IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 8, 2021, accepted July 28, 2021, date of publication July 30, 2021, date of current version August 10, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3101639

An Infrared and Visible Image Fusion Method Guided by Saliency and Gradient Information

QINGQING LI[1,2], GUANGLIANG HAN[1], PEIXUN LIU[1], HANG YANG[1], JIAJIA WU[1,2], AND DONGXU LIU[1,2]

[1]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China
[2]School of Optoelectronics, University of Chinese Academy of Sciences, Beijing 100049, China

Corresponding authors: Guangliang Han (hangl@ciomp.ac.cn) and Peixun Liu (liupx@ciomp.ac.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61602432 and Grant 61401425.

ABSTRACT Infrared and visible image fusion is a hot topic due to the perfect complementarity of their information. There are two key problems in infrared and visible image fusion. One is how to extract significant target areas and rich texture details from the source images, and the other is how to integrate them to produce satisfactory fused images. To tackle these problems, we propose a novel fusion framework in this paper. A multi-level image decomposition method is used to obtain the base layer and detail layer of the source image. For the fusion of base layer, an ingenious fusion strategy guided by the saliency map of source image is designed to improve the intensity of salient targets and the visual quality of the fused image. For the fusion of detail layer, an efficient approach by introducing the enhanced gradient information is presented to boost the detail features and sharpen the edges of the fused image. Experimental results demonstrate that, compared with fifteen classical and advanced fusion methods, the proposed image fusion framework has better performance in both subjective and objective evaluation.

INDEX TERMS Image fusion, base layer, detail layer, saliency map, gradient information.

I. INTRODUCTION

Multi-sensors image fusion is an enhancement technology that integrates the image information obtained by different kinds of sensors into one image, and it plays a vital role in computer vision tasks, such as target recognition, remote sensing, and surveillance [1]–[3].

Infrared images can distinguish targets from the background according to the thermal radiation information and work well in day/ night and all-weather conditions, but they have low resolution and weak details. Visible images contain detailed texture information and high spatial resolution, but they are easily affected by the weather and brightness [4]. Therefore, the infrared and visible image fusion becomes a very hot topic due to the perfect complementarity of their information. There are two key problems in infrared and visible image fusion. One is how to extract significant target areas and rich texture details from the source images, and the other is how to combine them to produce excellent fused images. This paper aims to explore an effective method to achieve satisfactory fusion of infrared and visible images.

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

Multifarious infrared and visible image fusion methods have been proposed in recent years, and they can be grouped into three dominant types, including multi-scale transform, sparse and low-rank representation learning-based, and deep learning-based methods [5]–[7].

Multi-scale transforms have developed for decades in the field of infrared and visible image fusion. Discrete wavelet transform is a typical multi-scale transform method [8], [9], which decomposes the input image into high and low frequency sub-images. Then, these sub-images are fused to a single image through appropriate fusion rules. The dual-tree complex wavelet transform (DTCWT) is proposed to over-come the shift variance and lack of directionality problems of the discrete wavelet transform [10]. For capturing the abundant directional information of source images, contourlet transform is proposed [11]. Nonsubsampled contourlet transform (NSCT) is a modified form based on contourlet transform, which is widely used in infrared and visible image fusion due to its flexibility and shift-invariance [12], [13]. However, the above-mentioned fusion methods need to transform images to frequency domain, which increases the computational complexity [14].

In order to avoid image transformation, representation learning-based methods have attracted the attention of

researchers. The most common methods are on the basis of sparse representation (SR) [15] and dictionary learning [16], [17], which are consistent with the physiological mechanism of the human visual system. Nevertheless, image fusion methods based on SR suffer from high sensitivity to misregistration. To tackle this drawback, Liu et al introduce the convolutional sparse representation (CSR) model to fuse multi-modal images. Even so, fused images obtained by fusion methods based on SR have insufficient texture details [18].

With the development of deep learning, many innovative image fusion methods based on deep learning are designed [19]–[21]. The convolutional neural network (CNN) attracts much focus due to its ability of powerful feature representation. Liu et al utilize CNNs to finish the fusion of infrared and visible images [22]. Whereas, CNN model usually requires the ground truth of training images. It is not considered to build the ground truth in infrared and visible image fusion because it is unrealistic to define a standard for fused images. To solve this issue, Ma et al. construct an end-to-end model named generative adversarial network for infrared and visible image fusion (FusionGAN) [4]. On this basis, they ameliorate the loss function of FusionGAN to increase the detail information and sharpen the edge of fused images [23]. There are mainly two drawbacks of deep learning-based fusion methods. One is that the network model is difficult to train when the number of images is limited, especially for infrared and visible image fusion, the other is that a good network often depends on GPUs with good performance.

Recently, the latent low-rank representation (LatLRR) model has been gradually attracted attention in image fusion. Fusion methods based on LatLRR can decompose source images into base and detail parts without transformation, which is beneficial to design fusion rules so that generate high quality fused images. In addition, these methods can fuse infrared and visible image without complex training process and good performance GPUs. Thus, fusion methods based on LatLRR are widely used in image fusion [14], [24], [25].

Inspired by this, we propose a novel framework for infrared and visible image fusion. The main challenge of infrared and visible image fusion is to generate a single image containing salient target areas and texture details. This paper proposed a novel infrared and visible image fusion framework, which improving the fusion image quality by introducing the saliency and gradient information to the fusion strategy. To facilitate the design of the fusion strategy, we decompose the source image into the detail and base layers. To improve the intensity of salient targets and the visual quality of the fused image, the saliency information of the source image is introduced to fuse the base layer. To boost texture details of the fused image, the gradient information is used to assist the fusion of detail layer. Experimental results prove that the proposed image fusion framework outperforms than other traditional and deep learning fusion methods.

The main contributions of this paper are summarized as follows.

1. A novel method based on multi-level image decomposition is proposed for infrared and visible image fusion.
2. For base layer fusion, an excellent strategy guided by the saliency map is designed to increase the salient information and improve the visual quality of the fused image.
3. For detail layer fusion, an efficient method with enhanced gradient information is presented to increase the detail information of the fused image.
4. Compared with the classical and state-of-the-art fusion methods, our proposed fusion framework has a better performance in terms of both subjective and objective evaluation.

This paper is organized as follows. In Section II, the proposed infrared and visible image fusion framework is introduced in detail. In Section III, information about the dataset and parameter settings is given. In Section IV, the experimental results and analyses are provided. In Section V, conclusions and the future work are presented.

II. THE PROPOSED INFRARED AND VISIBLE IMAGE FUSION METHOD

The proposed fusion framework is schematically presented in Figure 1, which is mainly composed of four parts: (1) image decomposition, (2) the fusion of base layer, (3) the fusion of detail layer, and (4) image reconstruction.

Image decomposition: a multi-level image decomposition method based on LatLRR is employed to obtain the base layer and detail layer of source images, which is beneficial to design fusion strategies. Infrared and visible images are decomposed to base layer images b_A^l and b_B^l, detail layer matrixes $M_A^{1:l}$ and $M_B^{1:l}$ after l levels decomposition. The base layer includes the primary intensity and brightness information of the input image. The detail layer contains texture details and feature information of source images.

The fusion of base layer: an ingenious fusion strategy guided by the saliency map of the source image is designed. As shown in Figure 1, salient regions of infrared and visible images are different. Because of the different imaging ways, the infrared image mainly highlights the thermal radiation information of objects, whereas the visible image focuses on the spectral information reflected by different objects. Inspired by this, the weight matrix used to fusion the base layer is calculated depends on the saliency map, which aims to retain suitable intensity information and improve the visual quality of the fused image, simultaneously.

The fusion of detail layer: an effective approach is presented to achieve the fusion of detail layer. First, the detail matrixes of infrared and visible images are fused based on weighted-average rule. Then the enhanced gradient information is added to the detail layer to improve the contrast and sharpness of the fusion result.

Image reconstruction: the final step of the proposed fusion framework is reconstructing fusion results of base layer and detail layer to obtain the fused image.

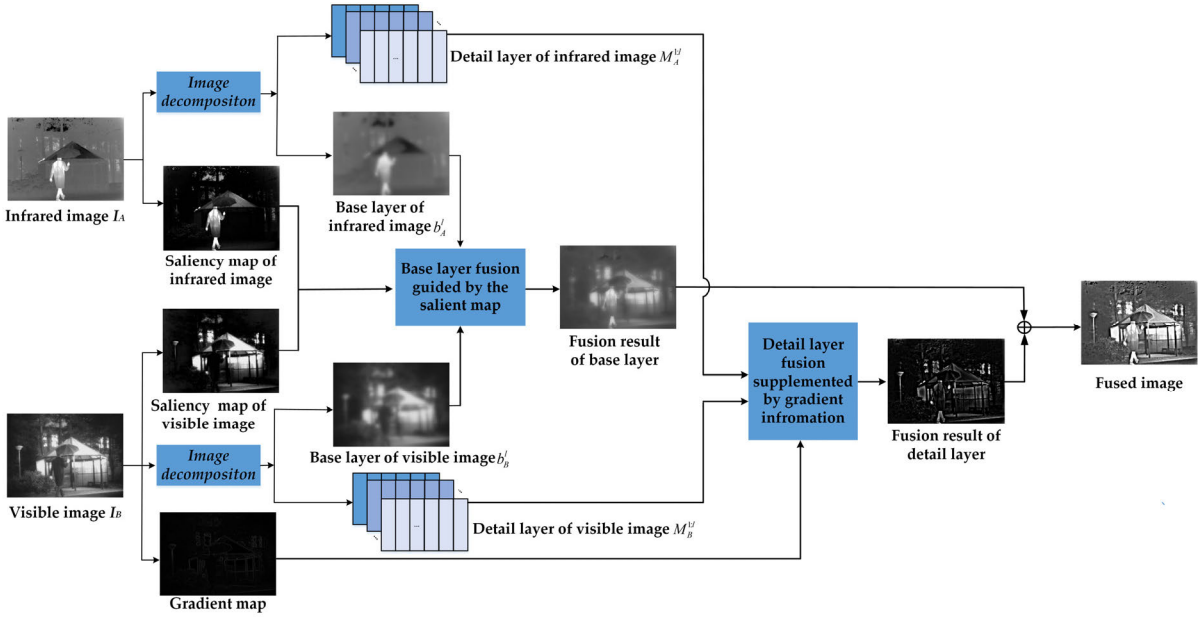

FIGURE 1. The framework of the proposed image fusion method.

