IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received March 7, 2020, accepted March 17, 2020, date of publication March 23, 2020, date of current version April 8, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2982712

Improving the Performance of Image Fusion Based on Visual Saliency Weight Map Combined With CNN

LEI YAN[1], JIE CAO[1], SAAD RIZVI[1], KAIYU ZHANG[1], QUN HAO[1,2,3], AND XUEMIN CHENG[2,3]

[1]Key Laboratory of Biomimetic Robots and Systems, Ministry of Education, School of Optics and Photonics, Beijing Institute of Technology, Beijing 100081, China
[2]Graduate School at Shenzhen, Tsinghua University, Beijing 518055, China
[3]Department of Precision Instrument, Tsinghua University, Beijing 518055, China

Corresponding author: Jie Cao (ajieanyyn@163.com)

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 61875012 and Grant 61871031, and in part by the Natural Science Foundation of Beijing Municipality under Grant 4182058.

ABSTRACT Convolutional neural networks (CNN) with their deep feature extraction capability have recently been applied in numerous image fusion tasks. However, the image fusion of infrared and visible images leads to loss of fine details and degradation of contrast in the fused image. This deterioration in the image is associated with the conventional ''averaging'' rule for base layer fusion and relatively large feature extraction by CNN. To overcome these problems, an effective fusion framework based on visual saliency weight map (VSWM) combined with CNN is proposed. The proposed framework first employs VSWM method to improve the contrast of an image under consideration. Next, the fine details in the image are preserved by applying multi-resolution singular value decomposition (MSVD) before further processing by CNN. The promising experimental results show that the proposed method outperforms state-of-the-art methods by scoring the highest over different evaluation metrics such as Q_0, multiscale structural similarity (MS_SSIM), and the sum of correlations of differences (SCD).

INDEX TERMS Convolutional neural network, image fusion, visual saliency weight map, multi-resolution singular value decomposition.

I. INTRODUCTION

Image information captured by multiple sensors can provide accurate complementary information through image fusion [1]. Compared to an image acquired by a single sensor, the composite image generated by image fusion provides good visualization and rich information. Hence, image fusion is widely employed in many fields, such as remote sensing [2], [3], pattern recognition [4], [5], medical imaging [6], [7], and military [8], [9].

Numerous fusion methods have been proposed in the past which achieve good fusion performance. Typical fusion methods include multiscale based methods [7], [10]–[15], sparse representation based fusion methods [16]–[18], and hybrid transformation methods [19]. Recently, deep learning has been successfully applied in many image processing tasks, such as image matting [20], [21], space clustering [22], image super-resolution [23]–[25] and face processing [26]. Particularly, the application of deep learning in image fusion has attracted considerable scholarly attention [27]–[32]. For example, Yu *et al.* [33] first used the convolutional neural network (CNN) for multi-focus image fusion by solving a pixel based classification problem. Although the CNN-based method achieves better performance, it is only suitable for multi-focus image fusion. A powerful CNN was recently used by Gatys *et al.* [34] for image style transfer. The method specifically employed the VGG network to extract deep features at various layers from ''content'' and ''style'' to generate images of high perceptual quality. Inspired by high-level image synthesis and manipulation capability of VGG, Hui *et al.* [35] applied the VGG-19 network to achieve infrared (IR) and visible (VIS) image fusion, obtaining better performance than conventional methods. Their approach decomposed images into base and detail layers. The base layer was obtained by applying the ''averaging'' rule, while the detail layer fusion was based on

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

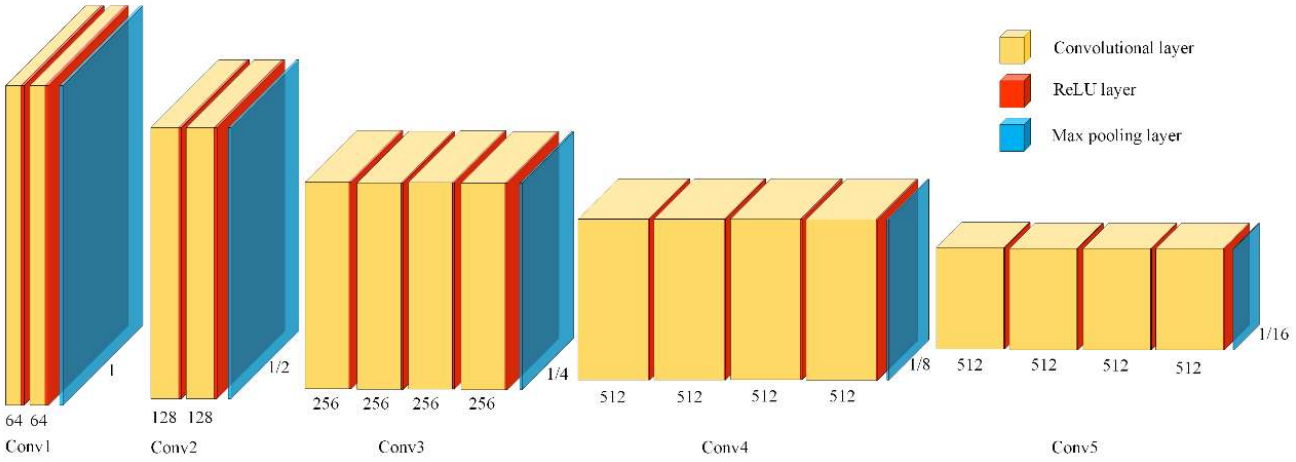

FIGURE 1. Schematic diagram of the convolution structure of the VGG19 network.

deep feature extraction using the VGG-19 network. However, for practical application, this method has two shortcomings: 1) the ''averaging'' fusion scheme for the base layer effectively reduces the image contrast, and 2) the image features extracted by VGG-19 network are relatively large, thereby missing fine details in the image.

