IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 30 January 2024, accepted 5 February 2024, date of publication 7 February 2024, date of current version 15 February 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3364050*

RESEARCH ARTICLE

# Two-Stream Edge-Aware Network for Infrared and Visible Image Fusion With Multi-Level Wavelet Decomposition

**HAOZHE WANG[1], CHANG SHU[1], XIAOFENG LI[1], YU FU[2], ZHIZHONG FU[1], AND XIAOFENG YIN[3]**

[1]School of Information and Communication Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China
[2]Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University, Wuxi, Jiangsu 214122, China
[3]China International Marine Containers (Group) Company Ltd., Nantong, Jiangsu 226001, China

Corresponding author: Xiaofeng Li (xfli@uestc.edu.cn)

**ABSTRACT** Infrared and visible image fusion (IVIF) aims to generate a fused image with both salient target and rich textures from two different complementary modality images. To better integrate valuable edge information into the fused image, we first propose a novel two-stream network based on Auto-Encoder (AE) framework, which extracts deep hierarchical detail information at coarse scale from base stream by multi-level wavelet decomposition progressively and incorporates them into detail stream for information compensation. The aggregation of edge information ranging from coarse to fine facilitates a more comprehensive representation of contours and textures. Then, we propose a new feature fusion strategy, termed as Structural Feature Map Decomposition (SFMD). The first step is to decompose local patches of feature map with each modality into three independent components by Structural Patch Decomposition (SPD). In the second step, appropriate fusion rules are carefully designed for each component and the fused patch can be derived by inverse SPD. Our extensive experiments on several benchmark datasets show that our method outperforms seven compared state-of-the-art methods, especially in human visual perception.

**INDEX TERMS** Image fusion, wavelet decomposition, edge information, multi-scale analysis.

## I. INTRODUCTION

Image fusion plays an important role in the field of image process [1]. The definition of image fusion is to integrate multiple images acquired from the same or different modalities into one single image which carries all complementary information from all images [2]. The fused image is more informative and accurate than any of the source images. Image fusion not only reduces the amount of data [3], but also facilitates subsequent tasks, such as semantic segmentation [4] and object tracking [5]. In recent years, sensor technology has been rapidly developed [6] and researchers are increasingly interested in acquiring comprehensive descriptions of a scenario using multiple sensors, which distinctive information from various modalities can be provided.

Different types of images, such as multi-exposure images, multi-focus images and infrared/visible images are all typical images for fusion. Among these types, IVIF is the most promising and widely used for civilian and military applications, like fruit detection [7] and night-vision object tracking [8]. Source images can provide different properties of the same scene since they come from two modalities. Infrared images capture thermal radiation from objects, which are robust to illumination and all weather conditions. Targets can usually be clearly distinguished from the background in infrared images, as indicated by the high gray value [2]. However, infrared images often have low resolution and poor textures due to the limitations of imaging sensors.On another hand, visible images provide high-resolution and

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

enriched texture details, making them suitable for human visualization. But visible images are highly sensitive to illumination conditions and harsh environments, which makes it difficult to observe targets in low-light or foggy environments. By combining the merits of each source image, a fused image can highlight targets while retain the most edge information, resulting in a comprehensive and accurate description of the image scene.

Key challenges in IVIF task are how to extract features from different source images effectively and design appropriate fusion strategies. To solve these challenging issues, various fusion methods have been proposed and can be categorized into traditional and deep-learning methods. In the early stages, traditional methods relied on some priors, such as multi-scale [9], [10], [11], sparsity [12], [13], [14], [15] and saliency [16], [17], [18], to extract features from infrared and visible source images and maximum or average fusion rules are employed. However, too many priors challenge the effectiveness and robustness of such methods. Recently, deep-learning methods [19], [20], [21], have been gradually applied to IVIF task. Noteworthy, the high-quality, large-scale and paired infrared and visible images are uneasily accessible, thus limiting the performance of deep learning methods. On the other hand, most deep-learning methods solely focus on the spatial domain, whereas ignoring that the transform domain can capture structural information in compact way and provide better visual perception [22], [23]. For instance, GAN-based methods like FusionGAN [24] only constrained the discriminator to judge whether the fused image is similar with visible image in the style. Some inherited work [25], [26] designed special loss functions of discriminator to preserve textures in the fused image. But the results are still blurred and unsatisfactory. The structural details of features are not fully exploited in the existing methods. Consequently, it is worthwhile to investigate how to integrate the structural details in the feature level to the neural networks, which further improves the quality of the fused image.

In this paper, we introduce a two-stream edge-aware network with multi-level decomposition for feature extraction. Additionally, we propose a novel fusion strategy to strategically fuse these features. The overall framework of our method is illustrated in Fig. 1. Specifically, we adopt Auto-Encoder (AE) processing framework and initially decompose the input image into base and detail bands by discrete wavelet transform (DWT). Different network modules are designed in two streams to deal with base and detail image bands separately. The prior structural knowledge can be naturally embedded into the network architecture. Subsequently, multi-level DWT is utilized to progressively extract edge information at coarser scales in deep network layers from base stream and interact them into detail stream since the base bands still cover most information from the image and coarse contour edges are salient after one-level decomposition. By fully extracting latent edge details and enabling the detail stream to derive comprehensive edge

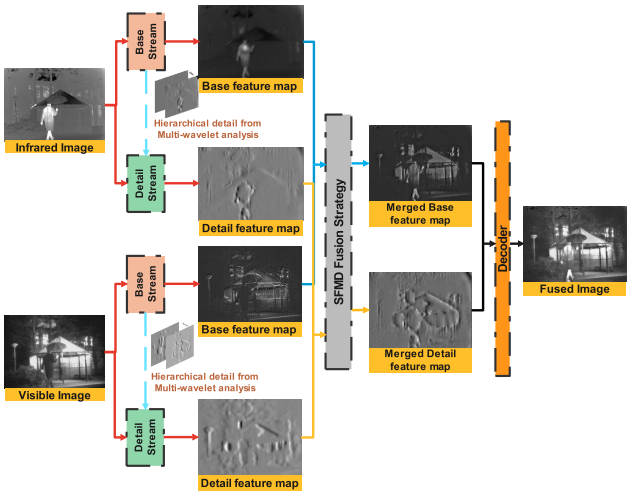


**FIGURE 1.** Framework of the proposed method.

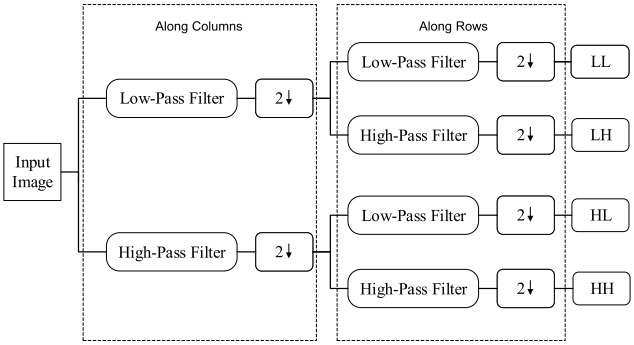


**FIGURE 2.** The procedure of 2D-DWT.

information across multiple scales from shallow to deep, the network capacity for representing detail features is greatly improved in our method. Consequently, it retains abundant edge information in the fused image and achieves good fusion performance with high visual quality.

Given that it is too simple to merge features by common average or $L_1$-norm fusion strategies [27], this article decomposes each patch of feature map into three independent components by Structural Patch Decomposition (SPD) theory [28]. We carefully design appropriate rules to fuse each component according to their modality characteristics. Experimental results show that salient targets can be better highlighted and textural edges is more natural in the fused image by our proposed fusion strategy.

In summary, the main contributions are listed as follows.

- We present a two-stream edge-aware network that incorporates deep hierarchical edge feature information at coarse scales and deep network layers in base stream into detail stream through multi-level wavelet decomposition. Only one level decomposition is insufficient to capture coarse details such as contours in the transform domain. Meanwhile, information loss is also avoided by invertible DWT operation.

- A novel fusion strategy, Structural Feature Map Decomposition (SFMD), is proposed to fuse the extracted features. Different from naive weight-averaged, chosen-max and $L1$-norm fusion strategies, SFMD explores inherent statistical relationships between each pair of extracted features from infrared and visible modalities and divides each patch into three independent components. Appropriate fusion rules are designed for each component to highlight thermal salient targets and keep natural textures in the fused image.
- Extensive ablation and comparative experiments have demonstrated that our proposed method is effective and transcends most of the state-of-the-art (SOTA) fusion methods. Qualitative and quantitative results on TNO and RoadScene datasets validate the strong generalization ability of our proposed model and feature fusion strategy.

The rest of this paper is organized as follows. In Section II, the related work is briefly described. In Section III, the proposed network model and the fusion strategy will be explained in detail. In Section IV the fusion performance of the proposed method is analyzed and compared with other methods extensively. In Section V, the effectiveness of our proposed method is also demonstrated in the field of medical image fusion. Limitations of our method are discussed in Section VI. Finally, conclusions are drawn in Section VII.

## II. RELATED WORKS

### A. MULTI-SCALE DECOMPOSITION FUSION METHODS

Multi-scale decomposition is one of the most popular techniques in the image fusion [7]. The scale refers to the spatial resolution of the image. The core idea is that different scale spaces represent different information in the image. Fine scales represent more local textural information of the image and coarse scales represent global and semantic information. The coarse to fine characteristic of multi-scales has been demonstrated that it is consistent with human visual system, enabling a good visual effect [22], [29]. Classical multi-scale decomposition methods are usually pyramid transform [30], [31], discrete cosine transform [32] and non-subsampled contourlet transform [33], [34]. Generally, the basic steps of multi-scale decomposition fusion are as follows. Firstly, a group of filter banks are used to decompose the image to transform domain with different scales. Secondly, the coefficients of different scales in transform domain are fused by given fusion rules. Finally, the inverse transform is applied to reconstruct the fused image.

