

# MDLatLRR: A novel decomposition method for infrared and visible image fusion

Hui Li, Xiao-Jun Wu, and Josef Kittler, *Life Member, IEEE*

**Abstract**—Image decomposition is crucial for many image processing tasks, as it allows to extract salient features from source images. A good image decomposition method could lead to a better performance, especially in image fusion tasks. We propose a multi-level image decomposition method based on latent low-rank representation(LatLRR), which is called MDLatLRR. This decomposition method is applicable to many image processing fields. In this paper, we focus on the image fusion task. We build a novel image fusion framework based on MDLatLRR which is used to decompose source images into detail parts(salient features) and base parts. A nuclear-norm based fusion strategy is used to fuse the detail parts and the base parts are fused by an averaging strategy. Compared with other state-of-the-art fusion methods, the proposed algorithm exhibits better fusion performance in both subjective and objective evaluation.

**Index Terms**—image fusion, latent low-rank representation, multi-level decomposition, infrared image, visible image.

## I. INTRODUCTION

IN the field of multi-sensor image fusion, a common task is to combine infrared and visible images. It arises in many applications, such as surveillance [41], object detection and target recognition [16] [15] [11]. The key assumption is that different modalities convey complementary information. The main purpose of image fusion is to generate a single image which combines the complementary sources of information from multiple images of the same scene [22]. The key problem is to extract the salient image content from both modalities and then fuse it to produce enhanced output. Many fusion methods have been proposed for this purpose. The methods most commonly used for image fusion are multi-scale transforms [4] [44] and representation learning based methods [19] [3].

In the multi-scale transform domain, there are many fusion tools, such as the discrete wavelet transform(DWT) [2], contourlet transform [53], shift-invariant shearlet transform [45], quaternion wavelet transform [35] and nonsubsampling shearlet transform [29] etc. These decomposition methods project source images to the frequency domain, which increases the computational complexity. Hence, researchers try to find other

methods to process source images without a transform, such as representation learning based methods.

In the representation learning domain, the most common methods are based on sparse representation(SR) [46] and dictionary learning [5] [6]. For instance, Zong et al. [57] proposed a novel medical image fusion method based on SR and Histogram of Oriented Gradients(HOG). There are also many algorithms based on combining SR and other tools including pulse coupled neural network(PCNN) [28], low-rank representation(LRR) [18] and shearlet transform [54]. Other options include the joint sparse representation [23] and cosparse representation [7].

Although SR-based fusion methods achieve good fusion performance, these methods are complex and dictionary learning is a very time-consuming operation, especially online learning. These problems have motivated a growing interest in deep learning to replace the dictionary learning phase in SR.

In deep learning-based fusion methods, deep features of the source images are used to generate the fused image. In [27], Liu et al. proposed a fusion method based on convolutional sparse representation(CSR), in which multi-scale and multi-layer features are used to construct the fused output. A follow up work was presented in Liu et al. [26]. Image patches which contain different blur versions are used to train a network to produce a decision map. The fused image is obtained by using the decision map and source images. In ICCV 2017, Prabhakar et.al. [36] proposed a novel CNN-based fusion framework for a multi-exposure image fusion task.

However, these deep learning-based methods still have drawbacks. The network is difficult to train when the training data is insufficient, especially in infrared and visible image fusion tasks. CNN-based methods tend to work for specific image fusion tasks. To overcome this drawback, Li et.al. [21] [20] proposed two novel fusion frameworks based on a pretrained network(VGG-19 [42] and ResNet-50 [9]). The pretrained network is used as a feature extraction tool. They compared this approach with a strategy involving training a novel network architecture (DenseFuse) [19]. Ma et.al. [32] [31], on the other hand, applied Generative Adversarial Networks to the image fusion task which also achieves impressive fusion performance. A simple yet effective end-to-end fusion framework was proposed by Zhang et. al [56]. Interestingly, in these deep learning based methods, a very little attention is paid to image decomposition.

The latent low rank representation (LatLRR) [25] is commonly used in clustering analysis tasks. The authors [25] also mentioned that LatLRR can be used to extract salient

This work was supported by the National Key Research and Development Program of China (Grant No. 2017YFC1601800), the National Natural Science Foundation of China (61672265, U1836218), the 111 Project of Ministry of Education of China (B12018), and the UK EPSRC (EP/N007743/1, MURI/EPSC/DSTL, EP/R018456/1).

Hui Li and Xiao-Jun Wu(*Corresponding author*) are with the School of Internet of Things Engineering, Jiangnan University, Wuxi 214122, China. (hui\_li\_jnu@163.com, xiaojun\_wu\_jnu@163.com)

Josef Kittler is with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, GU2 7XH, UK. (e-mail: j.kittler@surrey.ac.uk)

This paper has supplementary downloadable material available at <http://ieeexplore.ieee.org>, provided by the authors.

features from the input data. In this paper we follow this suggestion and propose a novel multi-level decomposition method(MDLatLRR) as a basis of a fusion framework for infrared and visible image fusion.

In our fusion framework, MDLatLRR is used to decompose source images, so as to extract salient detail parts and the base parts. The detail parts are reconstructed by an adaptive fusion strategy and by a reshape operator. The base parts are fused by averaging strategy. Finally, the fused image is constructed by combining the detail parts and the base part. Compared with the state-of-the-art fusion methods, our fusion framework achieves better fusion performance in both subjective and objective evaluations.

The main contributions are summarized as follows,

(1) A LatLRR based multi-level decomposition method (MDLatLRR) is proposed for the image fusion task. In contrast to another LRR-based fusion method [18] which calculates the low-rank coefficients for each image, in our method, the LatLRR problem is only solved once in the training phase to learn a projection matrix  $L$ . As the size of  $L$  is only related to the image patch size, thus the fusion method can process the source images with an arbitrary size.

(2) With the multi-level decomposition operation, the proposed fusion method captures more detail information from the source images. As the base part contains preponderance of contour features, it makes the weight-average strategy more effective.

(3) Contrasting it with the  $l_1$ -norm and  $l_2$ -norm, nuclear-norm(the fusion strategy for detail parts) calculates the sum of singular values, which preserves 2D information from the source images. It will generate a larger weight when an input patch contains more texture and structural information. This behaviour is more preferable in image processing tasks.

This paper is structured as follows. In Section II, we introduce preliminaries. In Section III, we present our multi-level decomposition method(MDLatLRR) in detail. In Section IV, the proposed fusion framework based on MDLatLRR is introduced. The experimental results are presented in Section V. Finally, the conclusions are drawn in Section VI.

## II. PRELIMINARIES

### A. Latent Low-Rank Representation

In 2010, Liu et al. [24] proposed LRR theory in which the input data matrix is chosen as the dictionary. This method does not achieve good performance when the input data is either insufficient or corrupted. So in 2011, the author [25] proposed the LatLRR theory, according to which the low-rank structure and salient structure are extracted from the raw data. More specifically, the LatLRR problem is formulated as the following optimization problem,

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

where  $\lambda > 0$  is a balance coefficient,  $\|\cdot\|_*$  denotes the nuclear norm which is the sum of the singular values of the matrix and  $\|\cdot\|_1$  is  $l_1$ -norm.  $X$  denotes the observed data matrix,  $Z$  is a matrix of low-rank coefficients,  $L$  is a projection matrix

named a salient coefficients matrix, and  $E$  is a sparse noisy matrix. Eq 1 is solved by the inexact Augmented Lagrangian Multiplier (ALM) [25] algorithm. Then the salient component  $LX$  is obtained by Eq 1.

### B. Projection Matrix $L$ in LatLRR

Assume that a source image is divided into  $M$  image patches, as shown in Figure 1. The size of an image patch is  $N = n \times n$ .  $X$  indicates the observed matrix and each column represents an image patch. The size of  $Z$  is related to the number of image patches, which depends on the size of the source image. It is time consuming to calculate the low-rank coefficients for every image in the test phase, if  $Z$  is applied to the fusion framework. In contrast, the size of the project

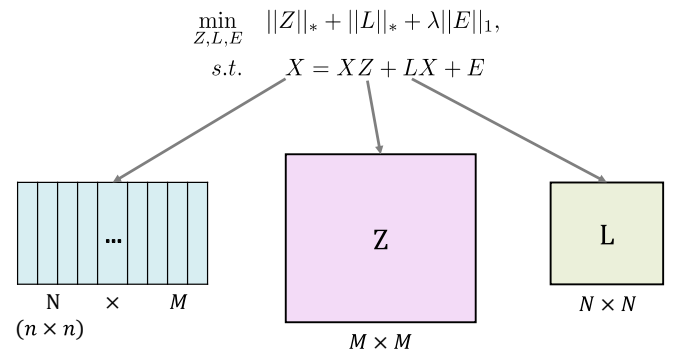


Fig. 1. Latent low-rank representation. The observed data matrix  $X$ , low-rank coefficients  $Z$  and projection matrix  $L$ .

matrix  $L$  is related to the image patch size. In this case, once the projection matrix is learned by LatLRR, it can be used to process other images which are of arbitrary size. In terms of the visual content [25],  $LX$  contains more detail information and salient features.

## III. MULTI-LEVEL DECOMPOSITION BASED ON LATLRR(MDLATLRR)

In this section, our MDLatLRR method is presented. Firstly, we introduce the LatLRR-based decomposition method(DLatLRR). Then, a multi-level version of DLatLRR will be introduced.

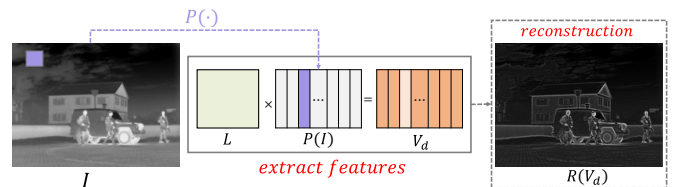


Fig. 2. The process of image decomposition based on LatLRR(DLatLRR) method.

