Infrared Physics & Technology 131 (2023) 104701

Contents lists available at ScienceDirect

Infrared Physics and Technology

journal homepage: www.elsevier.com/locate/infrared

Infrared and visible image fusion based on domain transform filtering and sparse representation

Xilai Li [a], Haishu Tan [a,b], Fuqiang Zhou [c], Gao Wang [d], Xiaosong Li [a,*]

[a] Guangdong-Hong Kong-Macao Joint Laboratory for Intelligent Micro-Nano Optoelectronic Technology, School of Physics and Optoelectronic Engineering, Foshan University, Foshan 528225, China
[b] Ji Hua Laboratory, Foshan 528000, China
[c] Key Laboratory of Precision Opto-mechatronics Technology, Ministry of Education, Beihang University, 100191, China
[d] State Key Laboratory of Dynamic Measurement Technology, North University of China, Taiyuan 030051, China

ARTICLE INFO

Keywords:
Infrared and visible image fusion
Sparse representation
Domain transform filtering
Salient structure extraction

ABSTRACT

Infrared and visible image fusion is an important branch in the field of information fusion. Fused images can provide more comprehensive and objective interpretations of complex scenes with confusing or hidden information. This paper proposes a novel infrared and visible image fusion method based on domain transform filtering (DTF) and sparse representation (SR). First, infrared and visible images are decomposed using a low-pass filter into base and detail layers, respectively. An SR-based rule is designed to fuse the detail layers. To improve the feature expression ability of a redundant dictionary, we introduce a multiscale detail enhancement technique to preprocess the training data. Additionally, a fast but effective rule is proposed based on salient structure extraction and DTF base layer fusion. The structure and border information of source images can be successfully transformed into the fused image. The proposed method can effectively retain the salient structure and edge information of the source images in the fused image. Experimental results demonstrate that the proposed method can outperform state-of-the-art methods in terms of both subjective visual and objective quantitative evaluations. The source code of the proposed method is released at https://github.com/ixilai/DTF-SR.

1. Introduction

Generally, a comprehensive description of a target scene cannot be obtained from a single image captured using one sensor owing to the differences in imaging mechanisms. One method to obtain a more accurate description is to combine images from different scenes; this approach led to the emergence of multisource image fusion technology [1]. Infrared and visible image fusion (IVIF) is an important branch of this technology. Infrared (IR) imaging sensors can capture the thermal radiation emitted from objects, which is less affected by dark or adverse weather conditions. However, IR images generally lack sufficient background information from the scene. By contrast, visible (VIS) images contain more details and texture information. IVIF can integrate heat source and visible details from source images into a fused image, which can provide a more comprehensive interpretation than that obtained from any source image. In recent years, IVIF has been widely used in applications such as military surveillance, resource detection, security surveillance, and target identification.

Based on the level at which fusion is performed, image fusion can be categorized into three: decision-level fusion (DLF), feature-level fusion (FLF), and pixel-level fusion (PLF). DLF relies on the recognition results of multiple sensors based on the independent execution of a decision or classification by each sensor to provide a globally optimal decision. FLF first extracts the feature information from the source image, then processes the feature information, and finally integrates the processed feature information to obtain a fused feature map. In contrast to DLF and FLF, PLF directly processes the pixels of the source image; the fused results of PLF yield better fusion performance than FLF and DLF. Thus, IVIF and the proposed method in this study primarily focus on PLF. Current PLF methods can be approximately classified into four categories: multiscale transform (MST)- [2], deep learning (DL)- [3], spatial domain (SPD)- [4], and sparse representation (SR)-based methods [5].

MST-based methods include three steps: multiscale decomposition (MSD), fusion rules, and inverse MSD. Popular MSDs include the discrete wavelet transform [6], Laplacian pyramid [7], contrast pyramid [8], and dual-tree complex wavelet transform [9]. To better capture the intrinsic

* Corresponding author.
E-mail address: lixiaosong@fosu.edu.cn (X. Li).

https://doi.org/10.1016/j.infrared.2023.104701
Received 12 July 2022; Received in revised form 13 April 2023; Accepted 24 April 2023
Available online 28 April 2023

geometrical structures of images, various multiscale geometrical analysis methods have been developed; these include the shearlet transform [10], curvelet transform [11], and nonsubsampled contourlet transform [12]. Nevertheless, MST-based methods inevitably lose details of the source image in the process of decomposition and reconstruction. DL-based methods [14,15] generally constrain the fused images to have desired distribution characteristics by constructing an objective function. In addition, convolutional neural networks (CNNs) are becoming very common in the field of image fusion because of the powerful feature extraction capability of neural networks. DL-based methods generally implement the fusion process in an end-to-end manner, which can effectively reduce preprocessing, parameter adjustment, and post processing. However, the training data of DL-based methods is insufficient. Compared with target recognition, detection, and other tasks, less amounts of data are available for image fusion, which has an impact on the performance of DL models.

Compared with the MST and DL methods, SPD-based methods are simpler, more direct, and faster. These methods deal with pixels and operators based on pixel intensities. They consider sections with greater energy or significant changes in pixel intensity as being in salient sections. A popular SPD-based method is pixel-based activity detection. First, the activity of each source image is detected separately by a specific function or algorithm, and the corresponding activity saliency map is obtained. Then, according to a given rule (e.g., the "max-absolute" rule), it can generate an activity decision map. Finally, the source image is reconstructed to obtain the fused image under the guidance of the decision map. Although SPD-based methods can preserve the edge features of the source images, they still have shortcomings. For example, incorrect decisions may appear in the optimization process (i.e., some wrong information is introduced in the boundary of the fused image), leading to a smoothing effect.

SR uses an overcomplete dictionary to represent signal flexibly and atoms, and SR coefficients in the dictionary can effectively express the intrinsic structural features of the image. SR has been widely used in face recognition [27], action recognition [28], person re-identification [29,30], object tracking [31], and image fusion [32–38]. Although SR-based methods can achieve satisfactory fusion performance to some extent, their computational efficiency remains a challenge. More importantly, the small-scale structures in the source images could be further refined.

To further improve the performance of SR-based methods, some scholars [36] have combined SR with MSD tools. The powerful feature extraction capability of SR can identify useful pixel information extracted in different directions and at different scales; however, this also means that these models have high computational complexity. Moreover, in the absence of sufficient training samples in the redundant dictionary, it may also introduce wrong pixel information in the fusion process, which reduces the clarity and contrast of the fusion results. In addition, differences in training data have a significant impact on the quality of the redundant dictionary. Moreover, boosting the redundancy of the dictionary by adding a large number of training samples will inevitably introduce duplicate or redundant information into the fusion results, leading to color distortion or loss of details. Although the mechanism of ASR considering blocks directly on the source image for dictionary learning can promote the fusion quality to an extent, not all significant information on the source image must be retained in the fusion result, and using this mechanism is not conducive to maintaining the correct contrast of the fusion result and introducing a large amount of redundant information while possibly obscuring some important texture details.

To solve the above problems, in this paper, we propose an IVIF method based on DTF and SR and design a detail layer fusion strategy based on SR. We preprocess the samples of the training dictionary with multiscale enhancement so that the compactness and feature representation of the dictionary can be maintained without a large amount of training data. Moreover, the multiscale enhanced dictionary is able to identify more useful pixel information in the detail layer than before preprocessing. For the fusion of the base layer, our aim is to identify the pixel information with more energy quickly and efficiently and to be able to maintain a reasonable contrast level. Therefore, we introduce a fast fusion rule based on DTF, which leverages the edge-holding filter's ability to retain high-contrast edges and remove low-contrast or subtly varying edge details, effectively reducing the effect of unnecessary texture information in low-frequency images and accurately identifying the location of useful pixel information in low-frequency images. In addition, we introduce an iterative mechanism in the domain transform filter that is able to further control the level of detail of the filtering. This strategy is able to detect prominent structures of the base layer well and retain them in the fused base layer. The main contributions of this study can be summarized as follows:

We propose a new IVIF method based on DTF and SR that can effectively identify the useful pixel information in IR and visible images, and the obtained fusion results have high color fidelity and detail retention.
We design an SR-based detail layer fusion rule and introduce a multiscale enhancement technique to increase the feature extraction capability of the redundant dictionary, and this technique can improve the compactness of the dictionary without a large increase in training samples. For base layer fusion, we design a simple and effective fusion rule based on DTF and introduce iterative processing to DTF to improve its performance in recovering large edge structures.
To verify the effectiveness of the proposed algorithm, we conduct extensive experiments on public datasets using 11 state-of-the-art comparison methods and experimentally demonstrate that the proposed algorithm presents superior fusion performance than others.

The remainder of the paper is organized as follows. In Section 2, we briefly review current image fusion methods and focus on two previous approaches, DTF and SR. The proposed IVIF algorithm is presented in Section 3. The experimental details, including parameter setting and fusion results analysis, and discussion are given in Section 4. Finally, the conclusions are presented in Section 5.

2. Related research

In this section, we introduce existing image fusion methods, and because our method is primarily implemented through DTF and SR, we also provide a brief review of these two approaches.

2.1. Existing image fusion methods

Existing image fusion methods can be roughly divided into four categories: MST-based methods, DL-based methods, SPD-based methods, and SR-based methods. Next, we briefly discuss the main features and some classical techniques of MST-based and DL-based methods, following which we focus on the status of SPD-based and SR-based methods and some cutting-edge techniques.

MST-based fusion methods are computationally expensive owing to multilayer and multidirectional decomposition. Bavirisetti and Dhuli [13] proposed an image fusion method based on saliency detection and two-scale image decomposition to reduce the complexity of multiscale fusion algorithms. DL-based methods generally utilize the powerful feature extraction ability of CNNs to identify useful information from images. For example, Liu et al. [14] introduced a deep CNN into multifocus image fusion and treated the fusion task as a classification problem. A classification model for focused and out-of-focus pixels can be obtained by learning. To improve the generalization ability of the CNN model, Zhang et al. [15] proposed a CNN-based general image fusion framework. This framework utilizes two convolutional layers to extract the salient image features from multiple input images. Then,

different convolutional features of the input images are reconstructed after they are fused by an appropriate fusion rule to form a fully connected CNN fusion framework. Notably, the limitation of high-quality datasets has always been a key concern in DL methods. To address this, Hou et al. [39] proposed an unsupervised end-to-end learning framework for IVIF, using visible light and infrared frames to construct a benchmark dataset for training purposes. To effectively explore the potential correlation between internal features of infrared and visible light images, Ding et al. [40] proposed an unsupervised cross-modal feature aggregation network that fuses the affinity features of cross-modal images through the combination of l1-norm and attention mechanisms. To make the neural network more sensitive to global spatial information in source images and effectively extract the thermal radiation information of infrared images and the detailed information of visible images, Li et al. [41] incorporated a dual-branch multiperception field neural network into the IVIF task. In addition to CNNs, generative adversarial networks (GANs) [16] have been applied to the field of image fusion [17], where the source image is input to the generator of the GAN to generate a fused image containing the features of the source image. GANs have also been widely used in multifocus image fusion (MFF)- [18], multiexposure image fusion (MEF)- [19], and remote sensing image fusion (PAN)-GAN [20].