The four parts of our proposed method will be described in detail in the following sections.

A. IMAGE DECOMPOSITION

A successful fusion of infrared and visible images should integrate as much as target prominent and detail information as possible to the fusion image. Therefore, it is a significant process to adequately extract the intensity and texture content of source images. Recently, a multi-level image decomposition method based on latent rank representation (LatLRR) is presented, which can divide input images into base layer and detail layer [14], [24]. The base layer mainly contains intensity and brightness information of source images, and the detail layer principally includes texture and structure information of source images. This decomposition way is beneficial for designing the fusion strategies. Inspired by this, we utilize this approach to decompose the source images. The flowchart of image decomposition is shown in Figure 2.

The LatLRR model is described as Equation (1).

$$\min_{Z,L,E} \|Z\|_* + \|L\|_* + \lambda \|E\|_1 \quad s.t.,\ X = XZ + LX + E \tag{1}$$

where λ is a positive balance factor, λ is an empirical value, we set the value to 0.4 in this paper according to reference [26]. $\|.\|_*$ represents the nuclear norm, and $\|.\|_1$ is the l_1-norm.

X denotes the observed data, which is the input image in this paper. Z and L are the low-rank and saliency coefficients, respectively. E expresses the sparse noise. The low-rank part XZ, saliency part LX and sparse noise part E can be obtained from Equation (1). The noise part is removed in image fusion operation.

The multi-level image decomposition method is constructed according to the LatLRR model in reference [14], which can obtain the base layer and detail layer of the input image. The detail layer and base layer images are expressed as follows:

$$M^i = P \times W\left(b^{i-1}\right) \tag{2}$$

$$d^i = R(M^i) \tag{3}$$

$$b^i = b^{i-1} - d^i, \quad i = 1, 2, 3, \ldots, l \tag{4}$$

where l is the highest decomposition level, M^i and b^i represent the detail matrix and the base image at level i. Note that b^0 is the source image. After l levels LatLRR decomposition, the source image derives l detail images $d^{1:l}$ and a base image b^l. P is the projection matrix learned by LatLRR. The size of the projection matrix P is 16×16, and the decomposition level is set to $\{1, 2, 3, 4\}$. $W(\cdot)$ denotes a two-stage operator including the sliding window technology and reshuffling, and $R(\cdot)$ means the function reconstructing the detail image d^i according to the detail matrix M^i.

The process of $W(\cdot)$ operator is shown in Figure 3. The window size is 16×16, the stride of sliding window is 1. Before sliding window, the source image is resized to $(\tilde{m}, \tilde{n})$ to guarantee the width and height of the resized image are both integer multiples of 16. $\tilde{m}$ and $\tilde{n}$ are calculated as Equations (5)-(8).

$$\tilde{m} = \begin{cases} m, & \Re_m = 0 \\ m + 16 - \Re_m, & others \end{cases} \tag{5}$$

$$\tilde{n} = \begin{cases} n, & \Re_n = 0 \\ n + 16 - \Re_n, & others \end{cases} \tag{6}$$

$$\Re_m = \mathrm{mod}(m/16) \tag{7}$$

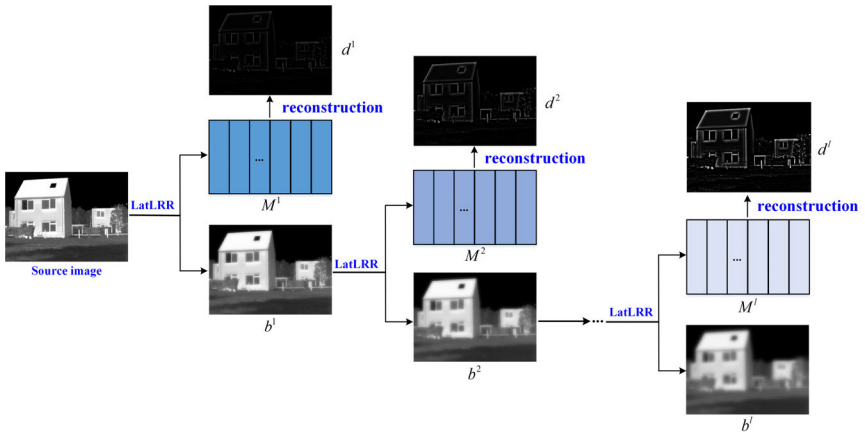

FIGURE 2. The flowchart of image decomposition.

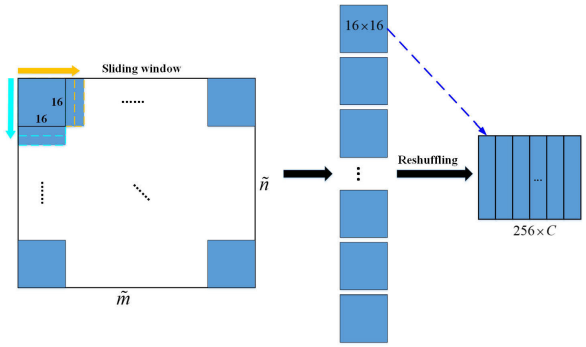

FIGURE 3. The process of the operator $W(\cdot)$.

$$\Re_n = \mathrm{mod}(n/16) \tag{8}$$

where, (m, n) is the size of the source image, mod is the function to calculate the remainder.

Then, a window with size of 16×16 is used to slide over the source image, which aims to get a series of image patches with size 16×16.

After sliding window, the image patch is reconstructed into the matrix $\hbar_k$ with size $S_p \times 1$. Next, these matrixes $\hbar_k$ are reshuffled to one matrix H with size $S_p \times C$.

$$H = [\hbar_1, \hbar_2, \hbar_3, \ldots, \hbar_C] \tag{9}$$

where, $S_p = 16 \times 16 = 256$. C is calculated as follows:

$$C = \prod_{z=1}^{2} (U_z - sqrt(S_p) + S_d) \tag{10}$$

where, $sqrt(\cdot)$ is the operation to get the square root. S_d is the stride value of sliding window, $S_d = 1$, $U_1 = m$, $U_2 = n$.

The process of $R(\cdot)$ is similar to the inverse process of $W(\cdot)$.

B. FUSION OF BASE LAYER

The base layer can be obtained through image decomposition. As exhibited in Figure 4, with the increase of decomposition level, the image of base layer becomes more and more smooth, and the contrast between pixels gets lower and lower. As a result, the salient object regions cannot be retained integrally, the salient object edges are blurred, and the intensity information is decreased at a great degree. Traditional average-rule only averages the pixel values of infrared and visible base images, which cannot satisfy the requirement of sufficiently integrating intensity information from the base layer to the fused image. However, as exhibited in Figure 4, the salient regions of infrared and visible images are highlighted in their saliency maps. Besides, the saliency map possesses strong contrast and clear object edge. Therefore, an innovative fusion strategy guided by the saliency map of source image is designed to improve the intensity of salient targets and the visual quality of the fused image. This base layer fusion method consists of two steps: extraction of saliency map and design of fusion strategy. The two parts are described in detail next.

1) EXTRACTION OF SALIENCY MAP

The human visual system always pays more attention to the salient area than background in an image to reduce the difficulty in some tasks, such as object detection, tracking and recognition [27]. Pixels of salient structure, region and object stand out from the surrounding neighbor pixels. Saliency detection is on the purpose of extracting visually salient regions of images. For image fusion, a suitable saliency detection method should simultaneously meet two requirements: clearly extract the salient region and preserve the edge and background information as far as possible. These two conditions guarantee the rich intensity information and integrated structure of the fused images. Inspired by [28], [29], a pixel-level saliency detection method named visual saliency map (VSM) is employed to obtain the saliency map of images in this step. The saliency map of an image is described as follows:

$$S(x, y) = \sum_{q=1}^{m} \sum_{t=1}^{n} \|I(x, y) - I(q, t)\| \tag{11}$$

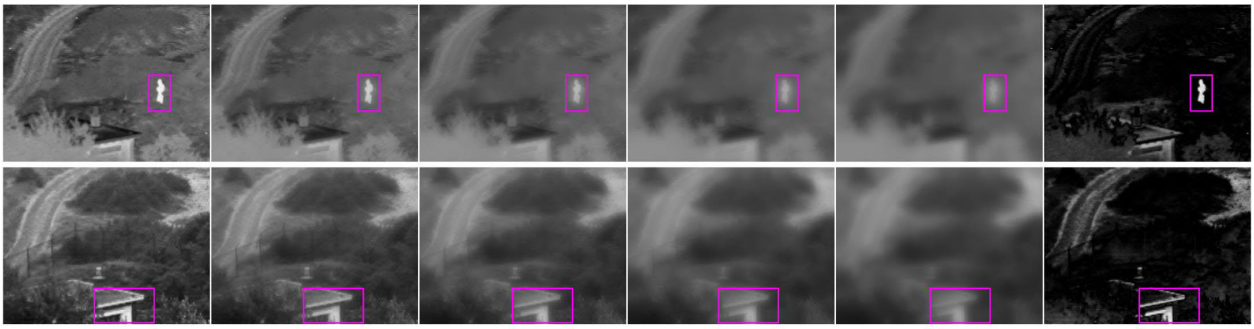

FIGURE 4. Base layer images and their saliency maps. The carmine box denotes the salient regions. (a) source images; (b) base layer at level 1; (c) base layer at level 2; (d) base layer at level 3; (e) base layer at level 4; (f) saliency map. From up to down: infrared image, visible image.

where, the size of the input image is $m \times n$. $\|\cdot\|$ represents the distance between the intensity values of two pixels. $I(x, y)$ and $I(q, t)$ are intensity values at the pixels (x, y) and (q, t), respectively.

In order to accelerate the computation and make saliency maps of infrared and visible images in the same order of magnitudes, $S(x, y)$ is normalized to [0,1], which is rewritten by:

$$\Phi(x, y) = \frac{S(x, y) - \min(S)}{\max(S) - \min(S)} \tag{12}$$

For proving the advantages and reasonability of the above-mentioned method which can extract the saliency map, we compare it with four classical saliency detection algorithms: GS [30], SF [31], GBMR [32], RBD [33]. As shown in Figure 5, saliency maps of GS, SF, GBMR and RBD methods all can detect the salient regions. However, these methods have a common deficiency that they overemphasize salient areas so that filter out plenty of background information, which will result in missing a large amount of information in fused images. Compared with the above four methods, VSM can not only perfectly extract the salient object (such as people of infrared image) but also reserve the major content in infrared and visible images. In summary, VSM is more suitable for image fusion.

