Multimedia Tools and Applications (2023) 82:36303–36323
https://doi.org/10.1007/s11042-023-14967-0

LatLRR-CNN: an infrared and visible image fusion method combining latent low-rank representation and CNN

Yong Yang[1] · Chengrui Gao[1] · Zhangqiang Ming[1] · Jixiang Guo[1] · Edou Leopold[1] · Junlong Cheng[1] · Jie Zuo[1] · Min Zhu[1]

Received: 3 March 2022 / Revised: 1 June 2022 / Accepted: 22 February 2023 /
Published online: 17 March 2023

Abstract
While infrared images have prominent targets and stable imaging, it can hardly maintain such detailed information or quality as texture or resolution. In contrast, although visible images have rich texture information and high resolution, the imaging is easily disturbed by the circumstance. Therefore, it is desirable to make up for shortcomings and integrate the advantages of the two images into one. In this paper, we propose an infrared and visible image fusion method that combines latent low-rank representation(LatLRR) and convolutional neural network(CNN), termed as LatLRR-CNN. This method can prevent loss of information, lack of imaging quality, and designing complex fusion rules or networks. Firstly, LatLRR is used to decompose infrared or visible images into low-rank parts and salient parts. Secondly, these two parts are fused separately using CNN. Finally, the fused low-rank part and the fused salient part are merged to obtain the fused image. Experiments on publicly accessible datasets reveal that our method outperforms state-of-the-art methods in terms of objective metrics and visual effects. Specifically, the average of our method on the Nato sequence, EN reaches 7.59, MI reaches 2.89, SD reaches 57.77, and VIf reaches 0.51.

Keywords Image fusion · Latent low-rank representation · Convolutional neural network · Infrared image · Visible image

1 Introduction

Nowadays, with the pace of informatization, more and more sensors are used to obtain data. Multiple sensors of different types are often deployed in one place to capture more

Project supported by the National Key Research and Development Project of China (JG2018190).

Chengrui Gao, Zhangqiang Ming, Jixiang Guo, Edou Leopold, Junlong Cheng and Jie Zuo are contributed equally to this work.

✉ Min Zhu
zhumin@scu.edu.cn

Extended author information available on the last page of the article.

accurate and comprehensive information. However, there is plenty of redundant information in the images acquired by these sensors, and storing these images at the same time wastes storage space [2]. Moreover, the data obtained by different sensors is distributed in different pictures, which is not conducive to subsequent work. Therefore, image fusion technology has received more and more attention. It can extract the most significant parts from the images acquired by different sensors and then combines them into a picture. Hence, the image contains more descriptive information of the scene, which is more convenient for subsequent applications [45].

Among different sensors, the combination of infrared and visible sensors has unique advantages [38]. The infrared sensor captures the object's thermal radiation, so the infrared images are not easily affected by the complex environment. They can better distinguish the thing from the background but lack detailed information. The visible sensor captures the reflected light of the circumstance, imaging is easily affected by the event. Still, the visible images cover a wealth of detailed information, in line with the human eye's perception. The fusion of these two images can simultaneously obtain almost all inherent attributes of the scene, including detailed information and target information, so the fusion of infrared and visible images can be applied to more fields than other image fusion categories. Typical infrared and visible image fusion applications have object recognition [32], detection [5], image enhancement [29], surveillance [10], and remote sensing [31].

The existing infrared and visible image fusion methods can be roughly divided into seven types according to the fusion theory, ie., multi-scale transform- [15, 25, 42], sparse representation- [16, 34], neural network- [9, 37], subspace-[1, 8], saliency-based [44, 46] methods, hybrid models [19, 24], and other methods [21, 47]. For the multi-scale transform-based method, the source images are firstly decomposed into different parts. Then the parts are fused according to the corresponding fusion rules. The fusion parts ultimately reconstruct the fusion image. Most multi-scale transform-based methods use artificially formulated fusion strategies and highly rely on multi-scale decomposition methods, making it more difficult to propose a new method. Sparse representation-based methods aim to learn an over-complete dictionary from many high-quality natural images to represent the source images sparsely. The fusion rules act on the sparse coefficient, and then, the learned over-complete dictionary is used to restore the fused images according to the fused sparse coefficient. Coefficient coding technology plays a vital role in fusion performance in sparse representation-based methods. Meanwhile, it is still challenging to construct a suitable over-complete dictionary with good representation ability. The neural network-based method uses the powerful learning ability of the neural network to fuse images. But designing a neural network suitable for image fusion is challenging. The subspace-based method projects the images into a low dimensional space or subspace by removing redundant information in the images because the low-dimensional space helps capture the original images' inherent structure. But there is still a problem that it is difficult to find a powerful representation ability subspace. The saliency-based method retains the salient target area according to the mechanism of the human visual system and can improve the quality of the fused images. However, there has been no accurate answer for how to make full use of the salient area. The hybrid model combines the advantages of the above methods to improve the quality of the fusion images. In addition to the abovementioned methods, some fusion methods use other theories, such as fuzzy theory-based methods [27], Markov random fields-based methods [7], and so on.

Although existing approaches can produce a good fusion effect, keeping both target and detailed information remains challenging. On the one hand, it is difficult for experts to

design comprehensive fusion rules because the performance of multi-scale transform-based methods is heavily reliant on expert experience. Meanwhile, the loss function limits the performance of neural network-based methods, and developing a good loss function is difficult. Considering the shortcomings of these methods, we combine the multi-scale transformation method with the neural network method, using the neural network's learning capabilities to improve the multi-scale transform-based method. Specifically, we combine latent low-rank representation(LatLRR) and convolutional neural network(CNN) to propose a new fusion model, termed LatLRR-CNN. Our method uses a neural network to substitute fusion rules in multi-scale methods, allowing us to avoid manually creating complicated fusion rules while also maximizing the benefits of decomposed images. Furthermore, because we decompose the image once before using the neural network, the features in the decomposed image are rather simple, making the network structure and loss function design easier. In our framework, the source images are decomposed into two parts through LatLRR, the features of the images are initially separated, and then each part is fused with a CNN. Finally, fused parts are added together to get the fusion images.

The contributions of our work include the following aspects:

- We combine LatLRR and CNN to propose a new infrared and visible image fusion model. After the source images are decomposed, the fusion results with good performance can be obtained without a complicated neural network. Our method can obtain a transparent background while retaining the target information.
- The proposed LatLRR-CNN is an end-to-end model, where the fusion images can automatically generate from the input images without manually design complex fusion rules.
- We conduct qualitative and quantitative comparison experiments with state-of-the-art methods on the publicly obtained infrared and visible image fusion datasets. Compared with previous methods, our method can get comparable results.

The rest of this paper is organized as follows. Section 2 briefly introduces the LatLRR theory and then describes several infrared and visible image fusion methods based on deep learning. Section 3 will show the detailed framework of LRR-CNN. The experimental results are exhibited in Section 4. Finally, we conclude in Section 5.

2 Related work

In this section, we briefly introduce the development and basic principles of LatLRR and then review some recent works using deep learning in infrared and visible image fusion.

2.1 Latent low-rank representation

In 2010, Liu et al. [18] proposed a low-rank representation for extracting data from the union of multiple linear subspaces. But in the case of severely damaged or insufficient data, the performance of this method will decrease. Therefore, the author Liu et al. [20] proposed latent low-rank representation theory in 2011 and added unobserved hidden data to construct a dictionary based on the original method. Li et al. [12] presented an infrared and visible picture fusion approach based on this.