SPD-based methods generally design rules that analyze the activity of image pixels to construct weights or decision maps to achieve image fusion. Image activity can reflect the level of image clarity and structural saliency, which has an important effect on the quality of fusion results. Li et al. [21] proposed an image fusion method based on residual removal. In this scheme, a multiscale focus detection strategy is designed, and it can effectively detect focus pixels and use focus detection to determine the focused regions outside the boundary. Moreover, a transform domain-based method is used to determine the boundary. This method combines the superiorities of transform domain and spatial domain methods. Recently, in addition to pixel activity detection-based methods, edge-preserving filtering algorithms have gained popularity. The benefit of edge-preserving filtering is that the image structure can be precisely separated at different scales while maintaining the spatial consistency of the image structure and reducing the blurring effect around edges. Kumar et al. [22] proposed a cross bilateral filter and used it for image fusion. Tan et al. [23] used a Gaussian curvature filter to decompose an image in IVIF and effectively preserved the fine scale structure. Li et al. [24] proposed a two-layer decomposition image fusion model based on a joint bilateral filter, which decomposes the image into an energy layer that comprises intensity information and a structural layer containing details. Li et al. [25] developed a guided filter-based fusion method, which can be effectively applied to multifocus, medical, and IVIF. Ma et al. [26] used a rolling guidance filter to decompose source images. Unlike the guided filter, the rolling mechanism can automatically refine the edges to maintain large-scale structure, and the texture information of the image can be preserved well.

SR is capable of learning an overcomplete dictionary from a set of training images used for image fusion, providing a more stable and meaningful representation of the source images. Liu et al. [5] demonstrated that for image fusion tasks, the pixel activity measure in the sparse domain is reliable, and it can yield good translation invariance when combined with spatial sliding window technology. Li and Yang [32] first introduced SR to the field of image fusion, using discrete cosine transform (DCT) to construct a redundant dictionary. Subsequently, they [33] constructed another hybrid dictionary comprising a DCT basis, wavelet “db1” basis, Gabor basis, and ridgelet basis. A hybrid dictionary can reflect several specific structures efficiently. However, because of the diversity and complexity of image features, dictionaries in the form of the above proposed basis cannot easily represent different images adaptively. The learning dictionary obtained from training samples can effectively solve this problem. Liu et al. [34] proposed an adaptive SR (ASR)-based image fusion method. This model learns a group of compact sub-dictionaries from a large number of high-quality image patches. In the fusion process, one of the sub-dictionaries is adaptively selected for a given source image patch so that it can effectively achieve image fusion and denoizing. ASR [34] also revealed that the differences in training data have a significant impact on the quality of redundant dictionaries. Nevertheless, directly exploiting source image patches for dictionary learning will inevitably introduce significant amounts of repetitive or redundant information. To solve this problem, Zhu et al. [35] proposed a dictionary learning approach for multimodal medical image fusion based on K-singular value decomposition (K-SVD). This scheme constructs a compact dictionary through information sampling and local density peak clustering. Except for promoting the quality of fused images by improving the redundant dictionary, a general image fusion framework based on MST and SR has been proposed in [36]. For simultaneous fusion and denoizing, Li et al. [38] proposed a three-layer decomposition image fusion and denoizing method, where the source image is decomposed into a high-frequency layer, low-frequency cartoon layer, and low-frequency texture layer. The SR rule was employed to fuse and denoize the high-frequency layer. Li and Wu [42] proposed an IVIF method based on latent low-rank representation, which could express global structural information and local information efficiently.

2.2. Dtf

The pioneering task of anisotropic diffusion is a structure-aware smoothing approach that can effectively preserve image structures [43–45]. Traditional anisotropic diffusion has the disadvantage of structural over-sharpening and high computational complexity. Therefore, some improved models of smoothing filters that preserve structures have emerged [25,46,47]; these include the guided filter [25], bilateral filter (BF) [46], and DTF [47]. The BF is a generalized form of the joint BF [48]. These spatial domain filters obtain the output image through joint filtering by using two inputs: guidance and input image. Given an input image I and a guidance image T, the output of the BF for pixel p is

$$L_p = \frac{1}{U_p}\sum_{q\in\Omega_p} I_q\left(exp\left(-\left(\frac{\|p-q\|_2^2}{2\sigma_s^2}+\frac{\|T_p-T_q\|_2^2}{2\sigma_r^2}\right)\right)\right) \tag{1}$$

where Ω_p is the sliding window centered at the pixel p, and q represents the pixel points in Ω_p except p. σ_s is the space domain scaling parameter, σ_r is the range domain scaling parameter, and U_p is the normalized term calculated as

$$U_p = \sum_{q\in\Omega_p} I_q\left(\exp\left(-\left(\frac{\|p-q\|_2^2}{2\sigma_s^2}+\frac{\|T_p-T_q\|_2^2}{2\sigma_r^2}\right)\right)\right). \tag{2}$$

When the intensity $T_q \in \Omega_p$ of the pixel q is similar to the intensity T_p at the center of Ω_p and the two pixels are on the same side of an edge, $\exp\left(\frac{\|T_p-T_q\|_2^2}{2\sigma_r^2}\right)$ tends to one, and the weight of I_q is determined by $\exp\left(-\frac{\|p-q\|_2^2}{2\sigma_s^2}\right)$. However, if T_q is very different from T_p and the two pixels lie on different sides of an edge, $\exp\left(\frac{\|T_p-T_q\|_2^2}{2\sigma_r^2}\right)$ tends to zero, and the weight of I_q also tends to zero. The correlation between I_p and I_q is very weak, and the intensity of I_q has almost no effect on the filter result L_p. In this manner, similar intensities in the local neighborhood Ω_p are smoothed, while the intensities on different sides of edges do not affect each other. Therefore, we can consider the joint filters as a type of smoothing filter with structure-aware capability. In [47], Gastel and Olivieria proved that DTF can achieve a comparable effect to joint bilateral filtering.

A distance metric d_{pq} can be defined by the power term in Eq. (1) as

$$d_{pq} = \frac{\|p-q\|_2^2}{2\sigma_s^2}+\frac{\|T_p-T_q\|_2^2}{2\sigma_r^2} \tag{3}$$

which can be rewritten as

$$2\sigma_s^2 \frac{d_{pq}}{\|p-q\|_2^2} = 1 + \frac{\sigma_s^2 \|T_p - T_q\|_2^2}{\sigma_r^2 \|p-q\|_2^2}. \tag{4}$$

Gastel et al. [47] improved the l_2-norm in the above equation by using the l_1-norm to achieve real-time efficiency. According to [47] and Eq. (4), the derivative of the transformed signal δ_p can be defined as

$$\delta'_p = 1 + \frac{\sigma_s}{\sigma_r} T'_p \tag{5}$$

where δ'_p and T'_p are the derivatives of δ_p and T_p, respectively. Because Eq. (5) is integrated and discretized, the distance δ_{pq} can be defined as

$$\delta_{pq} = \sum_{i=p}^{q} \left(1 + \frac{\sigma_s}{\sigma_r} |T_i - T_{i-1}|\right). \tag{6}$$

From the above analysis, it is clear that the DTF has two input images I_p and T_p, one output image L_p, and two parameters σ_s and σ_r, which can generally be defined as

$$L_p = DTF(I_p, T_p, \sigma_s, \sigma_r). \tag{7}$$

Further details on DTF can be found in [44,50].

2.3. SR theory

For a given signal $y = (y_1, y_2, \cdots y_n), y \in R^r$, the basic assumption of the SR theory is that y can be approximated as a linear combination of a set of elementary signals $\{d_i\}_{i=1}^{c}$ in a redundant dictionary $D \in R^{r \times c} (r < c)$, expressed as

$$y = \sum_{i=1}^{f} d_i \alpha_i = D\alpha, \tag{8}$$

where $\alpha = (\alpha_1, \alpha_2, \cdots \alpha_N)$ is a vector of unknown sparse coefficients, and d_i is an atom of D. Because the dictionary is redundant, the vector α is not unique; hence, the goal of SR is to calculate the sparsest α, which contains the fewest non-zero entries. This method is expressed mathematically as

$$\widehat{\alpha} = \underset{\alpha}{argmin} \|\alpha\|_0, s, t. \|y - D\alpha\|_2^2 < \varepsilon, \tag{9}$$

Where $||\cdot||_0$ denotes the l_0-norm that counts the number of non-zero entries, and $\varepsilon > 0$ is an error tolerance. The above formula is solved by the greedy approximation approach of orthogonal matching pursuit (OMP) [49].

3. Proposed method

In this study, an IVIF method based on DTF and SR is proposed (DTF-SR), the specific process is shown in Fig. 1. An image is first decomposed into a detail layer and base layer using a low-pass filter. The detail layer represents the detailed information of the image, which is fused by SR. This approach consists of three steps: dictionary learning, sparse coding, and sparse coefficient fusion. In this study, multiscale detail enhancement was performed on the samples of the training dictionary to effectively improve the fusion results. The base layer can be approximated as a complete image and contains partial structural information. Therefore, we first detect the gradient features of the image to obtain the large-scale structure of the source image and then transfer the small-scale details from the source image to the fused output using improved DTF. The DTF output is used as a weighted map that is combined with the source images to obtain a fused base layer by a weighted-sum rule. Finally, the fused detail layer is superimposed with the fused base layer to obtain the fused image.