2) DESIGN OF FUSION STRATEGY

To maintain sufficient intensity information of salient targets in the source image and improve the visual quality of the fused image, the saliency map is considered as an important reference to calculate the weight matrix and guide the base layer fusion. The weight matrixes of infrared and visible images fusion are provided by:

$$w_{bA}(x, y) = \begin{cases} 0.5, & if\ \Phi_A(x, y) = \Phi_B(x, y) = 0 \\ \dfrac{\Phi_A(x, y)}{\Phi_A(x, y) + \Phi_B(x, y)}, & others \end{cases} \tag{13}$$

$$w_{bB}(x, y) = 1 - w_{bA}(x, y) \tag{14}$$

where, $\Phi_A(x, y)$ and $\Phi_B(x, y)$ are the saliency values of infrared and visible images at the pixel (x, y), $w_{bA}(x, y)$ and $w_{bB}(x, y)$ are the weight values of infrared and visible images at the pixel (x, y). Note that, at the pixel (x, y), if $\Phi_A(x, y)$ is 0 and $\Phi_B(x, y)$ is also 0, the formula $\frac{\Phi_A(x,y)}{\Phi_A(x,y)+\Phi_B(x,y)}$ will be unusable. In this condition, the saliency values of infrared and visible images are equal at pixel (x, y). Therefore, we set the weight value $w_{bA}(x, y)$ to 0.5.

Therefore, the fused image of base layer F_{base} is generated as follows:

$$F_{base} = w_{bA} b_A^l + w_{bB} b_B^l \tag{15}$$

The details of the proposed base layer fusion strategy are described as **Algorithm** 1.

Algorithm 1 The Proposed Base Layer Fusion Strategy in This Paper

Input: a pair of aligned infrared and visible images;
Output: the fused image of base layer F_{base};

1. Input images are decomposed by Equations (1)-(4) to obtain base layer images b_A^l and b_B^l;
2. Normalized saliency maps Φ_A and Φ_B are obtained by Equations (11) and (12);
3. Weighted matrixes w_{bA} and w_{bB} are computed by Equations (13) and (14);
4. Fused image of base layer F_{base} is acquired by Equation (15).

C. FUSION OF DETAIL LAYER

As shown in Figure 2, the detail layers of input images are acquired by decomposition. Generally, the thermal radiation information produced by targets is contained in infrared images, which can be emphasized in the fused parts of base layer. Nevertheless, texture detail information of objects is included in visible images, which is beneficial to improve target tracking and recognition due to its high spatial resolution. The gradient map of an image contains contour and edge information, which has strong sharpness and contrast.

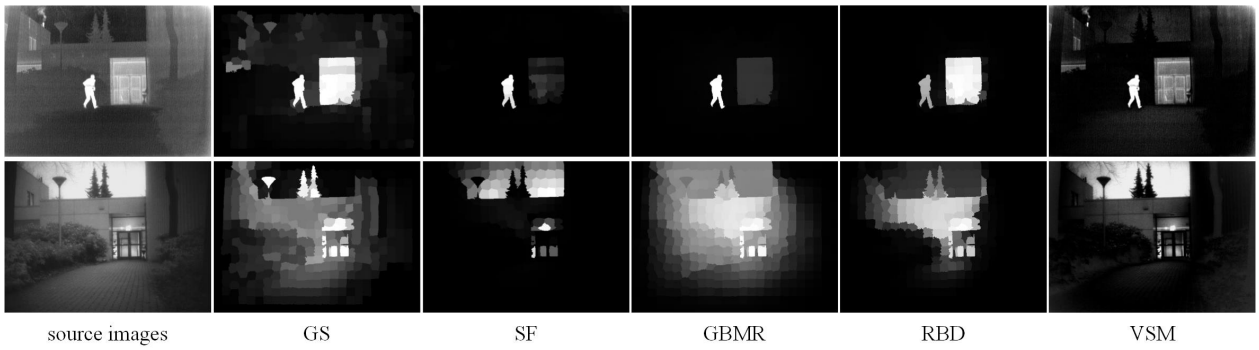

FIGURE 5. The comparison of different saliency detection methods. From up to down: infrared image, visible image.

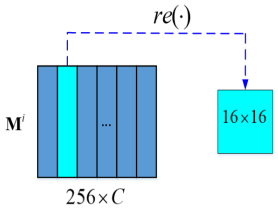

FIGURE 6. The process of *re*(·) function.

Some researchers have utilized the gradient information by different ways to enhance the image quality [34], [35]. To get fused images with rich texture details and sharpened edges, an efficient approach by introducing the enhanced gradient information is presented to finish the fusion of detail layer, which is specifically introduced as follows.

After l levels decomposition, detail matrixes $M^{1:l}$ of an image with size $m \times n$ is obtained. The size of a detail matrix M^i is $S_p \times C$. $S_p = 256$, C is calculated by Equation (10).

The weight for each pair of corresponding image patches can be written as follows:

$$w_{dA}^{i,k} = \frac{\left\| re(M_A^{i,k}) \right\|_*}{\left\| re(M_A^{i,k}) \right\|_* + \left\| re(M_B^{i,k}) \right\|_*} \quad i = 1, 2, 3, \cdots, \quad l, k = 1, 2, 3, \cdots, C \tag{16}$$

where, A and B are infrared and visible images, respectively, $\|\cdot\|_*$ denotes the nuclear norm to calculate the sum of singular values of the matrix. $re(\cdot)$ indicates the function that reorganizes the matrix $M^{i,k}$ with size 256×1 into the image patch with size 16×16. The process of $re(\cdot)$ is shown in Figure 6.

Based on Equation (16), the fused detail matrix is provided as below:

$$M_{dFused}^{i,k} = w_{dA}^{i,k} \times M_A^{i,k} + w_{dB}^{i,k} \times M_B^{i,k} \tag{17}$$

$$M_{dFused}^{i} = [M_{dFused}^{i,1}, M_{dFused}^{i,2}, \cdots, M_{dFused}^{i,k}, \cdots, M_{dFused}^{i,C}] \tag{18}$$

The fused image of detail layer at level i is expressed by Equation (19):

$$d_{Fused}^{i} = R(M_{dFused}^{i}) \tag{19}$$

Here, $R(\cdot)$ is the same function as Equation (3), which reconstructs the detail image d^i according to the detail matrix M^i.

For improving the quality of the fused image, enhanced gradient information of the visible image is added to the fusion of detail layer image as the supplementary content. Gamma transformation function is used to increase the contrast of gradient map, which is denoted as follows:

$$G = (g + \varepsilon)^{\gamma} \tag{20}$$

where, g and G are the input matrix and the matrix after Gamma transformation, respectively. ε is a very small constant term called compensation factor, which makes sure $(g + \varepsilon)$ is non-zero, γ is the Gamma coefficient. In this paper, the input of Equation (20) is the gradient map of visible image, which can be solved by:

$$\nabla_{I_B}(x, y) = \sqrt{[I_B(x+1, y) - I_B(x, y)]^2 + [I_B(x, y+1) - I_B(x, y)]^2} \tag{21}$$

where I_B is the visible image. According to Equations (20) and (21), the gradient map $\tilde{\nabla}_B$ after Gamma transformation can be expressed as follows:

$$\tilde{\nabla} I_B = (\nabla I_B + \varepsilon)^{\gamma} \tag{22}$$

Based on Equations (19) and (22), the fused image of detail layer F_{detail} can be derived as follows:

$$F_{detail} = \sum_{i=1}^{l} d_{Fused}^{i} + \tilde{\nabla} I_B \tag{23}$$

The details of the proposed detail layer fusion strategy are given as **Algorithm** 2.

D. IMAGE RECONSTRUCTION

Image reconstruction contains mainly two parts in this paper. One part is to reconstruct the fused image of detail layer and the other part is to reconstruct the final fused image depend on the fused images of base layer and detail layer.

FIGURE 7. Five pairs of source images. From up to down: infrared images, visible image.

Algorithm 2 The Proposed Detail Layer Fusion Strategy in This Paper

Input: a pair of aligned infrared and visible images;
Output: the fused image of detail layer F_{detail};

1. Input images are decomposed by Equations (1)-(4) to obtain detail layer matrixes $M_A^{1:l}$ and $M_B^{1:l}$;
2. The fused matrix M_{dFused}^{i} of detail layer at level i is calculated by Equations (16)-(18);
3. The fused image d_{Fused}^{i} of detail layer at level i is reconstructed by Equation (19);
4. The gradient map ∇_{I_B} of visible image is acquired by Equation (21);
5. The gradient map ∇_{I_B} of visible image is enhanced based on Gamma transformation so that the enhanced gradient map $\tilde{\nabla} I_B$ is acquired by Equation (22);
6. The fused image of detail layer is obtained by Equation (23);

The fused image F of infrared and visible image is summarized as follows:

$$F(x, y) = F_{base}(x, y) + F_{detail}(x, y) \tag{24}$$

III. EXPERIMENTAL DATASET AND SETTINGS

A. EXPERIMENTAL DATASET

In this paper, we test our method on TNO Image Fusion Dataset [36] and KAIST Dataset [37]. TNO dataset contains many registered infrared and visible images under different scenes, which can freely be used for research purpose. Therefore, the TNO dataset is widely used for infrared and visible image fusion research. A sample of these image pairs is shown in Figure 7. KAIST is a multispectral pedestrian dataset, which contains abundant registered infrared-visible image pairs.

