IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 10 January 2023, accepted 9 February 2023, date of publication 22 February 2023, date of current version 6 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3246481

RESEARCH ARTICLE

A Target-Aware Fusion Framework for Infrared and Visible Images

YINGMEI ZHANG AND HYO JONG LEE, (Member, IEEE)
Division of Computer Science and Engineering, CAIIT, Jeonbuk National University, Jeonju 54896, Republic of Korea

Corresponding author: Hyo Jong Lee (hlee@jbnu.ac.kr)

This work was supported by the Joint Demand Technology Research and Development of Regional SMEs through the Korea Ministry of SMEs and Startups, 2023, under Grant RS-2023-00207672.

ABSTRACT Infrared and visible image fusion aims to obtain an image that can retain the prominent infrared target and the detailed texture information from the source images. Most scale filter-based decomposition methods mainly attempt to extract more detailed features by increasing decomposition layers, and they ignore fully considering the intrinsic properties of the original images and the influence of noises. To solve this issue and obtain better fusion & detection performance, this paper proposes a fusion method via a globalet filter and detail enhancement model to construct a target-aware fusion framework. The globalet filter is first comprehensively considered from three perspectives based on global point-to-point optimization: target brightness, removal of gradients of large-small scales (noise reduction), and preservation of texture details. Mathematically, three limited error measure equations are constructed between the target output and the source image in the shape of the L_2-norm and the first-order derivative difference. Next, a weighted average operator and a detail enhancement model are proposed to guide the corresponding sub-layers. This model creates connections between the detail layers and the input images so that the fused detail layer contains the best pixel region from these images to the greatest extent possible after the ''maximum absolute'' rule. As a result, the fused image is reconstructed by adding the previously obtained sub-images. Extensive experimental results demonstrate that our method outperforms state-of-the-art fusion methods, particularly in highlighting infrared targets, preserving substantial details, and producing average detection accuracy of exceeds 98%.

INDEX TERMS Infrared and visible image fusion, target-aware fusion framework, a globalet filter, detail enhancement model.

I. INTRODUCTION

Image fusion technology is a popular method that can ignore the difference in the generation mechanisms of the sensors to integrate multiple images into one comprehensive and informative image. In general, a visible image captured by the visible measure sensor generally describes the texture details of a scene under light reflection. Still, it is easily affected by illumination conditions and bad weather. Similarly, infrared images that are professionally captured by an infrared measure sensor reflect the thermal information on the surface of the object through the principle of thermal radiation imaging, but they typically suffer from a lack of necessary structure detail information [1], [2], [3], [4]. In this condition, infrared and visible image fusion (IVIF) is actively studied, and its resultant images can then be further processed and applied to high-level vision tasks such as object detection [5], [6], [7], pedestrian re-identification [8], [9], and semantic segmentation [10], [11].

Over the past decade, a number of IVIF algorithms have been proposed, and they can mainly be divided into the following categories according to the different adopted theories: sparse representation-based methods, subspace-based methods, neural network-based methods, and multi-scale transform-based methods [1]. Sparse representation can employ as few sparse atoms as possible to represent

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

the significant information of an image by establishing the relationship between image features and sparse coefficients, *i.e.*, learning an overcomplete dictionary, as has been widely used in the field of image fusion [12], [13], [14]. Although dictionary learning has some advantages in adaptive feature selection and description, its generation typically requires a large number of image samples and a long training time, and it may not fully express all spatial morphologies [15]. Subspace-based methods, such as principal component analysis (PCA) [16] and independent component analysis (ICA) [17], aim to project a high-dimensional input image into a low-dimensional space or subspace because processing low-dimensional subspace data requires less time and memory. Because these two methods only focus on a specific aspect of the image, the details of the fused image will be lost to some extent [18].

The neural network-based method can be divided into two categories: traditional-based and deep learning (DL)-based. The traditional-based method, such as pulse coupled neural networks, can imitate the perceptual behavior mechanism of the human brain through a large number of intricate neurons within it. The interaction between neurons represents the transmission and processing of neuron information, thus making the neural network highly adaptable, with fault tolerance and anti-noise capabilities [19], [20], [21]. The DL-based method has been substantially improved and achieved good research results in the past five years [3], [4], [9], [22], [23], [24], [25], [26], [27], [28], [29], [30]. Liu et al. [24] applied the deep convolutional neural network (CNN) to the IVIF for the first time, which can deal with activity level measurement and weight assignment to overcome the difficulty of manual design. Li et al. [25] adopted a well-trained auto encoder-decoder, DenseFuse, to extract features from input images. Ma et al. [26] first employed a generative adversarial network (GAN) to guide the IVIF. Their content loss function measurement enforced the fused image with both the thermal radiation and the gradient information of the two source images, and the maximum-minimum competition between the generator and the discriminator helps preserve more visible details. Ma et al. [27] proposed an end-to-end model, STDFusionNet, that first calculate salient target mask and then design a specific loss function to guide the feature extraction and reconstruction. Tang et al. [28] proposed a SuperFusion that registration network and a Lovasz-Softmax loss are combined used to implement a practical network to meet registration, fusion, and semantic requirements. Several transformer-based fusion algorithms [29], [30], in recently years, have been developed. In [29], both self-attention and cross-attention were utilized using pure transformers without the help of CNNs to capture the inter-domain long-range dependencies. Nevertheless, transformer-based methods typically require significant computational resources, thereby restricting their applicability to high-resolution image. All in all, DL-based methods have achieved better performance, but there are still some objective limitations [4]: i) with existing GAN-based methods, it is difficult to fuse diverse semantic objects in specific ways, which leads to the loss of some semantic information; ii) more importantly, DL-based methods require strong computing resources to train tremendous image datasets.

Multi-scale transform-based (MST) is one of the earliest researched and most effective methods. It usually consists of three parts: i) image decomposition: decompose the source images into base layers and detail layers of multiple scales using existing or newly constructed decomposition algorithms. Discrete wavelet transforms (DWTs), dual-tree complex wavelet transforms (DTCWTs), non-subsampled contourlet transforms (NSCTs), and non-subsampled shearlet transform [31] have been widely and efficiently used in the IVIF. Nevertheless, the fused image may suffer from distortions or artifacts as a result of neglect of saliency with spatial consistency. To solve this problem, Li et al. [32] proposed a guided filter-based fusion method that combines saliency with spatial consistency through average filtering, which results in a better human visual effect. This method can also efficiently preserve edges during the decomposition process, thus leading to the fused image having clear edges without artifacts. Therefore, many single or hybrid edge-preserving filters-based fusion methods have been proposed, and these have proven to be popular decomposition tools in image fusion [33], [34], [35], [36], [37]; ii) fusion rules: design corresponding rules to fuse the base layers and detail layers respectively to obtain the merged results. The common fusion rules are 'choose-maximum' and 'weighted average.' However, this method typically suffers from low contrast, blurred edges, and loss of image details caused by noise. Specifically, most decomposition methods and fusion rules are designed from the perspective of scale, which divides the information in an image into two parts: brightness information and structure detail information. That is, 'brightness information + structure detail information = an image'. Accordingly, the fusion algorithms typically focus on retaining detailed information at various scales rather than target brightness information. As a result, the fused images exhibit low contrast and edge divergence, ultimately resulting in poor visual effects and making the targets inconspicuous.

To address the abovementioned limitations of traditional and DL-based methods, we design an IVIF via the proposed globalet filter (GLF) and detail enhancement model to construct a target-aware fusion framework. Conceptually, an image is composed of brightness distribution, gradient structure changes at large-small scales, and texture detail from the global perspective; mathematically, three corresponding limited error measure equations (E_1, E_2, E_3) are constructed between the target output and the source image in the shape of the L_2-norm and the first-order derivative difference to separate the brightness, gradient structure at different scales, and texture detail of an image. At this point, GLF is used as a decomposition tool to decompose source images into base layers and detail layers. The base layers

reflect the overall brightness distribution of fused images, so it is reasonable to employ the ''weighted average'' strategy to dominate the fusion. As the detail layers represent the gradient structure change of source images, a detail enhancement model (DEM) is proposed. Then, the merged sub-images are obtained. Finally, the fused image is obtained by adding the previously obtained sub-images. Extensive experiments on the public TNO and RoadScene datasets have proven the proposed method have better fusion and detection performance.

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

1) An effective image filter based on global point-to-point optimization is designed, which is comprehensively considered from three limited error measure equations: target brightness, noise reduction, and retention of texture details. Compared to other image filters, this filter succeeds in maintaining overall brightness, preserving spatial structures and target edges, and reducing noise.
2) A detail enhancement model based on F-norm and gradient operator is constructed to guide the fusion of detail layers, which creates the connection between the detail layers and the input images (M_1, M_2) such that the final detail image contains the best pixel region from these images to the greatest extent possible after the ''maximum absolute'' rule, then further enriches detail distribution.
3) A novel target-aware fusion framework is proposed, where the globalet filter and two fusion rules are specifically designed to generate fused images with prominent infrared targets and abundant structure details. And then put them into the detector to obtain detection results with the greater detection range and average detection accuracy of exceeds 98%.

The rest of this paper is organized as follows. Section II briefly introduces the development of MST. Section III shows the construction process of the proposed filter and proposes the IVIF method based on this filter. Section IV provides the parameters setting and experimental results, followed by the conclusion, which is presented in Section V.

II. RELATED WORK

Over the past few years, researchers have proposed many methods for IVIF tasks. As our method also relies on MST framework, we briefly introduce its four improvements: pyramid transform-based, wavelet transform-based, nonsubsampled contourlet transform-based, and edge-preserving filter-based [1].

A. PYRAMID TRANSFORM-BASED

Laplacian pyramid transformation (LPT) is the earliest MST, and it is an iterative process that requires four operations of low-pass filtering, sampling, interpolating, and differencing [38]. For example, Bulanon et al. [39] proposed an IVIF method for fruits based on LPT and fuzzy logic and thus obtained a better fruit detection method. Then, the steerable pyramid transform (SPT) [40], [41] have been proposed, which has the advantages of self-inversion, aliasing-free, translation and rotation invariance. Liu et al. [41] proposed an image fusion method based on the SPT and expectation-maximization algorithm, and the experimental results show that their approach could outperform the traditional SPT. After that, the contrast pyramid transform (CPT), a derivative of the LPT, is proposed to calculate the ratio of two adjacent low-pass filtering images of the Gaussian pyramid and consider local contrast [42].

B. WAVELET TRANSFORM-BASED

Inspired by the tower algorithm for signal decomposition and reconstruction, the wavelet transform theory has become more mature and has also been used in MST decomposition. Unlike the pyramid transform, the multi-scale representation coefficients of the wavelet transform are uncorrelated, which are embodied in wavelet transform including discrete wavelet transforms (DWTs), dual-tree complex wavelet transforms (DTCWTs), quaternion wavelet transforms (QWTs) [43]. For example, Liu et al. [31] proposed a general framework based on MST to process the task of image fusion, including DWTs-based, DTCWTs-based, and NSCT-based. Chai et al. [43] proposed an image fusion method using QWT and multiple features and applied it to multi-modal image pairs.