To address the abovementioned problems of the existing scheme in Ref [35], we propose an effective fusion framework based on visual saliency weight map (VSWM) [36] and CNN for IR and VIS image fusion. Compared to the 'averaging rule', we employ a better VSWM-based strategy to obtain the base layer fusion, achieving better contrast in the fused images. For the detail layer fusion, the multi-resolution singular value decomposition (MSVD) method is used [37]. In the past, most methods have opted to apply singular value decomposition (SVD) method for image fusion tasks [38], [39], which differs from MSVD. Specifically, in Ref [38], SVD is used for base layer fusion by first decomposing and reconstructing the base layer, and then summing up all the base layers. The fusion method in Ref [39] employs a DCT dictionary learning method. The singular value of the image is used as a reference for the coefficients fused in the DCT dictionary. During fusion, image coefficients with larger singular values are retained. However, the MSVD method used in our work is quite different from the above two methods. In principle, MSVD is very similar to a wavelet transform. The idea behind MSVD is to replace the filters in a wavelet transform with SVD. Therefore, the images obtained by MSVD are multiple image groups at different scales, retaining information during decomposition and reconstruction. According to Ref [37] the fusion by MSVD shows better performance than wavelets. Therefore, the image detail layer is first subjected to MSVD decomposition, and then the image features are extracted using the CNN. By doing this, the proposed method effectively preserves fine details in the reconstructed image. Experimental results (both qualitative and quantitative) demonstrate the superior fusion performance of the proposed method compared to existing state-of-the-art schemes.

The remainder of this paper is organized as follows. Section 2 explains the feature map extraction process by the convolutional layers of the VGG19 network, pointing out that the network is insensitive to low-frequency information (such as background) and sensitive to high-frequency information. Section 3 presents the proposed fusion framework in detail. Through experiments, it is concluded that the VSWM fusion scheme enhances contrast in image fusion, and the MSVD combined with CNN preserve information at fine scales. Section 4 provides the experimental results and comparisons. Section 5 concludes the paper.

II. FEATURE MAPS OF VGG19 NETWORK

A. CONVOLUTIONAL LAYER STRUCTURE OF VGG19

The VGG architecture was initially proposed in Ref. [40]. The VGG19 network comprises of five blocks of convolutional layers followed by three fully-connected layers. The convolution layers are mainly responsible for constructing (extracting) feature maps of an image, and the fully-connected layers are used as classification function. In order to better explain the image features learned by the deep network, it is necessary to conduct an in-depth study on the output of the convolutional layer (the fully-connected layers are not within the scope of this study). Specifically, the convolution structure of the VGG19 network can be represented by Fig. 1.

The yellow colored block represents the convolution layer, the red color indicates the ReLU layer [40], and the max pooling layer is represented by the blue color. For each convolutional layer, the number of channels used are shown against the corresponding layer in Fig. 1. For example, in Fig. 1, the Conv1 block contains two convolutional layers (each with 64 channels), two ReLU Layers, and a max pooling layer. The purpose of employing max-pooling layer is to down-scale the

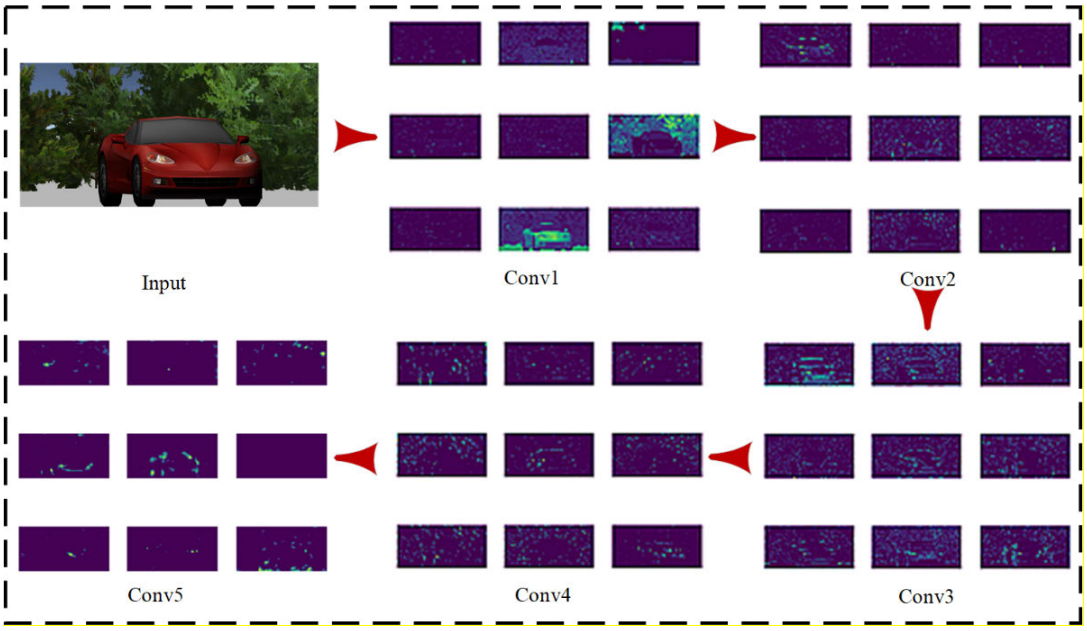

FIGURE 2. VGG19 network convolutional layer output visualization.

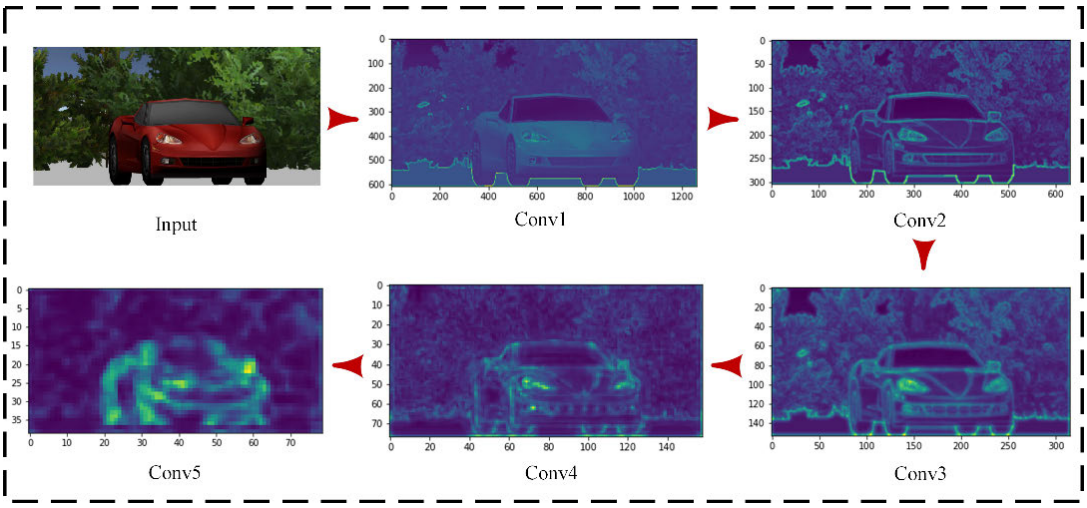

FIGURE 3. Output channel summation for each block of convolutional layers.

output from convolutional layer (to half) and analyze the data at different scales.

