Multi-focus image fusion based on non-negative sparse representation and patch-level consistency rectification

Qiang Zhang[a,b], Guanghe Li[b], Yunfeng Cao[b], Jungong Han[c*]

[a] *Key Laboratory of Electronic Equipment Structure Design, Ministry of Education, Xidian University, Xi'an, Shaanxi 710071, China*

[b] *Center for Complex Systems, School of Mechano-electronic Engineering, Xidian University, Xi'an Shaanxi 710071,China*

[c] *WMG Data Science, University of Warwick, Coventry CV4 AL7, U.K.*

Abstract Most existing sparse representation-based (SR) fusion methods consider the local information of each image patch independently during fusion. Some spatial artifacts are easily introduced to the fused image. A sliding window technology is often employed by these methods to overcome this issue. However, this comes at the cost of high computational complexity. Alternatively, we come up with a novel multi-focus image fusion method that takes full consideration of the strong correlations among spatially adjacent image patches with *NO* need for a sliding window. To this end, a non-negative SR model with local consistency constraint (CNNSR) on the representation coefficients is first constructed to encode each image patch. Then a patch-level consistency rectification strategy is presented to merge the input image patches, by which the spatial artifacts in the fused images are greatly reduced. As well, a compact non-negative dictionary is constructed for the CNNSR model. Experimental results demonstrate that the proposed fusion method outperforms some state-of-the art methods. Moreover, the proposed method is computationally efficient, thereby facilitating real-world applications.

Keywords: Multi-focus image fusion, non-negative sparse representation, compact non-negative dictionary construction, patch-level consistency rectification, high computational efficiency

---

1. Introduction

Multi-focus image fusion is a process of combining several images with different focus points into a composite image with full-focus [1]. So far, numerous multi-focus image fusion methods have been presented [1,2]. One of the critical components in these methods is to determine a decision map by using

*Corresponding author. Address: University of Warwick, Coventry CV4 AL7, UK. Email address: jungonghan77@gmail.com (J. Han).

some measure of focus (MOF). This decision map helps to select the focused regions in various input images and preserve those regions on the fused image. High computational efficiency is also desirable in many real-time applications. In this paper, we will address such issues by using a non-negative sparse representation (NNSR) model with some local spatial consistency priors.

As a result of their successful applications in many computer vision and image processing tasks, spare representation (SR) [3] as well as its variants have been introduced to multi-sensor image fusion, including multi-focus image fusion, in recent years [1,2,4-9]. In these SR-based fusion methods, the traditional SR model [3] seems to be the most popular one used to achieve the sparse coding of the input image patches [10]. However, the traditional SR model just performs a sparsity constraint on the representation coefficients with the consequence that the representation coefficients for each image patch contain both positive and negative values. This apparently contradicts the non-negative property of image patches, i.e., the intensity of each pixel in an image patch is non-negative. Therefore, it is questionable if such representation coefficients are really meaningful and reasonable [11].

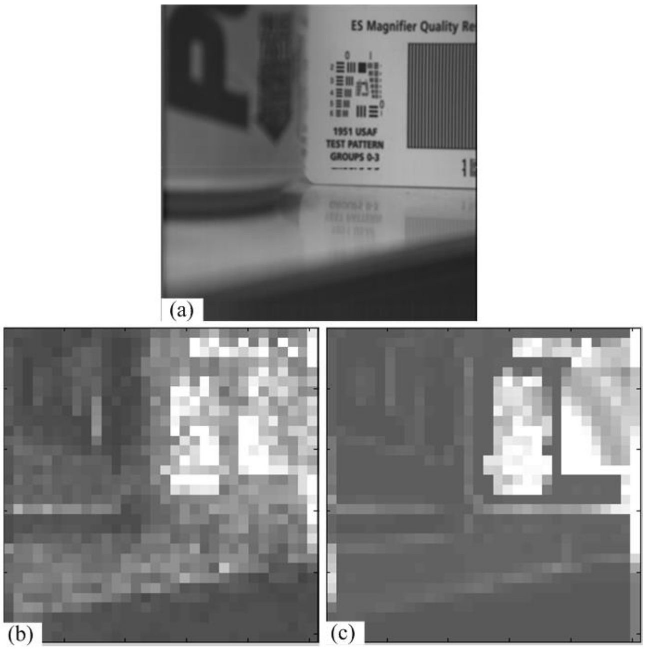

Fig. 1. Superiority of NNSR over SR when applied to multi-focus image fusion. (a) An image with focus on the right part; (b) Representation coefficients obtained by SR; (c) Representation coefficients obtained by NNSR. As shown in (b), the representation coefficients for the left part have high absolute values in addition to those for the right part. While, as shown in (c), only the representation coefficients for the right part have high values. This demonstrates that the representation coefficients obtained by NNSR can more accurately determine the focused and defocused regions in a multi-focus image than those obtained by SR.

Different from the traditional SR model, the non-negative sparse representation (NNSR) jointly imposes the sparsity and non-negativity constraints on the representation coefficients. As discussed in

[11], the source images can be efficiently encoded by using "few" components with the sparsity constraint. In addition, the representation for each image is purely additive because of the non-negativity constraint. When applied to multi-focus images, the non-negative representation coefficients obtained by using NNSR can better capture the focus information of the input image than the coefficients obtained by the traditional SR model. This is shown in Fig.1. Therefore, in this paper, we will employ NNSR in our proposed fusion method.

It should be noted that the input images are needed to be divided into a set of patches in most SR-based fusion methods prior to being sparsely coded and fused. As well, these image patches are independently considered during the fusion process. Some spatial artifacts are thus easily introduced to the fused image. In order to address such issue, the sliding window technology [4] is often used in these fusion methods. However, this greatly increases the computational complexity of a fusion method. In addition, some detailed information in the fused image may also be lost during the fusion process [12, 13].

In fact, there exists strong correlations or spatial consistency among these spatially adjacent patches Specifically, these spatial adjacent image patches have similar focus pattern, i.e., they are either all in-focus or all out-focus in most cases. In view of this, we will employ such spatial consistency prior among the image patches, instead of the sliding window, in our proposed fusion method to reduce the spatial artifacts in the fused image. Furthermore, it is desirable to improve the computational efficiency of the fusion method.

To achieve this goal, we first present a new non-negative sparse representation model with local consistency constraint (CNNSR) that adds a Laplacian regularization term on the representation coefficient matrix, when encoding the input image patches. The intention of adding such a Laplacian regularization term is to enforce the spatially-adjacent patches with similar features to have similar representation coefficients and thus similar focus information. In the subsequent fusion process, we will present a patch-level consistency rectification strategy, further ensuring each input image patch to have similar focus information with most of its spatial neighbors. Apart from its simplicity, the proposed patch-level consistency rectification strategy can significantly suppress the spatial artifacts in the fused image. In addition, it can also increase the computational efficiency of the fusion method due to: 1) The proposed patch-level consistency rectification strategy allows input images to be divided into a set of non-overlapped patches, rather than a set of overlapped patches, during the fusion process; and 2) A compact non-negative dictionary is constructed for the CNNSR model when encoding the image patches, which will further reduce the computational complexity of the fusion method. Several sets of experimental results demonstrate the validity of the proposed fusion method.

Our main contributions are summarized as follows:

(1) We propose a non-negative sparse representation (CNNSR) model with local consistency constraint imposed onto the representation coefficients for multi-focus image fusion, taking advantage of the strong correlations among spatially-adjacent patches.

(2) We present a compact non-negative dictionary learning (CNNDL) method for the proposed CNNSR model, which employs an orthogonality constraint as well as a non-negativity constraint to reduce the redundancy among dictionary atoms.

(3) We propose a patch-level consistency rectification strategy during the fusion process, instead of the sliding window technology, to reduce the spatial artifacts in the fused images and increase the computational efficiency of the proposed method.

The rest of the paper is organized as follows. Section 2 briefly reviews the related work. Section 3 details the dictionary construction method for NNSR. Section 4 elaborates the proposed fusion method. Experimental results and conclusions are provided in Section 5 and Section 6, respectively.

2. Related work

So far, numerous fusion methods for multi-focus images have been presented, which may be simply categorized into two groups, i.e., transform-domain-based and spatial-domain-based. Among the former, most methods follow the idea of multi-scale transform-based (MST) fusion algorithm [14], including those based on wavelet transform [15], contourlet transform [16], neighbor distance [17], and so on.

The earlier spatial-domain-based fusion methods are generally pixels or blocks based ones, which easily introduce spatial artifacts to the fused images. Recently, some advanced fusion methods based on image matting [18, 19], dense scale invariant transform (DSIFT) [20], and even convolutional neural network (CNN) [21, 22], are presented to suppress the spatial artifacts.