3.1. Base layer and detail layer decomposition

The source image primarily contains detail information and energy information. In image fusion tasks, the source image can generally be regarded as a superposition of two components [24] (i.e., it can be decomposed into a base layer and detail layer) and then fused separately using different rules. Using the low-pass filter optimization function composed of the following discrete gradient operators, the basic layer of

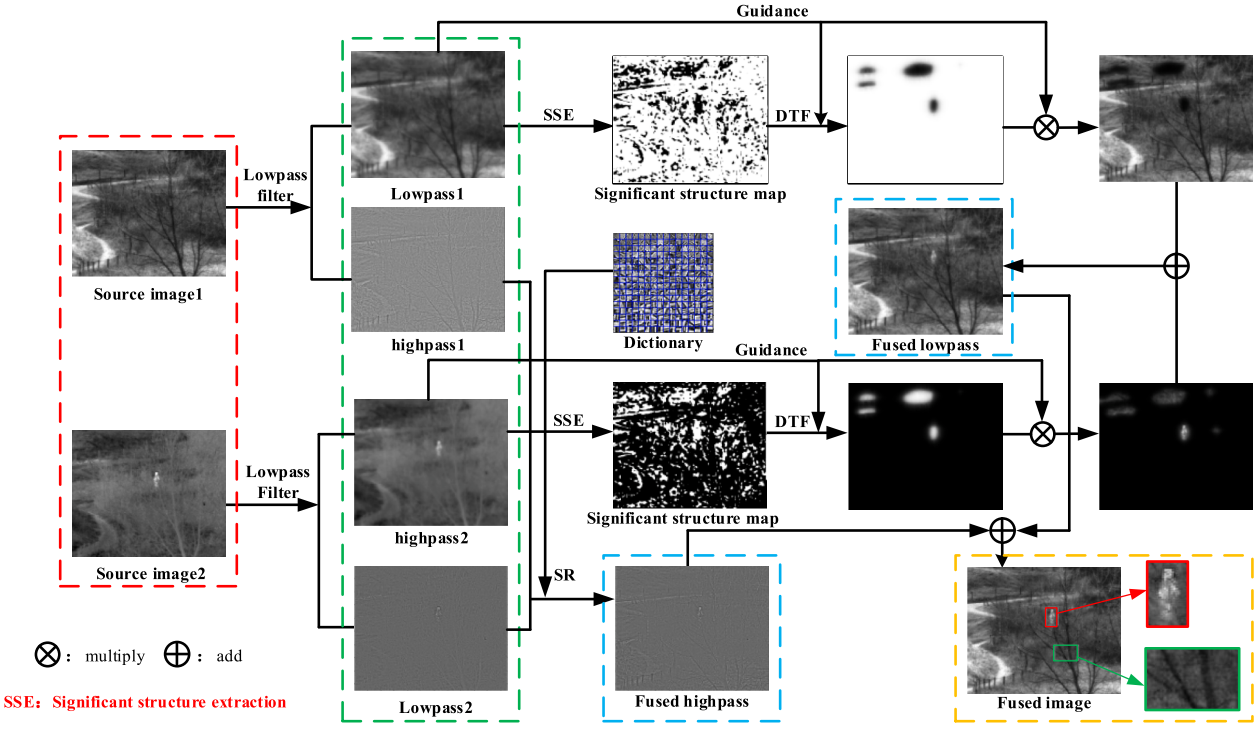

Fig. 1. Schematic of the proposed IVIF method.

the image can be obtained:

$$\arg\min_{I_b} \|I_m - I_b\|_F^2 + \beta\|g_p * I_b\|_F^2 + \|g_q * I_b\|_F^2 \tag{10}$$

where I_m is the input source image, and I_b is the base layer of I_m. $g_p = [-1, 1]$ and $g_q = [-1, 1]^T$ are the gradient operators in the vertical and horizontal directions, respectively. * denotes the convolution operation. β is the regularization parameter controlling the low-pass filter. We can solve this objective function in the discrete Fourier domain to obtain the base layer I_b and then subtract it from the source image to obtain the detail layer I_d:

$$I_d = I_m - I_b. \tag{11}$$

3.2. Fusion rules

3.2.1. Detail layer fusion

The detail layer contains rich detail information such as edges and textures. Generally, the "max-absolute" rule is employed for fusion, but this approach only considers the energy information of the pixels, which may yield sub-optimal results. For this purpose, detail layer fusion rules are designed based on SR, which contains dictionary learning and SR.

(1) Dictionary learning.

For methods based on SR, redundant dictionaries have a significant impact on the fusion results. Considering that the quality of the training images may influence the performance of SR, a multiscale detail enhancement technique was designed using a low-pass filter to perform multiple decompositions of the images to extract high-frequency information at different levels; finally, they were merged into the training data. Eq. (10) represents a low-pass filter, indicated by *Low-pass*. The multiscale detail enhancement is expressed as follows:

$$\widehat{h}_t(i,j) = h_t(i,j) + \sum_{l=1}^{l} H_t^l(i,j), \tag{12}$$

$$H_t^l = h_t - Lowpass * h_t \tag{13}$$

where h_t denotes the t-th source image, $\widehat{h}_t(i,j)$ is the image after detail enhancement, H_t^l represents the high-frequency information in the l-th layer of image h_t, and * denotes the convolution operation.

Further, the enhanced data $\widehat{h}_t$ are used as training samples. Image blocks of size 8×8 are collected using the sliding window technique to construct the final training set, and the redundant dictionary D is obtained using the K-SVD algorithm [35]. As shown in Fig. 2, the enhanced training samples can highlight more active pixel information, which is beneficial to the extraction and retention of details in the SR process.

(2) SR.

SR generally consists of three steps: block extraction and vectorization, sparse coding, and sparse vector fusion.

Step 1: Block extraction and vectorization.

Each source image is split into 8×8 patches from top to bottom and left to right. To ensure better shift invariance, the number of overlapping pixels is set to seven, and then, each patch is normalized. Finally, we pull each image block into the corresponding column vector $\vec{I}_{d,l}^k (k = 1, 2, ..., K, l = 1, 2, ..., n)$, where K and n represent the numbers of image blocks and source images, respectively. $\vec{I}_{d,l}^k$ represents the column vector of the k-th layer image block of image $I_{d,l}$.

Step 2: Sparse coding.

OMP [46] is used to optimally solve the sparse coefficient vectors $\{\vec{\alpha}_1^k, \vec{\alpha}_2^k, ..., \vec{\alpha}_n^k\}$ of the detail layer images.

$$\widehat{\alpha} = \underset{\alpha}{argmin}\{\|y - D\alpha\|_2^2 + \|\alpha\|_1\} \tag{14}$$

where $\widehat{\alpha} = \{\vec{\alpha}_1^k, \vec{\alpha}_2^k, ..., \vec{\alpha}_n^k\}$, and y is the sparse coefficient vector $\{\vec{I}_{d,1}^k, \vec{I}_{d,2}^k, ...\vec{I}_{d,n}^k\}$.

Step 3: Sparse vector fusion.

The "max-absolute choosing" rule is employed to obtain the fused sparse vector $\vec{\alpha}_F^k$:

$$\vec{\alpha}_F^k = \max\{\vec{\alpha}_1^k, \vec{\alpha}_2^k, ..., \vec{\alpha}_n^k\} \tag{15}$$

By integrating D and $\vec{\alpha}_F^k$, the column vector $\vec{F}_k$ of the *k-th* layer image block of the fused image can be obtained:

$$\vec{F}_k = D\vec{\alpha}_F^k. \tag{16}$$

All column vectors $\vec{F}_k$ are reconstructed into image patches F_k, and then, each F_k is inserted into its corresponding original position to finally obtain the fused image $\widehat{I}_d$:

$$\widehat{I}_d = \bigcup_{K=1}^{K} F_k. \tag{17}$$

3.2.2. Base layer fusion

Step 1. Significant structure extraction (SSE).

The base layer contains useful structural information from the source image, and salient structure detection and extraction [50–52] of the source image are of great importance to the fusion of the base layer. Inspired by [53], we introduce a DFT-based base layer fusion rule.

I_b^A and I_b^B are the base layers of the VIS and IR images, respectively. Assuming that each source image I_b with intensity in the range of [0, 255] has eight bits, I_b is normalized to

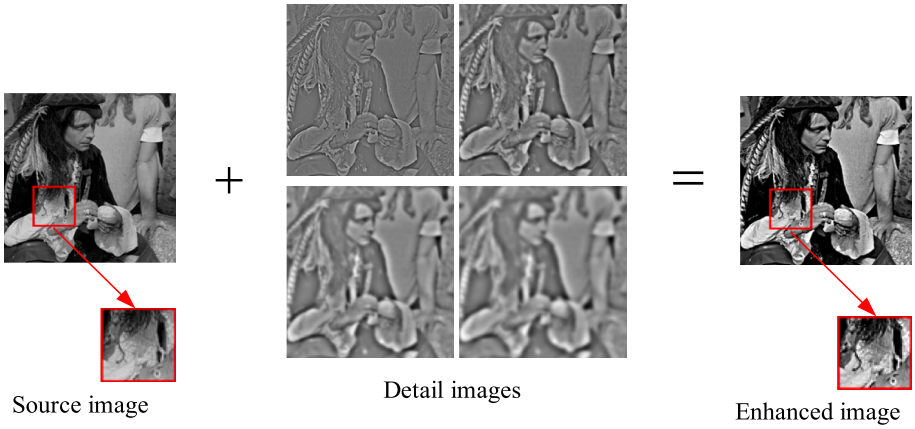

Fig. 2. Enhancement of training samples.

$$S = normalised(I_b) = \frac{I_b}{255}, \tag{18}$$

where S is the normalized image of I_b^A in the range of [0, 1].

To lower the effect of artifacts and unnecessary information, we use a rolling guidance filter (RGF) [54] to process low-frequency images. The RGF is an edge-preserving filter that introduces a scale measure following the scale space theory. This definition enables control over the level of detail during filtering. Furthermore, its rolling guidance mechanism enables automatic edge refinement, which can preserve large-scale structures optimally. The RGF consists of three steps: small structure removal, edge recovery, and structure detection.

(1) Small structure removal.

First, it acquires the original guidance image G through a Gaussian filter. The Gaussian filtering of images I_b^A and I_b^B at central pixel m can be expressed as follows:

$$G_A(m) = \frac{1}{U}\sum_{n\in S}\exp\left(-\frac{\|m-n\|^2}{2\sigma_s^2}\right)\mathrm{I}_b^A(n), \tag{19}$$

$$G_B(m) = \frac{1}{U}\sum_{n\in S}\exp\left(-\frac{\|m-n\|^2}{2\sigma_s^2}\right)\mathrm{I}_b^B(n) \tag{20}$$

where $U = \sum_{n\in S}\exp\left(-\frac{\|m-n\|^2}{2\sigma_s^2}\right)$ is used for normalization, S is the set of neighboring pixels at m, n denotes a neighbor pixel, and structural scale parameter σ_s can be defined as the smallest standard deviation of the Gaussian kernel.