B. CONVOLUTION VISUALIZATION OF VGG19

Convolutional neural networks exhibit remarkable performance in target classification and recognition. However, there is no clear understanding of why they perform so well, or how they might be improved. In this regard, Zeiler and Fergus [41] first attempted to visualize convolutional network by using the deconvolution method to display different features learned by the network and explain what each layer learned from the image. Their research paved the way for better understanding neural networks through visualization.

For the visualization of VGG19 network's convolutional layers, instead of using the deconvolution method, the output for each convolution layer is directly visualized. This is done to avoid the loss of information associated with deconvolution. Fig. 2 shows the output of each block of convolutional layers in VGG19.

Fig. 2 shows the 'car' image used as an input to the network. Since each block of convolutional layers contains many channels, only 9 feature maps from each convolutional stage are displayed. These images are the first 9 output images (feature maps) of each convolution block. In addition, Fig. 3 shows the superimposed channel output for every block to clearly represent the output from convolutional blocks. In Fig. 2, it can be found that not every feature map effectively learns the features of an image. Since there are many feature

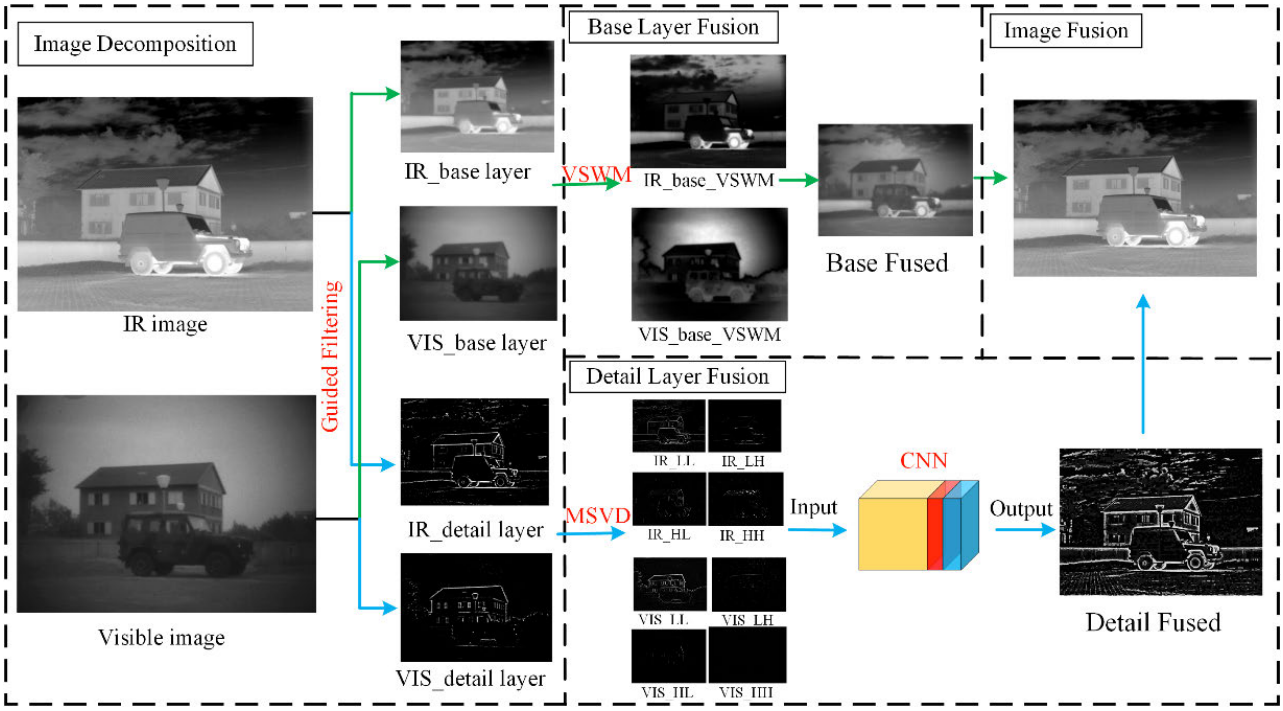

FIGURE 4. Schematics of the proposed IR and VIS images fusion framework.

maps at each convolutional block, the overall feature learning (as seen in the channel output) is not affected. For example, the image in Fig. 2 (second row, third column-Conv5) is blank with no features, but its impact on the overall channel output (shown in Fig. 3 for Conv5) is negligible.

From Fig. 3, it can be seen that the shallow network (in Fig. 1) has rich activation information in each output channel, and tends to detect the edges and contours of the image. The detected content is comprehensive, and some background information is also preserved (the conv1 and conv2 in Fig. 3 have significant edge strength, and the background can be clearly identified). As the hierarchy deepens, the number of white spaces in the output channel of the network increase because the convolution kernel does not learn the required features. The feature map encodes more abstract information, ignoring many details (the conv4 and conv5 convolution networks in Fig. 3 extract front and rearview mirror, and tire as the main features of the car, but the background, edge and other information is lost).

From the analysis of feature extraction, it can be seen that as the image propagates down the network, different types of features are extracted by different convolutional layers. However, in the image fusion process, it is necessary to retain maximum information about the source image, such as background and high frequency features. Since the deep convolutional layers are insensitive to background information, the proposed method decomposes the source image into base layer (containing background and contrast information) and detail layer (containing edge and texture feature information), and individually processes them through VSWM and MSVD method respectively, to take the advantage of both methods and improves image fusion performance.

III. FUSION FRAMEWORK BASED ON VSWM COMBINED WITH CNN

Fig. 4 presents the schematics of the proposed IR and VIS image fusion framework, which comprises of four steps, namely: image decomposition, base layer fusion, detail layer fusion, and final image fusion. Step I: source images (VIS and IR) are decomposed into base and detail layers using a guided filter [35]. Step II: the base layers are fused using the VSWM strategy. Step III: MSVD is employed to decompose detail layers further into sub-layers before CNN operations to preserve further details. The CNN can use the architecture of any one of the following networks: VGG-19, VGG-16, AlexNet, and RexNet. Step IV: the composite fused image is finally obtained by combining information fused in detail and base parts. These four steps are further explained as follows.