### A. DLatLRR

Once the projection matrix  $L$  is learned by LatLRR, it can be used to extract the detail part and the base part of the input

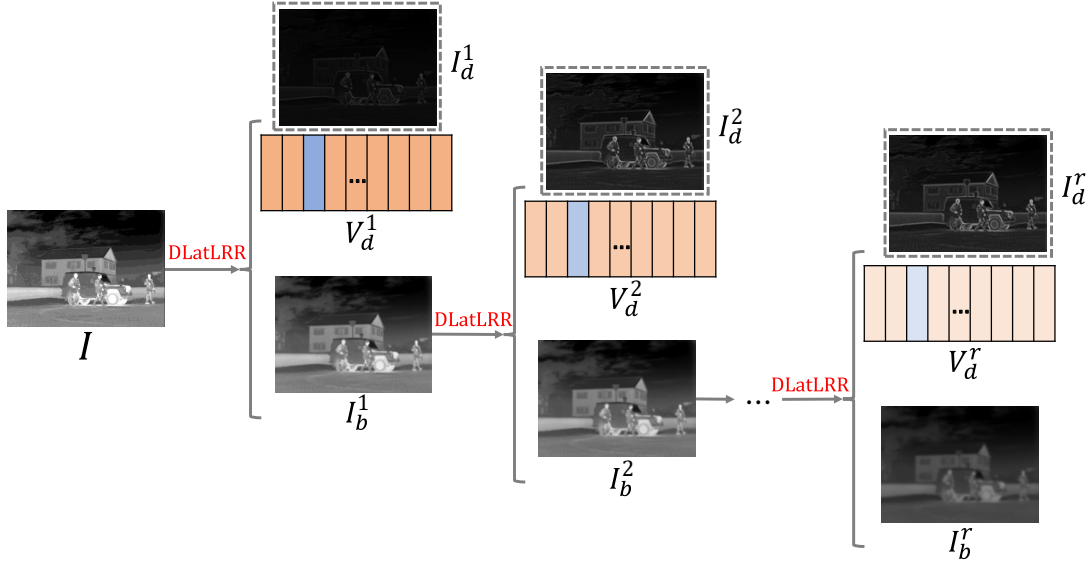


Fig. 3. MDLatLRR, the deep version of DLatLRR.  $r$  denotes the number of decomposition levels.

image by DLatLRR. The process of DLatLRR is illustrated in Figure 2.

In Figure 2, the input image( $I$ ) is divided into many image patches by the sliding window technique with an overlap. A sliding window of size  $n \times n$  accents pixel data in the image covered by the window. Stride  $s$  means that the window is moved by  $s$  pixels in the  $x$  and  $y$  direction at a time.

These image patches are reshuffled into a source matrix. Each column of this matrix corresponds to an image patch. The detail part and the base part are determined by Eq 2.

$$\begin{aligned} V_d &= L \times P(I), I_b = I - I_d \\ s.t., I_d &= R(V_d) \end{aligned} \quad (2)$$

where  $L$  denotes the projection matrix learned by LatLRR.  $V_d$  is the result of decomposition of the input image( $I$ ),  $P(\cdot)$  denotes a two-stage operator which consists of the sliding window technique and reshuffling.  $R(\cdot)$  signifies the operator which reconstructs the detail image from the detail part. When restoring  $V_d$ , the averaging strategy is used to process the overlapping pixel. Firstly, we count the number of the overlapping pixels in each position. Then, for each position in the restored image, we calculate the average pixel value.

As shown in Eq 2, the detail part is generated by  $L$ ,  $P(\cdot)$  and the input image  $I$ . The base part( $I_b$ ) is obtained by subtracting the detail image( $I_d$ ) from the input image.

#### B. MDLatLRR

Using the DLatLRR decomposition method, a multi-level decomposition strategy is applied to each base part. As a result, a multi-level version of DLatLRR is obtained, which is named MDLatLRR. The framework of MDLatLRR is presented in Figure 3.

In Figure 3,  $I$  is the input image,  $V_d^i$  and  $I_b^i$  denote the detail part matrix and the base part of the input image decomposed by DLatLRR at level  $i$ .  $r$  denotes the highest decomposition level and  $i = [1, 2, \dots, r]$ .

As shown in Figure 3, each previous base part is decomposed by DLatLRR. Then, for  $r$  level decomposition, we will get  $r$  detail part matrices  $V_d^{1:r}$  and one base part  $I_b^r$ .  $I$  can be reconstructed by adding  $I_b^r$  and  $r$  detail images. The detail parts  $V_d^{1:r}$  and base part  $I_b^r$  are obtained by Eq 3,

$$\begin{aligned} V_d^i &= L \times P(I_b^{i-1}), \quad I_b^i = I_b^{i-1} - I_d^i \\ I_d^i &= R(V_d^i), \quad I_b^0 = I, i = [1, 2, \dots, r]. \end{aligned} \quad (3)$$

#### IV. MDLATLRR-BASED FUSION METHOD

In this section, the MDLatLRR-based fusion method is presented in detail.

##### A. Learning the Projection Matrix $L$

The training images are available at [17]<sup>1</sup>. These images are shown in Figure 4, where the first row contains infrared images and the second row presents visible images. All the images are divided into image patches by the sliding window technique.



Fig. 4. Five pairs of infrared and visible images, which are used to learn  $L$  by LatLRR.

We obtain 10090 to 40654 image patches, depending on the image patch size, which is  $n \times n$  and  $n \in \{8, 16\}$ . 2000 image patches are randomly chosen to generate an input matrix  $X$

<sup>1</sup>The path is “/learning\_projection\_matrix/training\_data/”. For more experimental results on extensive training dataset, please refer to our supplementary materials.

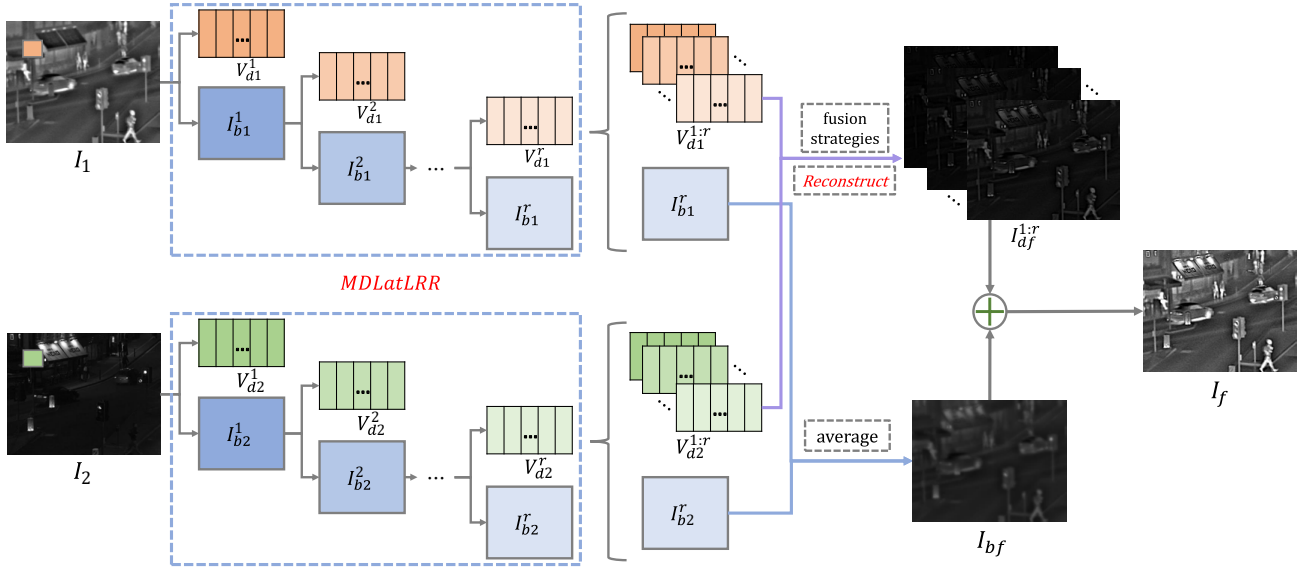


Fig. 5. The framework of the proposed fusion method. Source images are decomposed into detail parts( $V_{dk}^{1:r}$ ) and base part( $I_{bk}^r$ ). Then, with adaptive fusion strategies, the fused image( $I_f$ ) is reconstructed from fused detail parts( $I_{df}^{1:r}$ ) and base part( $I_{bf}$ ).

in which each column indicates all pixels of one image patch. The size of  $X$  is  $N \times M$ , where  $N = n^2$  and  $M = 2000$ .

As we discussed in Section II, the projection matrix  $L$  is learned by LatLRR and ALM. In LatLRR,  $\lambda$  is set to 0.4 [25]. With different image patch sizes, two types of  $L$  ( $64 \times 64$ ,  $256 \times 256$ ) are obtained.

In the fusion framework,  $L$  is used to extract salient features from the test images with the same patch size ( $8 \times 8$ ,  $16 \times 16$ ). The details of how to train the projection matrix  $L$  will be introduced in our ablation study (Section V-B).

### B. Fusion Method

Given two input images  $I_1$  and  $I_2$  (infrared and visible images), we use  $I_k$  ( $k \in \{1, 2, \dots, K\}$ ,  $K = 2$ ) if it is unimportant what type the input image is. The fusion algorithm is the same when the number of input images  $K > 2$ . The framework of our fusion algorithm is presented in Figure 5.

The input image( $I_k$ ) is first decomposed by MDLatLRR to  $r$  decomposition levels. Then,  $r$  pairs of the detail part matrices( $V_{dk}^{1:r}$ ) and one pair of base part( $I_{bk}^r$ ) are determined.

For each pair of detail parts, an adaptive fusion strategy is used to fuse these parts column by column. Consequently,  $r$  fused detail images( $I_{df}^{1:r}$ ) are obtained. The fused detail images are calculated by Eq 4,

$$I_{df}^i = R(FS(V_{d1}^i, V_{d2}^i)), i = 1, 2, \dots, r \quad (4)$$

where  $R(\cdot)$  denotes the reconstruction operator as discussed in Section III-A.  $FS(\cdot)$  is the fusion strategy which will be introduced in the next subsection.

As the base part contains contour and brightness information, in our fusion method, a weighted average strategy is utilized to obtain the fused base part.

Once the fused detail images and base part are obtained, the fused image is reconstructed.

In the next subsections, the details of the deep decomposition method, the fusion strategies and the reconstruction process will be presented.

### C. Fusion Strategies and Reconstruction

Once the input image is decomposed by MDLatLRR, we adopt adaptive strategies to fuse these parts.

1) *Fusion of base parts*: The base parts of input images contain more common features, redundant and brightness information. So, in our fusion method, we use weighted averaging to obtain the fused base part. It is calculated by Eq 5,

$$I_{bf}(x, y) = w_{b1}I_{b1}^r(x, y) + w_{b2}I_{b2}^r(x, y) \quad (5)$$

where  $(x, y)$  denotes the corresponding position in the base parts( $I_{b1}^r, I_{b2}^r$ ) and the fused base part( $I_{bf}$ ).  $w_{b1}$  and  $w_{b2}$  denote the weights of the base parts.

2) *Fusion of detail parts*: In contrast to the base parts, the detail parts preserve more structural information, as well as salient features. So, the fusion strategy( $FS(\cdot)$ ) for detail parts requires a more careful design.

In our method, the nuclear norm is used to calculate the weights for each pair of corresponding image patches which are obtained by  $P(\cdot)$  from the input images. An overview of this fusion strategy is shown in Figure 6.

In Figure 6,  $V_{dk}^{i,j}$  and  $V_{df}^{i,j}$  are the vectors corresponding to the  $j$ -th column in the detail part matrix  $V_{dk}^i$  and the fused detail part matrix  $V_{df}^i$ , respectively.  $i$  is the decomposition level,  $k \in \{1, 2\}$ .