(2) Edge recovery.

The main role of the RGF is reflected in the edge recovery of the blurred image G and, in turn, the iterative update of images G_A and G_B. In Eqs. (19) and (20), K_A^1 and K_B^1 are initially set as G_A and G_B, respectively, which are the outputs of Gaussian filtering. Denoting K^{N+1} as the result of the N-th iteration, the iterative processing can be defined as follows:

$$K_A^{N+1}(\mathrm{m}) = \frac{1}{U}\sum_{n\in S}\exp\left(-\frac{\|m-n\|^2}{2\sigma_s^2} - \frac{\|K^N(m)-K^N(n)\|^2}{2\sigma_r^2}\right)\mathrm{I}_b^A(n), \tag{21}$$

$$K_B^{N+1}(\mathrm{m}) = \frac{1}{U}\sum_{n\in S}\exp\left(-\frac{\|m-n\|^2}{2\sigma_s^2} - \frac{\|K^N(m)-K^N(n)\|^2}{2\sigma_r^2}\right)\mathrm{I}_b^B(n) \tag{22}$$

where $U = \sum_{n\in S}\exp\left(-\|m-n\|^2/2\sigma_s^2\right) - \left(\|K^N(m)-K^N(n)\|^2/2\sigma_r^2\right)$ is used for normalization, and σ_s and σ_r control the domain and range weights, respectively.

Iterative operations have excellent performance in preserving maximum scale edges and smoothing irrelevant details. In this paper, the output of the RGF represented by K^{N+1} is an augmented result that consists of enhanced features such as edges, sharp surfaces, and textures.

(3) Structure detection.

A commonly used approach for structure detection is to approximate the gradient magnitude by absolute values:

$$W = \left|\frac{\partial K}{\partial x}\right| + \left|\frac{\partial K}{\partial y}\right| \tag{23}$$

where $\times$ and y are the spatial domain coordinates, and the digital difference is

$$\left|\frac{\partial K}{\partial x}\right| = K(x+1) - K(x). \tag{24}$$

By comparing the magnitudes of K_A and K_B, we can obtain a decision diagram:

$$D = W_A - W_B \tag{25}$$

where W_A and W_B are the sizes of images K_A and K_B, respectively. Typically, regions with significant variation produce larger D values, whereas flatter regions produce smaller D indicators. Therefore, in the decision map D, if the pixel in K_A changes drastically, it is proved to have a large positive value, whereas a smaller negative value indicates a drastic change in the pixel in K_B.

Sharply changing pixels represent the large-scale structure of an image; that is, the pixels have a larger gradient magnitude. To ensure that a pixel with a larger gradient magnitude influences the decision of its neighborhood more, we introduce a mean filter $\overline{D}_m$ as a means of spreading the sharply changed pixel to its neighborhood:

$$\overline{D}_m = \frac{1}{|\Omega|}\sum_{m\in\Omega_m} D_m,\ \forall \mathrm{m} \in \Omega_m \tag{26}$$

The saliency structure L^A is a binary matrix

$$L_m^A = step(\overline{D}_m) \tag{27}$$

where if the function step $(\cdot)$ returns one for an element of L^A, then the element corresponding to $\overline{D}$ is positive; otherwise, it is negative. In the structure matrix L^A, a returned value of one indicates a sharp change in the pixels in K_A; otherwise, it indicates a sharp change in the pixels in K_B.

Step 2. Iterative DTF.

The DTF can be used to retrieve small-scale information adjacent to the large-scale structure, making the final fused image sharper and producing more realistic patches distributed along the structure. Because the details recovered by filtering the image only once are limited, Li et al. [55] improved the performance of the DTF by modifying Eq. (7) to a domain transform filter with an iterative mechanism:

$$L_m^{t+1} = DTF\left(L_m^t, K_m^{N+1}, \sigma_s, \sigma_r\right), \tag{28}$$

where t represents the number of iterations.

The output of the domain transform filter $\widehat{I}_b$ after T iterations is denoted by L^T, and the selected base layer component can be obtained by

$$\widehat{I}_b = L^T I_b^A + \left(1 - L^T\right) I_b^B. \tag{29}$$

3.3. Fused result

The final fusion result F can be obtained by superimposing the above fused base and detail layers:

$$F = \widehat{I}_b + \widehat{I}_d. \tag{30}$$

4. Experiments

4.1. Experimental setup

To test the effectiveness of the proposed algorithm, infrared and visible images were extracted from the TNO Image Fusion Dataset (figshare.com), which contains 105 sets of images of different scenes. In addition, we randomly selected six classical image pairs from this dataset as the display images for our qualitative comparison (see Fig. 3).

All experiments were implemented in a MATLAB 2021b environment on a PC with an AMD Ryzen 5 4600H with Radeon Graphics processor and NVIDA GeForce GTX 1650 graphics card. Additionally, the redundant dictionary in SR was obtained by pre-training on eight high-quality

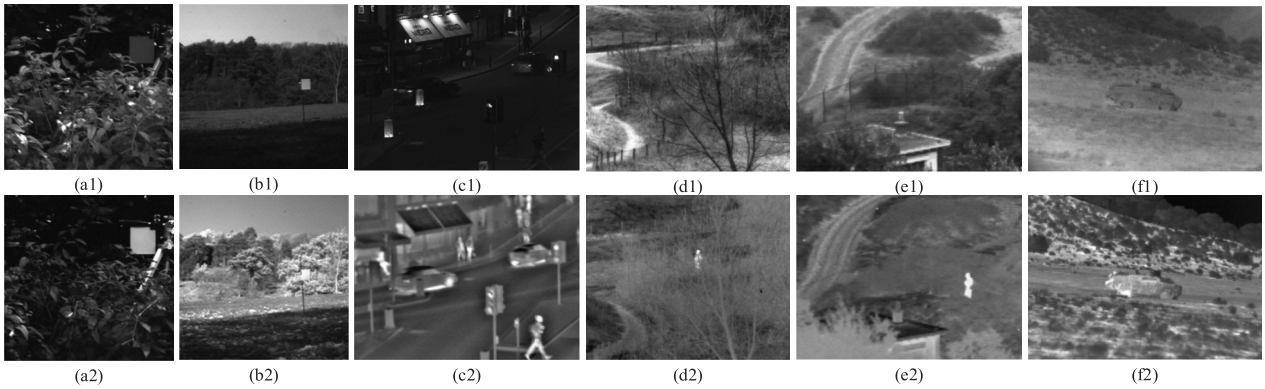

Fig. 3. Source images. (a1)–(f1) are visible images: (a1) leaves, (b1) The woods, (c1) Bristol Queens Road, (d1) field, (e1) UN Camp, and (f1) car. (a2)–(f2) are the infrared images corresponding to (a1)–(f1), respectively.

natural images (see Fig. 4).

4.1.1. Quantitative evaluation metrics

The quality evaluation of fused images has always been a difficult problem because a single metric can only reflect limited aspects of the fused images. Therefore, the general practice is to use multiple metrics for a comprehensive evaluation. Liu et al. [56] conducted a detailed analysis and comparison of quality evaluation metrics for fused images and classified the mainstream metrics into four categories: (1) information theory-based, (2) image feature-based, (3) image structural similarity-based, and (4) human perception inspired fusion metrics. In our experiments, we selected eight representative metrics for evaluation, which cover these four major categories: (1) nonlinear correlation information entropy Q_{NCIE} [57], (2) gradient-based fusion performance Q_G [58], (3) image fusion metric based on a multiscale scheme Q_M [59], (4) image fusion metric based on phase congruency Q_P [60], (5) Piella's metric Q_S [61], (6) Chen–Blum metric Q_{CB} [62], (7) entropy EN [63], and (8) visual information fidelity VIF [64,65].

Among them, Q_{NCIE} measures the nonlinear correlation entropy in the fused image. Q_S aims to calculate the extent of structural information preservation between the source image and the fused image. Q_G is generally used to evaluate the amount of edge information. Q_M is used to measure the degree of edge preservation in different scale spaces of the image. Q_{CB} can evaluate the fused image from the perspective of visual saliency. EN measures the amount of information contained in the fused image, and VIF can evaluate the visual information fidelity of fused images based on the statistical theory of natural scenes. In addition, the larger the values of these fusion quality metrics, the better the fusion performance of the evaluated method.

4.1.2. Comparison methods

We compared the proposed algorithms qualitatively and quantitatively with 11 state-of-the-art image fusion methods. For better comparisons, we divided the compared methods into two main categories, that is, traditional methods: ASR [34], CBF [22], LatLRR [42], MLGCF [23], SD [13], and VSM-WLS [26] and deep learning-based methods: GANMcC [66], MFEIF [67], Perception-GAN [68], SDDGAN [69], and U2Fusion [70]. The parameter settings of all comparison methods were consistent with those in the corresponding published literature.

4.2. Parameter analysis

For our proposed method, the iteration number t of the DTF in Eq. (28), ∂_s and ∂_r in Eq. (21), and error tolerance ε in Eq. (9) were set to $\{t = 5, \partial_s = 3, \partial_r = 0.05, \varepsilon = 0.01\}$. ∂_s and ∂_r in Eq. (28) were set to $\{\partial_s = 10,\ \partial_r = 0.8\}$. Two key parameters in the proposed model were

Fig. 4. Training data of dictionary learning.

discussed: the number of RGF iterations N in Eq. (21) and regularization parameter β in Eq. (10). We used the control variable method for parameter analysis (i.e., the values of other parameters are fixed when analyzing one of the parameters). To verify the quality of the fusion image under different parameters, three groups of IR and VIS source images of different types were selected for testing.

4.2.1. Low-pass filter regularization parameter β

To analyze the optimal value of the regularization parameter β in the low-pass filter, the number of iterations of the RGF $N = 2$ was first fixed, and then, β was changed from one to six to generate a series of fusion results. Because it is difficult to distinguish the quality differences of these fusion results clearly using subjective vision alone, the above eight metrics were used to evaluate the fusion results quantitatively. The evaluation results are listed in Table 1. As shown in Table 1, when β was set to one, the metrics in 14 places achieved the maximum value, which is much higher than the ranking of metrics corresponding to other β values. This indicates that the information content in the fused image produced at this time is the richest, and the fusion quality is the best. Therefore, the value of the parameter β can be set to one.

