IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 20, 2020, accepted June 3, 2020, date of publication June 12, 2020, date of current version June 24, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3001974

Infrared and Visible Image Fusion Based on a Latent Low-Rank Representation Nested With Multiscale Geometric Transform

SHEN YU, (Member, IEEE), AND XIAOPENG CHEN
School of Electronic and Information Engineering, Lanzhou Jiaotong University, Lanzhou 730070, China

Corresponding author: Xiaopeng Chen (3064683191@qq.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61861025, Grant 61562057, Grant 61761027, and Grant 51669010.

ABSTRACT To solve the problems of low image contrast and low feature representation in infrared and visible image fusion, an image fusion algorithm based on latent low-rank representation (LatLRR) and non-subsampled shearlet transform (NSST) methods is proposed. First, infrared and visible images are decomposed into base subbands, saliency subbands and sparse noise subbands by the LatLRR model. Then, the base subbands are decomposed into low-frequency and high-frequency coefficients by NSST, and a feature extraction algorithm based on VGGNet and a logical weighting algorithm based on filtering are proposed to merge the coefficients. An adaptive threshold algorithm based on the regional energy ratio is proposed to fuse the saliency subbands. Finally, the fused base subbands are reconstructed, the sparse noise subbands are discarded, and a fused image is obtained by combining the subband information after fusion. Experimental results show that for the fused image produced, the algorithm performs well in both subjective and objective evaluation.

INDEX TERMS Image fusion, latent low-rank representation, non-subsampled shearlet transform, VGG net, logical weight, energy adaptation.

I. INTRODUCTION

Image fusion involves extracting and integrating the effective information contained in two or more images collected by different sensors from the same scene through specific methods to obtain composite images with rich information and excellent visual effects and meet the needs of subsequent research and processing steps. Based on the reflection spectrum imaging of a scene, the image generated by a visible light sensor has a high spatial resolution and contains abundant background information. However, infrared sensor imaging is easily affected by the external environment. If the external lighting environment is bad, the amount of information contained in the image will decrease sharply; The infrared sensor images the radiation difference or temperature difference of the scene. Although the resulting image is of poor quality and low resolution and lacks detailed information, infrared sensor images are relatively stable and can accurately capture the hidden heat source targets in scenes, even in harsh environments [1], [2]. The fusion of infrared and visible images uses the spatiotemporal correlation between two source images and the complementary information in scene descriptions so that the fusion image can describe the scene in a detailed and comprehensive way; this approach is conducive to human visual interpretation and automatic machine-based detection. This method has been widely used in video surveillance [3], object detection [4], face recognition [5] and other processes.

Image fusion is generally performed at three levels: the pixel level, feature level and decision level. The fusion of infrared and visible images is generally performed at the pixel level, and such methods can be divided into two categories: data-driven and model-driven methods. The former directly fuses images by manipulating the pixel values of the source image, and the latter involves indirect model transformation [6]. Model-driven multiscale geometric analysis has been widely studied because of its unique multiscale analysis characteristics. Feng [7] proposed a unique image fusion method based on Tetrolet transform that uses the activity level to guide the sparse coefficients and accurately fit the decomposed low-frequency subbands; then, the number of firing

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

times of the neurons in the pulse-coupled neural network (PCNN) is used to select the coefficients of high-frequency subbands. The fused image effectively retains the edge information and detail features of the source image. Zhang and Maldague [8] proposed an effective image fusion method based on non-subsampled contourlet transform (NSCT). The decomposed low-frequency subbands were fused based on a regional adaptive energy criterion, the highest frequency coefficients in the high-frequency subbands were fused based on the absolute value method, and the remaining subband coefficients were fused with the adaptive Gaussian regional standard deviation criterion. In this approach, fusion target is clearer and the contrast is improved compared to these features in traditional methods. Chen *et al*. [9] decomposed a source image by using Laplacian pyramid transformation. At low frequencies, the weight coefficient is obtained from the pixel intensity distribution information for an infrared image, and the coefficients at high frequencies are fused by taking the maximum absolute value of the coefficient. The fusion image has abundant details, and the prominent objects in the scene are clearly visible. Liu *et al*. [10] proposed a robust image fusion algorithm based on the complex shearlet transformation, and the low-frequency coefficients were fused with two-dimensional guided filtering. Additionally, the high-frequency coefficients were fused based on the Laplace energy and maximum combined guided filtering. The resulting image performed well in subjective vision and objective evaluation tasks. Liu *et al*. [11] proposed a novel image fusion method based on non-subsampled shearlet transform (NSST). The pilot low-frequency subbands were decomposed using a pilot filter algorithm to calculate the significant mapping relations, and the high-frequency subbands were used to extract the edge information considering phase consistency; then, according to the scene consistency, the weight coefficient matrix was constructed for the fusion of subbands. The fused image could better retain the information of the source image and was smoother than the original image. Deng *et al*. [12] proposed an image fusion method based on non-subsampled double-tree complex contourlet transform. The decomposed low-frequency subbands were fused by the adaptive size segmentation method. The image block size was optimized and determined by the improved fruit fly algorithm, and the low-frequency fusion results were refined to obtain an accurate label map. The neighborhood coefficient difference of the high-frequency subbands was combined with the label graph to fuse the high-frequency subband information, and the fused image overcame the block effect generated during spatial block fusion.

With the continuous improvement of sparse representation theory, the sparse representation capability obtained through learning can increasingly improve the fusion effect of infrared and visible images. Liu *et al*. [13] proposed an image fusion method based on a convolutional sparse representation. A two-scale image decomposition and difference method was used to obtain the base subbands and detailed subbands. The different base subbands were fused using the maximum method and mean method. The convolutional sparse representation of the detailed subbands was used to encode the weighted fusion result. The superimposed reconstructed image overcome effectively preserved important details in sparse representation fusion. Chang *et al*. [14] proposed an image fusion method based on a joint sparse representation. The low-frequency subbands after quaternion wavelet transform decomposition were fused using the rules of the joint sparse representation, and the high-frequency subbands were fused using the absolute maximum value of the fusion coefficients. The reconstructed image avoided excessively smooth boundary processing and time consumption issues in traditional sparse representation fusion.

With the increase in the popularity of deep learning, new convolutional neural networks have gradually penetrated the field of image fusion [15]. Ma *et al*. [16] built and trained an end-to-end generative adversarial network model to fuse infrared and visible images. Through the continuous iterative adversarial relation between the generator and discriminator, fused images with high definition and rich details can be obtained, thus effectively avoiding the design problem of fusion rules in traditional algorithms. Liu *et al*. [17] combined a convolutional neural network and Laplace pyramid to design a fusion algorithm and constructed a weight graph to guide the image fusion process by extracting the image features from Siamese networks with shared parameters and the same structure; this approach reduced the complexity of the weighting strategy. Li and Wu [18] built a DenseNet deep learning model based on the method of dense connection and performed the fusion of infrared and visible images through encoding and decoding, thus reducing the loss of information in the fusion process.

At present, the fusion of infrared and visible images still has some problems; for example, the prominent features of fusion images are often not prominent, and noise interference can be serious. In view of the above problems, this paper proposes a fusion algorithm that combines a latent low-rank representation (LatLRR) and NSST to achieve the multilevel decomposition and fusion of infrared and visible images. The specific algorithm flow is shown in Fig. 1. By combining the feature extraction and denoising capability of LatLRR and the sparse representation capability of NSST, the optimal decomposition of the image can be obtained, and the interference caused by noise in the fused image can be reduced. The fusion rules based on the regional energy ratio reflect the significant areas and targets in a scene and improve scene recognition. A feature extraction algorithm based on VGGNet through VGG-16 is used in model training to extract the image features for deep fusion and avoid complex operations; this tool is combined with weighting rules based on a filtering algorithm to retain the maximum amount of background information and edge information, resulting in high contrast and rich details for significant features in fused images.

The main contributions of this paper are as follows.