Firstly, for each corresponding column, the weights  $w_{dk}^{i,j}$  are calculated by Eq 6,

$$w_{dk}^{i,j} = \frac{\hat{w}_{dk}^{i,j}}{\sum_{q=1}^K \hat{w}_{dq}^{i,j}} \quad (6)$$

$$\hat{w}_{dk}^{i,j} = \|re(V_{dk}^{i,j})\|_*$$



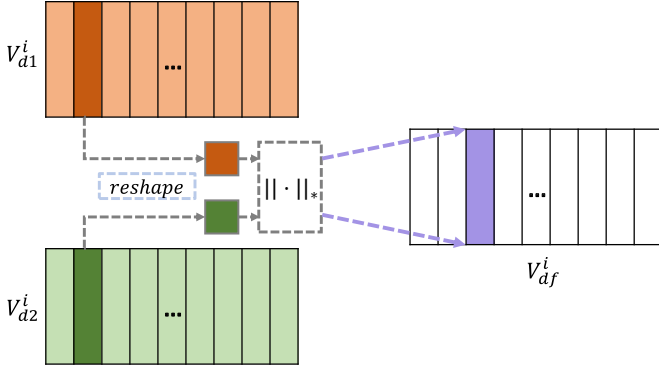


Fig. 6. Fusion strategy based on nuclear norm.  $V_{d1}^i$  and  $V_{d2}^i$  denote detail part matrices at the  $i$ -th level. “reshape” means the reconstruction operator( $re(\cdot)$ ) which is used to reconstruct an image patch from the corresponding vectors.  $\|\cdot\|_*$  indicates nuclear-norm.  $V_{df}^i$  denotes the fused detail part.

where  $re(\cdot)$  denotes the operator which is used to reconstruct the image patch from vector  $V_{dk}^{i,j}$ , and  $\|\cdot\|_*$  indicates the nuclear-norm which calculates the sum of singular values of the matrix.  $K$  is the number of input images.

Then, the fused detail part vector  $V_{df}^{i,j}$  is obtained by weighting the detail part vectors  $V_{dk}^{i,j}$ , as shown in Eq 7,

$$V_{df}^{i,j} = \sum_{k=1}^K w_{dk}^{i,j} \times V_{dk}^{i,j} \quad (7)$$

This fusion strategy is applied to all pairs of detail part matrices  $V_{dk}^{1:r}$ . Each fused detail image( $I_{df}^i$ ) is calculated by Eq 8,

$$I_{df}^i = R(V_{df}^i), \quad i = [1, 2, \dots, r] \quad (8)$$

where  $R(\cdot)$  indicates the reconstruction operator.

3) *Reconstruction*: Once we have the fused detail images and the fused base part, the fused image is generated by Eq 9,

$$I_f(x, y) = I_{bf}(x, y) + \sum_{i=1}^r I_{df}^i(x, y) \quad (9)$$

## V. EXPERIMENTAL RESULTS AND ANALYSIS

The aim of our experiments is to provide a supporting evidence for the key ideas underpinning the proposed method, namely the use of the nuclear-norm and multi-level decomposition. We evaluate the fusion performance of our method and compare it to other existing fusion methods.



Fig. 7. Five pairs of source images. The top row contains infrared images, and the second row contains visible images.

### A. Experimental Setting

The test data is collected from [43], it contains 21 pairs infrared and visible images. A sample of these infrared and visible images is shown in Figure 7. The code of MDLatLRR and the test data are available at [17]<sup>2</sup>.

Firstly, the sliding window technique is used to divide the source images into image patches, using the stride set to 1. The window size depends on the size of  $L$  ( $8 \times 8, 16 \times 16$ ).

In our MDLatLRR-based fusion method,  $w_{b1} = w_{b2} = 0.5$  for the fusion strategy of the base part, since the base parts contain more common features and redundant information.

For comparison, thirteen recent and classical fusion methods are chosen to conduct the same experiment. They include: cross bilateral filter method(CBF) [14], discrete cosine harmonic wavelet transform method(DCHWT) [13], joint sparse representation-based method(JSR) [55], the JSR model with saliency detection method(JSRSD) [23], the gradient transfer fusion method(GTF) [30], weighted least square optimization-based method(WLS) [33], convolutional sparse representation(ConvSR) [27], VGG-19 and multi-layer fusion strategy-based method(VggML) [21], ResNet50 and zero-phase component analysis fusion framework (ResNet-ZCA<sup>3</sup>) [20], a CNN-based method(DeepFuse) [36], GAN-based fusion algorithm(FusionGAN) [32], dense block based fusion method(DenseFuse<sup>4</sup>) [19] and an end-to-end fusion network (IFCNN) [56].

For the purpose of a quantitative comparison between our method and the other existing fusion methods, twelve quality metrics are utilized. These are: entropy( $En$ ) [38]; mutual information( $MI$ ) [10] indicates how much features are preserved in fused images; the standard deviation ( $SD$ ) [37];  $Q_{abf}$  [52] reflects the quality of visual information obtained from the fusion of the input images;  $FMI_p$ ,  $FMI_w$  and  $FMI_{det}$  [8] calculate fast mutual information ( $FMI$ ), discrete cosine and wavelet features, respectively;  $SCD$  [1] calculates image quality metric value based on the sum of the correlations of differences; a modified structural similarity  $SSIM_a$  [21]; MS-SSIM [34] calculates a modified structural similarity which only focuses on structural information; visual information fidelity( $VIF$ ) [40]; and edge preservation index ( $EPI$ ) [39]. The fusion result is better with the increasing magnitude of these metric values.

All the experiments are implemented in MATLAB R2017b on 2.8 GHz Intel(R) Core(TM) i5-8400 CPU with 16 GB RAM.

### B. Ablation Study

1) *Projection Matrix  $L$* : In this section, we introduce how to choose the training image patches when learning the projection matrix  $L$ .

As discussed in Section IV-A, the sliding window technique is used to divide training images into image patches. If the

<sup>2</sup>The test data path is “/IV\_images/”. For more results on additional test data (new test data containing 20 pairs of images), please refer to supplementary materials

<sup>3</sup>The deep features were extracted by ‘Conv5’.

<sup>4</sup>The fusion strategy is ‘addition’ and  $\lambda = 1e2$ .

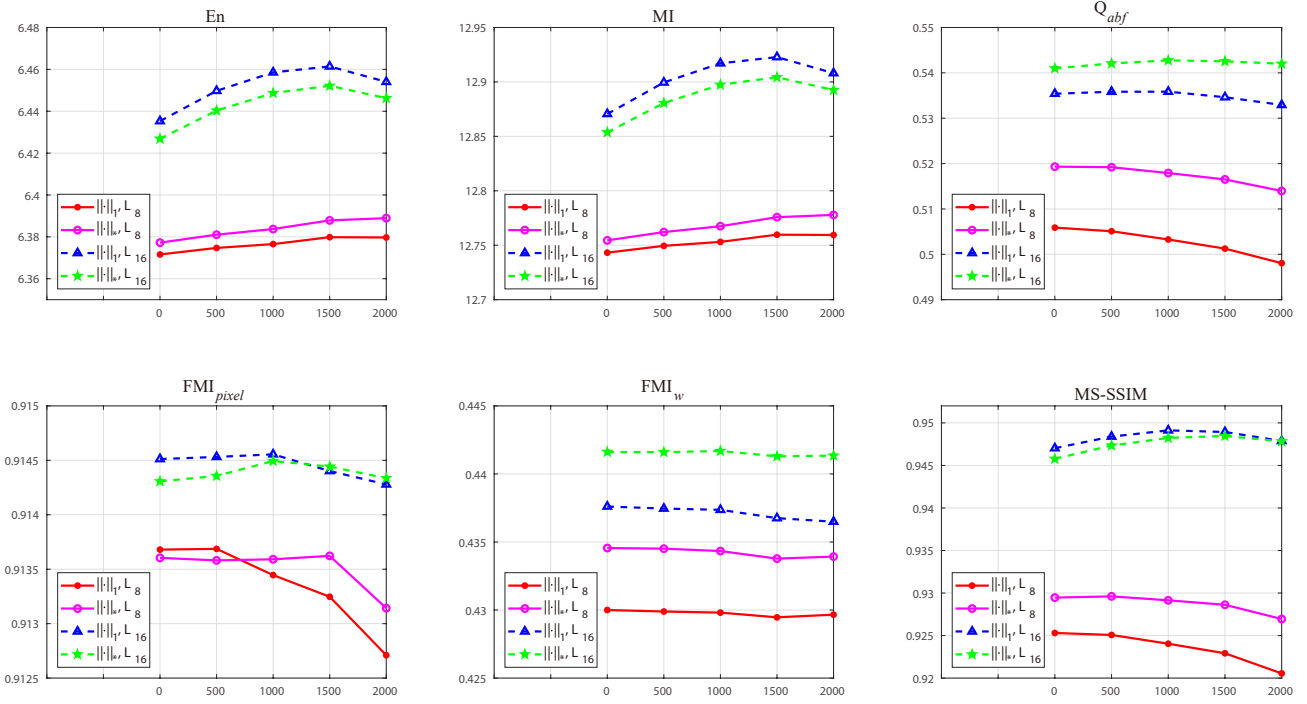


Fig. 8. The fusion performance of different projection matrices. The number of training image patches is 2000. The horizontal axis indicates the number of smooth patches in all training patches. The vertical axis denotes the average values of the six quality metrics computed on all test images.  $\|\cdot\|_1$  and  $\|\cdot\|_2$  indicate  $l_1$ -norm and nuclear norm, respectively.  $L_8$  and  $L_{16}$  denote projection matrices trained by different image patch size ( $8 \times 8$  and  $16 \times 16$ ).

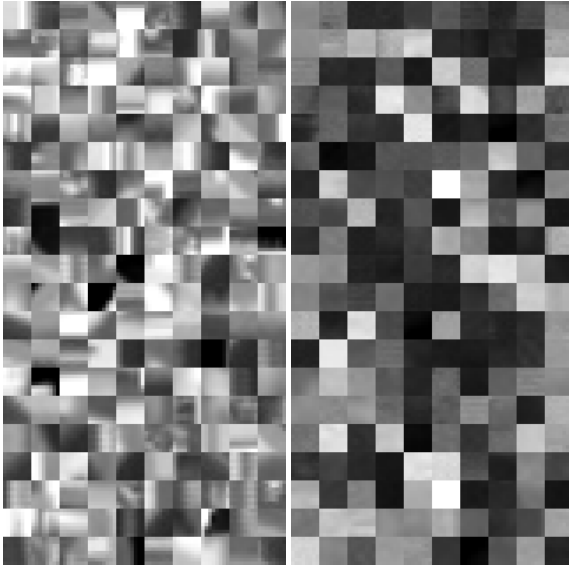


Fig. 9. Training image patches selection ( $8 \times 8$ ). Left part is 200 *detail* image patches and right part is 200 *smooth* patches, which are all randomly chosen from categorized image patches.

