. The detail content is fused by dynamic level measurement with maximum value. Section 4 presents the structure of the proposed image fusion algorithm. The experimental results are discussed and presented in Section 5. Section 6 summarizes this paper.

2. Multilevel Decomposition Latent Low-Rank Representation

In this section, the method of MDLatLRR is introduced. Liu et al. [30] proposed the method of low-rank representation (LRR) which can extract features from the input data. LRR is a method to explore the structure of data multisubspace by finding the lowest rank representation among the data.

However, this method can not work well when the input data is inadequate and damaged. In order to obtain good performance, the latent low-rank representation (LatLRR)

of theory [31] is proposed. The method utilizes more data to acquire the dictionary. In addition, the salient features can be extracted from the source data [31] by using the method of LatLRR. More specially, the single-level decomposition LatLRR (DLatLRR) problem is formulated as

$$V_d = S \cdot Q(I), \tag{1}$$

$$I_d = R(V_d), \tag{2}$$

$$I_b = I - I_d, \tag{3}$$

where I is the source image. $Q(\cdot)$ represents the two-stage operator, which composes of reshuffling and the sliding window technique. S denotes the projection matrix which is obtained by LatLRR. V_d means the decomposed result of the source image. $R(\cdot)$ is the operator which reconstructs the detail image based on detail content. I_d and I_b, respectively, signify detail content and the base part from the source image.

Due to DLatLRR, a multilevel latent low-rank representation (MDLatLRR) [14] is formed which is able to extract saliency features from the source image. The method of MDLatLRR is formulated as

$$V_d^i = S \cdot Q\left(I_b^{i-1}\right), \tag{4}$$

$$I_d^i = R\left(V_d^i\right), \tag{5}$$

$$I_b^i = I_b^{i-1} - I_d^i, I_b^0 = I, i = [1, 2, \cdots, r], \tag{6}$$

where i and r represent the present and the highest decomposition layers, respectively. V_d^i means the ith-level decomposition result of the source image. I_d^i and I_b^i, respectively, signify the ith-level detail content the base part the source image. I_b^0 indicates the source image. In the end, a base part and r detail contents are obtained in different decomposition levels.

The framework of MDLatLRR is described in Figure 1. The source image I is decomposed base part I_b^1 and detail content I_d^1 by DLatLRR. In order to obtain more feature information from the base part, the I_b^1 is further decomposed I_b^2 and I_d^2 by DLatLRR. If the decomposition layer is r, it will get r detail contents and a base part. As a result, the fused image can show more information from the source image. Nevertheless, with the decomposition layer increasing, the artifacts will introduce more. An important problem was how to select a suitable decomposition layer. The detailed description is in Section 5.1. Next, we will introduce the fusion method of the base part and the detail content, respectively.

3. Fusion Method

The source images are decomposed base parts and detail contents using the method of MDLatLRR. The base part contains edge information and basic contour information. Simonyan and Zisserman [25] employed the VGG network for the first time to extract features from images of different levels and obtained excellent results. As the level of decomposition increases, the amount of information contained in the base layer becomes less and less. Using the VGG network to extract the base part, more helpful information will be identified and integrated. The generated weight map will contain more useful information. Then, in order to contain more edge information, guided filtering [32] is used to smooth the weight map. Finally, the fused base part can be acquired by multiplying the refined weight and the source of images. In contrast to the base part, the detailed content contains more structural and textural information. The fused detail content is obtained by using the rule of taking the maximum for dynamic measurement [33].

3.1. Fusion of Base Parts. VGG-19 is a convolutional neural network with 19 layers, including 16 convolutional layers and three fully connected layers [34]. The structure of VGGNet is straightforward, using the same size convolutional kernel size (3×3) and maximum pooling size (2×2) for the whole network. The performance can be improved by continuously deepening the network structure. The structure diagram is shown in Figure 2. For the fused base part to contain more information, a five-layer VGG-19 network is used to extract the base part to form the feature map.

For the base part I_{b1}^i and I_{b2}^i, $\{\phi_1^m\}_{m=5}^{512}$ and $\{\phi_2^m\}_{m=5}^{512}$ indicate the deep features extracted from base part by the fifth convolutional level of VGG-19. As shown in Figure 2, the 5th convolutional layer is conv3-512, so there are 512 deep features in the each base part. In addition, the pooling process will resize the the feature map, which is $1/2^5$ times of the original size. Impacted by [18], the l_1-norm of $\{\phi_k^m\}_{m=5}^{512}(x, y)\}$ can transform into the activity level survey of the original detail content, where $k \in \{1, 2\}$. Hence, the activity level map C_k^i is shown in

$$C_k^i(x, y) = \left\| \{\phi_k^m\}_{m=5}^{512}(x, y) \right\|_1. \tag{7}$$

The soft-max operator is used to obtain the initial weight maps $\widehat{W}_k$, which is shown in

$$\widehat{W}_k(x, y) = \frac{C_k(x, y)}{\sum_{n=1}^{j} C_n(x, y)}, \tag{8}$$

where j is the amount of weight map, which is set to $j = 2$. W_k denotes the value of the weight map.

Using the upsampling operator, the final weight map is obtained that is consistent with the size of the detail content. The final weight is shown in

$$W_k(x, y) = \widehat{W}_k(x + p, y + q), p, q \in \{0, 1, 2, \cdots, 15\}. \tag{9}$$

In order to retain more edge information in the base part, guided filtering is used to smooth the final weight map W_k. The detailed calculation procedure for the guided filtering is described in [32]. First, W_k is processed to obtain the binary image, which is calculated by

$$P_k^n = \begin{cases} 1, \text{if } W_k^n = \max\left(W_1^n, W_2^n\right), \\ 0, \text{others}, \end{cases} \tag{10}$$

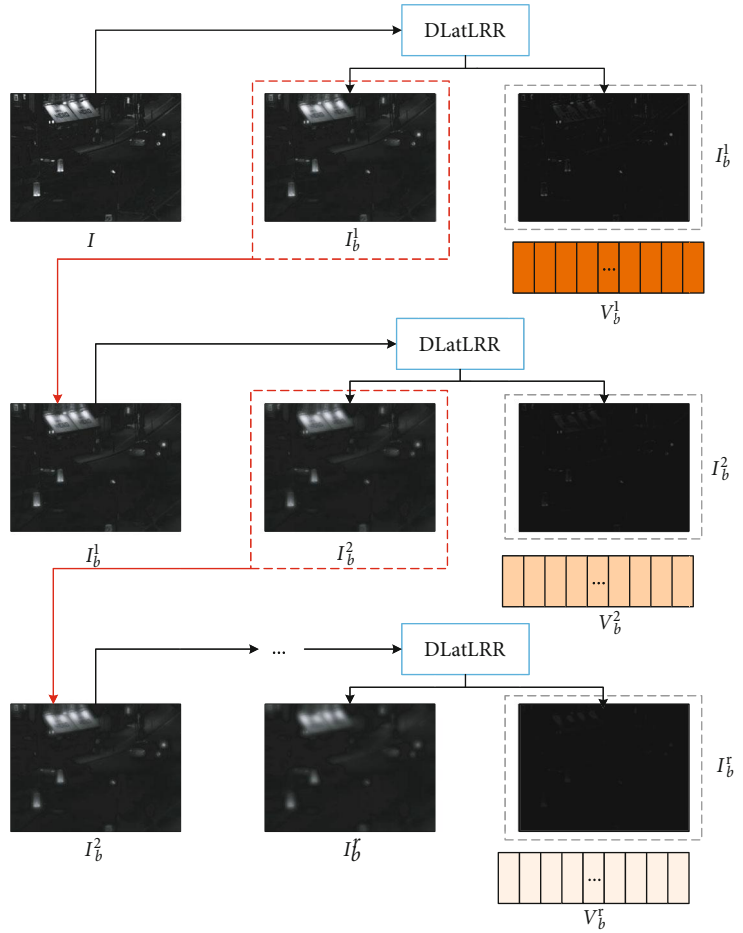


Figure 1: MDLatLRR decomposition diagram.

where P_k^n denotes the value of the nth pixel of the kth image in the binary image, and W_k^n means the value of the nth pixel of the kth image in the weight image.

Then, using the source image I_1 and I_2 as guided image, the guided filtering is applied to P_1 and P_2, as shown in

$$\bar{W}_1 = G_{r,\varepsilon}(P_1, I_1), \tag{11}$$

$$\bar{W}_2 = G_{r,\varepsilon}(P_2, I_2), \tag{12}$$

where $\bar{W}_1$ and $\bar{W}_2$ denotes the refined weight map, which is smoothed by the guided filtering. G is the guided filtering function. r and ε represent the parameters of guided filtering. If it is too smooth, it will cause the image of the edge and feature to be inconspicuous. The values of r and ε parameters are set in the experimental Section 5.1.

The fused base part I_b is calculated by

$$I_{bf} = \bar{W}_1 \cdot I_{b1}^r + \bar{W}_2 \cdot I_{b2}^r. \tag{13}$$

3.2. Fusion of Detail Content. In general, the greater the level of coefficient activity, the more information is contained in the image. To make the fused image include rich information, we use a fusion method called the dynamic level measurement with maximum value to fuse the detail content. The variance of each image patch over 3×3 or 5×5 windows is calculated as a measure of activity. The activity measure is associated with the pixel in the center of that window. The active measurements at the corresponding position are either taken as the maximum or the average, which is closed to each other. Since the activity measure in [35] corresponds to the cascading of a linear high-pass filter with a nonlinear high-pass filter, it has no clear physical meaning. In our implementation, we use the maximum absolute value within the window as the activity measure associated with the center pixel. Consistency verification can be understood as a switch to image B in the transform domain if the central pixel value is from image A and most of the surrounding pixel values are from image B. The fusion strategy diagram of the detailed content matrices is shown in Figure 3.