(1) An image cascading decomposition framework with LatLRR and NSST is constructed. Compared with a single

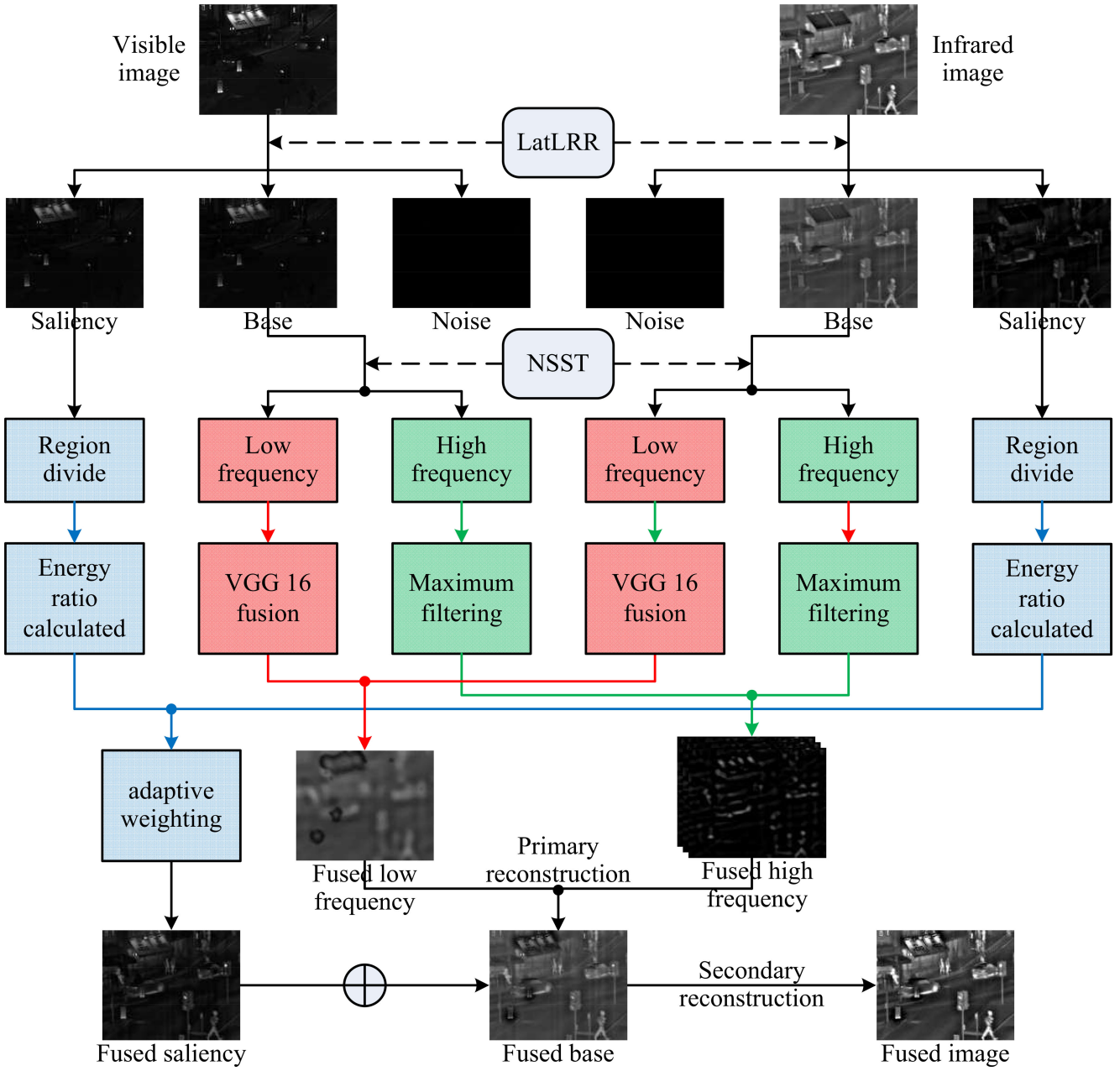


FIGURE 1. Algorithm flowchart.

LatLRR or NSST, this framework can not only better separate the important information and noise in an image but also obtain a multidirection sparse representation of the image.

(2) To fully retain the background, details, and other information from the base subbands, a feature extraction algorithm based on VGG-16 is proposed to fuse the low-frequency coefficients of the base subbands, and a logic weighting algorithm based on filtering is proposed to fuse the high-frequency coefficients of the base subbands.

(3) To highlight significant areas and targets in a scene, a threshold adaptive weighting algorithm based on the area energy ratio is proposed for subband fusion.

The remainder of this paper is organized as follows: Section II introduces the LatLRR model and the image decomposition model of NSST; Section III discusses the saliency of subbands, the fusion method of the low-frequency and high-frequency coefficients of the base subbands, and image reconstruction; Section IV explains the experimental setup used in this paper, and the results of the experiment are analyzed; Section V gives the conclusions of this study.

II. IMAGE DECOMPOSITION MODEL

A. LATENT LOW-RANK REPRESENTATION

In 2010, Liu *et al*. [19] proposed the low-rank representation (LRR) theory. In the case of determining the learning dictionary, the original data matrix Q is expressed as a linear combination of the dictionary matrix G, and the coefficient matrix is expressed as a low-rank matrix to separate data and

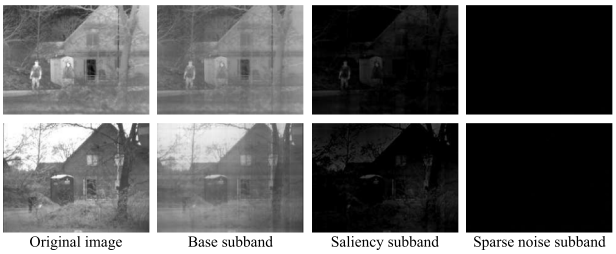


FIGURE 2. LatLRR decomposition.

noise. The mathematical model of the LRR approach is:

$$\min_{Z} ||Z||_* \quad s.t.\ Q = GZ \tag{1}$$

where, $||\cdot||_*$ is the kernel norm and Z is the optimal LRR matrix of the original data. Generally, the original data are selected to form the dictionary, i.e. $G = Q$; thus, the above equation is transformed into:

$$\min_{Z} ||Z||_* \quad s.t.\ Q = QZ \tag{2}$$

Although the LRR can represent the overall structure of the data, the local structure information of the image cannot be retained when this approach is used for image processing. In 2011, Liu and Yan [20] proposed the LatLRR approach based on LRR. LatLRR inherits the benefits of LRR and can extract the global structure and local structure of an image by considering the influence of hidden information on the learning dictionary; therefore, LatLRR has strong feature information extraction ability and denoising ability. The mathematical model of LatLRR can be expressed as:

$$\min_{Z,L,E} ||Z||_* + ||L||_* + \lambda||E||_1 \quad s.t.\ Q = QZ + LQ + E \tag{3}$$

where $\lambda > 0$ is the equilibrium coefficient and $||\cdot||_1$ is the L1 norm. The meanings of Q and Z are consistent with those in Equation (1), L is the significance coefficient matrix, and E is the sparse noise. Equation (3) can be regarded as a convex optimization problem with a kernel norm that can be solved by the inexact augmented Lagrangian multiplier (ALM) method. When LatLRR is used for image decomposition, QZ, LQ and E correspond to the base subband, salient subband and sparse noise subband of the image, respectively. Taking one group of infrared and visible images as an example, the decomposition effect of LatLRR is shown in Fig. 2.

B. NON-SUBSAMPLED SHEARLET TRANSFORM

To represent images in a sparse manner, Easley *et al.* [21] proposed the shearlet transform (ST) approach in 2007 by combining multiscale analysis with geometric analysis using the theory of affine systems. This transform has excellent multiscale, localization and orientation properties. However, the ST process involves an extraction operation, shifting occurs, which can cause Gibbs distortion in image processing. The emergence of the NSST approach [22] has effectively overcome this deficiency. NSST not only inherits the

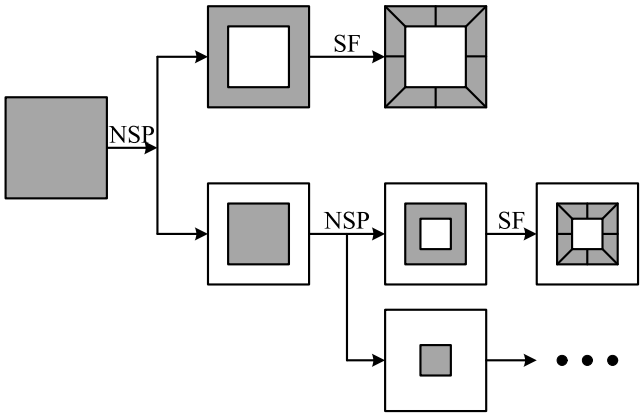


FIGURE 3. NSST decomposition process.

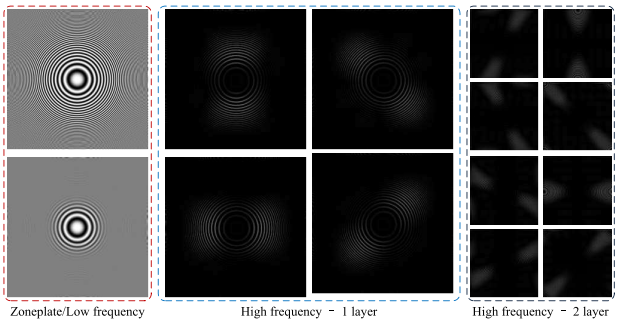


FIGURE 4. NSST decomposition subbands.

excellent characteristics of the traditional ST but also has good translation invariance for effectively extracting the edge details of images. In addition, in the process of orientation localization, the number of decomposition directions can be customized by selecting Meyer windows with variable aspect ratios, thus overcoming the limitation related to the number of decomposition directions.

