Multimedia Tools and Applications (2024) 83:68931–68957
https://doi.org/10.1007/s11042-024-18141-y

A multi-weight fusion framework for infrared and visible image fusion

Yiqiao Zhou[1] · Kangjian He[1] · Dan Xu[1] · Hongzhen Shi[1] · Hao Zhang[1]

Received: 6 May 2023 / Revised: 17 October 2023 / Accepted: 3 January 2024 /
Published online: 30 January 2024

Abstract

Infrared and visible image fusion (IVF) aims to generate a fused image with important thermal target and texture information from infrared and visible images. However, the existing advanced fusion methods have the problem of insufficient extraction of visible image details, and the fused image is not natural and does not conform to human visual perception. To solve this problem, we propose an effective infrared and visible image fusion framework inspired by the idea of multi-exposure fusion. First, we design an adaptive visible light exposure adjustment module to enhance the low-brightness pixel area information in the visible image to obtain an adaptive exposure image. Secondly, three feature weight maps of the input infrared, visible light and adaptive exposure images are extracted through the multi-weight feature extraction module: DSIFT map, saliency map and saturation map, and then the feature weight maps are optimized through the Mutually Guided Image Filtering (MuGIF). Then, we use the Gaussian and Laplacian pyramids to decompose and reconstruct the feature weight map and input image to obtain the pre-fused image. Finally, to further enhance the contrast of the pre-fused image, we use a Fast Guided Filter to enhance the pre-fused image to obtain the final fusion result. Qualitative and quantitative experiments show that the proposed method exhibits better fusion performance on public datasets compared with 11 state-of-the-art methods. In addition, this method can retain more visible image details, and the fusion result is more natural. Our code is publicly available at https://github.com/VCMHE/MWF_VIF.

Keywords Image fusion · Multi-weight feature extraction · Adaptive exposure adjustment · Human visual perception

1 Introduction

Image fusion aims to extract valuable features from source images of different modalities and integrate them into a unified image [1]. One common application of this technique is IVF, which provides more comprehensive scene information for a variety of downstream

✉ Kangjian He
hekj@ynu.edu.cn

[1] School of Information Science and Engineering, Yunnan University, Kunming 650091, China

tasks [2, 3]. Infrared images can distinguish objects well by capturing the thermal radiation information emitted by objects during the day or at night. However, it has the disadvantages of poor texture and low spatial resolution. In contrast, visible images offer rich texture details but are susceptible to environmental factors like low illumination, making it challenging to identify targets in the scene. Therefore, considering their complementary nature, IVF can yield a more comprehensive scene understanding, serving advanced applications such as object detection and recognition [4–6].

Existing methods for IVF can be broadly categorized into traditional fusion methods [7] and deep learning-based fusion methods [8]. Traditional image fusion algorithms transform the source image into the transformation domain and retain the characteristics of the source image by manually designing feature extraction and fusion rules to achieve image fusion. According to their corresponding theories, traditional fusion methods can be classified into those based on multiscale transformation, representation learning, subspace-based approaches, hybrid models, and other methods [9]. Among them, methods based on multiscale transformation [10] and representation learning [11] are more common.

Classic algorithms based on multi-scale transformations include pyramid transform, wavelet transform, edge-preserving filtering, and non-subsampling contourlet transform. Inspired by biological vision, Zhou et al. [12] proposed a new method for infrared and visible image fusion. This method converts image intensity into the visual response space of the human visual system (HVS), achieving a better fusion effect. At the same time, some new algorithms have also been proposed, such as Li et al. [10] and MDLatLRR [13]. However, these decomposition methods project the source image into the frequency domain, increasing computational complexity.

Representation learning-based methods have been widely used in image fusion. Algorithms based on representation learning include sparse representation (SR) [14] and latent low-rank representation (LatLRR) [15]. Liu et al. [16] proposed an adaptive sparse representation (ASR) model. In addition, some new algorithms have also been proposed, such as IVFusion [17] and Zhang et al. [18]. Although SR-based fusion methods directly extract features in the spatial domain, reducing the loss of image information and thus achieving good fusion performance, these methods are complex and time-consuming.

In recent years, with the continuous development of artificial neural networks, image fusion based on deep learning has attracted much attention. Auto-encoder (AE) based fusion methods usually consist of an encoder, a decoder, and a fusion strategy. Ren et al. proposed a variational autoencoder method [19] to compensate for the loss of infrared image information during the fusion process. Qu et al. combined the converter module with the CNN network and proposed a fusion framework known as TransMEF [20]. In addition, AE-based fusion methods include [21, 22]. However, AE-based methods have low fusion performance due to the manual design of fusion rules.

Convolutional neural network (CNN)-based fusion methods can avoid the limitations of hand-crafted fusion rules. Xu et al. [23] proposed a fusion network capable of handling multiple image fusion tasks within a unified framework, known as U2Fusion. Ma et al. define the information required for IVF by introducing a salient object mask [24]. However, due to the absence of basic facts, the potential performance of the converged network is only partially utilized in some cases.

Infrared and visible image fusion can be directly generated using the generative adversarial network (GAN) method. Ma et al. [25] introduced FusionGAN, a GAN-based fusion technique that showed remarkable performance in combining infrared and visible images. DDcGAN [26] utilizes two discriminators to guide the generator in fusing information from the original images. However, these deep learning-based methods still face challenges. The scarcity of adequate benchmark datasets for IVF tasks hinders the training of deep learning networks. This limitation complicates the application of deep learning methods in this domain.

Interestingly, both traditional and deep learning methods often overlook the challenge of sufficiently extracting detailed information from visible images. Moreover, most methods fail to ensure that the fusion image is both natural and consistent with human visual perception. To address this issue, this paper introduces an effective framework for the fusion of infrared and visible images. Initially, an adaptive visible light exposure adjustment module is devised to enhance the extraction of fine details from visible images. Subsequently, a multi-weighted feature extraction module is proposed to achieve a more natural fusion effect by combining the features extracted from infrared images, visible images, and adaptive exposure images. This approach outperforms many existing methods, ensuring superior results in image fusion. Figure 1 shows the advantages of the proposed fusion method on the Road example: the most left is an infrared image, the second one is a visible image, and the last three ones are the results of the GTF method [27], SDNet method [28], and the proposed method respectively. By comparison, the fusion results of our approach can simultaneously preserve the rich texture information of the visible image and the critical thermal target information of the infrared image (as shown by the yellow and green arrows). In addition, our fusion results are more natural.

The main contributions of this work can be summarized as follows:

(1) In IVF, visible images often suffer from unclear image details and textures in low-exposure scenes, similar to the challenges encountered in multi-exposure image fusion. To address this issue, we proposed an adaptive exposure module specifically designed for visible images. This module enhances the brightness of low-luminance pixel areas in visible images, enriching the details of visible images.

(2) We have proposed a multi-weight feature extraction module. Traditional methods typically use multi-scale techniques to decompose infrared and visible images into base layers and detail layers, and then apply different weight fusion strategies for each layer. In contrast to conventional methods, our approach is unique. Through our multi-weight feature extraction module, we directly extract weight maps from input infrared, visible, and adaptive exposure images. This novel approach enables us to capture more intricate information from the source images through multi-weighted feature extraction operations.

(3) The proposed multi-weight fusion framework is applicable to both grayscale and color visible and infrared image fusion tasks. We conduct qualitative and quantitative comparisons of 11 state-of-the-art fusion methods on two public datasets. The fusion results show that our results have richer texture details and prominent infrared targets, more consistent with human visual perception.

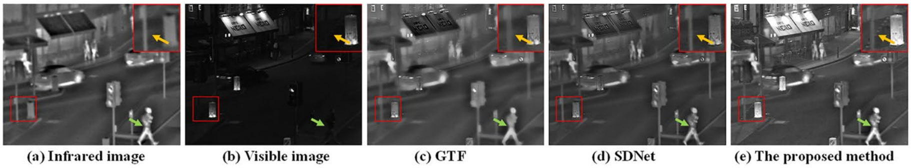
(a) Infrared image (b) Visible image (c) GTF (d) SDNet (e) The proposed method

Fig. 1 An example of infrared and visible image fusion on the scene of Road from TNO dataset

The structure of this paper is as follows: Section 2 provides the necessary preliminaries. In Section 3, the proposed method is elaborated in detail. Section 4 presents comprehensive experimental results and discussions based on the public dataset. Finally, we conclude our work in Section 5.

2 Preliminaries

2.1 Camera response model

The camera response model includes the Camera Response Function (CRF) model and the Brightness Transform Function (BTF) model. The mathematical description of the BTF model is as follows:

$$\mathbf{P}_1 = g\left(\mathbf{P}_0, k\right) = \beta \mathbf{P}_0^{\gamma}, \tag{1}$$

where $\mathbf{P}_0$ and $\mathbf{P}_1$ represent images captured under the same scene with different exposure levels, and k denotes the exposure ratio. β and γ are parameters related to the exposure ratio, g stands for BTF. In the BTF model, β and γ are determined by camera parameters and exposure ratio k. To establish their relationship, the corresponding CRF model is proposed as follows [29]:

$$f(kE) = \beta f(E)^{\gamma}, \tag{2}$$

where E is the irradiance of the picture, the closed-form solution for f is presented below:

$$f(E) = \begin{cases} e^{b(1-E^a)}, & \text{if } \gamma \neq 1 \\ E^c, & \text{if } \gamma = 1 \end{cases} \tag{3}$$

where a and b are model parameters in the given scenario of $\gamma \neq 1$

$$a = \log_k \gamma, \ b = \frac{\ln \beta}{1 - \gamma}; \tag{4}$$

And c is a model parameter in the given scenario of $\gamma = 1$

$$c = \log_k \beta, \tag{5}$$

since the BTF is non-linear for most cameras, the BTF g is defined as follows when $\gamma \neq 1$.

$$g(\boldsymbol{P}, k) = e^{b(1-k^a)} \boldsymbol{P}^{(k^a)}. \tag{6}$$

where $\mathbf{P}$ is the input exposure image.

2.2 Dense scale invariant feature transform (DSIFT) descriptor

The (scale invariant feature transform) SIFT descriptor introduced by Lowe [30] encapsulates vital information for measuring activity levels in image fusion tasks. This descriptor is formed by detailing local gradient features for identified points of interest. In pixel-level image fusion, activity-level measurements are needed for each pixel or possibly for each adjacent block. However, the SIFT descriptor cannot determine points of interest in sparse regions, preventing its direct application in pixel-level image fusion. To address this limitation, Liu et al. [31] proposed a solution: DSIFT descriptor.

DSIFT descriptor is an approach for dense feature extraction, computing feature descriptors for every pixel in an image without the need to identify specific points of interest. This method utilizes a technique similar to the SIFT descriptor [30] to calculate the pixel descriptor. Initially, the immediate surrounding of a pixel is divided into smaller cells. The gradient details of each cell are then determined using a directional histogram with a specified number of bins. Finally, the resulting descriptor is normalized to ensure consistency and reliability.