Firstly, the energy E_1 and E_2 are calculated for the corresponding local regions of the infrared and visible detail content, as shown in

$$E_k(x,y) = \sum_{m=-(M_1-1)/2}^{(M_1-1)/2} \sum_{n=-(N_1-1)/2}^{(N_1-1)/2} \left| V_{dk}^{i,j}(x+m, y+n) \right|^2, \tag{14}$$

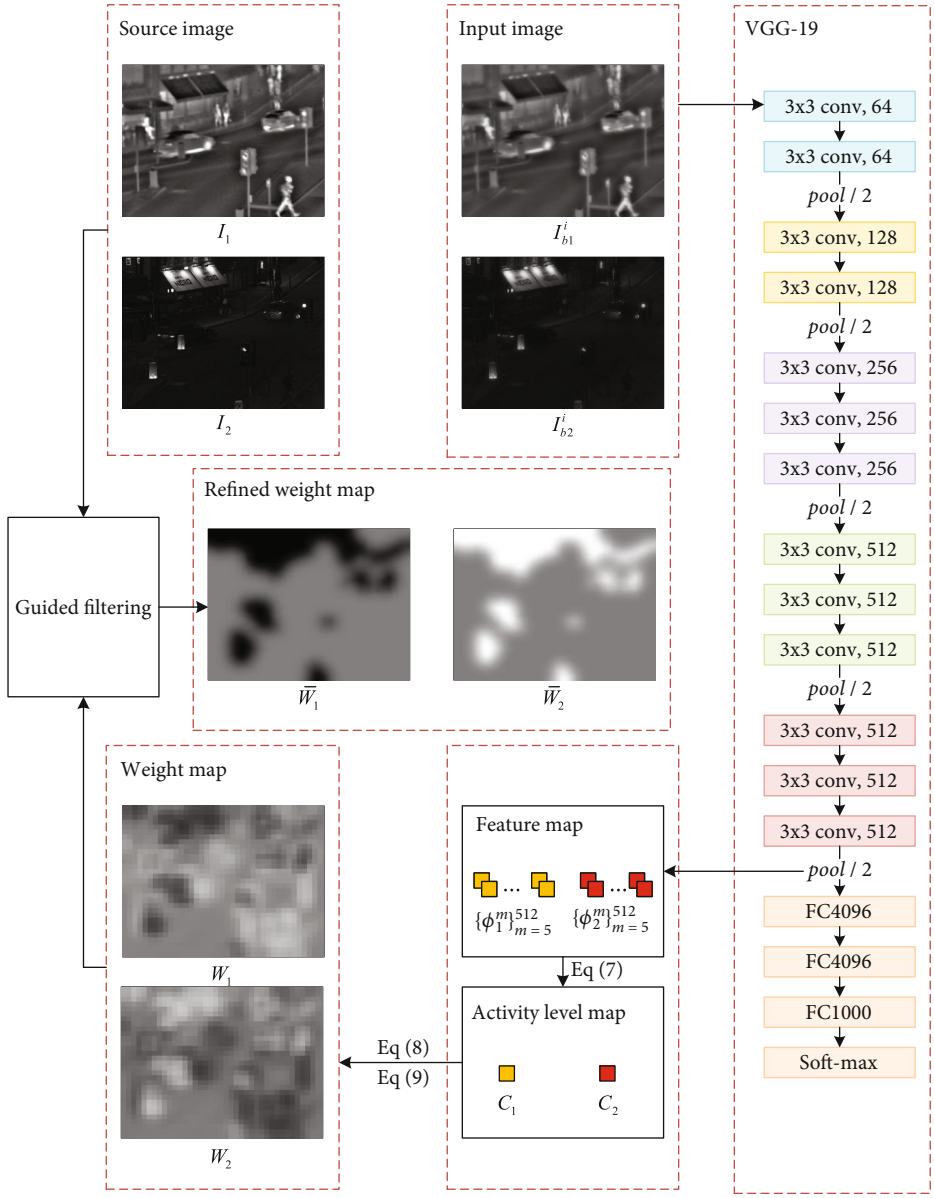


Figure 2: The procedure of base part fusion.

where $E_k(x, y)$ denotes the magnitude of the local energy, $k \in \{1, 2\}$. $m \times n$ defines the size of the local area, which is set to $m = n = 3$.

Then, the local area matching degree S_k is calculated by

$$S(x,y) = \frac{2\sum_{m=-(M_1-1)/2}^{(M_1-1)/2}\sum_{n=-(N_1-1)/2}^{(N_1-1)/2}\left|V_{d1}^{i,j}(x+m,y+n)V_{d2}^{i,j}(x+m,y+n)\right|^2}{E_1(x,y)+E_2(x,y)}. \tag{15}$$

When the two images are strongly correlated, the weighted average is used. Conversely, the coefficient with higher local energy is used. The fused detail content vector $V_{dk}^{i,j}$ is acquired by

$$V_{dk}^{i,j}(x,y) = \begin{cases} w^{\max}(x,y)V_{d1}^{i,j}(x,y) + w^{\min}(x,y)V_{d2}^{i,j}(x,y) & E_1(x,y) \geq E_2(x,y), \\ w^{\min}(x,y)V_{d1}^{i,j}(x,y) + w^{\max}(x,y)V_{d2}^{i,j}(x,y) & E_1(x,y)E_2(x,y), \end{cases} \tag{16}$$

$$w^{\min}(x,y) = \frac{1}{2} - \frac{1}{2}\left[\frac{1-S(x,y)}{1-\alpha}\right], w^{\max}(x,y) = 1 - w^{\min}(x,y), \tag{17}$$

where α is the matching threshold, which is set to $0.5 \sim 1$. $w^{\min}$ and $w^{\max}$ are the weighting factors.

The strategy is used to all detail content vector $V_{dk}^{1:r}$. The detailed content fusion procedure is shown in Figure 3.

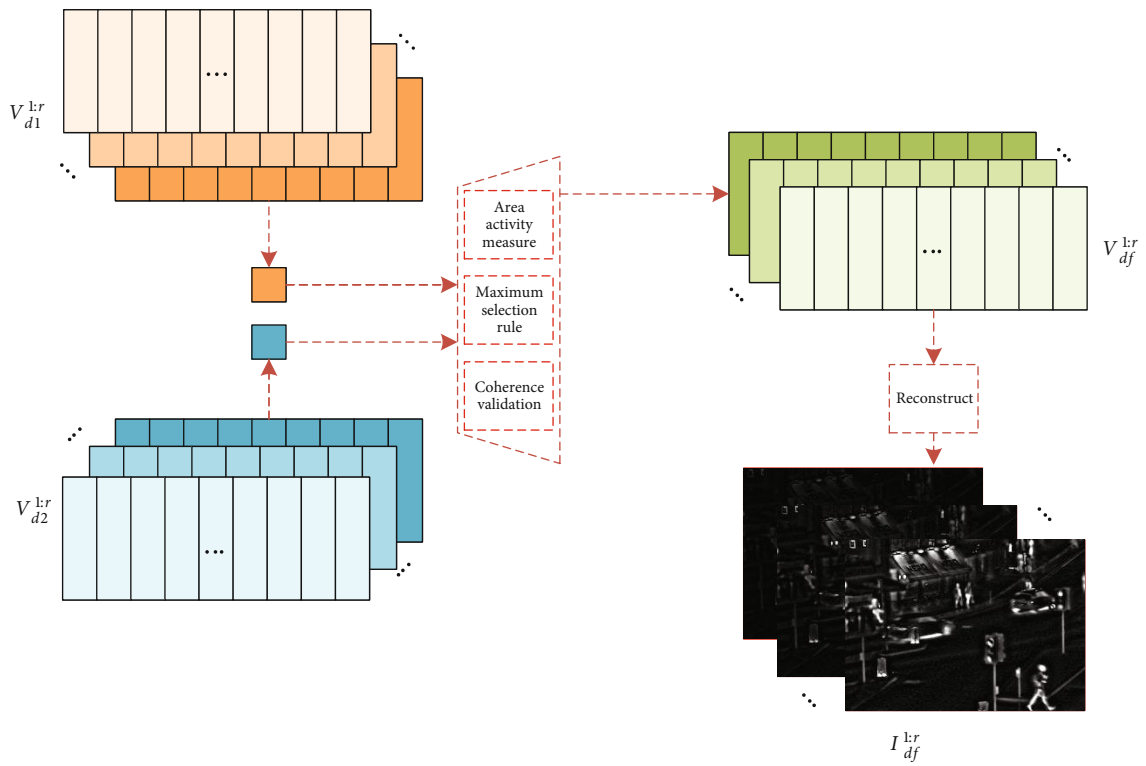


Figure 3: The procedure of detail content fusion.

Every detail contentI_{df}^{i} is obtained by

$$I_{df}^{i} = R\left(V_{df}^{i}\right) \quad i = [1, 2, \cdots r], \tag{18}$$

where $R(\cdot)$ denotes the refactor operator, which is mainly used to reorganize vectors into image blocks.

3.3. Reconstruction. The fused base part I_{bf} and detail content I_{df}^{i} is superposed to reconstruct the fused image I_f, as shown in

$$I_f = I_{bf} + \sum_{i=1}^{r} I_{df}^{i}. \tag{19}$$

4. Structure of Fusion Algorithm

We develop a novel infrared and visible image fusion method called a multilevel low-rank decomposition framework based on guided filtering and feature extraction. The source images are denoted as I_1 andf I_2, which are preregistered. The proposed algorithm framework in this paper is shown in Figure 4.