The image decomposition process of NSST is divided into two parts: multiscale decomposition and directional localization. Multiscale decomposition involves the source image and is based on the non-subsampling pyramid filter bank (NSPFB) approach. One low-frequency subband and one high-frequency subband can be obtained after each decomposition. If the source image is decomposed into g levels, $g + 1$ subbands with the same size as the source image can be obtained. The focus of directional localization is the low-frequency subband after the multiscale decomposition of the source image, which is achieved with a shearlet filter (SF). If a subband is decomposed in the z direction, $2z + 2$ subbands with the same size as the atomic band can be obtained. Two-layer NSST decomposition was performed on a zone plate image, and the decomposition subband image with the direction coefficient of [4 8] is shown in Fig. 4.

III. IMAGE FUSION

A. SALIENCY SUBBAND FUSION

It is assumed that the infrared image to be fused is I and the visible image is V. Although traditional weighted fusion

can produce a low-noise and stable image, the characteristics of image (including statistical characteristics and amplitude characteristics) are randomly distributed and have a direct impact on the weighting coefficient. When the infrared image I and the visible image V are fused, if the features of image I are more prominent than those of image V, the corresponding weighting coefficient of I will be relatively large, and vice versa. Therefore, simple weighted fusion cannot perfectly integrate the characteristics of the images to be fused and fully retain the significant details.

A fusion method based on regions can mitigate this situation to a certain extent. In image fusion, fusion methods based on regions can be divided into three categories: methods based on the region energy, methods based on the region gradient and methods based on the region variance. Image fusion methods based on the region gradient and region variance do not fully consider the correlations among adjacent pixels and thus cannot reflect the local features of the image. The fusion methods based on region energy assume that the local features of an image are represented by multiple pixels in the region. The pixels in the same region have a strong correlation that reflects the local features of the image; therefore, a the fusion method based on region energy is chosen as the basis of the saliency subband fusion method [23]. To make overcome the above shortcomings, this paper proposes a new adaptive weighted fusion method based on the threshold of the regional energy ratio. To fuse images and fully retain the relevant details, the adaptive changes in the weighted coefficient are adjusted according to the continuous changes in the pixel at the region center and the corresponding region energy. The detailed process is as follows.

First, for the saliency subbands S_I and S_V of the infrared and visible images after LatLRR decomposition, the region energy levels $E_I(m, n)$ and $E_V(m, n)$ centered at pixel (m, n) can be obtained, respectively. The corresponding formula is as follows:

$$E_I(m,n) = \sum_{m' \in X, n' \in Y} \omega \times [S_I(m+m', n+n')]^2 \quad (4)$$

$$E_V(m,n) = \sum_{m' \in X, n' \in Y} \omega \times [S_V(m+m', n+n')]^2 \quad (5)$$

where m' and n' are the offsets of the pixels in the region window relative to the center pixel. X and Y represent the maximum row and column coordinates of the regional window, and the size of the regional window is generally 3×3; $\omega = \frac{1}{16} \times \begin{pmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{pmatrix}$ is the window coefficient.

Second, according to the regional energy, the regional energy ratio $E_{ratio}(m, n)$ is calculated as follows:

$$E_{ratio}(m,n) = \frac{E_I(m,n)}{E_V(m,n)} \quad (6)$$

Finally, the fused saliency subband S_F is calculated by a weighting method, and the corresponding formula is as follows

$$S_F(m,n) = w_1 \times S_I(m,n) + w_2 \times S_V(m,n) \quad (7)$$

where w_1 and w_2 are weighting coefficients. The specific formulas for these variables are as follows:

$$\begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{cases} [1 \quad 0]^T, & E_{ratio} < th1 \\ \begin{bmatrix} E_I/(E_I+E_V) \\ E_V/(E_I+E_V) \end{bmatrix}, & th1 < E_{ratio} < th2 \\ [0 \quad 1]^T, & E_{ratio} > th2 \end{cases} \quad (8)$$

where $[\cdot]^T$ is the matrix transpose operator and $th1$ and $th2$ are threshold coefficients, which are determined according to the overall energy distribution of the image. From equations (7) to (8), if the energy ratio of the region is too small or too large, the two energy values corresponding to the region will greatly differ. In such cases, the weight of the region with the higher energy level will be set to 1, and the weight of the region with the lower energy level will be set to 0. If the energy ratio of the region is within the threshold range, the two energy values of the region will be close to each other. In this case, the adaptive weight calculation is based on the energy proportion; that is, the larger the regional energy is, the larger the corresponding weighting coefficient will be and the higher the proportion will be in the combined result. Conversely, the smaller the energy level is, the smaller the contribution to the fusion result.

B. LOW-FREQUENCY COEFFICIENT FUSION FOR BASE SUBBANDS

The low-frequency component is similar to the smoothed version of the base subbands, which contain most of the important image information. To better integrate this information, the depth feature auxiliary fusion rules of image extraction are introduced by pretraining VGGNet.

VGGNet was proposed by the Visual Geometry Group of Oxford University. Compared with previous networks, VGGNet can explore the relationship between network depth and performance, and networks of 16 to 19 layers can be built [24]. VGG-16 has five convolutional groups, including 2, 2, 3, 3 and 3 convolution layers, with a total of 13 convolutional layers. Each convolution group is followed by a max pooling layer, with a total of 5 pooling layers. The fifth pooling layer is followed by 3 fully connected layers, and the final fully connected layer is followed by a softmax classifier. For a VGG-16 image input size of $N \times N$, the parameters are shown in Table. 1.

Convolutional layers can extract image features through convolution operations, and as the number of convolutional layers increases, the extracted image features become increasingly abstract. According to Table. 1, the image features extracted by the fifth convolution group of VGG-16 are too abstract. Notably, the output feature map is too different from the detailed content map of the source image. Therefore, the output of the four convolution groups in this algorithm is used as the basis to build the weight map and guide the fusion of the low-frequency coefficients. The acquisition process of the weight map is shown in Fig. 5. The specific fusion steps are as follows.

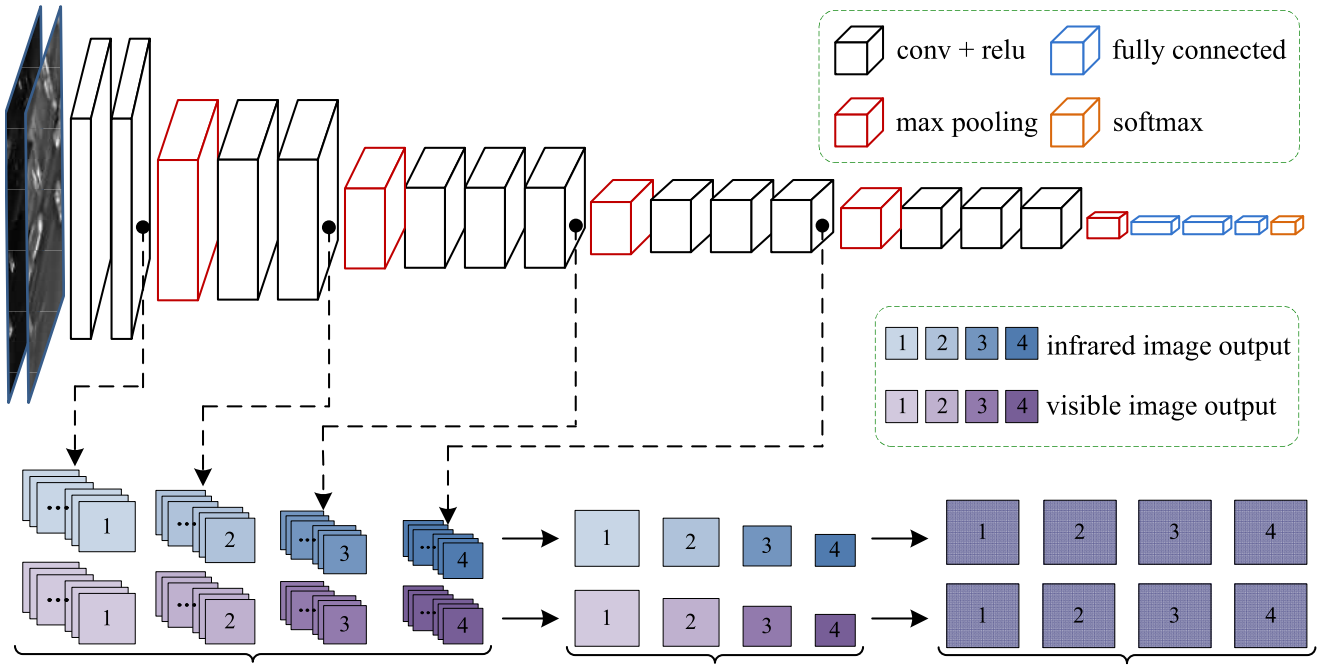


FIGURE 5. Acquisition process for the joint feature weight map.