The key advantage of DSIFT in the context of image fusion lies in its ability to simultaneously address two crucial challenges: measuring the activity level of each source image and determining the local similarity among various source images. The former aspect is specifically designed to extract local contrast information from diverse source images, playing a pivotal role in scenarios involving multiple image fusion types such as multi-exposure, multi-focus, and multi-mode fusions. The latter aspect becomes significant when source images require precise registration. Moreover, compared to quality measures based solely on image gradient [32], the activity level measure derived from the DSIFT descriptor proves more reliable and robust against noise.

3 The proposed method

In this section, we introduce a novel multi-weight fusion framework for infrared and visible image fusion. The schematic representation of the proposed fusion algorithm is illustrated in Fig. 2, comprising an adaptive exposure adjustment module for visible images, a multi-weight feature extraction module, and a weight map optimization and image fusion module.

To begin, we devise an adaptive exposure adjustment module tailored to optimize the quality and enhance visible image details. Subsequently, the adaptive exposure image, visible image, and infrared image are fed into the multi-weight feature extraction module,

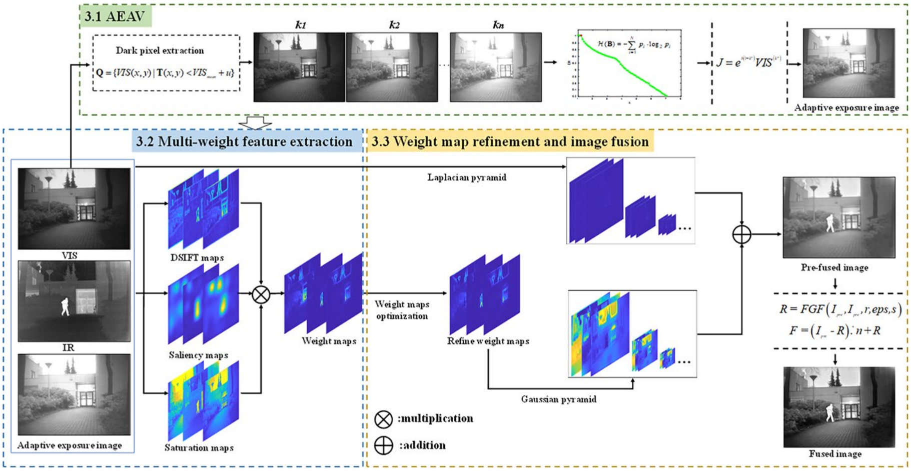
3.1 AEAV
Dark pixel extraction
$\mathbf{Q} = \{VIS(x,y) \mid \mathbf{T}(x,y) < VIS_{mean} + u\}$
k1
k2
kn
$H(\mathbf{B}) = -\sum_{i=1}^{N} p_i \cdot \log_2 p_i$
$J = e^{b(1-k^*)} VIS^{(k^*)}$
Adaptive exposure image
3.2 Multi-weight feature extraction
3.3 Weight map refinement and image fusion
VIS
IR
Adaptive exposure image
DSIFT maps
Saliency maps
Saturation maps
Weight maps
Laplacian pyramid
Weight maps optimization
Refine weight maps
Gaussian pyramid
Pre-fused image
$R = FGF(I_{pre}, I_{pre}, r, eps, s)$
$F = (I_{pre} - R) \cdot n + R$
Fused image
:multiplication
:addition

Fig. 2 Framework of the proposed method

generating feature weight maps for adaptive exposure adjustment, visible, and infrared images. To mitigate noise and artifacts in the fused image, we employ MuGIF [33] for denoising the feature weight maps. Gaussian and Laplacian pyramids are then utilized for decomposing and reconstructing the weight maps and input images, yielding pre-fused images. To further enhance the contrast of the pre-fused images, we leverage the Fast Guided Filter [34] to refine the pre-fused images, resulting in the final fusion outcome. The specific details of the proposed method are outlined as follows.

3.1 Adaptive exposure adjustment for visible image (AEAV)

Visible images are susceptible to environmental interference, such as low brightness, affecting the quality of subsequent fused images. As discussed in Section 2.1, we can use BTF to generate a series of visible images with different exposures in the same scene and then fuse these images to obtain an enhanced visible image. In this work, we adopt BTF to develop a new adaptive exposure adjustment module to obtain adaptive exposure images for visible images. The flow chart of the proposed method is shown in Fig. 3.

Initially, we extract the brightness component from the input visible image VIS. This extracted brightness component serves as our initial estimate of illumination $\mathbf{W}$.

$$\mathbf{W}(x, y) = \max_{c\in\{R,G,B\}} VIS_c(x, y), \tag{7}$$

where c represents the color channels. (x, y) represents the image pixel. Employing a structural texture decomposition method [35], we process the luminance component to derive the scene illumination map $\mathbf{T}$ of the image. This illumination map effectively preserves the meaningful structures within the image while eliminating textural details. As stated in reference [36], $\mathbf{T}$ is obtained by solving the following optimization equation:

$$\min_T \|\mathbf{T} - \mathbf{W}\|_2^2 + \lambda\|\mathbf{M}\circ\nabla\mathbf{T}\|_1, \tag{8}$$

where $\| * \|_2$ and $\| * \|_1$ are the ℓ_2 and ℓ_1 norm, respectively. ∇ is the first order derivative filter, which contains $\nabla_h\mathbf{T}$ (horizontal) and $\nabla_v\mathbf{T}$ (vertical). $\lambda\|\mathbf{M}\circ\nabla\mathbf{T}\|_1$ is used to

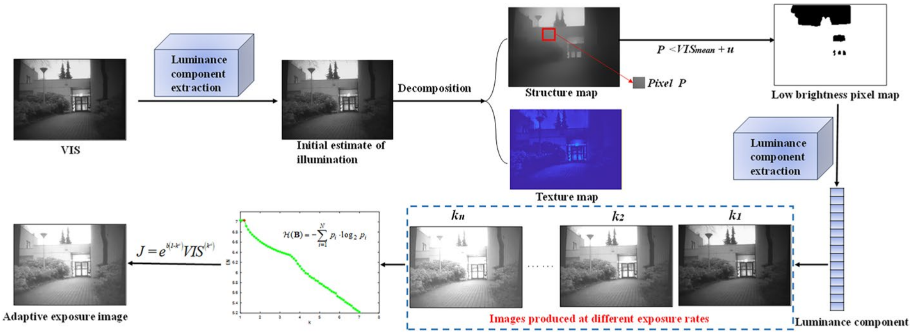
Luminance component extraction
VIS
Initial estimate of illumination
Decomposition
Structure map
Texture map
P <VISmean + u
Pixel P
Low brightness pixel map
Luminance component extraction
Luminance component
k1
k2
kn
Images produced at different exposure rates
$\mathcal{H}(\mathbf{B}) = -\sum_{i=1}^{N} p_i \cdot \log_2 p_i$
$J = e^{b(1-k^*)} VIS^{(k^*)}$
Adaptive exposure image

Fig. 3 Flowchart of adaptive exposure adjustment for visible image module

preserve the smoothness of $\mathbf{T}$, where $\mathbf{M}$ represents the weight matrix, and λ is the coefficient. Notably, the optimization of the scene lighting map $\mathbf{T}$ focuses on the design of $\mathbf{M}$. This is because the primary edges within the local window exhibit direction gradients that are more consistent than the texture. Consequently, windows containing edges should be relatively smaller. The design of the weight matrix is as follows.

$$\mathbf{M}_d(x,y) = \frac{1}{\left|\sum_{q\in\omega(x,y)} \nabla_d \mathbf{T}(q)\right| + \varphi}, d \in \{h, v\}, \tag{9}$$

here, $|*|$ represents the absolute value operator, d represents direction, including the horizontal direction h or the vertical direction v, and $\omega(x,y)$ is a local window centered at pixel (x,y). Additionally, φ is a very small constant used to avoid division by zero. For a comprehensive understanding of these operations, please refer to the detailed explanation in reference [29].

Based on the average pixel value VIS_{mean} of the input visible image VIS, the extracted low-luminance pixels are as follows:

$$\mathbf{Q} = \{VIS(x,y)|\mathbf{T}(x,y) < VIS_{mean} + u\}, \tag{10}$$

where $\mathbf{Q}$ represents pixels in the low-brightness pixel area, and u is a constant. As shown in Fig. 4, as the u value increases, the exposure of the low-brightness pixel area gradually decreases (shown in the red enlarged area). To increase the visibility of low-brightness pixel areas while maintaining well-exposed regions, we set the value of u to 0.2.

In order to minimize computational overhead, it is crucial to determine the optimal exposure rate k for the camera response model, ensuring adequate exposure in low-brightness pixel areas of the original visible image. To achieve this, it is necessary to identify the optimal exposure ratio, denoted as $\hat{k}$, while focusing solely on the brightness component. The brightness component $\mathbf{B}$ is defined as the geometric mean of the three channels (red, green, and blue), represented as:

$$\mathbf{B} := \sqrt[3]{\mathbf{Q}_r \circ \mathbf{Q}_g \circ \mathbf{Q}_b}, \tag{11}$$

where $\mathbf{Q}_r$, $\mathbf{Q}_g$ and $\mathbf{Q}_b$ are the red, green and blue channels of input image, respectively. High-exposure images generally offer superior visibility compared to low or overexposed images, thereby providing richer information. Therefore, the optimal value for k can be determined based on the entropy value of the image:

$$\mathcal{H}(\mathbf{B}) = -\sum_{i=1}^{n} p_i \cdot \log_2 p_i, \tag{12}$$

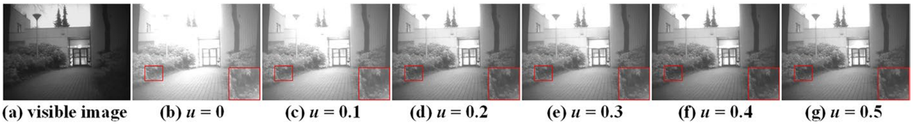
(a) visible image (b) $u = 0$ (c) $u = 0.1$ (d) $u = 0.2$ (e) $u = 0.3$ (f) $u = 0.4$ (g) $u = 0.5$

Fig. 4 The adaptive exposure adjustment results by using different u

where $\mathcal{H}(\cdot)$ represents the entropy value of the image, where p_i is the i-th bin of the histogram of $\mathbf{B}$ which counts the number of data valued in $\left[\frac{i}{m}, \frac{i+1}{m}\right)$, and m is the number of bins (usually set to 256). According to (12), the image entropy under different exposure ratios can be computed. The optimal exposure ratio $\widehat{k}$ is determined by maximizing the image entropy of the enhancement brightness as follows:

$$\widehat{k} = \underset{k}{\mathrm{argmax}}\ \mathcal{H}(g(\mathbf{B}, k)), \tag{13}$$

like [29], our method employs an exposure ratio k ranging from 1 to 7. Finally, based on the determined optimal exposure ratio $\widehat{k}$, the adaptive exposure image of VIS can be obtained as follows:

$$J = e^{b\left(1-\widehat{k}^a\right)} VIS^{\left(\widehat{k}^a\right)}. \tag{14}$$

To ensure compatibility with a wide range of cameras, fixed camera parameters ($a = -0.3293, b = 1.1258$) are used in our method.