A. IMAGE DECOMPOSITION

Multi-scale decomposition (MSD) is a well-known method for image decomposition. Shutao *et al.* [42] proposed a rapid and effective image decomposition method with guided filtering. The base parts through this method can be obtained

FIGURE 5. VSWM results of base layers.

from (1)

$$I_b = \arg\min_{I_b} ||I - I_b||_F^2 + \lambda \nabla I_b, \tag{1}$$

where

$$\nabla I_b = sqrt(||\nabla_p^h I_b||_F^2 + ||\nabla_p^v I_b||_F^2), \tag{2}$$

where $\nabla_P = (\nabla_p^h, \nabla_p^v)$ denotes the image gradient at pixel p with ∇^h and ∇^v as the linear operators corresponding to the horizontal and vertical first-order differences, respectively. λ is the regularization parameter and is set to 5 [35].

After base parts I_b, the detailed content is obtained by (3)

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

B. FUSION OF THE BASE LAYER

In image decomposition, the base layer contains a wealth of information, such as image texture, contrast, edges, and other background information. The purpose of base layer fusion is to transfer information from the base layer of the IR and VIS images to the fused image. However, the ''averaging'' strategy applied over base layer during its fusion by the conventional methods leads to the loss of critical IR and VIS base layer information. For example, the IR images contain strong contrast information, while the VIS images have rich texture information. The application of ''averaging'' process reduces the image contrast and blurs the texture. On the contrary, the VSWM method calculates the importance of each pixel relative to the original image [43]. As a result, the contrast and texture information in the source image can be well preserved and a better base layer fusion effect can be achieved.

VSWM defines pixel-level saliency on the basis of a pixel's contrast to all other pixels. The saliency value $V^k(p)$ of pixel p is defined as follows:

$$V^k(p) = \sum_{\forall q \in I^k} |I_p^k - I_q^k|, \tag{4}$$

where k denotes the source images and $k = \{\text{IR}, \text{VIS}\}$, I_p denotes the intensity value of pixel p in image I, and q is each pixel of image I. The visual saliency of a particular pixel is computed by individually subtracting its intensity value with all the pixels in the image and then summing up those values.

For (4), the pixel by pixel expansion of $V^k(p)$ can be written as follows:

$$V^{\text{k}}(p) = |I_p^k - I_1^k| + |I_p^k - I_2^k| + \cdots + |I_p^k - I_N^k|, \tag{5}$$

where N is the number of pixels in I. The saliency values are equal if two pixels have the same intensity value, such that (5) can be rewritten as follows:

$$V^k(p) = \sum_{l=0}^{L-1} S_l |I_p^k - I_l^k|, \tag{6}$$

where l denotes pixel intensity, S_l represents the number of pixels whose intensities are equal to l, and L is the gray levels of images and $L = 256$ in this paper. Furthermore, the visual saliency weight map V^{k} will be obtained by calculating the visual saliency of other pixels in image using (6). Finally, the V^{k} is normalized to [0, 1].

In (6), we obtain a saliency map for the original image. Regions with large values of VSWM typically correspond to intensity and texture areas, whose information are useful and necessary for fusion. The base layer fusion rule is written as

$$\begin{aligned} I_b^F &= VSWM(I_b^{\text{IR}}, I_b^{\text{VIS}}) \\ &= \frac{(V^{\text{IR}} I_b^{\text{IR}} + (1 - V^{\text{IR}}) I_b^{\text{VIS}}) + (V^{\text{VIS}} I_b^{\text{VIS}} + (1 - V^{\text{VIS}}) I_b^{\text{IR}})}{2}. \end{aligned} \tag{7}$$

where V^{IR} and V^{VIS} denote the VSWM of the IR and VIS images, respectively. Fig. 5 shows the VSWM results of IR base layer and VIS base layer.

C. FUSION OF THE DETAIL LAYER

The detail layer contains the high frequency information about the source image. However, it is observed that if this image is directly fed to a CNN, the feature extraction process (of CNN more inclined towards large scale extraction) fails to register fine details in it. Therefore, the proposed method pre-processes this data through MSVD method, effectively preserving fine details in the image [37]:

$$I_{d,j}^k = MSVD(I_d^k, le), \tag{8}$$

where le denotes the number of decomposition levels, j denotes the different frequency information and $j = \{LL, LH, HL, HH\}$, and k denotes the source images and $k = \{\text{IR}, \text{VIS}\}$.

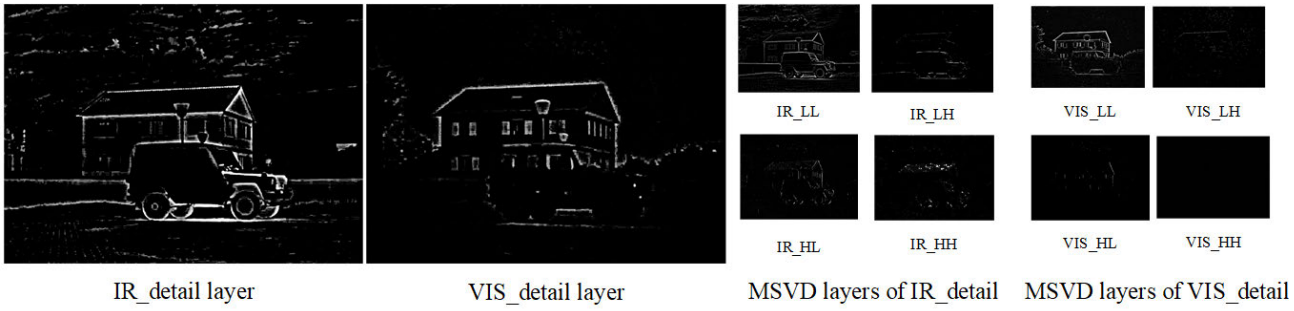

FIGURE 6. Intermediate images of MSVD processing.