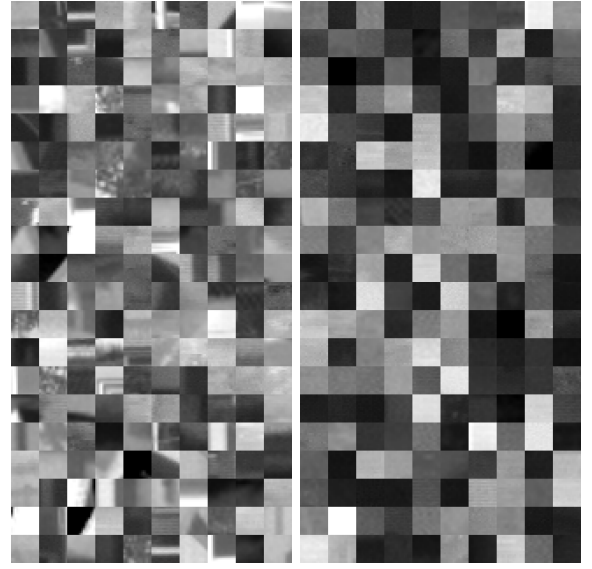


Fig. 10. Training image patches selection ( $16 \times 16$ ). Left part is 200 *detail* image patches and right part is 200 *smooth* patches, which are all randomly chosen from categorized image patches.

window size is  $8 \times 8$ , 40654 image patches are obtained. For  $16 \times 16$ , we will get 10090 image patches. These image patches are classified into two categories (*detail*, *smooth*) by standard deviation ( $SD$ ) [37]. The  $SD$  is defined as,

$$SD(P) = \sqrt{\sum_{i=1}^H \sum_{j=1}^W (P(i, j) - \mu)^2} \quad (10)$$

where  $P(i, j)$  is the pixel value of image patch,  $H$  and  $W$  denote the patch size ( $H \times W$ ).  $\mu$  is the mean value of the input image patch. The larger  $SD$  means image patch contains more detail information. In this experiment, the classification strategy ( $C$ ) is defined as,

$$C(P) = \begin{cases} detail & SD(P) > Th \\ smooth & otherwise \end{cases} \quad (11)$$

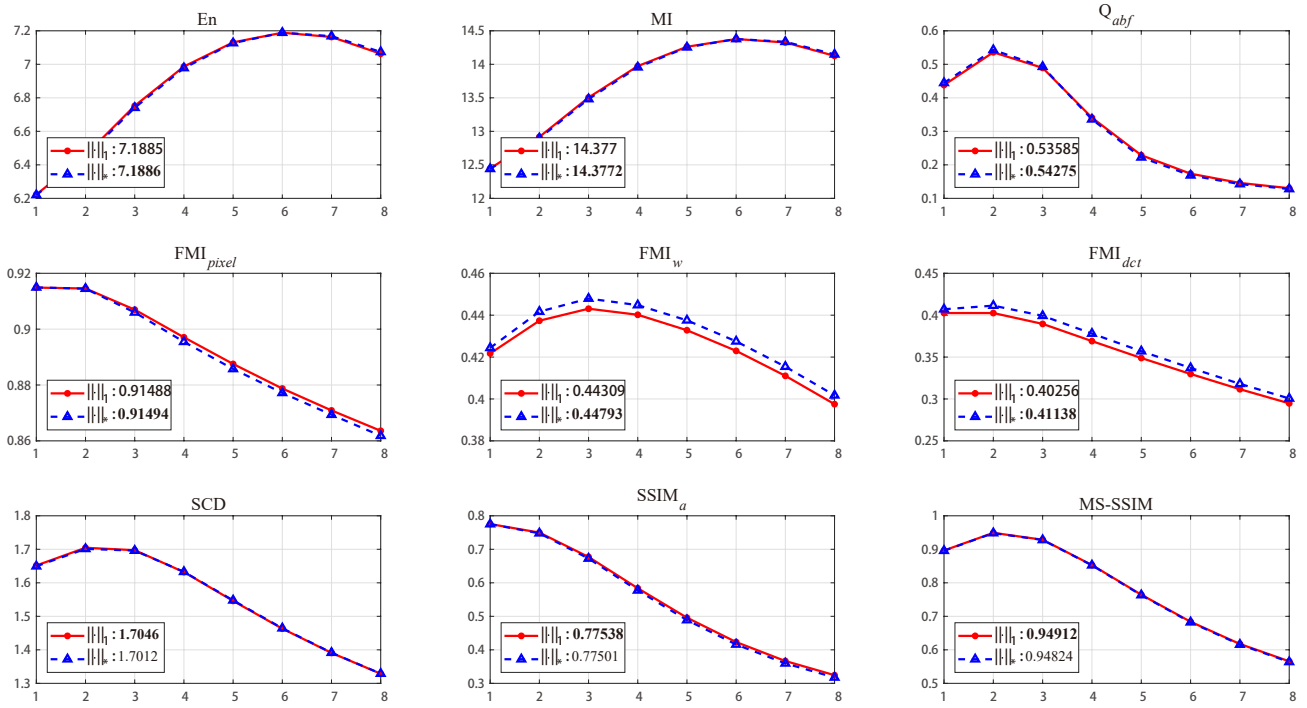


Fig. 11. Different decomposition levels and different norms. The average values of nine metrics ( $En$ ,  $MI$ ,  $Q_{abf}$ ,  $FMI_{pixel}$ ,  $FMI_w$ ,  $FMI_{dct}$ ,  $SCD$ ,  $SSIM_a$  and MS-SSIM) for all fused images obtained by MDLatLRR using multi-level (8 levels, horizontal axis) and different norms.  $\|\cdot\|_1$  denotes  $l_1$ -norm, and  $\|\cdot\|_*$  indicates nuclear-norm.

where  $P$  indicates the input image patch,  $Th$  is the threshold which determines the patch belongs to *detail* or *smooth*.  $Th$  is set to 0.5 in this experiment.

Following the classification strategy (C), for  $8 \times 8$  image patches, 2316 *detail* patches and 38338 *smooth* patches are obtained. For  $16 \times 16$ , we have 2646 *detail* patches and 7444 *smooth* patches, respectively. The examples of these patches are shown in Figure 9 and Figure 10, respectively.

We randomly choose 2000 image patches to learn the projection matrix  $L$ . Thus, the number of *detail* and *smooth* in 2000 training patches becomes an important issue.

We set the number of *smooth* to 0, 500, 1000, 1500, and 2000 to analyze the influence of the ratio *detail* to *smooth*. Then, five group ratios are obtained which are [2000, 0], [1500, 500], [1000, 1000], [500, 1500], [0, 2000]. The first value is the *detail* number and the second is the *smooth* number. The projection matrices are learned by the above training patches and the decomposition level of MDLatLRR is temporarily set to 2. For each scale ( $8 \times 8$ ,  $16 \times 16$ ), five matrices are learned by LatLRR. Six quality metrics ( $En$ ,  $MI$ ,  $Q_{abf}$ ,  $FMI_{pixel}$ ,  $FMI_w$  and MS-SSIM) are chosen to evaluate the performance for different projection matrices. The average values are shown in Figure 8.

From Figure 8, when the size of image patch is  $16 \times 16$ , the proposed method obtains better fusion performance no matter what the ratios are. The reason is that a larger image patch contains more useful information which is beneficial for generating fused image. However, if the image patch size is  $32 \times 32$ , the size of the projection matrix will become  $1024 \times 1024$ , and the MDLatLRR fusion method will become

time-consuming.

Focusing on  $L_{16}$  (with  $l_1$ -norm and nuclear norm), the proposed method obtains four best values ( $Q_{abf}$ ,  $FMI_{pixel}$ ,  $FMI_w$  and MS-SSIM) when the ratio is [1000, 1000]. For  $En$  and  $MI$ , the projection matrix  $L$  still achieves comparable performance.

The reason why 2000 *detail* patches can not learn a better projection matrix is that both positive and negative samples should be available for learning. In our learning process, *detail* patches are positive samples and *smooth* patches are negative samples. Hence, in our experiments, the projection matrix  $L$  is learned using 1000 *detail* patches and 1000 *smooth* patches with the size of image patch being  $16 \times 16$ .  $l_1$ -norm and nuclear norm are used to fuse detail parts.

**2) Decomposition Levels and Nuclear-norm:** In this section, multi-level decompositions,  $l_1$ -norm and nuclear-norm are utilized to make comparisons. When  $l_1$ -norm is used in fusion, Eq 6 is changed to Eq 12,

$$w_{dk}^{i,j} = \frac{w_{dk}^{\hat{i},j}}{\sum_{q=1}^K w_{dq}^{\hat{i},j}}, \quad w_{dk}^{\hat{i},j} = \|V_{dk}^{i,j}\|_1 \quad (12)$$

where  $\|\cdot\|_1$  indicates the  $l_1$ -norm.

The average values of the nine quality metrics are shown in Figure 11. The annotation contains the maximum values in each different situation and the best values are marked in **bold**.

In Figure 11, the levels between 1 to 4 obtain almost best values, but the increasing the number of decomposition levels (more than four levels) would cause fusion performance