Based on the above analysis, this method obtains the optimal exposure rate $\widehat{k}$ according to (13) and generates an adaptive exposure image based on $\widehat{k}$. Figure 5 displays the exposure images of select visible images captured under varying exposure ratios k and the optimal exposure rate $\widehat{k}$. These images demonstrate the method's effectiveness in exposing low-brightness pixel areas of visible images according to real-world conditions.

In Fig. 5, input image 1 was taken in a low-light environment, and most of the details in the image are not visible, so the value of the optimal exposure $\widehat{k}$ is larger. The input image 2 was taken in a normal light environment, and only a small part of the details in the image are not visible, so the value of the optimal exposure $\widehat{k}$ is small. It can be seen from Fig. 5 that both adaptive exposure image 1 and adaptive exposure image 2 restore the detailed information of the original visible image in the low-brightness pixel area very well.

The entire calculation process can be expressed by Algorithm 1.

```
Parameter descriptions:
$VIS$ is the visible image.
$J$ is the image after adaptive exposure adjustment.
Input: $VIS$  Output: $J$
The brightness component of the visible image is obtained by (7) as an initial estimate of illumination $\mathbf{W}$;
The scene light map $\mathbf{T}$ of $VIS$ is obtained via structural texture decomposition method and equations (8)-(9);
Obtain the pixel average value $VIS_{mean}$ of $VIS$;
According to $VIS_{mean}$, the pixel Q in the low-brightness pixel area is obtained via (10);
Equations (12)-(13) are used to find the optimal exposure ratio $\hat{k}$;
According to $\hat{k}$, $J$ is obtained via (14);
Return $J$
```

Algorithm 1 Main steps of the proposed AEAV module

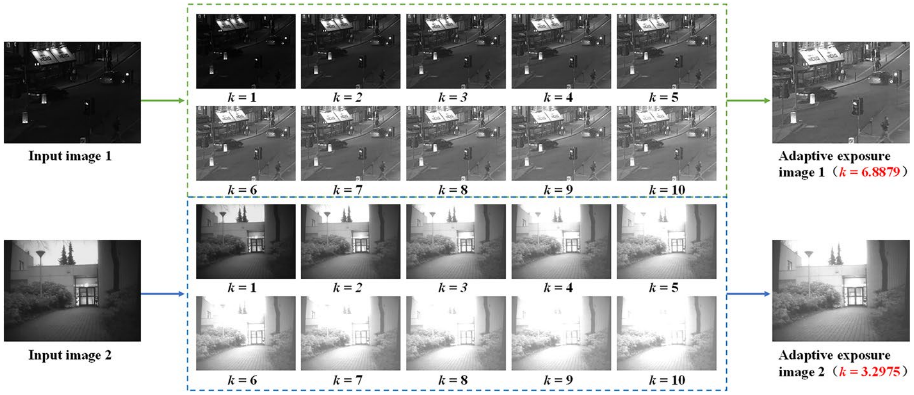

Fig. 5 Experimental result of adaptive exposure adjustment for visible image (AEAV) module

3.2 Multi-weight feature extraction module

The multi-weight feature extraction module is used to extract feature weights from the input visible, infrared, and adaptive exposure images. This module computes three metrics (DSIFT, saliency, and saturation features) from the input visible, infrared, and adaptive exposure images to estimate the feature weight map of the source image. By calculating the weight maps for each metric and multiplying all the weight maps together, the feature weight map is obtained. The following explains the process of getting corresponding weight maps through DSIFT, saliency, and saturation features.

3.2.1 Weight maps via DSIFT

As stated in Section 2.2, the SIFT descriptor contains crucial information for measuring activity levels in image fusion tasks. This is because it captures local gradient details of identified points of interest. However, SIFT descriptors alone cannot determine points of interest and sparse regions, making them unsuitable for direct pixel-level image fusion. To address this limitation, the DSIFT descriptor, which allows feature descriptors for each image pixel, is employed to extract spatial details from the input visible, infrared, and adaptive exposure images.

In this study, DSIFT scores are extracted and employed as weight maps, denoted as DSIFT maps. These maps are instrumental in preserving vital details, including texture and edges, in the images. Let I_i, where $i = 1, 2, 3, \ldots, n$, represent the input grayscale or color image sequence. The DSIFT feature can be measured as follows:

$$C_i(x,y) = \|\mathrm{DSIFT}\big(I_{\mathrm{g}}(x,y)\big)\|_1, \tag{15}$$

where DSIFT($*$) indicates the operator, which calculates the unnormalized DSIFT map of the source input image. Please refer to [30, 31] for more details about the calculation of DSIFT. For a specific pixel located at (x,y), $I_{\mathrm{g}}(x,y)$ and $C_i(x,y)$ respectively represent the grayscale version and DSIFT feature measure of the source input image sequence I_i. To optimize memory usage, an 8-bin orientation histogram and a 2 * 2 cell array are employed

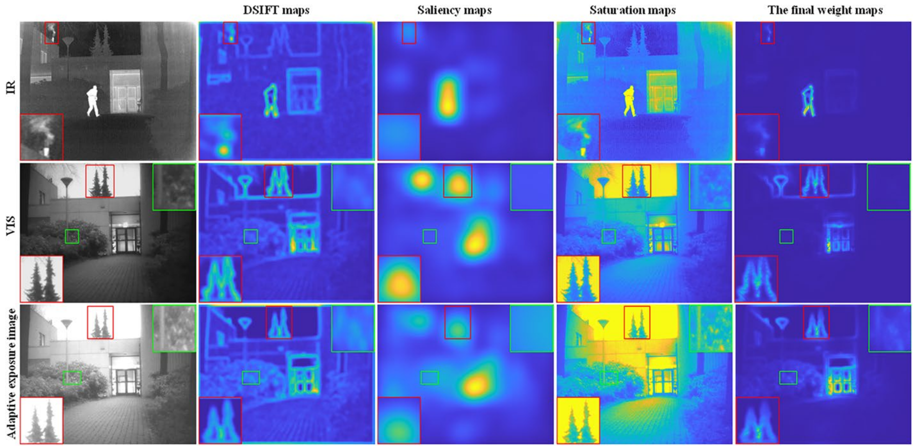

Fig. 6 Weight maps (pseudo-color) for the “Kaptein_1123”

in each cell to generate the descriptor. As a result, the dimension of each descriptor vector is limited to 32.

Figure 6 (second column) displays the DSIFT weight map C_i of “Kaptein_1123”. The Fig. 6 clearly illustrates the ability of DSIFT to capture spatial details from the source input image effectively. This capability enhances the transmission of intricate information to the fused image, thereby improving the overall fusion outcome.

3.2.2 Weight maps via saliency

In order to simulate the Human Visual System (HVS), various computational models have been devised to emphasize salient areas. Saliency maps are widely used in image processing tasks to improve image quality and generate visually appealing outputs.

In this study, highlighted image regions (more attractive to human observers) are given greater weights through saliency maps. Compared to other saliency algorithm methods, the distances between images caused by image features are closer to human perceptual distances. Therefore, we adopted the method proposed by Hou et al. [37] to obtain the saliency map. The method depends on the image signature descriptor. The descriptor is defined as follows:

$$ISig(I_g) = sign(DCT(I_g)), \tag{16}$$

$$I_g = F + Z, I_g,\ F, Z \in R^N. \tag{17}$$

where DCT stands for Discrete Cosine Transform, I_g represents the grayscale version of the source input image sequence I_i, and $sign(*)$ represents the entrywise sign operator. F represents the foreground and is assumed to have sparse support in the standard spatial basis. Z represents the background and is assumed to have sparse support in the basis of the DCT.

For the problem of foreground–background separation, it is very difficult to accurately separate F and Z when only I_g is given. Therefore, the focus is solely on the spatial support for F (F is a non-zero pixel set). Hou et al. [37] approximately isolate the support of F by taking the sign of the mixture signal I_{g} in the transformed domain, and subsequently reverted it back to the spatial domain through an inverse transformation. The reconstructed image is expressed as:

$$\overline{I_{\mathrm{g}}} = IDCT\left[sign(DCT(I_{\mathrm{g}}))\right], \tag{18}$$

where $IDCT$ represents the inverse DCT transform. Assuming that the foreground of an image is visually salient compared to its background, the saliency map S_i (where i represents the i-th image in the input image sequence) can be obtained by smoothing the reconstructed image by the square of $\overline{I_{\mathrm{g}}}$:

$$S_i = \widehat{G} * (\overline{I_{\mathrm{g}}} \circ \overline{I_{\mathrm{g}}}), \tag{19}$$

where $\widehat{G}$ is a Gaussian kernel, $*$ is the convolution operator and $\circ$ is the Hadamard (entry-wise) product operator. Readers can refer to Hou et al. [37] to learn more details.

In Fig. 6 (third column), the obtained saliency maps S_i are presented for the "Kaptein_1123".

3.2.3 Weight maps via saturation

Considering the saturation feature is crucial when estimating the weight map because saturation can make the images appear more vivid and vibrant. The method used in [38], which calculates the standard deviation of each pixel in the R, G, and B channels, is primarily employed to measure the dispersion of colors. However, in some cases, measuring only the color dispersion might not be sufficient to accurately reflect the image's saturation. In contrast, our proposed approach focuses more on the pixel count within the color channels rather than the dispersion of colors by calculating the sum of absolute values for each pixel in the R, G, and B channels.

In grayscale images, each pixel's brightness value is unique and is not influenced by color channels. Therefore, the saturation of grayscale images is primarily reflected in the brightness variation. Calculating the sum of absolute values for each pixel can more directly represent the overall intensity of brightness. By employing this method, the

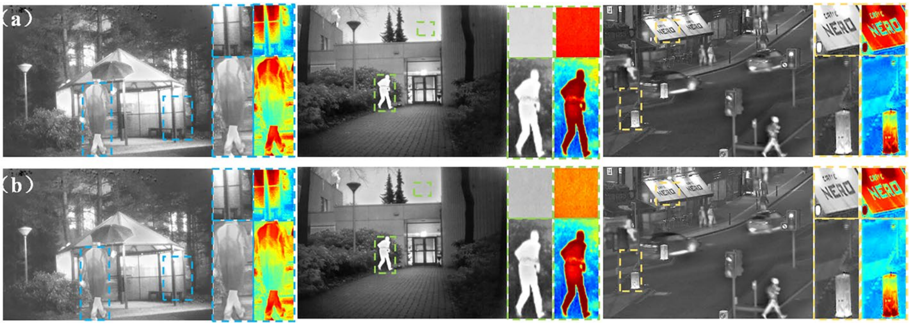

Fig. 7 The fusion results obtained using different saturation measurement methods. The first line represents the results obtained using our method, while the second line represents the fusion results obtained using the method described in reference [38]

saturation weight map A_i (where i represents the i-th image in the input image sequence) can better capture the changes in brightness in grayscale images. This further enhances the image contrast, making the details in the image more clearly visible.

Figure 6 (fourth column) shows the saturation weight map A_i obtained using the "Kaptein_1123" image. As shown in Fig. 7, compared with the method in [38], our saturation measurement method provides a higher contrast fusion result, making the critical information in the image more prominent, thereby improving the visual quality.