### B. 2D DISCRETE WAVELET TRANSFORMATION (2D-DWT)

Wavelet transform has long been a powerful tool in image processing due to its strong time-frequency analysis and energy compaction [35], [36]. As illustrated in Fig. 2, high-pass and low-pass filter banks are applied along with rows and columns to extract approximation coefficients (LL) and detail coefficients in horizontal (LH), vertical (HL) and diagonal directions (HH) in four sub-bands. The resolution of sub-bands are downsampled by two factor compared with that of input image, preventing information redundancy. An example of 2D-DWT decomposition in one level is as shown in Fig. 3. It should be noted that 2D-DWT is invertible, which four sub-bands images can be inverted into original image by 2D inverse discrete wavelet transform (2D-IDWT) without loss.

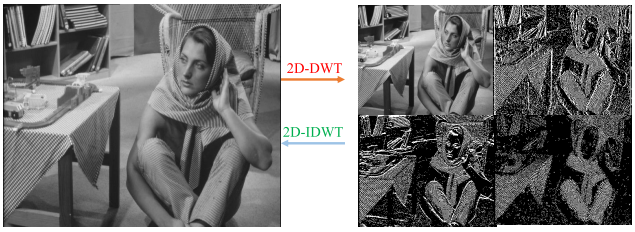


**FIGURE 3.** An example of 2D-DWT. From top left to bottom right are approximation coefficient,vertical detail coefficient, horizontal detail coefficient and diagonal coefficient respectively in the right part.

In image fusion, the input image is usually decomposed by 2D-DWT in multi-levels, which means the approximation and detail information separated in one level are decomposed again, forming a set of approximate and detail coefficients with different resolution. The reason is that much information still exists in the approximation coefficients in one-level [37]. Hence, multi-level 2D-DWT is used to extract approximation and detail information sufficiently. Then, approximate sub-band and detail sub-bands with same resolution are fused by given fusion rules respectively. Finally, the set of fused sub-bands are recovered to achieve fused image by 2D-IDWT. Wavelet-variants like curvelet [38], dual-tree wavelet [39], contourlet [40] and non-subsampled contourlet [33] were proposed later to improve the ability of anisotropy representation from image at the sacrifice of processing time. In this paper, the Haar wavelet kernel is adopted due to its efficiency and fast-implementation.

### C. DEEP LEARNING-BASED FUSION METHODS

Currently, methods based on AE is still very popular in the field of IVIF. The AE methods usually train an autoencoder to extract features. Then, the intermediate feature fusion is realized according to carefully-designed fusion rules. Finally, the autodecoder reconstructs the fused image by the fused features. Li and Wu [27] firstly utilized dense network as encoder for feature extraction. Except for introducing the dense connection, he also proposed $l_1-$norm fusion rule and attained better performance than the traditional fusion rules. Zhang et al. proposed a unified fusion network known as IFCNN, which dynamically selects the fusion strategy for the deep features according to different fusion tasks [41]. Recently, Zhao et al. [42] adopted transformer architecture as the encoder for better modeling the long-range dependence in the feature domain. Xu et al. [43] used two pairs of encoders to extract shallow and deep features and decompose

them into common and unique parts respectively. After that, different fusion rules can be applied according to the flexible requirements.

## III. PROPOSED METHOD

The network architecture of the proposed method is illustrated in Fig. 4. In the training phase, the encoder is trained to extract features from the input image and the decoder learns how to reconstruct the original image by corresponding features. To ensure that the encoder is well adapted to both modalities, we use an equal number of visible and infrared images as input for training. The encoder consists of base and detail stream, with each stream appropriately processing global and local detail information respectively. The resulting base and detail feature maps are then channel-concatenated and sent to the decoder. In the testing phase, two identical trained encoders are used to extract features of the visible and infrared images separately. SFMD fusion strategy is then applied to fuse base and detail features from visible and infrared images respectively. Finally, the trained decoder generates the fused image from the resulting fused feature maps. In the next part, we will analyze encoder-decoder architecture, loss function and SFMD fusion strategy.

### A. ENCODER-DECODER ARCHITECTURE

The architecture details of the encoder part are illustrated in the upper part of Fig. 4, consisting of base and detail streams. The input image is firstly decomposed into approximation and detail coefficients with 2D Haar wavelet transform which is given as follows,

$$\begin{cases} A = (a+b+c+d)/4 \\ B = (a-b+c-d)/4 \\ C = (a+b-c-d)/4 \\ D = (a-b-c+d)/4, \end{cases} \tag{1}$$

where *a, b, c, d* are four pixels in every $2 \times 2$ block, and *A* is the resulting approximation sub-band and *B, C* and *D* are the resulting detail sub-bands. The resolution of *A, B, C, D* is the quarter of the input image. Next, the approximation sub-band is the initial input for base stream and the three detail sub-bands are input for the detail stream.

U-Net architecture [44] is adopted in the base stream since it is proved that is very suitable for extracting global information of an image and its multi-scale characteristic caters for our multi-level wavelet decomposition design. Instead of simple downsampling and upsampling operations in a normal U-Net, pairs of forward and inverse Haar wavelet transforms are used in our scheme. The wavelet transform avoids information loss due to its invertibility and the sparse wavelet coefficients have more compact and directional structural edge representation than down-sample images directly. Different kernel sizes are used to enlarge the receptive filed and extract scale information more effectively.

A four layer denseblock [45] is designed for the detail stream to keep detail features from shallow to deep. The dense concatenation has three advantages. Firstly, it preserves more information by feature reuse. In low-level image processing, the shallow and deep detail information are both valuable. Secondly, the problem of gradient degradation is solved by channel concatenation to some extents. Therefore, the training is easy to convergence. Thirdly, the dense connection reduces the overfitting effect due to its regularity. Details in different decomposition scales and depth, which are separated from base stream, are concatenated with different feature maps in the detail stream for subsequent operations. In the detail stream, the kernel size of $3 \times 3$ is adopted, focusing on local detail information. And the size of feature maps is unchanged to avoid information loss.

The decoder has six sequential convolutional blocks to reconstruct the original image from the feature maps produced by base and detail streams. The last convolutional block, Conv without batchnorm and parametric rectified linear activation is adopted, generating the wavelet coefficients of reconstructed image. Finally, a 2D-IDWT is applied to reconstruct the original image, corresponding to the initial 2D-DWT in the encoder.

### B. MULTI-LEVEL WAVELET DECOMPOSITION

The motivation for us to interact hierarchical multi-scale edge information in base stream into the detail stream is that only one-level decomposition is not sufficient to separate all base and detail information of an image since the approximation sub-band includes most main information of the image. The initial detail sub-bands are at finest scales, which only carries some local details information. Therefore we progressively use DWT in base stream to extract latent feature edge information at coarse scales from shallow to deep layers, and interact them into detail stream for detail information aggregation. There are totally four-level wavelet decomposition and a set of hierarchical multi-scale feature edge information is shown in the green box in Fig. 4, which forms a comprehensive representation for contour and local details. It is worthy noted that the hierarchical edge feature information is copied, not destroying the feature integrity of image information in base stream.

Specifically, let *H* denote as the Haar wavelet decomposition process. The multi-level wavelet decomposition in the proposed encoder is formulated as follows,

$$H\left(I^{i,j}\right) = \left[I_{LL}^{i+1,j}, I_{LH}^{i+1,j}, I_{HL}^{i+1,j}, I_{HH}^{i+1,j}\right], \tag{2}$$

where *i* is the level order of decomposition and *j* is the $j^{th}$ channel from feature maps that to be decomposed in base stream. $I^{0,0}$ is the input grayscale image.

The approximation coefficients $I_{L,L}^{1,0}$ decomposed from $I^{0,0}$ is the input for base stream and it passes a convolution block with kernel size $7 \times 7$ . The $I_{L,L}^{2,0}$ is concatenated with the output of the first convolutional block. The resulting feature map is represented as $B_1 = \left[F\left(I_{L,L}^{1,0}\right), I_{LL}^{2,0}\right]$, where *F* represents the