Specifically, the LatLRR problem is described as the following optimization problem,

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

where $\lambda > 0$ is the balance coefficient, $\|\cdot\|_*$ represents the nuclear norm, which is the sum of the singular values of the matrix, and $\|\cdot\|_1$ symbolizes l_1-norm. X denotes the observed data matrix, Z is the low-rank coefficient matrix, L is a projection matrix, also named as the significant coefficient matrix. And E is the sparse noisy matrix. Equation (1) is solved by the inexact Augmented Lagrangian Multiplier (ALM) [20] algorithm. Then the salient part is obtained by LX in (1).

2.2 Deep learning-based fusion methods

Due to deep learning's powerful feature extraction ability, various fusion methods have been presented based on it in recent years. Deep learning plays two roles in these methods: one is to utilize deep learning methods to extract features and then manually create fusion rules, and the other is to design an end-to-end fusion network to acquire fusion results by inputting source images directly.

In terms of deep learning feature extraction, Li et al. [13] employed a pre-trained VGG network to extract deep features from source image detail layers in order to create appropriate fusion weight maps, and then performed a fusion of detail layers based on the weight maps. Zhao et al. [48] employed an auto-encoder to decompose the original image, fused the features extracted by the encoder according to certain rules, and then used the decoder to reconstruct the fused image. This is the first time a deep learning algorithm has been used to decompose an image. To better segregate the features of the source image, Fu et al. [3] trained a dual-branch AE network, utilized the channel attention approach to fuse the corresponding regions of the source images, and then used the decoder to acquire the fused image. Yang et al. [40] created a texture conditional generative adversarial network to construct a composite texture map with the source images' texture information. They used adaptive guided filtering to generate a series of decision maps based on the composite texture map and then used the multi-decision maps to rebuild the fused image. Zhang et al. [43] used a guided filter to decompose the source image into high-frequency and low-frequency parts, then processed the high-frequency part with ResNet-152 to obtain the maximum weight layer. Finally, they calculated the fused high-frequency part based on the maximum weight layer to reconstruct the fusion image.

In terms of an end-to-end fusion network, Ma et al. [23] proposed a generative adversarial network for image fusion for the first time, then moved to utilize two discriminators to train the generator in their future work [38], resulting in more stable results. They improved the fusion results in later work [22] by expanding the generative adversarial network and enhancing the loss function. Han et al. [39] used DenseNet as the backbone network for image fusion and applied image quality evaluation to calculate the weight of the corresponding source image in the fused image. Zhang et al. [45] employed two DenseNet networks with the same structure to extract the image's intensity and gradient information, respectively, and mixed the information extracted by the two networks using a route transfer block between the two networks. Zhao et al. [49] trained an AE network with two attention branches, then replaced the decoder with a new augmented fusion model, fixed the encoder parameters, retrained the model, and then used the new model to generate fused images. Wang et al. [35] built an AE network based on Res2Net and then used a dual non-local

attention model to fuse the corresponding feature maps. The fused feature map is sent to the decoder to get the fused image. Yang et al. [41] employed the Fourier transform to extract the image's high-frequency information and then utilized two encoders with the same structure to extract the source image's features and the high-frequency information. Finally, they merge the resulting feature maps and feed them to the decoder to get the final fusion result.

Although they have achieved good results, there are still some shortages: 1) The network is challenging to train when the training data is insufficient; 2) Because image fusion has no ground truth, it cannot directly apply supervised learning. Some approaches can circumvent by designing a content loss function, however, designing a comprehensive loss function is challenging; 3) In the training of GAN-based methods, the proportion of infrared image thermal radiation characteristics in the fusion images are getting lower and lower; 4) To improve the fusion effect, the loss function and network structure will become more and more complex. The above methods either utilize traditional methods or neural network methods, and few methods combine traditional methods and neural networks. Therefore, this paper proposes a new fusion method combining traditional LatLRR and neural network CNN. Our method preliminarily decomposes the features in the source image through the traditional method, which simplifies the task of the neural network and makes it easier for the network to fully learn the features that need to be retained. At the same time, the use of neural networks also avoids the performance degradation caused by inappropriate rule design in traditional methods.

3 Method

In this section, we introduce our proposed method in detail. Firstly, we present our problem formulation and then explain the framework exhaustively. Finally, we introduce the details of network training.

3.1 Problem formulation

To facilitate the neural network to learn the features that need to be preserved, we first decompose the source image using LatLRR. The LatLRR decomposition is shown in Fig. 1. The source image X is decomposed into three parts: low-rank part, salient part and noise. To avoid losing the information of the source image during the decomposition, we use (2) to obtain the salient part and low-rank part.

$$X_s = LX, \quad X_{lrr} = X - X_s \tag{2}$$

Figure 1 shows that the salient part mainly contains the target information of the source image, and the low-rank part is equivalent to the source image with reduced pixel intensity of the target. We utilize two CNNs to effectively integrate the information of distinct parts and assure the reversibility of the decomposition process. One is named the background model, which is used to fuse the low-rank part that mainly retains the gradient information of the image. The other is termed as saliency model, which is used to fuse the salient part principally, maintains the pixel intensity distribution of the image. In the input of the saliency model, we cascade the salient part I_r^s of the infrared image I_r and the salient part I_v^s of the visible image I_v in the channel dimension. So the output of the model is the fused salient part I_f^s. The low-rank parts are processed in the same way to get the fused low-rank part I_f^{lrr}. We consider that the fused salient part I_f^s and the fused low-rank part I_f^{lrr} are

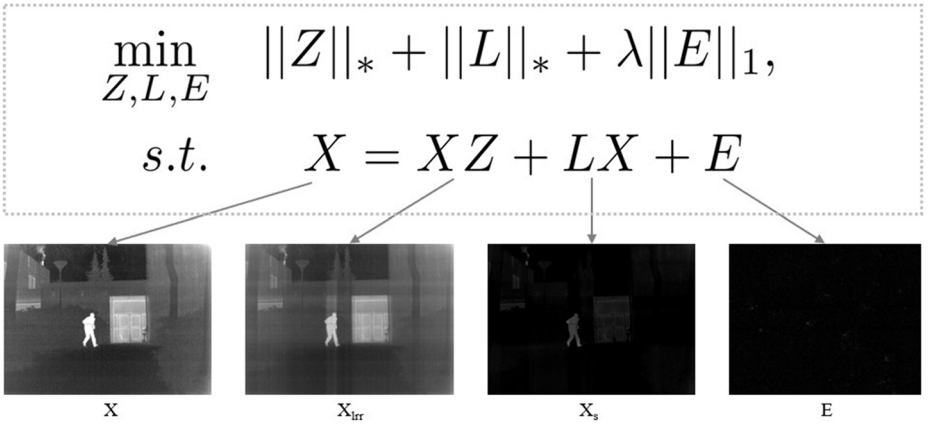

Fig. 1 The Latent low-rank representation

the salient part and low-rank part of the final fused image I_f. Therefore, according to the inverse transform of (2), we can add the fused salient part I_f^s and the fused low-rank part I_f^{lrr} to generate the final fusion image I_f. The overall framework of the method is shown in Fig. 2.