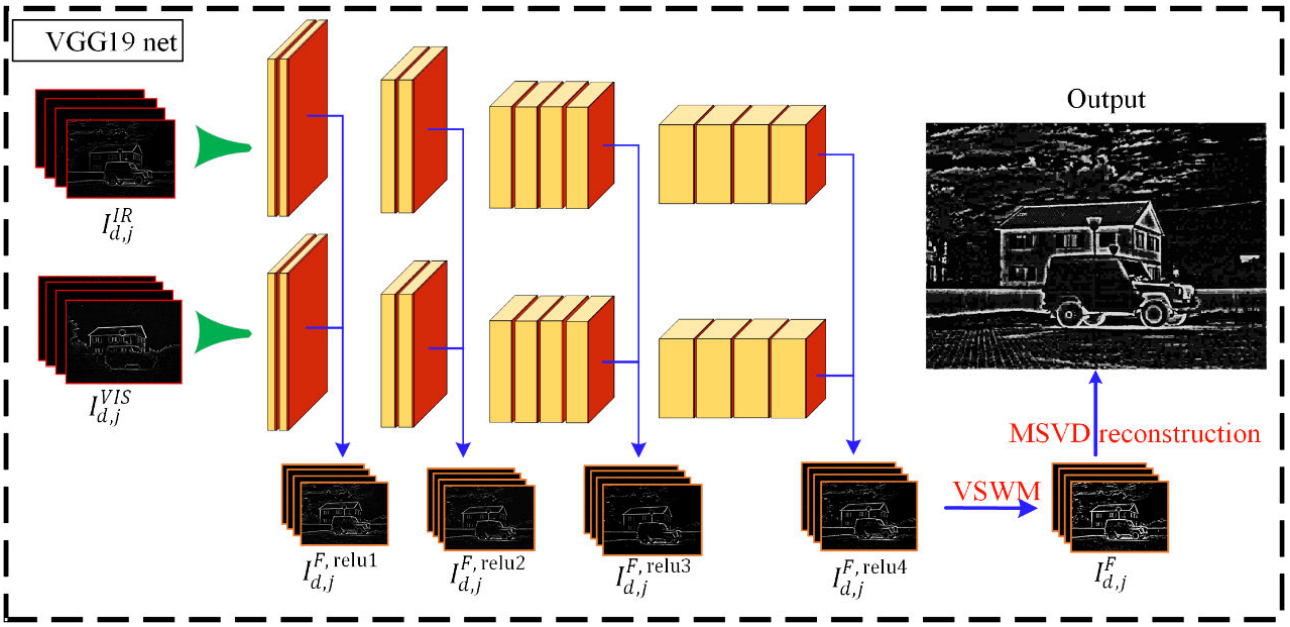

FIGURE 7. Procedure of the deep learning network in FIGURE 4.

As shown in the detail layer fusion part in Fig. 4, the detail layer obtained by image decomposition is preprocessed by the MSVD method to obtain a sub-set of images, which are similar to the results obtained by a wavelet transform. These sub-images can effectively retain fine details, which has been proven by the experiments in Ref [37]. Fig. 6 shows the intermediate images of MSVD processing.

As shown in Fig. 4, $I_{d,j}^{k}$ is viewed as an input to the CNN, and image information is extracted in depth from each of the hidden layers:

$$\phi_{d,j}^{k,net,name,m} = \Phi_{name}^{net}(I_{d,j}^{k}) \quad m = 1, 2, \cdots, M, \tag{9}$$

where *net* denotes the model of CNN, *name* refers to the hidden layer name in the model *net*, M is the number of output channels of name, and Φ denotes the operation of the hidden layer. The L1-norm is used to obtain the final detail image as:

$$I_{d,j}^{k,net,name}(x,y) = \frac{1}{w}\sum_{x,y\in w} ||\phi_{d,j}^{k,net,name,m}(x,y)||_1 \quad m = 1, 2, \cdots, M, \tag{10}$$

where w is a sliding window set to a size of 3×3 [35]. Then, we can obtain the detail fusion MSVD as:

$$I_{d,j}^{F,net} = VSWM(I_{d,j}^{\text{IR},net,name}, I_{d,j}^{\text{VIS},net,name}) \quad j \in \{LL, LH, HL, HH\}, \tag{11}$$

where $I_{d,j}^{IIR,net,name}$ and $I_{d,j}^{IIR,net,name}$ denote the output corresponding to *name* hidden layer in network *net*. The fusion of detail layer is obtained by MSVD reconstruction, given by:

$$I_{d}^{F,net} = MSVD^{-1}(I_{d,j}^{F,net}). \tag{12}$$

where the VGG-19 network is selected and *name* = {*relu1*, *relu2*, *relu3*, *relu4*}. Fig. 7 illustrates this procedure.

Specifically, there are three steps in Fig.7. The first step is to use the sub-images, obtained by the MSVD preprocessing of the image detail layer, as input to train the VGG-19 network. The second step is to extract (and visualize) images of different network depths (scales) from the convolution blocks of VGG19 network. The third step is to further process these images to reconstruct visualization images (through VSWM)

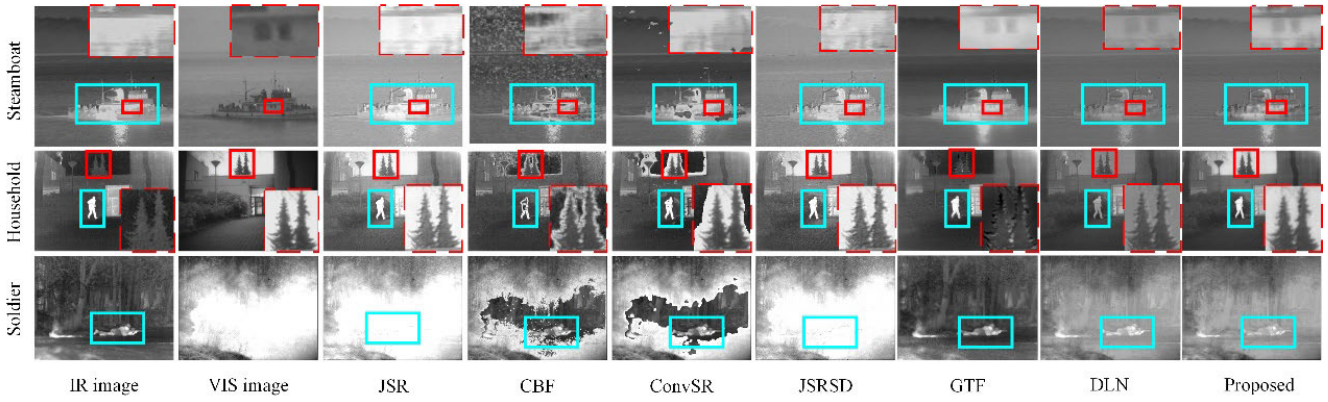

FIGURE 8. Comparison of fusion results from different methods. From top to bottom are "Steamboat," "Household," and "Soldier."

and finally obtain the final detail layer fusion result. This is consistent with the theoretical part.