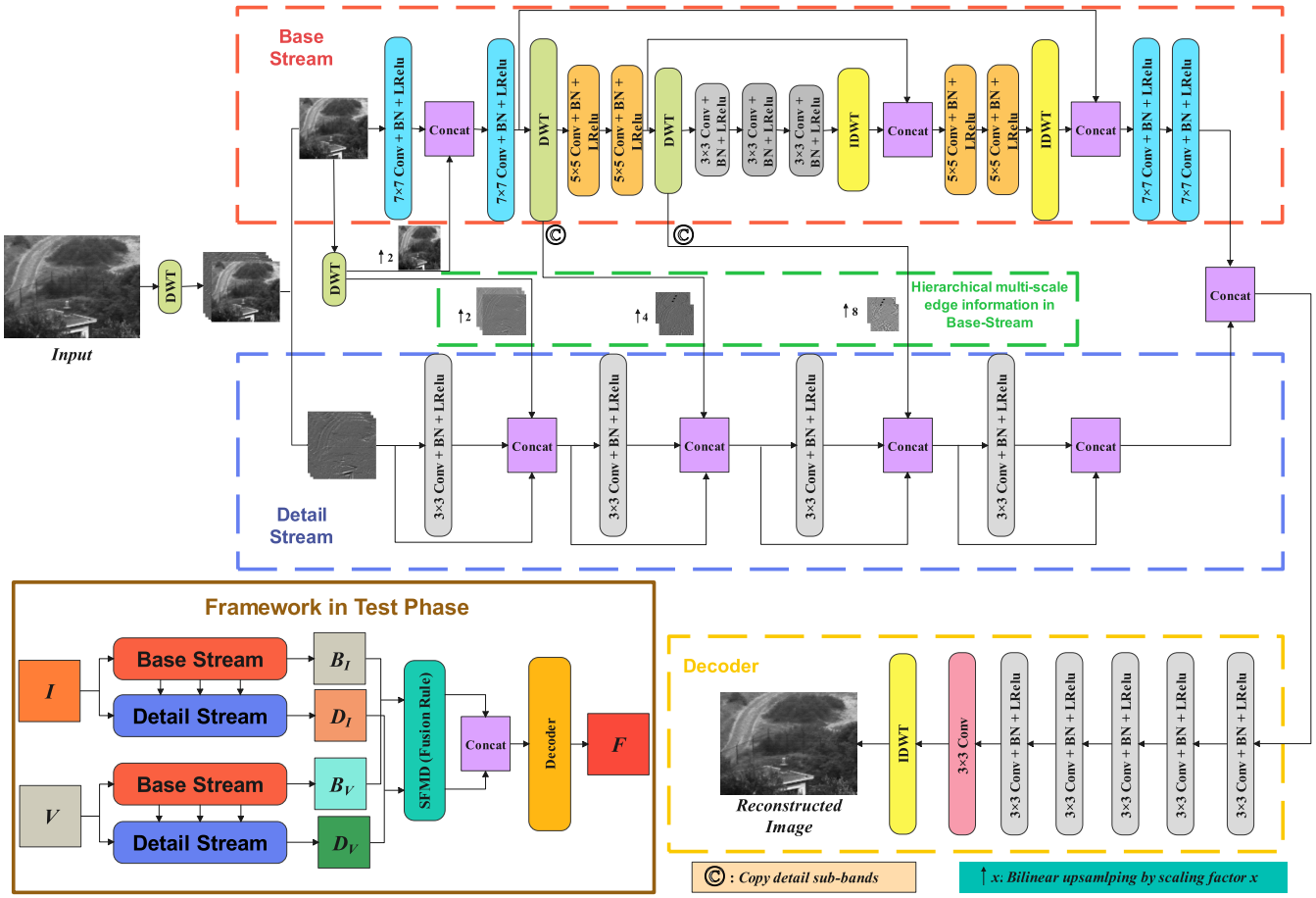


**FIGURE 4.** **Flowchart of our proposed model. In the training phase, we first adopt U-Net architecture for base stream to process image information at coarse scales. Then, downsampling and upsampling are replaced by DWT and IDWT operation to extract hierarchical multi-scale edge information in base stream. Finally, the hierarchical multi-scale edge information is incorporated into the detail stream for detail information aggregation.In the testing phase, we propose a SFMD fusion strategy to fuse base and detail feature maps from different modalities, highlighting thermal salient targets and keeping natural textures in the fused image.**

process of reflected padding, convolution, batch-norm and parametric rectified linear unit operations. Then, the feature maps pass another convolution block with kernel size $7 \times 7$ , denoted as $F(B_1) = \left[I^{2,0} \cdots I^{2,15}\right]$. A 2D-DWT is applied to $\left[I^{2,0} \cdots I^{2,15}\right]$, denoted as $H\{\left[I^{2,0}, \cdots, I^{2,15}\right]\}$. The results $\left[\left[I_{LL}^{3,0}, I_{LH}^{3,0}, I_{HL}^{3,0}, I_{HH}^{3,0}\right], \cdots, \left[I_{LL}^{3,15}, I_{LH}^{3,15} I_{HL}^{3,15}, I_{HH}^{3,15}\right]\right]$ are processed in the next level of the U-Net architecture. $\left[\left[I_{LH}^{3,0}, I_{HL}^{3,0}, I_{HH}^{3,0}\right], \cdots, \left[I_{LH}^{3,15}, I_{HL}^{3,15}, I_{HH}^{3,15}\right]\right]$, the three detail sub-bands generated in the base stream are copied, upsampled and incorporated into the detail stream. The similar steps are repeated in the second and third level of the U-Net. As the resolutions are decreased by quarter in the next levels, $5 \times 5$ and $3 \times 3$ kernel size are adopted respectively to keep the same receptive field. In the expanding path of U-Net, deconvolution operations are replaced with 2D-IDWT.

### C. LOSS FUNCTION

The loss function is designed for reconstructing original image in the training phase. Mean-Square-Error (MSE) loss is often the main part of loss function in the low-level image processing task, which evaluates average mean square difference of pixel intensity between reconstructed image and input image. However, the MSE loss function is not sensitive enough to detail information and it usually results in a smooth reconstructed image. Structural Similarity Index Measure (SSIM) measures similarity between two different images by luminance, structure and contrast, which is proven to be consistent with human visual perception on the image quality. Considering that infrared image may be kind of noisy, Total-Variation Loss (TV) is introduced as regularization term. The total loss function is defined as follows,

$$L_{\text{total}} = L_{MSE}(X, \hat{X}) + \lambda_1 L_{SSIM}(X, \hat{X}) + \lambda_2 L_{TV}(\hat{X}), \quad (3)$$

where X is the input visible/infrared image, and $\hat{X}$ is the reconstructed image and $\lambda_1$, $\lambda_2$ are hyperparameters which controls the trade-off.

$L_{MSE}$ computes the mean square intensity difference between input and reconstructed images, defined as follows,

$$L_{MSE} = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} \left(X_{ij} - \hat{X}_{ij}\right)^2, \quad (4)$$

where H is the number of pixel in each column and W is the number of pixels in each row. $X_{ij}$, $\hat{X}_{ij}$ denote the pixel value in the $i^{th}$ row and $j^{th}$ column in the input image and the reconstructed image respectively.

*SSIM* [46] is the structural similarity index whose value is usually between 0 and 1. The formula of SSIM is as follows.

$$SSIM(X, Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}, \quad (5)$$

where $\mu$ and $\sigma$ are mean intensity and standard deviation of the image. $\sigma_{XY}$ is the correlation coefficient of two images. $C_1$ and $C_2$ are small constants, which are set to avoid instability when $\mu$, $\sigma$ of two images are close to zero. A larger *SSIM* means the luminance, structure and contrast of two images are more similar. The term $L_{SSIM}$ is optimized in the loss function. It is the dominant loss function term in our model. The formula of $L_{SSIM}$ is as follows,

$$L_{SSIM} = 1 - SSIM(X, \hat{X}). \quad (6)$$

Total Variation Loss (TV loss) [48] is used to suppress the noise in the reconstructed image as a regularization term. The formulation of TV loss is as follows,

$$L_{TV} = \sum_{i,j} \Big( \|\hat{X}(i, j+1) - \hat{X}(i, j)\|_2 / W + \|\hat{X}(i+1, j) - \hat{X}(i, j)\|_2 / H \Big), \quad (7)$$

where $\|\|_2$ is $L_2$-norm, and $\hat{X}(i, j)$ denotes the pixel value in the $i^{th}$ row and $j^{th}$ column in the reconstructed image. H and W are the numbers of rows and columns respectively.

### D. FUSION STRATEGY

In the testing phase, the fusion layer is inserted to fuse feature maps of infrared and visible images from encoders. It fuses feature maps from base and detail stream separately. The mathematical representation is as follows.

$$B_F = \text{Fusion}(B_I, B_V), \quad (8)$$

$$D_F = \text{Fusion}(D_I, D_V), \quad (9)$$

where $B_I$ and $B_V$ are infrared and visible feature maps from base stream respectively. $D_I$, $D_V$ are infrared and visible feature maps from detail stream respectively and $B_F$, $D_F$ are fused results.

The common fusion strategies for feature maps include average, chosen-max and $L_1$-norm [27]. Average and chosen-max strategies process each pixel with same weight map, ignoring characteristics of feature maps from different modalities. They are too naive to deal with complicated fusion cases. $L_1$-norm strategy is based on channel-attention mechanism, simply designing weight maps by summing each channel. Most existing fusion strategies ignore the local structural relationship between different modalities. Therefore, it is necessary to develop an approach for fusing feature maps from infrared and visible images more reasonably.

Inspired by Structural Patch Decomposition (SPD) proposed by Ma et al. [28] and the inherited work by Li et al. [47] in multi-exposure fusion, we develop Structural Feature Map Decomposition (SFMD) for fusing feature maps from infrared and visible images. In SPD, a local image patch is decomposed into three parts: signal strength, signal structure and mean intensity.

$$\begin{aligned} \mathbf{x} &= \|\mathbf{x} - \mu\| \cdot \frac{\mathbf{x} - \mu}{\|\mathbf{x} - \mu\|} + \mu \\ &= \|\tilde{\mathbf{x}}\| \cdot \frac{\tilde{\mathbf{x}}}{\|\tilde{\mathbf{x}}\|} + \mu \\ &= c \cdot \mathbf{s} + \mu, \end{aligned} \quad (10)$$

where $\mu$ is the mean value of the signal patch, and $\|\cdot\|$ is the $l_2$ norm of the mean-removed signal patch, and $\tilde{\mathbf{x}}$ is the mean-removed signal patch, and $c$ is the scalar of signal strength and $\mathbf{s}$ is the unit-vector of the mean-removed signal patch respectively. The decomposition operation is also invertible, which means the original signal patch can be recovered by three independent components.

The mean intensity, signal strength and signal structure represent the luminance, contrast and texture information of an image patch. Similar concepts can be applied to the patches of feature maps. From a statistical perspective, the signal strength is analogous to the standard deviation of the patch which indicates the degree of dispersion of data distribution. The signal structure represents the mean-removed data distribution. In image processing, a larger standard deviation of the patch usually means that more enriched details are presented. Although CNN is still a black-box, the feature maps can also be disentangled by statistical information and appropriate fusion rules are designed for each component to achieve the goal of fusing infrared and visible images. This leads to our SFMD feature fusion strategy.

As shown in Fig 5, sliding window of size $31 \times 31$ is used to divide patches in each feature map. Then each patch is decomposed into three independent components by SPD. Different fusion rules are designed for fusing each component respectively. Finally, fused patch is derived by inverse SPD.