In [4], the spare representation theory was first introduced to multi-sensor image fusion. Since then, varieties of multi-sensor image fusion, including multi-focus image fusion, were presented based on different SR models, such as robust SR (RSR) [1, 13], joint SR (JSR) [23], group SR (GSR) [24] and NNSR [11]. However, in most of these fusion methods, each input image patch is independently encoded and fused. This ignores the strong correlations (or spatial consistency) among spatially-adjacent patches and easily introduces some undesirable spatial artifacts to the fused images.

Considering that, a multi-task RSR (MRSR) model [13] was proposed and applied to integrate multi-focus images, where the focus information of each image patch was jointly determined by its spatial contextual information as well as its local information. Despite its desirable fusion performance, the MRSR-based fusion method is at the cost of high computational complexity. For that, an improved multi-focus image fusion method based on RSR model was proposed in [1]. However, the computational complexity of the RSR-based fusion method in [1] is still high.

In addition to SR models, the constructed over-complete dictionaries also play an important role in improving fusion performance and computational efficiency of a fusion method [10]. These dictionaries may be directly constructed from some fixed (e.g., Discrete Cosine Transform (DCT) or Wavelet) basis [4]. They can also be learned from a set of auxiliary images (called *globally-trained* ones) [25] or input images themselves (called *adaptively-trained* ones) [2] by using various learning methods, such as K-Singular Value Decomposition (K-SVD) [26]. Generally, those learned dictionaries could achieve better fusion performance than those with a fixed basis.

However, most of these dictionary learning methods focus on enhancing the representation capability of the dictionary, but ignore the correlations among the dictionary atoms. As a result of that, those learned dictionaries may have good representation capability while highly redundant. This will not only increase the computational complexity of the subsequent fusion method but degrade the fusion performance. A compact dictionary with a small number of atoms maintaining high representation capability is greatly desirable in image fusion [10].

3. Compact non-negative dictionary learning (CNDL) for NNSR

As discussed in the previous Section 1, we will employ a NNSR model, more specifically the CNNSR model, to encode source image patches during the fusion process. For that, we will discuss how to construct a compact non-negative dictionary for the NNSR model in detail in this section.

Suppose that $\mathbf{Y}=\left[\mathbf{y}_1,\mathbf{y}_2,\cdots,\mathbf{y}_N\right]\in R_+^{n\times N}$ contains N data samples of dimension n. Each column $\mathbf{y}_i\in R_+^n$ in the matrix $\mathbf{Y}$ represents a data vector. A non-negativity dictionary $\mathbf{D}=\left[\mathbf{d}_1,\mathbf{d}_2,...,\mathbf{d}_M\right]\in R_+^{n\times M}$ with M dictionary atoms may be learned by [11, 27]

$$\left(\mathbf{D},\mathbf{X}\right)=\arg\min_{\mathbf{D},\mathbf{X}}\frac{1}{2}\left\|\mathbf{Y}-\mathbf{D}\mathbf{X}\right\|_F^2+\lambda\left\|\mathbf{X}\right\|_1 \qquad s.t.\ \mathbf{D}\geq\mathbf{0},\mathbf{X}\geq\mathbf{0} \quad . \tag{1}$$

Here, each column $\mathbf{d}_m\in R_+^n$ in the matrix $\mathbf{D}$ denotes a dictionary atom. $\mathbf{X}=\left[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_N\right]\in R_+^{M\times N}$ is the representation coefficient matrix. Each column $\mathbf{x}_i \in R_+^M$ ($i = 1, 2, ..., N$) in the matrix $\mathbf{X}$ denotes the representation coefficients for the data vector $\mathbf{y}_i$. $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius-norm and l_1-norm of a matrix, respectively. λ is a balance parameter. $\mathbf{D} \geq \mathbf{0}$ and $\mathbf{X} \geq \mathbf{0}$ mean that all the elements in $\mathbf{D}$ and $\mathbf{X}$ are non-negative.

However, as what discussed in the previous Section 2, Eq. (1) just pays attentions to the representation capability of the dictionary, and ignores the correlations among the dictionary atoms. In other words, the dictionary $\mathbf{D}$ learned from Eq. (1) may have a large number of redundant atoms, which will decrease the fusion performance and computational efficiency of the proposed fusion method.

In [28], an orthogonal enforcement term was introduced to minimize the redundancy among the dictionary atoms during the non-negative matrix factorization. In [29], a concept of mutual incoherence was defined to measure the correlations across the dictionary atoms, and an orthogonal dictionary was learned for the traditional SR model in image restoration. Motivated by these works, we also add a simple yet effective penalty term in Eq. (1), as suggested in [28], to reduce the redundancy among the learned dictionary atoms. Accordingly, the proposed compact non-negative dictionary learning (CNDL) method for NNSR is mathematically formulated by

$$\left(\mathbf{D}, \mathbf{X}\right) = \arg\min_{\mathbf{D},\mathbf{X}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \|\mathbf{X}\|_1 + \lambda_2 \sum_{i \neq j}\left(\mathbf{d}_i^T \mathbf{d}_j\right)^2 \qquad s.t.\ \mathbf{D} \geq \mathbf{0}, \mathbf{X} \geq \mathbf{0} \ . \tag{2}$$

By minimizing the last penalty term in Eq. (2), the atoms in the dictionary $\mathbf{D}$ are enforced to be as orthogonal as possible. As a result of that, the redundancy among the atoms in the dictionary $\mathbf{D}$ is greatly reduced.

Eq. (2) can be solved by using an alternating way with two steps: sparse coding and dictionary updating. In the sparse coding step, $\mathbf{D}$ is assumed to be fixed. Then Eq. (2) becomes

$$\mathbf{X} = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \lambda_1 \|\mathbf{X}\|_1 \qquad s.t.\ \mathbf{X} \geq \mathbf{0} \ , \tag{3}$$

which is a convex optimization problem. Many methods can solve such problem. Here, we adopt the alternative direction multiplier method (ADMM) [30] because of its fast convergence rate. For that, Eq. (3) is first reformulated into Eq. (4) by introducing an auxiliary variable $\mathbf{Z}$ and then solved by minimizing the augmented Lagrangian function in Eq. (5).

$$\mathbf{X} = \arg\min_{\mathbf{X}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{Z}\|_F^2 + \lambda_1 \|\mathbf{X}\|_1 \qquad s.t.\ \mathbf{X}=\mathbf{Z}, \mathbf{X} \geq \mathbf{0} . \tag{4}$$

$$J(\mathbf{X},\mathbf{Z},\mathbf{V},\mu)=\frac{1}{2}\|\mathbf{Y}-\mathbf{DZ}\|_F^2+\lambda_1\|\mathbf{X}\|_1+\langle\mathbf{V},\mathbf{X}-\mathbf{Z}\rangle+\frac{\mu}{2}\|\mathbf{X}-\mathbf{Z}\|_F^2 \quad s.t.,\ \mathbf{X}\geq\mathbf{0}\ . \tag{5}$$

In Eq. (5), the Lagrange multiplier $\mathbf{V}$ and the penalty parameter μ are introduced to remove the equality constraint in Eq. (4). $\langle\cdot\rangle$ denotes the Euclidean inner product of two matrices.

Solving Eq. (5) consists of the following alternative iterations:

$$\begin{aligned}&\mathbf{Z}^{(t+1)}=\arg\min_{\mathbf{Z}} J\left(\mathbf{X}^{(t)},\mathbf{Z},\mathbf{V}^{(t)},\mu^{(t)}\right)\\ &\mathbf{X}^{(t+1)}=\arg\min_{\mathbf{X}} J\left(\mathbf{X},\mathbf{Z}^{(t+1)},\mathbf{V}^{(t)},\mu^{(t)}\right) \quad s.t.,\ \mathbf{X}\geq\mathbf{0}\end{aligned}\ , \tag{6}$$

where t is the iteration number. The two sub-optimization problems have the following closed-form solutions, i.e.,

$$\mathbf{Z}^{(t+1)}=(\mathbf{D}^T\mathbf{D}+\mu^{(t)}\mathbf{I})^{-1}(\mathbf{D}^T\mathbf{Y}+\mu^{(t)}\mathbf{X}^{(t)}+\mathbf{V}^{(t)})\ , \tag{7}$$

$$\mathbf{X}^{(t+1)}=\left[S_{\lambda_1/\mu^{(t)}}\left(\mathbf{Z}^{(t+1)}-\frac{\mathbf{V}^{(t)}}{\mu^{(t)}}\right)\right]_+\ , \tag{8}$$

where $[\mathbf{A}]_+=\max(\mathbf{A},0)$, and the threshold function $S_\tau(x)$ is defined as [31]

$$S_\tau(x)=\begin{cases}x-\tau, & \text{if } x>\tau\\ x+\tau, & \text{if } x<-\tau\\ 0, & \text{otherwise}\end{cases}\ . \tag{9}$$