Comparing the detail layers and MSVD-processed images in Fig. 6 with the fused output image (obtained by MSVD reconstruction) in Fig. 7, it can be observed that the details fused in the output image (Fig. 7) contain enhanced features from both infrared and visible detail layers. Through further qualitative comparison, it can be seen that the VGG19 network is prone to loss of fine details (across different blocks in Fig. 3), whereas the fine details in the fused image of Fig. 7 (e.g., the ground texture on the bottom left) are retained. The enhanced fused image information in Fig. 7 reflects the advantages of the MSVD method.

D. FUSION OF THE IMAGE

The final fused image is obtained from (13) using the acquired fused detail contents I_b^F and I_d^F as:

$$I^F = I_b^F + I_d^F. \tag{13}$$

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETTINGS

To verify the effectiveness of the proposed framework, twenty-one pairs of images [35] are used to compare its performance with existing state-of-the-art fusion methods of cross-bilateral filter fusion (CBF) [44], joint SR (JSR) [16], JSR with saliency detection fusion (JSRSD) [17], convolutional SR model (ConvSR) [18], gradient transfer fusion (GTF) [45], and deep learning network (DLN) [35].

According to Ref [46], the qualitative performance of a fusion algorithm differs from its quantitative performance. Therefore, to effectively quantify fusion strategies (of visual saliency, MSVD, and CNN) in our algorithm, we use three quality metrics, namely, Q_0 [47], multiscale structural similarity (MS_SSIM) [48], and sum of correlations of differences (SCD) [49]. Q_0 measures how much salient information contained by input images is transferred into the fused image without introducing distortions. SCD computes quality by considering the source images and their impact on the fused image. MS_SSIM is based on the structural similarity, which provides more flexibility than a single-scale approach in incorporating the variations of image resolution and viewing conditions. For all three metrics, a large value indicates a better fused result.

B. COMPARISON WITH OTHER FUSION METHODS

The fused images processed by six existing methods and the proposed method are shown in Fig. 8. For qualitative comparison, the details in each image (marked by red box) are shown as zoomed inset (inside the red dotted box). The cyan box in each image highlights its salient area.

It can be seen from Fig. 8. that the fused images of CBF and ConvSR methods produce serious artifacts, which are not suitable for subsequent image processing. Similarly, in the case of 'Soldier' image, the salient target information is lost in the fused results of both JSR and JSRSD methods, which is an unacceptable error. In contrast, the fused images of GTF, DLN, and the proposed method efficiently retain salient information about the original image. In addition, the contrast of fused images for the proposed and GTF methods is higher than that of the DLN method (the cyan box in the 'Steamboat' and 'Household' images). Particularly, for the 'Soldier' image, the fused image of the proposed method contains more information than that of the DLN method (which is more like the source IR image).

The details (marked by red box) in Fig. 8 for the 'Steamboat' (its windows) and 'Household' (trees in the scene) images are further studied for comparison. It can be seen for the 'steamboat' image that the windows (zoomed details in the dotted red box) have disappeared in the fused images of JSR, CBF, ConvSR and JSRSD methods. In contrast, the fused images of the proposed, GTF and DLN methods accurately retain fine details (windows). Moreover, the fused images of the Proposed and GTF methods show better reconstruction compared to the DLN method. For the 'Household' image, the fused images of the CBF, ConvSR, GTF, and DLN