For fusion of mean intensity in SFMD, softmax rule is proposed to assign fusion weights between infrared and visible feature map patches. Mean intensity reflects average brightness in each patch, which is very closely related to thermal target region. Softmax function is used to enlarge the difference between fusion weights, which ensures the target region to be salient after fusion. Meanwhile, the extreme fusion rule like maximum is not adopted because brightness information of the other modality should also be retained in fused image. Denote $\mu_{IR}^i$ and $\mu_{VIS}^i$ as the local mean intensity of the $i^{th}$ channel of feature maps from infrared and visible images, $\hat{\mu}_{FUSE}^i$ as the local mean intensity of the fused $i^{th}$ channel in the feature map. The formulas are as follows.

$$\hat{\mu}_{FUSE}^i = \alpha_1 \cdot \mu_{IR}^i + \alpha_2 \cdot \mu_{VIS}^i, \quad (11)$$

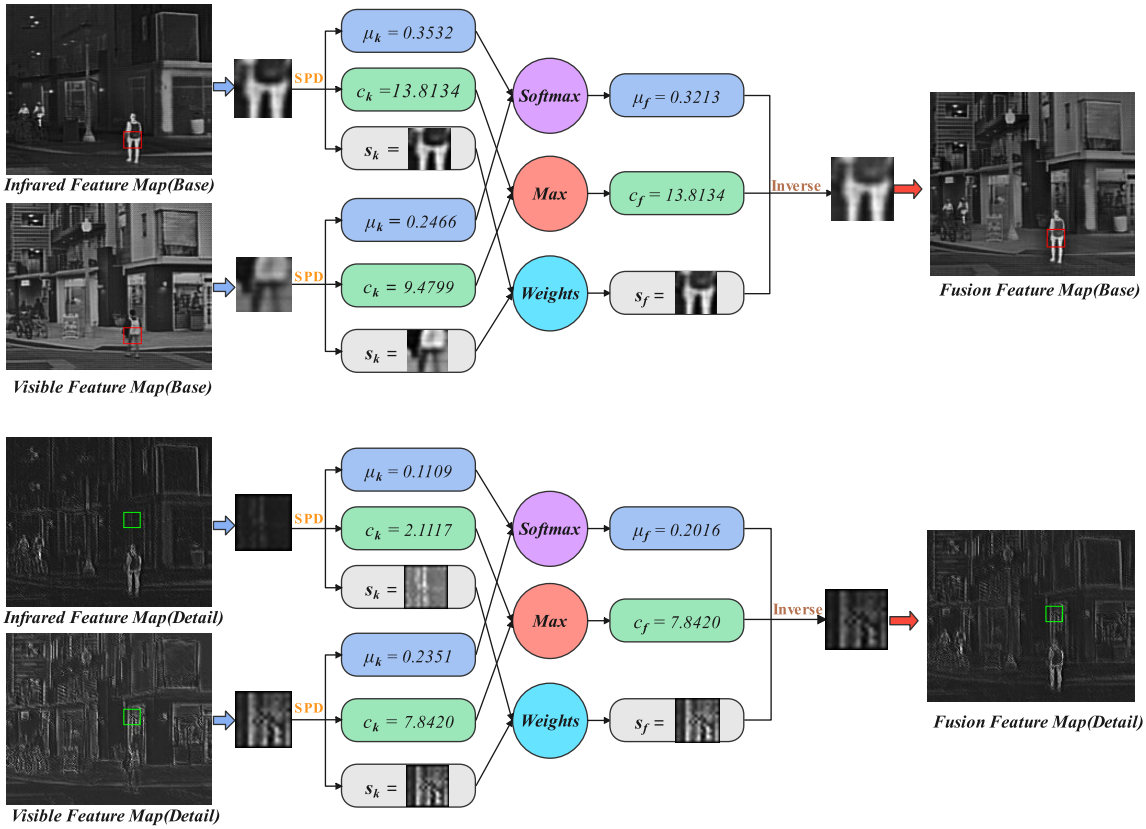


**FIGURE 5.** Flowchart of our SFMD Fusion Strategy. We firstly use 31 × 31 sliding window to divide the whole feature map into local patches. Secondly, we decompose each patch into signal strength $c_k$, signal structure $s_k$ and mean intensity $\mu_k$ three independent components by SPD [28]. Thirdly, we design appropriate rules for fusing each component. Finally, we derive the fused local patch by inverse SPD and the final fused feature map can be derived by patch aggregation [47].

$$\alpha_1 = \frac{\exp\left(k \cdot \mu_{IR}^i\right)}{\exp\left(k \cdot \mu_{IR}^i\right) + \exp\left(k \cdot \mu_{\text{VIS}}^i\right)}, \tag{12}$$

$$\alpha_2 = \frac{\exp\left(k \cdot \mu_{VIS}^i\right)}{\exp\left(k \cdot \mu_{IR}^i\right) + \exp\left(k \cdot \mu_{VIS}^i\right)}, \tag{13}$$

where $k$ is scaling factor, controlling the degree of magnifying the difference between weight coefficients. The value of $k$ is set as 8 empirically.

For fusion of signal strength in SFMD, maximum fusion rule is used to pick the larger one. Signal strength represents the contrast of local region in image. Corresponding to the feature map patch, it represents the standard deviation, i.e. the richness of variation in data distribution. The scalar signal strength does not involve the variability between different modalities, so the maximum value fusion strategy is used to ensure that the fluctuating changes in information are reflected in fused feature map to the greatest extent, bringing better visibility and higher contrast in fused image. The maximum fusion rule is as follows.

$$\hat{c}_{FUSE}^i = \max\left\{c_{IR}^i, c_{VIS}^i\right\} = \max\left\{\left\|\tilde{x}_{IR}^i\right\|, \left\|\tilde{x}_{VIS}^i\right\|\right\}, \tag{14}$$

where $c_{IR}^i$ and $c_{VIS}^i$ are the local signal strength of $i^{th}$ channel of feature maps from infrared and visible image and $\hat{c}_{FUSE}^i$ is the local signal strength of fused $i^{th}$ channel in feature map. $\tilde{x}_{VIS}^i$ is the mean-removed signal patch in the $i^{th}$ channel.

For fusion of signal structure in SFMD, an enhanced power function is adopted. There are two principles to follow. First, the fused signal structure should be a unit-vector. Second, the signal structure with larger signal strength should also be dominant in the fused signal structure. But it should be in a softer manner rather than softmax fusion since valuable detail information from two modalities is expected to be reflected in the fused image as much as possible. Therefore, a soft power function is used. Meanwhile, signal strength is used to guide the signal structure fusion. The formulas are shown as follows.

$$\tilde{s}_{FUSE}^i = \frac{\beta_1}{\beta_1 + \beta_2} \cdot s_{IR}^i + \frac{\beta_2}{\beta_1 + \beta_2} \cdot s_{VIS}^i, \tag{15}$$

$$\beta_1 = \left\|\tilde{x}_{\text{IR}}^i\right\|^p \quad \beta_2 = \left\|\tilde{x}_{VIS}^i\right\|^p, \tag{16}$$

$$\hat{s}_{FUSE}^i = \frac{\tilde{s}_{FUSE}^i}{\left\|\tilde{s}_{FUSE}^i\right\|}, \tag{17}$$

where $s_{IR}^i$ and $s_{VIS}^i$ are the local signal structure unit vector of $i^{th}$ channel of feature maps from infrared and visible image, and $\hat{s}_{FUSE}^i$ is the local signal structure unit vector of fused $i^{th}$ channel in feature map and $p \geq 0$ is an exponential parameter.

A greater $p$ means the patch with stronger signal strength is more transferred into the final fused signal structure. $p$ is set to 4 empirically in our paper.

The final fused patch can be derived by the inverse SPD operation, which is illustrated as follows. The fused channel can be derived by patch aggregation.

$$\hat{\mathbf{x}}^i_{FUSE} = \hat{c}^i_{FUSE} \cdot \hat{s}^i_{FUSE} + \hat{\mu}^i_{FUSE}. \tag{18}$$

## IV. EXPERIMENTS

In this section, we firstly introduce the experimental settings. Specifically, datasets and preprocessing details, seven state-of-the-art methods with evaluation metrics and hyperparameters setting are covered. Secondly, qualitative and quantitative experimental results of various fusion methods on two datasets are demonstrated. Finally, we conduct multiple ablation experiments to analyze the effectiveness of our model and the SFMD fusion strategy.

### A. EXPERIMENTAL SETTINGS

#### 1) DATASETS AND PRE-PROCESSING

The training dataset consists of 180 randomly selected pairs of infrared and visible image pairs from *RoadScene* [49] dataset. To augment the limited amount of training data, we randomly crop the images to a size of 256 × 256 and transform them into grayscale. The validation set consists of 52 pairs of infrared and visible images from NIR [50] dataset. The first test dataset includes 47 typical pairs of infrared and visible images from *TNO* [51] dataset, and the second test set includes 40 additional pairs of infrared and visible images from the *RoadScene* dataset. All these datasets are publicly available.

#### 2) COMPARISONS AND EVALUATION

Seven state-of-the-art methods are selected to compare with our proposed methods, namely, SDNet [52], SwinFusion [21], GANMcC [25], U2Fusion [53], UNFusion [19], IPLF [20] and CDDFuse [42]. The objective metrics selected are entropy (EN) [54], standard deviation (SD) [55], edge intensity (EI) [56], visual information fidelity (VIF) [57], Chen-Blum metric ($Q_{CB}$) [58] and MS-SSIM [59]. Larger EN mean the richer information transferred from source images to fused image. SD measures the visual contrast of fused image. EI measures the abundance of texture information. VIF and $Q_{CB}$ are consistent with human perception. MS-SSIM measures the similarity of source images and fused image. The fusion performance is better if all metrics are larger.

#### 3) HYPERPARAMETERS SETTING

The training epoch is set to 160 with the batch size of 32. The learning rate is $10^{-3}$ at first 80 epoches and decays by half for the last 80 epoches. The hyperparameters of loss function are adjusted by the fusion performance in the validation set. Experimental results of different groups of hyperparameters are illustrated in Table 1. The number in bold implies the best value in each metric. It is clearly reflected that when $\lambda_1$ is equal to 5 and $\lambda_2$ is equal to 0.1, the model achieves the best validation performance. Therefore, $\lambda_1$ and $\lambda_2$ are set to 5 and 0.1 in the experiment respectively.