3.3 Weight map refinement and image fusion

For each input image, the initial weight map is calculated by combining three weights: the DSIFT weight map C_i, the saliency weight map S_i, and the saturation weight map A_i.

As the estimated initial weights often contain discontinuities, a Mutually Guided Image Filtering technique is employed to remove noise and reduce these discontinuities from the initial weights. For every pixel, information from each metric is multiplied to obtain a scalar weight map. This scalar weight map is then optimized to yield the refined weight map W_i:

$$W_i = \mathrm{MuGIF}\left(\mathrm{C}_i \times S_i \times A_i\right),\ i = 1 \ldots n \tag{20}$$

where MuGIF represents the Mutually Guided Image Filtering. Subsequently, the optimized weight map W_i is normalized to ensure that the total exposure at each spatial position (x, y) adheres to the constraint that the sum is 1, thereby defining the final weight map:

$$\widehat{W}_i(x, y) = \left[\sum_{i=1}^{n}\left(W_i(x, y) + \varepsilon\right)\right]^{-1}\left(W_i(x, y) + \varepsilon\right), \tag{21}$$

where ε is a small positive value (e.g., 10^{-25}). The last column of Fig. 6 gives the final weight map of "Kaptein_1123". As shown in Fig. 6, the final weight map effectively preserves the significant target from the infrared image (as shown in the red box in the first row and fifth column in Fig. 6). Additionally, the visible image and adaptive exposure image complement each other, with the final weight of the visible image being relatively large in well-exposed areas (as shown in the red box in column 5 of the second and third rows in Fig. 6). In the under-exposed area, the final weight of the adaptive exposure image is relatively large (as shown in the green box in column 5 of the second and third rows in Fig. 6).

Due to variations in local intensity among the input images, using a direct-weighted mixing strategy might result in artifacts and unsatisfactory outputs. However, this problem can be avoided by adopting a pyramid-based multi-resolution method [39]. This approach involves converting the images into a pyramid structure, blending them at each level, and reconstructing the pyramid to obtain the fused image. Specifically, the method uses the Gaussian and Laplacian pyramids to decompose the obtained weights and input image sequences, respectively. Then, a weighted blending strategy is applied at each pyramid level to get a new Laplacian pyramid for the fused image. The entire process is as follows:

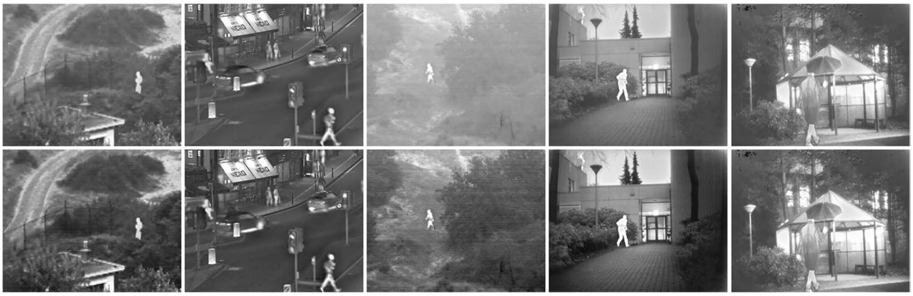

Fig. 8 Contrast enhancement of pre-fused image using Fast Guided Filter. The first row is the pre-fused image, and the second row is the final fused image after contrast enhancement

$$L\{F(x,y)\}^l = \sum_{i=1}^{n} G\left\{ \widehat{W}_i(x,y) \right\}^l L\left\{ I_i(x,y) \right\}^l, \tag{22}$$

where l denotes pyramid decomposition level, G represents the Gaussian pyramid, L represents the Laplacian pyramid. The fused pyramid $L\{F(x,y)\}^l$ is finally collapsed to yield the pre-fused image I_{pre}.

However, in all the operations mentioned above, our primary focus has been preserving the original images' details within the input image sequence. However, the saliency of the infrared target in the fused image and the final fusion effect are often neglected. For instance, the contrast in the preliminary fusion image we obtain is insufficient. There are two possible reasons for this.

Firstly, we calculate a perceived quality measure for each pixel in the input image sequence during the multi-weight feature extraction. Guided by these quality measures, we selectively choose "good" pixels from the input sequence and incorporate them into the final result. Here, our primary concern is detecting abrupt changes in pixel values within the image. Notably, there are very few pixels inside the infrared target, leading to the generation of pre-fused images containing many smooth regions within the infrared target pixels. Secondly, in the design of the weight fusion rule, we compromise the pixel intensity of certain prominent targets in the infrared image to preserve the details of both the infrared and visible images. To solve this problem, the Fast Guided Filter is used to further process the I_{pre}:

$$\widehat{R} = \mathrm{FGF}\left(I_{pre}, I_{pre}, r, eps, s\right), \tag{23}$$

$$F = \left(I_{pre} - \widehat{R}\right) \otimes \eta + \widehat{R}. \tag{24}$$

where FGF represents the Fast Guided Filter, I_{pre} is the guide image, $\widehat{R}$ is the filtering result of the FGF, $\otimes$ stands for multiplication, η is the enhancement coefficient, and F is the final fusion image. r, eps, s are the parameters, where $r = 32$, $eps = 0.05^2$, $s = 8$, $\eta = 3$. As shown in Fig. 8, based on the FGF, we obtain a fused image with higher contrast and highlighting the infrared target. The following are the steps of our proposed infrared and visible image fusion algorithm, as shown in Algorithm 2.

```
Parameter descriptions:
$VIS$ is the input visible image.
$IR$ is the input infrared image.
$G$ is the Gauss pyramid.
$L$ is the Laplace pyramid.
$l$ is the decomposition level of the pyramid.
MuGIF is a mutually guided image filtering.
Ä stands for multiplication.
$F$ is the final fusion image.

Input: $VIS$, $IR$ Output: $F$
$J$ is the image after adaptive exposure adjustment by Algorithm 1;
for $i = 1$ to $n$ do // Traverse the input sequence containing elements $VIS$, $IR$ and $J$
    obtain the DSIFT feature weight map $C_i$ of $VIS$, $IR$ and $J$ by (15);
    obtain the saliency weight map $S_i$ of $VIS$, $IR$ and $J$ by (19);
    obtain the saturation weight map $A_i$ of $VIS$, $IR$ and $J$ VI by Section 3.2.3;
    $W_i$ =MuGIF$(C_i$ ´ $S_i$ ´ $A_i)$, $i$ =1¼$n$; // Weight map optimization via MuGIF
    $W_{pyr} = G(W_i)$; // Gaussian pyramid decomposition
    $I_{pyr} = L(VIS, IR, J)$; // Laplace pyramid decomposition
        for $j = 1$ to $l$ do // process the features of each layer
        $L^l = W_{pyr}^l$ Ä $I_{pyr}^l$;
        end for
end for
the pre-fused image $I_{pre}$ is obtained by the inverse pyramid transformation;
obtain the final fusion image $F$ by (23) and (24);
return $F$
```

Algorithm 2 Main steps of the proposed infrared and visible image fusion algorithm

4 Experiment and analysis

In this section, we first introduce the experimental configuration. Then, some ablation studies were conducted to verify the effectiveness of our designed module. Next, we evaluate the fusion performance of the proposed method and compare it with other existing fusion methods to demonstrate the superiority of our algorithm. Finally, we analyze the operating efficiency of different methods.

Table 1 The comparison method used in this paper

Name/Reference	Year	Traditional/Deep learning	Category	Code
GTF [27]	2016	Traditional	Other	Matlab
FPDE [40]	2017	Traditional	Subspace-based	Matlab
MGFF [41]	2019	Traditional	Multi-scale	Matlab
MDLatLRR [13]	2020	Traditional	Multi-scale	Matlab
Bayesian [42]	2020	Traditional	Other	Matlab
MST [43]	2020	Traditional	Multi-scale	Matlab
SDNet [28]	2021	Deep learning	CNN	Tensorflow
RFN-Nest [44]	2021	Deep learning	AE	Pytorch
U2Fusion [23]	2022	Deep learning	CNN	pytorch
Luo et al. [45]	2022	Traditional	Multi-scale	Matlab
CMTFusion [46]	2023	Deep learning	CNN	pytorch

4.1 Experiment setting

We select two public datasets to evaluate the fusion performance of the proposed method. TNO [47, 48] and VIFB [49] are both public datasets. Specifically, 21 pairs of infrared and visible light images were selected for testing in TNO. The VIFB dataset also contains 21 pairs of visible and infrared images, which the authors collected from the fused tracking dataset [50], which covers indoor, outdoor, low-light, over-exposure, and other extensive environments.

To validate the performance of the proposed method, we conduct qualitative and quantitative experiments with eleven state-of-the-art infrared and visible image fusion methods, including GTF(2016) [27], FPDE(2017) [40], MGFF(2019) [41], Bayesian(2020) [42], MST(2020) [43], MDLatLRR(2020) [13], SDNet(2021) [28], RFN-Nest(2021) [44], U2Fusion(2022) [23], Luo et al.(2022) [45] and CMTFusion(2023) [46]. Table 1 lists more details about these algorithms. The relevant parameters of the algorithms above are set according to the original papers. In this paper, the decomposition level in MDLatLRR is 4. We implement all experiments using Matlab codes on a PC with an AMD Ryzen 7 5800H, 16G RAM, and CPU @3.20 GHz processor.

To objectively illustrate the performance of the proposed fusion method, six image quality evaluation indicators are adopted: Entropy (EN) [51], Mean value (ME), Pixel-based visual information fidelity (VIFP) [52], Gradient-based fusion metric ($Q^{AB/F}$) [53], information fidelity criterion (IFC), and Chen-Varshney metric (Q_{CV}) [54]. Among these indicators, EN can reflect the amount of detail and texture of the fused image, and ME calculates the arithmetic mean of all pixels, representing the average brightness that human eyes can perceive. VIFP measures the visual fidelity between the fused image and the source image. $Q^{AB/F}$ is used to estimate the edge information in the fused image. IFC calculates how much information is exchanged between the input source image and the output image. The larger these values, the better the quality of the fused image. Q_{CV} is a human perception-inspired indicator measuring the visual difference between the source and fused images. The smaller the Q_{CV} value, the better the fusion performance.