4.2.2. Number of RGF iterations N

To analyze the optimal value of the number of iterations N, we first fixed β as one, and then, N was changed from two to eight to generate a series of fusion results. Similarly, the above eight metrics were used to evaluate the fusion results, and the corresponding results are listed in Table 2. It can clearly be seen that when N was set to two, the metrics in seven places achieved the maximum value, indicating that the fusion is the best in this time. Thus, we take N to be two. In summary, the optimal values of these two parameters are$\{\beta = 1,\ N = 2\}$.

4.3. Ablation analysis

4.3.1. Multiscale enhancement techniques

In SR, we perform multiscale detail enhancement on the training samples to improve the feature extraction capability of SR, and to verify whether the scheme is effective, we perform an ablation study. We verify whether the multiscale enhancement technique can improve the fusion performance of the model by comparing the fusion results obtained using the data before and after enhancement. The scores of the objective evaluation metrics corresponding to the two methods are listed in Table 3, where we label the model without the multiscale enhancement technique as Original and the original model that uses the technique as the enhanced model. It can be clearly seen that the fused images obtained from the dictionary with enhancement training have a significant advantage in the eight metrics compared to the dictionary without enhancement training. From images (c) and (d) in Table 3, we can see that the enhanced dictionary has eight metrics with maximum values, while for images (a), (e), and (f), seven metrics have the maximum values. Combining the experimental results in Table 3, we can conclude that the multiscale approach can effectively enhance the details in different scales, which improves the algorithm's ability to capture and preserve more diverse and informative features. This is particularly beneficial in scenarios where the fused image must maintain high fidelity and retain as much detail as possible, such as in object detection or military reconnaissance applications. Overall, the results demonstrate the effectiveness and practicality of the proposed multiscale detail enhancement technique in enhancing fusion performance and preserving important visual information.

4.3.2. Dtf

We introduced DTF in the fusion of low-frequency components to further refine the obtained saliency map. To validate its effectiveness in optimizing the visual quality and clarity of the fused image, we conducted an ablation study on it, labeling the method that was not processed by DTF as SSE-SR. As shown in Fig. 5, two pairs of source images in the TNO dataset were used to compare the subjective visual quality of these two different methods. It can be observed that the fusion results obtained by the SSE-SR method exhibit many abrupt patches in the image, leading to inconsistent visual effects. This is because it directly uses the saliency map extracted from the salient structure as the decision map for the fusion of low-frequency components, without better handling the patches distributed along the structure and performing a refined classification of the pixels. Therefore, we can conclude that the addition of DTF can effectively improve the fusion performance of the proposed algorithm, making the patches distributed along the structure look more realistic and consistent with the human visual system.

Table 1
Comparison of the regularization parameters for group (a), group (c), and group (e). "NMV" represents the number of maximum values. The best results are presented in bold.

Images	β	Q_{NCIE}	Q_G	Q_M	Q_P	Q_S	Q_{CB}	EN	VIF	NMV
(a)	1	**0.8206**	0.6622	1.4371	**0.8238**	0.9006	**0.7177**	6.7681	**0.5228**	4
	2	0.8196	0.6618	1.4366	0.8216	0.9030	0.7138	6.7718	0.5108	0
	3	0.8191	0.6631	1.4388	0.8224	0.9040	0.7116	6.7738	0.5059	0
	4	0.8187	0.6642	1.4438	0.8217	0.9047	0.7104	6.7736	0.5015	0
	5	0.8185	0.6646	1.4501	0.8216	0.9051	0.7089	6.7758	0.4989	0
	6	0.8184	**0.6656**	1.4598	0.8210	**0.9053**	0.7083	6.7791	0.4972	1
	7	0.8180	**0.6656**	1.4498	0.8203	**0.9053**	0.7080	6.7790	0.4955	2
	8	0.8176	**0.6656**	**1.4600**	0.8198	**0.9053**	0.7083	**6.7792**	0.4940	4
(c)	1	0.8222	**0.6988**	1.4602	**0.5989**	0.8607	**0.6009**	6.7380	**0.4113**	4
	2	0.8213	0.6957	1.5073	0.5833	**0.8611**	0.5959	6.7352	0.3973	1
	3	0.8207	0.6926	1.5696	0.5838	0.8610	0.5936	6.7525	0.3920	0
	4	0.8205	0.6929	1.5812	0.5857	0.8609	0.5922	6.7575	0.3888	0
	5	0.8205	0.6938	1.6130	0.5898	0.8610	0.5919	6.7669	0.3872	0
	6	0.8204	0.6943	1.6278	0.5909	0.8608	0.5910	6.7764	0.3855	0
	7	0.8204	0.6940	1.6340	0.5919	0.8607	0.5907	6.7769	0.3842	0
	8	**0.8203**	0.6941	**1.6572**	0.5929	0.8606	0.5905	**6.7775**	0.3834	3
(e)	1	**0.8077**	**0.4738**	1.0060	**0.3719**	0.7825	**0.6149**	**7.1107**	**0.3883**	6
	2	0.8071	0.4613	0.9904	0.3411	0.7887	0.5938	7.1017	0.3655	0
	3	0.8067	0.4594	0.9941	0.3280	0.7925	0.5850	7.0967	0.3541	0
	4	0.8065	0.4603	0.9951	0.3221	0.7949	0.5799	7.0921	0.3480	0
	5	0.8063	0.4612	0.9915	0.3194	0.7962	0.5760	7.0899	0.3439	0
	6	0.8062	0.4619	1.0002	0.3182	0.7971	0.5734	7.0877	0.3407	0
	7	0.8061	0.4637	1.0063	0.3184	0.7975	0.5724	7.0863	0.3386	0
	8	0.8060	0.4641	**1.0065**	0.3186	**0.7981**	0.5711	7.0844	0.3364	2

Table 2
Comparison of number of iterations for group (a), group (c), and group (e). "NMV" represents the number of maximum values. The best results are in bold.

Images	*N*	Q_{NCIE}	Q_G	Q_M	Q_P	Q_S	Q_{CB}	EN	VIF	NMV
(a)	2	0.82064	0.66216	**1.43706**	0.82385	**0.90065**	0.71773	6.76805	0.52284	2
	3	0.82072	0.66235	1.43350	0.82433	0.90062	**0.71774**	6.76812	0.52283	1
	4	0.82078	0.66251	1.43241	0.82432	0.90061	0.71770	**6.76824**	**0.52289**	2
	5	0.82078	0.66249	1.43492	0.82448	0.90061	0.71759	6.76816	0.52287	0
	6	**0.82079**	0.66246	1.43526	0.82438	0.90059	0.71761	6.76802	0.52282	1
	7	**0.82079**	**0.66253**	1.43277	0.82434	0.90059	0.71765	6.76789	0.52284	2
	8	0.82077	0.66243	1.43561	**0.82452**	0.90058	0.71769	6.76764	0.52288	1
(c)	2	**0.82224**	**0.69880**	1.46022	0.59890	0.86074	0.60085	6.73800	0.41125	2
	3	0.82208	0.69757	1.46084	**0.60125**	0.86063	0.60063	6.74013	0.41122	1
	4	0.82198	0.69693	1.46721	0.60043	0.86062	0.60046	**6.74087**	0.41112	1
	5	0.82202	0.69722	1.46292	0.60072	0.86071	0.60057	6.73899	0.41115	0
	6	0.82192	0.69668	1.47543	0.60057	0.86074	0.60057	6.73877	0.41125	0
	7	0.82181	0.69656	1.48375	0.60079	0.86089	0.60091	6.73656	0.41138	0
	8	0.82176	0.69658	**1.48485**	0.60080	**0.86092**	**0.60099**	6.73575	**0.41141**	4
(e)	2	0.80773	0.47384	1.00598	**0.37189**	**0.78254**	0.61492	**7.11074**	0.38827	3
	3	0.80781	**0.47421**	1.03642	0.36996	0.78188	0.61426	7.10966	0.39014	1
	4	**0.80788**	0.47396	1.04735	0.36954	0.78218	0.61504	7.10970	**0.39058**	2
	5	**0.80788**	0.47379	1.05096	0.36985	0.78236	0.61508	7.10890	0.39047	1
	6	0.80787	0.47363	**1.05159**	0.36961	0.78246	**0.61531**	7.10768	0.39055	2
	7	0.80784	0.47382	1.05128	0.36964	0.78229	0.61502	7.10476	0.39050	0
	8	0.80782	0.47375	1.04788	0.36919	0.78231	0.61495	7.10373	0.39031	0

Table 3
Comparison of the dictionaries for groups (a) to (f). "NMV" represents the number of maximum values. The best results are in bold.

Images	Dictionaries	Q_{NCIE}	Q_G	Q_M	Q_P	Q_S	Q_{CB}	EN	VIF	NMV
(a)	Original	0.8183	0.6539	1.2402	0.8193	**0.9012**	0.7120	6.7615	0.5217	**1**
	Enhanced	**0.8206**	**0.6622**	**1.4371**	**0.8239**	0.9007	**0.7177**	**6.7681**	**0.5228**	**7**
(b)	Original	0.8388	0.6716	2.3287	**0.7938**	0.9199	0.6894	**7.5713**	**0.5275**	**3**
	Enhanced	**0.8389**	**0.6719**	**2.3399**	0.7935	**0.9200**	**0.6895**	**7.5713**	0.5273	**6**
(c)	Original	0.8218	0.6972	1.4201	0.5971	0.8605	0.5997	6.7375	0.4095	**0**
	Enhanced	**0.8222**	**0.6988**	**1.4602**	**0.5989**	**0.8607**	**0.6009**	**6.7380**	**0.4113**	**8**
(d)	Original	0.8203	0.5653	1.4562	0.5385	0.7800	0.7340	7.1289	0.4546	**0**
	Enhanced	**0.8212**	**0.5696**	**1.5119**	**0.5449**	**0.7806**	**0.7358**	**7.1305**	**0.4547**	**8**
(e)	Original	0.8075	0.4664	0.9429	0.3663	0.7807	0.6117	7.1089	**0.3884**	**1**
	Enhanced	**0.8077**	**0.4738**	**1.0060**	**0.3719**	**0.7825**	**0.6149**	**7.1107**	0.3883	**7**
(f)	Original	0.8134	0.5166	0.7608	0.5213	0.8557	0.6823	**7.1802**	0.4526	**1**
	Enhanced	**0.8144**	**0.5261**	**1.0926**	**0.5243**	**0.8577**	**0.6829**	7.1712	**0.4619**	**7**