**TABLE 1. Validation performance of different groups of hyperparameters.**

| $\lambda_1$ | $\lambda_2$ | EN | SD | EI | VIF | $Q_{CB}$ | MS-SSIM |
|---|---|---|---|---|---|---|---|
| 0.5 | 0 | 7.35 | 56.95 | **75.94** | 1.13 | 0.47 | 0.83 |
| | 0.01 | 7.21 | 55.07 | 65.44 | 1.18 | 0.48 | 0.87 |
| | 0.1 | 7.22 | 58.47 | 67.80 | 1.03 | 0.49 | 0.84 |
| 5 | 0 | 7.22 | 58.98 | 73.14 | 1.18 | 0.49 | 0.89 |
| | 0.01 | 7.24 | 59.31 | 66.20 | 1.07 | 0.49 | 0.83 |
| | 0.1 | **7.41** | 57.22 | 71.57 | **1.22** | **0.51** | **0.91** |
| 50 | 0 | 7.31 | 59.99 | 75.50 | 1.15 | 0.50 | 0.86 |
| | 0.01 | 7.40 | 61.65 | 73.81 | 1.19 | 0.46 | 0.82 |
| | 0.1 | 7.24 | **63.43** | 72.49 | 1.20 | 0.44 | 0.88 |

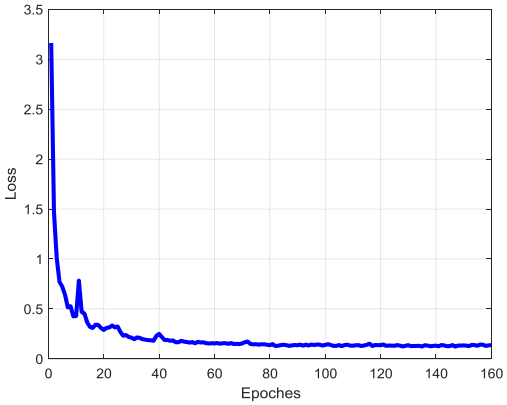


**FIGURE 6. The cost evaluation over training process.**

**TABLE 2. The average values of six metrics of 47 source image pairs from *TNO* dataset with different fusion methods.**

| Methods | EN | SD | EI | VIF | $Q_{CB}$ | MS-SSIM |
|---|---|---|---|---|---|---|
| GANMcC | 6.704 | 32.805 | 25.371 | 0.608 | 0.438 | 0.870 |
| SDNet | 6.654 | 32.937 | 40.195 | 0.708 | 0.481 | 0.877 |
| UNFusion | **6.987** | 41.073 | 37.299 | 0.836 | 0.477 | 0.876 |
| U2Fusion | 6.391 | 25.433 | 34.451 | 0.511 | 0.506 | 0.904 |
| CDDFuse | 6.851 | 42.596 | 40.072 | 0.986 | 0.488 | 0.875 |
| SwinFusion | 6.870 | 39.424 | 42.063 | 0.752 | 0.481 | 0.894 |
| IPLF | 6.957 | 39.414 | **72.681** | 0.754 | 0.504 | 0.849 |
| Ours | 6.976 | **44.845** | 41.913 | **1.019** | **0.512** | **0.907** |

#### 4) CONVERGENCE SPEED

As illustrated in Fig. 6, the network converges after about 80 epoches. The time of convergence is about 2 hours in one GPU. The hardware platform is NVIDIA TITAN X PASCAL GPU of memory size 32G and Intel i7-9700K CPU.

### B. EXPERIMENTS ON TNO DATASET

#### 1) QUALITATIVE EVALUATION

Three typical infrared and visible image pairs from *TNO* dataset, named *Nato Camp*, *Solider Behind Smoke* and *Bunker* are selected for qualitative evaluation. We marked

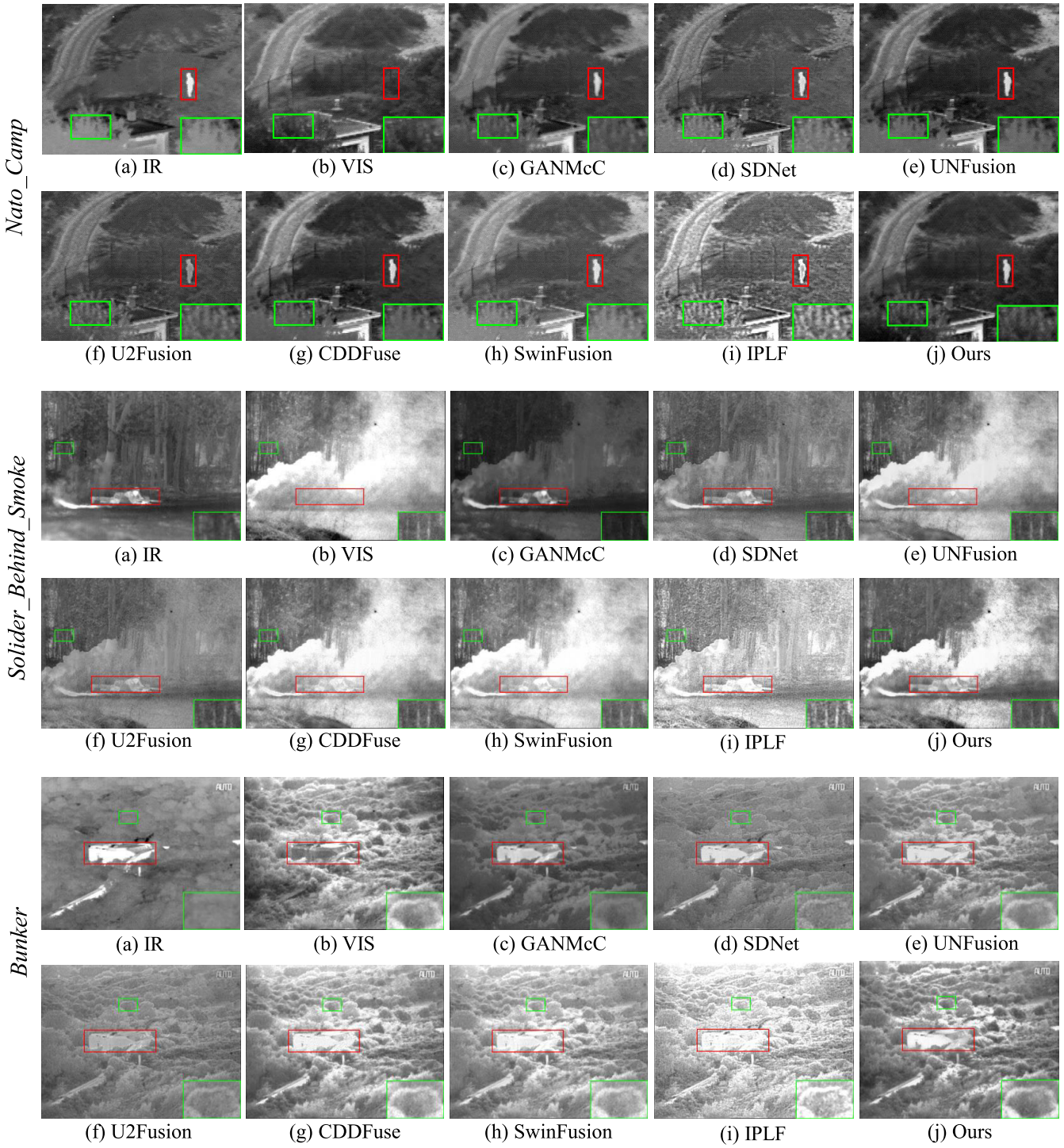


**FIGURE 7.** Qualitative comparisons on three typical image pairs from *TNO* dataset. The first two rows are the fusion results of *Nato Camp* image pair. The third and fourth rows are the fusion results of *Solider Behind Smoke* image pair. The last two rows are the fusion results of *Bunker* image pair.

salient target and texture region with red and green boxes respectively. The texture region is zoomed in the bottom right corner of the image for better visual comparison.

Fig. 7 shows the visual results of different methods. In *Nato Camp* image, U2Fusion loses the salient thermal target information which should be clearly reflected in the fused image. The fused image of GANMcC seems to be blurred in target and some local areas, such as the texture of the road. The fused images of SDNet, SwinFusion and IPLF are over-sharpened, and the image style bias to infrared modality. Most fusion methods fail to recover the contour and