The general steps of the proposed algorithm in this paper are shown in Algorithm 1.

5. Experiments

The aim of experiment is to give a supporting evidence for the proposed method. The experiment in this paper is composed of experimental settings, ablation experiment, subjective evaluation, and objective evaluation.

5.1. Experimental Settings. In our experiment, our infrared and visible images were collected from [36], which contains a lot of registered infrared and visible images from a different scene. We randomly selected six pairs of images to compare the fusion results is shown in Figure 5. From left to right, these pairs, respectively, named *Men in front of house*, *Bench*, *Bunker*, *Man_in_doorway*, *Soldier_in_trench_1*, and *Lake*.

The parameter setting for GF. According to [37], the value of r and ε is set to 45 and 0.3. The stride of the sliding window is set to 1, which can decompose the source images into patches. The window size is set to 16×16. The number of decomposition layers of the MDLatLRR and the number of network layers of the VGG-19 network to extract the base part of the feature map will be obtained from the subsequent ablation experiments.

Six classical infrared and visible image fusion methods are applied to conduct the same experiment for comparison, containing a generative adversarial network for image fusion(FusionGAN) [28], the joint-sparse representation model (JSR) [38], the JSR model with the method of saliency detection(JSR_SD) [39], multilevel decomposition method of MDLatLRR [14], two approaches based on deep learning VGG-19 [40], and ResNet50 [41].

In order to get a quantitative comparison at different methods, four quality metrics are used for the fused images. These are as follows: entropy [42] measures the amount of information contained in the fused image based on

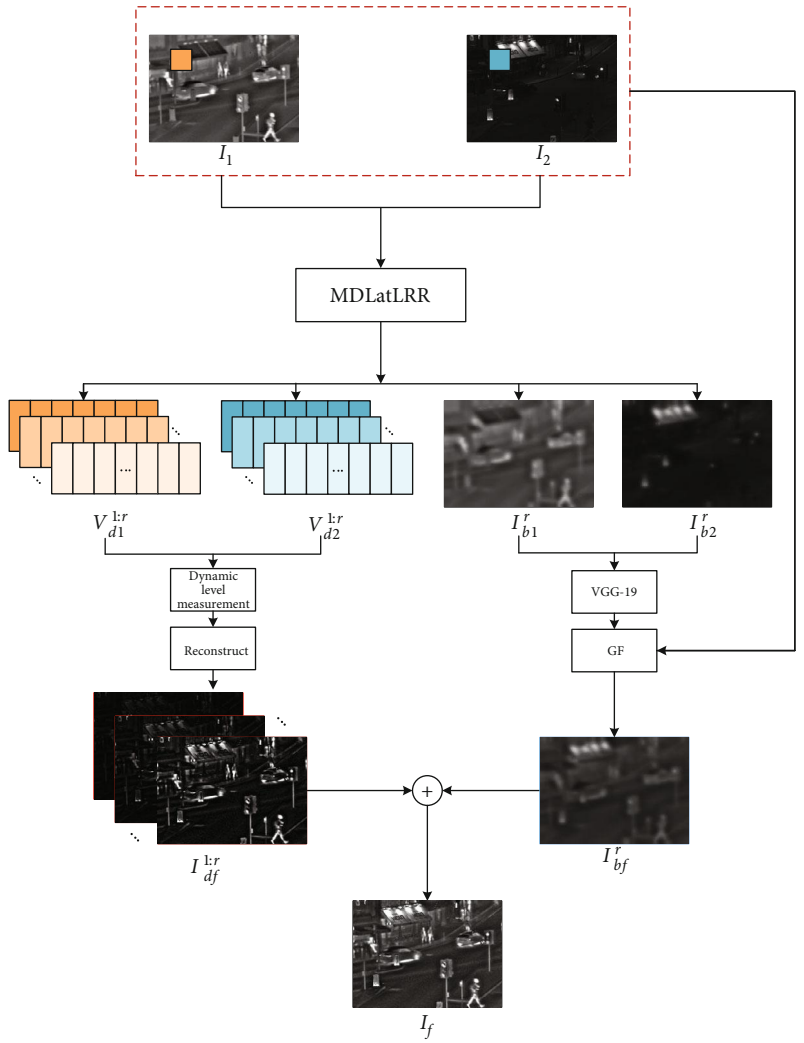


Figure 4: The proposed algorithm framework.

information theory; mutual information [43] represents a measure of the amount of information transferred from the source image to the fused image; Q_{abf} [44] indicates the quality of edge information acquired from the source images; MS-SSIM [45] only counts the structural information based on the refined structural similarity. The larger these metrics are, the better result of fusion quality will be.

All the fusion algorithms experiments are prosecuted in MATLAB R2020a on 3.95 GHz AMD(R) Ryzen(R) 5 3500X 6-Core Processor with 16 GB RAM and Win 10 64-bit operating system. The graphics card is GeForce RTX 2070 SUPER 8G.

5.2. Ablation Experiment

5.2.1. Ablation Experiments for Decomposing Layers. To select the best decomposition layer for the proposed method, five pairs of images in Figure 5 are implemented in the proposed algorithm in different decomposition layers. To test the decomposition layers of MDLatLRR, the layer is set from 1 to 4. The decomposition level of fused results for five pairs of the source image is shown in Figure 6. With the increase of MDLatLRR decomposition level, the fused image luminance and contour are improved. However, it introduces the artifact around the object and makes some detailed information degradation. To obtain better fusion quality, the fewer artifacts, the better.

The experimental results are shown in Figure 7, which is obtained by the above quality metric. As can be seen, it is not the case that the greater the number of layers, the greater the value of the quality evaluation index. When the number of decomposition layers is at the first level, the value of EN and MI is more prominent than other layers. It suggests that the first layer can make the fused image contain more information from the source image. In addition, there are several images value of Q_{abf} that is best at the second layer. That indicates that more edge information is preserved in the fused image with the increasing decomposition layers. As for MS-SSIM, the first two layers show better values. It shows that the structure of the fused image is similar to the source image. However, when the decomposition layer is more than two layers, the fusion performance will decline.

Input:
The source of image I_1 and I_2.
Output:
Fused image I_f.
/∗ Part 1: multilevel DLatLRR decomposition. ∗/.
1: **for** each $K \in [I_1, I_2]$**do**
2: **for** each $i \in [1, r]$**do**
3: Run DLatLRR decomposition on K to obtain $\{I_{b1}^i, V_{d1}^{1:i}\}$, $\{I_{b2}^i, V_{d2}^{1:i}\}$
4: **end for**
5: **end for**
/∗ Part 2: fusion of base parts. ∗/
6: **for** each $k \in \{1, 2\}$**do**
7: Input image I_{bk}^i is extracted by the 5th layer of VGG-19 network to acquire $\{\phi_k^m\}_{m=5}^{512}$;
8: Transform the l_1-norm of $\{\phi_k^m\}_{m=5}^{512}$ into the activity level map C_k by the Equation (7);
9: Calculate the final weight map W_k via Equations (8) and (9);
10: Use the guided filtering to smooth the final weight to obtain the refined weight $\bar{W}_k$ via Equations (11) and (12).
11: **end for**
12: Calculate the fused base parts I_{bf} via Equation (13).
/∗ Fusion of detail content.
13: **for** each $i \in \{1, 2, \cdots, r\}$**do**
14: Apply the dynamic activity level with the maximum value on $\{V_{d1}^{1:i}, V_{d2}^{1:i}\}$ to obtain the fused vector $\{V_{df}^{1:i}, V_{d2}^{1:i}\}$ as Equation (16);
15: Reconstruct the vector $\{V_{df}^{1:i}, V_{d2}^{1:i}\}$ to I_{df}^i via Equation (18).
16: end for
/∗ Reconstruction ∗/.
17: Superpose the fused base part I_{bf} and detail content I_{df}^i to reconstruct the fused image I_f, as shown in Equation (19).