In the dictionary updating step, $\mathbf{X}$ is assumed to be fixed, and the non-negative dictionary $\mathbf{D}$ is updated by

$$\mathbf{D}=\arg\min_{\mathbf{D}}\frac{1}{2}\|\mathbf{Y}-\mathbf{DX}\|_F^2+\lambda_2\sum_{i\neq j}\left(\mathbf{d}_i^T\mathbf{d}_j\right)^2 \quad s.t.\ \mathbf{D}\geq\mathbf{0}\,. \tag{10}$$

Similar to that in [26], the sub-optimization problem in Eq. (10) can be solved in an iterated way. In each iterate, $M-1$ dictionary atoms in the dictionary $\mathbf{D}$ are supposed to be fixed and only one atom $\mathbf{d}_m$ is updated, i.e.,

$$\mathbf{d}_m=\arg\min_{\mathbf{d}_m}\frac{1}{2}\left\|\mathbf{Y}-\sum_{i\neq m}\mathbf{d}_i\bar{\mathbf{x}}_i-\mathbf{d}_m\bar{\mathbf{x}}_m\right\|_F^2+\lambda_2\sum_{i\neq m}\left(\mathbf{d}_i^T\mathbf{d}_m\right)^2 \quad s.t.,\ \mathbf{d}_m\geq 0\ . \tag{11}$$

Here $\bar{\mathbf{x}}_m$ denotes the m-th row of the representation coefficient matrix $\mathbf{X}$. The sub-optimization in Eq. (11) has the following closed-solution

$$\mathbf{d}_m=\left[\left(\bar{\mathbf{x}}_m\left(\bar{\mathbf{x}}_m\right)^T\mathbf{I}_n+2\lambda_2\tilde{\mathbf{D}}_m\left(\tilde{\mathbf{D}}_m\right)^T\right)^{-1}\mathbf{E}_m\left(\bar{\mathbf{x}}_m\right)^T\right]_+\ , \tag{12}$$

where $\tilde{\mathbf{D}}_m = [\mathbf{d}_1, \cdots, \mathbf{d}_{m-1}, \mathbf{d}_{m+1}, \cdots, \mathbf{d}_M]$ and $\mathbf{E}_m = \mathbf{Y} - \sum_{i \neq m} \mathbf{d}_i \overline{\mathbf{x}}_i$. $\mathbf{I}_n$ is an identity matrix of size $n \times n$.

Algorithm 1 summarizes the optimization of the proposed CNDL method. As shown in Eq. (12), a non-negative constraint is employed during the updating of the dictionary atoms, which may force some atoms in the constructed dictionary $\mathbf{D}$ to be zero ones. Accordingly, these zero atoms should be removed from the constructed dictionary $\mathbf{D}$ in Algorithm 1.

Algorithm 1: Compact Non-negative Dictionary Learning (CNDL)

Input: Observed data $\mathbf{Y}$ and parameters λ_1 and λ_2

Initialization: $\mathbf{D}^0$, $\mu = 0.07$, $\rho = 1.25$ $\mu_{\max} = 10^{10}$, $\varepsilon = 0.005$, $\mathbf{X}^0 = \mathbf{B}^0 = \mathbf{0}$, $Oiter_{\max} = 1 \times 10^3$, $Iiter_{\max} = 100$

Outer Loop: $j = 1$

while not converged **do**

(1) Fix $\mathbf{D}$ and update $\mathbf{X}$:

Inner Loop: $t = 1$

while not converged **do**

(1.1) Fix $\mathbf{X}$ and update $\mathbf{Z}$ via Eq. (7);

(1.2) Fix $\mathbf{Z}$ and update $\mathbf{X}$ via Eq. (8);

(1.3) Update the multiplier $\mathbf{V}$: $\mathbf{V}^{(t+1)} = \mathbf{V}^{(t)} + \mu^{(t)}(\mathbf{X}^{(t+1)} - \mathbf{Z}^{(t+1)})$;

(1.4) Update μ : $\mu^{(t+1)} = \min(\rho\mu^{(t)}, \mu_{\max})$;

(1.5) Update t : $t = t + 1$;

(1.6) Check the convergence condition:

$t > Iiter_{\max}$, or $\left\|\mathbf{X}^{(t+1)} - \mathbf{X}^{(t)}\right\|_\infty < \varepsilon$, or $\|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F / \|\mathbf{Y}\|_F < \varepsilon$.

end while

(2) Fix $\mathbf{X}$ and update $\mathbf{D}$:

for $m = 1,2,\ldots,M$

Update $\mathbf{d}_m$ via Eq. (12);

end for

(3) Update j : $j = j + 1$;

(4) Check the convergence condition:

$j > Oiter_{\max}$ or $\left\|\mathbf{Y} - \mathbf{D}\mathbf{X}^{(t+1)}\right\|_F / \|\mathbf{Y}\|_F < \varepsilon$

end while

Output: Remove the zero columns in $\mathbf{D}$ and output the compact non-negative dictionary $\mathbf{D}$.

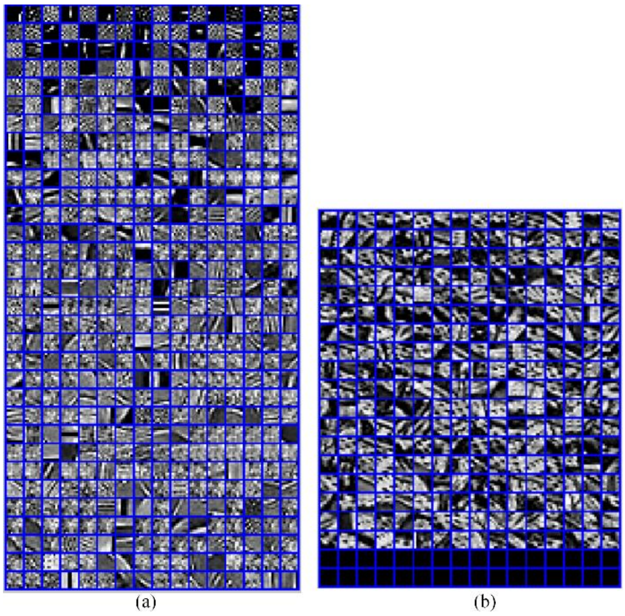

Fig. 2. Constructed dictionaries by using different methods. (a) Traditional dictionary learning method [27]; (b) Proposed CNDL.

Fig.2 illustrates the constructed dictionaries by using the traditional non-negative dictionary learning method [27] (Fig.2(a)) and the proposed CNDL method (Fig.2(b)). The initial numbers of atoms in the two dictionaries are both set to 512. As shown in Fig. 2, the finally constructed dictionary in Fig. 2 (a) still has 512 atoms, but the dictionary in Fig. 2(b) just consists of 288 atoms. This demonstrates that the dictionary constructed by using CNDL is more compact than the one constructed by using the traditional method. However, the compactness does not reduce and even improves the representation capability of the dictionary and the subsequent fusion performance of the fusion method, which will be verified in the latter experiment part (i.e., Section 5).

4. NNSR model with local consistency constraint and its application to multi-focus image fusion

In this section, we will first present a non-negative sparse representation model (CNNSR, for short) with a local consistency prior. Then we will discuss the proposed CNNSR-based fusion method in detail.

4.1 NNSR model with local consistency constraint

Given an over-complete non-negative dictionary $\mathbf{D} \in R_{+}^{n \times M}$, the traditional NNSR model can be computed by [11]

$$\mathbf{X} = \underset{\mathbf{X}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha \|\mathbf{X}\|_1 \qquad s.t.\ \mathbf{X} \geq \mathbf{0}, \tag{13}$$

where $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_N] \in R_+^{n \times N}$ denotes the observed data to be sparsely coded, i.e., the input image patches here. $\mathbf{y}_i \in R_+^n$ in the matrix $\mathbf{Y}$ denotes an input image patch. $\mathbf{X} \in R_+^{M \times N}$ is the representation coefficient matrix.

The traditional NNSR model may be directly adopt to fuse multi-focus images. However, as shown in Eq. (13), the image patches are independently coded by using NNSR without taking the local consistency among image patches into consideration, so that the representation coefficients for those spatial-adjacent image patches may look different even if these image patches have similar features. Subsequently, these image patches will be determined to have different focus information, which will introduce some obvious block artifacts to the fused image.