3.2 Network architecture

The background model and the saliency model have the same network architecture based on convolutional neural networks. As shown in Fig. 3, our network is a simple five-layer convolutional neural network. A 5x5 convolution kernel is applied in the first layer. A 3x3

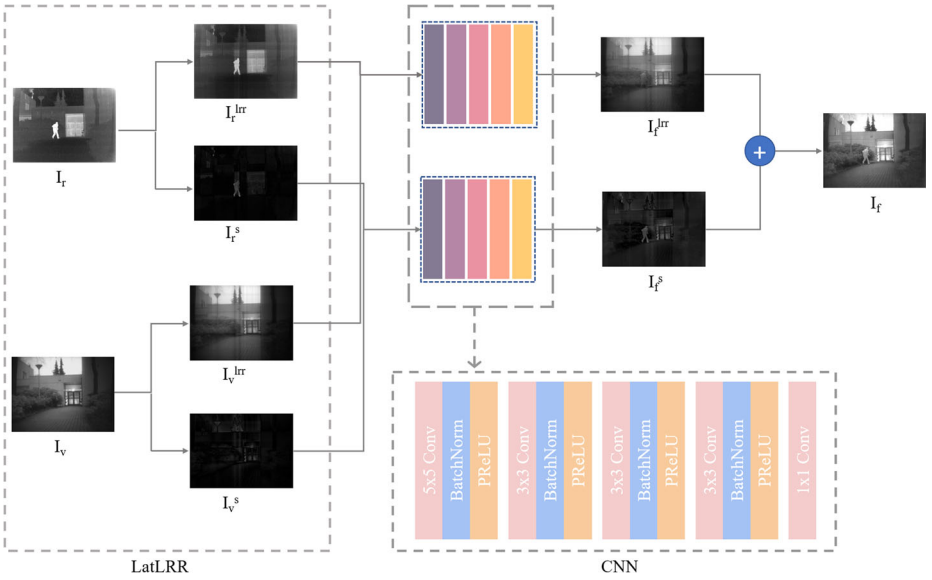

Fig. 2 Framework of the proposed method. The background model and the saliency model use the same network architecture, and the only difference between the two is the different input

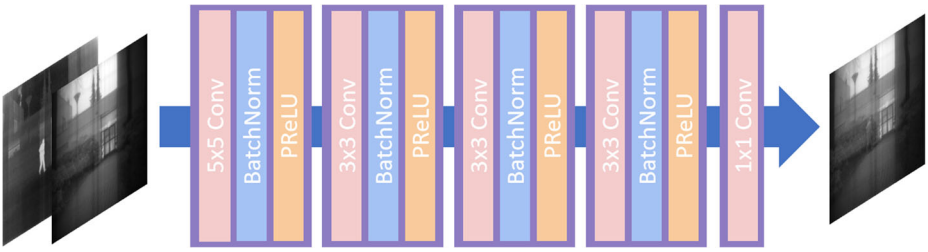

Fig. 3 The network architecture of the salient model and the background model

convolution kernel is used in the second, third and fourth layers, while a 1x1 convolution kernel is used in the last layer. Each layer's stride is set to one. A padding operation is added into each convolution to keep the input and output image sizes consistent, and the reflection filling is used to reduce the influence on the results. For infrared and visible image fusion, downsampling will lose the information of the source images, which is not conducive to obtaining an information-rich fusion result. As a result, we utilize a convolutional layer with no downsampling, which also assures that the input and output images are the exact sizes. We implement batch normalization behind the first four convolution layers to prevent the gradient from vanishing, which can help our network perform more consistently during the training process. Since the last layer of convolution is for dimensionality reduction, batch regularization is not required. We chose PReLU for the first four layers as the activation function and the tanh activation function for the last layer. All parameter settings are shown in Table 1.

3.3 Loss function

The background model and the saliency model have different loss functions. The loss function of the background model consists of two parts:

$$\mathcal{L}_B = \lambda \mathcal{L}_{gradient} + \mathcal{L}_{ssim}^{lrr} \tag{3}$$

where $\mathcal{L}_B$ symbolizes the total loss, the first term $\mathcal{L}_{gradient}^{lrr}$ on the right side of the equation represents the gradient loss, and λ is used to balance $\mathcal{L}_{gradient}$ and $\mathcal{L}_{ssim}^{lrr}$. Since the low-rank part of the source image mainly embodies the background information of the scene, and the gradient contains the texture information of the image, we force the output I_f^{lrr} of

Table 1 Network configuration. KernelS denotes the size of convolutional kernel

InC and OutC are the numbers of input and output channels, respectively

Layer	Size	InC	OutC	Padding	Activation
conv1	5	2	258	Reflection	PReLU
conv2	3	258	128	Reflection	PReLU
conv3	3	128	64	Reflection	PReLU
conv4	3	64	32	Reflection	PReLU
conv5	1	32	1	–	–

the background model to be as similar as the input in gradient. Specifically, $\mathcal{L}_{gradient}$ can be expressed as follows:

$$\mathcal{L}_{gradient} = \frac{1}{HW}(\xi \|\nabla I_f^{lrr} - \nabla I_v^{lrr}\|_F^2 + \|\nabla I_f^{lrr} - \nabla I_r^{lrr}\|_F^2) \tag{4}$$

where H and W represent the height and width of the input images, respectively, $\|\cdot\|_F$ stands for the matrix Frobenius norm, and ∇ means the gradient operator. ξ is a positive parameter to control the trade-off of two terms.

The second term $\mathcal{L}_{ssim}^{lrr}$ depicts the structural similarity index measure(SSIM) loss. The final result needs not only an excellent quantitative evaluation index but also an excellent visual performance. The SSIM [33] index of the fusion result can better meet the human eye's perception, so we expect the SSIM index of the fusion result to be as high as possible. Mathematically, $\mathcal{L}_{ssim}^{lrr}$ is defined as follows:

$$\begin{aligned} \mathcal{L}_{ssim}^{lrr} &= (1 - SSIM_{I_f^{lrr}, I_v^{lrr}}) + \omega(1 - SSIM_{I_f^{lrr}, I_r^{lrr}}) \\ SSIM_{X,F} &= \sum_{x,f} \frac{2\mu_x\mu_f + C_1}{\mu_x^2 + \mu_f^2 + C_1} \cdot \frac{2\sigma_x\sigma_f + C_2}{\sigma_x^2 + \sigma_f^2 + C2} \cdot \frac{\sigma_{xf} + C_3}{\sigma_x\sigma_f + C3} \end{aligned} \tag{5}$$

where ω is used to weigh the similarity between the fused image and the source images, μ is the mean value of the image, σ_x and σ_f denote the variance of the source image and the fused image, and σ_{xf} represents the covariance of the source image and the fused image. C1, C2, and C3 are constants that can avoid the instability caused when the denominator is close to 0.

As for the saliency model, the loss also consists of two parts:

$$\mathcal{L}_D = \beta\mathcal{L}_{intensity} + \mathcal{L}_{ssim}^s \tag{6}$$

where $\mathcal{L}_{ssim}^s$ describes the SSIM loss whose mathematical formula is the same as (5), $\mathcal{L}_{intensity}$ expresses intensity loss, and β is used to equilibrate $\mathcal{L}_{intensity}$ and $\mathcal{L}_{ssim}^s$. The target information of the salient part is more prominent than the low-rank part. Pixel intensity can better characterize the target information. Therefore, we expect the model's output to be as similar as possible to the salient part in terms of pixel intensity. Explicitly, $\mathcal{L}_{intensity}$ can be formulated as follows:

$$\mathcal{L}_{intensity} = \frac{1}{HW}(\|I_f^s - I_v^s\|_F^2 + \gamma\|I_f^s - I_r^s\|_F^2) \tag{7}$$

where the meaning of the symbol is the same as that of (4), and γ is used to control the ratio of the two terms.