C. NONSUBSAMPLED CONTOURLET TRANSFORM-BASED

Because the wavelet transform cannot capture the abundant directional information of images, the contourlet transform is proposed to solve this weakness. However, the contourlet transform suffers from shift variance; therefore, the nonsubsampled contour transformation (NSCT) has been proposed. NSCT is a flexible and fully shift-invariant model with implementation so fast that it is widely used in IVIF tasks [44], [45]. Meng et al. [44] proposed an IVIF method based on NSCT and object region detection; this method first obtains the target region from the infrared image, then combines the sub-images decomposed by NSCT to obtain the final fusion image with good retention of targets and details. Zhao et al. [45] have applied NSCT and pixel information estimation to improve fusion performance and reduce the computational burden. These algorithms perform efficiently in the decomposition task; however, they may cause distortions of the fused image, produce artifacts, or lose edges due to ignoring the spatial consistency during the fusion processing.

D. EDGE-PRESERVING FILTER-BASED

The edge-preserving filter has been widely used in IVIF as an effective decomposition tool. As the name suggests, it preserves edges and retains the spatial consistency of texture and reduces artifacts. Yang et al. [12] proposed a multi-scale decomposition method based on a rolling guided filter and a fast bilateral filter to decompose the input images into the base and detail layers and then used two fusion rules—the

sparse representation-based and a detail injection model—to obtain the fused base and detail layers for further processing. Kumar [46] proposed a cross bilateral filter to decompose the source image into several sub-images, calculate its weight, then combine the ''weighted average'' to obtain the fusion image. Hu et al. [47] proposed an IVIF method that uses a guided filter to decompose the input images to obtain two sub-layers and combines the cumulative distribution of gray levels and the entropy to preserve infrared targets and visible textures adaptively. Tan et al. [48] proposed a multi-level Gaussian curvature filtering method for decomposing the original images into three types of layers, small-large scales and base layers, followed by three fusion measures wherein max-value measure, integrated measure, and energy-based measure are used to combine these layers; the experimental results demonstrate the superiority and effectiveness of the algorithm. Chen et al. [49] proposed an IVIF method based on a rolling guided filter and multi-directional filter banks, where the former filter is used to separate the source image into base layers and detail layers whereas the latter filter is applied to fuse the base layers; this combination achieves better fusion performance.

III. THE PROPOSED METHOD

The proposed IVIF method consists of three parts: the proposed decomposition filter, the fusion rules (weighted average and detail enhancement model), and the reconstruction algorithm shown in Figure. 1. First, the input images are decomposed by the proposed globalet filter into base (B_{ir}, B_{vis}) and detail layers (D_{ir}, D_{vis}). According to the characteristics of the different layers, a weighted average fusion rule is introduced to deal with the base layers. A detail enhancement model is proposed to guide the detail layers, with the goal of obtaining two fused sub-layers (FB, FD). Finally, an addition operation is used to reconstruct the fused image. The following sections will describe the construction process of the proposed globalet filter and the IVIF method based on this filter.

A. THE GLOBALET FILTER

Given an input image, we want to decompose the image into sub-images with different information from a global perspective. Three goals are considered: there should be high brightness, noise should be eliminated, and texture details should be preserved to the greatest extent possible. To achieve the first goal, the L_2-norm is introduced between the output image and the input image to make the squared error measure (E_1) as small as possible. In other words, E_1 is designed to ensure that the brightness distribution of the output image is as similar as possible to the input image, and it is defined as follows:

$$E_1 = \sum_i \left(\|O_i - I_i\|_2^2 \right), \tag{1}$$

where I denote the input image, O is the output image, i is the pixel coordinates in the form of (i_x, i_y), and $\|*\|_2$ represents the L_2-norm.

As we all know, noise is high-frequency information that typically appears in regions where pixel values vary in an extensive range. When the brightness distribution is as similar as possible to the source image, it means that the original noise is also preserved, which violates the original intention of designing a filter. To make the noise as small as possible or even zero, the noise limit error measure E_2 is defined as:

$$E_2 = \sum_i \left(|\nabla_x O_i|^2 + \left|\nabla_y O_i\right|^2 \right), \tag{2}$$

where ∇_x and ∇_y are the first-order difference in the x and y directions, respectively.

When both luminance information and noise are limited, we also want the texture detail information to be preserved to the greatest extent. In general, a visible source image is widely used as a reference object due to contained rich texture details. However, in this paper, we also take the infrared source image into account to further absorb the indispensable infrared details. The texture detail error measure E_3 is expressed as:

$$E_3 = \sum_i \left(\frac{1}{2} \left(O_i^*\right)^2 + \lambda \cdot \alpha_i \cdot \left(O_i^* - O_i\right)^2 \right), \tag{3}$$

where λ represents the texture fidelity factor, which is used to balance the degree of detail preservation of the two source images. O^* is the target result, and α_i is the spatial-varying weight, which is defined as follows:

$$\alpha_i = \left(\left| \sum\nolimits_{j \in \omega_i} I_j \right| + \varepsilon \right)^{-1}, \tag{4}$$

where ε is a small constant (typically 10^{-6}) preventing division by zero. ω_i is a square window centered at pixel i, $j \in \{x, y\}$ denote a pixel coordinate belonging to the ω_i.

By combining Eqs. (1)-(3), the final filtered image O^r can be defined as Equation (5) while minimizing errors:

$$\begin{aligned} O^r &= arg \min_0 E_1 + argE_2 + argE_3, \\ &= arg \min_0 \sum_i \left\{ \|O_i - I_i\|_2^2 + \left(|\nabla_x O_i|^2 + \left|\nabla_y O_i\right|^2 \right) \right\} \\ &\quad + arg \min_0 \sum_i \left\{ \frac{1}{2} \left(O_i^*\right)^2 + \lambda \cdot \alpha_i \cdot \left(O_i^* - O_i\right)^2 \right\}. \end{aligned} \tag{5}$$

In this function, we must pay special attention to the noise limit error measure E_2. In general, mild noise is unlike strong noise surrounding the edge region, which is easily distinguished due to its large gradient. Mild noise and texture details are highly similar. Once they are indiscriminately removed, it is difficult to determine whether noise or texture details are removed, leading to unsatisfactory edge-preserving properties of the filter. Fortunately, there is a characteristic that edge gradients are larger than gradients for mild noise or texture details. With the aid of this characteristic, a penalty term $P_{i,j}$ based on the threshold function

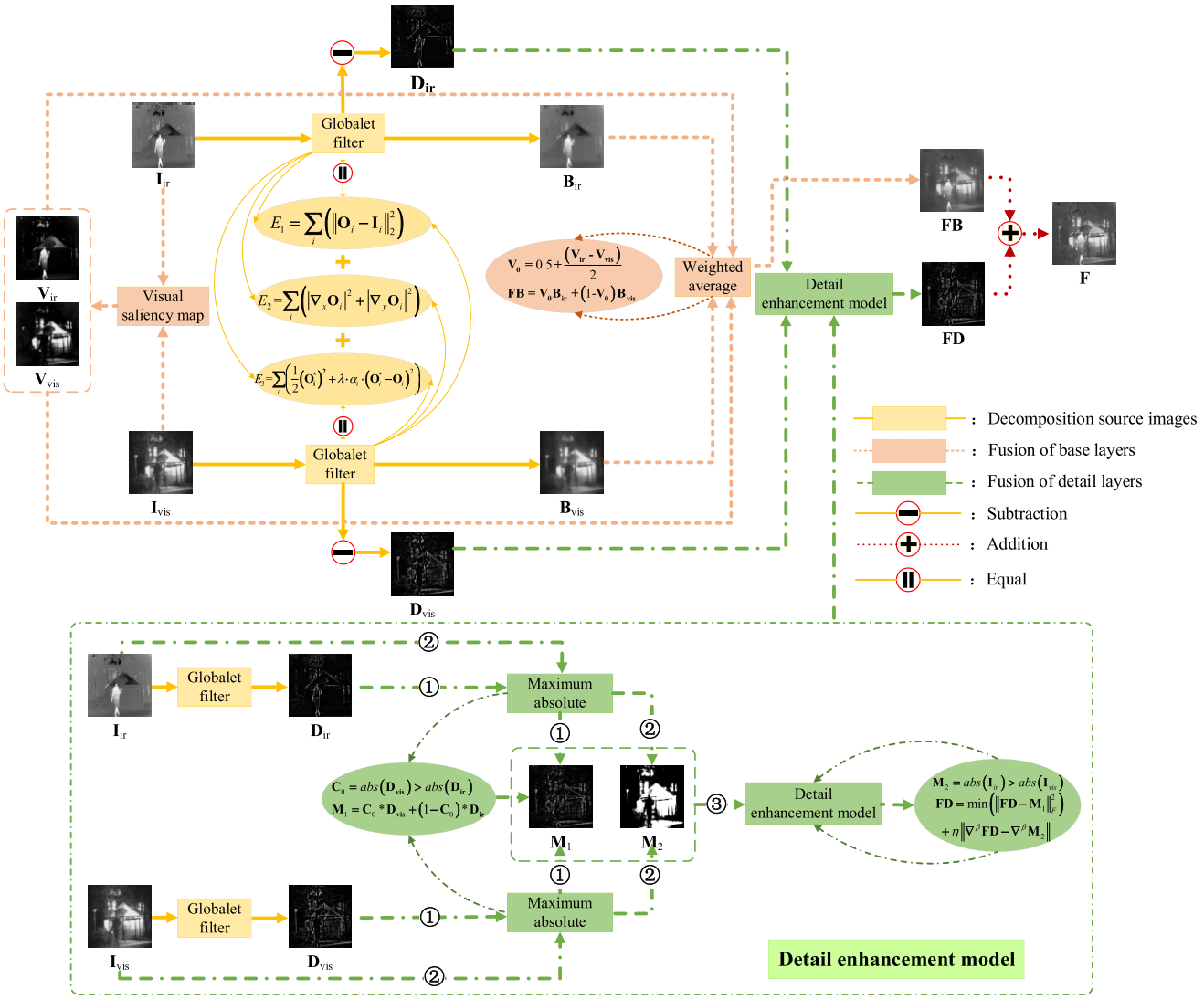


FIGURE 1. Flowchart of the IVIF method based on the globalet filter. Different line formats represent different sections.