B. COMPARISON METHODS

Fifteen classical and state-of-the-art image fusion methods are chosen to evaluate the fusion performance of our proposed fusion framework, including: curvelet transform fusion method (CVT) [38], dual-tree complex wavelet transform fusion method (DTCWT) [10], multi-resolution singular value decomposition fusion method (MSVD) [39], cross bilateral filter fusion method (CBF) [40], guided filter fusion method (GFF) [41], gradient transfer and total variation minimization fusion method (GTF) [42], hybrid multi-scale decomposition with Gaussian and bilateral filters fusion method (HMSD-GF) [43], infrared feature extraction and visual information preservation fusion method (IFEVIP) [44], convolutional neural networks fusion method (FCNN) [22], gradient filter fusion method (GF) [45], visual saliency map and weighted least square optimization-based fusion method (WLS) [28], latent low-rank representation fusion method (LatLRR) [24], GAN based fusion method (FusionGAN) [4], dense block based fusion method (DenseFuse) [46], and Nest Connection and Spatial/Channel Attention fusion method (NestFuse) [47]. All above comparison methods are conducted based on their publicly available codes, and their parameters are set according to their papers.

Experiments of these deep learning methods (including FusionGAN, DenseFuse and NestFuse) are finished on a GPU (NVIDIA GeForce GTX 1070). Experiments of our method and other comparison methods are implemented in MATLAB 2018a on a computer (Intel Core i7, 2.20-GHz CPU).

C. EVALUATION METRICS

For objectively analyzing the fusion results of the proposed method, seven quality metrics are utilized.

1) ENTROPY(EN)

EN calculates the information richness of the fused image. The higher the EN is, the more information is contained in the fused image, and the better quality of the fused image is [48]. The definition of EN is provided by:

$$EN = -\sum_{z=0}^{Z-1} p_z \log_2 p_z \tag{25}$$

where Z is the number of gray values, p_z is the normalized histogram of the corresponding gray level in the fused image.

2) MUTUAL INFORMATION(MI)

MI computes the amount of information that is integrated from the source images to the fused image. The larger MI means the higher quality of the fused image [49]. MI is defined as below:

$$MI = MI_{A,F} + MI_{B,F} \tag{26}$$

$$MI_{S,F} = \sum_{s,f} p_{S,F}(s,h) \log \frac{p_{S,F}(s,h)}{p_S(s)p_F(h)} \tag{27}$$

where $p_S(s)$ and $p_f(f)$ are the marginal histograms of source image S and fused image F, respectively. $p_{S,F}(s,f)$ are the joint histogram of the source image and the fused image.

3) AVERAGE GRADIENT(AG)

AG measures the degree of sharpness and clarity in the fused image. Large AG means the fused image has much details information and clear edges [50]. AG is calculated as below:

$$AG = \sqrt{\frac{\sum_{x=1}^{m}\sum_{y=1}^{n}(F(x,y)-F(x+1,y))^2+F(x,y)-F(x,y+1)^2}{mn}} \tag{28}$$

In which, $F(x,y)$ is the pixel value of the fused image with the size $m \times n$.

4) SPATIAL FREQUENCY(SF)

SF calculates the distribution of the intensity information and structure features in the fused image. Larger SF means the fused image has more texture details and more sharpening edges. It contains two parts: spatial Row Frequency (RF) and spatial Column Frequency (CF) [51]. SF is expressed as follows:

$$SF = \sqrt{RF^2 + CF^2} \tag{29}$$

$$RF = \sqrt{\frac{\sum_{x=1}^{m}\sum_{y=1}^{n}(F(x,y) - F(x,y-1))^2}{mn}} \tag{30}$$

$$CF = \sqrt{\frac{\sum_{x=1}^{m}\sum_{y=1}^{n}(F(x,y) - F(x-1,y))^2}{mn}} \tag{31}$$

where $F(x,y)$ is the pixel value of the fused image.

5) STANDARD DEVIATION(SD)

SD indicates the spread of the information in the image. Larger SD means that the fused image has higher contrast, wide distribution of the gray value, and richer information [52]. SD is defined as follows:

$$SD = \sqrt{\frac{\sum_{x=1}^{m}\sum_{y=1}^{n}(F(x,y) - F_{mean})^2}{mn}} \tag{32}$$

Here, $F(x,y)$ is the pixel value of the fused image with the size $m \times n$, F_{mean} is the mean pixel value of the fused image.

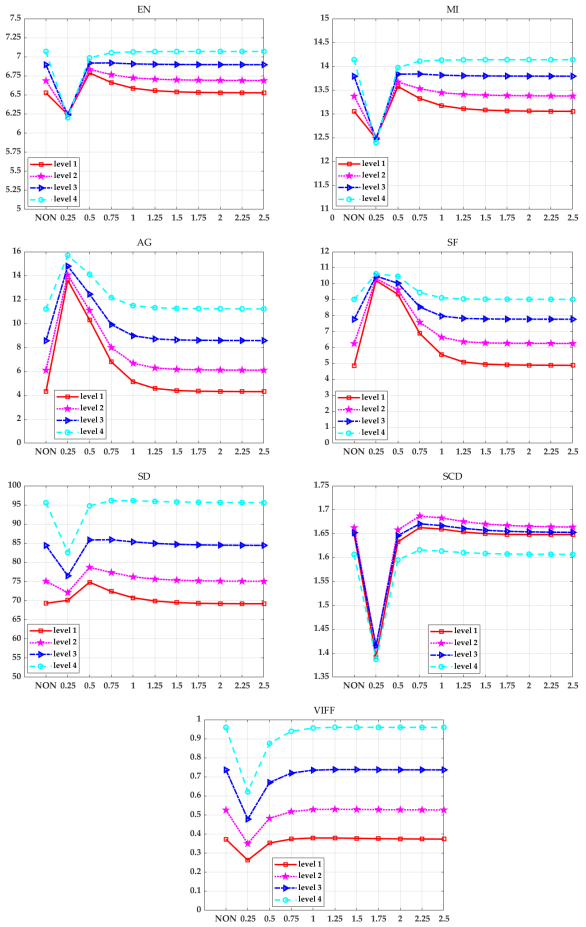

FIGURE 8. The average values of evaluation metrics for all fused images obtained by the proposed fusion framework with different γ and different level I. NON on the X-axis means the proposed fusion framework without γ enhanced visible gradient information.

6) SUM OF THE CORRELATIONS OF DISSERENCES (SCD)

SCD means the amount of complementary information contained in the fused image. The larger SCD is, the higher quality of the fused image is [53]. SCD is defined as follows:

$$SCD = \Upsilon(D_{I_A,I_f}, I_A) + \Upsilon(D_{I_B,I_f}, I_B) \tag{33}$$

where D_{I_A,I_f} and D_{I_B,I_f} are the difference between the source image and the fused image. $\Upsilon(D_I, I)$ is the correlation between D_I and I. It is represented as follows:

$$\Upsilon(D_I, I) = \frac{\sum_{x=1}^{m}\sum_{y=1}^{n}(D_I(x,y) - D)(I(x,y) - a)}{\sqrt{(\sum_{i=1}^{m}\sum_{j=1}^{n}(D_I(x,y) - D)^2)(\sum_{i=1}^{m}\sum_{j=1}^{n}((I(x,y) - a))^2)}} \tag{34}$$

where D and a are the average value of the pixel values of D_I and I, respectively.

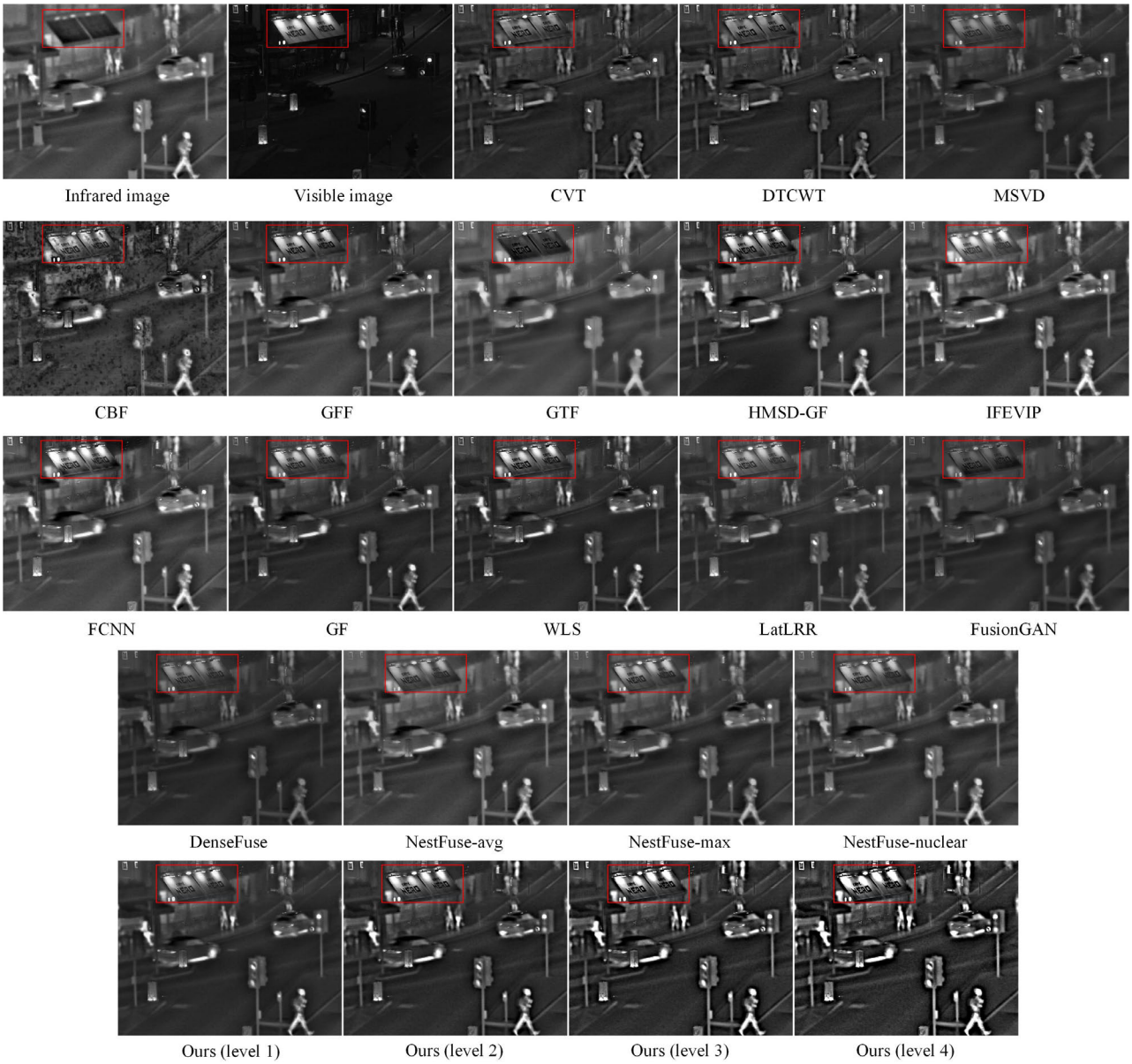

FIGURE 9. Experiments on "street" images of TNO dataset.