4.2 Ablation experiments

4.2.1 Experimental validation of the AEAV module

To verify the validity of AEAV in this paper, corresponding ablation experiments are conducted here. Taking the final fusion image as an example, we mainly use AEAV to enrich the texture details of the fusion image. The low-exposure pixels in the visible image are enhanced to reveal

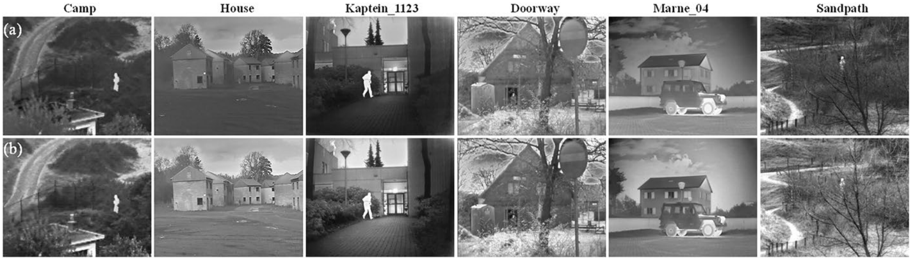

Fig. 9 Fusion image comparison: (**a**) row is the fused image without the AEAV. Row (**b**) is the fused image after processing by our method

Table 2 Six different parameters are used to compare the advantages of fusion results. The first row of each comparison parameter is the data obtained from the fused image in row (a) in Fig. 9, and the second row is the experimental data from the fused image in row (b). The best values are indicated in bold and underlined

	Camp	House	Kaptein_1123	Doorway	Marne_04	Sandpath
EN↑	6.6540	6.6379	7.1139	7.2053	7.1736	6.7968
	7.2158	**6.9412**	**7.5664**	**7.3856**	**7.3008**	**7.3098**
ME↑	75.5390	102.8500	72.6080	**135.2700**	121.3100	79.7100
	93.7730	**143.9500**	**93.1190**	133.8300	**133.1000**	**125.2900**
VIFP↑	0.2931	0.3393	0.4137	0.3461	0.4256	0.3351
	0.4175	**0.4727**	**0.5350**	**0.3904**	**0.4660**	**0.4912**
$Q^{AB/F}$↑	0.4752	0.6023	**0.4988**	0.5503	**0.5479**	0.5045
	0.5300	**0.6893**	0.4184	**0.5930**	0.4801	**0.5773**
IFC↑	2.2982	2.9758	3.7442	3.5434	**3.7657**	3.3729
	3.4402	**4.6163**	**3.9817**	**4.1772**	3.3960	**5.3350**
Q_{CV}↓	**318.1100**	**329.8400**	**101.2000**	712.0800	803.7200	**142.2100**
	566.7800	454.1100	261.2600	**654.0300**	**370.9400**	226.5500

the details hidden in the dark so that the final fusion image has the rich texture details of the visible image. We compare the fusion images obtained with and without an AEAV module. The results are shown in Fig. 9. It can be seen that the fusion image processed by the AEAV module is superior to the fusion image without AEAV in terms of detail richness and brightness.

Additionally, it maintains the significance of the infrared target. In Table 2, the quantitative experimental scores of the six fusion image evaluation measures further illustrate the validity of the proposed AEAV module. The first row of each comparison parameter is the data obtained from the fused image in row (a) in Fig. 9, and the second row is the experimental data from the fused image in row (b).

4.2.2 Experimental validation of the multi-weight maps

To validate the DSIFT, Saliency, and Saturation features, the weight maps are utilized individually and in pairwise combinations to guide the fusion process. The experiments are

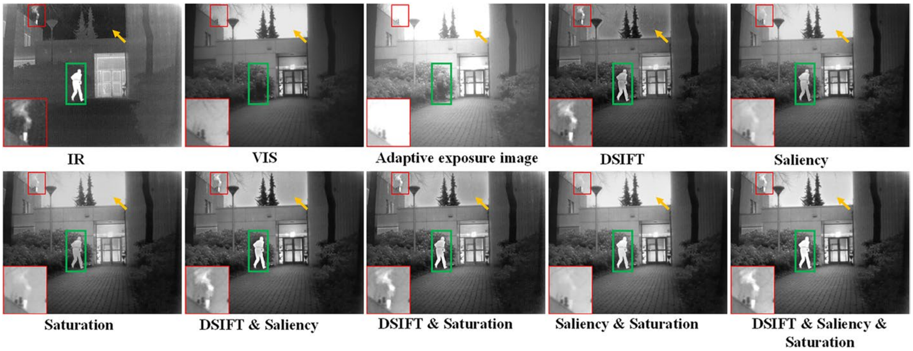

Fig. 10 Experimental validation results of single or multiple weight maps

Table 3 Comparison of the weights used in the fusion process. The best values are indicated in bold and underlined

	EN↑	ME↑	VIFP↑	$Q^{AB/F}$↑	IFC↑	Q_{CV}↓
DSIFT	6.8252	116.6525	0.4130	0.5371	3.5243	471.9323
Saliency	7.0096	118.3390	0.4340	0.4896	3.5072	508.4009
Saturation	7.0552	113.9920	0.4216	0.4780	3.1657	**389.3341**
DSIFT & Saliency	6.9419	120.1027	0.4517	**0.5476**	4.0616	479.6632
DSIFT & Saturation	7.0202	115.1073	0.4489	0.5330	3.7686	416.3226
Saliency & Saturation	7.0646	**124.4305**	0.4594	0.4990	3.8149	400.2803
DSIFT & Saliency & Saturation	**7.0839**	119.8910	**0.4792**	0.5362	**4.2732**	435.9792

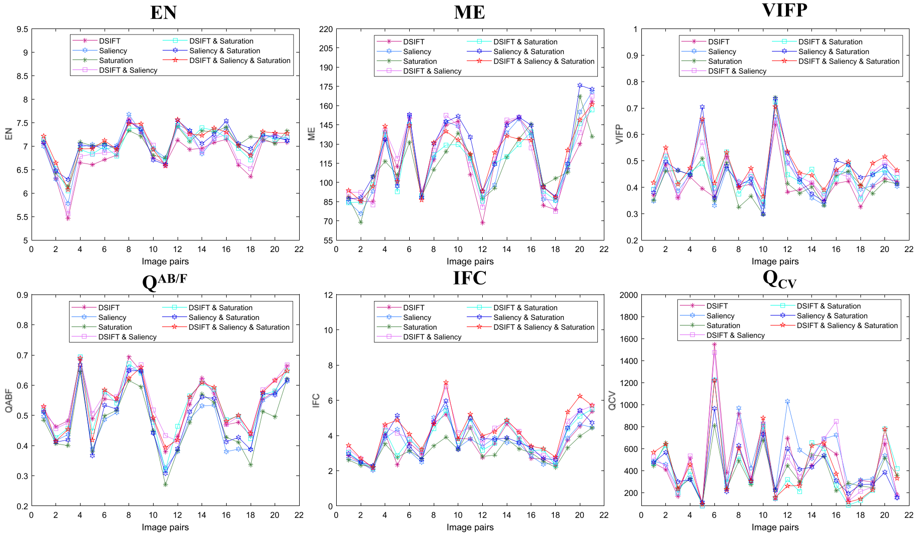

Fig. 11 Quantitative comparison of the different ablation experiments on TNO dataset

conducted on the TNO dataset, and all comparative experiments are performed under consistent settings. In Fig. 10, we illustrate the fusion results of the proposed method using different weight maps, taking “Kaptein_1123” as an example. When employing the DSIFT weight map alone, artifacts are noticeable in the fusion result, particularly in the sky area (indicated by the yellow arrow). If only the Saliency weight map is used, the fused image will lack the details of the smoke, which are clearly visible in the enlarged red box area. Similarly, using the saturation weight map alone also leads to the loss of smoke details while reducing the contrast of the fused image.

When combining any two weight maps, “DSIFT & Saliency” and “DSIFT & Saturation” introduce spatial inconsistencies, particularly in the sky area, causing it to appear unusually dim (indicated by the yellow arrow). Although the overall visual effect is improved in the “Saliency & Saturation” combination, there remains an issue with the loss of smoke details. Moreover, the infrared target saliency in the “Saliency &

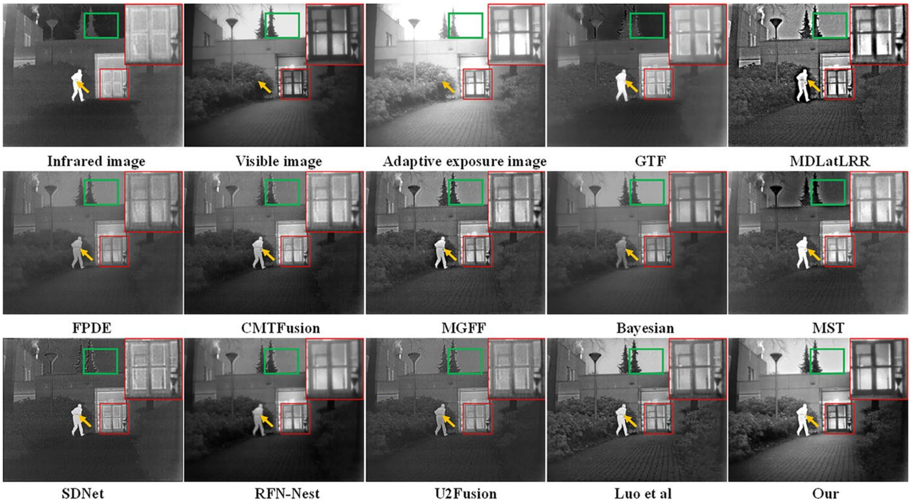
Infrared image
Visible image
Adaptive exposure image
GTF
MDLatLRR
FPDE
CMTFusion
MGFF
Bayesian
MST
SDNet
RFN-Nest
U2Fusion
Luo et al
Our

Fig. 12 The fusion results on TNO dataset ("Kaptein_1123" image)

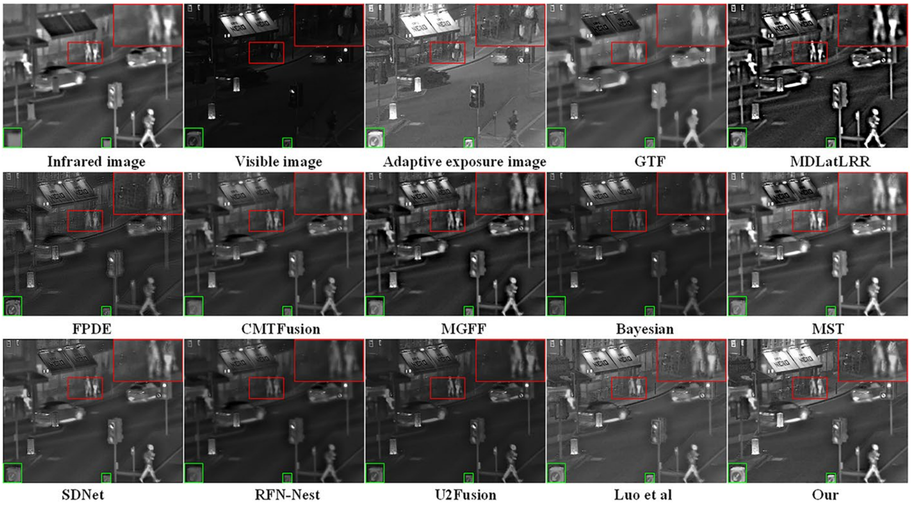
Infrared image
Visible image
Adaptive exposure image
GTF
MDLatLRR
FPDE
CMTFusion
MGFF
Bayesian
MST
SDNet
RFN-Nest
U2Fusion
Luo et al
Our

Fig. 13 The fusion results on TNO dataset ("Road" image)

Saturation" combination is less effective than when utilizing all three weight maps in combination (as demonstrated in the green box).