3.4 Training details

By removing some image pairs with harsh noise and repeated scenes in the TNO dataset, we get 28 pairs of original images. We trim the authentic images with a step size of 14 to create 30,014 pairs of 120x120 image pairs, which we utilize as our training set to make the network fully trained. During training, the image pairs are directly input into the entire model, and the Adam optimizer is used in the network part, in the testing phase, the model straight outputs the fused images. During training, we set the batch size to 32 and the epoch to 20. The learning rate is set to 10^{-4} and the Adam optimizer is used. Empirically, we set the parameters in the loss function as follows: $\lambda = 8 \times 10^3$, $\xi = 1$, $\omega = 0.4$, $\beta = 75$, $\gamma = 1$. The training procedure is summarized in Algorithm 1.

for *number of training iterations* **do**
  for *steps* **do**
    Select m visible patches $I_v^{(1)}, \cdots, I_v^{(m)}$;
    Utilize LatLRR to get $I_v^{s(1)}, \cdots, I_v^{s(m)}$ and $I_v^{lrr(1)}, \cdots, I_v^{lrr(m)}$;
    Select m infrared patches $I_r^{(1)}, \cdots, I_r^{(m)}$;
    Utilize LatLRR to get $I_r^{s(1)}, \cdots, I_r^{s(m)}$ and $I_r^{lrr(1)}, \cdots, I_r^{lrr(m)}$;
    Update saliency model by AdamOptimizer: $\bigtriangledown \mathcal{L}_D$;
    Update background model by AdamOptimizer: $\bigtriangledown \mathcal{L}_B$;
  end
end

Algorithm 1 Training procedure of LatLRR-CNN.

4 Experiments

To verify the effectiveness of our method, we conduct experiments on the public dataset TNO.[1] The TNO dataset contains multispectral (intensified visual, near-infrared, and longwave infrared or thermal) nighttime imagery of different military relevant scenarios, registered with different multiband camera systems. Base on it, we compare our method with eight state-of-the-art methods, namely, the gradient transfer fusion method(GTF) [21], infrared and visible image fusion based on multi-scale transformation and norm optimization(MTANO) [11], infrared and visible image fusion using deep learning framework (IVIFDLF) [13], MDLatLRR [14], infrared and visible image fusion based on target-enhanced multiscale transform decomposition(IVIFTMTD) [2], FusionGAN [23], a total variation global optimization framework and its application on infrared and visible image fusion(TVGOF) [4], Res2Fusion [36]. All comparison methods are based on public codes and set to the optimal parameters in the original paper. All experiments are implemented with Intel Xeon CPU E5-2680, GPU Tesla P100, and 125GB memory.

4.1 Fusion metrics

Due to the limitations of the human visual system, the perception of each person is different; the qualitative comparison is difficult to explain the pros and cons of the fusion results, so we consider four indicators for quantitative comparison. They are entropy(EN) [30], mutual information(MI) [26], standard deviation(SD) [28], and visual information fidelity(VIf) [6]. EN is based on information theory, which can measure the total information contained in the fusion image. MI describes the degree of correlation between the fusion image and the source image, how much information in the fusion image is converted from the source image. SD is based on statistical concepts, reflecting the distribution and contrast of the fusion image. VIf calculates the distortion of the fused image compared to the source image and can measure the information fidelity of the fused image, which is consistent with the human visual system. The above four indicators are the greater the value, the better the performance.

[1] https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.

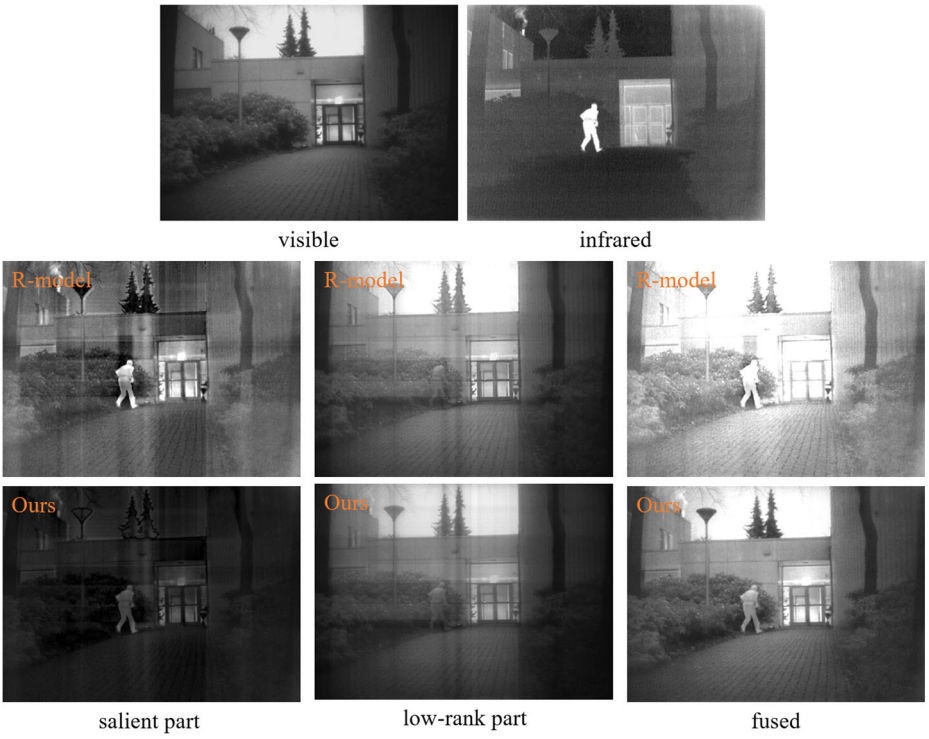

Fig. 4 The fusion results of the R-model and our method. The first line is the source images. The second and third rows are the results of the R-model and our method, respectively, from left to right are the fused salient parts, the fused low-rank parts, and the fusion images

4.2 Validation of LatLRR

LatLRR is an integral part of our method, and it is also an essential difference between our method and other CNN-based methods such as [17]. Therefore, in this section, we focus on verifying the role of LatLRR (Fig. 4).

To begin, we attempt to substitute LatLRR with other classic decomposition techniques. Considering the requirements for image semantics in subsequent processing, we choose the Laplacian pyramid to replace the LatLRR and adjust it to a two-layer pyramid, and retrain

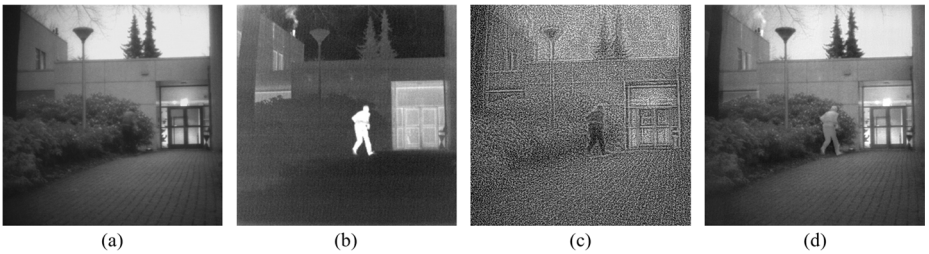

Fig. 5 The fusion results of the L-model and our method. (a-d) correspond to the original visible and infrared images, and the fusion results of L-model and LatLRR-CNN

Fig. 6 Before and after fusion of the L-model detail part. (a-c) corresponding to the infrared image detail part, visible image detail part, and fused detail part, respectively

to obtain L-model. The L-model does not increase the performance of our method, but introduces a lot of noise into the results, as seen in Fig. 5. Because the Laplacian pyramid preserves the image's difference, the fusion of the pyramid's various portions is likely to generate a mismatch, resulting in the occurrence in the figure. Besides, we believe another reason is that the noise included in the detail part of the infrared image is likewise fused when fusing the detail part, as shown in Fig. 6. Above all, LatLRR provides irreplaceable advantages over other standard decomposition methods in our method.