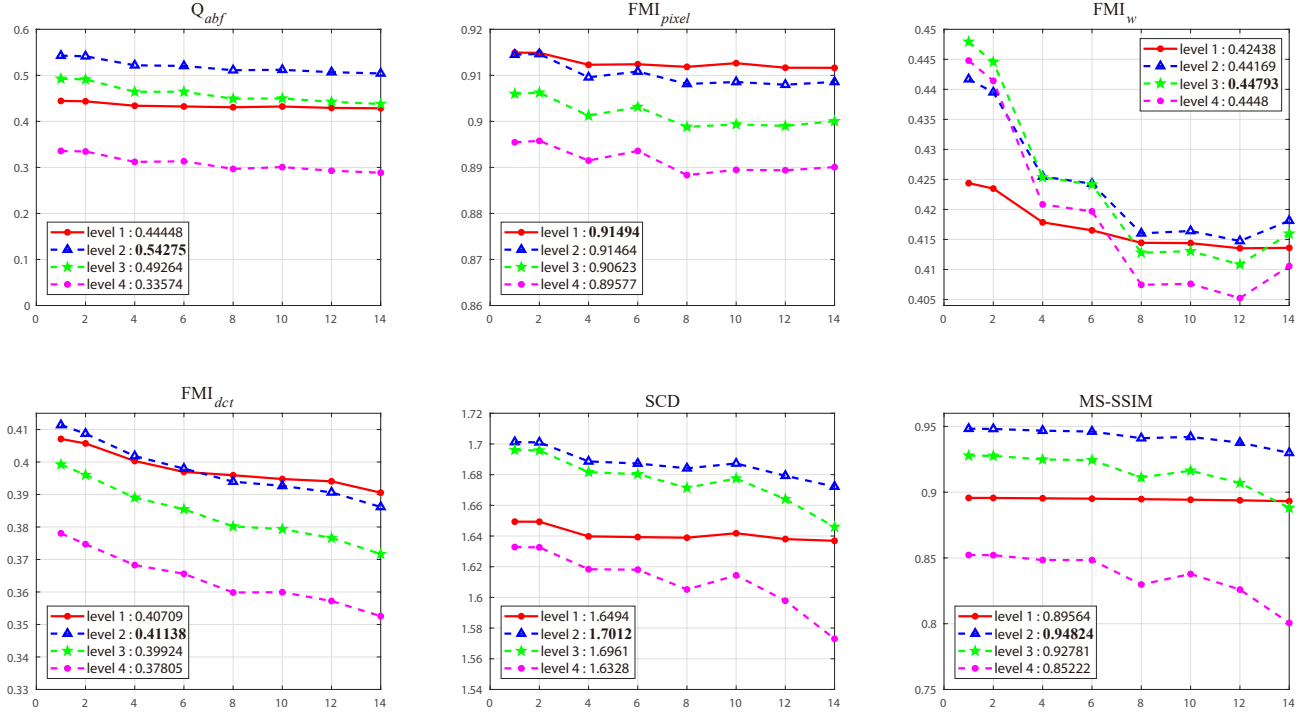


Fig. 12. The sliding window technique with different strides ( $s = 1, 2, 4, 6, 8, 10, 12, 14$ ). We present the average values of six metrics ( $Q_{abf}$ ,  $FMI_{pixel}$ ,  $FMI_w$ ,  $FMI_{dct}$ ,  $SCD$  and  $MS-SSIM$ ) for all fused images obtained by MDLatLRR with different stride (8 strides, horizontal axis) and different levels (1 to 4).

degradation. With the growing number of decomposition levels, the detail parts will contain more base information (e.g. luminance, contour). However, the fusion strategy for detail parts is not equipped to fuse them. So, in image fusion tasks, the decomposition levels should be chosen carefully. From Figure 11, when the number of decomposition levels is set to 1 to 4, our proposed fusion method achieves seven best values (except  $En$  and  $MI$ ). Even for  $En$  and  $MI$ , the results are still comparable when the level is set to 3 and 4. The reason is that the source images are successfully divided into the detail parts and base part by our MDLatLRR with level 1 to 4, and this is beneficial for image fusion.

On the other hand, the nuclear norm achieves six best values in all nine quality metrics. For  $SCD$ ,  $SSIM_a$  and  $MS-SSIM$ , the nuclear norm still reports comparable values. These results indicate the fusion strategy with the nuclear-norm achieves better performance than other option ( $l_1$ -norm). Recall, the nuclear-norm, which calculates the sum of singular values of a matrix, represents the structural information of an image patch.

In our nuclear-norm based fusion strategy, when the detail patch contains more texture and structural information, which means larger values of nuclear-norm, this type of patch will be endowed larger weight in the fusion process. Thus, the nuclear-norm based strategy leads to better fusion performance.

Based on this analysis, on our MDLatLRR fusion framework, the projection matrix is  $L_{16}$ , the decomposition level varies between 1 to 4 (means  $r = 1, 2, 3, 4$ ) and nuclear-norm is used in the detail parts fusion strategy.

**3) The Stride of Sliding Window Technique:** In the sliding window technique, different strides could cause different fusion performance. In our experiment, the stride ( $s$ ) is set to  $\{1, 2, 4, 6, 8, 10, 12, 14\}$  to analyze the influence.

As discussed before, the image patch size is  $16 \times 16$ , the projection matrix is  $L_{16}$  and the decomposition level is set to  $\{1, 2, 3, 4\}$ . We choose six quality metrics ( $Q_{abf}$ ,  $FMI_{pixel}$ ,  $FMI_w$ ,  $FMI_{dct}$ ,  $SCD$  and  $MS-SSIM$ ) to evaluate the fusion performance with different strides in the sliding window technique. The average metric values of all test images are shown in Figure 12.

In Figure 12, when the decomposition level is set to 1, the quality metrics (except  $FMI_w$  and  $FMI_{dct}$ ) values are quite stable. The reason is that the detail information is not fully extracted by MDLatLRR with the shallow decomposition level (such as level 1), while the base parts contain more source image information. Thus, in this case, the stride has less influence.

However, with the increase in the number of decomposition levels, the detail parts preserve more texture and structural information, and our proposed method achieves better fusion performance. Thus, the stride of the sliding window should be chosen carefully when the decomposition level deepens.

In all six metrics, increasing the stride leads to fusion performance degradation no matter what the decomposition levels are (*level-2*, *level-3* and *level-4*). When the stride is large (more than 2), less feature information is processed by the projection matrix and the detail parts contain much less detail information. We can observe that more usefull information is extracted by a small stride, and the overall fusion performance



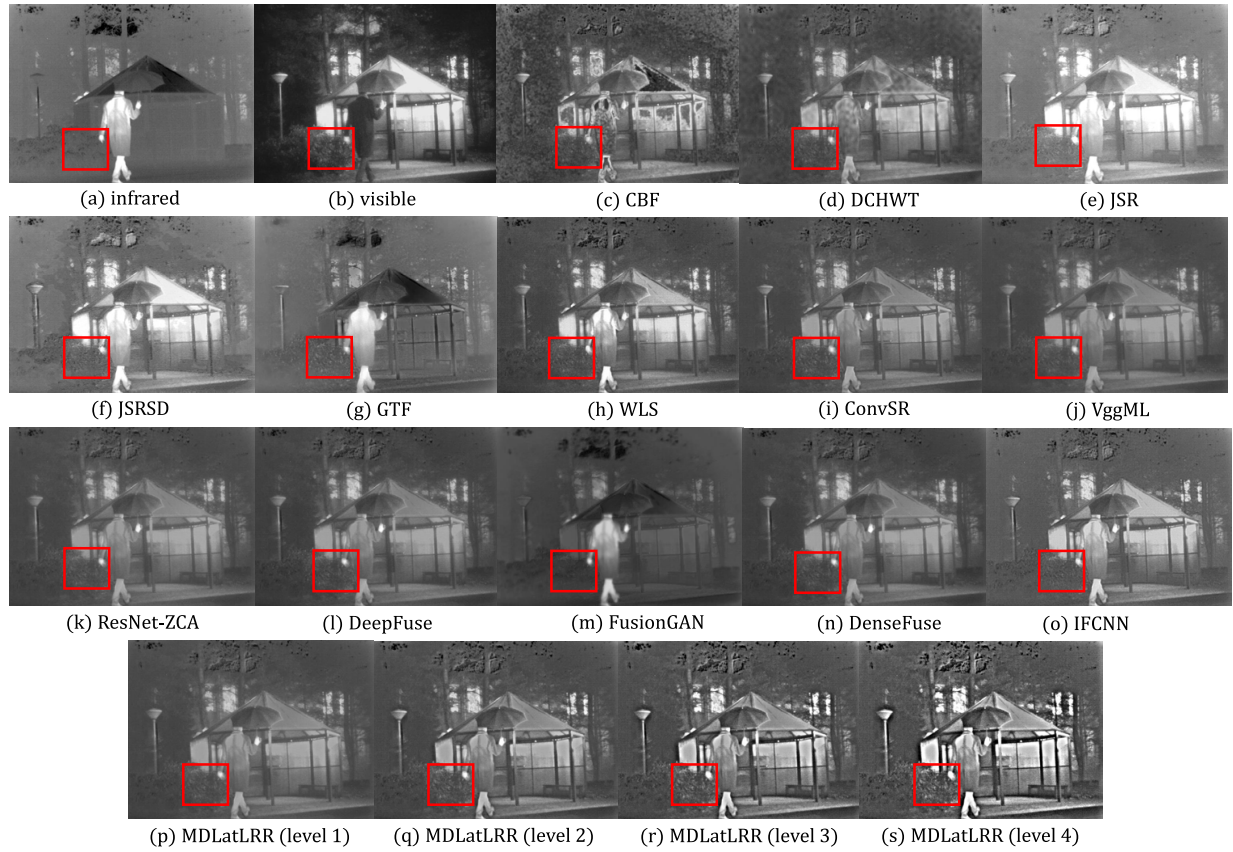


Fig. 13. Experiment on “umbrella” images. (a) Infrared image; (b) Visible image; (c) CBF; (d) DCHWT; (e) JSR; (f) JSRSD; (g) GTF; (h) WLS; (i) ConvSR; (j) VggML; (k) ResNet-ZCA; (l) DeepFuse; (m) FusionGAN; (n) DenseFuse; (o) IFCNN; (p-s) MDLatLRR ( $L_{16}$ ) with levels 1 to 4.

is improved. However, the small stride also makes our fusion method become time-consuming. Thus, the choice of stride should be optimized for different tasks.

In our experiment, the stride is set to 1 to obtain the best fusion performance. When the proposed method is applied to other computer vision tasks (such as RGBT object tracking), the time efficiency is an important factor, it is also acceptable that the stride is set to 4. This experiment is discussed in Section V-E.

### C. Subjective Evaluation

The fused results for three pairs of source images (“umbrella”, “soldier” and “man”)<sup>5</sup> are shown in Figure 13-15. These results are obtained by thirteen existing fusion methods and our algorithm (MDLatLRR), which uses  $L_{16}$ , nuclear-norm and decomposition levels (1 to 4). For sliding window technique, the stride is set to 1.

In Figure 13-15, the fused images obtained by CBF and DCHWT contain more artefacts and noise, and their salient features are not clear. The fused images obtained by JSR, JSRSD and GTF contain many artefacts, manifesting in ringing around the salient features.

In contrast, the fused images obtained by WLS, ConvSR, VggML, ResNet-ZCA, DeepFuse, FusionGAN, DenseFuse,

IFCNN and the proposed method capture more salient features and the fused images have a better visual quality.

Specifically for the bushes (red boxes) in Figure 13-15, comparing with the state-of-the-art fusion methods, MDLatLRR preserves more detailed information.

On the other hand, with the increasing number of decomposition levels (between 1 to 4), the salient features are better enhanced by our fusion methods and the detail information becomes more clear, as shown in Figure 13-15(p-s).

From these figures, our fusion method preserves more salient features from the infrared image and more detail information from the visible image in the subjective evaluation.

### D. Objective Evaluation

In order to demonstrate the attractive properties of our fusion method, ten quality metrics are used to compare the fusion performance of thirteen existing fusion methods and our algorithm.

The average values of the ten quality metrics for the set of test images are presented in Table I. The best and the second-best values are marked in **bold** and *red italic*, respectively.

In Table I, the proposed method achieves the best results in seven cases ( $En$ ,  $MI$ ,  $SD$ ,  $Q_{abf}$ ,  $FMI_w$ ,  $MS$ -SSIM and  $VIF$ ) and second-best values in four cases ( $FMI_p$ ,  $FMI_w$ ,  $VIF$  and  $EPI$ ). Although in  $SSIM_a$ , our method does not achieve the best value, it obtains a comparable result. These

<sup>5</sup>More experiments are presented in our supplementary materials.