TABLE 1. VGG 16 structure parameter.

Conv group	Size	Channel	Stride	Pooling	Output
Group 1	3×3	64	1	Max, 2×2	$N \times N$
Group 2	3×3	128	1	Max, 2×2	$\frac{N}{2} \times \frac{N}{2}$
Group 3	3×3	256	1	Max, 2×2	$\frac{N}{4} \times \frac{N}{4}$
Group 4	3×3	512	1	Max, 2×2	$\frac{N}{8} \times \frac{N}{8}$
Group 5	3×3	512	1	Max, 2×2	$\frac{N}{16} \times \frac{N}{16}$

Step 1. Input the low-frequency components $\{L_I, L_V\}$ into the VGG-16 network, and extract the output $\{O_V^{i,m}, O_I^{i,m}\}$ of the i-th convolution group, where $i = 1, 2, 3, 4$ represent the first, second, third and fourth convolution groups, respectively. m is the number of channels in the output characteristic map of the i-th convolution group, which is determined by the number of convolution kernels; notably, $m \in 1, 2, \ldots, M, M = 64 \times 2^{i-1}$. According to the number of convolution kernels in each convolutional layer of the VGG-16 network, as shown in Table. 1, $O^{i,1:M}$ is an M-dimensional vector, and $O^{i,1:M}(x, y)$ represents the value of $O^{i,m}$ at position (x, y). The output multichannel feature map is compressed according to the L1 norm, and the single-channel feature map $\{C_V^i, C_I^i\}$ is obtained. The formula is as follows:

$$\begin{aligned} C_V^i(x, y) &= ||O_V^{i,1:M}(x, y)||_1 \\ C_I^i(x, y) &= ||O_I^{i,1:M}(x, y)||_1 \end{aligned} \tag{9}$$

Step 2. To improve the fused image, an average filter of size 3×3 is introduced to smooth the single-channel feature map, and the normalized adaptive weight $\{w_V^i, w_I^i\}$ is then calculated. The corresponding formula is as follows:

$$\begin{aligned} w_V^i(x, y) &= \frac{C_V^i(x, y)}{C_V^i(x, y) + C_I^i(x, y)} \\ w_I^i(x, y) &= 1 - w_V^i(x, y) \end{aligned} \tag{10}$$

Step 3. The pooling layer in the convolutional neural network employs a data sampling operation. The size of the feature map after pooling is changed to $\frac{1}{s}$, and s is the step of the pooling operator. The value of s in VGG-16 is fixed at 2, and the size of the feature map output by different convolution groups is $\frac{1}{2^{i-1}}$ of the original image size. According to the size consistency principle of image fusion, upsampling is used for the size reconstruction of the weight map $\{w_V^i, w_I^i\}$, and four groups of weight maps $\{W_V^i, W_I^i | i = 1, 2, 3, 4\}$ with sizes consistent with those of the original base subbands are obtained; then, the low-frequency coefficients are fused. The formula is as follows:

$$\begin{aligned} L_F^i(x, y) = W_{VIS}^i(x, y) \times L_V(x, y) \\ + W_{IR}^i(x, y) \times L_I(x, y) \end{aligned} \tag{11}$$

Step 4. The fused low-frequency coefficients have four values at each position (x, y). Four values at the same position are selected according to the maximum absolute value criterion to obtain the fused low-frequency coefficients L_F:

$$L_F(x, y) = \max\{L_F^1(x, y), L_F^2(x, y), L_F^3(x, y), L_F^4(x, y)\} \tag{12}$$

According to the above equation, compared with the use of a single network model of the final output of a fusion algorithm, this algorithm can minimize the loss of image detail, provide comprehensive and effective information, improve

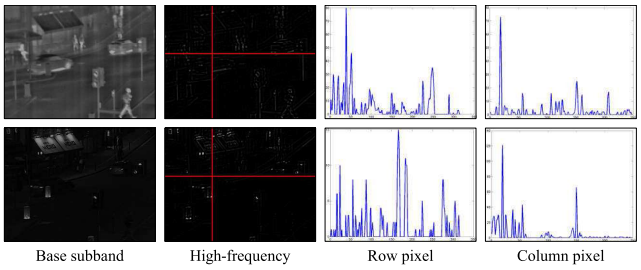


FIGURE 6. Comparison of row pixels and column pixels.

the depth of image feature extraction, and promote the fusion effect.

C. HIGH-FREQUENCY COEFFICIENT FUSION FOR BASE SUBBANDS

The low-frequency coefficients of an image concentrate most of the energy of the image and reflect the brightness of the image, and the high-frequency coefficients represent the details and edges of the image and reflect the texture characteristics of the image. For the high-frequency coefficients of the base subband, fusion is mainly performed to get clear edge areas and rich texture information to improve image clarity. Fig. 6 shows the pixel value curves of the high-frequency coefficients of the base subband and the randomly selected high-frequency coefficients in row 106 and column 99 (marked by the crossing of the red lines).

Fig. 6 shows that the high-frequency parts of the infrared and visible images are quite different from each other, so the fusion strategy for the high-frequency coefficients should focus on enhancing the detailed information in the fused image. Therefore, for the high-frequency coefficients $\{H_I^j, H_V^j\}$ of the base subbands $\{B_I, B_V\}$, a logical weighting method based on maximum filtering is proposed in this paper. The specific steps are as follows.

First, the absolute value of the high-frequency coefficients is obtained, and a local window w of size 3×3 is used to filter the maximum value. The corresponding mathematical expression is as follows:

$$\begin{aligned} Z_I^j(\ddot{m}, \ddot{n}) &= \max_w\{|H_I^j(\ddot{m}, \ddot{n})|\} \\ Z_V^j(\ddot{m}, \ddot{n}) &= \max_w\{|H_V^j(\ddot{m}, \ddot{n})|\} \end{aligned} \tag{13}$$

where $\{Z_I^j, Z_V^j\}$ is the maximum filtering result of the high-frequency coefficients $\{H_I^j, H_V^j\}$.

Second, a logic map is obtained by comparing the difference value $U(\ddot{m}, \ddot{n})$ and threshold value T of $Z_I^j(\ddot{m}, \ddot{n})$ and $Z_V^j(\ddot{m}, \ddot{n})$:

$$Map(\ddot{m}, \ddot{n}) = \begin{cases} 1, & U(\ddot{m}, \ddot{n}) > T \\ 0, & U(\ddot{m}, \ddot{n}) \leq T \end{cases} \tag{14}$$

Finally, the fused high-frequency coefficients are obtained through fusion according to the logic map. The applied formula is as follows:

$$\begin{aligned} H_F^j(\ddot{m}, \ddot{n}) = Map(\ddot{m}, \ddot{n}) \times H_I^j(\ddot{m}, \ddot{n}) \\ +[\sim Map(\ddot{m}, \ddot{n})] \times H_V^j(\ddot{m}, \ddot{n}) \end{aligned} \tag{15}$$

where the elements of $Map(\ddot{m}, \ddot{n})$ are logical values of 0 or 1, $\sim Map(\ddot{m}, \ddot{n})$ is the inverse of the previous variable, and $(\ddot{m}, \ddot{n})$ are the position coordinates. From equations (14) to (15), after the high-frequency coefficients are filtered to identify the maximum, if the coefficient difference at the corresponding position is greater than the preset threshold, then the logical value corresponding to that point in the logical graph is 1; otherwise, the value is 0. In the process of logical weighted fusion, the points with relatively obvious features correspond to a logical weight of 1, and the weights of the corresponding points are then reversed to preserve details.

D. IMAGE RECONSTRUCTION

Inverse NSST was performed on the fused low-frequency coefficients L_F and high-frequency coefficients H_F^j to obtain the fused base subband B_F, and the sparse noise subband $\{N_I, N_V\}$ was discarded. Secondary reconstruction was performed by combining the fused saliency subband S_F to obtain the fusion image F. The corresponding formula is as follows:

$$F = B_F + S_F \tag{16}$$

IV. EXPERIMENTAL EVALUATION

A. PARAMETER SETTING

To verify the feasibility and effectiveness of the proposed algorithm, infrared and visible images were selected from the TNO Image Fusion Dataset for fusion experiments. The experimental simulation platform was equipped with an Intel Core i7-9700k CPU with a main frequency of 3.6 GHz, 64 GB running memory, and a 64-bit Windows 10 system; the programming environment was MATLAB 2018b.