Algorithm 1: Framework of the proposed algorithm for image fusion

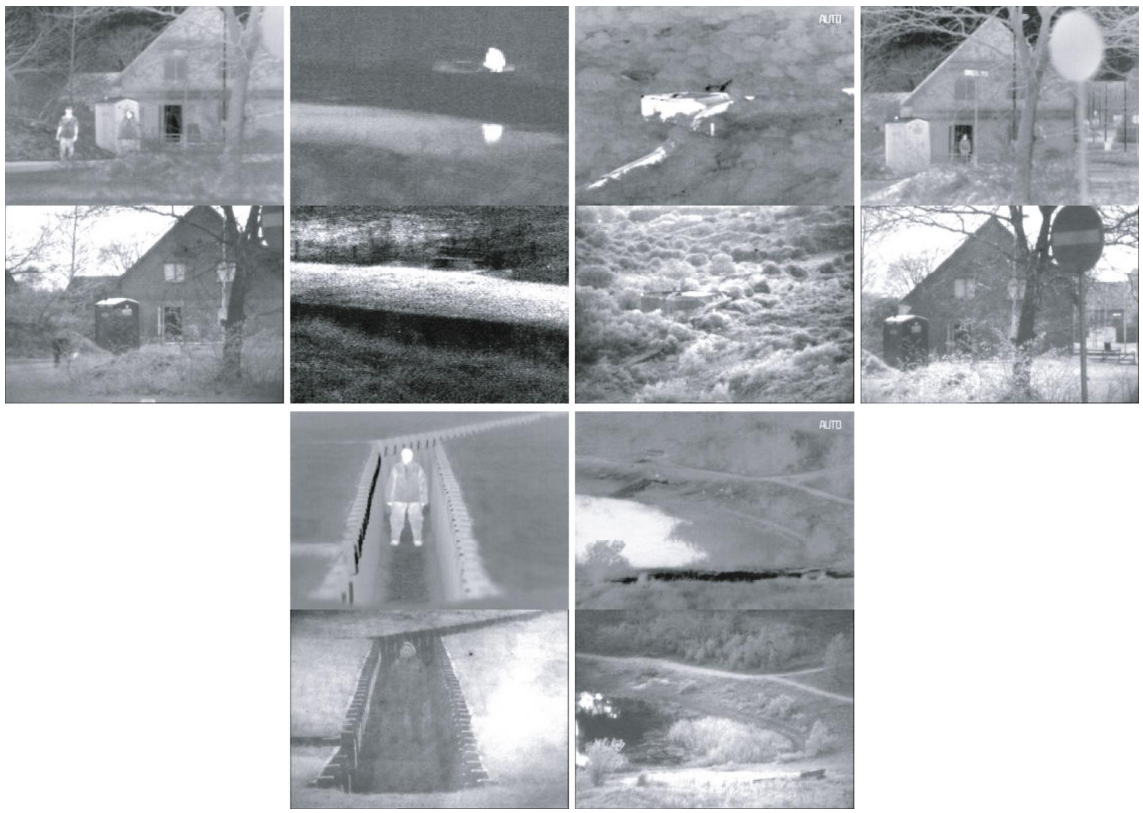


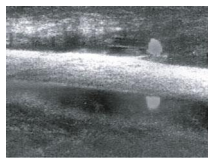
Figure 5: Six pairs of multimodal infrared and visible images. The top is infrared images, and the bottom is visible images.

1 layer (a)

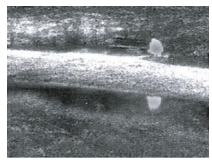
2 layers (b)

3 layers (c)

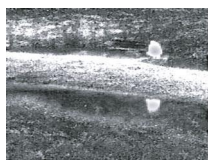
4 layers (d)



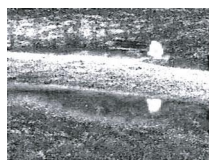
1 layer (e)



2 layers (f)



3 layers (g)



4 layers (h)

1 layer (i)

2 layers (j)

3 layers (k)

4 layers (l)

1 layer (m)

2 layers (n)

Figure 6: Continued.

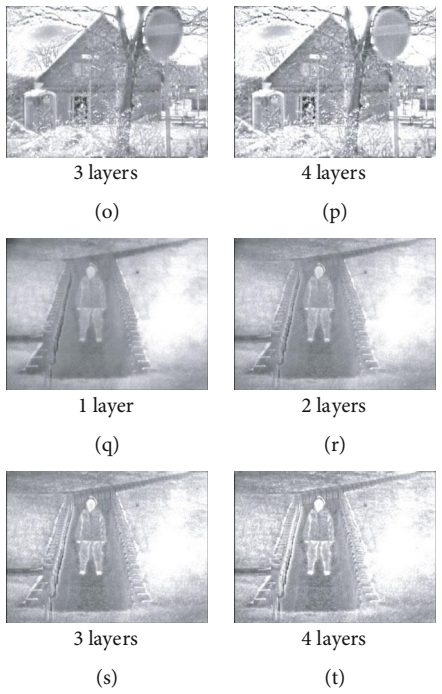


Figure 6: The each level fusion result is decomposed by MDLatLRR.

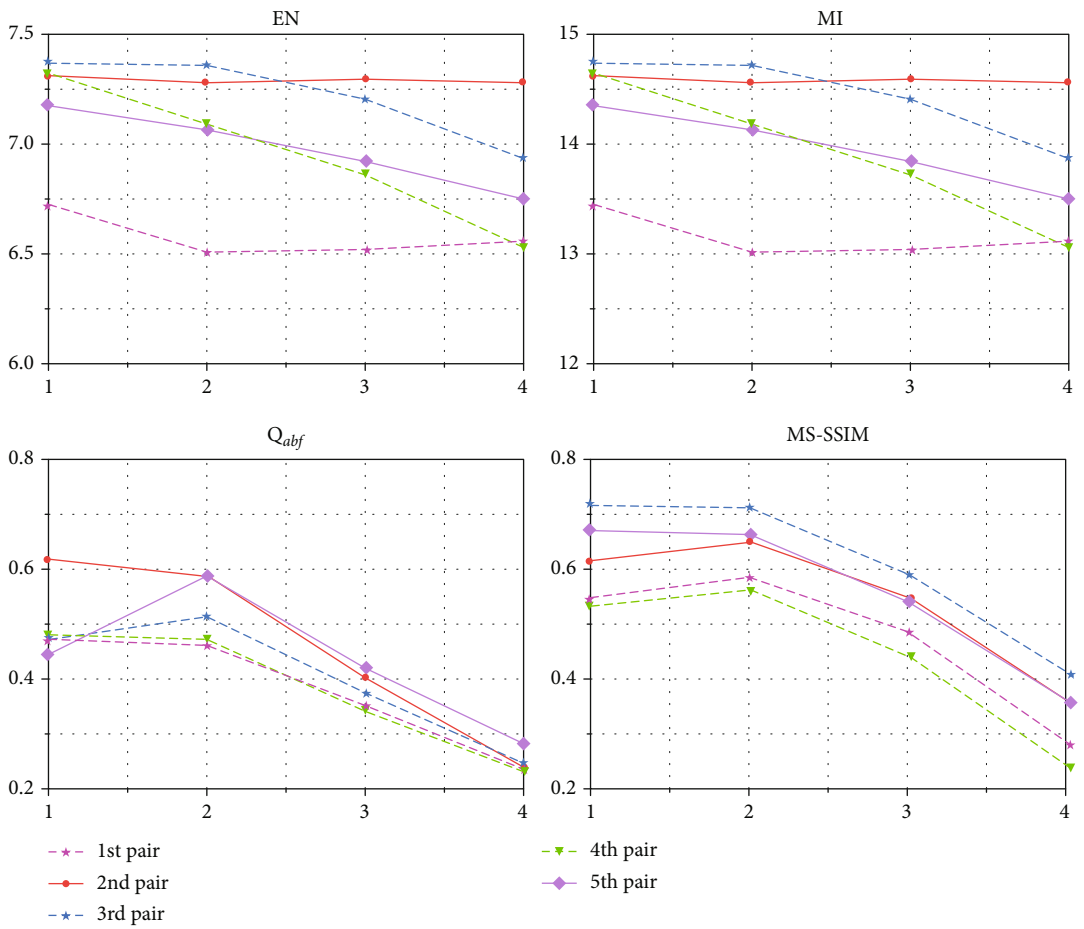


Figure 7: The decomposition layer is set from 1 to 4. The values of four evaluation metrics are acquired by MDLatLRR with different layers.

1 layer

(a)

2 layers

(b)

3 layers

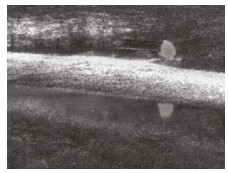
(c)

4 layers

(d)

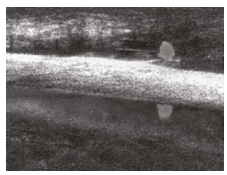
5 layers

(e)



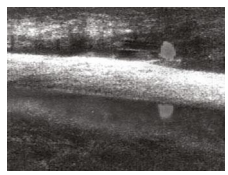
1 layer

(f)



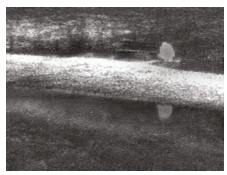
2 layers

(g)



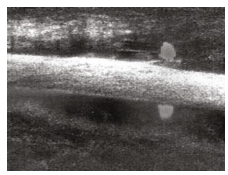
3 layers

(h)



4 layers

(i)



5 layers

(j)

1 layer

(k)

2 layers

(l)

3 layers

(m)

4 layers

(n)

Figure 8: Continued.

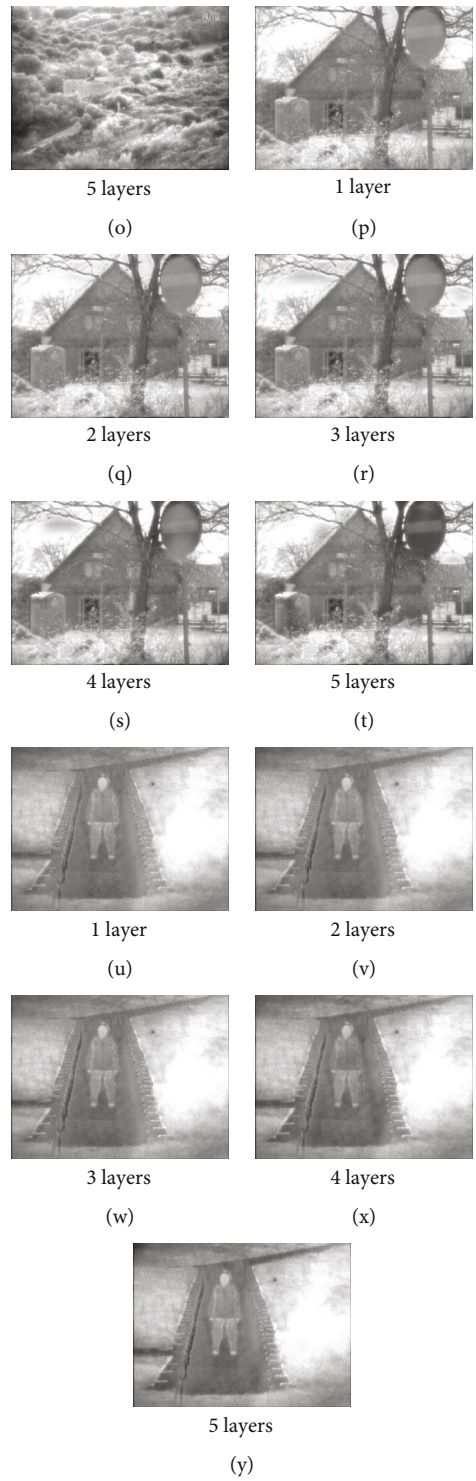


Figure 8: The fusion results for different VGG-19 network layers.