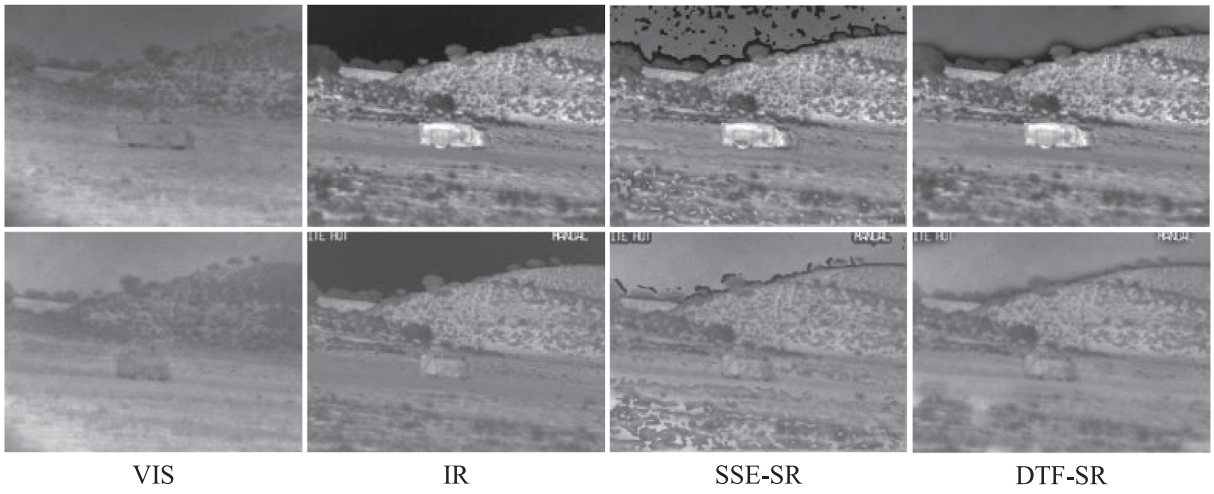

Fig. 5. Fusion results of different methods in the ablation study of DTF.

4.4. Subjective quality assessment

Here, we compare the results of the proposed method with those of the state-of-the-art methods. We first conducted a qualitative comparison of the fusion results. Figs. 7 to 9 show the fusion results produced by six groups of IR and VIS images using 12 different methods.

In Fig. 6, the results of applying the proposed method on the "leaves" and "The woods" source images are shown and compared with those of other methods. As we can see, all methods preserve the main information and characteristics of the source image, but significant differences still exist in the details. Observing the red enlarged area in Fig. 6, it can be found that the fusion results obtained by both CBF and LatLRR

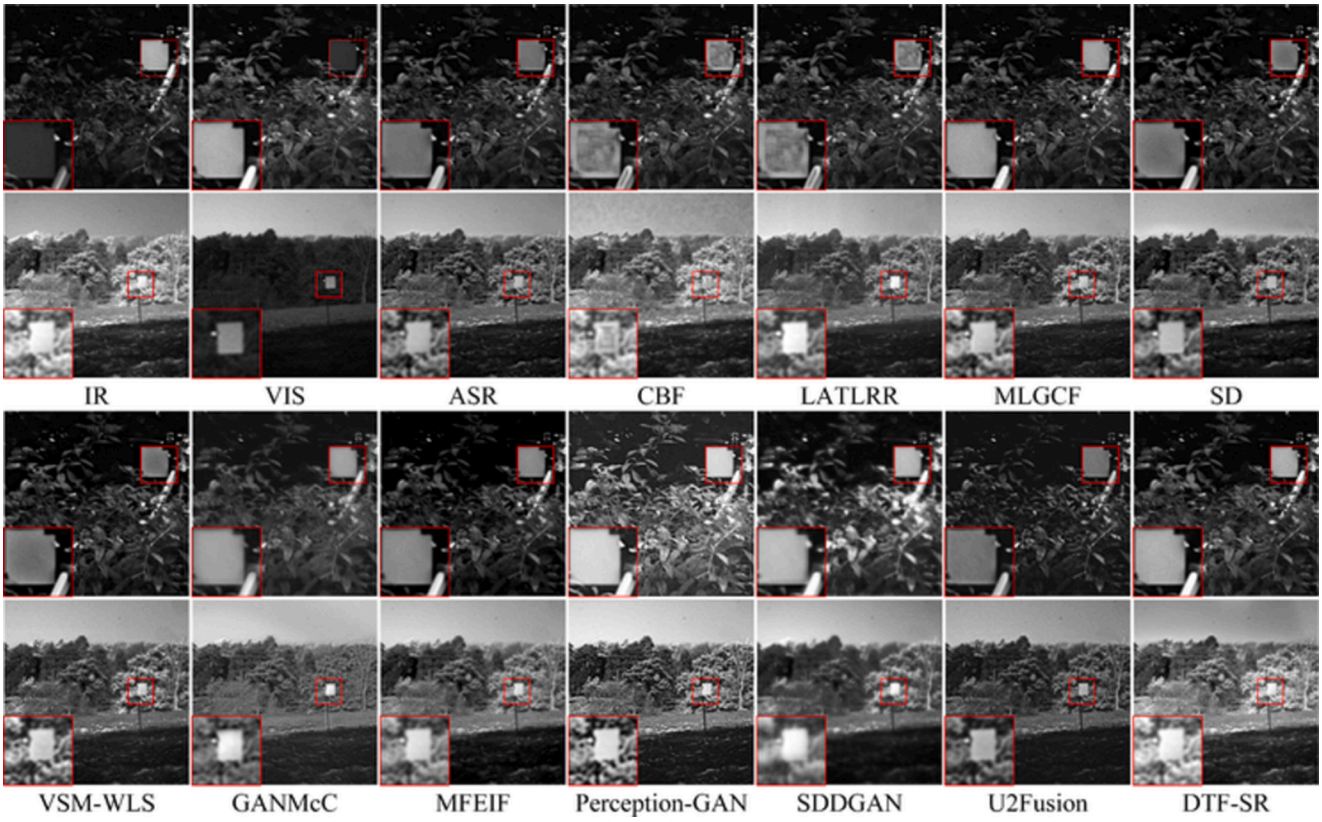

Fig. 6. Fusion results of different algorithms for "leaves" and "The woods" image sets.

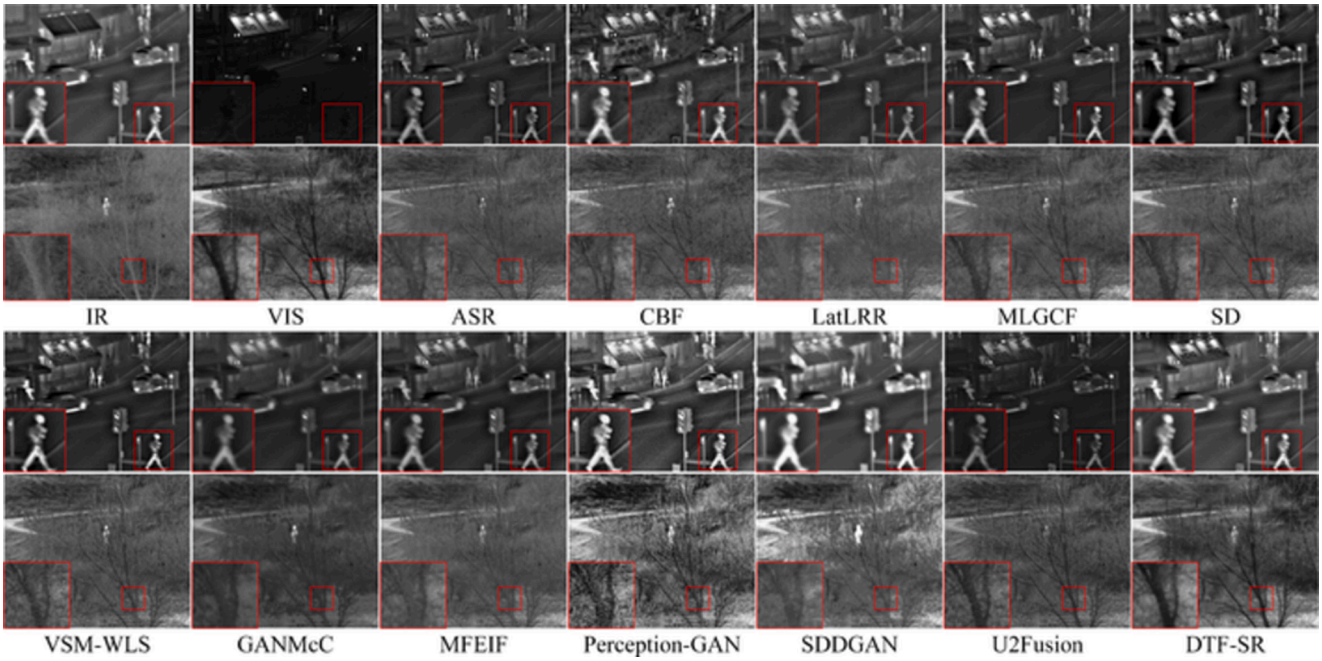

Fig. 7. Fusion results of different algorithms for "Bristol Queens Road" and "field" image sets.

methods generate blocking artifacts and incorrect grayscale information, which reduces the visual quality of the fusion results and is not conducive to the understanding of the scene. In addition, it can be seen that the five methods ASR, MLGCF, SD, VSM-WLS, and U2Fusion are weak in retaining the luminance information on the source image and do not transfer some bright regions and objects in the IR image well into the fusion results, which reduces the contrast of the image. The fusion results obtained by the four methods GANMcC, MFEIF, SDDGAN and Perception-GAN can effectively retain the details in the visible image; however, they have limited ability in recognizing the energy information in the infrared image, which leads to a specific degree of color distortion in the fusion results. Compared with these methods, the