Then, after removing LatLRR and keeping all other settings, we retrain a model called R-model. The fusion result of the R-model and our method on scene kaptein_1123 is shown in Fig. 4. From top to bottom are the original images, the results of the R-model, and the results of our method. From left to right, the second and third rows are the fused salient parts, the fused low-rank parts, and the fused images, respectively. In comparison, the fused salient part of the R-model retains most of the target intensity in the infrared image, but the overall brightness is too high and some information is lost. The fused low-rank part of the two models is not much different, in the scene, only the brightness of the sky part is slightly different. The fusion result of the R-model is like an overexposed image, the brightness of the white part in the visible image is further enhanced, causing the image to lose a lot of information and not be harmonized with the human vision. These prove that LatLRR helps our model capture the target area of the image at lower brightness so as not to cause the loss of detailed information in the fused image.

All in all, LatLRR is an essential part of our method. On the one hand, it can aid in the initial separation of the source image's features, and on the other, it can prevent losing details while maintaining the target information.

4.3 Effectiveness of learning-based fusion rules

To verify the effectiveness of the learning-based fusion rules, we design four variants, namely: max-max, max-mean, mean-mean, mean-max. The max-mean model preserves the latent low-rank representation, fuses the salient parts with the choose-max strategy, and fuses the background parts with the weighted-average strategy. The other three models have a similar structure. Although the maximum absolute value strategy and the weighted average strategy can preserve some features to some amount, the detailed information in the visible image is lost heavily. As is shown in Fig. 7, all models except the mean-mean model generate artifacts in the fused images, while the mean-mean model itself fails to keep detailed

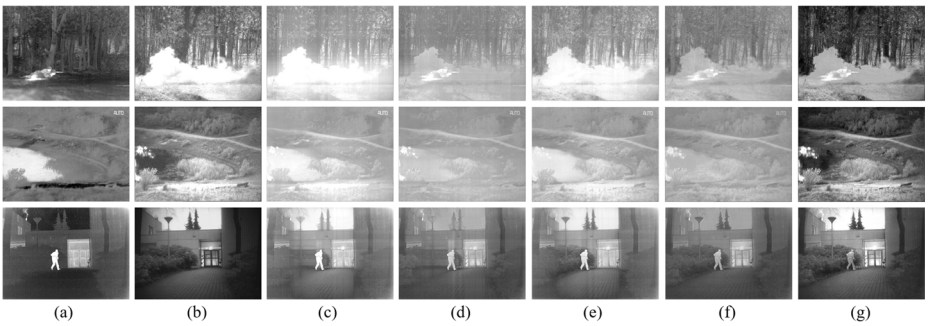

Fig. 7 Subjective comparison of fusion results of LatLRR-CNN and four variants on three representative image pairs. (a-g) correspond to the original infrared and visible images, and the fusion results of max-max, max-mean, mean-max, mean-mean and LatLRR-CNN

information well. On Nato sequences, we analyze the four models quantitatively. As indicated in Table 2, max-max, max-mean, and mean-max all have a rapid increase in only one index while other indicators remain low. The mean-mean model has all indicators lower than our method. The aforementioned four variants' qualitative and quantitative evaluations show that our learning-based rule is effective in improving fusion performance.

4.4 Validation of loss function

We design two different loss functions based on the input characteristics of the models. In this section, we will verify the correctness of the settings.

We train a model dubbed S-model by swapping the loss of the background model with the loss of the saliency model. In addition, we combine the losses of the two models and retrain another model called C-model. The results of the S-model, C-model and our method on scene kaptein_1123 are shown in Fig. 8. From top to bottom in the figure are the source images, the results of the C-model, S-model, and our method. The fused salient parts, the fused low-rank parts and the fused images are from left to right. From the results, it is evident that the results of the S-model are the worst, and there are apparent artifacts. The results of the C-model and our method are similar at first glance, but the results of our method have a crisper edge, in the lamp and tree as you zoom in on the images. However, the brightness of the target area in our method's result is still a little lower than that of the C-model. For the fused salient part and the fused low-rank part, the fused salient part of the S-model retains the target area not enough. The fused low-rank part of our method has a slightly lower brightness of the target area. This is why the brightness of the target area is slightly lower in the results of our method. We can consider this point to do a bit of targeted

Table 2 Effectiveness validation of learning-base fusion rules on Nato sequence

	max-max	max-mean	mean-max	mean-mean	LatLRR-CNN
EN	6.6282	6.3718	6.54	6.2513	7.5954
MI	4.6983	1.9444	2.2865	1.6657	2.8973
SD	29.8491	25.8744	27.1201	23.4816	57.7775
VIf	0.2736	0.2714	0.2811	0.3193	0.507

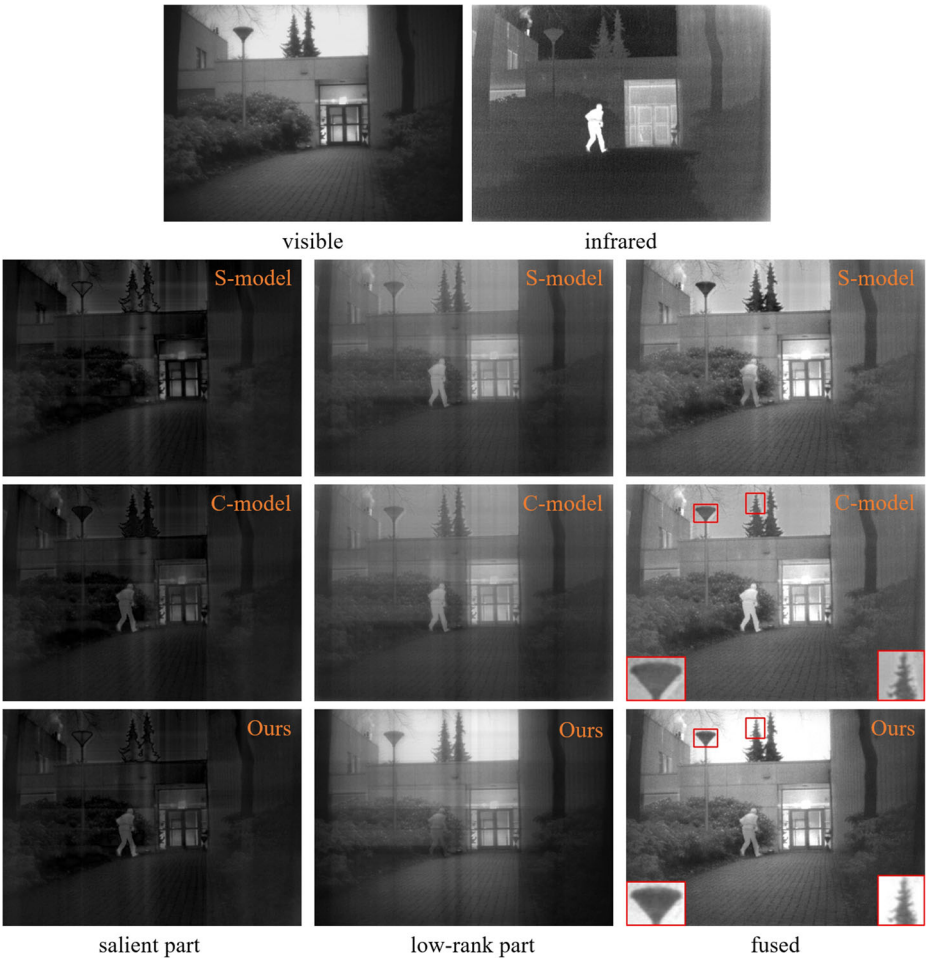

Fig. 8 The fusion results of the S-model, C-model, and our method. The first line is the source images. The second to fourth lines are the results of the S-model, C-model, and our method. The fused salient parts, the fused low-rank parts, and the fusion images are left to right

work in the future. To further illustrate the superiority of our settings, we make a quantitative comparison of four metrics on the Nato sequence, as shown in Fig. 9. Although our method is narrowly lower than C-model and S-model on VIf, our method has obvious advantages in the other three metrics. The human vision system is more sensitive to VIf and MI. Only the combined observation of VIf and MI is meaningful. In other words, our results retain as much information in the source image as possible without affecting the perception of the human eyes. All of the above shows that our settings are the best at the moment.