Six quality metrics (EN, ME, VIFP, $Q^{AB/F}$, IFC and Q_{CV}) were selected to evaluate the fusion performance of different weight maps on the TNO dataset. The average values are shown in Table 3 and Fig. 11, the best values are indicated in bold and underlined. As can be seen from Table 3 when the weight is only "Saturation", the index Q_{CV} gets the best score. When the weight combination is "DSIFT & Saliency" or "Saliency & Saturation", $Q^{AB/F}$ or ME score is the best. However, it was observed that with the combination of all three weights producing the greatest EN, VIFP, and IFC scores. For $Q^{AB/F}$ and ME, the fusion image obtained by combining the three weight maps still achieves comparable performance. Therefore, the method of combining three types of weight maps proves to be more effective in guiding the fusion process and can produce better fusion results.

4.3 Qualitative results

4.3.1 Fusion results of TNO

We compare the fusion performance of 11 existing fusion methods and the proposed algorithm on the TNO dataset. The fusion result of two pairs of source images ("Kaptein_1123" and "Road") is shown in Figs. 12 and 13.

In Figs. 12 and 13, FPDE, Bayesian, and U2Fusion cannot highlight infrared salient targets, and GTF and RFN-Nest blur the edges of infrared targets. As can be seen from the red box in Fig. 12, MDLatLRR, FPDE, CMTFusion, MGFF, SDNet, and U2Fusion introduce noise into the fused image while retaining texture. As can be seen from the red box in Fig. 13, except for FPDE, Luo et al., and our method, the details and textures of other methods are almost invisible. In addition, except that our method and Luo et al. maintain the spatial consistency of the sky area (shown as the green box in Fig. 12), other methods suffer from spectral pollution to varying degrees. However, compared with Luo et al.'s method, our fusion results contain more apparent texture details and more prominent infrared targets (such as yellow arrows and highlighted green rectangles in Figs. 12 and 13). Furthermore, the proposed method produces more natural fused images than other fusion methods.

4.3.2 Fusion results of VIFB

In this paper, many advanced algorithms are initially designed to fuse grayscale images. The VIFB data set used for testing has some visible images that are RGB

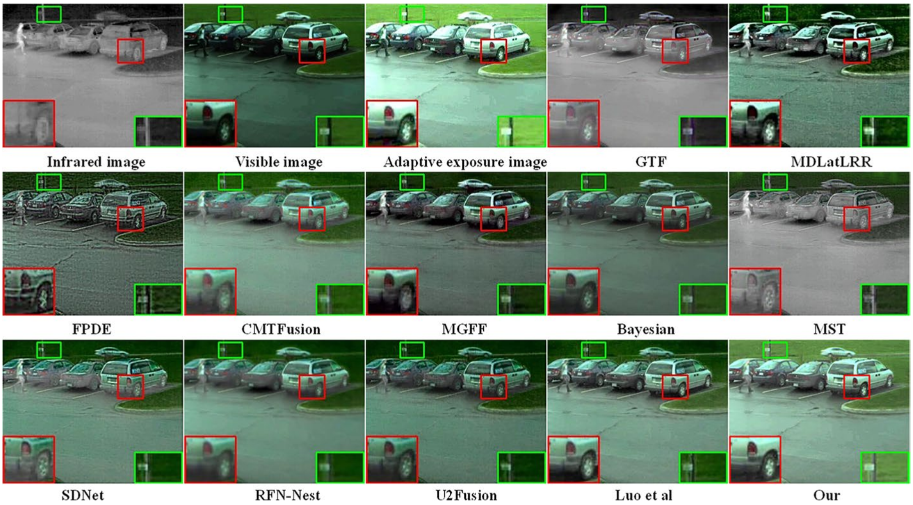

Fig. 14 The fusion results on VIFB dataset ("walking2" image)

images. Therefore, as in VIFB [49], we improve each channel of RGB images by fusing them with the corresponding infrared image, achieving color image fusion. The fusion results of two image pairs from the VIFB dataset are shown in Figs. 14 and 15.

In Fig. 14, GTF and MST cannot preserve the color and rich texture detail information of the visible image (as shown in the green area in the Fig. 14). The fusion images generated by FPDE, CMTFusion, and RFN-Nest are relatively blurry, especially FPDE, which causes the deformation of the car tires. Although MDLatLRR, MGFF, Bayesian, SDNet, U2Fusion, and Luo et al. can retain the detailed information of visible images to a certain extent, the contour edges of the car are not clear enough, and the overall image fusion result is dark, which is not convenient for people to observe. On the contrary, compared with other advanced methods, our fusion result retains the clear edges and texture of the rear tire and presents a more natural fusion effect (as shown in the enlarged area of the red rectangle).

In Fig. 15, the people around the car are not visible in the visible light image due to overexposure. As shown in the enlarged area of the green box in Fig. 15, the people around the car are still invisible or unclear in the fusion results of many methods, such as the fusion results of Bayesian, FPDE, RFN-Nest, and Luo et al. The infrared targets are not noticeable, although one can see the people around the car from the fusion results of GTF, MST, SDNet, and U2Fusion. MDLatLRR, CMTFusion, MGFF, and our method can well maintain the saliency of infrared targets even under intense light, but the fused image of MDLatLRR has the problem of over-enhancement, resulting in infrared target distortion. In addition, in terms of visible light detail preservation, compared with other methods, our method can maintain a clearer and more natural outline of the visible image (shown in the red enlarged area).

The fusion results of other groups are shown in Fig. 16, which shows the same phenomenon (as shown in the enlarged red box area).

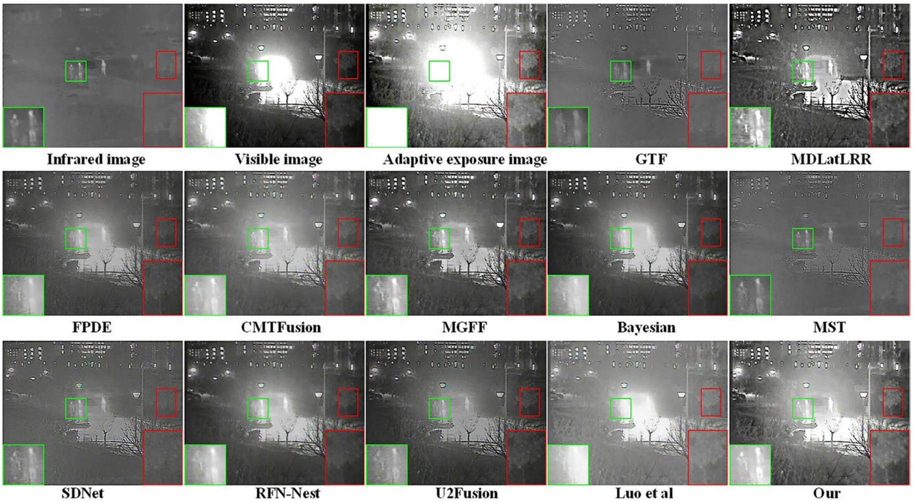

Fig. 15 The fusion results on VIFB dataset ("manlight" image)

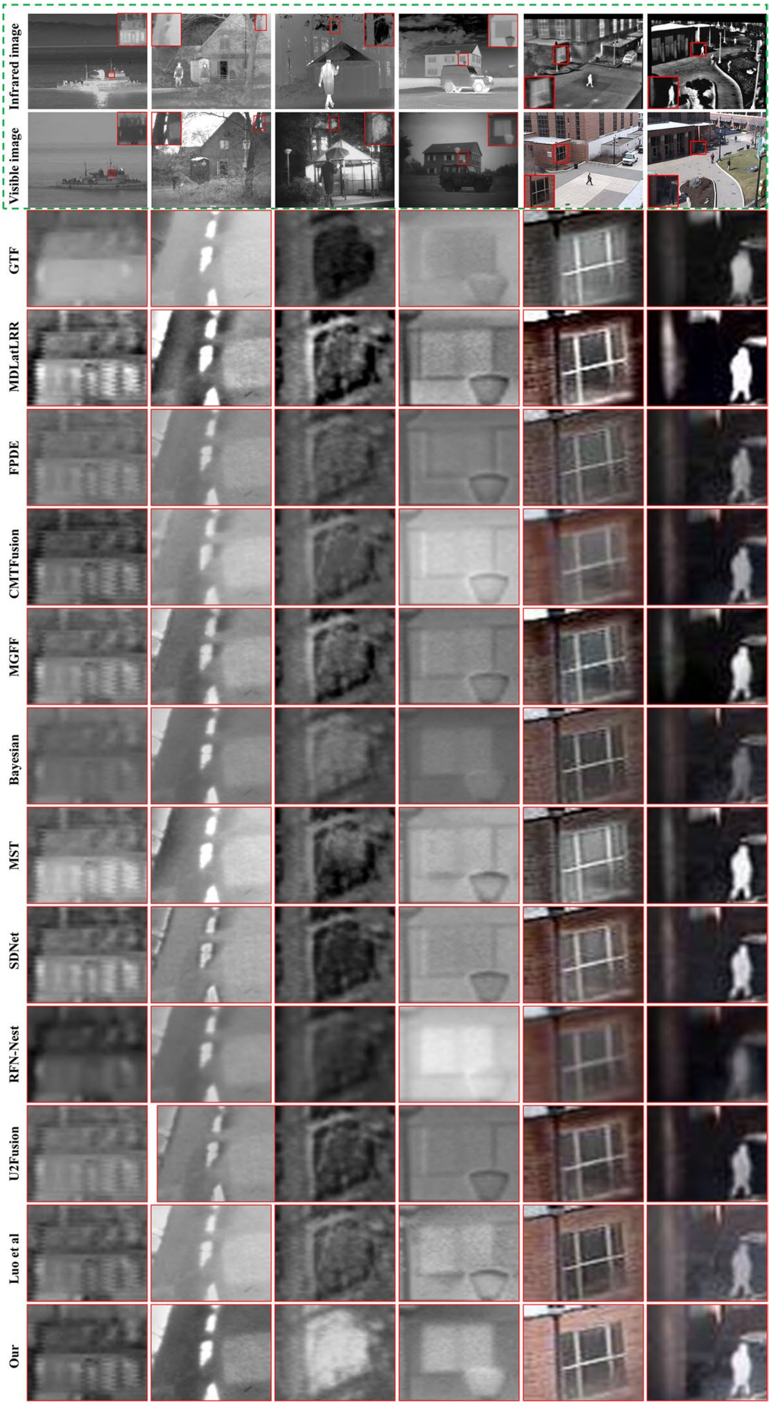

Fig. 16 Fusion results of six pairs of infrared and visible images in TNO and VIFB datasets by different methods

Table 4 Average quantitative comparison results on TNO dataset, the optimal values for each metric are shown in bold and underlined, and the suboptimal values are underlined and italicized