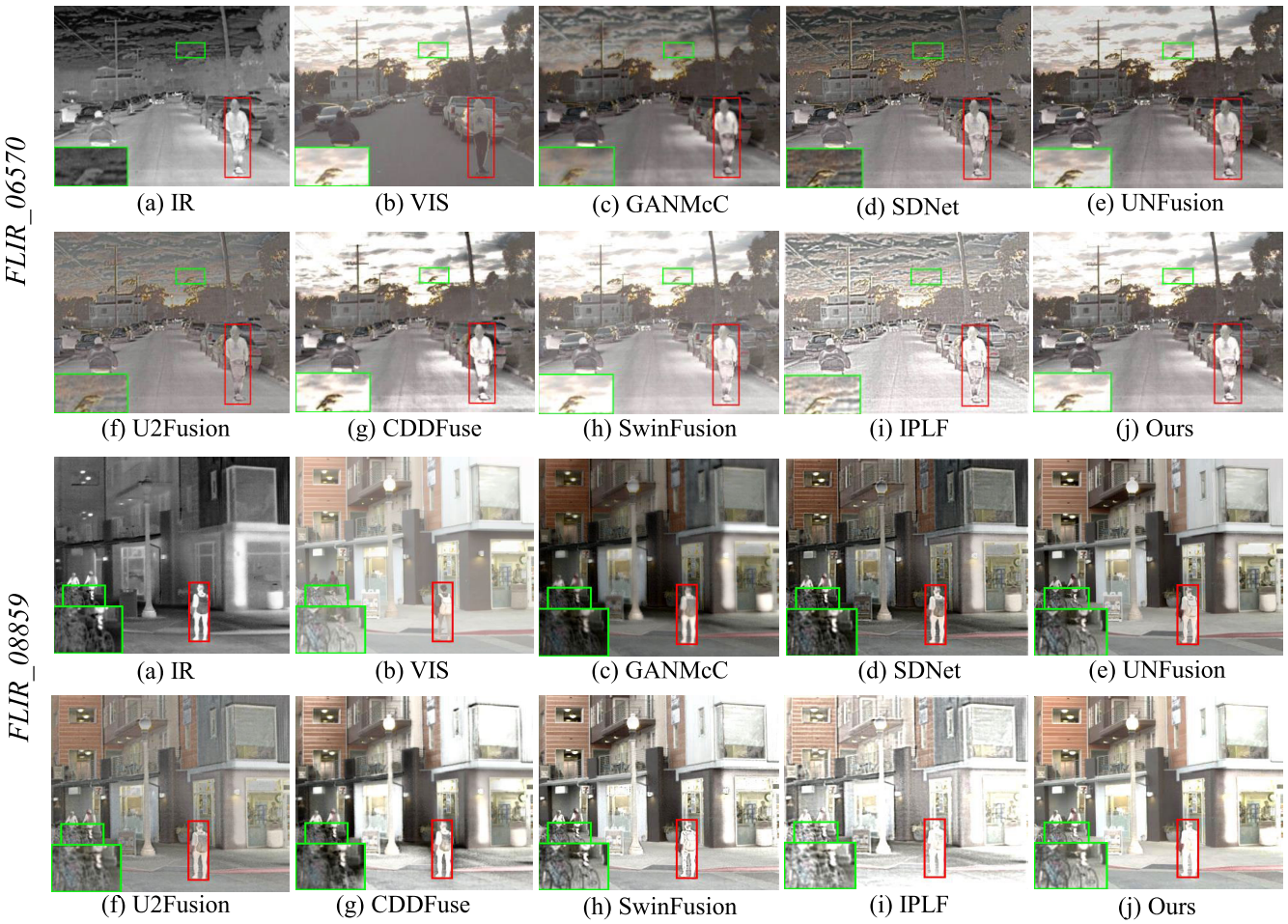


**FIGURE 8.** Qualitative comparisons on examples from ***RoadScene*** dataset. The first two rows are the fusion results of ***FLIR_06570*** image pair. The third and fourth rows are the fusion results of ***FLIR_08859*** image pair.

**TABLE 3.** The average values of six metrics of 40 source image pairs from ***RoadScene*** dataset with different fusion methods.

| Methods | EN | SD | EI | VIF | $Q_{CB}$ | MS-SSIM |
|---|---|---|---|---|---|---|
| GANMcC | 7.208 | 42.449 | 39.391 | 0.589 | 0.477 | 0.849 |
| SDNet | 7.292 | 43.461 | 64.390 | 0.707 | 0.493 | 0.829 |
| UNFusion | 7.372 | 49.286 | 53.015 | 0.826 | 0.497 | 0.843 |
| U2Fusion | 6.859 | 32.139 | 50.511 | 0.483 | **0.507** | 0.890 |
| CDDFuse | 7.293 | 50.351 | 55.306 | 0.890 | 0.486 | 0.865 |
| SwinFusion | 6.921 | 43.187 | 43.637 | 0.673 | 0.495 | 0.873 |
| IPLF | 6.696 | 45.733 | **79.683** | 0.860 | 0.489 | 0.838 |
| Ours | **7.447** | **51.327** | 58.908 | **0.894** | 0.501 | **0.895** |

edge details in the bushes next to the deck, as illustrated in the green marked area. Thanks for strong ability to preserve edge details at coarse scales in our method, distinct edge hierarchy of the bushes are well reflected in our fused image. Our method also retains the target to be salient well. In *Solider Behind Smoke* image, GANMcC, SDNet, and IPLF lose most smoke information from source visible image, though they keep target to be salient well. CDDFuse, SwinFusion and our method transfer sufficient information from both modalities. However, the target is a little hard to identified by human perception in the fused image of CDDFuse and SwinFusion. In *Bunker* image, how to recover the complex and abundant edge levels of bushes in the visible image is a challenge for fusion methods. The fused image of GANMcC, SDNet, U2Fusion and IPLF are heavily contaminated by the infrared image. By comparison, our method not only integrates useful information from sour images, but also reflects the hierarchy of edge information and the brightness well.

### 2) QUANTITATIVE EVALUATION

We evaluated the performance of different fusion methods in *TNO* dataset by six objective metrics, as shown in Table 2. Bold and underlined value are the best and second-best results in each metric respectively. Our method is superior to all other methods in terms of SD, VIF, $Q_{CB}$ and MS-SSIM. It indicates that our method can not only high contrast and good visual effects, but also preserves the representative structures of both modalities. Our method also ranks the second in EN, lagging behind UNFusion with a narrow margin. It implies that our method can transfer useful and complementary information to the fused image effectively. It is noted that the EI value

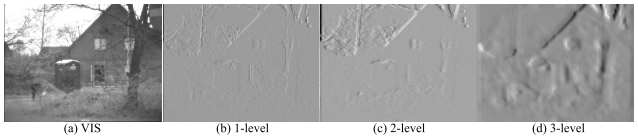


**FIGURE 9.** Hierarchical detail layers from shallow to deep. (a): original visible image; (b): detail features after 1-level wavelet decomposition; (c): detail features after 2-level wavelet decomposition; (d): detail features after 3-level wavelet decomposition. (Different scales are interpolated to same size for visualization).

of IPLF is too high, corresponding with the over-sharpened visual result in the qualitative evaluation. By contrast, our method retains abundant edge information with better visual perception.

### C. EXPERIMENTS ON ROADSCENE DATASET

#### 1) QUALITATIVE EVALUATION

Fig. 8 shows two examples of different fusion methods in *RoadScene* dataset. In "FLIR_06570" image pair, only CDDFuse, SwinFusion and our method preserve the sky region to be bright. The fused images of GANMcC, SDNet and UNFusion are bias to infrared modality, losing much detail information from the visible image. Salient targets in the fused image are not highlighted in U2Fusion. Compared with SwinFusion, our method retains high image contrast, presenting better visual results. In "FLIR_08859" image pair, UNFusion, SwinFusion and our method recover the details of vehicles well, as shown in green boxes. Similar to TNO dataset, the fused image based on GANMcC is blurred again. Fusion method based on IPLF carries excessive brightness information, which does not conform to human eye perception. By contrast, our method not only integrates useful complementary information from different modalities into the fused image, but is also more suitable for human observation.

#### 2) QUANTITATIVE EVALUATION

Table 3 shows the objective performance of different fusion methods in RoadScene dataset. The proposed method achieves the optimal value in EN, SD, VIF and MS-SSIM and lags the optimal of $Q_{CB}$ by a narrow margin, manifesting that our method generates the fused image with high dependence without introducing of noise and artifacts, consistent with human visual perception. Although the proposed method ranks third in terms of EI, it is still within acceptable range. Quantitative evaluation in two datasets validates the superior performance and strong robustness of our method.

### D. ABLATION EXPERIMENTS

#### 1) EFFECT OF MULTI-LEVEL WAVELET DECOMPOSITION

In this section, we discuss the impact of decomposition levels on the entire network. Specifically, 1-level means that only detail features after one-level decomposition in base stream are interacted with detail stream. 2-level means details after one and two-level decomposition are interacted with detail stream and so on.

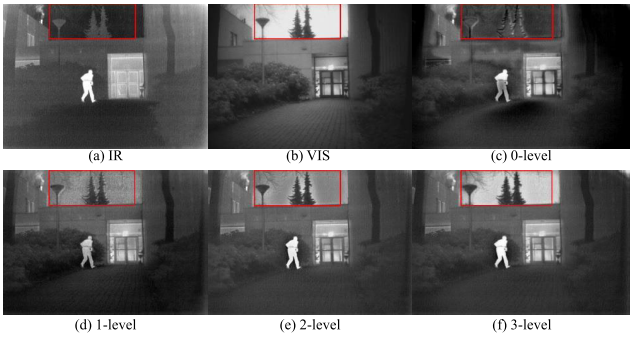


**FIGURE 10.** Different level decomposition on "*Man*" image.

**TABLE 4.** Ablation study on 47 image pairs from *TNO* dataset.

| | EN | SD | EI | VIF | $Q_{CB}$ | MS-SSIM |
|---|---|---|---|---|---|---|
| 0-level | 6.714 | 40.718 | 40.909 | 0.915 | 0.491 | 0.848 |
| 1-level | 6.949 | 41.307 | 41.487 | 0.965 | 0.493 | 0.876 |
| 2-level | 6.920 | 42.583 | **42.799** | *0.988* | 0.495 | **0.921** |
| 3-level | **6.976** | **44.845** | 41.913 | **1.019** | **0.512** | 0.907 |

A typical group of hierarchical coarse-to-fine detail layers is shown in Fig. 9 for intuitive visualization. From left to right are input image and detail features from shallow to deep, which are derived by multi-level wavelet decomposition in base stream. The granularity of image features is larger with the increment of decomposition level and network depth. In (d), features reflect more rough details of main objects, such as the contour of house, the shape of windows and the trunk. In (b), shallow features reflect more details on common information in the image, such as leaves and grass. The detail features from shallow to deep with multiple levels form a complete structural representation for the image.

We present the qualitative and quantitative fusion results of different decomposition levels in Fig. 10 and Table 4 respectively. In (c), it is clearly shown that the quality of fused image is close to that of the infrared image and the target is not salient enough if we do not apply wavelet decomposition completely. With the addition of decomposition levels, the textures in the fused image are more abundant and the target becomes more salient. The sky also gets brighter progressively, which proves each level details contributes to the fusion network. While there exists noise in the sky and artifacts around the tree regions in 1-level and 2-level decomposition, these artifacts are removed with the incorporation of more details at coarse scales. Most objective indicators exhibit upward trend as the decomposition levels increase, which indicates that details at coarse scales and deep network layers are vital to image contrast, entropy and visual fidelity.