It is because that the detailed content obtains more luminance and contour information from the base part. This information can not be fused well by the detail content fusion method. By the way, the larger the value of the above evaluation index, the better the effect of fused images. On the basis of the above analysis, the decomposition of MDLatLRR is set one in our proposed algorithm.

5.2.2. Ablation Experiments for VGG-19 Network Layers. In order to select an appropriate number of layers for the VGG-19 network of the proposed method, five pairs of images in Figure 5 are used for ablation experiments of VGG-19 network layers. The layer is set from 1 to 5, which represents *relu_1_1*, *relu_2_1*, *relu_3_1*, *relu_4_1*, and *relu_5_1*, respectively. The fusion results for different

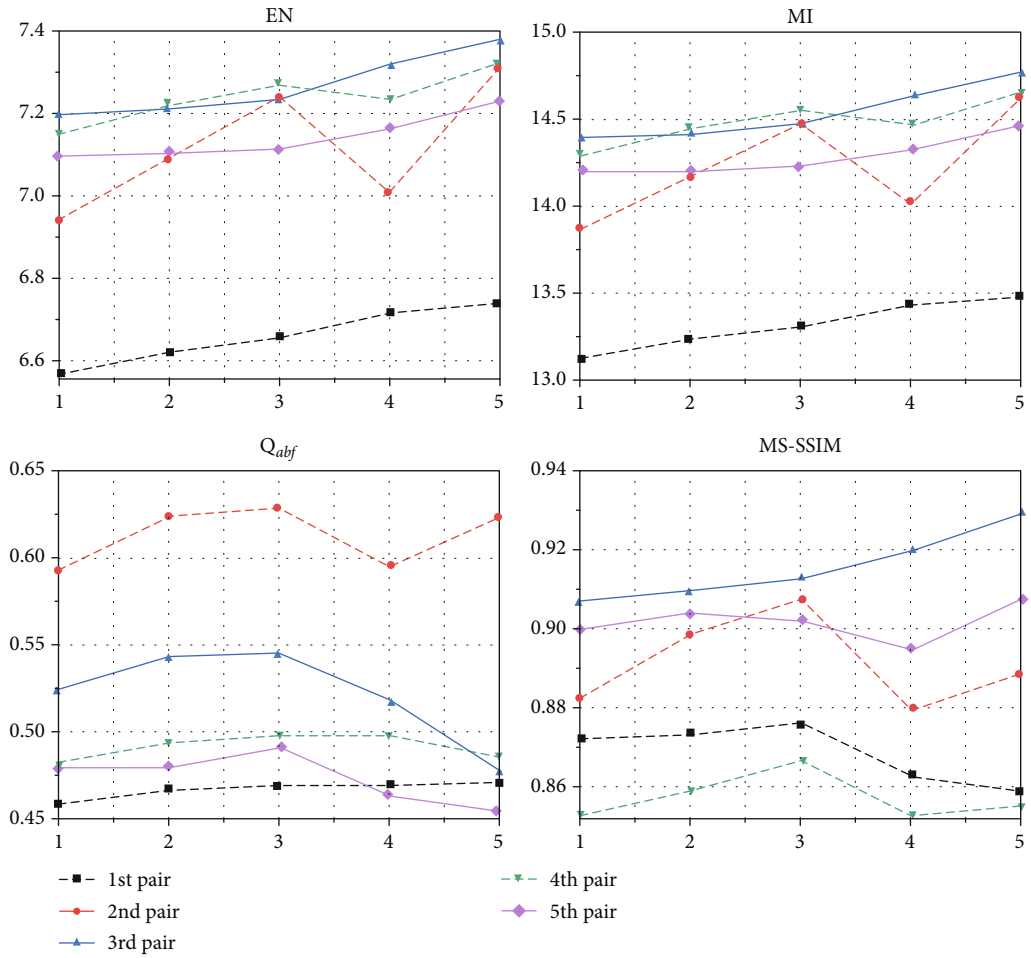


Figure 9: The values of four evaluation metrics are acquired by VGG-19 with different network layers.

VGG-19 network layers are shown in Figure 8. As can be seen from the Figure 8, compared with the others layers of VGG-19 network, the fifth layers of VGG-19 network extract more detail features and salient target information from the source image. For example, Figure 8(t) of the extraction of traffic sign in the fifth layer is better than the others layers, so we choose 5-layer VGG-19 network to extract features.

The experimental results of different network layers are shown in Figure 9. As can be seen, with the number of network layers increases, the values of evaluation metrics EN and MI become larger. It represents that the five-layer VGG-19 network can extract more feature information from the source images. The evaluation metric Q_{abf} is basically the best when the source images are extracted using a three-layer VGG-19 network. It indicates that the three-layer VGG-19 network can extract edge information well. As for the evaluation of MS-SSIM, the MS-SSIM values of the third and fifth pairs are the best when using the five-layer VGG-19 network. The MS-SSIM values of the first, second, and fourth pairs are the best when using the three-layer VGG-19 network. To sum up the above, we select a 5-layer VGG-19 network to extract features.

5.3. Subjective Evaluation. Figure 10 shows the subjective fusion results of the first pair of images. Figures 10(a) and 10(b) are the original images. The object of man in the red box and the grass in the green box obtained by JSR and JSR_SD are fuzzy, and the fused images obtained by JSR and JSR_SD have significantly more the visible components than the infrared ones. The fused images obtained by FusionGAN have more the infrared components than the visible ones. In addition, the fused images obtained by MDLatLRR, VGG-19, and ResNet-ZCA are less artifact but the detailed texture information in the visible image is not well preserved. As shown from Figure 10(i), the object of man in the red box and the grass in the green box are the clearest compared with other methods. The proposed method adds more detailed texture information to make the same as a visible image while containing the infrared image of thermal radiation information. It has excellent visibility. Figure 11 shows the subjective fusion results of the second pair of images. Figures 11(a) and 11 (b) are the original images. It can be seen from Figure 11(i) that the pixel consistency of the object edge structure is the best for the fused images. The objects of red and green boxes obtain more texture information. Figure 12 shows the subjective fusion results

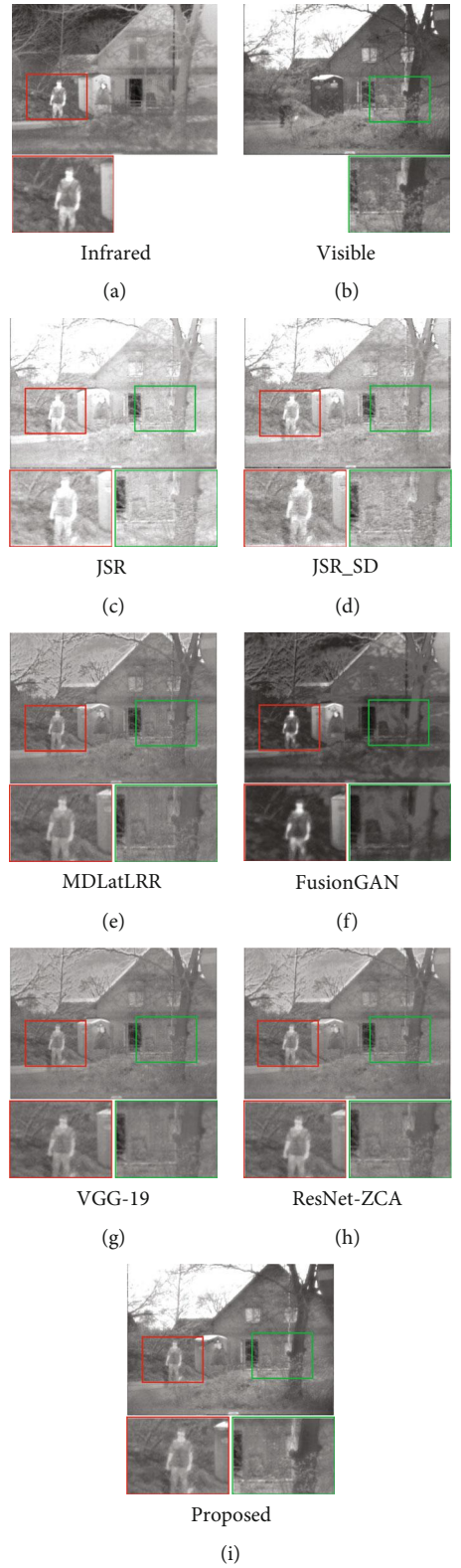


Figure 10: Comparison of subjective fusion results using different methods in the first pair images.

of the third pair images. Figures 12(a) and 12 (b) are the original images. In Figure 12(i) of the proposed method, the building in the red box contrasts with its surroundings, and the chromatic aberration is consistent with the visible image. The grass in the green box has more texture information. Figure 13 shows the subjective fusion results of the fourth pair of images. Figures 13(a) and 13(b) are the original images. In Figure 13(i), the brand