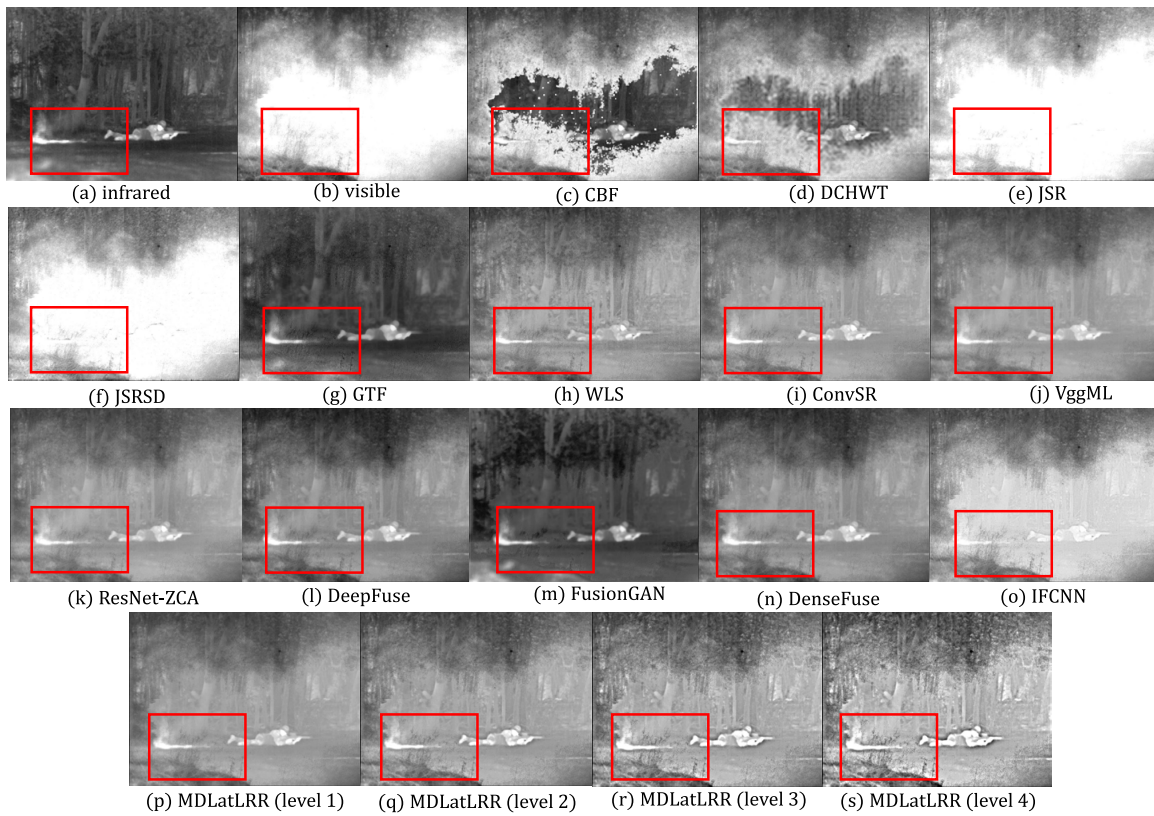


Fig. 14. Experiment on “soldier” images. (a) Infrared image; (b) Visible image; (c) CBF; (d) DCHWT; (e) JSR; (f) JSRSD; (g) GTF; (h) WLS; (i) ConvSR; (j) VggML; (k) ResNet-ZCA; (l) DeepFuse; (m) FusionGAN; (n) DenseFuse; (o) IFCNN; (p-s) MDLatLRR ( $L_{16}$ ) with levels 1 to 4.

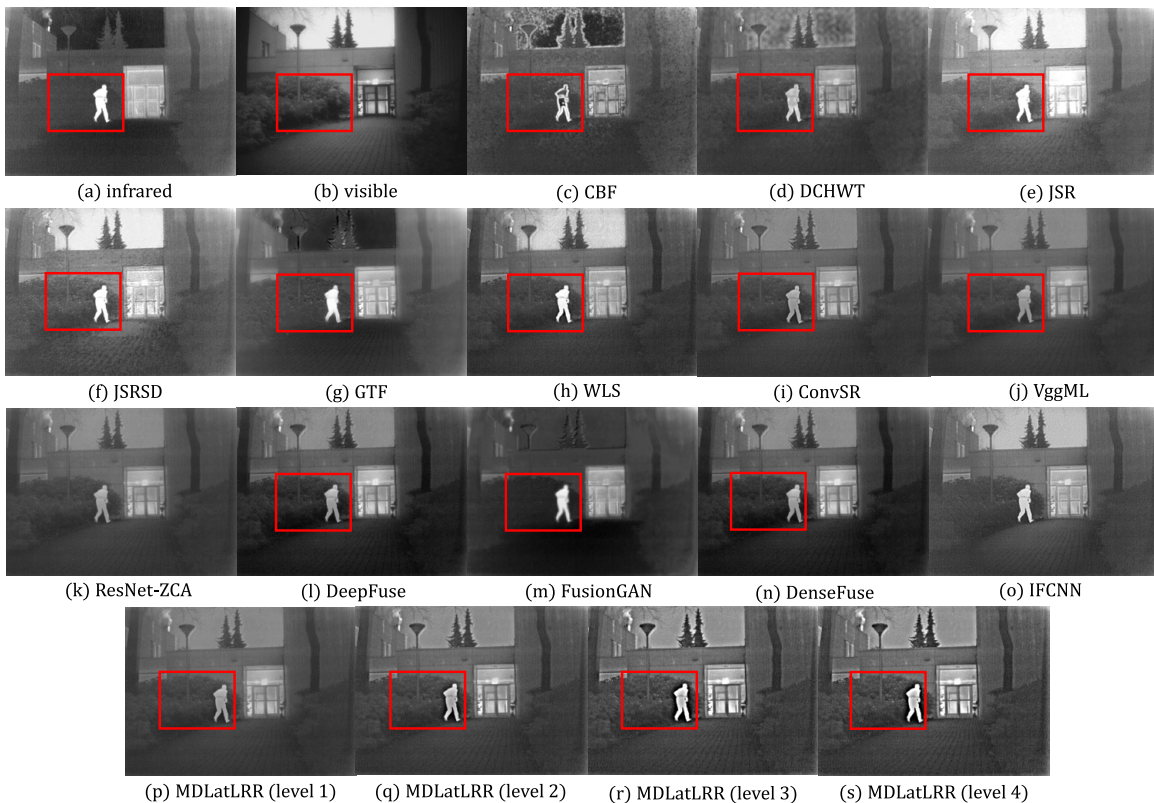


Fig. 15. Experiment on “man” images. (a) Infrared image; (b) Visible image; (c) CBF; (d) DCHWT; (e) JSR; (f) JSRSD; (g) GTF; (h) WLS; (i) ConvSR; (j) VggML; (k) ResNet-ZCA; (l) DeepFuse; (m) FusionGAN; (n) DenseFuse; (o) IFCNN; (p-s) MDLatLRR ( $L_{16}$ ) with levels 1 to 4.



TABLE I  
THE AVERAGE VALUES OF TEN QUALITY METRICS FOR ALL SOURCE IMAGES.

		En	MI	$SD$	$Q_{abf}$	$FMI_p$	$FMI_w$	$SSIM_a$	MS-SSIM	VIF	EPI
CBF [14]		6.85749	13.71498	76.82410	0.43961	0.87203	0.32350	0.59957	0.70879	0.71849	0.57240
DCHWT [13]		6.56777	13.13553	64.97891	0.46592	0.91377	0.40147	0.73132	0.84326	0.50560	0.70181
JSR [55]		6.72263	12.72654	74.10783	0.32306	0.88463	0.18506	0.54073	0.75523	0.63845	0.59644
JSRSD [23]		6.72057	13.38575	79.19536	0.32281	0.86429	0.18498	0.54127	0.75517	0.67071	0.47473
GTF [30]		6.63433	13.26865	67.54361	0.41037	0.90207	0.41038	0.70016	0.80844	0.41687	0.67011
WLS [33]		6.64071	13.28143	70.58894	0.50077	0.89788	0.37662	0.72360	0.93349	0.72874	0.67837
ConvSR [27]		6.25869	12.51737	50.74372	0.53485	0.91691	0.34640	0.75335	0.90281	0.39218	0.71130
VggML [21]		6.18260	12.36521	48.15779	0.36818	0.91070	0.41684	0.77799	0.87478	0.29509	0.77923
ResNet-ZCA [20]		6.19527	12.39054	48.71803	0.35098	0.90921	0.41689	0.77825	0.87324	0.29249	0.78366
DeepFuse [36]		6.69935	13.39869	68.79312	0.43797	0.90470	0.42477	0.72882	0.93353	0.65773	0.73795
FusionGAN [32]		6.36285	12.72570	67.57282	0.21834	0.89061	0.37083	0.65384	0.73182	0.45355	0.68470
DenseFuse [19]		6.67158	13.34317	54.35752	0.44009	0.90847	0.42767	0.73150	0.92896	0.64576	0.73321
IFCNN [56]		6.59545	13.19090	66.87578	0.50329	0.90092	0.40166	0.73186	0.90527	0.59029	0.73767
MDLatLRR	level-1	6.21986	12.43972	49.65098	0.44448	0.91494	0.42438	0.77501	0.89564	0.34773	0.76350
	level-2	6.44870	12.89740	57.69733	0.54275	0.91449	0.44169	0.74783	0.94824	0.71472	0.71142
	level-3	6.74064	13.48128	69.91578	0.49264	0.90599	0.44793	0.67248	0.92781	1.39009	0.66903
	level-4	6.97774	13.95548	83.77407	0.33574	0.89546	0.44480	0.57736	0.85222	2.27705	0.63077



Fig. 16. Examples of RGBT frames in the VOT-RGBT2019 benchmark. Each cloumn denotes a pair of frames which come from a video. First row shows RGB frames and the infrared frames are shown in the second row. The groundtruth of each frame is indicated by the blue box.

values indicate that the fused images obtained by our method are more natural and contain fewer artefacts. The objective evaluation shows that our fusion method delivers better fusion performance than the existing methods.

Specifically for the quality metrics of  $EN$  and  $MI$ , we note that CBF achieved the second-best values. Its fused images contain more noise and artefacts, as shown in Figure 13-15(c). However, our fusion method achieves the best values on  $EN$  and  $MI$ , thanks to the enhancement of the salient features, instead of amplifying noise and generating artefacts. Thus our fusion method also exhibits a feature enhancement capability.