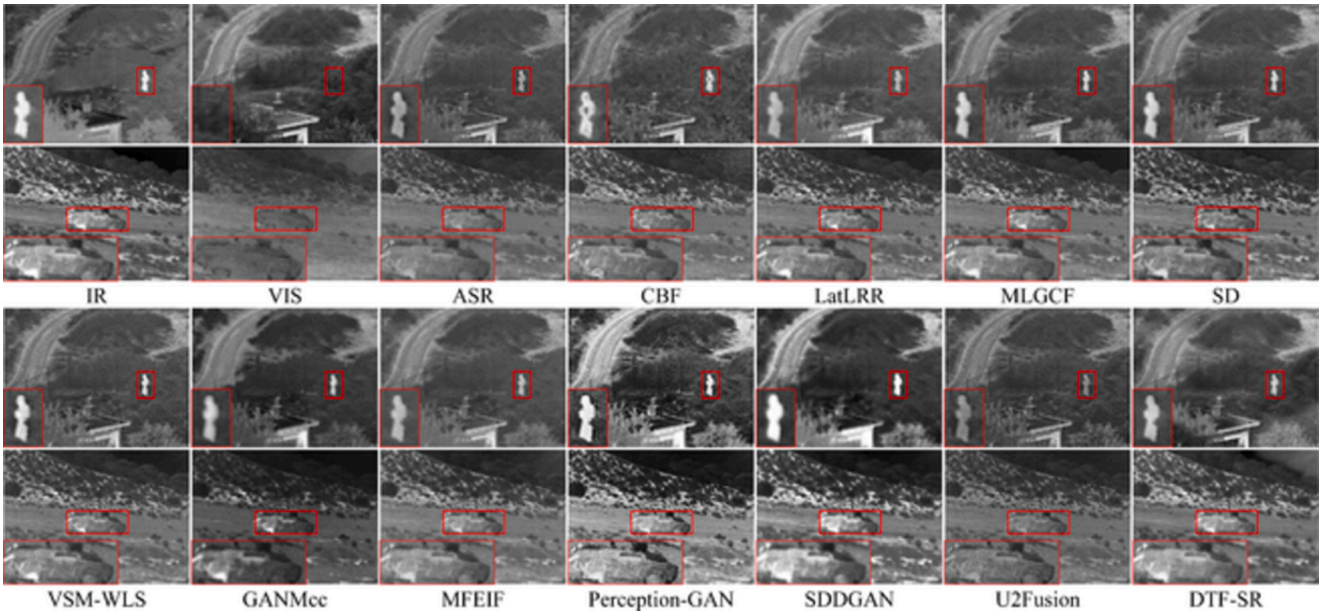

Fig. 8. Fusion results of different algorithms for "UN Camp" and "car" image sets.

proposed approach not only effectively preserves the brightness information of the IR image, but the useful information, such as edges, structures, and details, in the VIS image is also well transferred to the fused image. This demonstrates that the proposed DTF-SR approach performs optimally in this group of fusion experiments.

A visual comparison of different algorithms on the "Bristol Queens Road" and "field" image sets is shown in Fig. 7. For the "Bristol Queens Road" image pair, only the lights and the billboards are visible in the VIS image, whereas the infrared image contains most of the luminance information such as the cars and pedestrians on the road. In the "field" image pair, we can see people hiding in the grass only in the IR image, whereas the visible image contains more detail and texture information. Observing the red magnified region, we can find that the five methods ASR, LatLRR, GANMcC, MFEIF, and U2Fusion are weak in extracting the brightness information of infrared images; for example, the brightness information of pedestrians in the magnified region is lost. Similar to the previous set of experiments, the CBF method still showed blocking artifacts that affected the visual quality, and the Perception-GAN method showed some degree of color distortion. Continuing to observe the tree branches in the red magnified region, it can be found that the five methods LatLRR, MLGCF, SD, VSM-WLS, and SDDGAN do not effectively identify the detail and texture information in the visible image, resulting in the fusion results losing some detail information and not clearly showing the tree trunk outline. The DTF-SR shows that the proposed algorithm can not only effectively retain the luminance region in the IR image, but also some details and textures in the visible image that are difficult to identify can be well transferred to the fusion result. The above analysis proves that the proposed algorithm has a better fusion performance than the 11 state-of-the-art algorithms.

In Fig. 8, the results of applying the proposed method on the "UN Camp" and "car" source images are shown and compared with those of other methods. The two sets of infrared and visible source images shown experimentally demonstrate that the visible images depict the environmental information well, whereas the infrared images emphasize the target information such as people in the thicket and cars in the field. Therefore, the fusion goal of this group of experiments is to retain the luminance information of people and cars and extract as much useful details as possible from the visible images. An observation of the red magnified region reveals that all six methods, that is, ASR, LatLRR, MLGCF, GANMcC, MFEIF, and U2Fusion, are unable to fully extract the luminance information of the salient targets in the infrared images, which reduces the contrast of the fusion results. Although the two methods Perception-GAN and SDDGAN can effectively retain the luminance information, they have limited ability to extract detail information in visible images, resulting in excessive contrast in the fusion results. The CBF method has inappropriate grayscale information in the zoomed region, which affects the visual effect. In summary, the proposed algorithm has better fusion performance in terms of detail retention and color fidelity compared with the current state-of-the-art contrast methods.

4.5. Objective quality assessment

We used eight popular metrics and infrared and visible source images from the TNO dataset to quantitatively compare the different methods, and we present the average scores of the different methods on each metric in Table 4. In addition, in Table 4, the best scores are bolded. The second ranked scores are marked in red, and the third ranked scores are marked in blue. By observing Table 4, we can find that the proposed algorithm obtains the best scores in four metrics, Q_{NCIE}, Q_M, Q_P, and Q_{CB}, which proves that the proposed algorithm has better performance in extracting useful information on the source image, reducing structural distortion, and maintaining reasonable contrast. In addition, the proposed algorithm ranks in the top three for Q_G, EN, and VIF, which indicates that our method has better edge information retention and visual information fidelity than most state-of-the-art image fusion algorithms. The metric EN is used to quantify the amount of information in an image, and a larger EN means that the fused image contains more information. However, if a method generates noise or a large amount of residual information during the fusion process, resulting in undesirable effects such as color distortion in the fusion result, this can also lead to a higher EN value, which explains why the Perception-GAN method has the highest EN value.

Among the traditional methods, the ASR method has four metrics in the top three, and the CBF and SD methods have two metrics in the top three. It can be seen that both the ASR and CBF methods can well maintain the edges in different scale spaces, and the ASR method can provide fusion results with better visual effects. However, although ASR demonstrates its excellent performance in quantitative comparisons, it has a longer running time and possesses a higher computational complexity compared to the proposed algorithms. Among the deep learning-based methods, all four methods, that is, MFEIF, Perception-

Table 4
Comparison of proposed fusion method with 11 competing methods in the TNO dataset.

Metrics	Q_{NICE}	Q_G	Q_M	Q_P	Q_S	Q_{CB}	EN	VIF
ASR	0.8035	**0.4979**	0.9852	0.2961	0.8085	0.5075	6.2155	0.4269
CBF	0.8045	0.3527	0.8605	0.1448	0.7386	0.5047	6.4605	0.1976
LatLRR	0.8036	0.4068	0.5774	0.2542	0.8313	0.5344	6.3531	0.3799
MLGCF	0.8041	0.4336	0.6413	0.2601	0.8286	0.4996	6.4717	0.3390
SD	0.8034	0.4576	0.7356	0.2828	0.8128	0.5579	6.4365	0.2965
VSM-WLS	0.8039	0.4526	0.6312	0.2817	0.8174	0.5541	6.5191	0.3046
GANMcC	0.8042	0.3380	0.5011	0.2263	0.7122	0.4890	6.6752	0.3563
MFEIF	0.8046	0.4177	0.5648	0.3081	0.8261	0.5358	6.4392	**0.4352**
Perception-GAN	0.8035	0.1968	0.2232	0.2395	0.5132	0.5043	**7.6176**	0.1795
SDDGAN	0.8037	0.3096	0.4611	0.1797	0.7040	0.5108	7.2932	0.2617
U2Fusion	0.8039	0.4225	0.5412	0.2657	**0.8333**	0.5592	6.2677	0.3365
DTF-SR	**0.8078**	0.4679	**1.2080**	**0.3098**	0.8241	**0.5626**	6.9043	0.4247

GAN, SDDGAN, and U2Fusion, have different metrics in the top three, also demonstrating the superiority of deep learning-based methods. However, we can conclude that although deep learning-based methods have been commonly used for image fusion in recent years, the results summarized in Table 4 show that such deep learning-based methods (e.g., GANMcC, MFEIF, Perception-GAN, SDDGAN, and U2Fusion) do not perform better than our proposed algorithm Therefore, the above analysis leads to the same conclusion as our qualitative comparison that the proposed algorithms outperform the 11 state-of-the-art algorithms compared for fusion on IR and VIS images.

5. Conclusion

In this paper, we proposed a fusion method of IR and VIS images based on DTF and SR. This method effectively integrates the advantages of spatial domain and SR and overcomes the limitations of existing image fusion methods—inability to retain weak structures and fine edges of source images. First, the source image is decomposed to obtain the base and detail layers. For the fusion of the detail layer, we designed SR fusion rules based on a multiscale detail-enhanced dictionary to improve the compactness of the dictionary and feature recognition ability. For the fusion of the base layer, we introduced a strategy based on SR with DTF. In this manner, the fused image can better retain the heat source information in the IR image, along with the detail and edge information in the VIS image. The experimental results show that the DFT-SR method has significant advantages in both subjective visual evaluation and objective quantitative evaluation compared with 11 state-of-art methods, demonstrating the effectiveness of our method.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The source code of the proposed method is released at https://github.com/ixilai/DTF-SR.

Acknowledgment

This research was supported by the National Natural Science Foundation of China under Grant (Nos. 62201149, 62271148), Research and Development Projects in Key Areas of Guangdong Province (No. 2020B1111040001), China, the Ji Hua Laboratory Grant (No. X200051UZ200) of Guangdong province, China, and the Research Fund of Guangdong-Hong Kong-Macao Joint Laboratory for Intelligent Micro-Nano Optoelectronic Technology (No. 2020B1212030010), China. The authors would like to thank the editors and anonymous reviewers for their constructive and valuable comments and suggestions.