The parameters of this algorithm were set as follows: the number of LatLRR decomposition layers was 1, the number of NSST decomposition layers was 4, the filter parameter was ''maxflat'', and the number of decomposition directions was [8, 8, 16, 16]. The regional energy ratio threshold coefficients were $th1 = 2$ and $th2 = 4$, the image patch size of the sparse representation was 8×8, the sliding step was $s = 1$, the error tolerance was $\epsilon = 0.1$, and the threshold was $T = 4$.

B. FEASIBILITY ASSESSMENT

For feasibility assessment, 32 groups of infrared and visible images of the Camp sequence were selected for experiments. By observing the image effect after fusion and calculating the entropy (EN) [25], average gradient (AvG), spatial frequency (SF) [26] and standard deviation (Std) of the source image of the sequence and the fused image, four quality evaluation indexes of the unreferenced image were comprehensively evaluated. EN is used to measure the information contained in the image, AvG is used to measure the gradient of the image, SF is used to measure the rate of change of the image gray level, and Std is used to measure the dispersion degree of the

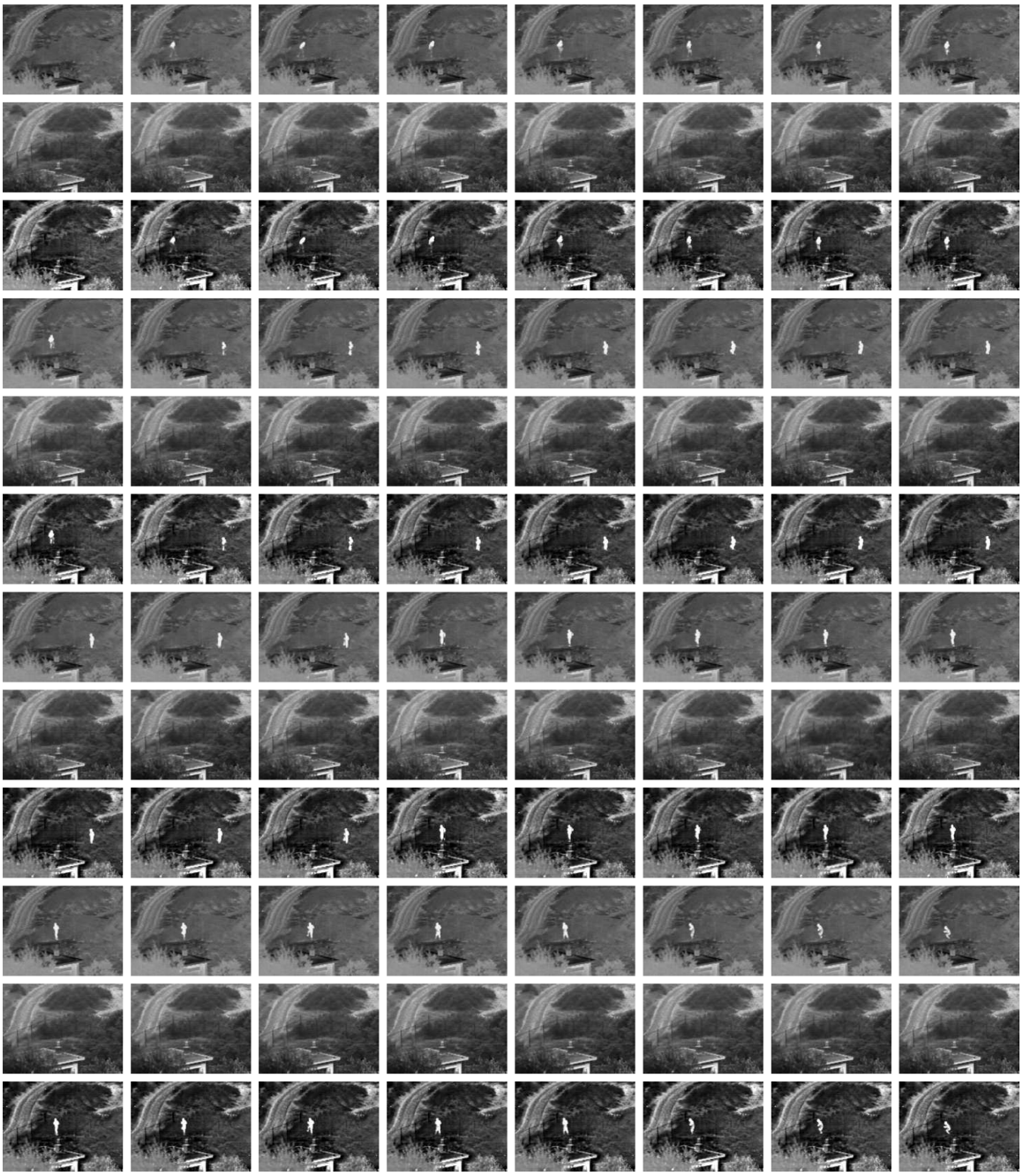


FIGURE 7. Image sequence fusion.

gray level of pixels. These metrics are positive indicators, and the larger the value is, the better the result.

The Camp sequence image fusion results are shown in Fig. 7. Among them, rows 1, 4, 7 and 10 are infrared images used to capture heat source objects in the scene, and rows 2, 5, 8 and 11 are visible images used to capture detailed information related to houses, fences, and trees in the scene. Rows 3, 6, 9 and 12 are the fused images. According to the fusion results, the fusion algorithm in this paper can effectively integrate the contour features of heat source objects in the infrared image with the background details of the visible image, thus making the fused image more clear and the contrast enhanced, which reflects the observation capabilities of human vision.

The four indexes of Camp sequence source image and fusion image are shown in Fig. 8, including EN, AvG, SF and Std from top to bottom. As shown in the line chart, since the scene in the entire sequence of images does not change much, the curve of a single image index does not considerably fluctuate. Compared with the source image, the fused image

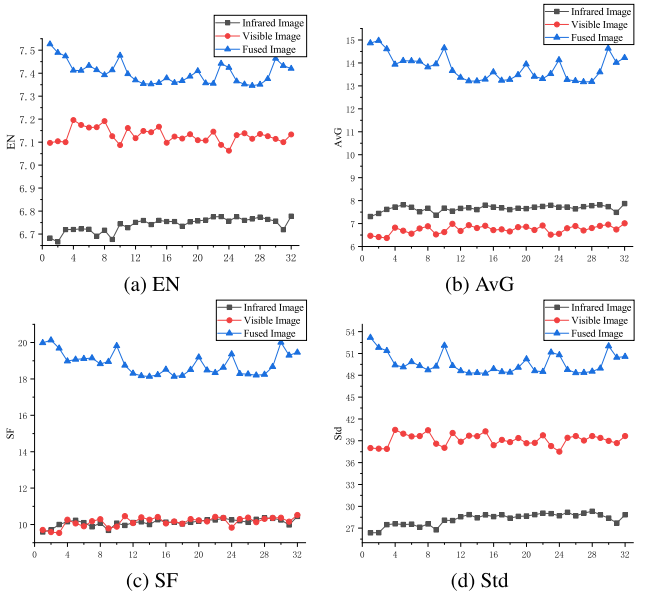


FIGURE 8. Image sequence evaluation indexes.

has different degrees of superiority for the four indicators, indicating that the fused image includes richer information and achieves a better overall visual effect, which is consistent with the subjective visual evaluation of human observation. In summary, this algorithm can integrate effective information in infrared and visible images to achieve image fusion, and the visual effect of the fused image is good; therefore, this approach can feasibly further improve the recognition of a scene.

C. VALIDITY ANALYSIS

Effectiveness analysis was performed to select 7 different groups of infrared and visible images. The fusion results of the algorithm in this paper are compared with those of guided filtering based fusion (GFF) [27], a convolution sparse representation (CSR) method [12], a joint sparse representation (JSR) method [28], a joint sparse representation method based on saliency detection (JSRSD) [29], cross-bilateral filtering (CBF) [30], a Siamese convolution neural network (SCNN) [17], and DenseFuse [18]. The results were compared by subjective and objective evaluations.