TF is designed, with the goal of removing mild noise while preserving texture details. Based on the above considerations, $P_{i,j}$ can be expressed as:

$$P_{i,j} = \mu \cdot TF\left(\nabla_j O_i, n, \beta\right), \tag{6}$$

where μ denotes the maximum penalty coefficient of $P_{i,j}$, ranging from 0 to μ. n is the slope of TF, β represents the gradient threshold, and TF is defined as:

$$TF\left(\nabla_j O_i, n, \beta\right) = 1 - \left(1 + \left(\frac{\beta}{\nabla_j O_i}\right)^n\right)^{-1}. \tag{7}$$

By observing Eqs. (6) and (7), it can be seen that only $\nabla_j O_i$ is variable while the others are constants. Based on these two equations, we draw a trend graph, as shown in Figure. 2. From the trend chart, the following rules can be concluded:

a) when $|\nabla_j O_i| < \beta$, $P_{i,j}$ become larger, and maximum $= \mu$;

b) when $|\nabla_j O_i| > \beta$, $P_{i,j}$ become smaller, and minimum $= 0$;

c) when $|\nabla_j O_i|$ is closer to β, $P_{i,j}$ and $|\nabla_j O_i|$ are inversely related. In other words, $P_{i,j}$ decreases as $|\nabla_j O_i|$ increases, and the magnitude of this decrease is controlled by n. The larger n is, the steeper TF is, and the closer it is to the step function (SF). In pace with n, it continues to approach positive infinity, $TF = SF$. However, this is an ideal situation because neither state can completely and clearly separate (strong or mild) noise and texture details, as they can only achieve this with maximum probability instead.

Finally, we integrate $P_{i,j}$ to rewrite Eq. (5) as follows:

$$\begin{aligned} O^r &= arg\min_O E_1 + arg\min_O E_2 + arg\min_{O^*} E_3 \\ &= arg\min_O \underbrace{\sum\nolimits_i \left\{\|O_i - I_i\|_2^2 + P_{i,j} \cdot \left(|\nabla_x O_i|^2 + |\nabla_y O_i|^2\right)\right\}}_{M} \end{aligned}$$

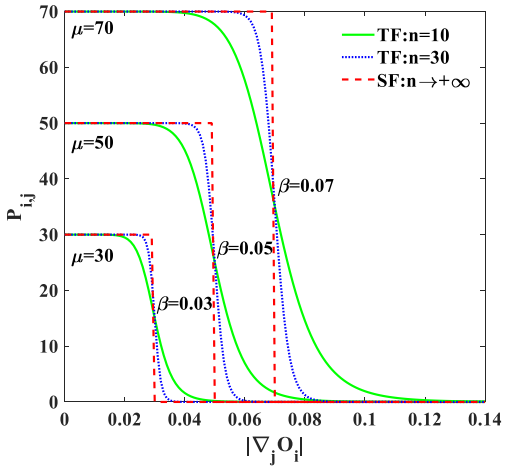


FIGURE 2. Trend graph between $P_{i,j}$ and $|\nabla_j O_i|$.

$$+ \underbrace{arg\min_{O^*} \sum_i \left\{ \frac{1}{2} \left(O_i^*\right)^2 + \lambda \cdot \alpha_i \cdot \left(O_i^* - O_i\right)^2 \right\}}_{N}. \quad (8)$$

According to the difference in the target outputs, Eq. (8) is divided into two terms (M, N) to solve it. For the M term, we subdivide $P_{i,j}$ into x and y directions, and the new formula is as follows:

$$arg\min_{O} \sum_i \left\{ \|O_i - I_i\|_2^2 + P_{i,x} \cdot |\nabla_x O_i|^2 + P_{i,y} \cdot \left|\nabla_y O_i\right|^2 \right\}, \quad (9)$$

where $P_{i,x}$ and $P_{i,y}$ are the penalty metrics in the x and y directions, respectively.

Based on the optimization theory, Eq. (9) can be converted into matrix form:

$$(O - I)^T (O - I) + O^T \left(D_x^T P_x D_x + D_y^T P_y D_y\right) O, \quad (10)$$

where O and I are represented as a column vector. T is the transpose symbol. D_x and D_y are the first-order forward difference operators in the x and y directions, respectively. P_x and P_y are diagonal matrices whose diagonal column vectors follow the form of the column vectors of the $P_{i,x}$ and $P_{i,y}$, respectively.

The final O is obtained by minimizing Eq. (10), that is, by taking the first derivative with respect to O in Eq. (10) and making it equal to zero. The derivation result is as follows:

$$O = \left\{E + \left(D_x^T P_x D_x + D_y^T P_y D_y\right)\right\}^{-1} \cdot I, \quad (11)$$

where E is the identity matrix and $D_x^T P_x D_x + D_y^T P_y D_y$ is a Laplacian homogeneous matrix.

Similarly, the N term is also rewritten in matrix form for minimization:

$$\frac{1}{2} \cdot \left(O^*\right)^T O^* + \lambda \cdot A \cdot \left(O^* - O\right)^T \left(O^* - O\right) \quad (12)$$

where O^* and O are represented as a column vector. A is a diagonal matrix containing the weights α_i of all pixels.

The target result O^* can be obtained by minimizing the Eq. (13) as follows:

$$\frac{1}{2} \cdot 2O^* + \lambda \cdot \left(A^T + A\right)\left(O^* - O\right) = 0, \quad (13)$$

where A^T is the transposed expression of A.

Rearranging Eq. (13), we obtain:

$$O^* + \left[\lambda \cdot \left(A^T + A\right)\right] \cdot O^* = \lambda \cdot \left(A^T + A\right) \cdot O, \quad (14)$$

$$\left[E + \lambda \cdot \left(A^T + A\right)\right] \cdot O^* = \lambda \cdot \left(A^T + A\right) \cdot O, \quad (15)$$

$$O^* = \left[E + \lambda \cdot \left(A^T + A\right)\right]^{-1} \cdot \lambda \cdot \left(A^T + A\right) \cdot O. \quad (16)$$

Combining the above derivation process, we conclude that the solution of Eq. (8) becomes the addition of the results of Eq. (12) and Eq. (17) as follows:

$$O^r = O + O^*. \quad (17)$$

To make the designed filter iterative, we respectively rewrite Eq. (11) and Eq. (16) as follows:

$$O^{t+1} = \left\{E + \left(D_x^T P_x^t D_x + D_y^T P_y^t D_y\right)\right\}^{-1} \cdot I, \quad (18)$$

$$\left(O^*\right)^{t+1} = \left[E + \lambda \cdot \left(A^T + A\right)\right]^{-1} \cdot \lambda \cdot \left(A^T + A\right) \cdot O^{t+1}, \quad (19)$$

where t denotes the iteration index.

Finally, the designed filter is named the globalet filter, and the mathematical expression is $GLF\ (I, \beta, \mu, t, n)$.

B. IVIF BASED ON GLOBALET FILTER

1) IMAGE DECOMPOSITION

After constructing the filter, it can be used as a decomposition tool to decompose the source images (I_X) to obtain the corresponding sublayers, as shown in Figure. 1. The base layers (B_X) are calculated as follows:

$$B_X = GLF\left(I_X, \beta, \mu, t, n\right), \quad (20)$$

where $X \in \{ir, vis\}$, and where ir and vis denote infrared image and visible image, respectively.

Further, the detail layers (D_X) are obtained by subtracting the base layers from the original images as follows:

$$D_X = I_X - B_X. \quad (21)$$

As a result, several sub-layers representing different types of information are obtained, and the next step is to design fusion rules to fuse them.

2) FUSION RULES

After obtaining the base layers and detail layers, it is extremely challenging to construct suitable and efficient fusion rules to fuse the two types of sub-layers.

a: FUSION OF BASE LAYERS

The base layers reflect the overall brightness distribution and spatial structure information of the source images. Therefore, it is reasonable to apply a weighted average operator to adaptively select the image feature from the source images to fuse the base layer, where the visual saliency map (*VSM*) [50] is introduced to calculate the weight V_0 as follows:

$$V_0 = 0.5 + \frac{V_{ir} - V_{vis}}{2}, \tag{22}$$

where V_{ir} and V_{vis} are the *VSMs* for an infrared image and a visible image, respectively. Moreover, for V_0, it needs to conform to:

$$V_0 = 0.5 + \frac{Y}{2} \begin{cases} Y = V_{ir}, & if\ V_{ir} - V_{vis} > 0. \\ Y = 0, & if\ V_{ir} - V_{vis} = 0. \\ Y = V_{vis}, & if\ V_{ir} - V_{vis} < 0. \end{cases} \tag{23}$$

where Eq. (23) indicates that the *VSMs* control the fusion performance of the base layers. If V_{ir} is greater than V_{vis}, it means that more infrared information is transmitted into V_0 and then fused into the fused base image (*FB*), and vice versa. When V_{ir} is equal to V_{vis}, which is the general weighted average rule,

$$FB = V_0 B_{ir} + (1 - V_0) B_{vis}, \tag{24}$$

b: FUSION OF DETAIL LAYERS

Compared to the base layers, the detail layers reflect the small gradient variations of source images. To highlight more detailed information, a detail enhancement model is proposed, the specific steps of which are as follows.

In the beginning, the pre-fused image M_1 is obtained by sequentially following three operators: ''maximum absolute'', ''Gaussian smoothing'', and ''weighted average''. These steps can be respectively expressed as:

$$C_0 = abs\left(D_{vis}\right) > abs\left(D_{ir}\right), \tag{25}$$

$$C_0' = Gaussian\left(C_0, \sigma_0\right), \tag{26}$$

$$M_1 = C_0' * D_{vis} + \left(1 - C_0'\right) * D_{ir}.\mathrm{d} \tag{27}$$

where standard deviation σ_0 is empirically set to 2.

Afterward, the mathematical expression of *DEM* is constructed as follows:

$$\min_{FD} \|FD - M_1\|_F^2 + \eta \left\|\nabla^\delta FD - \nabla^\delta M_2\right\|_F^2, \tag{28}$$

$$M_2 = abs\left(I_{vis}\right) > abs\left(I_{ir}\right), \tag{29}$$

where M_2 denotes the binary image through the ''maximum absolute'' operator between I_{vis} and I_{ir}. *FD* is the fused detail image, η denotes the trade-off parameter between the two terms, and ∇^δ is the discrete fractional-order gradient operator. $\left[\nabla^\delta M_2\right]_k = \left(\left[D_x^\delta M_2\right]_k, \left[D_y^\delta M_2\right]_k\right)^T \in R^2$($R$ denote set of real numbers), wherein $\left[D_x^\delta M_2\right]_k$ and $\left[D_y^\delta M_2\right]_k$ denote the k-th column discrete fractional-order gradient matrixes in the x and y directions [51]. Empirically, δ is set to 1.2.