7) VISUAL INFORMATION FIDELITY FUSION(VIFF)

VIFF expresses the visual information fidelity of the fused image, the higher the VIFF is, the better visual quality of the fused image is [54]. VIFF is a metric with complexity calculation, which can be simplified as follows:

$$VIFF = \frac{VID}{VIND} \tag{35}$$

where VID is the visual information with distortion information, whereas, VIND is the visual information without distortion information.

All above-mentioned evaluation indicators are positively correlated with fusion quality, that is, the greater the metric value is, the better the fusion quality is.

D. PARAMETER SETTINGS

In this paper, the image decomposition level l is set to $\{1, 2, 3, 4\}$. In Section II-C, for the fusion of detail layer, an efficient approach by introducing the enhanced gradient information is presented to increase the texture detail information and sharpen the edges of the fused image. A vital parameter of this approach is γ, which is a factor of Gamma transform and determines the enhancement degree to the gradient map. Different γ value will result in different fusion performance. In this part, we set the γ from 0.25 to 2.5. The interval is 0.25. It is necessary to select one appropriate γ value for our proposed fusion framework based on the test images. Thus, we use seven metrics to evaluate the performance of the proposed fusion method with different γ.

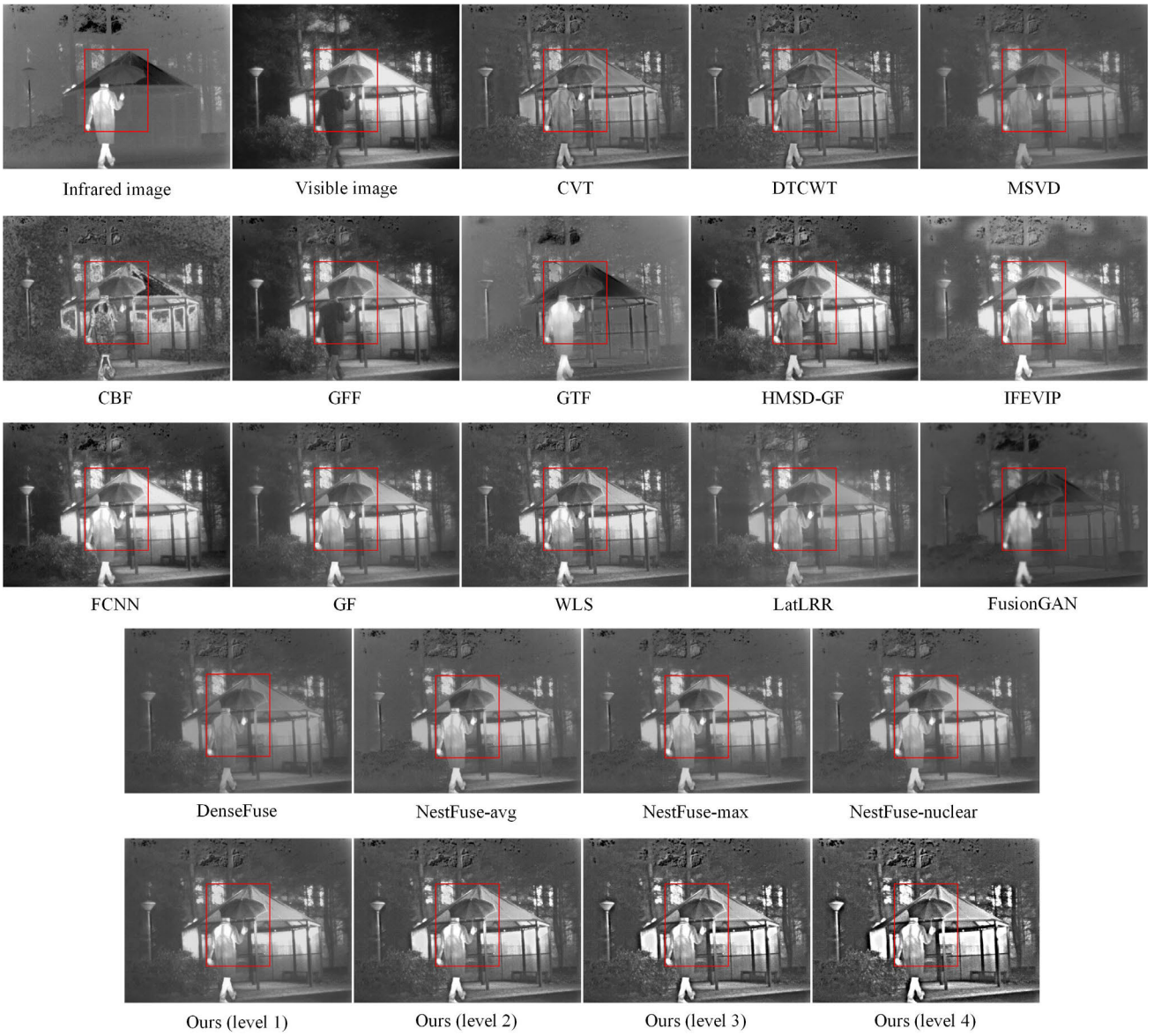

FIGURE 10. Experiments on "umbrella" images of TNO dataset.

The average values of seven metrics for all fused images obtained by the proposed fusion framework with different γ and different level l are shown in Figure 8.

In Figure 8, NON on the X-axis means the proposed fusion framework without γ enhanced visible gradient information, which is considered as the reference to estimate the fusion quality of the proposed fusion framework with different γ. Compared with NON, when $\gamma = 0.25$, the average values of metrics (except AG and SF) are observably decreased, which illustrates that the fusion performance of the proposed framework has not been improved. When γ belongs to [0.5, 1.5], the maximum average values of metrics (including EN, MI) are still smaller than NON condition. When $\gamma > 1.5$, the average values of metrics at levels 1 to 4 are all larger than NON, which proves that the fusion quality of the proposed framework has been significantly increased.

In summary, if γ is set to an appropriate value, the fusion quality of the proposed fusion framework will be enhanced.

For intuitively and concretely evaluating the fusion performance of the proposed fusion method with different γ, we propose a max-comparison method to score γ, which is described as follows:

$$Score(\gamma) = \sum_{nm=1}^{NM} \mathrm{sgn}(\max(\mathrm{E}_\gamma(nm)) - \max(\mathrm{E}(nm))) \quad (36)$$

$$E_\gamma(nm) = (E_\gamma(nm, 1), E_\gamma(nm, 2), E_\gamma(nm, 3), E_\gamma(nm, 4)) \quad (37)$$

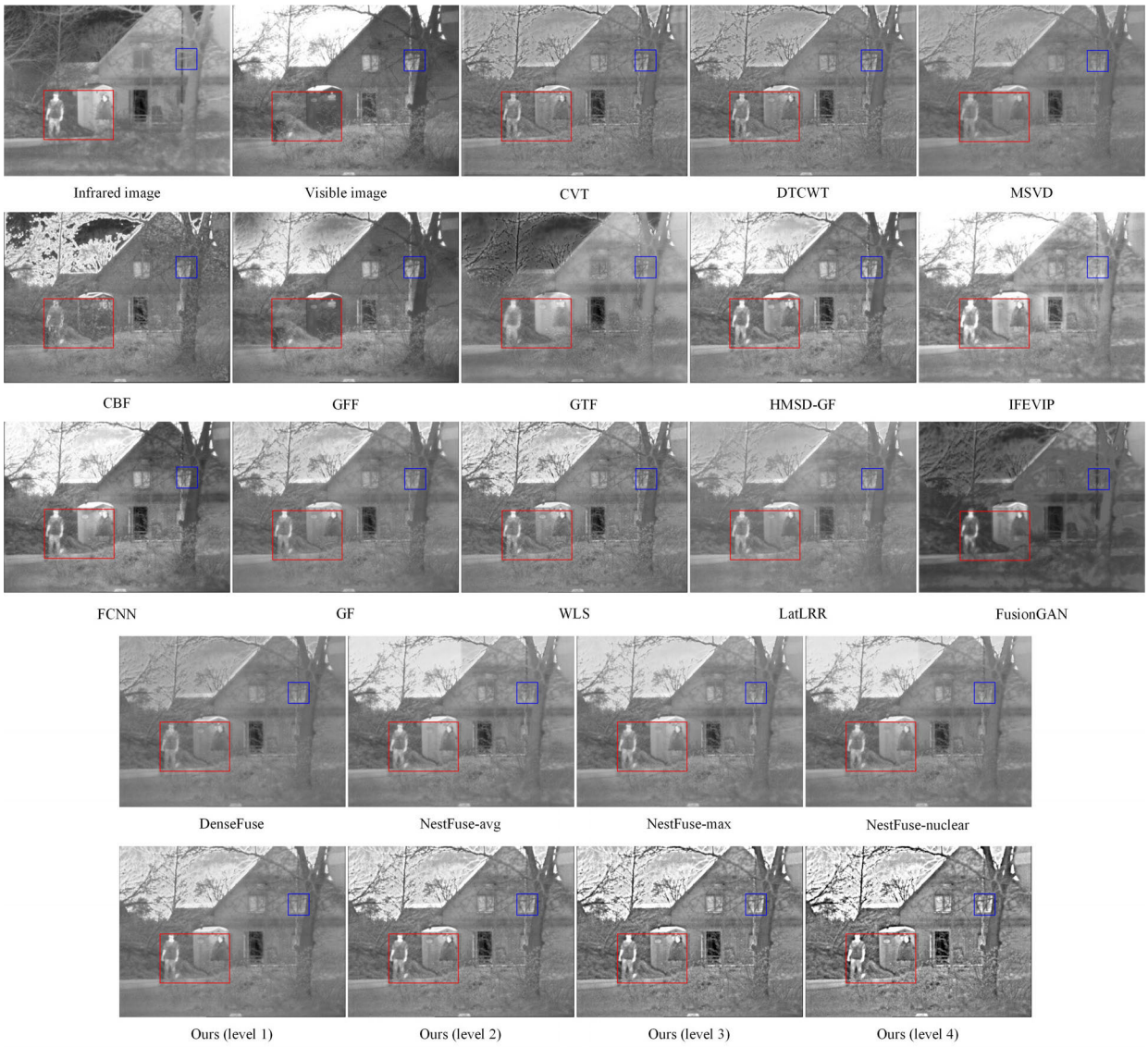

FIGURE 11. Experiments on "men" images of TNO dataset.