To address such problem, we present a new non-negative representation (CNNSR, for short) model by adding a Laplacian regularization term into the traditional NNSR model as

$$\mathbf{X} = \underset{\mathbf{X}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{Y} - \mathbf{D}\mathbf{X}\|_F^2 + \alpha_1 \|\mathbf{X}\|_1 + \alpha_2 \operatorname{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^T) \qquad s.t.\ \mathbf{X} \geq \mathbf{0}, \tag{14}$$

where α_1 and α_2 are two positive trade-off parameters. The regularization term $\operatorname{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^T)$ in Eq. (14) is defined by

$$\operatorname{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^T) = \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \omega_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 . \tag{15}$$

The weight $\omega_{i,j}$ indicates the similarity between two image patches and is simply defined by

$$\omega_{i,j} = \exp\left(-\frac{\|\mathbf{y}_i - \mathbf{y}_j\|_2^2}{2} \right). \tag{16}$$

The Laplacian matrix $\mathbf{L} \in R^{N \times N}$ is computed by $\mathbf{L} = \mathbf{\Gamma} - \mathbf{W}$, where the affinity matrix $\mathbf{W} \in R^{N \times N}$ and the diagonal matrix $\mathbf{\Gamma} \in R^{N \times N}$ are defined by $\mathbf{W}(i, j) = \omega_{i,j}$ and $\mathbf{\Gamma}(i,i) = \sum_j \omega_{i,j}$, respectively [1].

As shown in Eq. (16), a large value will be assigned to the weight $\omega_{i,j}$ if $\mathbf{y}_i$ and $\mathbf{y}_j$ have similar features. Accordingly, $\mathbf{y}_i$ and $\mathbf{y}_j$ will be enforced to have similar representation coefficients by minimizing Eq. (15). Subsequently, the two patches will be both determined to be in-focus (or out-focus) during the fusion.

4.2 Optimization of CNNSR model and its computational complexity

Eq. (14) can be efficiently solved by jointly adopting ADMM [30] and a modified Sparse Reconstruction by Separable Approximation (SpaRSA)-based method [32]. For that, an auxiliary variable $\mathbf{H}$ is first introduced to make the objective function in [13] separable, i.e.,

$$\mathbf{X} = \underset{\mathbf{X}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{Y}-\mathbf{DH}\|_F^2 + \alpha_1\|\mathbf{X}\|_1 + \alpha_2 \operatorname{tr}(\mathbf{XLX}^T) \qquad s.t.\ \mathbf{X}=\mathbf{H}, \mathbf{X}\geq \mathbf{0} \tag{17}$$

In order to remove the equality constraint in Eq. (17), a Lagrangian multiplier $\mathbf{S}$ is further introduced by

$$\begin{aligned} J(\mathbf{X},\mathbf{H},\mathbf{S},\eta) &= \frac{1}{2}\|\mathbf{Y}-\mathbf{DH}\|_F^2 + \alpha_1\|\mathbf{X}\|_1 + \alpha_2 \operatorname{tr}(\mathbf{XLX}^T) + \frac{\eta}{2}\|\mathbf{X}-\mathbf{H}\|_F^2 + \langle \mathbf{S}, \mathbf{X}-\mathbf{H}\rangle \\ &= \frac{1}{2}\|\mathbf{Y}-\mathbf{DH}\|_F^2 + \alpha_1\|\mathbf{X}\|_1 + \alpha_2 \operatorname{tr}(\mathbf{XLX}^T) + \frac{\eta}{2}\left\|\mathbf{X}-\mathbf{H}+\frac{\mathbf{S}}{\eta}\right\|_F^2 \\ & s.t.\ \mathbf{X}\geq \mathbf{0} \end{aligned}, \tag{18}$$

where η is a penalty parameter. Finally, the problem is minimized with respect to $\mathbf{X}$, $\mathbf{H}$ and $\mathbf{S}$, respectively, by fixing the others. The optimization of CNNSR is summarized in Algorithm 2. Appendix A provides more details.

Algorithm 2: Optimization of CNNSR

Input: Observed data $\mathbf{Y}$, over-compete dictionary $\mathbf{D}$, and parameters α_1 and α_2

Initialization: $\mathbf{X}^0=\mathbf{H}^0=\mathbf{0}$, $\eta=0.035$, $\rho=1.25$, $\eta_{\max}=10^{10}$, $\varepsilon=0.005$, $iter_{\max}=10^3$, $t=1$

while not converged **do**

(1) Fix $\mathbf{H}$ and update $\mathbf{X}$ via Eq. (A4);

(2) Fix $\mathbf{X}$ and update $\mathbf{H}$ via Eq. (A6);

(3) Update the multiplier $\mathbf{S}$: $\mathbf{S}^{(t+1)}=\mathbf{S}^{(t)}+\eta^{(t)}(\mathbf{X}^{(t+1)}-\mathbf{H}^{(t+1)})$;

(4) Update η : $\eta^{(t+1)}=\min(\rho\eta^{(t)},\eta_{\max})$;

(5) Update t : $t=t+1$;

---

(6) Check the convergence condition:

$$t > Iter_{\max}\ \text{, or } \left\| \mathbf{X}^{(t+1)} - \mathbf{X}^{(t)} \right\|_{\infty} < \varepsilon \text{ , or } \left\| \mathbf{Y} - \mathbf{D}\mathbf{X}^{(t+1)} \right\|_F / \left\| \mathbf{Y} \right\|_F < \varepsilon \ .$$

end while

Output: The representation coefficient matrix $\mathbf{X}$.

4.3 Proposed multi-focus image fusion method

In this subsection, we will present a multi-focus image fusion method based on CNNSR. Furthermore, we will employ a simple yet effective patch-level consistency rectification strategy to reduce the spatial block artifacts during the fusion process. By virtue of the proposed rectification strategy, each image patch and most of its spatial neighbors are simultaneously determined to be in-focus or out-focus. Moreover, because of the proposed rectification strategy, the input images may be divided into a set of non-overlapped patches, rather than a set of overlapped ones, in the proposed fusion method. This makes the proposed fusion method have high computational efficiency.

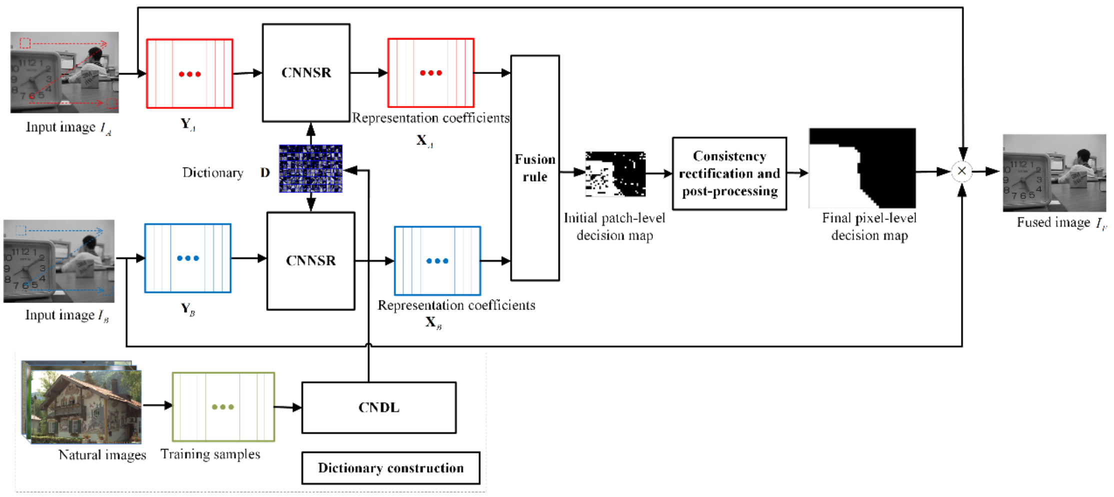

Fig. 3. Diagram of the proposed multi-focus image fusion method.

The diagram of the proposed multi-focus image fusion method is shown in Fig. 3. To simplify the discussion, we assume that the fused image is generated from a pair of well-registered images of size $N_1 \times N_2$, denoted by I_A and I_B, respectively. The proposed fusion method consists of the following steps.

(1). The input images I_A and I_B are divided into N non-overlapped patches of size $b_x \times b_y$ from left-top to right-bottom, respectively. Two sets of image patches $\{I_i^A \mid i = 1, 2, ..., N\}$ and $\{I_i^B \mid i = 1, 2, ..., N\}$ are then obtained. Here, $N = N_1' \times N_2'$, $N_1' = \left\lceil \frac{N_1 - b_x + 1}{b_x} \right\rceil$ and $N_2' = \left\lceil \frac{N_2 - b_y + 1}{b_y} \right\rceil$. $\lceil x \rceil$ denotes the smallest integer that is greater than or equal to x.