Algorithm 1 IVIF method based on the globalet filter

Input: source images $\boldsymbol{I}$
Output: Fused image $\boldsymbol{F}$
Initialize: $t = 1$, $\beta = 0.08$, $\mu = 100$, $n = 30$, $\eta = 10^{-4}$, $\lambda = 10^{-4}$, $\omega = 13 * 13$, $\sigma_0 = 2$, $\delta = 1.2$
Globalet filter: $\boldsymbol{GLF(I, \beta, \mu, t, n)}$
  For i = 1: t
    (1) Construct objective function using Eq. (5).
    (2) Calculate $\boldsymbol{P_x}$ and $\boldsymbol{P_y}$ using Eqs. (6)-(7) to obtain $\boldsymbol{O}$ in Eq. (11).
    (3) Obtain $\boldsymbol{O^*}$ in Eq. (12) using Eqs. (13)-(16).
    (4) Obtain target result $\boldsymbol{O^r}$ using Eq. (17).
  End For
Decomposition:
    (1) Obtain the base layers $\boldsymbol{B_X}$ using Eq. (20).
    (2) Obtain the detail layers $\boldsymbol{D_X}$ using Eq. (21).
Fusion:
    (1) Calculate the $\boldsymbol{VSMs}$ using Eqs. (22)-(23).
    (2) Obtain the fused base layer (***FB***) using Eq. (24).
    (3) Obtain the pre-fused image $\boldsymbol{M_1}$ using Eqs. (25)-(27).
    (4) Obtain the fused detail layer (***FD***) using Eqs. (28)-(31).
Reconstruction:
    (1) Obtain the final fused image (***F***) using Eq. (32).

FIGURE 3. IVIF method based on the globalet filter.

The purpose of designing *DEM* is to create connections between the detail layers and the input images (M_1, M_2) so that the *FD* contains the best pixel region from these images to the greatest extent possible after the ''maximum absolute'' rule. This model consists of two constraint terms, where the first term forces *FD* to contain similar detail information of M_1, and the second term is a global structure fidelity term, which aims to constrains the fractional-order global structure feature of *FD* to be consistent with that of the M_2 image.

Eq. (28) is actually a typical convex optimization problem that can be solved by the sliding window technique and the gradient descent algorithm. It can be written in matrix form as follows:

$$FD = (E + \eta B)^{-1} (M_1 + \eta B M_2), \tag{30}$$

$$B = \left[D_x^\delta\right]^T D_x^\delta + \left[D_y^\delta\right]^T D_y^\delta, \tag{31}$$

where E is the identity matrix. *FD*, M_1, and M_2 denote the column vectorization when they are calculated.

3) IMAGE RECONSTRUCTION

As discussed above, the final fused image, F, can be obtained by summing *FB* and *FD* as follows:

$$F = FB + FD. \tag{32}$$

We summarize the IVIF method based on the GLF filter in Algorithm 1 in Figure. 3.

IV. EXPERIMENTS

A. EXPERIMENTAL SETTING

This section conducts a large number of experiments to verify the effectiveness of the proposed method, and these experiments typically use the TNO datasets [52] and the

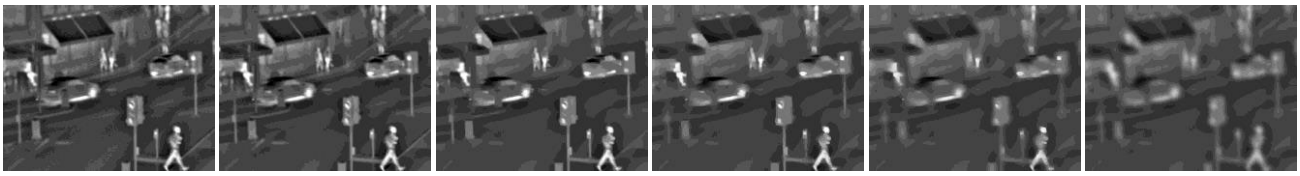


FIGURE 4. Filtered results when β is set to 0, 0.01, 0.03, 0.05, 0.08, and 1, where $n = 30$, $\mu = 50$.

RoadScene real dataset [53]. Meanwhile, nine state-of-the-art fusion methods are used for comparison with the proposed method, followed by six metrics assisting in validation. These methods include the VGG-19 and multi-layer fusion strategy-based method (VggML) [54], ResNet and zero-phase component analysis (Resnet50) [55], coupled feature learning (CFL) [56], Bayesian fusion (BayF) [57], algorithm unrolling image fusion (AUIF) [58], classification saliency-based (CSF) [59], dual-discriminator conditional generative adversarial network (DDcGAN) [60], detail injection model (DIM) [12], and semantic-aware real-time fusion network (SeAFusion) [61]. Six indicators are also introduced: edge-based metric (Q_{abf}) [62], structure-based metric (SSIM) [63], multiscale feature-based metric (Q_m) [64], phase congruency-based metric (Q_P) [65], human perception-based metric (Q_{CB}) [66] and feature mutual information (FMI_w) [67]. Higher values of these indicators indicate better fusion results. We also select eight pairs of source images, comprising six—"Bunker," "Kaptein_1654(K_1654)," "Jeep," "Sand_path," "Road," "Kaptein_1123(K_1123),"—from the TNO dataset and two—"FLIR_04215," "FLIR_04726"—from the RoadScene dataset, to observe the subjective results.

B. PARAMETERS DISCUSSION

There are several parameters that need to be discussed to obtain the optimal fusion performance. Five parameters (gradient threshold β, maximum penalty coefficient μ, threshold function slope n, texture fidelity factor λ, and sliding window size ω) are generated when designing the filter and the trade-off parameter η in *DEM*, respectively. It is noted that the bolded value represents the best, and the underlined value indicates the second best.

1) GRADIENT THRESHOLD β

Eqs. (6)-(7) show the relationship between the three parameters β, μ, and n: one of them changes whereas the other two remain unchanged. To further illustrate the effects of β on the filtering results, we take an example, and the results are shown in Figure. 4 when $n = 30$, $\mu = 50$, and β is set to 0, 0.01, 0.03, 0.05, 0.08, and 1, respectively. As β increases, an increasing number of large-scale gradients are removed, and the filtered image becomes smoother. When $\beta = 0.08$, gradient removal reaches a certain level, and the edges are preserved well.

2) MAXIMUM PENALTY COEFFICIENT μ, THRESHOLD FUNCTION SLOPE n

As we know from Eqs. (6)-(7), the maximum penalty coefficient μ controls the smoothness of large-scale gradients, and the threshold function slope n controls how much the penalty coefficient drops around the gradient threshold β. The larger the μ, the higher the smoothness; the larger the n, the more obvious the drop, and the clearer the distinction between texture details and mild noise.

Figure. 5 shows the filtered results for μ ranging from 10 to 500 and for n ranging from 0 to 50 when the gradient threshold β is 0.08. From left to right, the larger μ is, the more large-scale gradients (mild noise) are removed, and the smoother the filtering results; from top to bottom, the larger n is, the more obvious the preservation of the edge and texture details. However, as n increases from 0 to 10 and μ increases from 10 to 500, the edges of the filtering results are lost, which can be verified from the filtered images in the first three rows in Table 1. When n reaches 30 or 50 and the range of μ is unchanged, the filtering effect is visually indistinguishable. In this case, we employ the average of objective indicator values to assist in the distinction of six pairs of source images from the TNO dataset, and the results are listed in Table 1. It should be noted that there are four exceptions ($n = 30$, $\mu = 10$ or 500, and $n = 50$, $\mu = 10$ or 500) because their subjective filtering effect is either too low or too high. From these results, we can observe that the filtering performance is the best at $n = 30$, $\mu = 100$ when $\beta = 0.08$.

3) TEXTURE FIDELITY FACTOR λ AND SLIDING WINDOW SIZE ω

Eqs. (3)-(4) constitute the texture detail error E_3, where parameter λ is used to balance the degree of detail preservation of the two source images, and parameter ω controls how many irrelevant infrared details are removed. To identify the best value, we still average the six metrics, and the results are presented in Tables 2 and 3. According to these two tables, we can conclude that the filtering effect is best when λ is set to 10^{-4} and ω is set to $13 * 13$.

4) TRADE-OFF PARAMETER η

Eq. (28) is the main component of the detail enhancement model, wherein the trade-off parameter η that controls the proportion of image information absorbed from M_1 and M_2 into FD needs to be discussed. Table 4 lists the results of

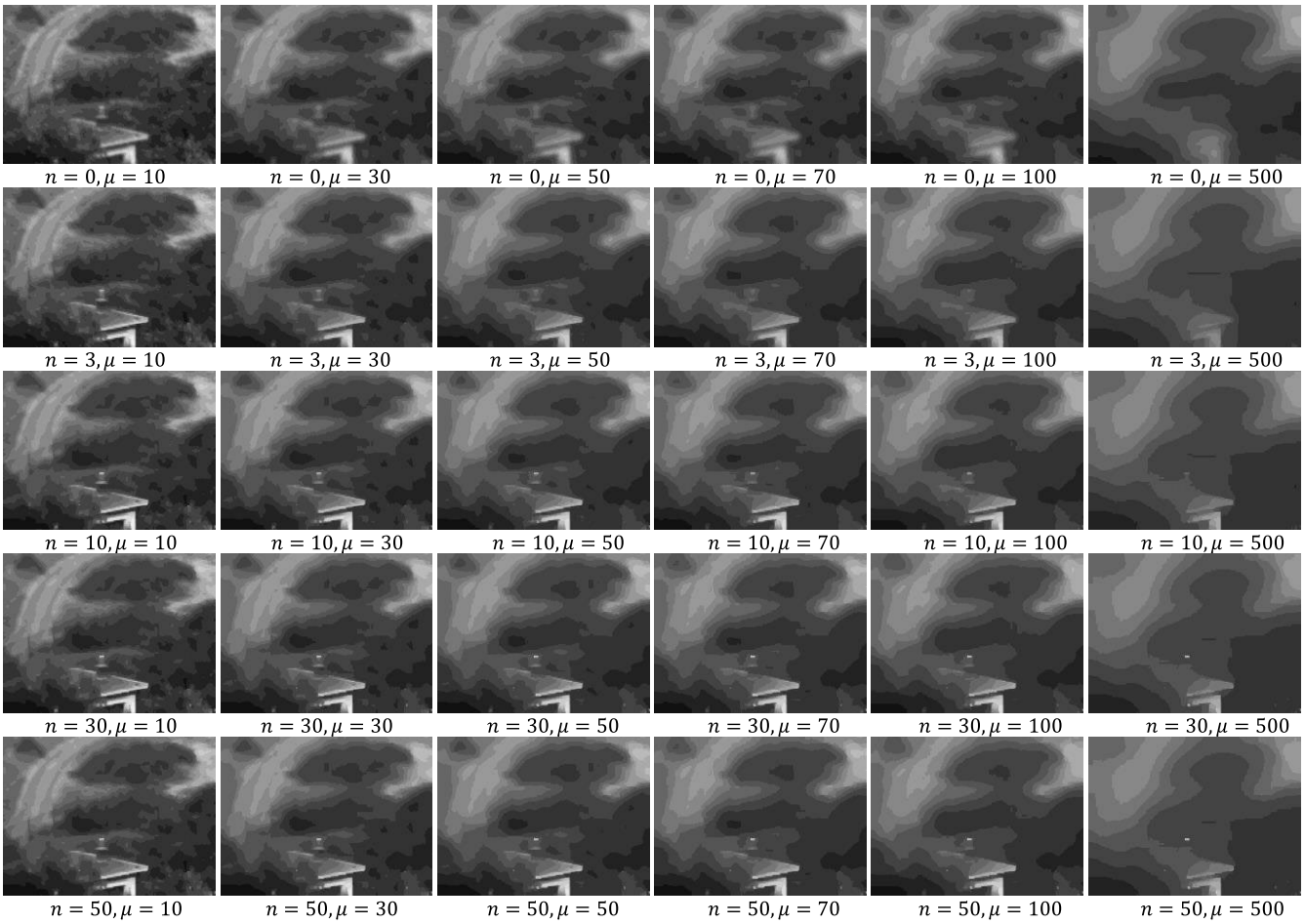


FIGURE 5. Filtered results for μ ranging from 10 to 500 and n ranging from 0 to 50 when gradient threshold β is set to 0.08.