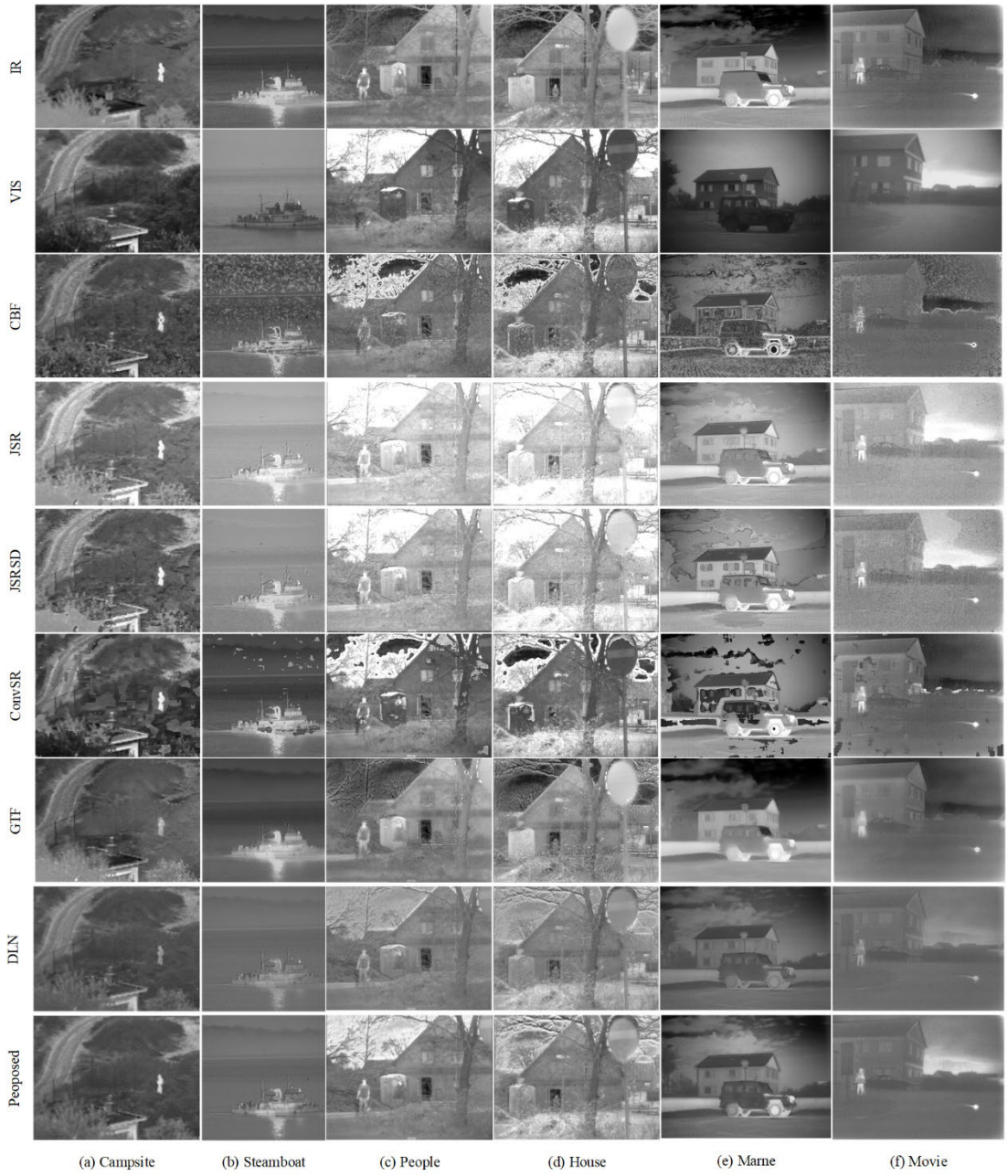

FIGURE 9. Comparison of fusion results from different methods on the (a) Campsite, (b) Steamboat, (c) People, (d) House, (e) Marne, and (f) Movie.

show obvious artifacts in the tree area. Whereas, the results for the proposed, JSRSD, and JSR methods are very clear and accurate. Therefore, it can be concluded from the qualitative comparison of Fig. 8 that the proposed method outperforms other methods by achieving better fusion performance in terms of reconstructing both low-frequency (contrast) and high-frequency (detail) information.

The purpose of image fusion is to fuse the information of different images to obtain an image with rich information. Considering the case of ''household'' image in Fig. 8, the contrast between the trees and sky in the IR image is not high. Opposite to that, the visible image can better distinguish between the trees and sky, with a high a contrast. Thus, to retain this significant information (of trees and sky) in the fused image, the proposed method reconstructs an image where the sky and trees are detected more like a visible image, along with all the significant details of IR image. This result further highlights the effectiveness of our fusion method in clearly distinguishing different targets and fusing with high efficiency.

Further qualitative comparison of the proposed framework with six other methods is provided in Fig. 9. The first two

TABLE 1. Quantitative comparison of the results in figure 9.

Images	Metric	CBF	JSR	JSRSD	ConvSR	GTF	DLF	Proposed
Campsite	Q_0	0.49731	0.49451	0.39413	0.52621	0.50205	0.61904	**0.63024**
	MS_SSIM	0.74380	0.87330	0.75190	0.69413	0.78433	0.86994	**0.89339**
	SCD	1.30442	**1.76007**	1.43823	1.06102	0.96967	1.48417	1.64180
Steamboat	Q_0	0.33991	0.37841	0.20729	0.60608	0.40572	0.63312	**0.64327**
	MS_SSIM	0.55578	0.91099	0.80706	0.82619	0.87704	0.91257	**0.92504**
	SCD	1.56290	1.93985	1.86197	1.18330	1.14071	1.90960	**1.95626**
People	Q_0	0.55712	0.39119	0.32115	0.62930	0.61427	0.64622	**0.65493**
	MS_SSIM	0.69862	0.81463	0.77459	0.74059	0.82603	0.86705	**0.90317**
	SCD	1.39012	1.79960	1.77973	1.29490	1.07672	1.76600	**1.86055**
House	Q_0	0.60931	0.35879	0.35081	0.67044	0.63607	0.65818	**0.67271**
	MS_SSIM	0.74049	0.72583	0.74387	0.77700	0.79269	0.84765	**0.88517**
	SCD	1.45324	1.61120	1.62061	1.30351	1.11126	1.69595	**1.78333**
Marne	Q_0	0.29592	0.49591	0.42452	0.53150	0.41571	**0.57750**	0.50774
	MS_SSIM	0.51051	0.88728	0.74463	0.64136	0.82672	0.87670	**0.91045**
	SCD	1.60600	1.90012	1.67706	1.38201	1.05059	1.81941	**1.91735**
Movie	Q_0	0.49906	0.32083	0.20815	0.59333	0.36068	**0.63350**	0.53411
	MS_SSIM	0.72340	0.80631	0.67891	0.78348	0.77054	0.88228	**0.90464**
	SCD	1.56524	1.91651	1.80115	1.09542	1.02887	1.87376	**1.93195**

rows in Fig. 9 present the original IR and VIS images, whereas the remaining seven rows correspond to the fusion results of seven different methods. From Fig. 9 it can be observed that the fused results of GTF, DLN, and proposed methods have few artifacts. The fused images of the proposed method tend to preserve the thermal radiation distribution in the IR images. Hence, the targets can be easily detected. Meanwhile, the details of the backgrounds in the VIS images are also retained in the fused results.

The quantitative analysis for the results in Fig. 9 is presented in table 1. The proposed method outperforms other fusion methods (except for JSR and DLN) over different metrics (of Q_0, MS_SSIM, and SCD). The JSR method achieves the highest value of the metric SCD on the ''Campsite'' source image. The DLN method achieves the best performance in terms of the metric Q_0 on the ''Movie'' and ''Marne'' source images. This result is basically consistent with the results of our subjective observations. This finding indicates that the fused images obtained by the proposed method contain more information and provide better fusion effect.

V. CONCLUSION

An effective IR and VIS fusion method based on VSWM and CNN is proposed in this study to solve the problems of low contrast and loss of fine details in fused images. We simultaneously retain the thermal radiation information of the IR image and preserve the appearance information of the VIS image. The quantitative comparisons between six state-of-the-art fusion methods and our proposed method demonstrate that the latter captures the most important information and retains approximately the largest amount of information in the source images. Furthermore, the experiments show that the proposed method is more stable and versatile than the existing state-of-the-art fusion methods.