Subjective evaluation involves intuitively assessing the advantages and disadvantages of the fusion results based on human vision. The fusion results of seven groups of infrared images and visible images are shown in Fig. 9, where row (a) shows the infrared images, (b) shows the visible images, (c) (i) present the fusion image of compared methods, and (j) gives the fusion image produced by the algorithm developed in this paper. From the fusion results, the images of the comparative methods have two problems: first, the fused images are often incomplete, resulting in the loss of image information; second, the fused images are generally dark, and regional fusion distortion and the loss of detailed information occur. For the above reasons, the visual effect of the fused images is generally not good, and the images do not effectively improve the visualization of the scenes. The two methods based on deep learning produce an image with moderate brightness and rich details, but the recognition of significant objects in the scene is poor. The fused image obtained by the algorithm proposed in this paper has high contrast and can fully integrate the effective information from the source image, making the contours of the target objects in the scene clear; therefore, this approach is conducive to understanding the scene.

To observe the differences among the fused images, local areas were selected from the third, fourth, fifth and seventh groups of images for amplification, as shown in Fig. 10. Clearly, the subjective effect is largely consistent with the previous description. Based on the subjective evaluation results of the above seven groups of images, it can be concluded that the algorithm proposed in this paper provides prominent image features and a good visual effect; therefore, the fused image is consistent with the observation behaviors of the human eye vision system.

Subjective evaluation can be used to directly determine the merits of the fusion result, but there are differences in visual sensitivity among different people, and the evaluation result can be one sided. Therefore, comprehensive evaluation should be based on comprehensive and objective evaluation indexes. Based on the SF and EN indexes, four forward evaluation indexes were calculated: the peak signal-to-noise ratio (PSNR), $Q^{ab/f}$ [31], the sum of the correlations of differences (SCD) [32] and structural similarity (SSIM) [33].

► Peak signal-to-noise ratio (PSNR)

The PSNR is the ratio between the important information in an image and the noise, and it is used to measure the distortion degree of a fused image in the fusion process. The PSNR of image X and reference image R is defined as:

$$PSNR_{X,R} = 10\log\frac{k^2}{\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(X(i,j)-R(i,j))^2} \quad (17)$$

where M and N represent the image size dimensions, (i, j) represents the pixel position, and k represents the maximum grayscale level of the image. The PSNR formula used in this experiment is as follows:

$$PSNR = (PSNR_{I,F} + PSNR_{V,F})/2 \quad (18)$$

where $PSNR_{I,F}$ and $PSNR_{V,F}$ are the PSNRs of fused image F when infrared image I and visible image V are used as reference images, respectively. The larger the value of the PSNR is, the better the image fusion effect.

► $Q^{ab/f}$

$Q^{ab/f}$ is used to measure the transfer of edge information from the source image to the fused image in the fusion

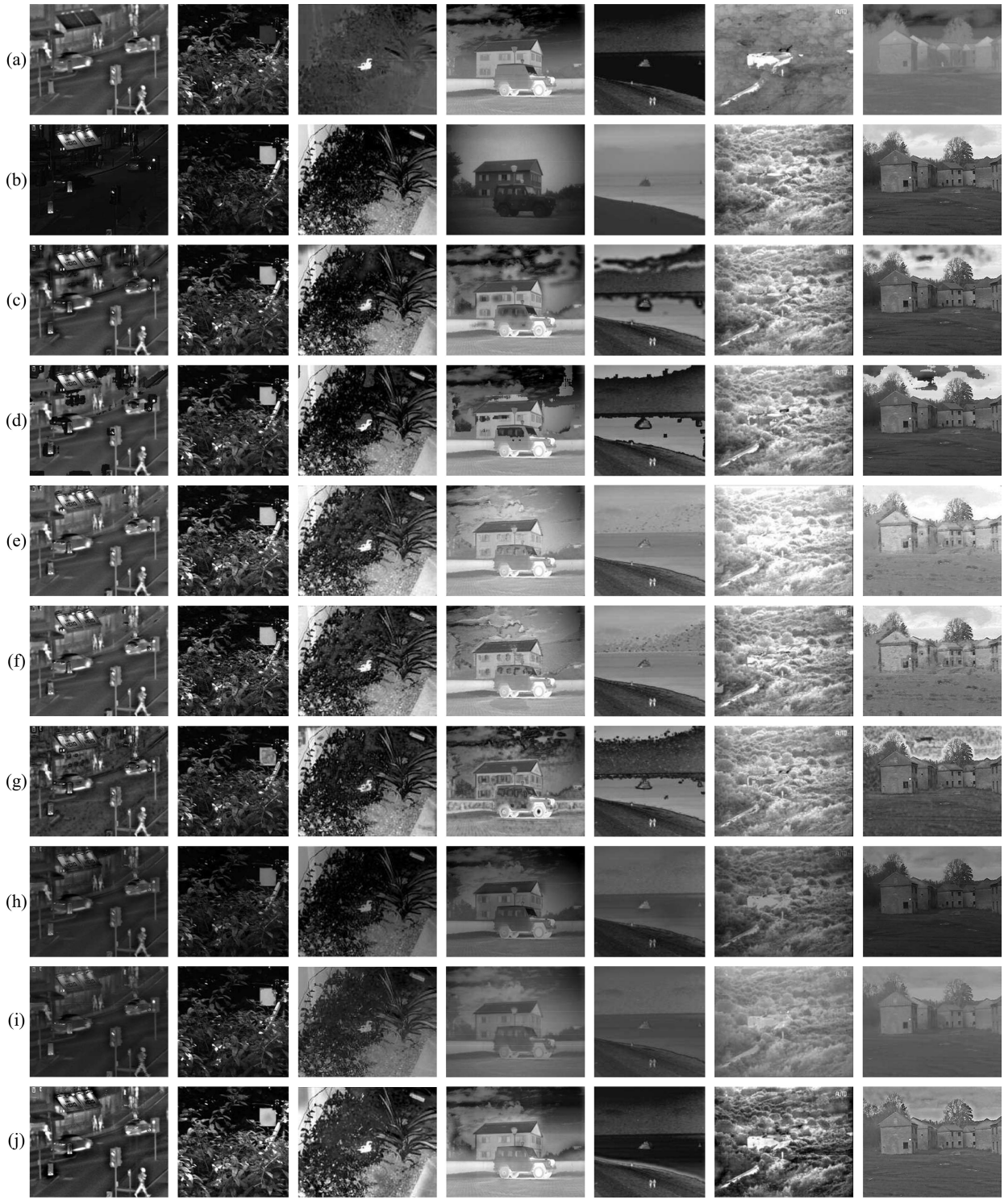


FIGURE 9. Image fusion. (a) Infrared image. (b) Visible image. (c) GTF. (d) CSR. (e) JSR. (f) JSRSD. (g) CBF. (h) SCNN. (i) DenseFuse. (j) Our method.

process, and it is defined as follows:

$$Q^{ab/f} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N}(Q^{AF}(i,j)\omega^{A}(i,j) + Q^{BF}(i,j)\omega^{B}(i,j))}{\sum_{i=1}^{M}\sum_{j=1}^{N}(\omega^{A}(i,j) + \omega^{B}(i,j))} \quad (19)$$

where M and N represent the image size dimensions, (i, j) represents the pixel position, and $\{Q^{AF}, Q^{BF}\}$ are the edge strength and orientation preservation values, respectively. $\{\omega^{A}, \omega^{B}\}$ are the weights that express the importance of each source image to the fused image. The closer the value of $Q^{ab/f}$ is to 1, the better the retention effect of edge information.

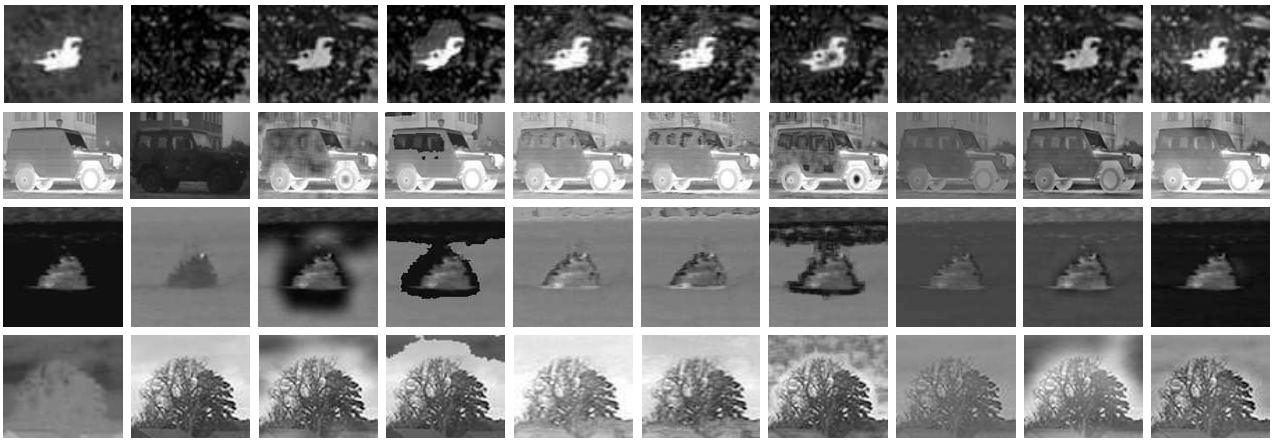


FIGURE 10. Image block from left to right: Infrared image, Visible image, GFF, CSR, JSR, JSRSD, CBF, SCNN, DenseFuse, Our method.