(2). Each image patch is transformed into a vector of dimension $n = b_x \times b_y$ via lexicographic ordering. Two data matrices $\mathbf{Y}_A = \left[\mathbf{y}_1^A, \mathbf{y}_2^A, ..., \mathbf{y}_N^A\right]$ and $\mathbf{Y}_B = \left[\mathbf{y}_1^B, \mathbf{y}_2^B, ..., \mathbf{y}_N^B\right]$ are then constructed for the two input images, respectively. $\mathbf{y}_i^A$ ($\mathbf{y}_i^B$) corresponds to the *i*-th image patch I_i^A (I_i^B) of image I_A (I_B).

(3). The two data matrices $\mathbf{Y}_A$ and $\mathbf{Y}_B$ are encoded via CNNSR. Their representation coefficient matrices $\mathbf{X}_A = \left[\mathbf{x}_1^A, \mathbf{x}_2^A ..., \mathbf{x}_N^A\right]$ and $\mathbf{X}_B = \left[\mathbf{x}_1^B, \mathbf{x}_2^B ..., \mathbf{x}_N^B\right]$ are, respectively, obtained by using Algorithm 2. Here, a compact non-negative dictionary $\mathbf{D} \in R_+^{n \times M}$ is learned in advance from a set of training images with high resolution by using Algorithm 1.

(4). A patch-level decision map (i.e., a matrix) $\mathbf{\Psi}_{patch}$ of size $N_1' \times N_2'$ is defined, whose elements $\mathbf{\Psi}_{patch}(p,q)$ are determined by

$$\mathbf{\Psi}_{patch}(p,q) = \begin{cases} 1, & \text{if } \left\|\mathbf{x}_i^A\right\|_2 \geq \left\|\mathbf{x}_i^B\right\|_2 \\ 0, & \text{otherwise} \end{cases}, \tag{19}$$

where the relationship between (p,q) and i is computed by

$$p = \left\lceil \frac{i}{N_1} \right\rceil, \quad q = i - p \times N_1 . \tag{20}$$

(5). A refined patch-level decision map $\mathbf{\Psi}'_{patch}$ is obtained by performing consistency rectification on $\mathbf{\Psi}_{patch}$, which is similar to that in [33]. However, each element in $\mathbf{\Psi}_{patch}$ represents an image patch rather than a pixel. Therefore, this step can be seen as a patch-level consistency rectification strategy. Mathematically, $\mathbf{\Psi}'_{patch}$ is computed by

$$\boldsymbol{\Psi}'_{patch}(p,q)=\begin{cases}1, & \text{if } C^{1}_{\Psi_{patch}}(p,q)\geq C^{0}_{\Psi_{patch}}(p,q)\\ 0, & \text{otherwise}\end{cases}, \tag{21}$$

where $C^{1}_{\Psi_{patch}}(p,q)$ and $C^{0}_{\Psi_{patch}}(p,q)$ denote the numbers of "1" and "0" in a region of size 3×3 centered the element (p,q) in the decision map $\boldsymbol{\Psi}_{patch}$, respectively. $C^{1}_{\Psi_{patch}}(p,q)\geq C^{0}_{\Psi_{patch}}(p,q)$ means that most patches around the current (p,q)-patch in image I_A are initially determined to be focused ones. Accordingly, the current (p,q)-patch in image I_A will also be seen as to be focused one, and vice versa. By using Eq. (21), each image patch and most of its spatial neighbors will be simultaneously determined to be focused regions or defocused regions.

(6). A pixel-level decision map $\boldsymbol{\Psi}_{pixel}$ of size $N_1\times N_2$ constructed by

$$\boldsymbol{\Psi}_{pixel}(x,y)=\boldsymbol{\Psi}'_{patch}(p,q),\quad \text{if } p=\left\lceil\frac{x}{b_x}\right\rceil \ \&\ q=\left\lceil\frac{y}{b_y}\right\rceil. \tag{22}$$

(7). The final pixel-level decision map $\boldsymbol{\Psi}^{Final}_{pixel}$ is obtained by performing some further post-processing on $\boldsymbol{\Psi}_{pixel}$. In spite of the validity of the proposed patch-level consistency rectification strategy in Eq. (21), some small regions may be still mistakenly marked. For that, a small region removal strategy as in in [1] is performed on $\boldsymbol{\Psi}_{pixel}$ to obtain the final pixel-level decision map $\boldsymbol{\Psi}^{Final}_{pixel}$ is obtained. Specifically, those connected regions in $\boldsymbol{\Psi}_{pixel}$ whose numbers of entries are less than 5% of the total number of pixels in the input images are first taken as isolated regions $\boldsymbol{\Omega}_{isolated}$. Then the element values within these isolated regions are re-assigned as 1 minus their original values, i.e.,

$$\boldsymbol{\Psi}^{Final}_{pixel}(x,y)=\begin{cases}1-\boldsymbol{\Psi}_{pixel}(x,y) & ,(x,y)\in\boldsymbol{\Omega}_{isolated}\\ \boldsymbol{\Psi}_{pixel}(x,y) & ,\text{otherwise}\end{cases}. \tag{23}$$

(8). The fused image I_F is finally constructed by using the decision map $\boldsymbol{\Psi}^{Final}_{pixel}$, i.e.,

$$I_F(x,y)=\boldsymbol{\Psi}^{Final}_{pixel}(x,y)I_A(x,y)+\left(1-\boldsymbol{\Psi}^{Final}_{pixel}(x,y)\right)I_B(x,y). \tag{24}$$

Fig.4 illustrates the validity of fusion strategies in the proposed method. As shown in Fig. 4(d), some isolated regions are in the decision map, while they are greatly reduced in Fig. 4(e) when the patch-level rectification consistency strategy is performed. As shown in Fig. 4(f), these isolated regions are further reduced by using the small region removal strategy, and the final decision map is more close to the ‘ideal’ one. Accordingly, some spatial artifacts will be greatly reduced in the fused image.

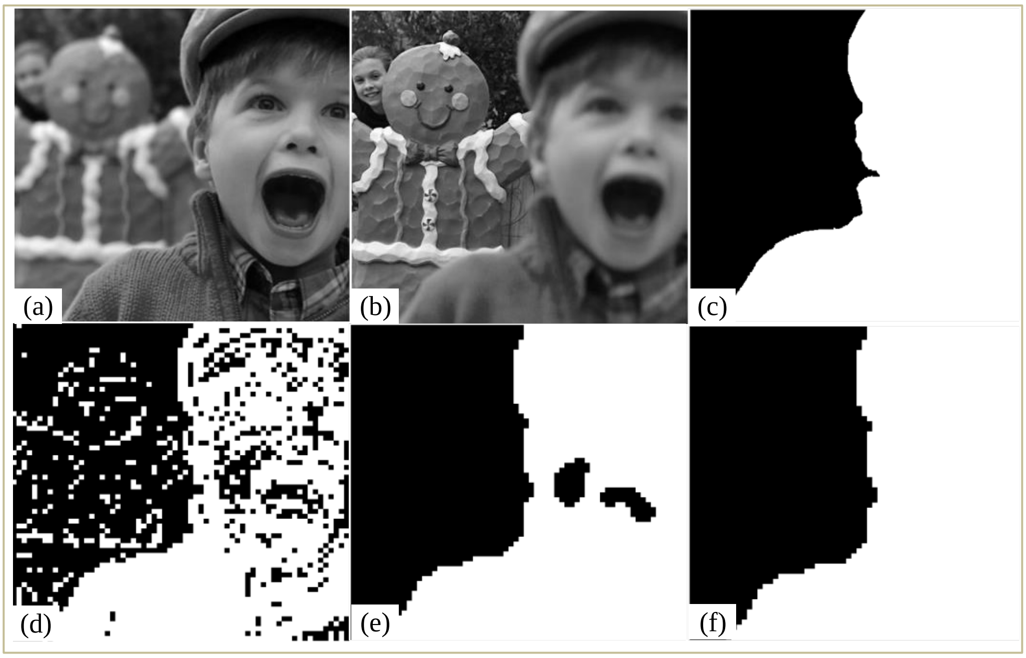

Fig. 4. Illustration of the validity of fusion strategies in the proposed method. (a) and (b) A pair of multi-focus images with focus on the right part and on the left part, respectively; (c) ‘Ideal’ decision map; (d) Decision map $\mathbf{\Psi}_{patch}$ without patch consistency rectification; (e) Decision map $\mathbf{\Psi}'_{patch}$ with patch consistency rectification; (f) Decision map $\mathbf{\Psi}_{pixel}^{Final}$ with small region removal. It should be noted that the patch-based decision maps in (d) and (e) have been transformed to pixel-based ones for better displaying.