TABLE 1. Averages of the six metrics for the different n and μ on six pairs of source images from TNO dataset.

Images	Dimensions	Metrics	$n = 30$ $\mu = 30$	$n = 30$ $\mu = 50$	$n = 30$ $\mu = 70$	$n = 30$ $\mu = 100$	$n = 50$ $\mu = 30$	$n = 50$ $\mu = 50$	$n = 50$ $\mu = 70$	$n = 50$ $\mu = 100$
Bunker	768*576	Qabf	0.5220	0.5272	0.5294	**0.5306**	0.5221	0.5268	0.5293	0.5303
Road	632*496	SSIM	0.8191	0.8231	0.8249	**0.8263**	0.8191	0.8228	0.8248	0.8261
K_1654	620*450	Q_m	0.6782	0.6884	0.6932	**0.6996**	0.6769	0.6856	0.6910	0.6971
Jeep	917*678	Q_P	0.3864	**0.3888**	0.3884	0.3873	0.3859	0.3878	0.3879	0.3865
Sand-path	575*475	Q_{CB}	0.5092	0.5146	0.5172	**0.5186**	0.5088	0.5143	0.5169	0.5185
K_1123	620*450	FMI_w	0.4177	0.4189	0.4195	**0.4204**	0.4176	0.4185	0.4193	0.4200

the indicators, and we can see that the fused performance is optimal when η is equal to 10^{-4}.

C. EXPERIMENTS ON DATASETS

Figure. 6 shows the fused results of six pairs of input images from the TNO dataset and two image pairs from the RoadScene dataset with ten fusion methods. Based on these results, we analyze it from three perspectives: the brightness of the infrared target, the preservation of visible details, and the overall contrast. In terms of brightness, we can observe that the infrared targets of the five methods are dim based on VggML, Resnet50, BayF, CSF, and AUIF because of insufficient infrared information extraction whereas the infrared target obtained by our method is highly consistent with the source image; these targets can be verified by the red regions in Figure. 6. Although the brightness of infrared target based on the three methods of DDcGAN, SeAFusion, and CFL is significantly improved, there are some problems with the visible details: the image details of DDcGAN are largely sharpened, resulting in over-exposed fused images; the visible details of SeAFusion and CFL are partially lost, and these can be observed from the blue regions. After a comprehensive analysis, the target brightness, and visible details of DIM-based and our method are considerable

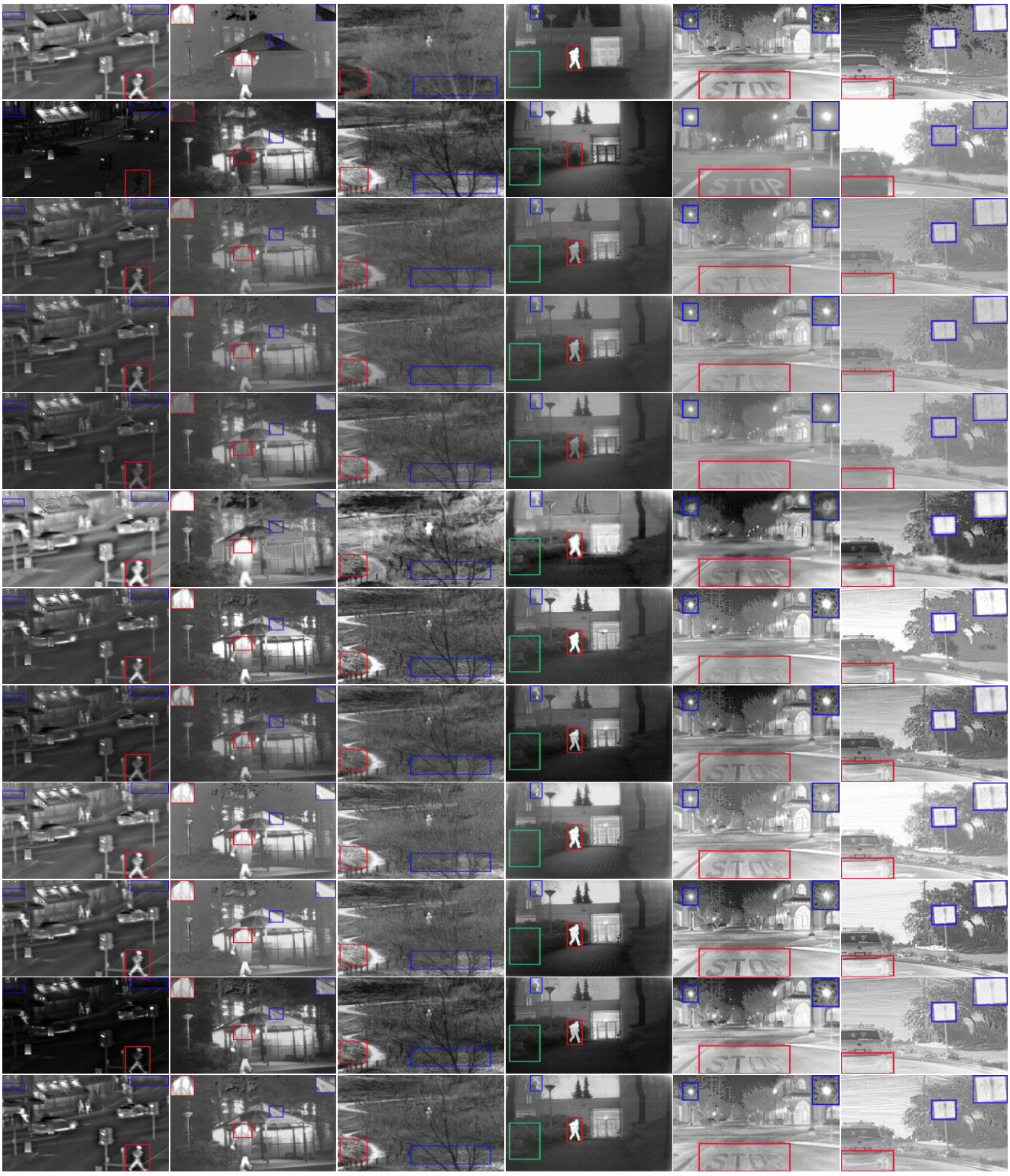


FIGURE 6. Subjective result images of ten fusion methods. From top to bottom in order: infrared images (IR), visible images (VIS), VggML, Resnet50, BayF, DDcGAN, DIM, CSF, CFL, SeAFusion, AUIF, and our method (Ours). From left to right in order: "Road", "Kaptein_1654", "Sand-path", "Kaptein_1123", "FLIR_04215", "FLIR_04726". The red, blue, and green boxes show the enlarged local regions.

overall; however, we find that the DIM-based method has flaws with careful observation, such as the overexposed target object on the "Bunker" image and the missing smoke on the "Kaptein_1123" image. By contrast, the proposed method can not only better maintain the prominent target, *e.g.,* the persons in "Road" and "Kaptein_1123", but it can also

TABLE 2. Averages of the six metrics for the different λ on six pairs of source images from TNO dataset.

Images	Dimensions	Metrics	λ			
			$= 10^{-1}$	$= 10^{-2}$	$= 10^{-3}$	$= 10^{-4}$
Bunker	768*576	Qabf	0.4883	0.5016	0.5306	**0.5321**
Road	632*496	SSIM	0.7859	0.8018	0.8263	**0.8281**
K_1654	620*450	Q_m	0.6077	0.6712	0.6996	**0.7017**
Jeep	917*678	Q_P	0.3067	0.3395	0.3873	**0.3898**
Sand-path	575*475	Q_{CB}	0.4579	0.5061	**0.5186**	0.5181
K_1123	620*450	FMI_w	0.4119	0.4204	0.4204	**0.4207**

preserve the visible information together, which can also make the image contrast conform to the human visual system.

For further convincing comparison, two groups of image pairs, called "FLIR_04215" and "FLIR_04726" from the RoadScene dataset, are tested, as shown in the last two columns of Figure. 6. By observing the fusion results in column of Figure. 6 carefully, we can see that the infrared brightness based on VggML, Resnet50, DIM, and BayF is dim, resulting in the letters in the red rectangle being indistinguishable from the background; that the DDcGAN-based method produces saturated infrared information and also introduces noise, thus making the fused image interweaved in black and white; that the CSF-based and CFL-based methods improve the shortcoming of the first five methods but lose details, such as street lights and tree branches in blue areas, which become blurred; AUIF-based, SeAFusion-based, and our method look natural, but a closer observation reveals that the sky region of the first two methods is black and that the sky of the source image is almost gray, which means that the fusion image exceeds the pixel value of the source image. By contrast, our method has higher image quality due to the prominent infrared targets and well-preserved visible details. Similar comparison directions are applied in column 8 of Figure. 6. The infrared information extraction based on VggML, Resnet50, and BayF is insufficient; the tail of the car (red rectangle area) based on the three methods AUIF, DDcGAN, and CFL produces artifacts and even becomes blurred. The arrows on the sign obtained based on the three fusion methods of CSF, SeAFusion, and DIM are missing to a certain extent, as can be verified from the magnified blue area. In conclusion, our method perfectly overcomes the shortcomings of the methods mentioned above and obtains the best image performance—prominent infrared targets, abundant visible details and natural contrast.