#### E. Fusion Method for Visual Object Tracking(RGB-T)

As described in VOT2019 [12], the visual object tracking challenges address short-term or long-term, causal and model-free tracking [51] [50] [49]. With the wide application of multiple sensors in computer vision, it is important to carefully process multi-modal data. Thus, in VOT2019, the committee

introduced two new sub-challenges (VOT-RGBT and VOT-RGBD). The VOT-RGBT sub-challenge focuses on short-term tracking. It contains two modalities (RGB and thermal infrared). The RGBT tracking problem is ideally suited to take advantage of our fusion method to improve the tracking performance.

In this experiment, we apply our fusion method to the RGBT visual object tracking task to verify the efficiency of our proposed fusion method in other computer vision task. The VOT-RGBT benchmark is available at [12] which contains 60 video sequences. The examples of these video frames are shown in Figure 16.

Firstly, the RGB and infrared frames are fused by our MDLatLRR fusion method. The parameters of MDLatLRR are the same as in the fusion task except the stride ( $s$ ) is set to 4. Then, the fused frame is fed into different trackers<sup>6</sup>. To evaluate the effect of our fusion method on RGBT tracking, we

<sup>6</sup>The combination framework of our fusion method and trackers is shown in supplementary materials.

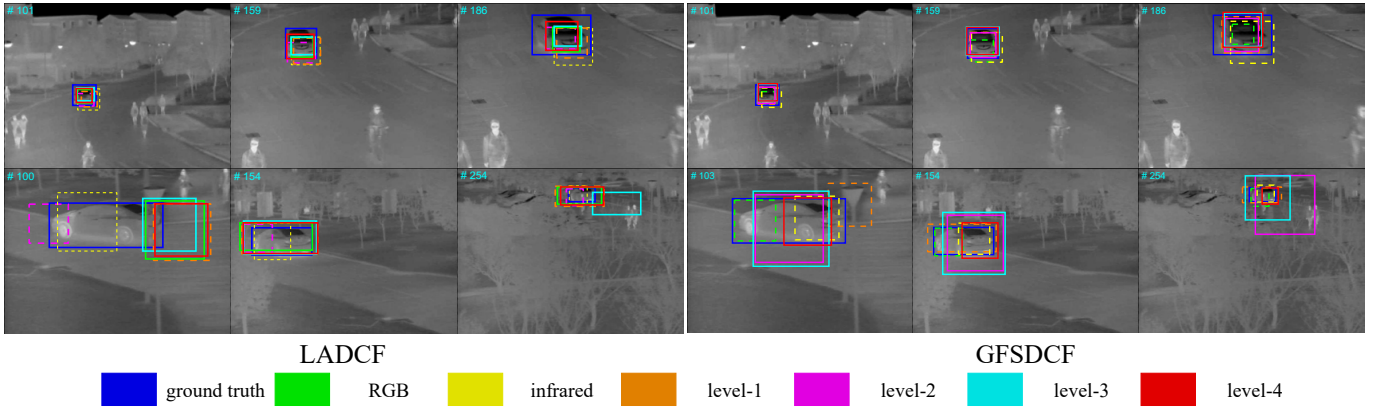


Fig. 17. The tracking results on the VOT-RGBT2019 benchmark. The frames in first row and second row are selected from ‘car10’ and ‘car41’, respectively. First three columns are the results of LADCF. The last three columns are the tracking results of GFSDCF. The *RGB* and *infrared* denote the case when the input of the trackers is just one modality (RGB or infrared). The *level-1* to *level-4* present the cases when the input of the trackers are the fused frames generated by MDLatLRR.

TABLE II

TRACKING RESULTS USING LADCF AND MDLatLRR ON THE VOT-RGBT2019 DATASET. MDLatLRR IS USED TO FUSE RGB AND INFRARED FRAMES.

	Type	EAO	Accuracy	Failures
LADCF [51]	<i>RGB</i>	0.2518	0.5406	41.2745
	<i>infrared</i>	0.1613	0.5109	57.6843
<b>LADCF</b> + <b>MDLatLRR</b>	<i>level-1</i>	0.2553	<i>0.5650</i>	40.4008
	<i>level-2</i>	<b>0.2633</b>	<b>0.5740</b>	42.4033
	<i>level-3</i>	0.2608	0.5604	<i>38.8786</i>
	<i>level-4</i>	<i>0.2618</i>	0.5331	<b>37.8026</b>

TABLE III

TRACKING RESULTS USING GFSDCF AND MDLatLRR ON THE VOT-RGBT2019 DATASET. MDLatLRR IS USED TO FUSE RGB AND INFRARED FRAMES.

	Type	EAO	Accuracy	Failures
GFSDCF [50]	<i>RGB</i>	<i>0.3104</i>	0.5832	<i>31.9099</i>
	<i>infrared</i>	0.1941	0.5608	55.5163
<b>GFSDCF</b> + <b>MDLatLRR</b>	<i>level-1</i>	0.3032	<b>0.6098</b>	35.3073
	<i>level-2</i>	0.3092	0.5916	32.1395
	<i>level-3</i>	0.2963	<i>0.5989</i>	34.3755
	<i>level-4</i>	<b>0.3107</b>	0.5824	<b>29.0613</b>

choose two trackers (LADCF [51], GFSDCF [50]) to evaluate the tracking performance on the VOT-RGBT2019 benchmark.

The LADCF [51] obtained the best tracking performance on the public dataset of VOT-ST2018 sub-challenge. Based on LADCF, the GFSDCF [50] explored group feature selection strategies to reduce the redundancy of deep features. We use the VOT2018 version code of LADCF. The parameters of GFSDCF were learned on OTB-2015 [47] dataset. The codes of these trackers were provided by the original authors and could be downloaded from [48].

As in VOT2019 [12], we choose three measures to analyze the tracking performance: accuracy, failures and expected average overlap (EAO). The accuracy indicates the average overlap between the predicted and ground truth bounding boxes during successful tracking periods. The failure measures the robustness of a tracker. The third primary measure, EAO, is an estimator of the average overlap a tracker manages to attain on a large collection with the same visual properties as the ground-truth. The larger the values of accuracy and EAO, the better tracking performance, and the lower the failure value the better robustness.

The evaluation measure values of LADCF and GFSDCF with our fusion method computed on VOT-RGBT are shown in Table II and Table III, respectively. The best values and the second-best values are marked in **bold** and *red italic*, respectively. We also choose several frames to demonstrate

the tracking results<sup>7</sup> on VOT-RGBT2019. These are shown in Figure 17. The ground truth is specified in the infrared-channel since this is the primary modality. The infrared frames are shown in Figure 17.

In Table II, Table III and Figure 17, the *RGB* and *infrared* denote that only one modality (RGB or infrared) is fed into the trackers. The *level-1* to *level-4* indicate that the RGB and infrared frames are firstly fused by our fusion method, then fed into the trackers. From these two tables, it is observed that with the pre-processing of our fusion method, both LADCF and GFSDCF trackers achieve better performance in three measures. Furthermore, when the decomposition level is increased, the trackers become more robust (lower values of Failures).

In VOT2019, the EAO is the primary measure. With the fused frame fed into the trackers, the value of EAO is increased. However, for GFSDCF, the increase is slight. The reason is that the fusion strategy in this experiment is pixel-level fusion which is particularly suited for extracting handcrafted features. Since LADCF has a greater proportion of handcrafted features, LADCF with *level-2* decomposition obtains a higher EAO value. In GFSDCF, the deep features are more powerful, thus the influence of the fused frames is very limited (the EAO value obtained by fused frame is close to RGB frame).

<sup>7</sup>More tracking results are shown in supplementary materials.



In conclusion, even with the simplest way of feeding our fusion results into the trackers, the tracking performance is consistently improved. The fused images enhanced by our fusion method have a positive impact on this computer vision (RGBT tracking) application. Furthermore, if the tracker engages a greater proportion of deep features for the data representation, the fusion method should focus on feature-level fusion and decision-level fusion. Thus, in the future, we will focus on these directions to improve RGBT tracker performance.

## VI. CONCLUSIONS

In this paper, we proposed a novel multi-level decomposition method (MDLatLRR) for image decomposition. We also developed an MDLatLRR-based fusion framework for fusing infrared and visible images. Firstly, a projection matrix  $L$  is learned by LatLRR. This matrix is then utilized to extract detail parts and base parts of the input images at several representation levels. With MDLatLRR, multi-level salient features are extracted. The final fused image is reconstructed by adaptive fusion strategies designed specifically for dealing with the detail parts and the base parts, respectively. The MDLatLRR framework is general and can be used to provide an efficient decomposition approach for extracting multi-level features for an arbitrary number of input images. It can also be applied to other image processing fields with different projection matrix. The proposed method was evaluated both subjectively and objectively in a number of experiments. The experimental results demonstrate that the performance of the proposed method is superior to that of existing methods.

Furthermore, we also applied our proposed fusion method to a RGBT object tracking task. Even with the simplest combination of our fusion method and different trackers, the tracking performance was improved.

In the future, we will investigate the link between the proposed decomposition method and other decomposition methods. Other potential applications of the proposed decomposition method to other tasks of image processing will be considered.