REFERENCES

[1] G. Pajares and J. Manuel de la Cruz, ''A wavelet-based image fusion tutorial,'' *Pattern Recognit.*, vol. 37, no. 9, pp. 1855–1872, Sep. 2004.
[2] C. Chen, Y. Li, W. Liu, and J. Huang, ''Image fusion with local spectral consistency and dynamic gradient sparsity,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2760–2765.
[3] L. Dong, Q. Yang, H. Wu, H. Xiao, and M. Xu, ''High quality multi-spectral and panchromatic image fusion technologies based on curvelet transform,'' *Neurocomputing*, vol. 159, pp. 268–274, Jul. 2015.
[4] J. Ma, J. Zhao, Y. Ma, and J. Tian, ''Non-rigid visible and infrared face registration via regularized Gaussian fields criterion,'' *Pattern Recognit.*, vol. 48, no. 3, pp. 772–784, Mar. 2015.
[5] N. Wang, Y. Ma, and K. Zhan, ''Spiking cortical model for multifocus image fusion,'' *Neurocomputing*, vol. 130, pp. 44–51, Apr. 2014.
[6] G. Bhatnagar, Q. M. J. Wu, and Z. Liu, ''A new contrast based multimodal medical image fusion framework,'' *Neurocomputing*, vol. 157, pp. 143–152, Jun. 2015.
[7] H. Li, B. S. Manjunath, and S. K. Mitra, ''Multi-sensor image fusion using the wavelet transform,'' in *Proc. 1st Int. Conf. Image Process.*, Nov. 1994, pp. 51–55.
[8] F. Meng, B. Guo, M. Song, and X. Zhang, ''Image fusion with saliency map and interest points,'' *Neurocomputing*, vol. 177, pp. 1–8, Feb. 2016.

[9] Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 2, pp. 157–161, Feb. 2016.

[10] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.

[11] A. Toet, "A morphological pyramidal image decomposition," *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 255–261, May 1989.

[12] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.

[13] Q. Zhang and B.-L. Guo, "Multifocus image fusion using the nonsubsampled contourlet transform," *Signal Process.*, vol. 89, no. 7, pp. 1334–1346, Jul. 2009.

[14] G. Guorong, X. Luping, and F. Dongzhu, "Multi-focus image fusion based on non-subsampled shearlet transform," *IET Image Process.*, vol. 7, no. 6, pp. 633–639, Aug. 2013.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2970–2979.

[21] B. Liu, J. Liu, X. Bai, and H. Lu, "Regularized hierarchical feature learning with non-negative sparsity and selectivity for image classification," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Stockholm, Sweden, Aug. 2014, pp. 24–28.

[22] Z. Lei, W. Shuai, B. Xiao, J. Zhou, and E. Hancock, "Iterative deep subspace clustering," in *Proc. Joint IAPR Int. Workshops Stat. Techn. Pattern*, 2018, pp. 42–51.

[23] K. Jiang, Z. Wang, and P. Yi, "Edge-enhanced GAN for remote sensing image super-resolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5799–5812, Mar. 2019.

[24] Z. Wang, P. Yi, K. Jiang, J. Jiang, Z. Han, T. Lu, and J. Ma, "Multi-memory convolutional neural network for video super-resolution," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2530–2544, May 2019.

[25] K. Jiang, Z. Wang, P. Yi, J. Jiang, J. Xiao, and Y. Yao, "Deep distillation recursive network for remote sensing imagery super-resolution," *Remote Sens.*, vol. 10, no. 11, p. 1700, 2018.

[26] K. Jiang, Z. Wang, P. Yi, G. Wang, K. Gu, and J. Jiang, "ATMFN: Adaptive-threshold-based multi-model fusion network for compressed face hallucination," *IEEE Trans. Multimedia*, early access, Dec. 18, 2019, doi: 10.1109/TMM.2019.2960586.

[27] D. Liu, D. Zhou, R. Nie, and R. Hou, "Infrared and visible image fusion based on convolutional neural network model and saliency detection via hybrid l0-l1 layer decomposition," *Proc. SPIE J. Electron. Imag.*, vol. 27, no. 6, Dec. 2018, Art. no. 063036.

[28] A. Azarang, H. E. Manoochehri, and N. Kehtarnavaz, "Convolutional autoencoder-based multispectral image fusion," *IEEE Access*, vol. 7, pp. 35673–35683, 2019.

[29] 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.

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

[31] H. Xu, P. Liang, W. Yu, J. Jiang, and J. Ma, "Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3954–3960.

[32] Z. Shao, J. Cai, P. Fu, L. Hu, and T. Liu, "Deep learning-based fusion of Landsat-8 and Sentinel-2 images for a harmonized surface reflectance product," *Remote Sens. Environ.*, vol. 235, Dec. 2019, Art. no. 111425.

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

[34] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

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

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

[37] V. P. S. Naidu, "Image fusion technique using multi-resolution singular value decomposition," *Defence Sci. J.*, vol. 61, no. 5, pp. 479–484, 2011.

[38] S. Yin and Y. Zhang, "Singular value decomposition-based anisotropic diffusion for fusion of infrared and visible images," *Int. J. Image Data Fusion*, vol. 10, no. 2, pp. 146–163, 2019.

[39] M. Amin-Naji, P. Ranjbar-Noiey, and A. Aghagolzadeh, "Multi-focus image fusion using singular value decomposition in DCT domain," in *Proc. 10th Iranian Conf. Mach. Vis. Image Process. (MVIP)*, Nov. 2017, pp. 45–51.

[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, Apr. 2015. [Online]. Available: https://arxiv.org/abs/1409.1556

[41] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, Zürich, Switzerland, 2014, pp. 818–833.

[42] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.

[43] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th Annu. ACM Int. Conf. Multimedia*, 2006, pp. 815–824.

[44] 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.

[45] 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.

[46] 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.

[47] G. Piella and H. Heijmans, "A new quality metric for image fusion," in *Proc. Int. Conf. Image Process.*, Sep. 2003, p. 173.

[48] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[49] 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.

LEI YAN received the B.Sc. degree from the Hebei University of Technology, in 2016. He is currently pursuing the Ph.D. degree with the Beijing Institute of Technology. His main research interests include multimodality image fusion and virtual reality.

JIE CAO received the Ph.D. degree in optical instruments from the Beijing Institute of Technology, China, in 2015. He completed his Postdoctoral studies from the National University of Singapore. He is currently an Associate Research Fellow with the Beijing Institute of Technology. His research interests include 3D imaging and bionic inspired vision.

SAAD RIZVI received the master's degree in electrical and computer engineering from the University of Manitoba, Canada. He is currently pursuing the Ph.D. degree with the Beijing Institute of Technology. His research interests include quantum optics, ghost imaging, and deep learning. He was a recipient of various funding and awards from the Government of Pakistan, Canada, and China.

KAIYU ZHANG, photograph and biography not available at the time of publication.

QUN HAO, photograph and biography not available at the time of publication.

XUEMIN CHENG, photograph and biography not available at the time of publication.