Aside from subjective visual observations, we quantitatively evaluate the fusion performance of the ten methods by using six objective metrics with 60 groups of images, including 30 groups of images from the TNO dataset and 30 pairs of images from the RoadScene dataset, and the results are presented in Tables 5-6. Table 5 demonstrates that our method achieves the best average values of Qabf, SSIM, Q_P, Q_{CB}, and FMI_w along with the second-best value of Q_m; Table 6 manifests that proposed method performs well as it achieves the four highest averages and the two second highest averages among the six metrics. The excellent performance of our method on both datasets proves that it is robust and outperforms other state-of-the-art methods.

D. DISCUSSION AND ANALYSIS

1) ITERATION NUMBER

Most decomposition methods are well known to be iterative, the number of iterations is generally four or five, and it is recognized that more decomposition layers lead to more detail on the extracted image information as well as better fusion quality [12], [36], [46], [47]. Therefore, the proposed globalet filter is also iterative, but the filtering effect, or even the fusion quality, may have little to do with the number of iterations. The main reasons are as follows: i) In the solution of the globalet filter, the number of iterations t is generally an irrelevant variable, or even a constant, while the three variables μ, β, and n are related to the solution. ii) There is a special case, namely the maximization or minimization of the μ, where the increase of t will affect the filtering results and fusion results. As μ is close to the maximum, the curve resembles a straight line, meaning that the magnitude varies little or remains almost constant, thus resulting in an almost constant penalty coefficient and threshold function, which greatly reduces the removal of large-scale gradients. In this case, the increase of t will extract the image information that was not originally obtained into the corresponding decomposition layer in turn, so as to ensure the integrity of the information as much as possible, and vice versa.

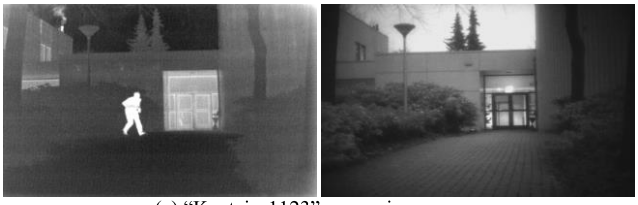
Considering the discussions in the above subsections have discussed the value of μ and reduced the running time, t was set to one in this experiment.

2) COMPARISON BETWEEN DIFFERENT DECOMPOSITION METHODS

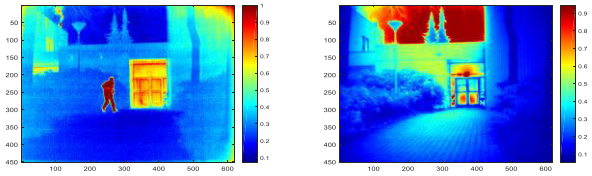
To verify the effectiveness of the decomposition method proposed in this paper, we have selected four recognized fusion algorithms of hybrid decomposition filters for comparison: HMSD (Gaussian and bilateral filters) [36], DIM (rolling guided filter and fast bilateral filter) [12], MLGCF (multi-level Gaussian curvature filtering) [48], and RGF-MDFB (rolling guided filter and multi-directional filter banks) [49]. The comparison results are shown in Figure. 7. Based on the filtering results, as can be seen in Figure. 7(c)-(d), we can observe that the filtering degrees of HMSD and MLGCF are too low, resulting in a lack of necessary detail information and the introduction of a certain amount of noise in the fused results; this can be proven by the fused images in the corresponding position in Figure. 7(e) (see the red and magnified blue areas). Conversely, DIM and RGF-MDFB are filtered so much that even infrared objects and visible details are completely filtered, thus causing reduced edge preservation along with the appearance of artifacts and noises at the edges. For example, the image information in the magnified blue area in RGF-MDFB can hardly show where

TABLE 3. Averages of the six metrics for the different ω on six pairs of source images from TNO dataset.

Images	Dimensions	Metrics	$\omega = 3*3$	$\omega = 5*5$	$\omega = 7*7$	$\omega = 9*9$	$\omega = 11*11$	$\omega = 13*13$
Bunker	768*576	Qabf	0.5318	0.5320	0.5321	0.5322	0.5323	**0.5325**
Road	632*496	SSIM	0.8278	0.8280	0.8281	0.8281	0.8281	**0.8282**
K_1654	620*450	Q_m	0.7016	0.7015	0.7017	0.7019	0.7019	**0.7020**
Jeep	917*678	Q_P	0.3895	0.3897	0.3898	**0.3899**	**0.3899**	**0.3899**
Sand-path	575*475	Q_{CB}	**0.5181**	**0.5181**	**0.5181**	0.5180	0.5180	0.5179
K_1123	620*450	FMI_w	0.4206	0.4206	0.4207	0.4207	0.4208	**0.4222**



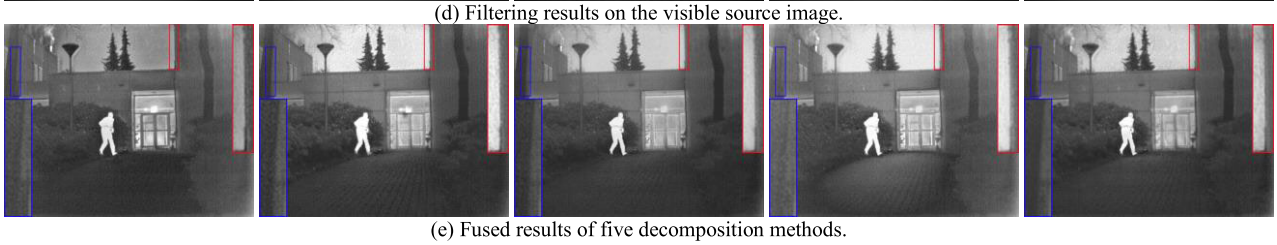
(a) "Kaptein-1123" source image



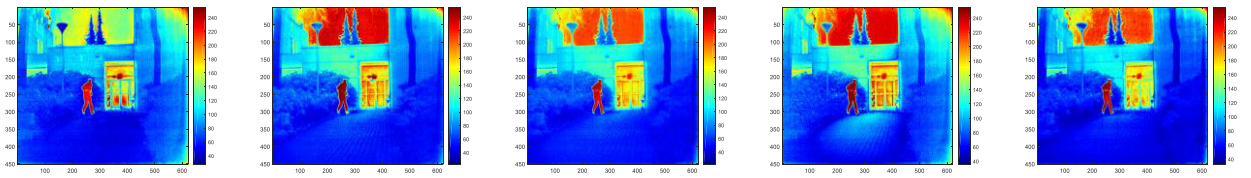
(b) Heatmaps of "Kaptein-1123"

(c) Filtering results on the infrared source image.

(d) Filtering results on the visible source image.



(e) Fused results of five decomposition methods.



(f) Heatmaps of fused results by five decomposition methods.

FIGURE 7. Illustrations of different decomposition methods. The decomposition methods from left to right of the second row to the fifth row are: HMSD, DIM, MLGCF, RGF-MDFB, and Ours.

the boundaries of objects are. By contrast, our decomposition method automatically determines the image region that needs to be filtered, maximally preserving edges, minimally introducing noises, and maintaining proper sharpening of details and edges. To further illustrate the superiority of the proposed decomposition method, Figure. 7(f) presents the heatmaps of the fusion results. From these heatmaps, we can observe that the color of the sky area of DIM and RGF is very red, which indicates that the absorption of infrared information is more sufficient, thus resulting in a certain degree of loss of visible details. Moreover, the color balance of our heatmap shows that the important information of the source images is transferred into the fusion image and has a natural contrast.

TABLE 4. Averages of the six metrics for the different η on six pairs of source images from TNO dataset.

Images	Dimensions	Metrics	η				
			$=10^{-1}$	$=10^{-2}$	$=10^{-3}$	$=10^{-4}$	$=10^{-5}$
Bunker	768*576	Qabf	0.4704	0.5266	0.5324	**0.5325**	**0.5325**
Road	632*496	SSIM	0.7272	0.8191	**0.8282**	**0.8282**	**0.8282**
K_1654	620*450	Q_m	0.5865	0.6763	0.7021	**0.7023**	0.7020
Jeep	917*678	Q_P	0.2984	0.3834	**0.3900**	**0.3900**	0.3897
Sand-path	575*475	Q_{CB}	0.5013	0.5200	**0.5180**	0.5179	0.5179
K_1123	620*450	FMI_w	0.3440	0.3979	0.4208	**0.4222**	**0.4222**

TABLE 5. Average of the six metrics on TNO dataset. The best and the second-best metrics are shown in bold and underline, respectively.

Methods	Qabf	SSIM	Q_m	Q_P	Q_{CB}	FMI_w
VggML	0.3632	0.6989	0.5758	0.3038	0.4869	0.3928
Resnet50	0.3526	0.6897	0.5572	0.3013	0.4848	0.3920
BayF	0.4212	0.7828	0.6340	0.3315	0.5128	0.4127
DDcGAN	0.3264	0.6191	0.4733	0.1485	0.4478	0.3998
DIM	0.4452	0.7578	0.5736	0.3030	0.5115	0.3717
CSF	0.3998	0.6741	0.5399	0.2713	0.5046	0.3048
CFL	0.4617	0.7916	**0.8019**	0.2798	0.4964	0.4062
SeAFusion	0.4478	0.7671	0.5926	0.3392	0.4931	0.3286
AUIF	0.4127	0.6893	0.6214	0.2999	0.5251	0.3981
Ours	**0.5134**	**0.8134**	0.7676	**0.3739**	**0.5274**	**0.4128**

TABLE 6. Average of the six metrics on Roadscene dataset. The best metrics are presented in bold and the second best metrics are underlined.

Methods	Qabf	SSIM	Q_m	Q_P	Q_{CB}	FMI_w
VggML	0.3759	0.6953	0.4488	0.3271	0.4906	0.3495
Resnet50	0.3590	0.6788	0.4130	0.3095	0.4925	0.3351
BayF	0.3102	0.6782	0.4349	0.3432	0.4977	0.3326
DDcGAN	0.2902	0.5385	0.3844	0.1948	0.4619	0.3350
DIM	0.4094	0.7211	0.3794	0.2825	0.4641	0.3783
CSF	0.4677	0.7362	0.4739	0.3439	0.5187	0.3674
CFL	0.4668	0.7533	0.5142	0.2877	0.4766	0.3858
SeAFusion	0.4412	0.7350	0.3987	**0.4141**	0.4849	0.3820
AUIF	0.5072	0.7614	0.4965	0.3613	0.4972	**0.4198**
Ours	**0.5079**	**0.7992**	**0.5526**	0.3683	**0.5194**	0.4012

3) OBJECT MEASUREMENT DISCUSSION

This section explores the positive impact of IVIF on object detection, which is an advanced computer vision task. A widely recognized detector, i.e., YOLOV4 [68], is used to obtain the detection results on the fused images with different methods, respectively. The detection number and mean average precision (mAP) are utilized to measure the detection performance, where mAP@0.7 indicate the mAP values at IoU thresholds of 0.7.