## REFERENCES

- [1] V. Aslantas and E. Bendes. A new image quality metric for image fusion: the sum of the correlations of differences. *Aeu-international Journal of electronics and communications*, 69(12):1890–1896, 2015. 5
- [2] A. Ben Hamza, Y. He, H. Krim, and A. Willsky. A multiscale approach to pixel-level image fusion. *Integrated Computer-Aided Engineering*, 12(2):135–146, 2005. 1
- [3] R. A. Borsoi, T. Imbiriba, and J. C. M. Bermudez. Super-resolution for hyperspectral and multispectral image fusion accounting for seasonal spectral variability. *IEEE Transactions on Image Processing*, 29:116–127, 2019. 1
- [4] J. Chen, X. Li, L. Luo, X. Mei, and J. Ma. Infrared and visible image fusion based on target-enhanced multiscale transform decomposition. *Information Sciences*, 508:64–78, 2020. 1
- [5] Z. Chen, X.-J. Wu, and J. Kittler. A sparse regularized nuclear norm based matrix regression for face recognition with contiguous occlusion. *Pattern Recognition Letters*, 2019. 1
- [6] Z. Chen, X.-J. Wu, H.-F. Yin, and J. Kittler. Noise-Robust Dictionary Learning with Slack Block-Diagonal Structure for Face Recognition. *Pattern Recognition*, page 107118, 2019. 1
- [7] R. Gao, S. A. Vorobyov, and H. Zhao. Image fusion with cospase analysis operator. *IEEE Signal Processing Letters*, 24(7):943–947, 2017. 1
- [8] M. Haghighat and M. A. Razian. Fast-FMI: non-reference image fusion metric. In *Application of Information and Communication Technologies (AICT), 2014 IEEE 8th International Conference on*, pages 1–3. IEEE, 2014. 5
- [9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1
- [10] M. Hossny, S. Nahavandi, and D. Creighton. Comments on ‘Information measure for performance of image fusion’. *Electronics letters*, 44(18):1066–1067, 2008. 5
- [11] M.-x. Jiang, C. Deng, J.-s. Shan, Y.-y. Wang, Y.-j. Jia, and X. Sun. Hierarchical multi-modal fusion FCN with attention model for RGB-D tracking. *Information Fusion*, 50:1–8, 2019. 1
- [12] M. Kristan, J. Matas, A. Leonardis, M. Felsberg, R. Pflugfelder, J.-K. Kamarainen, L. Cehovin Zajc, O. Drbohlav, A. Lukežić, A. Berg, et al. The seventh visual object tracking vot2019 challenge results. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 11, 12
- [13] B. S. Kumar. Multifocus and multispectral image fusion based on pixel significance using discrete cosine harmonic wavelet transform. *Signal, Image and Video Processing*, 7(6):1125–1143, 2013. 5, 11
- [14] B. S. Kumar. Image fusion based on pixel significance using cross bilateral filter. *Signal, image and video processing*, 9(5):1193–1204, 2015. 5, 11
- [15] X. Lan, M. Ye, S. Zhang, H. Zhou, and P. C. Yuen. Modality-correlation-aware sparse representation for RGB-infrared object tracking. *Pattern Recognition Letters*, 2018. 1
- [16] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang. RGB-T object tracking: benchmark and baseline. *Pattern Recognition*, 96:106977, 2019. 1
- [17] H. Li. Codes for MDLatLRR. 2019. [https://github.com/hli1221/imagefusion\\_mdlatlr](https://github.com/hli1221/imagefusion_mdlatlr). 3, 5
- [18] H. Li and X.-J. Wu. Multi-focus image fusion using dictionary learning and low-rank representation. In *International Conference on Image and Graphics*, pages 675–686. Springer, 2017. 1, 2
- [19] H. Li and X.-J. Wu. DenseFuse: A Fusion Approach to Infrared and Visible Images. *IEEE Transactions on Image Processing*, 28(5):2614–2623, 2019. 1, 5, 11
- [20] H. Li, X.-J. Wu, and T. S. Durrani. Infrared and Visible Image Fusion with ResNet and zero-phase component analysis. *Infrared Physics & Technology*, page 103039, 2019. 1, 5, 11
- [21] H. Li, X.-J. Wu, and J. Kittler. Infrared and Visible Image Fusion using a Deep Learning Framework. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2705–2710. IEEE, 2018. 1, 5, 11
- [22] S. Li, X. Kang, L. Fang, J. Hu, and H. Yin. Pixel-level image fusion: A survey of the state of the art. *Information Fusion*, 33:100–112, 2017. 1
- [23] C. Liu, Y. Qi, and W. Ding. Infrared and visible image fusion method based on saliency detection in sparse domain. *Infrared Physics & Technology*, 83:94–102, 2017. 1, 5, 11
- [24] G. Liu, Z. Lin, and Y. Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, pages 663–670, 2010. 2
- [25] G. Liu and S. Yan. Latent low-rank representation for subspace segmentation and feature extraction. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1615–1622. IEEE, 2011. 1, 2, 4
- [26] Y. Liu, X. Chen, H. Peng, and Z. Wang. Multi-focus image fusion with a deep convolutional neural network. *Information Fusion*, 36:191–207, 2017. 1
- [27] Y. Liu, X. Chen, R. K. Ward, and Z. J. Wang. Image fusion with convolutional sparse representation. *IEEE signal processing letters*, 23(12):1882–1886, 2016. 1, 5, 11
- [28] X. Lu, B. Zhang, Y. Zhao, H. Liu, and H. Pei. The infrared and visible image fusion algorithm based on target separation and sparse representation. *Infrared Physics & Technology*, 67:397–407, 2014. 1
- [29] X. Luo, Z. Zhang, B. Zhang, and X. Wu. Image fusion with contextual statistical similarity and nonsubsampling shearlet transform. *IEEE Sensors Journal*, 17(6):1760–1771, 2017. 1
- [30] J. Ma, C. Chen, C. Li, and J. Huang. Infrared and visible image fusion via gradient transfer and total variation minimization. *Information Fusion*, 31:100–109, 2016. 5, 11
- [31] J. Ma, P. Liang, W. Yu, C. Chen, X. Guo, J. Wu, and J. Jiang. Infrared and visible image fusion via detail preserving adversarial learning. *Information Fusion*, 54:85–98, 2020. 1
- [32] J. Ma, W. Yu, P. Liang, C. Li, and J. Jiang. FusionGAN: A generative adversarial network for infrared and visible image fusion. *Information Fusion*, 48:11–26, 2019. 1, 5, 11
- [33] J. Ma, Z. Zhou, B. Wang, and H. Zong. Infrared and visible image fusion based on visual saliency map and weighted least square optimization. *Infrared Physics & Technology*, 82:8–17, 2017. 5, 11

- [34] K. Ma, K. Zeng, and Z. Wang. Perceptual quality assessment for multi-exposure image fusion. *IEEE Transactions on Image Processing*, 24(11):3345–3356, 2015. 5
- [35] H. Pang, M. Zhu, and L. Guo. Multifocus color image fusion using quaternion wavelet transform. In *Image and Signal Processing (CISP), 2012 5th International Congress on*, pages 543–546. IEEE, 2012. 1
- [36] K. R. Prabhakar, V. S. Srikar, and R. V. Babu. Deepfuse: A deep unsupervised approach for exposure fusion with extreme exposure image pairs. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, pages 4724–4732, 2017. 1, 5, 11
- [37] Y.-J. Rao. In-fibre Bragg grating sensors. *Measurement science and technology*, 8(4):355, 1997. 5, 6
- [38] J. W. Roberts, J. A. Van Aardt, and F. B. Ahmed. Assessment of image fusion procedures using entropy, image quality, and multispectral classification. *Journal of Applied Remote Sensing*, 2(1):023522, 2008. 5
- [39] F. Sattar, L. Floreby, G. Salomonsson, and B. Lovstrom. Image enhancement based on a nonlinear multiscale method. *IEEE transactions on image processing*, 6(6):888–895, 1997. 5
- [40] H. R. Sheikh and A. C. Bovik. Image information and visual quality. *IEEE Transactions on image processing*, 15(2):430–444, 2006. 5
- [41] V. Shrinidhi, P. Yadav, and N. Venkateswaran. IR and Visible Video Fusion for Surveillance. In *2018 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)*, pages 1–6. IEEE, 2018. 1
- [42] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1
- [43] A. Toet. TNO Image Fusion Dataset, 2014. [https://figshare.com/articles/TN\\_Image\\_Fusion\\_Dataset/1008029](https://figshare.com/articles/TN_Image_Fusion_Dataset/1008029). 5
- [44] A. Vishwakarma and M. Bhuyan. Image Fusion Using Adjustable Non-subsampled Shearlet Transform. *IEEE Transactions on Instrumentation and Measurement*, 2018. 1
- [45] L. Wang, B. Li, and L.-F. Tian. EGGDD: An explicit dependency model for multi-modal medical image fusion in shift-invariant shearlet transform domain. *Information Fusion*, 19:29–37, 2014. 1
- [46] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma. Robust face recognition via sparse representation. *IEEE transactions on pattern analysis and machine intelligence*, 31(2):210–227, 2008. 1
- [47] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015. 12
- [48] T. Xu. The codes of LADCF and GFSDCF, 2019. <https://github.com/XU-TIANYANG>. 12
- [49] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler. An Accelerated Correlation Filter Tracker. *Pattern Recognition accepted. arXiv preprint arXiv:1912.02854*, 2019. 11
- [50] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler. Joint group feature selection and discriminative filter learning for robust visual object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7950–7960, 2019. 11, 12
- [51] T. Xu, Z.-H. Feng, X.-J. Wu, and J. Kittler. Learning Adaptive Discriminative Correlation Filters via Temporal Consistency preserving Spatial Feature Selection for Robust Visual Object Tracking. *IEEE Transactions on Image Processing*, 2019. 11, 12
- [52] C. Xydeas, , and V. Petrovic. Objective image fusion performance measure. *Electronics letters*, 36(4):308–309, 2000. 5
- [53] S. Yang, M. Wang, L. Jiao, R. Wu, and Z. Wang. Image fusion based on a new contourlet packet. *Information Fusion*, 11(2):78–84, 2010. 1
- [54] M. Yin, P. Duan, W. Liu, and X. Liang. A novel infrared and visible image fusion algorithm based on shift-invariant dual-tree complex shearlet transform and sparse representation. *Neurocomputing*, 226:182–191, 2017. 1
- [55] Q. Zhang, Y. Fu, H. Li, and J. Zou. Dictionary learning method for joint sparse representation-based image fusion. *Optical Engineering*, 52(5):057006, 2013. 5, 11
- [56] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang. IFCNN: A general image fusion framework based on convolutional neural network. *Information Fusion*, 54:99–118, 2020. 1, 5, 11
- [57] J.-j. Zong and T.-s. Qiu. Medical image fusion based on sparse representation of classified image patches. *Biomedical Signal Processing and Control*, 34:195–205, 2017. 1



**Hui Li** received the B.Sc. degree in School of Internet of Things Engineering from Jiangnan University, China, in 2015. He is currently a PhD student at the Jiangsu Provincial Engineering Laboratory of Pattern Recognition and Computational Intelligence, Jiangnan University. His research interests include image fusion, machine learning and deep learning.



**Xiao-Jun Wu** received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991, and the M.S. degree and Ph.D. degree in pattern recognition and intelligent system from the Nanjing University of Science and Technology, Nanjing, in 1996 and 2002, respectively. From 1996 to 2006, he taught at the School of Electronics and Information, Jiangsu University of Science and Technology, where he was promoted to Professor.

He has been with the School of Information Engineering, Jiangnan University since 2006, where he is a Professor of pattern recognition and computational intelligence. He was a Visiting Researcher with the Centre for Vision, Speech, and Signal Processing (CVSSP), University of Surrey, U.K. from 2003 to 2004. He has published over 300 papers in his fields of research. His current research interests include pattern recognition, computer vision, and computational intelligence. He was a Fellow of the International Institute for Software Technology, United Nations University, from 1999 to 2000. He was a recipient of the Most Outstanding Postgraduate Award from the Nanjing University of Science and Technology.



**Josef Kittler** (M’74-LM’12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is a distinguished Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook *Pattern Recognition: A Statistical Approach* and over 700 scientific papers. His publications have been cited more than 60,000 times (Google Scholar).

He is series editor of Springer Lecture Notes on Computer Science. He currently serves on the Editorial Boards of *Pattern Recognition Letters*, *Pattern Recognition and Artificial Intelligence*, *Pattern Analysis and Applications*. He also served as a member of the Editorial Board of *IEEE Transactions on Pattern Analysis and Machine Intelligence* during 1982–1985. He served on the Governing Board of the International Association for Pattern Recognition (IAPR) as one of the two British representatives during the period 1982–2005, President of the IAPR during 1994–1996.