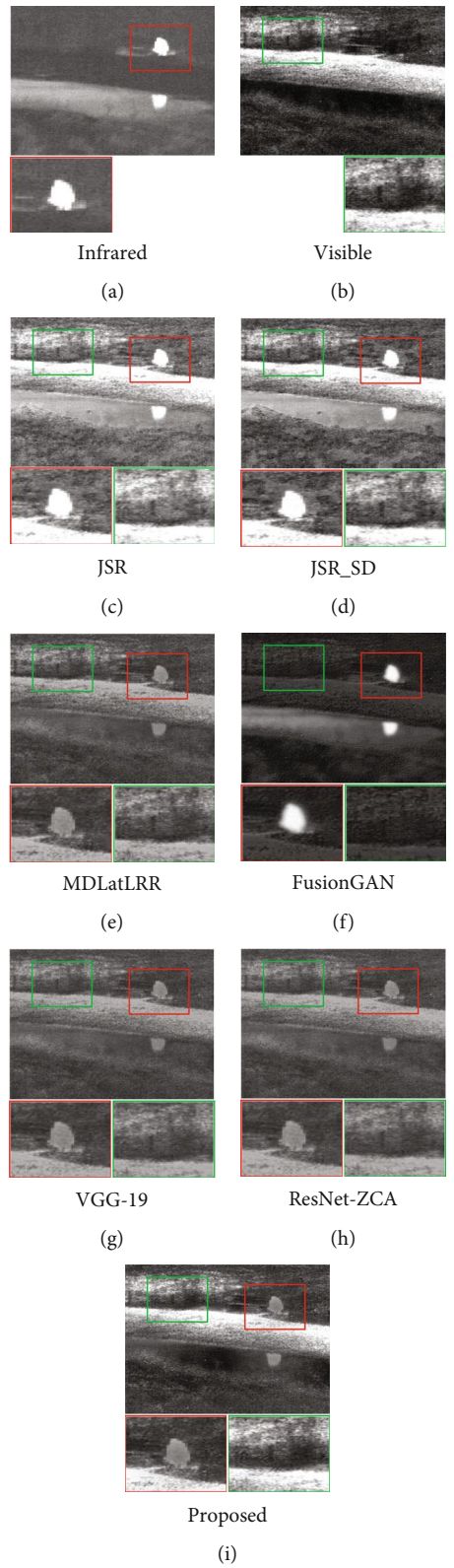


Figure 11: Comparison of subjective fusion results using different methods in the second pair images.

in the green box is the most recognizable compared with the results of other methods. The object in the red box contains more edge feature information. Besides, target object information is lost in some images, such as Figures 13(c), 13(d), and 13(f). The proposed method performs well and has good visibility. Figure 14 shows the subjective fusion results of the fifth pair images. Figures 14(a) and 14(b) are the original images.

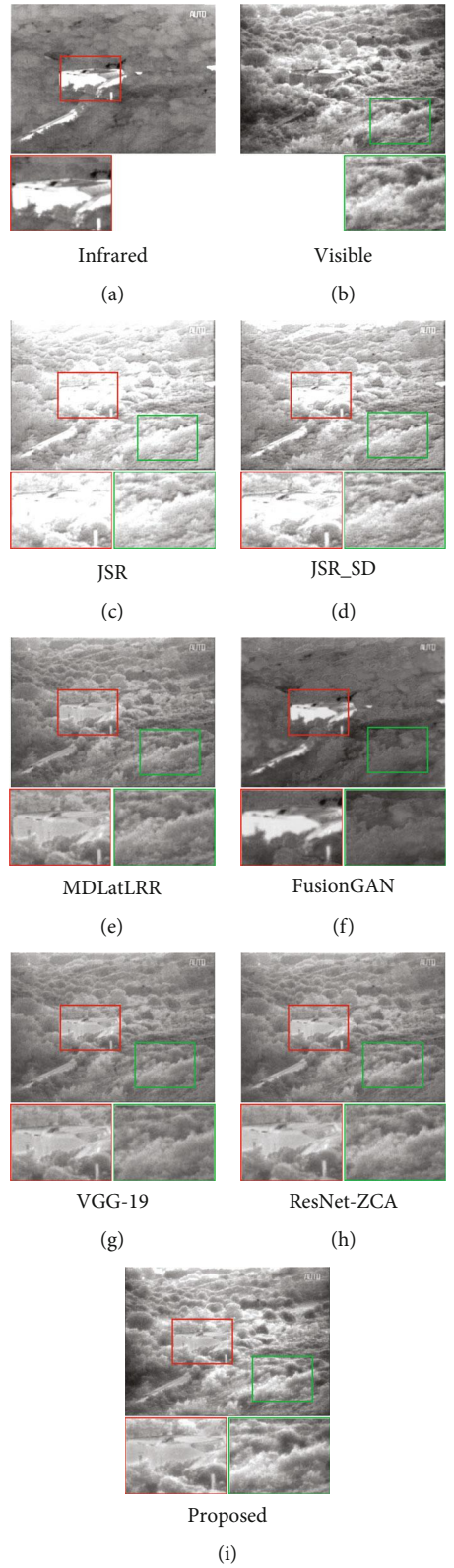


Figure 12: Comparison of subjective fusion results using different methods in the third pair images.

Figure 15 shows the subjective fusion results of the sixth pair of images. Figures 15(a) and 15(b) are the original images. From the target of red boxes and the detail features of green boxes in Figures 15(c)–15(i), the proposed method contains more details information from the source images compared with other methods. The proposed method adds more detail texture information to make the same as a visible image while containing the

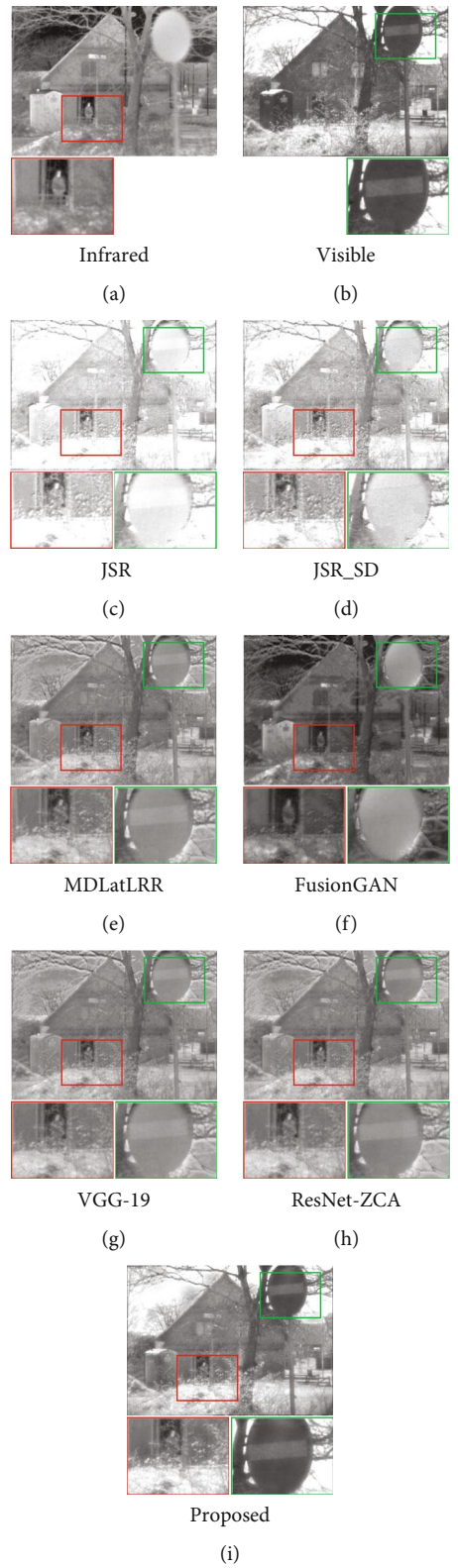


Figure 13: Comparison of subjective fusion results using different methods in the fourth pair images.

infrared image of thermal radiation information. Compared with other methods, the object in the red box and green box are more texture information, and the contrast between light and dark details is sharp. The structure is the most consistent with the original images. In addition, we randomly chose 20 pairs of images from [36] to verify the performance of the proposed method in Figure 16.

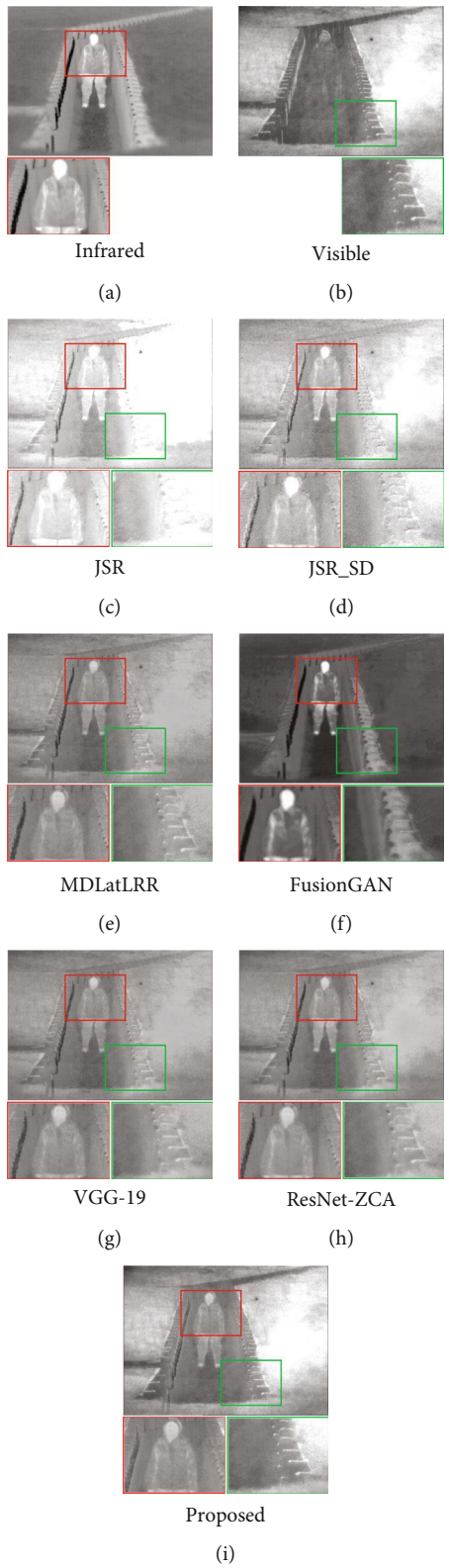


Figure 14: Comparison of subjective fusion results using different methods in the fifth pair images.