References

[1] A.A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Information Fusion 8 (2) (2007) 114–118.
[2] M. Yin, X. Liu, Y. Liu, et al., Medical image fusion with parameter-adaptive pulse coupled neural network in nonsubsampled shearlet transform domain, IEEE Trans. Instrum. Meas. 68 (1) (2019) 49–64.
[3] Y. Li, J. Zhao, et al., Medical image fusion method by deep learning, Int. J. Cognitive Computing in Eng. 2 (2021) 21–29.
[4] J. Du, W. Li, H. Tan, Three-layer medical image fusion with tensor-based features, Inf. Sci. 525 (2020) 93–108.
[5] Y. Liu, et al., Recent advances in sparse representation based medical image fusion, IEEE Instrum. Meas. Mag. 24 (2) (2021) 45–53.
[6] H. Li, B.S. Manjunath, S.K. Mitra, Multisensor image fusion using the wavelet transform, Graph Models and Image Process. 57 (3) (1995) 235–245.
[7] P.J. Burt, E.H. Andelson, The Laplacian pyramid as a compact image code, Reading in Computer Vision. 1987 671-679.
[8] H. Xu, Y. Wang, Y. Wu, Y. Qian, Infrared and multi-type images fusion algorithm based on contrast pyramid transform, Infrared Phys. Technol. 78 (2016) 133–146.
[9] J.J. Lewis, R.J. O'Callaghan, S.G. Nikolov, D.R. Bull, C.N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Information Fusion. 8 (2) (2007) 119–130.
[10] C. Shi, Q. Miao, P. Xu, A novel algorithm of remote sensing image fusion based on Shearlets and PCNN, Neurocomputing 117 (2013) 47–53.
[11] G.G. Bhutada, R.S. Anand, S.C. Saxena, Edge preserved image enhancement using adaptive fusion of images denoised by wavelet and curvelet transform, Digital Signal Process. 21 (1) (2011) 118–130.
[12] A.L. Da Cunha, J. Zhou, M.N. Do, The nonsubsampled contourlet transform: theory, design, and applications, IEEE Trans. Image Process. 15 (10) (2006) 3089–3101.
[13] D.P. Bavirisetti, R. Dhuli, Two-scale image fusion of visible and infrared images using saliency detection, Infrared Phys. Technol. 76 (2016) 52–64.
[14] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Information Fusion 36 (2017) 191–207.
[15] Y. Zhang, Y. Liu, et al., IFCNN: A general image fusion framework based on convolutional neural network, Information Fusion 54 (2020) 99–118.
[16] I. Goodfellow, J. Pougetabadie, M. Mirza, B. Xu, D. Wardefarley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets (2014) 2672–2680.
[17] J. Ma, W. Yu, P. Liang, C. Li, J. Jiang, Fusiongan: A generative adversarial network for infrared and visible image fusion, Information Fusion 48 (2019) 11–26.
[18] H. Zhang, Z. Le, Z. Shao, H. Xu, J. Ma, Mff-gan: An unsupervised generative adversarial network with adaptive and gradient joint constraints for multi-focus image fusion, Information Fusion 66 (2020) 40–53.
[19] H. Xu, et al, Mef-gan: Multi-exposure image fusion via generative adversarial networks, IEEE Trans. Image Process. (99) (2020) 1-1.
[20] J. Ma, et al., Pan-GAN: An unsupervised pan-sharpening method for remote sensing image fusion, Information Fusion 62 (2020) 110–120.
[21] X. Li, et al., Multi-focus image fusion based on nonsubsampled contourlet transform and residual removal, Signal Process 184 (4) (2021) 108–1062.
[22] S. Kumar, B. K, Image fusion based on pixel significance using cross bilateral filter, SIViP 9 (5) (2015) 1193–1204.
[23] Wei, et al., Infrared and visible image perceptive fusion through multi-level Gaussian curvature filtering image decomposition, Appl. Opt. 58 (12) (2019) 3064–3073.
[24] X. Li, et al., Multimodal medical image fusion based on joint bilateral filter and local gradient energy, Inf. Sci. 569 (2021).
[25] S.T. Li, X.D. Kang, J.W. Hu, Image fusion with guided fifiltering, IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc. 22 (7) (2013) 2864–2875.
[26] J. Ma, et al., Infrared and visible image fusion based on visual saliency map and weighted least square optimization, Infrared Phys. Technol. 82 (2017) 8–17.
[27] J. Wright, et al., Robust face recognition via sparse representation, IEEE Trans. Pattern Anal. Mach. Intell. 31 (2) (2009) 210–227.
[28] T. Guha, R.K. Ward, Learning sparse representations for human action recognition, IEEE Trans. Pattern Anal. Mach. Intell. 34 (8) (2012) 1576–1588.
[29] H. Li, et al, Attribute-Aligned Domain-Invariant Feature Learning for Unsupervised Domain Adaptation Person Re-Identification, IEEE Transactions on Information Forensics and Security. 99 2020 1-1.

[30] H. Li, et al, Jointly Learning Commonality and Specificity Dictionaries for Person Re-Identification, IEEE Transactions on Image Processing. 99 2020 1-1.
[31] X. Yuan, X. Liu, S. Yan, Visual classification with multitask joint sparse representation, IEEE Trans. Image Process. 21 (10) (2012) 4349–4360.
[32] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Meas. 59 (4) (2010) 884–892.
[33] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inform. Fusion 13 (1) (2012) 10–19.
[34] Y. Liu, Z. Wang, Simultaneous image fusion and denoising with adaptive sparse representation, Image Processing Iet. 9 (5) (2015) 347–357.
[35] Z. Zhu, Y. Chai, H. Yin, J. Zhou, Z. Liu, A novel dictionary learning approach for multi-modality medical image fusion, Neurocomputing 214 (2016) 471–482.
[36] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inform. Fusion 24 (2015) 147–164.
[37] Y. Liu, et al., Medical image fusion via convolutional sparsity based morphological component analysis, IEEE Signal Process Lett. 26 (3) (2019) 485–489.
[38] X. Li, et al., Joint image fusion and denoising via three-layer decomposition and sparse representation, Knowl.-Based Syst. 224 (1) (2021) 107–1087.
[39] R. Hou, et al., VIF-Net: an unsupervised framework for infrared and visible image fusion, IEEE Trans. Comput. Imaging. 6 (2020) 640–651.
[40] Z. Ding, et al., CMFA_Net: a cross-modal feature aggregation network for infrared-visible image fusion, Infrared Phys. Technol. 118 (2021), 103905.
[41] Z. Ding, et al., A robust infrared and visible image fusion framework via multi-receptive-field attention and color visual perception, Appl. Intell. 53 (7) (2023) 8114–8132.
[42] H. Li, X. Wu, Infrared and visible image fusion using Latent Low-Rank Representation, arXiv preprint arXiv:1804.08992 (2018).
[43] P. Perona, J. Malik, Scale-space and edge detection using anisotropic diffusion, IEEE Trans. Pattern Anal. Mach. Intell. 12 (7) (1990) 629–639.
[44] Y. Jiang, M. Wang, P-M equation based multiscale decomposition and its application to image fusion, Pattern Anal. Appl. 17 (1) (2014) 167–178.
[45] Q. Wang, et al., Robust multi-modal medical image fusion via anisotropic heat diffusion guided low-rank structural analysis, Information fusion 26 (2015) 103–121.
[46] C. Tomasi, R. Manduchi, Bilateral filtering for gray and color images, Int. Conference on Computer Vision IEEE. (1998).
[47] E. Gastal, M.M. Oliveira, Domain transform for edge-aware image and video processing, ACM Trans. Graph. 30 (4) (2011) 1–12.
[48] G. Petschnigg, et al., Digital photography with flash and no-flash image pairs, ACM Trans. Graph. 23 (3) (2004) 664–672.
[49] A.M. Bruckstein, D.L. Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images, SIAM Rev. 51 (1) (2009) 34–81.
[50] X. Chang, Z. Ma, M. Lin, Y. Yang, A.G. Hauptmann, Feature interaction augmented sparse learning for fast Kinect motion detection, IEEE Trans. Image Process. 26 (8) (2017) 3911–3920.
[51] X. Chang, Z. Ma, Y. Yang, Z. Zeng, A.G. Hauptmann, Bi-level semantic representation analysis for multimedia event detection, IEEE Trans Cybern. 47 (5) (2017) 1180–1197.
[52] X. Chang, Y. Yang, Semisupervised feature analysis by mining correlations among multiple tasks, IEEE Trans. Neural Network and Learning Systems. 28 (10) (2017) 2294–2305.
[53] B. Goyal, et al., Measurement and analysis of multi-modal image fusion metrics based on structure awareness using domain transform filtering, Measurement 182 (2021) 109–663.
[54] Q. Zhang, et al, Rolling Guidance Filter, European Conference on Computer Vision Springer International Publishing. 2014.
[55] W. Li, et al., Structure-aware image fusion, Optik – Int. J. Light and Electron Optics. 172 (2018) 1–11.
[56] Z. Liu, E. Blasch, Z. Xue, et al., Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision: a comparative study, Pattern Analysis and Machine Intelligence, IEEE Trans. on. 34 (1) (2012) 94–109.
[57] Q. Wang, Y. Shen, J. Jin, Performance evaluation of image fusion techniques, Image Fusion: Algorithms and Applications. 19 (2008) 469–492.
[58] C.S. Xydeas, V.S. Petrovic, Objective image fusion performance measure, Electron. 36 (4) (2000) 308–309.
[59] P. Wang, B. Liu, A novel image fusion metric based on multi-scale analysis, Int. Conference on Signal Processing IEEE. (2008) 965–968.
[60] J. Zhao, et al., Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement, Int. J. Innovative Computing Information & Control Ijicic. 3 (6) (2006).
[61] G. Piella, H. Heijmans, A new quality metric for image fusion, Int. Conference on Image Processing IEEE. (2003).
[62] C. Yin, R.S. Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (10) (2009) 1421–1432.
[63] J. Zhao, et al., Fusion of visible and infrared images using global entropy and gradient constrained regularization, Infrared Phys. Technol. 81 (2017) 201–209.
[64] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.
[65] Y. Han, et al., A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2) (2013) 127–135.
[66] J. Ma, et al, GANMcC: A Generative Adversarial Network With Multiclassification Constraints for Infrared and Visible Image Fusion, IEEE Transactions on Instrumentation and Measurement. PP.99 2020 1-1.
[67] J. Liu, et al, Learning a Deep Multi-scale Feature Ensemble and an Edge-attention Guidance for Image Fusion, IEEE Transactions on Circuits and Systems for Video Technology. PP.99 2021 1-1.
[68] F. Yu, et al., Image fusion based on generative adversarial network consistent with perception, Information Fusion. 72 (2021) 110–125.
[69] H. Zhou, et al., Semantic-supervised Infrared and Visible Image Fusion via a Dual-discriminator Generative Adversarial Network, IEEE Trans. Multimedia 1–1 (2021).
[70] H. Xu, et al., U2Fusion: a unified unsupervised image fusion network, IEEE Trans. Pattern Anal. Mach. Intell. 44 (1) (2020) 502–518.