4.5 Comparative experiments

We selected five general scenarios from the TNO dataset for qualitative comparison, namely soldier_behind_smoke, sandpath, kaptein_1123, lake, and kaptein_1654, and the results are shown in Fig. 10.

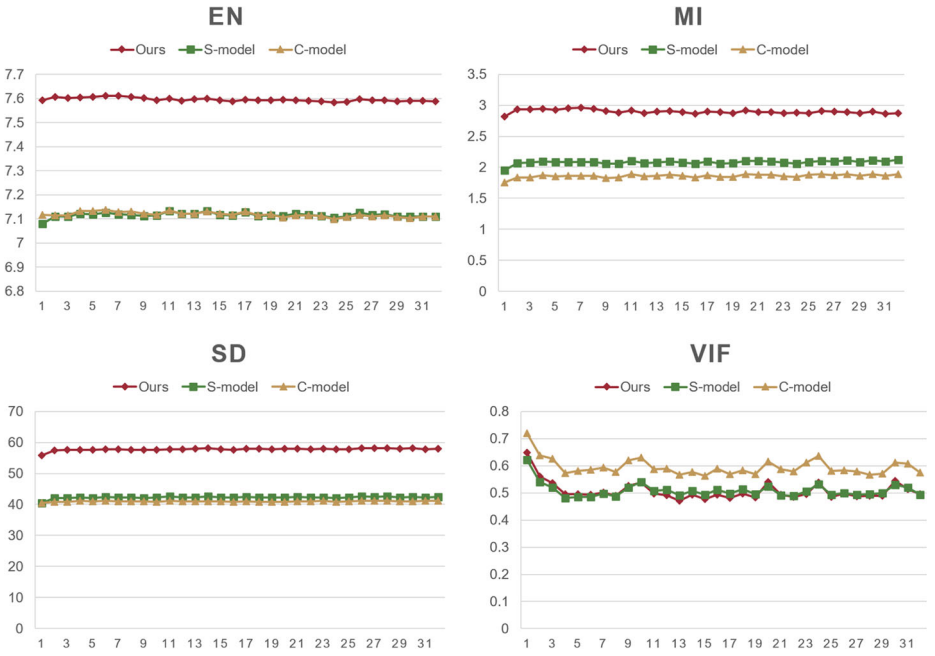

Fig. 9 Quantitative comparison of the four metrics on the Nato sequence. Participating in the comparison is the S-model, C-model, and our method

The first two rows of Fig. 10 are infrared and visible images, the last row is the results of our method, and the remaining rows are the results of the comparison method. On the whole, all methods can fuse the information of visible and infrared images to a certain extent, and it is difficult to evaluate the pros and cons of different methods based on this. However, considering the detailed information in the fusion images, it is obvious that GTF, IVIFDLF, and FusionGAN do not retain the detailed information of the visible images. Their overall images are blurry except for the target area, while the results of MTANO, TVGOF, Res2Fusion, IVIFTMTD, MDLatLRR, and our method are rich in texture information. Although MTANO can retain certain detailed information, its results are too sharp and not in line with human perception. As for TVGOF, the fusion results contain some noise, making the information obscure. Res2Fusion is capable at preserving target data, but it loses details. IVIFTMTD, on the other hand, keeps the specifics but not enough target information. Our method is slightly weaker than MDLatLRR in target prominence, but our results retain more detailed information. For example, in sandpath, the boundaries of wooden stakes are evident in our results, but it isn't easy to distinguish wooden stakes from the background in MDLatLRR. Similar phenomena can be observed in the other four scenes. In general, the results of our method do not lose the significance of the target while retaining detailed information. Our method outperforms the state-of-the-art methods in controlling target information and texture information at the same time.

Furthermore, we also perform a quantitative comparison on the Nato sequence in the TNO dataset, and the results of the four metrics are shown in Fig. 11, average metrics are