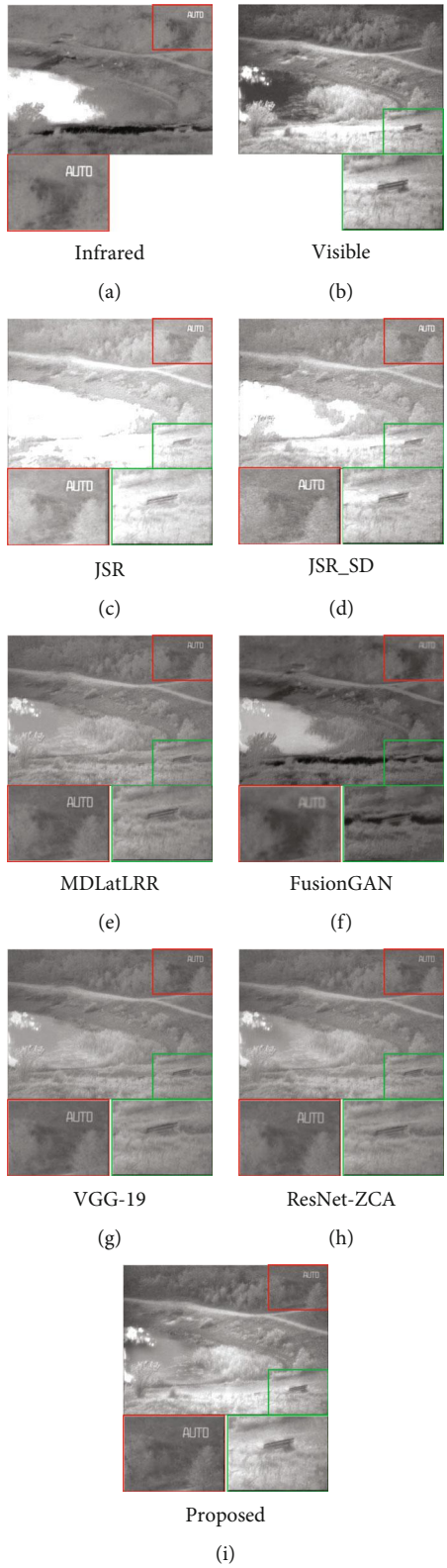
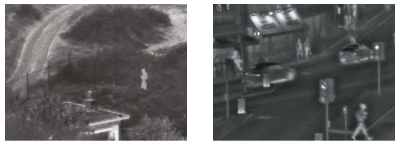


Figure 15: Comparison of subjective fusion results using different methods in the sixth pair images.



1st pair (a)

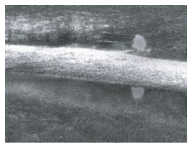
2nd pair (b)

3rd pair (c)

4th pair (d)

5th pair (e)

6th pair (f)



7th pair (g)

8th pair (h)

9th pair (i)

10th pair (j)

11th pair (k)

12th pair (l)

13th pair (m)

14th pair (n)

Figure 16: Continued.

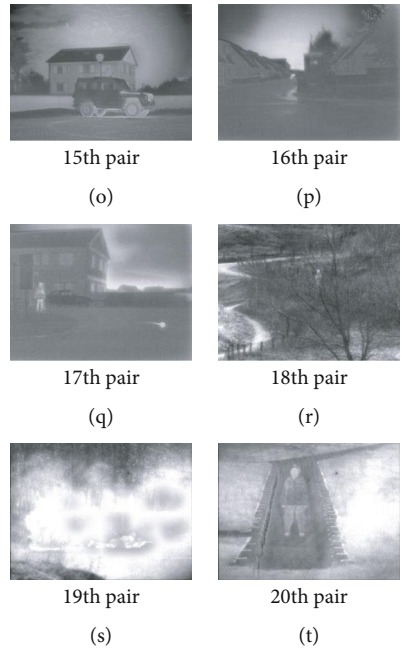


Figure 16: The fusion results using the proposed method on the twenty pairs of images.

5.4. Objective Evaluation. For exhibiting the attractive characteristic of our proposed method, four evaluation metrics are applied to compare the fusion property of six popular fusion methods and our proposed algorithm. In the tables, the best values are shown in italics.

In Table 1, the evaluation metrics EN, MI, and Q_{abf} are the best. It indicates that the proposed method contains more detailed information from the original images and edge information. In addition, the MS-SSIM is not the best, but the gap between the proposed method of MS-SSIM and the best value by MDLatLRR is tiny. As mentioned in Section 5.1, EN and MI measure the amount of information from the source image in the fused image. But EN is susceptible to noise. As shown in Figure 12(d), the object of the JSR_SD fusion image is distorted and has apparent artifacts. That is why EN and MI perform undeniable advantages in the third pair images. MS-SSIM counts the structural information based on the refined structural similarity, and the artifacts and the distortion of image structure will lower this metric. That will result in poor visibility. The proposed method has obvious advantages in the MS-SSIM index, which contains little noise and distortion of the structure. It is crucial for infrared and visible images. In Tables 2–5, the proposed method mostly performs the best in EN, MI, and MS-SSIM index. It shows that our algorithm makes the fused image contain more information from the source image and the structure of the fused image is similar to the source image. In addition, for objective evaluation, we provide Table 6 which contains the average values of all test images on different metrics. The evaluation metrics obtained by the proposed method are the best except the values of t/s. It indicates that our proposed method contains more feature information from the source image. And the structure of the fused image is similar to the source image, better than the others compared methods. Among all the compared methods, the proposed method is in the middle level in terms of time consumption in Tables 1–6. Based on the above analysis, our fusion algorithm is effective.

Table 1: Objective fusion results on the first pair images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	*6.7289*	*13.4578*	*0.4735*	*0.8586*	18.1600
FusionGAN	6.4955	12.9910	0.2303	0.7476	3.3096
JSR	6.1779	12.3558	0.2953	0.8146	*2.1414*
JSR_SD	6.4545	12.9090	0.2866	0.7746	130.4739
MDLatLRR	6.4837	12.9675	0.4261	*0.8907*	74.0113
VGG-19	6.4450	12.8901	0.3526	0.8700	7.2790
ResNet-ZCA	6.5132	13.0264	0.3640	0.8758	4.1634

Table 2: Objective fusion results on the third pair images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	*7.3758*	*14.7516*	*0.4807*	*0.9295*	17.6002
FusionGAN	6.4505	12.9010	0.1658	0.4494	3.1953
JSR	6.5060	13.0121	0.3221	0.8279	*2.1392*
JSR_SD	7.0878	14.1756	0.3023	0.8210	129.8800
MDLatLRR	6.7717	13.5433	0.4630	0.8631	73.3203
VGG-19	6.7090	13.4181	0.3185	0.8070	7.4550
ResNet-ZCA	6.7676	13.5352	0.3522	0.8280	4.1687

Table 3: Objective fusion results on the fourth pair images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	*7.3181*	*14.6363*	*0.4857*	0.8551	17.5412
FusionGAN	6.8485	13.6971	0.2294	0.6862	3.2032
JSR	5.5820	11.1640	0.2739	0.7258	*2.1193*
JSR_SD	6.4993	12.9987	0.2803	0.7439	129.8348
MDLatLRR	6.8080	13.6160	0.4395	*0.8770*	73.2556
VGG-19	6.7667	13.5333	0.3614	0.8516	8.1230
ResNet-ZCA	6.7676	13.5352	0.3522	0.8280	4.2020

Table 4: Objective fusion results on the fifth pair images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	*7.1776*	*14.3552*	0.4494	*0.9082*	17.5217
FusionGAN	6.3209	12.6418	0.2147	0.7218	3.2608
JSR	5.8003	11.6006	0.2944	0.8217	*2.1742*
JSR_SD	6.9258	13.8516	0.2944	0.8249	129.8357
MDLatLRR	6.5841	13.1682	*0.5273*	0.8954	73.3878
VGG-19	6.5430	13.0859	0.3988	0.8693	7.4692
ResNet-ZCA	6.6948	13.3896	0.4062	0.8813	4.2912

Table 5: Objective fusion results on the sixth pair images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	*7.1558*	*14.3115*	*0.4775*	0.8748	28.5617
FusionGAN	6.5194	13.0387	0.2329	0.7299	3.2456
JSR	6.1612	12.3224	0.2862	0.7932	*2.2352*
JSR_SD	6.9259	13.8518	0.2660	0.7287	143.9548
MDLatLRR	6.5695	13.1389	0.4522	*0.8823*	47.2279
VGG-19	6.5451	13.0901	0.3526	0.8565	11.6805
ResNet-ZCA	6.5782	13.1565	0.3584	0.8609	5.1654

Table 6: Objective fusion results of average value on the six pairs images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	*7.1786*	*14.3572*	*0.4981*	*0.8856*	17.0943
FusionGAN	6.5198	13.0395	0.2155	0.6521	2.8039
JSR	6.2254	12.4509	0.3211	0.7993	*1.8635*
JSR_SD	6.8701	13.7401	0.3098	0.7706	114.3775
MDLatLRR	6.6449	13.2899	0.4766	0.8763	58.9798
VGG-19	6.5957	13.1914	0.3658	0.8435	7.3742
ResNet-ZCA	6.6489	13.2978	0.3684	0.8500	3.8398

Table 7: Objective fusion results on the second pair images when using different algorithms.