TABLE 1. The max-comparison score results of different γ. The best score is denoted in red.

γ value	0.25	0.50	0.75	1.00	1.25
Score	-3	-3	1	1	3
γ value	1.50	1.75	2.00	2.25	2.50
Score	3	7	7	7	7

TABLE 2. The rank scores of different γ. The best score is denoted in red.

γ value	1.75	2.00	2.25	2.50
Rank Score	22	19	18	11

where NM is the number of evaluation metrics, in this paper, NM is equal to 7. The nm values (from 1 to 7) correspond to the metrics EN, MI, AG, SF, SD, SCD, and VIFF, respectively. $E_\gamma(nm)$ contains the nm-th evaluation metric values calculated according to the fusion result obtained from the proposed fusion method with γ enhanced visible gradient information, $E_\gamma(nm, i)$ represents the evaluation metric value at level i. $E(nm)$ is the nm-th evaluation metric values calculated according to the fusion result obtained from the proposed fusion method without γ enhanced visible gradient information. sgn$(\cdot)$ is the signum function, which is denoted as follows:

$$\mathrm{sgn}(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0 \\ 0, & x = 0 \end{cases} \tag{38}$$

The higher the $Score(\gamma)$ is, the better the fusion quality is. The max-comparison score results of different γ is shown in Table 1. When $\gamma = 1.75, 2, 2.25, 2.5$, the scores are all

TABLE 3. The average evaluation metric values obtained by different fusion methods on TNO dataset. The best three values in each metric are denoted in red, green, and blue, respectively.

Methods	EN	MI	AG	SF	SD	SCD	VIFF
CVT	6.857488	13.714975	5.417645	5.806767	76.824097	1.586156	0.315645
DTCWT	6.387783	12.775566	5.334281	5.734421	54.324967	1.590848	0.304397
MSVD	6.187836	12.375671	3.660290	4.445638	48.162419	1.583146	0.241612
CBF	6.857488	13.714975	6.784169	6.862700	76.824097	1.294660	0.265627
GFF	6.854095	13.708189	5.197526	5.621794	82.197150	1.261495	0.250367
GTF	6.635343	13.270686	4.543870	5.085128	67.626026	0.965453	0.188175
HMSD-GF	7.006695	14.013389	6.268162	6.491452	87.904729	1.645908	0.494975
IFEVIP	6.591954	13.183908	4.706830	5.410741	79.224212	1.631409	0.312994
FCNN	7.067761	14.135523	5.561077	5.932891	96.762089	1.572781	0.430778
GF	6.588266	13.176532	4.071189	4.611556	69.984434	1.660118	0.354707
WLS	6.637861	13.275722	6.338700	6.534948	71.484084	1.652197	0.443601
LatLRR	6.328579	12.657159	3.521720	4.252030	53.665687	1.702454	0.289102
FusionGAN	6.362867	12.725734	2.874152	3.612263	54.358016	1.013374	0.186125
DenseFuse	6.174034	12.348068	3.053937	3.711629	47.820403	1.592125	0.255122
NestFuse-a	6.919710	13.839419	4.728811	5.383777	82.752426	1.563750	0.344662
NestFuse-m	6.894208	13.788416	4.650401	5.307727	80.363710	1.561454	0.352859
NestFuse-n	6.904613	13.809227	4.736089	5.386202	82.925723	1.560105	0.353456
Ours (level1)	6.533856	13.067712	4.349085	4.911336	69.309516	1.648503	0.376077
Ours (level2)	6.693662	13.387324	6.133988	6.267771	75.181072	1.667130	0.528114
Ours (level3)	6.898194	13.796389	8.605596	7.780944	84.570625	1.654878	0.737768
Ours (level4)	7.070026	14.140051	11.246767	9.022772	95.688861	1.607489	0.960138

7 and higher than when γ is equal to other values. That is, the proposed fusion framework with $\gamma = 1.75, 2, 2.25, 2.5$ can achieve good performance.

To further select the best γ from these four values, we propose a novel rank-score method based on [55]. The modified rank-score method is expressed as follows:

$$RScore(k) = \sum_{nm=1}^{NM} Rank(nm, k) \tag{39}$$

$$Rank(nm, k) = K - k + 1 \tag{40}$$

where K is the number of γ value, and $K = 4$. K is the ranking of the fusion performance with γ.The rank scores of these Gamma $\gamma = 1.75, 2, 2.25, 2.5$ are presented in Table 2. When $\gamma = 1.75$, the rank score is the maximum among these four conditions, which means the fusion quality is the best. As a result, we set γ to 1.75 in our proposed fusion framework.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this part, fusion results of the proposed method are evaluated subjectively and objectively. We choose fifteen comparison methods and seven evaluation metrics introduced in Section III to demonstrate the fusion performance of our method. To test the fusion performance of our proposed method, we conduct our method and the fifteen comparison methods on TNO and KAIST datasets. In the field of image fusion, researchers generally select about 20 pairs of images to test the algorithm [46], [47]. Therefore, we select 20 infrared and visible image pairs from TNO dataset and the other 20 infrared and visible image pairs from KAIST dataset for testing. The subjective and objective analyses are provided as follows.

A. EXPERIMENTS ON TNO DATASET

1) SUBJECTIVE EVALUATION

Three examples of the fused results on TNO dataset are given in Figures 9-11. As shown in the red and blue boxes of Figures 9-11, the fused images of CBF method have much noise and unclear detail information. Compared with our proposed method, GTF, FusionGAN, and DenseFuse can only generate very little saliency features in the fused images. Although CVT, DTCWT, MSVD, and IFEVIP, LatLRR and DenseFuse can integrate some salient and texture information to the fused images, the edges are blurred and incomplete. By contrast, GFF, HMSD-GF, FCNN, GF, WLS, NestFuse and the proposed fusion framework can achieve better fusion than others. Furthermore, our method can simultaneously integrate enough luminance and detail structure information to the fused image. The visual quality of fused images obtained by our method is obviously better than others.

Especially, as shown in Figure 9, the words in the red boxes obtained by our methods are quite clear, have the

TABLE 4. The average evaluation metric values obtained by different fusion methods on KAIST dataset. The best three values in each metric are denoted in red, green, and blue, respectively.

Methods	EN	MI	AG	SF	SD	SCD	VIFF
CVT	6.153845	12.307690	3.291974	3.952871	54.422564	1.134236	0.621967
DTCWT	6.130921	12.261843	3.271274	3.930998	54.221100	1.135757	0.631665
MSVD	5.989234	11.978468	2.148555	2.904062	48.420646	1.235791	0.386309
CBF	6.623837	13.247734	3.979940	4.482026	77.467522	1.146893	0.808007
GFF	6.646380	13.293347	3.360617	4.027473	80.216184	0.987257	0.887131
GTF	5.692654	11.385316	2.652302	3.342401	35.670138	1.127157	0.313285
HMSD-GF	6.686621	13.384263	3.650204	4.319250	90.958343	1.349769	0.997625
IFEVIP	6.480973	12.967608	2.928441	3.669750	80.523631	1.301946	0.787849
FCNN	6.655659	13.321279	3.375459	4.045421	85.745091	1.064554	0.932067
GF	6.238149	12.476305	2.994091	3.621630	71.941797	1.256552	0.769432
WLS	6.266102	12.532716	3.620939	4.202929	77.564170	1.258300	0.890243
LatLRR	6.199751	12.399533	2.077855	2.864040	58.466047	1.381000	0.583438
FusionGAN	5.636667	11.273350	1.689583	2.259777	39.681618	1.193961	0.236257
DenseFuse	5.975567	11.951133	1.794913	2.381396	48.265991	1.171954	0.390878
NestFuse-a	6.537058	13.074139	2.772799	3.524377	78.594094	1.390104	0.741245
NestFuse-m	6.460591	12.921204	2.830861	3.571926	78.002794	1.419032	0.786303
NestFuse-n	6.476185	12.952395	2.850895	3.592109	79.390626	1.363164	0.786508
Ours (level1)	6.182518	12.365037	2.720172	3.485232	76.481881	1.182797	0.823638
Ours (level2)	6.318133	12.644489	4.109615	4.609438	84.175014	1.231801	1.095499
Ours (level3)	6.463926	12.939375	5.855435	5.996404	94.001882	1.219064	1.382719
Ours (level4)	6.558945	13.126298	7.602689	7.258955	103.958743	1.149147	1.650828

whole contour and sharpened edges. In addition, as exhibited in Figure 11, the men in red boxes and the window in blue boxes are retained well in the fused images of our method. With the increase of the image decomposition level (from 1 to 4), the saliency of the target regions is enhanced, and the contrast of the fused image is improved gradually.

In summary, compared with traditional and deep learning methods, our method can deliver fused images with stronger intensity of the salient targets, more detail information, more sharpened edges, and higher visual quality.

2) OBJECTIVE EVALUATION

The average evaluation metrics values obtained by the proposed method and comparison methods on TNO dataset are shown in Table 3. In general, the average values of our proposed method (decomposition level from 1 to 4) on these metrics are acceptable. Specifically, the average values (including AG, SF, VIFF) of our method (level 4) are much larger than those of other methods. These values show that the proposed method can produce the fused image with adequate details information, clear edges, and high visual quality.