#### 2) EFFECT OF SFMD FUSION STRATEGY

The impact of different feature fusion strategies on the fused image is discussed in this section. In Fig. 11, only

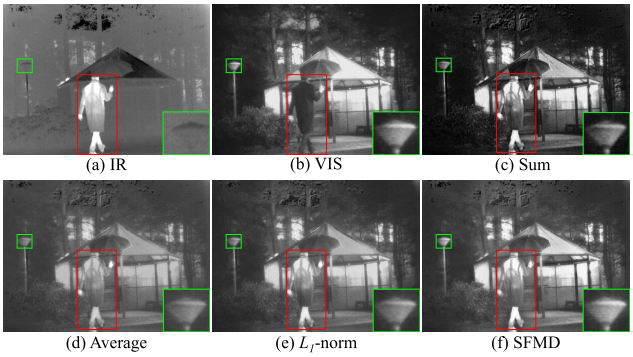


**FIGURE 11.** Different fusion strategies on "*Kaptein_1654*" image.

**TABLE 5.** Ablation study of fusion strategies on 47 image pairs from *TNO* dataset.

| | EN | SD | EI | VIF | $Q_{CB}$ | MS-SSIM |
|---|---|---|---|---|---|---|
| Sum | 6.728 | 40.718 | 34.910 | 0.808 | 0.460 | 0.887 |
| Average | 6.971 | 40.465 | 38.832 | 0.791 | 0.482 | **0.922** |
| $L_1$-norm | **6.995** | 43.782 | 39.559 | 0.805 | 0.496 | 0.898 |
| SFMD | 6.976 | **44.845** | **41.913** | **1.019** | **0.512** | 0.907 |

$L_1$-norm and our SFMD fusion strategies preserve the texture of street lamps well, which is reflected in the green boxes. Compared to $L_1$-norm, SFMD achieves more appropriate brightness information in the local areas and maintain high contrast in the fused image, benefiting from the local patch decomposition in SFMD. The thermal information is also highlighted well in the fused image. On another hand, SFMD outperforms other fusion strategies in four objective metrics, as illustrated in Table 5. SFMD also achieves the second rank in two other metrics. Combined with the subjective analysis in Fig. 11, it shows that SFMD is more suitable for feature fusion because it utilizes the inherent statistical relationship in each feature patch well.

## V. EXTENSION TO MEDICAL IMAGE FUSION

In this section, we apply our method to the medical image fusion to further demonstrate the effectiveness of our method. Medical images of different modalities provide information from different aspects [60]. Specifically, Magnetic Resonance Imaging (MRI) is structural imaging modality which shows soft tissues with high contrast, as illustrated in Fig. 12 (b). While Positron Emission computed Tomography (PET) is functional imaging modality that shows blood flow and metabolic activities occurring inside the body [61], which is as illustrated in Fig. 12 (a). Therefore, it is meaningful to merge structural data with functional data, which preserves salient functional information and anatomical information for a better diagnosis of the disease. We select 263 pairs of registered PET and MRI images from the Harvard dataset, set aside 31 pairs for testing.

### A. QUALITATIVE EVALUATION

We compare our method with GANMcC, SDNet, U2Fusion, CDDFuse and SwinFusion 5 state-of-the-art

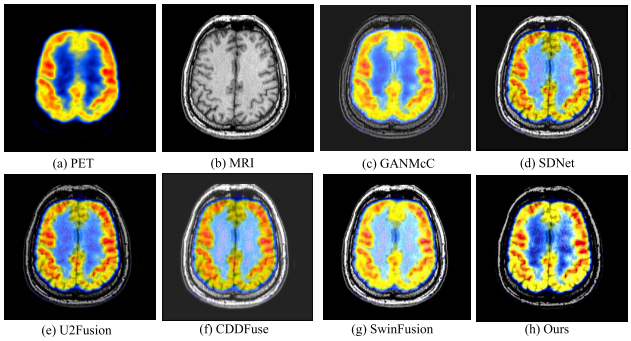


**FIGURE 12.** Visual comparison of our proposed method with 5 state-of-the-art fusion methods on the Harvard medical image fusion dataset (https://www.med.harvard.edu/aanlib/home.html).

**TABLE 6.** The average values of six metrics of 31 source image pairs from Harvard medical image fusion dataset with different methods.

| Methods | EN | SD | EI | VIF | $Q_{CB}$ | MS-SSIM |
|---|---|---|---|---|---|---|
| GANMcC | 5.298 | 45.031 | 42.671 | 0.557 | 0.465 | 0.705 |
| SDNet | 5.563 | 40.315 | 51.186 | 0.798 | **0.584** | 0.749 |
| U2Fusion | 4.939 | 47.140 | 38.118 | 0.644 | 0.418 | **0.812** |
| CDDFuse | 5.677 | 43.448 | **59.440** | 0.795 | 0.548 | 0.782 |
| SwinFusion | 5.465 | **48.192** | 56.972 | 0.783 | 0.566 | 0.804 |
| Ours | **5.770** | 47.512 | 49.233 | **0.801** | 0.516 | 0.794 |

fusion methods. The visual results are shown in Fig. 12. Intuitively, GANMcC and U2Fusion fail to preserve the brightness of anatomical structures in MRI images. SDNet and CDDFuse produce the fused image with low contrast. The details of the fused image are blurred in SwinFusion. By contrast, the fused image from our method not only attains high contrast, but also highlights the key details from the MRI source image.

### B. QUANTITATIVE EVALUATION

The objective performance evaluation is reported on the Harvard dataset is reported in Table 6. Clearly, our method ranks the first in terms of EN and VIF, indicating the fused image of our method contains abundant information and retains good visual quality. The suboptimal valve of SD is achieved by our method, implying our method can keep high contrast in the fused image. Other metrics of our method are still comparable with the optimal methods. In conclusion, our method can also achieves competitive performance in the medical image fusion.

## VI. DISCUSSION OF LIMITATIONS

The limitation of our method is the poor fusion performance in low-light or nighttime environments. We provide a typical example to illustrate this intuitively, as shown in Fig. 13. The reason is that our method utilizes multi-level wavelet transform to extract hierarchical details from source images and it fails to work when abundant details in the visible image are obscured in the darkness. Consequently, the fusion quality is unsatisfactory. A possible solution is that we can firstly enhance the low-light visible images by


(a) VIS (b) IR (c) Ours

**FIGURE 13.** Failure case. The fused image has weak texture details and poor visual perception due to the illumination degradation.

image enhancement algorithms and then fuse it with the infrared image. But the problem of color distortion may be introduced in the stage of enhancement. Another solution is that we can design an illuminance adjustment module to strip the illumination degradation in nighttime visible images while preserving informative features of source images in the future.

Another limitation is that we use Harr-wavelet transform tool only for computational efficiency in our method. However, the disadvantage of Harr-wavelet is that it lacks shift-invariance. Other wavelet transform tools like curvelet and contourlet can also be considered. In the future, we will explore how different wavelet transform tools influence the fusion performance of the network and provide some insights on how to select the most appropriate wavelet transform tool for decomposition according to the network architecture.

## VII. CONCLUSION

In this paper, we propose a novel two-stream edge-aware fusion network with multi-level wavelet decomposition. The network consists of base stream and detail stream, which deal with different levels' information of source images. The detail features at coarser scales from base stream are extracted and incorporated into detail stream by multi-level wavelet decomposition. More latent structural edge information can be processed in detail-stream. In the testing phase, the extracted features from the infrared and visible image pair are fused by the proposed Structural Feature Map Decomposition (SFMD) feature fusion strategy, outperforming than common fusion strategies. Ablation experiments demonstrate that the proposed network model and fusion strategy are both effective. The comparative experiments also show that our proposed model outperforms other SOTA methods. Furthermore, the proposed model can be applied to a series of other tasks like medical image fusion. The combination of multi-scale decomposition in wavelet domain with neural network can also inspire other researchers.

## DECLARATIONS

**Conflict of interest** The authors have no relevant financial or non-financial interests to disclose.

**Ethical Approval** This article does not contain any studies with human participants or animals performed by any of the authors. The medical image dataset can be publicly downloaded at https://www.med.harvard.edu/aanlib/home.html.

## REFERENCES

[1] S. Karim, G. Tong, J. Li, A. Qadir, U. Farooq, and Y. Yu, "Current advances and future perspectives of image fusion: A comprehensive review," *Inf. Fusion*, vol. 90, pp. 185–217, Feb. 2023.
[2] X. Zhang and Y. Demiris, "Visible and infrared image fusion using deep learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 10535–10554, Aug. 2023.
[3] Z. Pan and W. Ouyang, "An efficient network model for visible and infrared image fusion," *IEEE Access*, vol. 11, pp. 86413–86430, 2023.
[4] L. Tang, H. Zhang, H. Xu, and J. Ma, "Rethinking the necessity of image fusion in high-level vision tasks: A practical infrared and visible image fusion network based on progressive semantic injection and scene fidelity," *Inf. Fusion*, vol. 99, Nov. 2023, Art. no. 101870.
[5] Y. Lin, P. Chiang, and S. Miaou, "Enhancing deep-learning object detection performance based on fusion of infrared and visible images in advanced driver assistance systems," *IEEE Access*, vol. 10, pp. 105214–105231, 2022.
[6] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, Mar. 2021.
[7] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.
[8] X. Zhang, P. Ye, H. Leung, K. Gong, and G. Xiao, "Object fusion tracking based on visible and infrared images: A comprehensive review," *Inf. Fusion*, vol. 63, pp. 166–187, Nov. 2020.
[9] P. Chai, X. Luo, and Z. Zhang, "Image fusion using quaternion wavelet transform and multiple features," *IEEE Access*, vol. 5, pp. 6724–6734, 2017.
[10] Z. Ren, G. Ren, and D. Wu, "Fusion of infrared and visible images based on discrete cosine wavelet transform and high pass filter," *Soft Comput.*, vol. 27, no. 18, pp. 13583–13594, Sep. 2023.
[11] R. Singh, R. Srivastava, O. Prakash, and A. Khare, "Multimodal medical image fusion in dual tree complex wavelet transform domain using maximum and average fusion rules," *J. Med. Imag. Health Informat.*, vol. 2, no. 2, pp. 168–173, Jun. 2012.
[12] H. Li, X. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.
[13] 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.
[14] C. Gao, C. Song, Y. Zhang, D. Qi, and Y. Yu, "Improving the performance of infrared and visible image fusion based on latent low-rank representation nested with rolling guided image filtering," *IEEE Access*, vol. 9, pp. 91462–91475, 2021.
[15] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.
[16] Q. Li, G. Han, P. Liu, H. Yang, J. Wu, and D. Liu, "An infrared and visible image fusion method guided by saliency and gradient information," *IEEE Access*, vol. 9, pp. 108942–108958, 2021.
[17] 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.
[18] D. P. Bavirisetti and R. Dhuli, "Two-scale image fusion of visible and infrared images using saliency detection," *Infrared Phys. Technol.*, vol. 76, pp. 52–64, May 2016.
[19] Z. Wang, J. Wang, Y. Wu, J. Xu, and X. Zhang, "UNFusion: A unified multi-scale densely connected network for infrared and visible image fusion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3360–3374, Jun. 2022.
[20] D. Zhu, W. Zhan, Y. Jiang, X. Xu, and R. Guo, "IPLF: A novel image pair learning fusion network for infrared and visible image," *IEEE Sensors J.*, vol. 22, no. 9, pp. 8808–8817, May 2022.
[21] J. Ma, L. Tang, F. Fan, J. Huang, X. Mei, and Y. Ma, "SwinFusion: Cross-domain long-range learning for general image fusion via Swin Transformer," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 7, pp. 1200–1217, Jul. 2022.
[22] W. Zhang et al., "Underwater image enhancement via weighted wavelet visual perception fusion," *IEEE Trans. Circuits Syst. Video Technol.*, early access, doi: 10.1109/TCSVT.2023.3299314.
[23] P. Liu, H. Zhang, K. Zhang, L. Lin, and W. Zuo, "Multi-level wavelet-CNN for image restoration," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 773–782.