Method	EN↑	ME↑	VIFP↑	Qab/F↑	IFC↑	QCV↓	Time(/s)
GTF	6.6353	107.5899	0.2408	0.4265	2.9596	1281.2336	5.5181
MDLatLRR	*6.9777*	118.2601	0.3398	0.3362	2.9287	656.7194	164.9806
FPDE	6.2549	111.6192	0.2707	0.4879	3.1601	489.4176	2.1680
CMTFusion	6.7821	103.7051	0.3189	0.4129	2.4972	620.6525	0.6253
MGFF	6.5390	100.8223	0.3103	0.4633	2.3973	563.5695	1.2650
Bayesian	6.3045	111.2663	0.3176	0.4840	3.4435	*453.5838*	1.4870
MST	6.7596	108.6469	0.3141	*0.5218*	2.8328	931.7186	**0.0600**
SDNet	6.4906	100.8989	0.2684	0.4227	2.2122	886.3064	*0.1650*
RFN-Nest	6.8414	93.3474	0.2809	0.3598	2.4709	534.2287	6.4760
U2Fusion	6.2466	103.3921	0.2644	0.4273	2.2576	618.5916	1.2837
Luo et al	6.6463	**122.7836**	*0.3564*	0.5086	*3.1359*	460.9323	44.9021
Our	**7.0839**	*119.8910*	**0.4792**	**0.5362**	**4.2732**	**435.9792**	0.9779

Table 5 Average quantitative comparison results on VIFB dataset, the optimal values for each metric are shown in bold and underlined, and the suboptimal values are underlined and italicized

Method	EN↑	ME↑	VIFP↑	Qab/F↑	IFC↑	QCV↓	Time(/s)
GTF	6.4452	111.7971	0.2530	0.4552	2.2008	2085.6339	4.8901
MDatLRR	6.7739	108.7773	0.3148	0.5261	2.4529	*690.3659*	74.6554
FPDE	6.7615	106.4071	0.2776	0.4846	2.3319	766.2566	1.9892
CMTFusion	6.9572	118.7655	0.3017	0.3964	2.0047	1010.5483	1.1880
MGFF	7.1144	105.9986	*0.3391*	*0.5684*	2.4689	667.7371	0.7512
Bayesian	6.8780	107.7806	0.3360	0.5509	*2.6667*	607.0103	1.6585
MST	6.4391	116.3998	0.2724	0.4632	2.1509	1639.7280	**0.0130**
SDNet	6.6275	107.7328	0.2618	0.5056	2.1255	1243.2490	*0.1891*
RFN-Nest	*7.1428*	104.1336	0.2977	0.4160	2.1401	822.3278	0.4829
U2Fusion	6.7484	100.5146	0.2992	0.5212	2.3817	787.7494	2.4388
Luo et al	6.9803	*132.6387*	0.3277	0.5330	2.4983	926.4335	38.3791
Our	**7.2723**	**148.0075**	**0.4002**	**0.5823**	**3.1277**	**649.5506**	0.9965

4.4 Quantitative results

We quantitatively compared the proposed method with the above 11 comparison methods on the TNO and VIFB datasets. The average values of the six metrics (EN, ME, VIFP, $Q^{AB/F}$, IFC, and Q_{CV}) on the TNO and VIFB datasets are shown in Tables 4 and 5. The optimal values for each metric are shown in bold and underlined, and the suboptimal values are underlined and italicized. The quantitative comparison results corresponding to Tables 4 and 5 are shown in Figs. 17 and 18.

In Table 4, the proposed method achieves the highest values for EN, VIFP, $Q^{AB/F}$, IFC and Q_{CV}, and secures the second-best score in ME on the TNO dataset. Although our method does not reach the optimal value in ME, it still produces comparable results. In Table 5, the proposed method obtains the best values for EN, ME, VIFP, $Q^{AB/F}$, IFC and

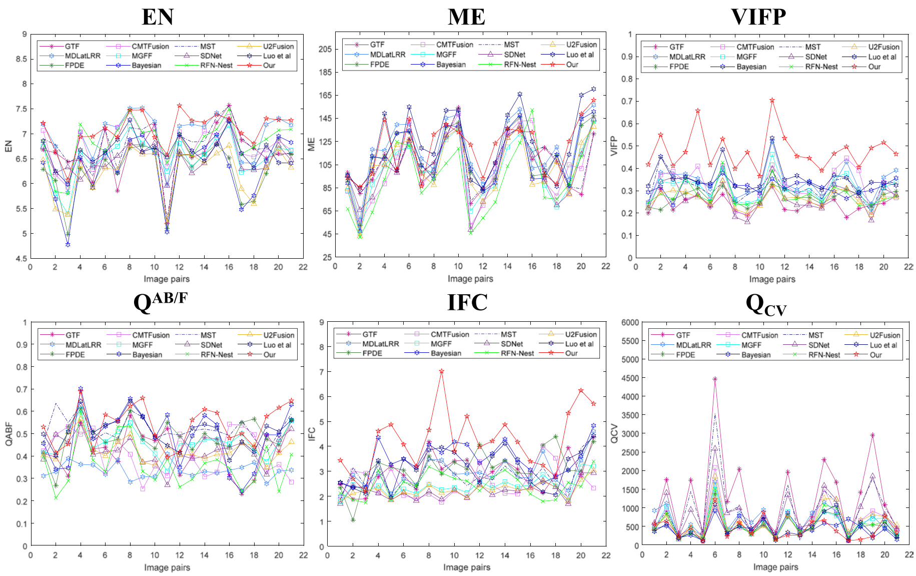

Fig. 17 Comparison of our method with eleven state-of-the-art methods on TNO dataset

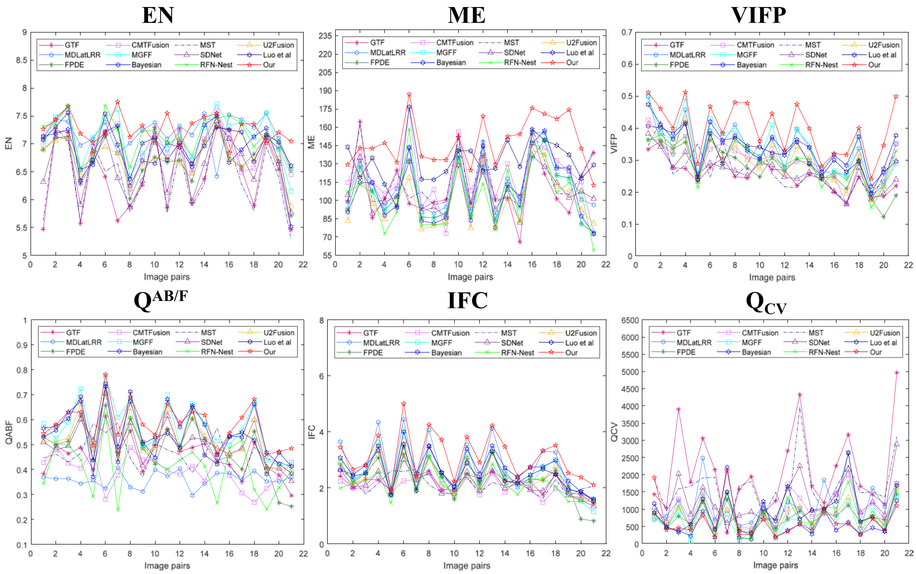

Fig. 18 Comparison of our method with eleven state-of-the-art methods on VIFB dataset

Q_{CV} on the VIIFB dataset. These metrics indicate that our algorithm delivers superior contrast and edge texture information, emphasizing infrared targets while enhancing visible details. In addition, the actual measurement results also show that the fusion image generated by the method is more natural.

4.5 Efficiency comparison

To comprehensively evaluate different algorithms, we have included the average running times of various methods on the TNO and VIFB datasets in the last columns of Tables 4 and 5.

In Table 4, our method ranks fourth in running time compared to other methods, outperforming specific deep learning methods like U2Fusion and RFN-Nest. In Table 5, our method ranks fifth in runtime when compared to other methods, surpassing some traditional methods such as GTF and MDLatLRR. This achievement is primarily due to our method eliminating traditional decomposition and fusion steps, significantly reducing the time required for image fusion. Specifically, methods like MDLatLRR, which obtain low-rank representation through multi-scale decomposition, tend to be highly time-consuming. Although our method might not have an absolute advantage in terms of runtime compared to other methods, the resulting fused image aligns more closely with human visual perception.

5 Conclusion

In this paper, we propose a novel multi-weight fusion framework for infrared and visible image fusion. Firstly, we propose an adaptive visible light exposure adjustment module to enhance the low-luminance pixel areas in the visible images, obtaining adaptive exposure images. This approach allows the fused image to contain more details and texture information from the visible images. Secondly, we employ a multi-weight feature extraction module to extract three feature weight maps from the input infrared, visible, and adaptive exposure images: DSIFT, saliency, and saturation maps. These feature weight maps are then optimized using MuGIF. Subsequently, we decompose and reconstruct the feature weight maps and input images using Gaussian and Laplacian pyramids, generating the pre-fusion images. Finally, to further enhance the contrast of the pre-fusion images, we apply a Fast Guided Filter to process the pre-fused images, producing the final fusion results. This method effectively preserves the saliency of infrared targets and background information from visible images, making the fusion results more consistent with human visual perception.

Qualitative and quantitative comparisons with 11 other state-of-the-art methods demonstrate the superiority of our method. However, as discussed in Section 4.5, our method has some limitations in terms of computational efficiency. In future research, we will explore more efficient fusion methods. In addition, with the continuous development of color visible and infrared image datasets, we will continue investigating more multi-channel color fusion image generation methods.

Acknowledgements This work was supported in part by the National Natural Science Foundation of China under Grant 62202416, Grant 62162068, Grant 62172354, Grant 62162065, in part by the Yunnan Province Ten Thousand Talents Program and Yunling Scholars Special Project under Grant YNWR-YLXZ-2018-022, in part by the Yunnan Provincial Science and Technology Department-Yunnan University “Double First Class” Construction Joint Fund Project under Grant No. 2019FY003012, in part by the Research Foundation of Yunnan Province No. 202105AF150011.

Author's contribution **Yiqiao Zhou:** Conceptualization, Methodology, Software, Writing – original draft. **Hongzhen Shi:** Visualization, Formal analysis. **Hao Zhang:** Validation, Data curation. **Kangjian He:** Supervision, Writing – review editing, Project administration, Funding acquisition. **Dan Xu:** Supervision, Project administration, Funding acquisition.

Data availability The datasets generated during and analyzed during the current study are available at: https://github.com/VCMHE/MWF_VIF.

Declarations

Competing interest The authors declare that there is no conflict of interest regarding the publication of the article.