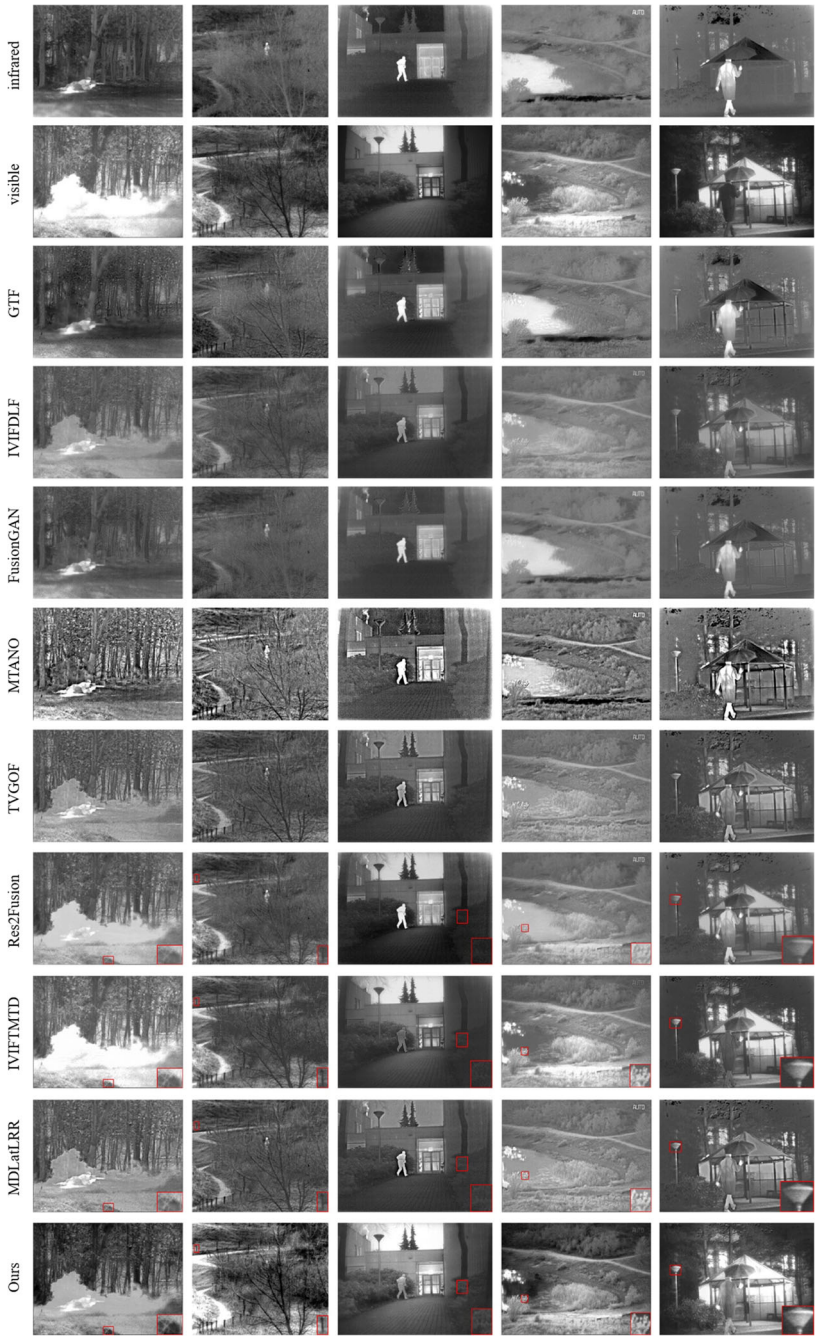

Fig. 10 The qualitative fusion results of five scenes. From left to right are soldier_behind_smoke, sandpath, kaptein_1123, lake, and kaptein_1654. From top to bottom are infrared images, visible images, the fusion results of GTF, IVIFDLF, FusionGAN, MTANO, IVIFTMTD, TVGOF, Res2Fusion, MDLatLRR, and our results. The last four lines provide a more detailed comparison. We enlarge the red box in the picture and place it in the lower right corner

Fig. 11 Quantitative comparison of the four metrics on the Nato sequence in the TNO dataset, namely EN, MI, SD, VIf. The eight state-of-the-art methods used for comparison are IVIFDLF, MDLatLRR, MTANO, GTF, IVIFTMTD, FusionGAN, TVGOF, Res2Fusion

shown in Table 3. Our method offers the highest value on EN, MI, and SD, and the average value is also the largest. Our method has absolute maximum values on EN and SD and is slightly higher than other methods on MI. For VIf, our method is just lower than MTANO, while MTANO has a lower value on the other three metrics. The most considerable EN value indicates that our results contain the most information. The most considerable MI value means that our result is the most relevant to the original image, with the most miniature artifacts. The largest SD shows that our results have the highest contrast, consistent with the qualitative comparison results. The large VIf proves that our results are the least distorted and most in line with the human eye's perception. Combining MI and VIf, the results of our method are most suitable for human vision. It is essential for tracking tasks in which compliance with the human visual system is a prerequisite. Table 4 summarizes the results of the preceding analysis, while demonstrating that our method achieves equivalent performance to state-of-the-art methods.

Because we separate the features of the source images by LatLRR, the learning task of the fusion network is relatively single. In this case, it can learn a good feature representation. At the same time, our method prevents the drawbacks of manually designing fusion rules in traditional methods, and does not result in performance degradation due to fusion rules. In conclusion, our method can simultaneously preserve thermal radiation information in infrared images and rich texture details in visible images. Compared with state-of-the-art fusion methods, our results align with human eye perception while containing more information.

Table 3 Average metrics for all methods on Nato sequences, the bolded ones indicate the maximum values

	GTF	IVIFDLF	FusionGAN	MTANO	TVGOF	Res2Fusion	IVIFTMTD	MDLatLRR	LatLRR-CNN
EN	6.611	6.2211	6.2783	7.3913	6.568	6.8631	6.6696	6.5419	**7.5954**
MI	1.9659	1.6425	2.1043	1.6481	1.5392	2.3715	2.7433	1.6032	**2.8973**
SD	25.8015	22.9834	21.6218	44.564	27.5855	35.7162	33.4077	27.8347	**57.7775**
VIf	0.2034	0.3095	0.1569	**0.8924**	0.4365	0.483	0.2718	0.4696	0.507

Table 4 Summary of our method and comparative methods

	GTF	IVIFDLF	FusionGAN	MTANO	TVGOF	Res2Fusion	IVIFTMTD	MDLatLRR	Ours
theory	gradient transfer+ total variation(TV)	guided filter+VGG	GAN	MDLatLRR+ L2	TV+CNN	Autoencoder+ Res2Net	laplace decomposi-tion+Max	LatLRR	LatLRR+CNN
qualitative	targets + basically no details	targets +little details	targets	targets + some details, images are too sharp	targets+some details	targets+details	some targets+ details	targets+some detials	some targets+rich details
quantitative	fifth EN,fifth MI,seventh SD,eighth VIf	ninth EN,seventh MI,eighth SD,sixth VIf	eighth EN,fourth MI,ninth SD,ninth VIf	second EN,sixth MI,second SD,highest VIf	sixth EN,ninth MI,sixth SD,fifth VIf	third EN,third MI,third SD,third VIf	fourth EN,second MI,fourth SD,seventh VIf	seventh EN,eighth MI,fifth SD,fourth VIf	highest EN,MI,SD,second VIf

Theory depicts the method's theoretical foundation, qualitative reflects the method's performance outcomes in the aforesaid scenarios, and quantitative provides the method's quantitative results on the Nato sequence

5 Conclusion

In this paper, we propose a novel infrared and visible image fusion method by combining latent low-rank decomposition and convolutional neural networks. It does not require artificially designed fusion rules or complex network structures to obtain better performance results. Compared with the state-of-the-art methods, the results of our method can preserve richer texture information and also better preserve thermal radiation information in infrared images. Moreover, our method is an end-to-end model, and fusion results can be automatically generated from the model without human intervention.

We conduct a large number of ablation experiments to select the optimal parameters. The quantitative comparison with the eight state-of-the-art methods on the four metrics confirms that our method has better visual effects, contains richer information and fewer artifacts.

Our method can obtain good fusion results without requiring complex fusion strategies and networks, but there is still some room for improvement. First, under the premise of ensuring fusion performance, different decomposition methods can be explored to optimize efficiency. Second, it can further optimize the background model's loss to enhance the target's prominence in the result.

Data Availability The datasets generated during and/or analysed during the current study are available in the TNO_Image_Fusion_Dataset repository, https://figshare.com/articles/TNO_Image_Fusion_Dataset/1008029.

Declarations

Conflict of Interests The authors declare that there are no conflicts of interest in this work.

References

1. Bavirisetti DP, Xiao G, Liu G (2017) Multi-sensor image fusion based on fourth order partial differential equations. In: 2017 20th international conference on information fusion (Fusion). IEEE, pp 1–9
2. Chen J, Li X, Luo L, Mei X, Ma J (2020) Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. Inf Sci 508:64–78
3. Fu Y, Wu X-J (2021) A dual-branch network for infrared and visible image fusion. In: 2020 25th international conference on pattern recognition (ICPR). IEEE, pp 10675–10680
4. Gao Z, Wang Q, Zuo C (2021) A total variation global optimization framework and its application on infrared and visible image fusion. SIViP 16(1):219–227
5. Han J, Bhanu B (2007) Fusion of color and infrared video for moving human detection. Pattern Recogn 40(6):1771–1784
6. Han Y, Cai Y, Cao Y, Xu X (2013) A new image fusion performance metric based on visual information fidelity. Information Fusion 14(2):127–135
7. Han J, Pauwels EJ, De Zeeuw P (2013) Fast saliency-aware multi-modality image fusion. Neurocomputing 111:70–80
8. Kong W, Lei Y, Zhao H (2014) Adaptive fusion method of visible light and infrared images based on non-subsampled shearlet transform and fast non-negative matrix factorization. Infrared Physics & Technology 67:161–172
9. Kong W, Zhang L, Lei Y (2014) Novel fusion method for visible light and infrared images based on nsst–sf–pcnn. Infrared Physics & Technology 65:103–112
10. Kumar P, Mittal A, Kumar P (2006) Fusion of thermal infrared and visible spectrum video for robust surveillance. In: Computer vision, graphics and image processing. Springer, pp 528–539
11. Li G, Lin Y, Qu X (2021) An infrared and visible image fusion method based on multi-scale transformation and norm optimization. Information Fusion 71:109–129
12. Li H, Wu X-J (2018) Infrared and visible image fusion using latent low-rank representation. arXiv:1804.08992
13. Li H, Wu X-J, Kittler J (2018) Infrared and visible image fusion using a deep learning framework. In: 2018 24th international conference on pattern recognition (ICPR). IEEE, pp 2705–2710