[24] 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.

[25] J. Ma, H. Zhang, Z. Shao, P. Liang, and H. Xu, "GANMcC: A generative adversarial network with multiclassification constraints for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5005014.

[26] J. Li, H. Huo, C. Li, R. Wang, and Q. Feng, "AttentionFGAN: Infrared and visible image fusion using attention-based generative adversarial networks," *IEEE Trans. Multimedia*, vol. 23, pp. 1383–1396, 2021.

[27] 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 2018.

[28] K. Ma, H. Li, H. Yong, Z. Wang, D. Meng, and L. Zhang, "Robust multi-exposure image fusion: A structural patch decomposition approach," *IEEE Trans. Image Process.*, vol. 26, no. 5, pp. 2519–2532, May 2017.

[29] J. Sebastian and G. R. G. King, "Comparative analysis and fusion of MRI and PET images based on wavelets for clinical diagnosis," *Int. J. Electron. Telecommun.*, vol. 68, pp. 867–873, Nov. 2022.

[30] J. Zhou, D. Zhang, P. Zou, W. Zhang, and W. Zhang, "Retinex-based Laplacian pyramid method for image defogging," *IEEE Access*, vol. 7, pp. 122459–122472, 2019.

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

[32] 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.

[33] A. L. Da Cunha, J. Zhou, and M. N. Do, "The nonsubsampled contourlet transform: Theory, design, and applications," *IEEE Trans. Image Process.*, vol. 15, pp. 3089–3101, 2006.

[34] H. Wei, Z. Zhu, L. Chang, M. Zheng, S. Chen, P. Li, G. Qi, and Y. Li, "A novel precise decomposition method for infrared and visible image fusion," in *Proc. Chin. Control Conf. (CCC)*, Jul. 2019, pp. 3341–3345.

[35] Y. Liu, L. Wang, J. Cheng, C. Li, and X. Chen, "Multi-focus image fusion: A survey of the state of the art," *Inf. Fusion*, vol. 64, pp. 71–91, Dec. 2020.

[36] A. O. Salau, S. Jain, and J. N. Eneh, "A review of various image fusion types and transform," *Indonesian J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, pp. 1515–1522, 2021.

[37] M. Jian, J. Dong, and Y. Zhang, "Image fusion based on wavelet transform," in *Proc. 8th ACIS Int. Conf. Software Eng., Artif. Intell., Netw., Parallel Distrib. Comput.*, vol. 1, Jul. 2007, pp. 713–718.

[38] J. Ma and G. Plonka, "The curvelet transform," *IEEE Signal Process. Mag.*, vol. 27, no. 2, pp. 118–133, Mar. 2010.

[39] 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.

[40] M. N. Do and M. Vetterli, "The contourlet transform: An efficient directional multiresolution image representation," *IEEE Trans. Image Process.*, vol. 14, pp. 2091–2106, 2005.

[41] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.

[42] Z. Zhao, H. Bai, J. Zhang, Y. Zhang, S. Xu, Z. Lin, R. Timofte, and L. van Gool, "CDDFuse: Correlation-driven dual-branch feature decomposition for multi-modality image fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5906–5916.

[43] H. Xu, M. Gong, X. Tian, J. Huang, and J. Ma, "CUFD: An encoder–decoder network for visible and infrared image fusion based on common and unique feature decomposition," *Comput. Vis. Image Understand.*, vol. 218, Apr. 2022, Art. no. 103407.

[44] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[45] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2017, pp. 4700–4708.

[46] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[47] H. Li, T. N. Chan, X. Qi, and W. Xie, "Detail-preserving multi-exposure fusion with edge-preserving structural patch decomposition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 11, pp. 4293–4304, Nov. 2021.

[48] T. Chan, S. Esedoglu, F. Park, and A. Yip, "Recent developments in total variation image restoration," *Math. Models Comput. Vis.*, vol. 17, pp. 17–31, Dec. 2005.

[49] H. Xu, J. Ma, Z. Le, J. Jiang, and X. Guo, "FusionDN: A unified densely connected network for image fusion," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, pp. 12484–12491.

[50] M. Brown and S. Süsstrunk, "Multi-spectral SIFT for scene category recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 177–184.

[51] A. Toet and M. A. Hogervorst, "Progress in color night vision," *Opt. Eng.*, vol. 51, no. 1, Feb. 2012, Art. no. 010901.

[52] H. Zhang and J. Ma, "SDNet: A versatile squeeze-and-decomposition network for real-time image fusion," *Int. J. Comput. Vis.*, vol. 129, no. 10, pp. 2761–2785, Oct. 2021.

[53] H. Xu, J. Ma, J. Jiang, X. Guo, and H. Ling, "U2Fusion: A unified unsupervised image fusion network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 502–518, Jan. 2022.

[54] J. van Aardt, "Assessment of image fusion procedures using entropy, image quality, and multispectral classification," *J. Appl. Remote Sens.*, vol. 2, no. 1, May 2008, Art. no. 023522.

[55] Y.-J. Rao, "In-fibre Bragg grating sensors," *Meas. Sci. Technol.*, vol. 8, no. 4, p. 355, 1997.

[56] B. Rajalingam, R. Priya, and R. Scholar, "Review of multimodality medical image fusion using combined transform techniques for clinical application," *Int. J. Sci. Res. Comput. Sci. Appl. Manag. Stud.*, vol. 7, no. 3, pp. 1–8, May 2018.

[57] 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.

[58] Y. Chen and R. S. Blum, "A new automated quality assessment algorithm for image fusion," *Image Vis. Comput.*, vol. 27, no. 10, pp. 1421–1432, Sep. 2009.

[59] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2003, pp. 1398–1402.

[60] J. Sebastian and G. R. G. King, "Fusion of multimodality medical images—A review," in *Proc. Smart Technol., Commun. Robot. (STCR)*, Oct. 2021, pp. 1–6.

[61] J. Sebastian and G. R. G. King, "A novel MRI and PET image fusion in the NSST domain using YUV color space based on convolutional neural networks," *Wireless Pers. Commun.*, vol. 131, no. 3, pp. 2295–2309, Aug. 2023.

**HAOZHE WANG** received the B.S. degree in communication engineering from Beijing Jiaotong University, Beijing, China, in 2020, and the M.S. degree in information and communication engineering from the University of Electronic Science and Technology of China, Sichuan, China, in 2023. His research interests include image fusion, image deblurring, and low-light image enhancement.

**CHANG SHU** received the B.S. degree in communication engineering from the University of Electronic Science and Technology of China, Sichuan, China, in 2004, and the Ph.D. degree in electronic information from Tsinghua University, Beijing, China, in 2011. From 2017 to 2018, he was Visiting Scholar with the University of Victoria. He is currently a Lecturer with the University of Electronic Science and Technology of China. His research interests include 3D reconstruction and image processing.

**XIAOFENG LI** received the B.S. degree in information theory from Xidian University, Xian, Shaanxi, in 1984, and the M.S. and Ph.D. degrees in signal and information processing from the University of Electronic Science and Technology of China, Sichuan, China, in 1987 and 2011, respectively. He was an Adjunct Professor of math with San Diego State University, USA, from 1995 to 1996. He is currently a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His research interests include multispectral imaging, image fusion, image deblurring, and image registration.

**YU FU** received the B.S. degree from the North China Institute of Science and Technology, China, in 2019, and the M.S. degree from Jiangnan University, China, in 2022. His research interests include image fusion, machine learning, and deep learning.

**ZHIZHONG FU** received the B.S., M.S., and Ph.D. degrees from the University of Electronic Science and Technology of China, Sichuan, China, in 1993, 1998, and 2002, respectively. He was a Visiting Researcher with the University of California, San Diego, from 2011 to 2012. He is currently a Professor with the School of Information and Communication Engineering, University of Electronic Science and Technology of China. His research interests include bit depth enhancement, image super-resolution, image fusion, and infrared sensor imaging.

**XIAOFENG YIN** received the B.S. degree in aircraft design and engineering from the Nanjing University of Aeronautics and Astronautics, Jiangsu, China, in 2019. He is currently pursuing the M.S. degree in professional accounting with the Jiangsu University of Science and Technology, Jiangsu, China. He is with China International Marine Containers (Group) Company Ltd. His research interests include intelligent manufacturing equipment, mechanical manufacturing automation, and management automation.