Fusion algorithm	EN	MI	Q_{abf}	MS-SSIM	*t/s*
Proposed method	7.3155	14.6310	*0.6220*	*0.8875*	3.1809
FusionGAN	6.4838	12.9677	0.2200	0.5777	0.6088
JSR	7.1251	14.2502	0.4546	0.8122	*0.3714*
JSR_SD	*7.3270*	*14.6541*	0.4292	0.7303	22.2860
MDLatLRR	6.6527	13.3053	0.5512	0.8494	12.6757
VGG-19	6.5654	13.1307	0.4109	0.8066	2.2384
ResNet-ZCA	6.5488	13.0976	0.3711	0.7990	1.0481

6. Conclusion

This paper proposes a multilevel low-rank decomposition method based on guided filtering and feature extraction for infrared and visible image fusion. The VGG-19 network and guide filtering are used in the base layer fusion to obtain the weight map. Then, the final base layer is acquired by multiplying the initial base layer and the weight map. As for detail content fusion, we are using the dynamic activity level with maximum value to obtain the final detail content. The results exhibit that our proposed method has an attractive performance in retaining the object detail features information and edge feature information compared with other fusion methods in both subjective and objective. The proposed method can be applied in target detection and recognition in daily computer vision. In addition, there are some drawbacks to our proposed algorithm. With decomposition increasing, more luminance and contour information are introduced, aggravating the fused performances. The artifacts are more bright to interfere with the targets. In the latter work, we will be committed to reducing artifacts' effect and enhancing the fusion performance with the number of decomposition layers increasing.

Data Availability

The figures data used to support the findings of this study are included within the article.

Conflicts of Interest

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

Acknowledgments

This work was partially supported by grants from the National Natural Science Foundation of China (Grant No. 22178036), Chongqing Nature Science Foundation for Fundamental Science and Frontier Technologies (Grant No. cstc2018jcyjAX0483), Science and Technology Research Program of Chongqing Education Commission of China (Grant Nos. KJQN201900821 and KJQN202000803), Innovative Research Group of Universities in Chongqing

(Grant No. CXQT21024), Graduate Innovation Project of Chongqing Technology and Business University (Grant No. yjscxx2021-112-45), and Major Science and Technology Funded Project of Chongqing Education Commission (KJZD-M201900802).

References

[1] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: a survey of the state of the art," *information Fusion*, vol. 33, pp. 100–112, 2017.

[2] V. Shrinidhi, P. Yadav, and N. Venkateswaran, "IR and visible video fusion for surveillance," in *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pp. 1–6, Chennai, India, March 2018.

[3] M. X. Jiang, C. Deng, J. S. Shan, Y. Y. Wang, Y. J. Jia, and X. Sun, "Hierarchical multi-modal fusion RCN with attention model for RGB-D tracking," *Information Fusion*, vol. 50, pp. 1–8, 2019.

[4] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: benchmark and baseline," *Pattern Recognition*, vol. 96, article 106977, 2019.

[5] Y. Zhu, B. Zhu, H. H. Liu, and K. Qin, "A model-based approach for measurement noise estimation and compensation in feedback control systems," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, pp. 8112–8127, 2020.

[6] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Information Sciences*, vol. 508, pp. 64–78, 2020.

[7] A. Vishwakarma and M. K. Bhuyan, "Image fusion using adjustable nonsubsampled shearlet transform," *IEEE Transactions on Instrumentation and Measurement*, vol. 68, no. 9, pp. 3367–3378, 2019.

[8] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez, "Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability," *IEEE Transactions on Image Processing*, vol. 29, pp. 116–127, 2020.

[9] G. Piella, "A general framework for multiresolution image fusion: from pixels to regions," *Information fusion*, vol. 4, no. 4, pp. 259–280, 2003.

[10] Z. Wang, X. Li, H. Duan, X. Zhang, and H. Wang, "Multifocus image fusion using convolutional neural networks in the discrete wavelet transform domain," *Multimedia Tools and Applications*, vol. 78, no. 24, pp. 34483–34512, 2019.

[11] K. Seethalakshmi and S. Valli, "A fuzzy approach to recognize face using contourlet transform," *International Journal of Fuzzy Systems*, vol. 21, no. 7, pp. 2204–2211, 2019.

[12] H. Wei, Z. Zhu, L. Chang et al., "A novel precise decomposition method for infrared and visible image fusion," in *2019 Chinese Control Conference*, pp. 3341–3345, Guangzhou, China, July 2019.

[13] W. Ahmad, S. Vagharshakyan, M. Sjöström, A. Gotchev, R. Bregovic, and R. Olsson, "Shearlet transform-based light eld compression under low bitrates," *IEEE Transactions on Image Processing*, vol. 29, pp. 4269–4280, 2020.

[14] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: a novel decomposition method for infrared and visible image fusion," *IEEE Transactions on Image Processing*, vol. 29, pp. 4733–4746, 2020.

[15] S. Maqsood and U. Javed, "Multi-modal medical image fusion based on two-scale image decomposition and sparse representation," *Biomedical Signal Processing and Control*, vol. 57, article 101810, 2020.

[16] Q. Hu, S. Hu, and F. Zhang, "Multi-modality medical image fusion based on separable dictionary learning and Gabor filtering," *Signal Processing: Image Communication*, vol. 83, article 115758, 2020.

[17] X. Li, F. Zhou, and H. Tan, "Joint image fusion and denoising via three-layer decomposition and sparse representation," *Knowledge-Based Systems*, vol. 224, article 107087, 2021.

[18] H. Li and X.-J. Wu, "Multi-focus image fusion using dictionary learning and low-rank representation," in *International Conference on Image and Graphics*, pp. 675–686, Springer, Cham, 2017.

[19] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Information Sciences*, vol. 432, pp. 516–529, 2018.

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

[21] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Information Fusion*, vol. 36, pp. 191–207, 2017.

[22] X. Ma, S. Hu, S. Liu, J. Fang, and S. Xu, "Multi-focus image fusion based on joint sparse representation and optimum theory," *Signal Processing: Image Communication*, vol. 78, pp. 125–134, 2019.

[23] R. Gao, S. A. Vorobyov, and H. Zhao, "Image fusion with cosparse analysis operator," *IEEE Signal Processing Letters*, vol. 24, no. 7, pp. 943–947, 2017.

[24] Y. Bin, Y. Chao, and H. Guoyu, "Efficient image fusion with approximate sparse representation," *Multiresolution and Information Processing*, vol. 14, no. 4, article 1650024, 2016.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for largescale image recognition," http://arxiv.org/abs/1409.1556.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, NV, USA, 2016.

[27] H. Li and X.-J. Wu, "Densefuse: a fusion approach to infrared and visible images," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2614–2623, 2018.

[28] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "Fusiongan: a generative adversarial network for infrared and visible image fusion," *Information Fusion*, vol. 48, pp. 11–26, 2019.

[29] J. Ma, P. Liang, W. Yu et al., "Infrared and visible image fusion via detail preserving adversarial learning," *Information Fusion*, vol. 54, pp. 85–98, 2020.

[30] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," *Icml*, vol. 1, article 8, 2010.

[31] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *2011 international conference on computer vision*, pp. 1615–1622, Barcelona, Spain, Nov 2011.

[32] K. He, J. Sun, and X. Tang, "Guided image filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 6, pp. 1397–1409, 2012.

[33] H. Li, B. Manjunath, and S. K. Mitra, "Multisensor image fusion using the wavelet transform," *Graphical Models and Image Processing*, vol. 57, no. 3, pp. 235–245, 1995.

[34] X.-C. Lou and X. Feng, "Multimodal medical image fusion based on multiple latent low-rank representation," *Computational and Mathematical Methods in Medicine*, vol. 2021, 16 pages, 2021.

[35] P. J. Burt and R. J. Kolczynski, "*Enhanced image capture through fusion*," in *1993 (4th) international Conference on Computer Vision*, pp. 173–182, Berlin, Germany, May 1993.

[36] A. Toet, "TNO image fusion dataset," *Data in Brief*, vol. 15, article 249, 2017.

[37] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing*, vol. 22, no. 7, pp. 2864–2875, 2013.

[38] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Optical Engineering*, vol. 52, no. 5, article 057006, 2013.

[39] C. Liu, Y. Qi, and W. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infrared Physics & Technology*, vol. 83, pp. 94–102, 2017.

[40] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *2018 24th international conference on pattern recognition (ICPR)*, pp. 2705–2710, Beijing, China, August 2018.

[41] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infrared Physics & Technology*, vol. 102, article 103039, 2019.

[42] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *Journal of Applied Remote Sensing*, vol. 2, no. 1, article 023522, 2008.

[43] M. Hossny, S. Nahavandi, and D. Creighton, "Comments on 'information measure for performance of image fusion'," *Electronics Letters*, vol. 44, no. 18, pp. 1066-1067, 2008.

[44] C. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electronics Letters*, vol. 36, no. 4, pp. 308-309, 2000.

[45] K. Ma, K. Zeng, and Z. Wang, "Perceptual quality assessment for multi-exposure image fusion," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3345–3356, 2015.