► Sum of the correlations of differences (SCD)

The SCD is used to measure the sum of the difference in complementary information between a fused image and a different image (including the difference between a fused image and an infrared image or a visible image); this variable is defined as follows:

$$SCD = r(F, F - I) + r(F, F - V) \tag{20}$$

$$r(F, X) = \frac{\sum_{i}^{M}\sum_{j}^{N}(F(i,j) - \bar{F})(X(i,j) - \bar{X})}{\sqrt{(\sum_{i}^{M}\sum_{j}^{N}(F(i,j) - \bar{F})^2)(\sum_{i}^{M}\sum_{j}^{N}(X(i,j) - \bar{X})^2)}} \tag{21}$$

where M and N represent the image size dimensions, (i, j) represents the pixel position, and $\bar{F}, \bar{X}$ are the average pixel values of F, X, respectively. The $r(F, X)$ function is used to calculate the correlation between fused image F and the image difference X. The higher the value of the SCD is, the higher the quality of the fused image.

► Structural similarity (SSIM)

SSIM is used to measure the structural similarity of two images, with a numerical range of [0, 1]. From the perspective of the image composition, the definition of SSIM reflects the distortion of an image and considers the image brightness, contrast and structure. The mean is the estimation of brightness, the standard deviation is the estimation of contrast, and the covariance is the measurement of SSIM.

The SSIM between image X and image F is defined as:

$$SSIM_{X,F} = \sum_{x,f} \frac{2\mu_x\mu_f + C_1}{\mu_x^2 + \mu_f^2 + C_1} \cdot \frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2 + \sigma_f^2 + C_2} \cdot \frac{\sigma_{xf} + C_3}{\sigma_x\sigma_f + C_3} \tag{22}$$

where x, f are the image blocks of images X and F, respectively; μ_x, μ_f are the mean values; σ_x, σ_f are the standard deviations; σ_{xf} is the covariance; and C_1, C_2, C_3 are the stable constants of the algorithm. The formula for calculating the SSIM in this experiment is as follows:

$$SSIM = 0.5 \times SSIM_{I,F} + 0.5 \times SSIM_{V,F} \tag{23}$$

where $SSIM_{I,F}$, $SSIM_{V,F}$ represent the structural similarity of infrared image I and fused image F and visible image V and fused image F, respectively. The larger the SSIM value is, the better the fusion effect. Six evaluation indexes for seven groups of images are shown in Fig. 11.

As shown in Fig. 11, the algorithm presented in this paper performs well based on the SF and EN indexes, and it is superior to most of the other methods to different degrees, indicating that the fused image contains abundant information and has an appropriate level of brightness and high definition. The SCD index of the proposed method is much higher than those of the other methods, suggesting that important information is transferred from the source image to the fused image. The $Q^{ab/f}$ and SSIM index differences are not large among models, and the fused image and overall structure of the source image are generally similar. The PSNR index margin is also small because the transfer between the original and fused images corresponding to pixel decisions did not consider the visual characteristics of human vision. Additionally, the sensitivity of the error is not absolute, and objective and subjective evaluation results can yield different outcomes. In summary, the objective evaluation results are generally consistent with the subjective evaluation results, and the algorithm in this paper has certain advantages over the other algorithm investigated; therefore, this new method is practical and effective.

The complexity of the algorithm is generally measured by the temporal complexity; therefore, the run times of different algorithms under the same conditions were determined. In this approach, the complexity of the initial algorithm can be reflected, and the computational efficiency of the algorithm can be evaluated. Therefore, to evaluate the performance of the algorithm proposed in this paper, the run time results of the algorithms discussed above are obtained and compared. The results are shown in the Fig. 12.

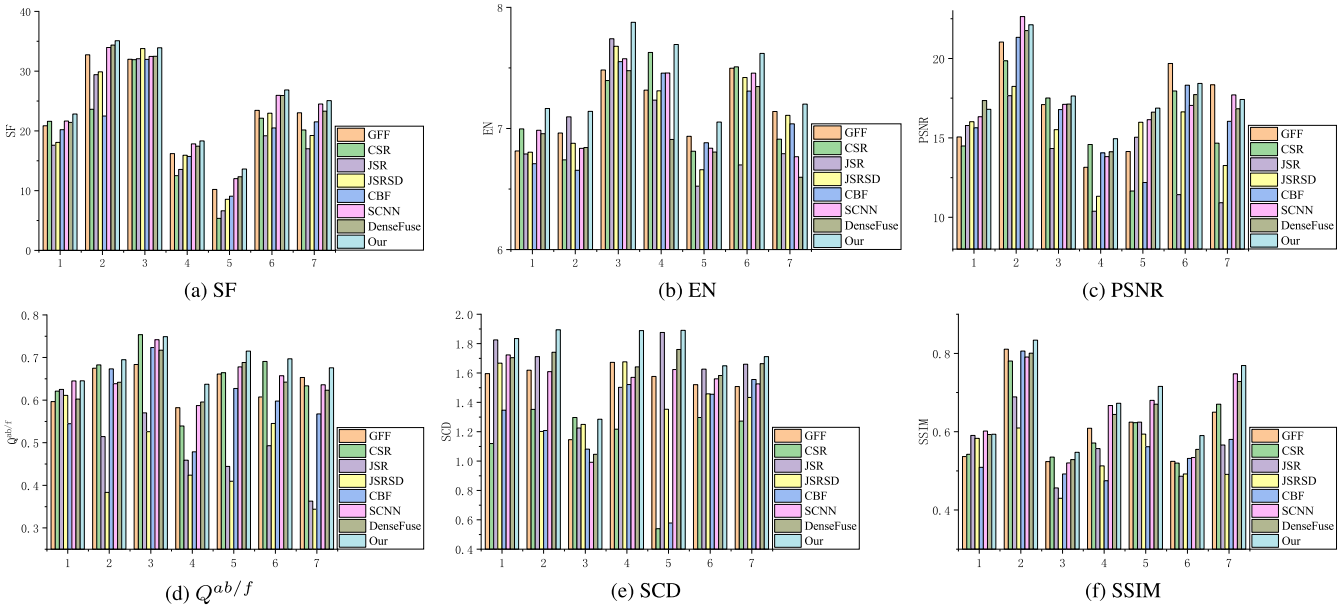


FIGURE 11. Image sequence evaluation indexes.

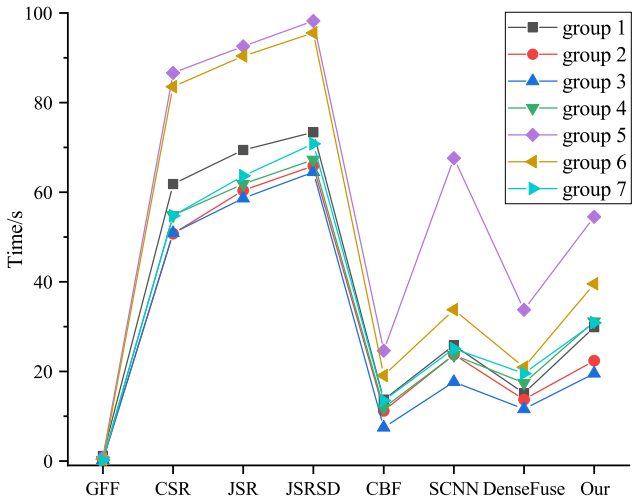


FIGURE 12. Time complexity comparison.

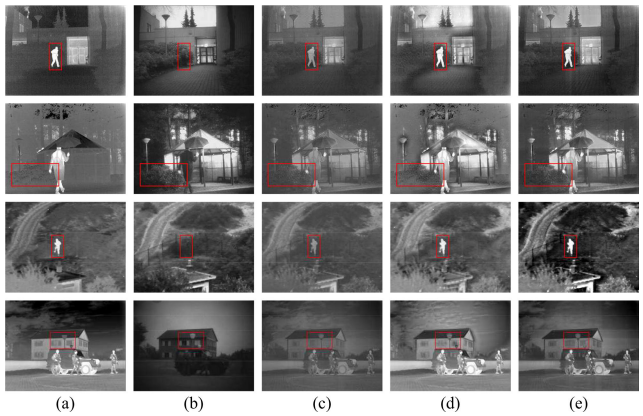


FIGURE 13. Image fusion; (a) Infrared image; (b) Visible image; (c) LatLRR; (d) NSST; (e) LatLRR+NSST.