Figure. 8 illustrates the visualized example of our fusion algorithm for the RoadScene real dataset. In the ''FLIR_03952'' scene, the bus disappears in the visible image while it is detected in the infrared image. The detector is unable to identify the bus on most of the fused methods owing to messy background and lower contrast. On the contrary, our method fully integrates the semantic information from the original images while maintaining optimal luminance ratio and preserving abundant structural details. At the same time, the object measurement trend is shown in Figure. 9. Most of the fused images have a detection number of four for the target ''person'' with a mAP value larger than 98%. The detection numbers of DCGAN are two because there are overlaps and artifacts in the objects of the fused image, resulting in a small distance between objects so that the detector cannot distinguish objects well, and a MAP value of 94% is obtained. In addition to being able to accurately detect close-range targets, the proposed method this can also detect target ''bus'' with 100% mAP values that are not detected by other methods, thereby expanding the detection range. Thus, our fused results lead the pack in terms of targets ''person'' and ''bus'' detection as well average detection accuracy.

TABLE 7. Average running time of on two image datasets.

Methods	Average running time (unit: s)	
	TNO dataset	RoadScene dataset
VggML	5.9308	3.5316
Resnet50	3.7758	2.7426
BayF	1.1611	0.8939
DDcGAN	3.2797	1.3940
DIM	33.9392	9.8214
CSF	14.8467	7.7464
CFL	33.9576	25.9295
SeAFusion	0.0033	0.0029
AUIF	12.4612	7.1988
Ours	7.1011	4.3709

4) DISCUSSION ON TIME EFFICIENCY

To compare the time efficiency of the proposed method with those of the other methods, the average running time of a total of 60 images on two image datasets is calculated and presented in Table 5. From Table 5, we can observe that our method runs faster than DIM and CFL, even up to 4 times faster. The reason for this is that our method uses a single iteration and the fusion rule is directly applied to the pixels, while the common feature of DIM and CFL is that both methods make use of the well-recognized time-consuming theory–sparse representation to generate an over-complete dictionary, resulting in a substantial increase in running time. A few methods require less running time than the proposed method mainly because: i) models are trained in advance for further testing, such as the VggML, Resnet50, and DDcGAN methods; ii) scale decomposition is not performed, such as the BayF-based. Although these methods are highly efficient, the proposed method achieves better performance combined with the results in terms of subjective and objective metrics.

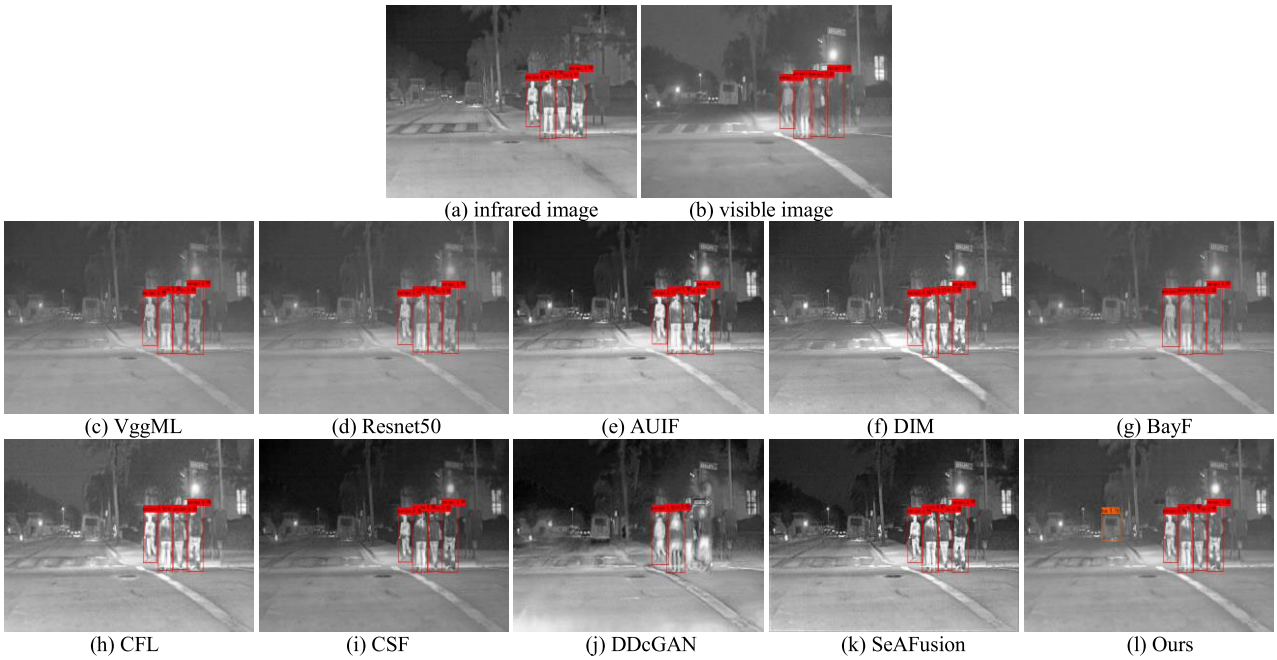


FIGURE 8. Object detection results for infrared, visible and fused images from the RoadScene dataset for the "FLIR_03952" scene.

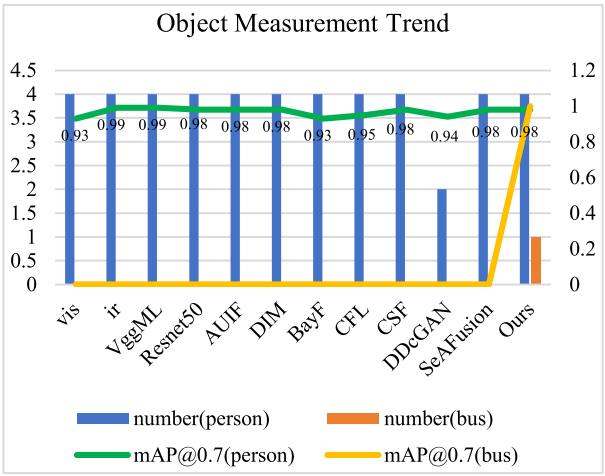


FIGURE 9. Show the object measurement trend of visible (vis), infrared (ir) and fused results on the RoadScene real dataset.

5) DISCUSSION ON ALGORITHM EFFECTIVENESS

After care analysis of the subjective visual results presented in Figure 6, the objective indicators shown in Tables 5-6, and the object detection results and probabilities displayed in Figures 8-9 on the TNO and RoadScene public datasets, it can be concluded that the proposed method is capable of producing prominent infrared objects, e.g., persons and signs, and capturing fine details, e.g., trees, as well as detecting the most objects, these cases are not produced simultaneously by other comparison methods. Attribute to the reason, we believe: i) the positive effect of GLF, which is constructed from three different aspects to extract necessary features from the root and reduce the influence of redundant information; 2) the design of DEM, which establishes connections between the detail layers and input images, thereby allowing the fused detail layer to contain the best pixel region from these images in accordance with the "maximum absolute" rule. Admittedly, this work also has limitations, namely the algorithm efficiency highlighted in Table 7. As can be seen from the table, the efficiency of the proposed method ranks in the middle, which is attributed to the fineness of the decomposition algorithm. In future work, we will further improve the efficiency and try to extend it to images of other modalities to verify its generality.

V. CONCLUSION

This paper proposes a novel target-aware fusion framework for infrared and visible images. First, the GLF is designed from three error measure equations with the aim of effectively removing mild noise while preserving the texture details and edges as well as brightness information. Next, the GLF is used as a decomposition tool to decompose the source images to obtain the based layers and the detail layers. Then, two fusion rules are presented. For the base layer, a "weighted average" scheme is applied to guide the fusion, and a DEM is proposed with respect to the detail layer. Finally, the fused image can be reconstructed by adding the sub-fused images. The experiment results prove that the proposed method based on GLF reaches a better fusion and detection performance and outperforms other existing mainstream fusion methods. In future work, we will consider extending this method to other image types to demonstrate the generalizability of this method.

REFERENCES

[1] J. Ma, Y. Ma, and C. Li, "Infrared and visible image fusion methods and applications: A survey," *Inf. Fusion*, vol. 45, pp. 153–178, Jan. 2019.

[2] Q. Li, G. Han, P. Liu, H. Yang, D. Chen, X. Sun, J. Wu, and D. Liu, "A multilevel hybrid transmission network for infrared and visible image fusion," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2022.3186048.

[3] Z. Wang, Y. Wu, J. Wang, J. Xu, and W. Shao, "Res2Fusion: Infrared and visible image fusion based on dense Res2Net and double nonlocal attention models," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022, doi: 10.1109/TIM.2021.3139654.

[4] H. Zhou, W. Wu, Y. Zhang, J. Ma, and H. Ling, "Semantic-supervised infrared and visible image fusion via a dual-discriminator generative adversarial network," *IEEE Trans. Multimedia*, vol. 25, pp. 635–648, 2023, doi: 10.1109/TMM.2021.3129609.

[5] Q. Zhang, T. Xiao, N. Huang, D. Zhang, and J. Han, "Revisiting feature fusion for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1804–1818, May 2021.

[6] W. Zhou, Y. Zhu, J. Lei, J. Wan, and L. Yu, "CCAFNet: Crossflow and cross-scale adaptive fusion network for detecting salient objects in RGB-D images," *IEEE Trans. Multimedia*, vol. 24, pp. 2192–2204, 2021.

[7] Y. Wang, Y. Xiao, J. Lu, B. Tan, Z. Cao, Z. Zhang, and J. T. Zhou, "Discriminative multi-view dynamic image fusion for cross-view 3-D action recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5332–5345, Oct. 2022.

[8] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "RGB-T salient object detection via fusing multi-level CNN features," *IEEE Trans. Image Process.*, vol. 29, pp. 3321–3335, 2020.

[9] L. Mou, C. Zhou, P. Xie, P. Zhao, R. Jain, W. Gao, and B. Yin, "Isotropic self-supervised learning for driver drowsiness detection with attention-based multimodal fusion," *IEEE Trans. Multimedia*, vol. 25, pp. 529–542, 2023, doi: 10.1109/TMM.2021.3128738.

[10] Z. Zeng, T. Wang, F. Ma, L. Zhang, P. Shen, S. A. A. Shah, and M. Bennamoun, "Probability-based framework to fuse temporal consistency and semantic information for background segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 740–754, 2022.

[11] L. C. Chen, G. Papandreou, and I. Kokkinos, "DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Jun. 2018.

[12] Y. Yang, Y. Zhang, S. Huang, Y. Zuo, and J. Sun, "Infrared and visible image fusion using visual saliency sparse representation and detail injection model," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–15, 2021, doi: 10.1109/TIM.2020.3011766.