14. Li H, Wu X-J, Kittler J (2020) Mdlatlrr: a novel decomposition method for infrared and visible image fusion. IEEE Trans Image Process 29:4733–4746
15. Li S, Yang B, Hu J (2011) Performance comparison of different multi-resolution transforms for image fusion. Information Fusion 12(2):74–84
16. Li S, Yin H, Fang L (2012) Group-sparse representation with dictionary learning for medical image denoising and fusion. IEEE Trans Biomed Eng 59(12):3450–3459
17. Liu Y, Chen X, Cheng J, Peng H, Wang Z (2018) Infrared and visible image fusion with convolutional neural networks. Int J Wavelets Multiresolut Inf Process 16(03):1850018
18. Liu G, Lin Z, Yu Y et al (2010) Robust subspace segmentation by low-rank representation. In: Icml, vol 1. Citeseer, p 8
19. Liu Y, Liu S, Wang Z (2015) A general framework for image fusion based on multi-scale transform and sparse representation. Information Fusion 24:147–164
20. Liu G, Yan S (2011) Latent low-rank representation for subspace segmentation and feature extraction. In: 2011 international conference on computer vision. IEEE, pp 1615–1622
21. Ma J, Chen C, Li C, Huang J (2016) Infrared and visible image fusion via gradient transfer and total variation minimization. Information Fusion 31:100–109
22. Ma J, Liang P, Yu W, Chen C, Guo X, Wu J, Jiang J (2020) Infrared and visible image fusion via detail preserving adversarial learning. Information Fusion 54:85–98
23. Ma J, Yu W, Liang P, Li C, Jiang J (2019) Fusiongan: a generative adversarial network for infrared and visible image fusion. Information Fusion 48:11–26
24. Ma J, Zhou Z, Wang B, Zong H (2017) Infrared and visible image fusion based on visual saliency map and weighted least square optimization. Infrared Physics & Technology 82:8–17
25. Pajares G, De La Cruz JM (2004) A wavelet-based image fusion tutorial. Pattern Recogn 37(9):1855–1872
26. Qu G, Zhang D, Yan P (2002) Information measure for performance of image fusion. Electron Lett 38(7):313–315
27. Rajkumar S, Mouli PC (2014) Infrared and visible image fusion using entropy and neuro-fuzzy concepts. In: ICT and critical infrastructure: proceedings of the 48th annual convention of computer society of India-Vol I. Springer, pp 93–100
28. Rao Y-J (1997) In-fibre bragg grating sensors. Measurement science and technology 8(4):355
29. Reinhard E, Adhikhmin M, Gooch B, Shirley P (2001) Color transfer between images. IEEE Comput Graphics Appl 21(5):34–41
30. Roberts JW, Van Aardt JA, Ahmed FB (2008) Assessment of image fusion procedures using entropy, image quality, and multispectral classification. J Appl Remote Sens 2(1):023522
31. Simone G, Farina A, Morabito FC, Serpico SB, Bruzzone L (2002) Image fusion techniques for remote sensing applications. Information Fusion 3(1):3–15
32. Singh R, Vatsa M, Noore A (2008) Integrated multilevel image fusion and match score fusion of visible and infrared face images for robust face recognition. Pattern Recogn 41(3):880–893
33. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP (2004) Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process 13(4):600–612
34. Wang J, Peng J, Feng X, He G, Fan J (2014) Fusion method for infrared and visible images by using non-negative sparse representation. Infrared Physics & Technology 67:477–489
35. Wang Z, Wu Y, Wang J, Xu J, Shao W (2022) Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models. IEEE Trans Instrum Meas 71:1–12
36. Wang Z, Wu Y, Wang J, Xu J, Shao W (2022) Res2fusion: Infrared and visible image fusion based on dense res2net and double nonlocal attention models. IEEE Trans Instrum Meas 71:1–12
37. Xiang T, Yan L, Gao R (2015) A fusion algorithm for infrared and visible images based on adaptive dual-channel unit-linking pcnn in nsct domain. Infrared Physics & Technology 69:53–61
38. Xu H, Liang P, Yu W, Jiang J, Ma J (2019) Learning a generative model for fusing infrared and visible images via conditional generative adversarial network with dual discriminators. In: IJCAI, pp 3954–3960
39. Xu H, Ma J, Le Z, Jiang J, Guo X (2020) Fusiondn: a unified densely connected network for image fusion. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12484–12491
40. Yang Y, Liu J, Huang S, Wan W, Wen W, Guan J (2021) Infrared and visible image fusion via texture conditional generative adversarial network. IEEE Trans Circuits Syst Video Technol 31(12):4771–4783
41. Yang Z, Zeng S (2022) TPFUsion: Texture preserving fusion of infrared and visible images via dense networks. Entropy 24(2):294
42. Zhang Z, Blum RS (1999) A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application. Proc IEEE 87(8):1315–1326
43. Zhang L, Li H, Zhu R, Du P (2022) An infrared and visible image fusion algorithm based on ResNet-152. Multimed Tools Appl 81(7):9277–9287

44. Zhang X, Ma Y, Fan F, Zhang Y, Huang J (2017) Infrared and visible image fusion via saliency analysis and local edge-preserving multi-scale decomposition. JOSA A 34(8):1400–1410
45. Zhang H, Xu H, Xiao Y, Guo X, Ma J (2020) Rethinking the image fusion: a fast unified image fusion network based on proportional maintenance of gradient and intensity. In: Proceedings of the AAAI conference on artificial intelligence, vol 34, pp 12797–12804
46. Zhao J, Chen Y, Feng H, Xu Z, Li Q (2014) Infrared image enhancement through saliency feature analysis based on multi-scale decomposition. Infrared Physics & Technology 62:86–93
47. Zhao J, Cui G, Gong X, Zang Y, Tao S, Wang D (2017) Fusion of visible and infrared images using global entropy and gradient constrained regularization. Infrared Physics & Technology 81:201–209
48. Zhao Z, Xu S, Zhang C, Liu J, Li P, Zhang J (2020) Didfuse: Deep image decomposition for infrared and visible image fusion. arXiv:2003.09210
49. Zhao F, Zhao W, Yao L, Liu Y (2021) Self-supervised feature adaption for infrared and visible image fusion. Information Fusion 76:189–203

Affiliations

Yong Yang[1] · Chengrui Gao[1] · Zhangqiang Ming[1] · Jixiang Guo[1] · Edou Leopold[1] · Junlong Cheng[1] · Jie Zuo[1] · Min Zhu[1]

Yong Yang
yangyong9809@163.com

Chengrui Gao
cr@stu.scu.edu.cn

Zhangqiang Ming
mingzhangqiang@stu.scu.edu.cn

Jixiang Guo
guojixiang@scu.edu.cn

Edou Leopold
ledou24@yahoo.fr

Junlong Cheng
cjl951015@stu.scu.edu.cn

Jie Zuo
zuojie@scu.edu.cn

[1] College of Computer Science, Sichuan University, Chengdu, 610065, China