The average values of our method on these metrics are all in the top three, which indicates that can achieve better image fusion than other fifteen methods. The average values (including EN, MI, AG, SF, VIFF) of our method (level 4) are the maximum among these comparison methods. These values

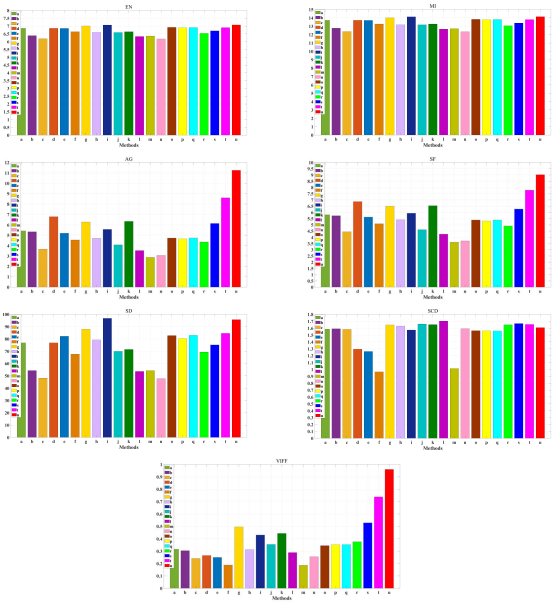

FIGURE 12. The bar charts of different fusion methods about seven evaluation metrics on TNO dataset. a∼u represent CVT, DTCWT, MSVD, CBF, GFF, GTF, HMSD-GF, IFEVIP, FCNN, GF, WLS, LatLRR, FusionGAN, DenseFuse, NestFuse-avg, NestFuse-max, NestFuse-nuclear, ours (level 1), ours (level 2), ours (level 3), ours (level 4), respectively.

demonstrate that our method can preserve sufficient information, enhance the features (such as edges), and improve

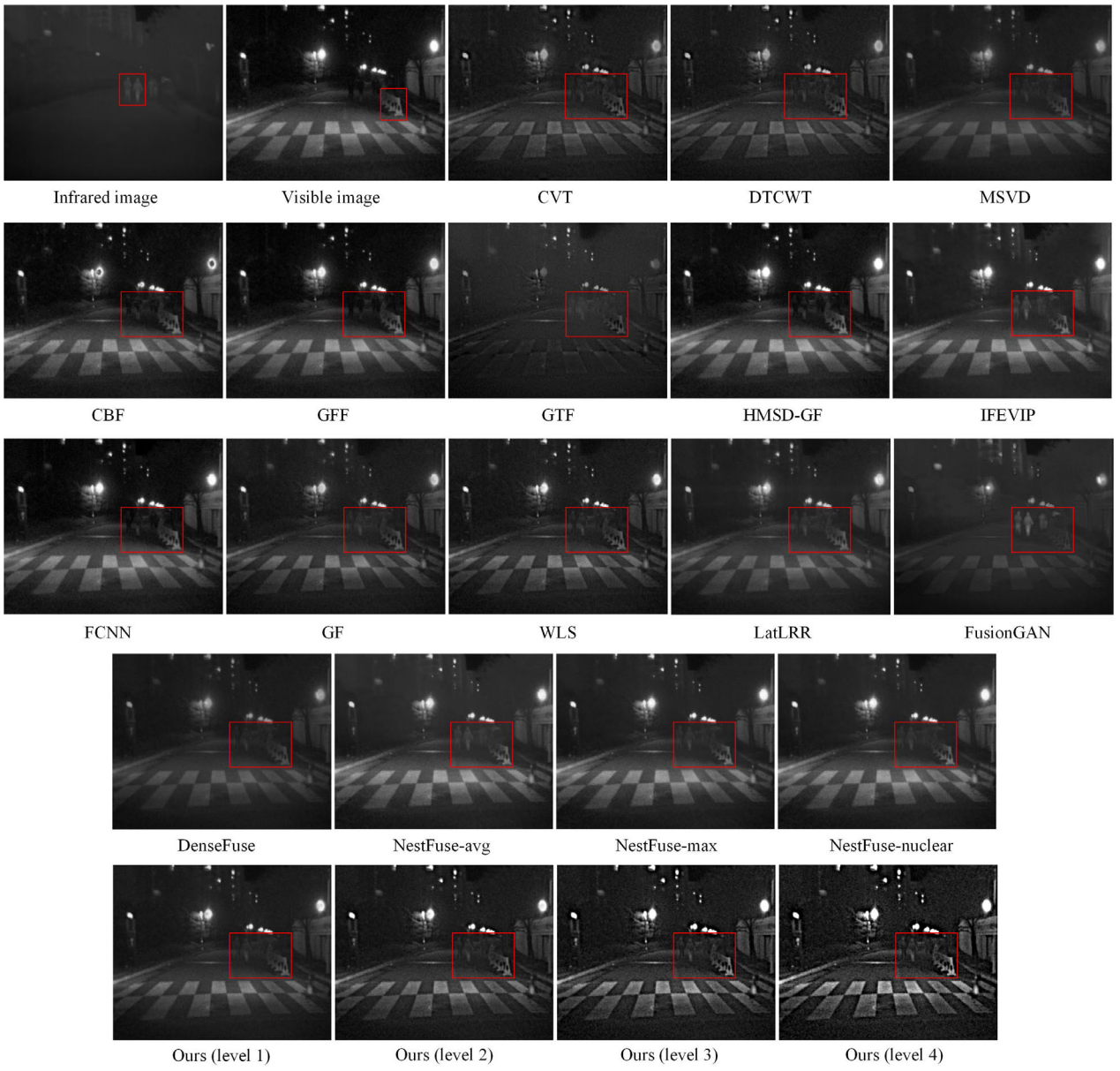

FIGURE 13. Experiments on KAIST dataset.

the visual quality of fused images. The average values on the metric SCD of our method are all bigger than most methods, which illustrates that the information of the fused images obtained by our method has credible complementarity.

For assessing the proposed fusion framework intuitively, we give the bar charts of different fusion methods about seven evaluation metrics on TNO dataset in Figure 12. The different color bars represent the average evaluation metrics values of different fusion methods. The average values of our method on metrics EN, MI, SD, and SCD are very high, these values illustrate that the proposed method can retain enough information. It can be seen that the average values on metrics AG, SF, and VIFF are significantly larger than those of other methods, which proves that the proposed fusion method can produce fused images with strong spatial structure and fine visual quality.

B. EXPERIMENTS ON KAIST DATASET

1) SUBJECTIVE EVALUATION

The example of the fused results on KAIST dataset is given in Figure 13. Pedestrians and fence are contained in the red boxes. As shown in Figure 13, the fused images of MSVD and GTF methods have poorer visual quality than our method. Specifically, the pedestrians and zebra crossings in MSVD and GTF fusion images are not obvious, and the edge structures of the fence and zebra crossings are not clear. Pedestrians in these fusion images (including CVT, CBF, GFF, HMSD-GF) is not salient. The texture information of

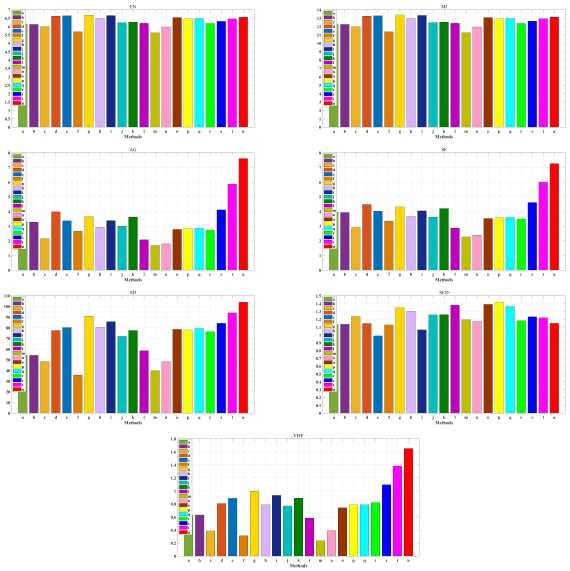

FIGURE 14. The bar charts of different fusion methods about seven evaluation metrics on KAIST dataset. a∼u represent CVT, DTCWT, MSVD, CBF, GFF, GTF, HMSD-GF, IFEVIP, FCNN, GF, WLS, LatLRR, FusionGAN, DenseFuse, NestFuse-avg, NestFuse-max, NestFuse-nuclear, ours (level 1), ours (level 2), ours (level 3), ours (level 4), respectively.

the fence in the fused image produced by FusionGAN is little. As exhibited in Figure 13, our method, IEFVIP, FCNN and LatLRR can generate the fused images with high quality. Especially, compared with other methods, the pedestrians and zebra crossings of the fused images obtained by our method are more salient, and the edges of the fence are more sharpened. In conclusion, the proposed fusion method can generate fused images with salient targets, clear detail textures, sharpened edges, and high visual quality.

2) OBJECTIVE EVALUATION

Table 4 gives the average evaluation metric values obtained by different fusion methods on KAIST dataset. As shown in Table 4, the average values (including AG, SF, SD and VIFF) of our method (level 4) are the best among these methods, which illustrates that the proposed infrared and visible image fusion method can produce fused images with clear edges and high visual quality. Although the average values (including EN, MI and SCD) of our method are not in the top three, they are all acceptable, which indicates that our proposed method can maintain sufficient source image information.

In order to compare the proposed fusion method with other methods more intuitively, Figure 14 provides the bar charts of different fusion methods about seven evaluation metrics on KAIST dataset. The different color bars represent the average evaluation metrics values of different fusion methods. As shown in Figure 14, the average values (including AG, SF, SD and VIFF) of our method are obviously larger than other methods. The average values (EN, MI and SCD) of our method are not the maximum, but they are still greater than many other methods. These values reflect the superiority of the proposed method.

TABLE 5. The average running time of our method and other comparison methods on TNO and KAIST datasets. (unit: seconds).

Methods	TNO	KAIST
CVT	0.88	0.90
DTCWT	0.28	0.37
MSVD	0.29	0.30
CBF	13.90	14.35
GFF	0.40	0.37
GTF	4.75	7.49
HMSD-GF	1.42	1.24
IFEVIP	0.15	0.13
FCNN	56.99	60.69
GF	0.53	0.66
WLS	3.15	3.36
LatLRR	83.60	89.35
FusionGAN	3.85	3.65
DenseFuse	3.11	3.08
NestFuse-a	0.45	0.26
NestFuse-m	0.55	0.39
NestFuse-n	37.94	39.57
Ours (level1)	33.26	36.22
Ours (level2)	69.09	76.17
Ours (level3)	103.67	116.17
Ours (level4)	138.39	156.47

After quantitative and qualitative analyses of the experimental results on TNO and KAIST datasets, we can draw a conclusion that the proposed method can realize satisfactory image fusion and outperforms than most of existing fusion methods.