It should be noted that the computational complexity of the proposed fusion method is mainly depended on the employed CNNSR model, whose major computation is the product of three matrices when updating $\mathbf{H}$ in Eq. (A6) and is about $O(nNM^2)$. Further considering the number of iterations r needed for convergence when encoding the input image patches, the proposed fusion method thus has a computational complexity of about $O(rnNM^2)$. As well, because of the non-overlapping division of input images in the proposed fusion method, N (i.e., the number of image patches) is much smaller than that in the traditional SR-based fusion method. For example, N is 1200 for an input image of size

240×320 in the proposed method. However, N is about 76800 for the same input image of size 240×320 in the traditional SR-based fusion methods. Moreover, as discussed in the previous Section 3, the compact non-negative dictionary employed by the proposed fusion method usually has a smaller number of dictionary atoms (e.g., $M=288$) than the traditional non-negative dictionary (e.g., $M=512$) under the same initial condition. These make the proposed fusion method have much high computational efficiency in the real applications, which will be verified in the following experimental parts.

5. Experiments and analysis

We perform several sets of experiments to validate the proposed fusion method in this section. First, we discuss the parameter settings for the proposed compact dictionary learning method (CNDL, for short) and the proposed fusion method; Secondly, we illustrate the effectiveness of the constructed compact non-negative dictionary as well as the proposed CNNSR model for multi-focus image fusion. Thirdly, we employ several pairs of multi-focus images to show the validity of the proposed fusion method. Finally, we extend our proposed method to multi-focus color image fusion. Before that, as suggested in [10], we also set the sizes of image patches to 8×8 in all of the following experiments for better fusion.

As well, some metrics are employed to evaluate different fusion methods subjectively, including mean square error (*MSE*), difference coefficient (*DC*), normalized mutual information (Q_{MI}) [34], gradient-based metric Q_G [35], structure similarity-based metric Q_Y [36] and human perception-based metric Q_{CB} [37].

The metrics *MSE* and *DC* reflect the errors between the fused image I_F and the 'ideal' fused image I_{IF}, and are computed by

$$MSE(I_F, I_{IF}) = \frac{1}{N_1\times N_2}\sum_{x,y}\left(I_F(x,y)-I_{IF}(x,y)\right)^2, \tag{25}$$

$$DC(I_F, I_{IF}) = \frac{1}{N_1 \times N_2} \sum_{x,y} \frac{|I_F(x,y) - I_{IF}(x,y)|}{I_{IF}(x,y)} . \quad (26)$$

Here, $N_1 \times N_2$ denotes the total number of pixels in the fused or 'ideal' fused image. $I_F(x,y)$ and $I_{IF}(x,y)$ are the intensity values of pixels at the position (x,y) in I_F and I_{IF}, respectively. Smaller *MSE* and *DC* values indicate better fusion performance and are more desirable.

The metrics Q_{MI}, Q_G, Q_Y and Q_{CB} evaluate the amount of different types of information that has been transferred from the input images to the fused image via a fusion method. Higher values of these metrics indicate better fusion performance and are more desirable.

Specifically, Q_{MI} measures the transferred information from source images I_A, I_B into the fused image I_F, and is defined by [34]

$$Q_{MI}(I_A, I_B, I_F) = 2\left[\frac{CE(I_A, I_F)}{E(I_A) + E(I_F)} + \frac{CE(I_B, I_F)}{E(I_B) + E(I_F)}\right], \quad (27)$$

where $CE(I_A, I_F)$ and $CE(I_B, I_F)$ denote the cross entropy between the source images and the fused image. $E(I_A)$, $E(I_B)$, and $E(I_F)$ denote the entropy of an image.

Q_G evaluates the amount of edge information that has been transferred from input images to the fused image and is computed by [35]

$$Q_G(I_A, I_B, I_F) = \frac{\sum_{(x,y)} \left(Q_G^{AF}(x,y)\omega_G^A(x,y) + Q_G^{BF}(x,y)\omega_G^B(x,y)\right)}{\sum_{(x,y)} \left(\omega_G^A(x,y) + \omega_G^B(x,y)\right)} . \quad (28)$$

Here, $Q_G^{AF}(x,y)$ and $Q_G^{BF}(x,y)$ are the edge information preservation values between the input images and the fused image. $\omega_G^A(x,y)$ and $\omega_G^B(x,y)$ are the edge strength-dependent weights for the input images.

Q_Y estimates how much information from the source images is preserved in the fused image and is computed by [36]

$$Q_Y(I_A, I_B, I_F) = \frac{1}{|W|} \sum_{w \in W} Q(I_A, I_B, I_F \mid w) \quad (29)$$

where $Q(I_A, I_B, I_F \mid w)$ denotes the quality measure in the local region w and is computed by

$$Q(I_A, I_B, I_F \mid w) = \begin{cases} \lambda(w) SSIM(I_A, I_F \mid w) + (1-\lambda(w)) SSIM(I_B, I_F \mid w) & , SSIM(I_A, I_B \mid w) \geq 0.75 \\ \max\{SSIM(I_A, I_F \mid w), SSIM(I_B, I_F \mid w)\} & , SSIM(I_A, I_B \mid w) < 0.75 \end{cases}. \quad (30)$$

Here, $SSIM(I_A, I_F \mid w)$ and $SSIM(I_B, I_F \mid w)$ are the structural similarities between the source images and the fused image under the local region w. $\lambda(w)$ is the local weight and W denotes the family of all sliding windows.

Finally, Q_{CB} is a perceptual quality measure based on contrast preservation calculation for image fusion, which is motivated by the process of human vision modeling and is computed by [37]

$$Q_{CB}(I_A, I_B, I_F) = \frac{1}{N_1 \times N_2} \sum_{(x,y)} \lambda_A(x,y) Q_{CB}^{AF}(x,y) + \lambda_B(x,y) Q_{CB}^{BF}(x,y), \quad (31)$$

where $Q_{CB}^{AF}(x,y)$ and $Q_{CB}^{BF}(x,y)$ calculate the contrast information preservation between the source images and the fused image on the spatial position (x,y). $\lambda_A(x,y)$ and $\lambda_B(x,y)$ are the contrast based weights for the input images. $N_1 \times N_2$ denotes the total number of pixels in the input or fused image. More details about these metrics are seen in [34], [35], [36], and [37], respectively.

5.1 Parameter settings

In this subsection, we will first discuss how to set the parameters λ_1 and λ_2 in Eq. (2) when constructing the dictionary. Then we will discuss how to set the parameters α_1 and α_2 in Eq. (17) for the proposed fusion method.

Fig. 5. Three natural images with high spatial resolution that are used to train the dictionary, which are downloaded from http://r0k.us/graphics/kodak. These images have been transformed from color images to gray-scale ones when constructing a dictionary for the fusion of gray-scale multi-focus images.

When constructing the dictionary, we first select three natural images with high spatial resolution, which are shown Fig. 5[1]. Then we divide the three images into a set of (more than 1000,000)) patches of size 8×8 and select those patches (about 20,000) with high local variance (larger than 0.05 in this paper) as the training samples. Finally, we construct two sets of dictionaries by using CNDL with the same initial number of atoms (i.e., 512). In the first set of dictionaries, λ_2 is set to the same value, i.e., $\lambda_2 = 10^{-4}$, and λ_1 is set to 0.0001, 0.001, 0.02, 0.025, 0.03, 0.035, 0.04, and 0.05, respectively. In the second set of dictionaries, λ_1 is set to the same value, i.e., $\lambda_1 = 0.04$, and λ_2 is set to 10^{-6}, 10^{-5}, 10^{-4}, 10^{-3} and 10^{-2}, respectively. Finally, we show the fusion performance of these dictionaries for the multi-focus input images in Fig. 6(a) and Fig. 6(b).

Fig. 6. A pair of multi-focus images that are used to test the impacts of different parameters on the fusion performance in the proposed dictionary learning method and the proposed fusion method. (a) Focus on the left part; (b) Focus on the right part; (c) 'Ideal' fused image.

Here, we employ the metrics *MSE* and *DC* to subjectively evaluate the fusion performance of these dictionaries. For that, the focused regions are manually selected from the input images in Fig.6(a) and Fig. 6(b) to construct the 'ideal' fused image in advance. Table 1 and Table 2 provide the fusion performance of the proposed method with the two sets of dictionaries mentioned above, respectively. Table 1 shows that the fusion performance achieves the best when λ_1 is within the range of $[0.03, 0.04]$.

---

[1] We also construct several dictionaries by using different numbers of training images and by using some training images with different visual qualities. We find that the quality of the training images seems more influential on the fusion performance of the proposed fusion method than the number of training images does. Mode details are seen in Supplementary materials.

Differently, Table 2 indicates that the proposed CNDL method is insensitive to the parameter λ_2 until it achieves 10^{-3}. In this paper, we set λ_1 and λ_2 to 0.04 and 10^{-4} in the proposed CNDL method, respectively.

Table 1. Fusion performance with the first set of dictionaries constructed by using different values of λ_1. The best scores are marked with bold in the table. As well, the final number of dictionary atoms *M* obtained by using different values of λ_1 are also provided in the table, which indicates that *M* obviously varies with λ_1.