References

1. Li J et al (2020) DRPL: deep regression pair learning for multi-focus image fusion. IEEE Trans Image Process 29:4816–4831. https://doi.org/10.1109/TIP.2020.2976190
2. Li H, Zhao J, Li J, Yu Z, Lu G (2023) Feature dynamic alignment and refinement for infrared-visible image fusion: Translation robust fusion. Inf Fusion 95:26–41. https://doi.org/10.1016/j.inffus.2023.02.011
3. Li J, Liang B, Lu X, Li M, Lu G, Xu Y (2023) From global to local: multi-patch and multi-scale contrastive similarity learning for unsupervised defocus blur detection. IEEE Trans Image Process 32:1158–1169. https://doi.org/10.1109/TIP.2023.3240856
4. Zhou H et al (2020) Feature matching for remote sensing image registration via manifold regularization. IEEE J Sel Top Appl Earth Obs Remote Sens 13:4564–4574. https://doi.org/10.1109/JSTARS.2020.3015350
5. Lin X et al (2022) Learning modal-invariant and temporal-memory for video-based visible-infrared person re-identification. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18–24, 2022, IEEE, pp 20941–20950. https://doi.org/10.1109/CVPR52688.2022.02030
6. Li S et al (2023) Logical relation inference and multiview information interaction for domain adaptation person re-identification. IEEE Trans Neural Netw Learn Syst 1–13. https://doi.org/10.1109/TNNLS.2023.3281504
7. Zhou Y, Xie L, He K, Xu D, Tao D, Lin X (2023) Low-light image enhancement for infrared and visible image fusion. IET Image Proc 17(11):3216–3234
8. Ma J et al (2020) Infrared and visible image fusion via detail preserving adversarial learning. Inf Fusion 54:85–98. https://doi.org/10.1016/j.inffus.2019.07.005
9. Ma J, Ma Y, Li C (2019) Infrared and visible image fusion methods and applications: a survey. Inf Fusion 45:153–178. https://doi.org/10.1016/j.inffus.2018.02.004
10. Li C et al (2023) Superpixel-based adaptive salient region analysis for infrared and visible image fusion. Neural Comput Appl 35:22511–22529
11. Borsoi RA, Imbiriba T, Bermudez JCM (2020) Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability. IEEE Trans Image Process 29:116–127. https://doi.org/10.1109/TIP.2019.2928895
12. Wang J, Xi X, Li D, Li F (2023) FusionGRAM: an infrared and visible image fusion framework based on gradient residual and attention mechanism. IEEE Trans Instrum Meas 72:1–12. https://doi.org/10.1109/TIM.2023.3237814
13. Li H, Wu X-J, Kittler J (2020) MDLatLRR: a novel decomposition method for infrared and visible image fusion. IEEE Trans Image Process 29:4733–4746. https://doi.org/10.1109/TIP.2020.2975984
14. Peng Y, Lu B-L (2017) Robust structured sparse representation via half-quadratic optimization for face recognition. Multim Tools Appl 76(6):8859–8880. https://doi.org/10.1007/s11042-016-3510-3
15. Liu G, Lin Z, Yan S, Sun J, Yu Y, Ma Y (2013) Robust recovery of subspace structures by low-rank representation. IEEE Trans Pattern Anal Mach Intell 35(1):171–184. https://doi.org/10.1109/TPAMI.2012.88
16. Liu Y, Wang Z (2015) Simultaneous image fusion and denoising with adaptive sparse representation. IET Image Process 9(5):347–357. https://doi.org/10.1049/iet-ipr.2014.0311
17. Li G, Lin Y, Qu X (2021) An infrared and visible image fusion method based on multi-scale transformation and norm optimization. Inf Fusion 71:109–129. https://doi.org/10.1016/j.inffus.2021.02.008
18. Zhang Q, Wang F, Luo Y, Han J (2021) Exploring a unified low rank representation for multi-focus image fusion. Pattern Recognit 113:107752. https://doi.org/10.1016/j.patcog.2020.107752

19. Ren L, Pan Z, Cao J, Liao J (2021) Infrared and visible image fusion based on variational autoencoder and infrared feature compensation. Infrared Phys Technol 117:103839
20. Qu L, Liu S, Wang M, Song Z (2022) TransMEF: A Transformer-Based Multi-Exposure Image Fusion Framework Using Self-Supervised Multi-Task Learning. In: Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022, AAAI Press, pp 2126–2134. https://doi.org/10.1609/aaai.v36i2.20109
21. Zhao H, Nie R (2021) Dndt: Infrared and visible image fusion via densenet and dual-transformer. In: 2021 International Conference on Information Technology and Biomedical Engineering (ICITBE), IEEE, pp 71–75
22. Qu L et al (2022) TransFuse: A Unified Transformer-based Image Fusion Framework using Self-supervised Learning. CoRR, vol. abs/2201.07451, [Online]. Available: https://arxiv.org/abs/2201.07451
23. Xu H, Ma J, Jiang J, Guo X, Ling H (2022) U2Fusion: a unified unsupervised image fusion network. IEEE Trans Pattern Anal Mach Intell 44(1):502–518. https://doi.org/10.1109/TPAMI.2020.3012548
24. Ma J, Tang L, Xu M, Zhang H, Xiao G (2021) STDFusionNet: an infrared and visible image fusion network based on salient target detection. IEEE Trans Instrum Meas 70:1–13. https://doi.org/10.1109/TIM.2021.3075747
25. Ma J, Yu W, Liang P, Li C, Jiang J (2019) FusionGAN: a generative adversarial network for infrared and visible image fusion. Inf Fusion 48:11–26. https://doi.org/10.1016/j.inffus.2018.09.004
26 Ma J, Xu H, Jiang J, Mei X, Steven Zhang X-P (2020) DDcGAN: a dual-discriminator conditional generative adversarial network for multi-resolution image fusion. IEEE Trans. Image Process 29:4980–4995. https://doi.org/10.1109/TIP.2020.2977573
27. Ma J, Chen C, Li C, Huang J (2016) Infrared and visible image fusion via gradient transfer and total variation minimization. Inf Fusion 31:100–109. https://doi.org/10.1016/j.inffus.2016.02.001
28. Zhang H, Ma J (2021) SDNet: A versatile squeeze-and-decomposition network for real-time image fusion. Int J Comput Vis 129:2761–2785. https://doi.org/10.1007/s11263-021-01501-8
29. Ying Z, Li G, Gao W (2017) A Bio-Inspired Multi-Exposure Fusion Framework for Low-light Image Enhancement. CoRR abs/1711.00591 [Online]. Available: http://arxiv.org/abs/1711.00591
30. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110. https://doi.org/10.1023/B:VISI.0000029664.99615.94
31. Liu C, Yuen J, Torralba A (2011) SIFT Flow: dense Correspondence across scenes and its applications. IEEE Trans Pattern Anal Mach Intell 33(5):978–994. https://doi.org/10.1109/TPAMI.2010.147
32. Zhang W, Cham W (2010) Gradient-directed composition of multi-exposure images. In: The Twenty-Third IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2010, San Francisco, CA, USA, 13–18 June 2010, IEEE Computer Society, pp 530–536. https://doi.org/10.1109/CVPR.2010.5540168
33. Guo X, Li Y, Ma J, Ling H (2020) Mutually guided image filtering. IEEE Trans Pattern Anal Mach Intell 42(3):694–707. https://doi.org/10.1109/TPAMI.2018.2883553
34. Li Z, Zheng J, Rahardja S (2012) Detail-enhanced exposure fusion. IEEE Trans Image Process 21(11):4672–4676. https://doi.org/10.1109/TIP.2012.2207396
35 Xu L, Yan Q, Xia Y, Jia J (2012) Structure extraction from texture via relative total variation. ACM Trans Graph 31(6):139:1-139:10. https://doi.org/10.1145/2366145.2366158
36. Guo X, Li Y, Ling H (2017) LIME: low-light image enhancement via illumination map estimation. IEEE Trans Image Process 26(2):982–993. https://doi.org/10.1109/TIP.2016.2639450
37. Hou X, Harel J, Koch C (2012) Image signature: highlighting sparse salient regions. IEEE Trans Pattern Anal Mach Intell 34(1):194–201. https://doi.org/10.1109/TPAMI.2011.146
38. Mertens T, Kautz J, Reeth FV (2009) Exposure fusion: a simple and practical alternative to high dynamic range photography. Comput Graph Forum 28(1):161–171. https://doi.org/10.1111/j.1467-8659.2008.01171.x
39. Ulucan O, Ulucan D, Türkan M (2023) Ghosting-free multi-exposure image fusion for static and dynamic scenes. Signal Process 202:108774. https://doi.org/10.1016/j.sigpro.2022.108774
40. Bavirisetti DP, Xiao G, Liu G (2017) Multi-sensor image fusion based on fourth order partial differential equations. In: 20th International Conference on Information Fusion, FUSION 2017, Xi'an, China, July 10–13, 2017, IEEE, pp 1–9. https://doi.org/10.23919/ICIF.2017.8009719
41. Bavirisetti DP, Xiao G, Zhao J, Dhuli R, Liu G (2019) Multi-scale guided image and video fusion: a fast and efficient approach. Circuits Syst Signal Process 38(12):5576–5605. https://doi.org/10.1007/s00034-019-01131-z

42. Zhao Z, Xu S, Zhang C, Liu J, Zhang J (2020) Bayesian fusion for infrared and visible images. Signal Process 177:107734. https://doi.org/10.1016/j.sigpro.2020.107734
43. Chen J, Li X, Luo L, Mei X, Ma J (2020) Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. Inf Sci 508:64–78. https://doi.org/10.1016/j.ins.2019.08.066
44. Li H, Wu X-J, Kittler J (2021) RFN-Nest: an end-to-end residual fusion network for infrared and visible images. Inf Fusion 73:72–86. https://doi.org/10.1016/j.inffus.2021.02.023
45. Luo Y, He K, Xu D, Yin W, Liu W (2022) Infrared and visible image fusion based on visibility enhancement and hybrid multiscale decomposition. Optik 258:168914
46. Park S, Vien AG, Lee C (2023) Cross-modal transformers for infrared and visible image fusion. IEEE Trans Circuits Syst Video Technol 1. https://doi.org/10.1109/TCSVT.2023.3289170
47. Li H, Wu X-J, Durrani TS (2020) NestFuse: an infrared and visible image fusion architecture based on nest connection and spatial/channel attention models. IEEE Trans Instrum Meas 69(12):9645–9656. https://doi.org/10.1109/TIM.2020.3005230
48. “TNO.” [Online]. Available: https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029
49. Zhang X, Ye P, Xiao G (2020) VIFB: A Visible and Infrared Image Fusion Benchmark. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR Workshops 2020, Seattle, WA, USA, June 14–19, 2020, Computer Vision Foundation / IEEE, pp 468–478. https://doi.org/10.1109/CVPRW50498.2020.00060
50. Li C, Liang X, Lu Y, Zhao N, Tang J (2019) RGB-T object tracking: Benchmark and baseline. Pattern Recognit 96:106977. https://doi.org/10.1016/j.patcog.2019.106977
51. Roberts W, van Aardt J, Ahmed F (2008) Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J Appl Remote Sens 2:1–28. https://doi.org/10.1117/1.2945910
52. Sheikh HR, Bovik AC (2006) Image information and visual quality. IEEE Trans Image Process 15(2):430–444. https://doi.org/10.1109/TIP.2005.859378
53. Petrovic V, Xydeas C (2005) Objective image fusion performance characterization. In: Tenth IEEE International Conference on Computer Vision (ICCV’05) Volume 1, pp 1866–1871 Vol. 2. https://doi.org/10.1109/ICCV.2005.175
54. Chen H, Varshney PK (2007) A human perception inspired quality metric for image fusion based on regional information. Inf Fusion 8(2):193–207. https://doi.org/10.1016/j.inffus.2005.10.001

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.