3) DISCUSSION ON TIME EFFICIENCY

Table 5 gives the average running time of the proposed fusion method and other comparison methods on TNO and KAIST datasets. As shown in Table 5, the LatLRR method consumes more time to fuse a pair of images than many algorithms since it needs sliding window during the fusion process. Our method is designed based on the LatLRR model, so it also takes a long time to fuse a pair of images. The average running time of these methods (including LatLRR, FCNN and NestFuse-n) are larger than our method (level 1). As presented in Table 5, many methods require less running time than the proposed method. However, there are still many methods that cannot meet the real-time requirement.

Our work focuses on the quality of the fused images. Although the proposed infrared and visible image fusion method has a certain complexity, it performs better than lots of classical and state-of-the-art methods. Considering the previous subjective and objective analysis results, we can

still say that the method proposed in this article has a good performance and is suitable for the condition without the real-time requirement. In the future work, we will try to accelerate the image fusion speed of our method.

V. CONCLUSION

In this paper, we propose an effective method for infrared and visible image fusion, which can produce fused images with strong intensity information, high visual quality, rich texture details, and sharpened edges. Source images are decomposed into base layer and detail layer by the decomposition method. For the fusion of base layer, an excellent strategy guided by the saliency map is designed, which can preserve suitable intensity information and improve the visual quality of the fused images. For the fusion of detail layer, an ingenious approach is constructed by utilizing the enhanced gradient information. This approach can increase the details information and sharpen the edges of the fused image. Moreover, lots of comparison experiments are conducted, which convincingly proves the effectiveness and advantages of the proposed fusion framework. In addition, image fusion technology has been widely used in many fields of computer vision. Therefore, in the future, we will try to apply the proposed fusion method to some computer vision tasks, such as target recognition.

ACKNOWLEDGMENT

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

REFERENCES

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.
[2] S. G. Simone, A. Farina, F. C. Morabito, S. B. Serpico, and L. Bruzzone, "Image fusion techniques for remote sensing applications," *Inf. Fusion*, vol. 3, no. 1, pp. 3–15, Mar. 2002.
[3] P. Kumar, A. Mittal, and P. Kumar, "Fusion of thermal infrared and visible spectrum video for robust surveillance," in *Proc. Indian Conf. Comput. Vis., Graph. Image Process.*, vol. 4338, Dec. 2006, pp. 528–539.
[4] 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.
[5] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, S.-J. Lee, and K. He, "Infrared and visual image fusion method based on discrete cosine transform and local spatial frequency in discrete stationary wavelet transform domain," *Infr. Phys. Technol.*, vol. 88, pp. 1–12, Jan. 2018.
[6] 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, p. 7006, May 2013.
[7] 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.
[8] J. Jinju, N. Santhi, K. Ramar, and B. S. Bama, "Spatial frequency discrete wavelet transform image fusion technique for remote sensing applications," *Eng. Sci. Technol., Int. J.*, vol. 22, no. 3, pp. 715–726, Jun. 2019.
[9] I. W. Selesnick, R. G. Baraniuk, and N. C. Kingsbury, "The dual-tree complex wavelet transform," *IEEE Signal Process. Mag.*, vol. 22, no. 6, pp. 123–151, Nov. 2005.
[10] 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.
[11] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2091–2106, Dec. 2005.
[12] J. Adu, J. Gan, Y. Wang, and J. Huang, "Image fusion based on nonsubsampled contourlet transform for infrared and visible light image," *Infr. Phys. Technol.*, vol. 61, pp. 94–100, Nov. 2013.
[13] X. Huang, G. Qi, H. Wei, Y. Chai, and J. Sim, "A novel infrared and visible image information fusion method based on phase congruency and image entropy," *Entropy*, vol. 21, no. 12, p. 1135, 2019.
[14] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, Feb. 2020.
[15] B. Yang and S. Li, "Multifocus image fusion and restoration with sparse representation," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 4, pp. 884–892, Apr. 2010.
[16] Z. Chen, X.-J. Wu, and J. Kittler, "A sparse regularized nuclear norm based matrix regression for face recognition with contiguous occlusion," *Pattern Recognit. Lett.*, vol. 125, pp. 494–499, Jul. 2019.
[17] Z. Zhu, G. Qi, Y. Chai, H. Yin, and J. Sun, "A novel visible-infrared image fusion framework for smart city," *Int. J. Simul. Process Model.*, vol. 13, no. 2, pp. 144–155, 2018.
[18] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
[19] X. Zhang, P. Ye, S. Peng, J. Liu, K. Gong, and G. Xiao, "SiamFT: An RGB-infrared fusion tracking method via fully convolutional Siamese networks," *IEEE Access*, vol. 7, pp. 122122–122133, Aug. 2019.
[20] Y. Cui, H. Du, and W. Mei, "Infrared and visible image fusion using detail enhanced channel attention network," *IEEE Access*, vol. 7, pp. 182185–182197, Dec. 2019.
[21] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, Mar. 2020.
[22] 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.
[23] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, and J. Wu, "Infrared and visible image fusion via detail preserving adversarial learning," *Inf. Fusion*, vol. 54, pp. 85–98, Feb. 2020.
[24] H. Li and X.-J. Wu, "Infrared and visible image fusion using latent low-rank representation," 2018, *arXiv:1804.08992*. [Online]. Available: http://arxiv.org/abs/1804.08992
[25] T. Nie, L. Huang, H. Liu, X. Li, and B. He, "Multi-exposure fusion of gray images under low illumination based on low-rank decomposition," *Remote Sens.*, vol. 13, p. 204, Jun. 2021.
[26] G. Liu and S. Yan, "Latent low-rank representation for subspace segmentation and feature extraction," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 1615–1622.
[27] Z. Yun and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia.*, Santa Barbara, CA, USA, Oct. 2006, pp. 815–824.
[28] 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.
[29] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5001715.
[30] Y. Wei, W. Fang, W. Zhu, and S. Jian, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis.*, vol. 7574, 2012, pp. 29–42.
[31] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.
[32] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.
[33] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.
[34] C. Zhao, Z. Wang, H. Li, X. Wu, S. Qiao, and J. Sun, "A new approach for medical image enhancement based on luminance-level modulation and gradient modulation," *Biomed. Signal Process. Control*, vol. 48, pp. 189–196, Feb. 2019.
[35] T. Zhao, J. Liu, J. Duan, X. Li, and Y. Wang, "Image quality enhancement via gradient-limited random phase addition in holographic display," *Opt. Commun.*, vol. 442, pp. 84–89, Jul. 2019.

[36] A. Toet. (2014). *TNO Image Fusion Dataset*. [Online]. Available: https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029
[37] S. Hwang, J. Park, N. Kim, Y. Choi, and I. S. Kweon, "Multispectral pedestrian detection: Benchmark dataset and baseline," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1037–1045.
[38] F. Nencini, A. Garzelli, S. Baronti, and L. Alparone, "Remote sensing image fusion using the curvelet transform," *Inf. Fusion*, vol. 8, no. 2, pp. 143–156, Apr. 2007.
[39] V. P. S. Naidu, "Image fusion technique using multi-resolution singular value decomposition," *Defence Sci. J.*, vol. 61, no. 5, pp. 479–484, 2011.
[40] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, 2015.
[41] 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.
[42] 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.
[43] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, Jul. 2016.
[44] Y. Zhang, L. Zhang, X. Bai, and L. Zhang, "Infrared and visual image fusion through infrared feature extraction and visual information preservation," *Infr. Phys. Technol.*, vol. 83, pp. 227–237, Jun. 2017.
[45] J. Ma and Y. Zhou, "Infrared and visible image fusion via gradientlet filter," *Comput. Vis. Image Understand.*, vols. 197–198, pp. 1077–3142, Aug. 2020.
[46] 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.
[47] H. Li, X.-J. Wu, and T. Durrani, "NestFuse: An infrared and visible image fusion architecture based on nest connection and spatial/channel attention models," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 12, pp. 9645–9656, Dec. 2020.
[48] W. J. Roberts, J. A. A. Van, and F. Ahmed, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, pp. 1–28, May 2008.
[49] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.
[50] G. Cui, H. Feng, Z. Xu, Q. Li, and Y. Chen, "Detail preserved fusion of visible and infrared images using regional saliency extraction and multi-scale image decomposition," *Opt. Commun.*, vol. 341, no. 15, pp. 199–209, Apr. 2015.
[51] 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.
[52] M. Volanthen, H. Geiger, K. J. Trundle, and J. P. Dakin, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, p. 355, Apr. 1997.
[53] 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.
[54] Y. Han, Y. Cai, Y. Cao, and X. Xu, "A new image fusion performance metric based on visual information fidelity," *Inf. Fusion*, vol. 14, no. 2, pp. 127–135, Apr. 2013.
[55] X. Han, T. Lv, X. Song, T. Nie, H. Liang, B. He, and A. Kuijper, "An adaptive two-scale image fusion of visible and infrared images," *IEEE Access*, vol. 7, pp. 56341–56352, 2019.

QINGQING LI received the B.E. degree from Hainan University, China, in 2017. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include image registration, image fusion, and deep learning.

GUANGLIANG HAN received the M.S. and Ph.D. degrees from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science, in 2000 and 2003, respectively. He is currently a Research Fellow with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His current research interests include computer vision, image processing, and object tracking.

PEIXUN LIU received the Ph.D. degree from Jilin University, in 2015. He is currently an Associate Research Fellow with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His research interests include image processing, object detection, and robot automation.

HANG YANG received the B.S. and Ph.D. degrees from Jilin University, in 2007 and 2012, respectively. He is currently an Associate Research Fellow with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Science. His research interests include image restoration and object tracking.

JIAJIA WU received the B.S. degree from Northeastern University, Qinhuangdao, in 2017. She is currently pursuing the Ph.D. degree with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China. Her current research interests include RGBD saliency detection and deep learning.

DONGXU LIU received the B.E. degree from Nanjing University of Information Science and Technology, China, in 2018. She is currently pursuing the Ph.D. degree with the University of Chinese Academy of Sciences and Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her research interests include hyperspectral image classification and deep learning.