Dictionary	$D_{\lambda_1=0.0001}$	$D_{\lambda_1=0.001}$	$D_{\lambda_1=0.02}$	$D_{\lambda_1=0.025}$	$D_{\lambda_1=0.03}$	$D_{\lambda_1=0.035}$	$D_{\lambda_1=0.04}$	$D_{\lambda_1=0.05}$
MSE	2.4988	2.3609	2.3960	2.3960	**2.3667**	**2.3667**	**2.3667**	2.3694
DC	0.0136	0.0128	0.0127	0.0127	**0.0126**	**0.0126**	**0.0126**	0.0126
M	512	510	486	432	392	339	288	266

Table 2. Fusion performance with the second set of dictionaries constructed by using different values of λ_2. The best scores are marked with bold in the table. Similarly, the final number of dictionary atoms *M* obtained by using different values of λ_2 are also provided in the table, which indicates that *M* keeps unchanged with λ_2.

Dictionary	$D_{\lambda_2=10^{-6}}$	$D_{\lambda_2=10^{-5}}$	$D_{\lambda_2=10^{-4}}$	$D_{\lambda_2=10^{-3}}$	$D_{\lambda_2=10^{-2}}$
MSE	**2.3667**	**2.3667**	**2.3667**	2.3667	2.3960
DC	**0.0126**	**0. 0126**	**0. 0126**	0. 0127	0. 0128
M	288	288	288	288	288

As discussed in the earlier Section 3, owing to the non-negativity and orthogonal constraints, the final number of dictionary atoms M will be smaller than the initial number of atoms (i.e., 512). Therefore, in addition to *MSE* and *DC*, the atom numbers of dictionaries constructed by using different parameters are also provided in Table 1 and Table 2, which demonstrate that λ_1 has a greater impact on the number of dictionary atoms than λ_2. The number of dictionary atoms increases with the decrease of

λ_1. As shown in Table 1 and Table 2, given the 512 initial dictionary atoms, the constructed dictionary with $\lambda_1 = 0.04$ and $\lambda_2 = 10^{-4}$ finally ends up with 288 atoms in this paper. And the dictionary, denoted by D_{288}, will be employed in the following experiments.

Similarly, parameters α_1 and α_2 in Eq. (17) are also set according to the fusion performance (i.e., *MSE* and *DC* values) of the proposed fusion method on the input images in Fig. 6(a) and Fig. 6(b). The fusion performance is shown to remain nearly unchanged when α_1 and α_2 are both in the range of $[10^{-9}, 10^{-4}]$. However, the fusion performance is shown to reduce greatly when α_1 or α_2 is larger than 10^{-4}. In the following experiments, α_1 and α_2 are both set to 10^{-6}.

5.2 Validity of the constructed dictionary and the proposed CNNSR model

Here, we will first illustrate the superiority of the compact non-negative dictionary D_{288} constructed by using CNDL over some dictionaries with 512 atoms, including a dictionary D_{512}^{DCT} with fixed cosine basis, a non-negative dictionary D_{512}^{Global} globally learned from a set of natural images by using the method in [27] and a non-negative dictionary $D_{512}^{Adaptive}$ adaptively learned from the input images by using the method in [38]. The superiority of CNNSR over NNSR [11] is also illustrated in this subsection.

For that, four fusion methods (CNNSR_ D_{512}^{DCT}, CNNSR_ D_{512}^{Global}, CNNSR_ $D_{512}^{Adaptive}$, and CNNSR_ D_{288}, for short, respectively) with the same CNNSR model but different dictionaries are first performed on the input images in Fig. 7(a) and Fig. 7(b). Then a fusion method (NNSR_ D_{288}, for short) with the traditional NNSR model and the dictionary D_{288} is also performed on the input images in Fig. 7(a) and Fig. 7(b). For simplification, the input images are divided by a non-overlapping way and a simple l_2-norm of representation coefficients based ‘maximum-selecting’ fusion rule [10] is employed in these fusion methods. Finally, the proposed fusion method (CNNSR_Pro, for short) is performed on the same pairs of input images, where the fusion rules described in Section 4.3 are employed.

Here, the four metrics Q_{MI}, Q_G, Q_Y and Q_{CB} are employed to evaluate these fusion methods subjectively, which are provided in Table 3. In addition, the computing time T of different methods are also provided in Table 3. From Table 3, it can be easily found that the fusion methods with those dictionaries learned from the natural images or input images significantly outperform the fusion method with the dictionary of fixed basis. Moreover, CNNSR_ D_{288} performs better than CNNSR_ D_{512}^{Global} and CNNSR_ $D_{512}^{Adaptive}$ do, although D_{288} has smaller number of atoms than D_{512}^{Global} and $D_{512}^{Adaptive}$. This indicates that the compactness of the constructed dictionary does not reduce the representation capability nor the subsequent fusion performance of a fusion method. In addition, as shown in Table 3, the compactness also makes CNNSR_ D_{288} have higher computational efficiency than CNNSR_ D_{512}^{Global} and CNNSR_ $D_{512}^{Adaptive}$.

Table 3. Fusion performance obtained by different sparse representation models and dictionaries. The best and second scores obtained by different methods are marked by red and blue colors with bold in the table, respectively.

Method	Q_{MI}	Q_G	Q_Y	Q_{CB}	T (in Seconds)
CNNSR_ D_{512}^{DCT}	1.1930	0.6832	0.9471	0.7023	4.0099
CNNSR_ D_{512}^{Global}	1.2075	0.7553	0.9675	0.7398	3.2053
CNNSR_ $D_{512}^{Adaptive}$	1.2073	0.7562	0.9681	0.7394	4.2856
CNNSR_ D_{288}	**1.2122**	**0.7564**	**0.9717**	**0.7447**	**2.2693**
NNSR_ D_{288}	1.1976	0.7539	0.9584	0.7304	2.9140
CNNSR_Pro	**1.2217**	**0.7608**	**0.9834**	**0.7548**	**2.0344**

From the experimental data in Table 3, it can also be found that CNNSR_ D_{288} significantly outperforms NNSR_ D_{288}. This demonstrates the superiority of CNNSR over NNSR when applied to the fusion of multi-focus images. The comparison between the performance obtained by CNNSR_ D_{288} and CNNSR_Pro further demonstrates the superiority of the fusion rules in our proposed fusion method.

In order to better demonstrate the validity of our proposed CNNSR model and fusion rules, the decision maps and fused images on Fig. 7(a) and Fig. 7(b) obtained by NNSR_ D_{288}, CNNSR_ D_{288} and CNNSR_Pro are illustrated in Fig. 7. By comprising Fig. 7(c) and Fig. 7(d), it can be easily found that the isolated patches in the decision map obtained by using CNNSR_ D_{288} are much fewer than those in the decision map obtained by using NNSR_ D_{288}. This demonstrates the superiority of the proposed CCNSR model over the traditional NNSR model in the reduction of spatial artifacts again. The isolated patches are further reduced and even eliminated by using CNNSR_Pro, as shown in Fig. 7(e). This owes to the fusion rules employed in CNNSR_Pro.