[13] E. Vargas, H. Arguello, and J.-Y. Tourneret, "Spectral image fusion from compressive measurements using spectral unmixing and a sparse representation of abundance maps," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 7, pp. 5043–5053, Jul. 2019.

[14] B. Wang, H. Niu, J. Zeng, G. Bai, S. Lin, and Y. Wang, "Latent representation learning model for multi-band images fusion via low-rank and sparse embedding," *IEEE Trans. Multimedia*, vol. 23, pp. 3137–3152, 2021, doi: 10.1109/TMM.2020.3020695.

[15] B. Wang, Q. Zhao, G. Bai, J. Zeng, S. Xie, and L. Wen, "LIALFP: Multi-band images synchronous fusion model based on latent information association and local feature preserving," *Infr. Phys. Technol.*, vol. 120, Jan. 2022, Art. no. 103975.

[16] S. Singh and R. S. Anand, "Multimodal medical image sensor fusion model using sparse K-SVD dictionary learning in nonsubsampled shearlet domain," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 2, pp. 593–607, Feb. 2020.

[17] P. Zhou, R. Zhang, J. Xie, J. Liu, H. Wang, and T. Chai, "Data-driven monitoring and diagnosing of abnormal furnace conditions in blast furnace ironmaking: An integrated PCA-ICA method," *IEEE Trans. Ind. Electron.*, vol. 68, no. 1, pp. 622–631, Jan. 2021.

[18] J. Jiang, J. Ma, C. Chen, Z. Wang, Z. Cai, and L. Wang, "SuperPCA: A superpixelwise PCA approach for unsupervised feature extraction of hyperspectral imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4581–4593, Aug. 2018.

[19] H. Jiang and Y. Tian, "Fuzzy image fusion based on modified self-generating neural network," *Exp. Syst. Appl.*, vol. 38, no. 7, pp. 8515–8523, 2011.

[20] S. Singh and D. Gupta, "Detail enhanced feature-level medical image fusion in decorrelating decomposition domain," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–9, 2021, doi: 10.1109/TIM.2020.3038603.

[21] M. Yin, X. Liu, Y. Liu, and X. Chen, "Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain," *IEEE Trans. Instrum. Meas.*, vol. 68, no. 1, pp. 49–64, Jan. 2019.

[22] Z. Wang, Y. Chen, W. Shao, H. Li, and L. Zhang, "SwinFuse: A residual Swin Transformer fusion network for infrared and visible images," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–12, 2022, doi: 10.1109/TIM.2022.3191664.

[23] H. Li, Y. Cen, Y. Liu, X. Chen, and Z. Yu, "Different input resolutions and arbitrary output resolution: A meta learning-based deep framework for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 30, pp. 4070–4083, 2021.

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

[25] H. Li and X.-J. Wu, "DenseFuse: A fusion approach to infrared and visible images," *IEEE Trans. Image Process.*, vol. 28, no. 5, pp. 2614–2623, May 2019.

[26] 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.

[27] J. Ma, L. Tang, M. Xu, H. Zhang, and G. Xiao, "STDFusionNet: An infrared and visible image fusion network based on salient target detection," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–13, 2021.

[28] L. Tang, Y. Deng, Y. Ma, J. Huang, and J. Ma, "SuperFusion: A versatile image registration and fusion network with semantic awareness," *IEEE/CAA J. Autom. Sinica*, vol. 9, no. 12, pp. 2121–2137, Dec. 2022.

[29] 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.

[30] W. Tang, F. He, and Y. Liu, "YDTR: Infrared and visible image fusion via Y-shape dynamic transformer," *IEEE Trans. Multimedia*, early access, Jul. 20, 2022, doi: 10.1109/TMM.2022.3192661.

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

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

[33] G. Wang, W. Li, X. Gao, B. Xiao, and J. Du, "Functional and anatomical image fusion based on gradient enhanced decomposition model," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–14, 2022, doi: 10.1109/TIM.2022.3170983.

[34] H. Li, X.-J. Wu, and J. Kittler, "MDLatLRR: A novel decomposition method for infrared and visible image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4733–4746, 2020.

[35] M. Das, D. Gupta, P. Radeva, and A. M. Bakde, "Optimized multimodal neurological image fusion based on low-rank texture prior decomposition and super-pixel segmentation," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–9, 2022, doi: 10.1109/TIM.2022.3165263.

[36] Z. Zhou, B. Wang, S. Li, and M. Dong, "Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with Gaussian and bilateral filters," *Inf. Fusion*, vol. 30, pp. 15–26, Jul. 2016.

[37] Y. Luo, K. He, D. Xu, W. Yin, and W. Liu, "Infrared and visible image fusion based on visibility enhancement and hybrid multiscale decomposition," *Optik*, vol. 258, May 2022, Art. no. 168914.

[38] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 2003.

[39] 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.

[40] Z. Liu, K. Tsukada, K. Hanasaki, Y. K. Ho, and Y. P. Dai, "Image fusion by using steerable pyramid," *Pattern Recognit. Lett.*, vol. 22, no. 9, pp. 929–939, 2001.

[41] G. Liu, Z. Jing, S. Sun, J. Li, Z. Li, and H. Leung, "Image fusion based on expectation maximization algorithm and steerable pyramid," *Chin. Opt. Lett.*, vol. 2, pp. 386–389, Jul. 2004.

[42] H. Jin, "Fusion of infrared and visual images based on contrast pyramid directional filter banks using clonal selection optimizing," *Opt. Eng.*, vol. 47, no. 2, Feb. 2008, Art. no. 027002.

[43] P. Chai, X. Luo, and Z. Zhang, "Image fusion using quaternion wavelet transform and multiple features," *IEEE Access*, vol. 5, pp. 6724–6734, 2017.
[44] F. Meng, M. Song, B. Guo, R. Shi, and D. Shan, "Image fusion based on object region detection and non-subsampled contourlet transform," *Comput. Electr. Eng.*, vol. 62, pp. 375–383, Aug. 2017.
[45] C. Zhao, Y. Guo, and Y. Wang, "A fast fusion scheme for infrared and visible light images in NSCT domain," *Infr. Phys. Technol.*, vol. 72, pp. 266–275, Sep. 2015.
[46] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, 2015.
[47] H. M. Hu, J. Wu, and B. Li, "An adaptive fusion algorithm for visible and infrared videos based on entropy and the cumulative distribution of gray levels," *IEEE Trans. Multimedia*, vol. 19, no. 12, pp. 2706–2719, Dec. 2017.
[48] W. Tan, H. Zhou, J. Song, H. Li, Y. Yu, and J. Du, "Infrared and visible image perceptive fusion through multi-level Gaussian curvature filtering image decomposition," *Appl. Opt.*, vol. 58, no. 12, pp. 3064–3073, Apr. 2019.
[49] L. Chen, X. Yang, L. Lu, K. Liu, G. Jeon, and W. Wu, "An image fusion algorithm of infrared and visible imaging sensors for cyber-physical systems," *J. Intell. Fuzzy Syst.*, vol. 36, no. 5, pp. 4277–4291, May 2019.
[50] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. 14th ACM Int. Conf. Multimedia*, Oct. 2006, pp. 815–824.
[51] P. Liu, L. Xiao, and T. Li, "A variational pan-sharpening method based on spatial fractional-order geometry and spectral–spatial low-rank priors," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 3, pp. 1788–1802, Mar. 2018.
[52] A. Toet. (2014). *TNO Image Fusion Dataset*. [Online]. Available: https://figshare.com/articles/dataset/TNO_Image_Fusion_Dataset/1008029
[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, doi: 10.1109/TPAMI.2020.3012548.
[54] H. Li, X.-J. Wu, and J. Kittler, "Infrared and visible image fusion using a deep learning framework," in *Proc. Comput. Vis. Pattern Recognit.*, Aug. 2018, pp. 2705–2710.
[55] H. Li, X.-J. Wu, and T. S. Durrani, "Infrared and visible image fusion with ResNet and zero-phase component analysis," *Infr. Phys. Technol.*, vol. 102, Nov. 2019, Art. no. 103039.
[56] F. G. Veshki, N. Ouzir, S. A. Vorobyov, and E. Ollila, "Coupled feature learning for multimodal medical image fusion," 2021, *arXiv:2102.08641*.
[57] Z. Zhao, S. Xu, C. Zhang, J. Liu, and J. Zhang, "Bayesian fusion for infrared and visible images," *Signal Process.*, vol. 177, Dec. 2020, Art. no. 107734.
[58] Z. Zhao, S. Xu, J. Zhang, C. Liang, C. Zhang, and J. Liu, "Efficient and model-based infrared and visible image fusion via algorithm unrolling," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1186–1196, Mar. 2022.
[59] H. Xu, H. Zhang, and J. Ma, "Classification saliency-based rule for visible and infrared image fusion," *IEEE Trans. Comput. Imag.*, vol. 7, pp. 824–836, 2021.
[60] J. Ma, H. Xu, J. Jiang, X. Mei, and X.-P. Zhang, "DDcGAN: A dual-discriminator conditional generative adversarial network for multi-resolution image fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4980–4995, 2020.
[61] L. Tang, J. Yuan, and J. Ma, "Image fusion in the loop of high-level vision tasks: A semantic-aware real-time infrared and visible image fusion network," *Inf. Fusion*, vol. 82, pp. 28–42, Jun. 2022.
[62] C. S. Xydeas and V. Petrović, "Objective image fusion performance measure," *Electron. Lett.*, vol. 36, no. 4, pp. 308–309, 2000.
[63] 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.
[64] P.-W. Wang and B. Liu, "A novel image fusion metric based on multi-scale analysis," in *Proc. 9th Int. Conf. Signal Process.*, Oct. 2008, pp. 965–968.
[65] J. Zhao, R. Laganière, and Z. Liu, "Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement," *Int. J. Innov. Comput., Inf. Control*, vol. 3, no. 6, pp. 1433–1447, Dec. 2007.
[66] 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.
[67] M. Haghighat and M. A. Razian, "Fast-FMI: Non-reference image fusion metric," in *Proc. IEEE 8th Int. Conf. Appl. Inf. Commun. Technol. (AICT)*, Oct. 2014, pp. 1–3.
[68] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

YINGMEI ZHANG received the M.S. degree from the School of Information Technology, Jiangxi University of Finance and Economics, Nanchang, China, in 2020. She is currently pursuing the Ph.D. degree in computer science and engineering with Jeonbuk National University, Jeonju, South Korea. Her research interests include multi-modal image processing, artificial intelligence, and pattern recognition.

HYO JONG LEE (Member, IEEE) received the Ph.D. degree in computer science from The University of Utah, in 1991. Since 1991, he has been a Professor with Jeonbuk National University. His research interests include computer graphics, image processing, parallel processing, and artificial intelligence.