In the comparison, the GFF and CBF algorithms are based on the theory of traditional filtering, and fusion rule is simple; therefore, the algorithm run time is relatively short, and the image fusion effect is limited. The CSR, JSR, and JSRSD methods had high run times, mainly because the three algorithms are all based on the sparse representation algorithm was improved in this paper. Notably, the number of required operations makes the complexity of the algorithm high. The SCNN and DenseFuse algorithms are based on deep learning models without training. The corresponding testing phase does not require much time, and the fusion effect is satisfactory. Although the nested-frame image decomposition algorithm in this paper takes more time than some other methods, the network model uses fusion rules in the training process to avoid information loss in the early stage of training. Overall, the complexity of the approach is moderate, and the fusion effect is the best among those of all the models considered, thus reflecting generally superior performance.

To verify the correlations obtained by the LatLRR and NSST algorithms were compared with that of the hybrid algorithm proposed in this paper. The parameter setting and fusion rules of the algorithms were consistent with those used for the model developed in this paper. The fusion results of the four groups of images are shown in Fig. 13.

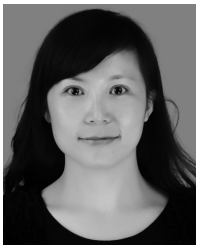
Notably, the LatLRR algorithm and NSST algorithm can achieve image fusion for a cognitive scene, but the advantages and disadvantages are also relatively obvious. Specifically, the expression of the target area is not outstanding, but detail processing is good, as most image details are retained. However, the edges of targets are largely insufficient.For the algorithm proposed in this paper, the fused image is clearer and has more details, contrast and definition than the other images. Thus, this fused image reflects human visual patterns, making image scenes easy to understand.

V. CONCLUSION

This paper proposes a model based on a combined LatLRR and NSST algorithm for image fusion. LatLRR is used to extract the image features and perform simple filtering. The multiscale characteristics of NSST and the feature extraction capability of VGG-16 can retain information in a source image; combined with the regional energy intensity ratio fusion rules, the characteristics of the original image can be maintained in the fused image. The proposed logical weighting method based on maximum filtering can enhance contour edge detail, thereby improving the image resolution, the richness of details, and target edges. By comparing source images and fused images based on four evaluation indexes, we verify that the proposed algorithm is feasible. The comparison of the fused images of the proposed algorithm and those of seven other methods shows that the algorithm developed in this paper can fully integrate the information from the source image and highlight the prominent features.

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] 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.
[3] Q. Zhang, Y. Wang, M. D. Levine, X. Yuan, and L. Wang, "Multisensor video fusion based on higher order singular value decomposition," *Inf. Fusion*, vol. 24, pp. 54–71, Jul. 2015.
[4] S. Gao, Y. Cheng, and Y. Zhao, "Method of visual and infrared fusion for moving object detection," *Opt. Lett.*, vol. 38, pp. 1981–1983, Jun. 2013.
[5] R. Singh, M. Vatsa, and A. Noore, "Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition," *Pattern Recognit.*, vol. 41, no. 3, pp. 880–893, Mar. 2008.
[6] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin, "Pixel-level image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 33, pp. 100–112, Jan. 2017.
[7] X. Feng, "Fusion of infrared and visible images based on Tetrolet framework," *Acta Photonica Sinica*, vol. 48, no. 2, pp. 76–84, Feb. 2019.
[8] Q. Zhang and X. Maldague, "An adaptive fusion approach for infrared and visible images based on NSCT and compressed sensing," *Infr. Phys. Technol.*, vol. 74, pp. 11–20, Jan. 2016.
[9] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma, "Infrared and visible image fusion based on target-enhanced multiscale transform decomposition," *Inf. Sci.*, vol. 508, pp. 64–78, Jan. 2020.
[10] S. Liu, M. Shi, Z. Zhu, and J. Zhao, "Image fusion based on complex-shearlet domain with guided filtering," *Multidimensional Syst. Signal Process.*, vol. 28, no. 1, pp. 207–224, Jan. 2017.
[11] Z. Liu, Y. Feng, H. Chen, and L. Jiao, "A fusion algorithm for infrared and visible based on guided filtering and phase congruency in NSST domain," *Opt. Lasers Eng.*, vol. 97, pp. 71–77, Oct. 2017.
[12] H. Deng *et al.*, "Fusion of infrared and visible images based on non-subsampled dualtree complex contourlet and adaptive block," *Acta Photonica Sinica*, vol. 48, no. 7, pp. 136–146, Jul. 2019.
[13] 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.
[14] L. Chang, X. Feng, and R. Zhang, "Image fusion scheme based on quaternion wavelet transform and sparse representation," *Syst. Eng. Electron.*, vol. 39, no. 7, pp. 1633–1639, Jul. 2017.
[15] Y. Liu, X. Chen, Z. Wang, Z. J. Wang, R. K. Ward, and X. Wang, "Deep learning for pixel-level image fusion: Recent advances and future prospects," *Inf. Fusion*, vol. 42, pp. 158–173, Jul. 2018.
[16] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang, "FusionGAN: A generative adversarial network for infrared and visible image fusion," *Inf. Fusion*, vol. 48, pp. 11–26, Aug. 2019.
[17] Y. Liu, X. Chen, J. Cheng, H. Peng, and Z. Wang, "Infrared and visible image fusion with convolutional neural networks," *Int. J. Wavelets, Multiresolution Inf. Process.*, vol. 16, no. 3, May 2018, Art. no. 1850018.
[18] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.
[19] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proc. 27th Int. Conf. Mach. Learn.*, Haifa, Israel, 2010, pp. 663–670.
[20] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, 2011, pp. 1615–1662.
[21] G. R. Easley, D. Labate, and W.-Q. Lim, "Optimally sparse image representations using shearlets," in *Proc. 14th Asilomar Conf. Signals, Syst. Comput.*, Pacific Grove, CA, USA, 2006, pp. 974–978.
[22] G. Easley, D. Labate, and W.-Q. Lim, "Sparse directional image representations using the discrete shearlet transform," *Appl. Comput. Harmon. Anal.*, vol. 25, no. 1, pp. 25–46, Jul. 2008.
[23] R. Srivastava, A. Khare, and O. Prakash, "Local energy-based multimodal medical image fusion in curvelet domain," *IET Comput. Vis.*, vol. 10, no. 6, pp. 513–527, Sep. 2016.
[24] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Representations*, San Diego, CA, USA, 2015, pp. 345–358.
[25] J. Van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.
[26] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[27] J. Ma, C. Chen, C. Li, and J. Huang, "Infrared and visible image fusion via gradient transfer and total variation minimization," *Inf. Fusion*, vol. 31, pp. 100–109, Sep. 2016.
[28] Q. Zhang, Y. Fu, H. Li, and J. Zou, "Dictionary learning method for joint sparse representation-based image fusion," *Opt. Eng.*, vol. 52, no. 5, May 2013, Art. no. 057006.
[29] C. H. Liu, Y. Qi, and W. R. Ding, "Infrared and visible image fusion method based on saliency detection in sparse domain," *Infr. Phys. Technol.*, vol. 83, pp. 94–102, Jun. 2017.
[30] B. K. Shreyamsha Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, Jul. 2015.
[31] C. S. Xydeas and V. Petrovic, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, p. 308, 2000.
[32] 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.
[33] Z. Wang and A. C. Bovik, "A universal image quality index," *IEEE Signal Process. Lett.*, vol. 9, no. 3, pp. 81–84, Mar. 2002.



SHEN YU (Member, IEEE) was born in Shandong, China, in 1982. She received the master's and Ph.D. degrees from the School of Electronics and Information Engineering, Lanzhou Jiaotong University, in 2008 and 2017, respectively.

She has presided over and participated in many projects, such as the National Natural Science Foundation of China, the Natural Science Foundation of Gansu Province, and the Science and Technology Program of the Youth Foundation. Her main research interests include deep learning and digital image processing. She is a member of the China Computer Society, ACM, and the Lanzhou Branch of China Computer Science and Technology (Youth Computer Technology Forum). She received the Title of Professor, in 2019.

XIAOPENG CHEN was born in Pingdingshan, Henan, China, in 1994. He received the bachelor's degree in communication engineering from the Luoyang University of Technology, in 2018. He is currently pursuing the master's degree with the School of Electronics and Information Engineering, Lanzhou Jiaotong University.

His major research interests include signal and information processing, deep learning, and image fusion.