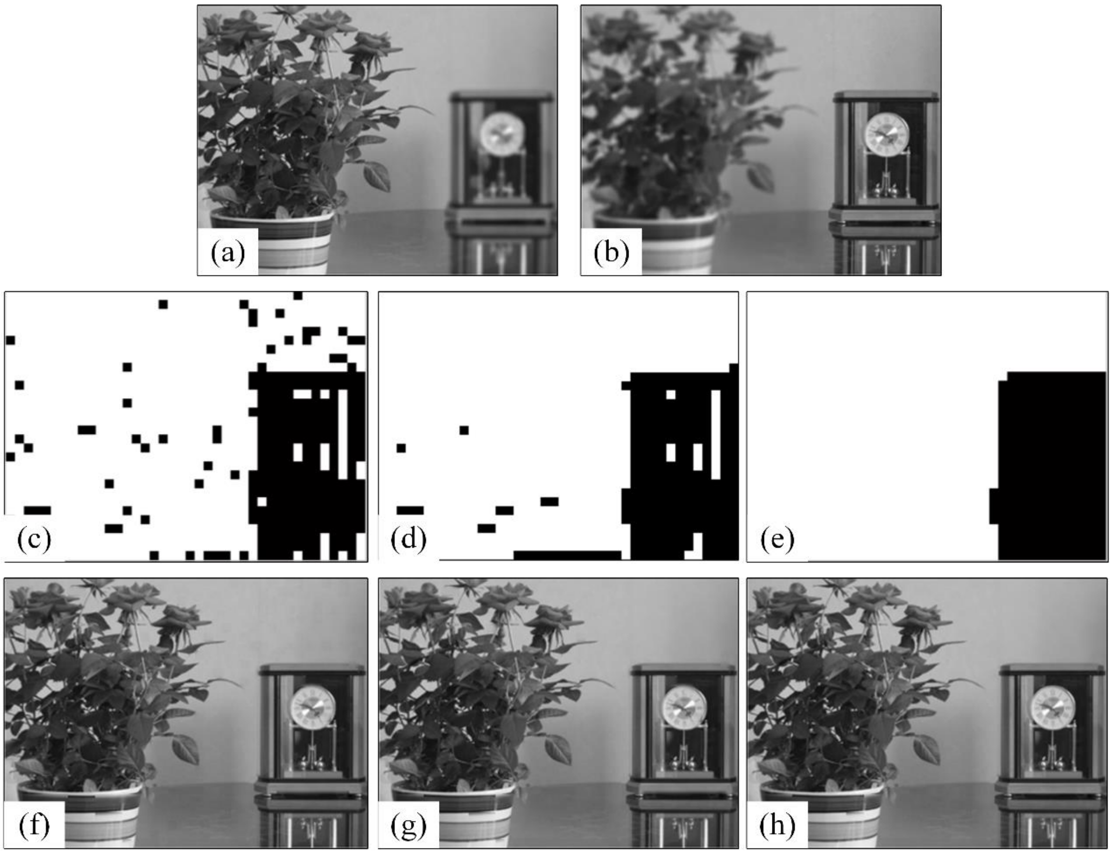

Fig. 7. Illustration of the validity of the proposed CNNSR model and fusion rules. (a) and (b) A pair of multi-focus images with the focus on the left part and the right part, respectively; (c), (d) and (e) The decision maps obtained by using NNSR_ D_{288},

CNNSR_ D_{288} and CNNSR_Pro, respectively; (f), (g) and (h) The fused images obtained by using NNSR_ D_{288} , CNNSR_ D_{288} and CNNSR_Pro, respectively.

5.3 Validity of the proposed fusion method

In order to thoroughly demonstrate the validity of the proposed fusion method, the multi-focus images, mentioned in Fig. 6 and Fig. 7 previously, and another several pairs of multi-focus images are employed in this subsection. These images are shown in Fig. 8[2]. In addition to the proposed fusion method (CNNSR_Pro, for short), some more fusion methods, including DSIFT [20], MF [39], DCNN [22], SR [4], MRSR [13], RSR_LR [1] and SRCF [2], are performed on these input images for comparisons. Specifically, DCNN is a deep convolutional neural network based fusion method.

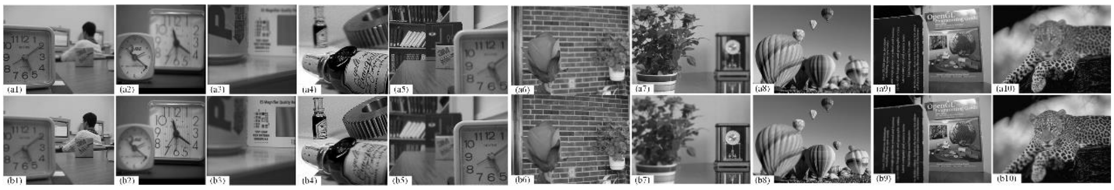

Fig. 8. 10 pairs of multi-focus input images. The input images in the top row focus on the left parts, and the corresponding input images in the bottom row focus on the right parts.

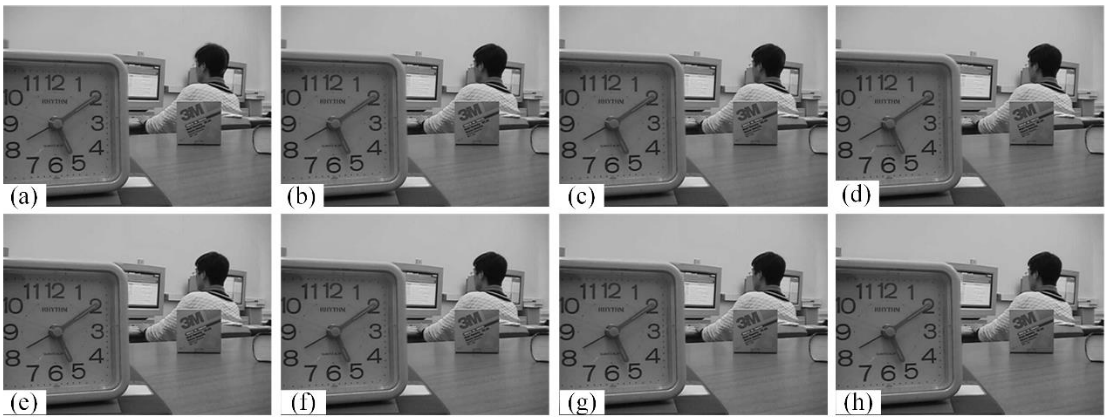

Fig. 9. Fusion images of Fig. 8(a1) and (b1) obtained by different fusion methods. (a) DSIFT; (b) MF; (c) SR; (d) MRSR; (e) RSR_LR; (f) DCNN; (g) SRCF; (h) CNNSR_Pro.

[2] These images are downloaded from http://home.ustc.edu.cn/~liuyu1. For better displaying, the input images in Fig. 6 and Fig. 7 are also shown in Fig. 8.

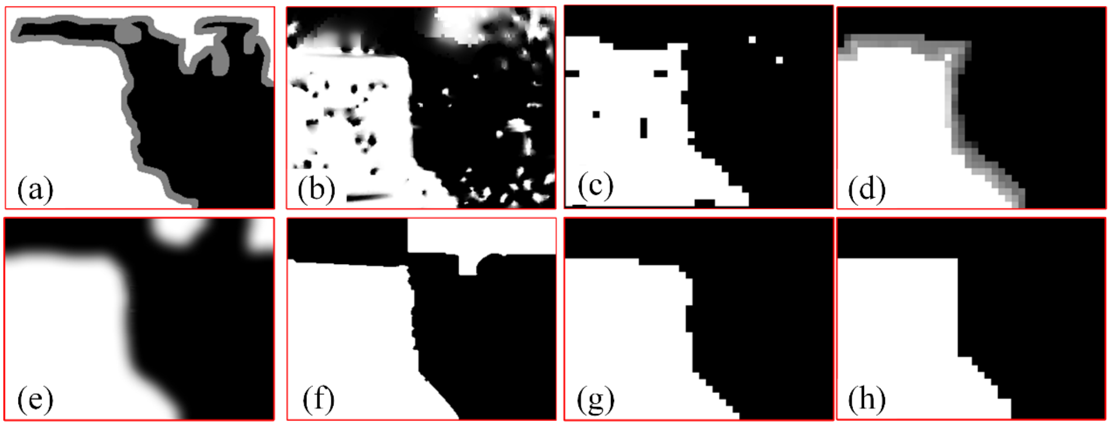

Fig. 10. Decision maps for the input images in Fig. 8(a1) and Fig. (b1) obtained by different fusion methods. (a) DSIFT; (b) MF; (c) MRSR; (d) RSR_LR; (e) DCNN; (f) SRCF; (g) CNNSR_Pro; (h) 'Ideal'. The 'white' ('black') regions in these decision maps denote that these regions in the fused images are directly selected from the input image in Fig. 8(a1) (Fig. 8(b1)), and the 'gray' regions denote that the regions in the fused images are the weighted average of the input images in Fig.8(a1) and Fig. 8(b1).

The fused images of Fig. 8(a1) and Fig. 8(b1) obtained by using different methods are illustrated in Fig. 9[3]. The decision maps obtained by different fusion methods are also provided in Fig. 10[4] for better visual comparisons. All of these methods mentioned here are shown to perform well for Fig. 8(a1) and (b1) from the fused images in Fig. 9. However, a more careful observation on Fig. 10 indicates that CNNSR_Pro performs the best among these fusion methods. It can be easily found that the decision map in Fig. 10(g) obtained by CNNSR_Pro is the closest to the 'ideal' one in Fig. 10(h). As shown in the right-top parts in Fig. 10 (a), (b), (e) and (f), some regions have been mistakenly determined to be in-focus. Owing to the use of spatial contextual information in MRSR, RSR_LR and CNNSR_Pro, those mistakenly determined regions are greatly reduced. Especially, there are few isolated patches in the decision maps obtained by using RSR_LR and CNNSR_Pro.

The quantitative results of different fusion methods in Table 4 coincide with the visual results

[3] The visual results of different fusion methods on the rest of input images in Fig. 8 are provided in Supplementary materials.
[4] Owing to the over-lapping division of input images, the decision map could not be obtained by using the SR fusion method. Therefore, in Fig. 10, we don't provide the decision map obtained by SR.

mentioned above, which also demonstrates that CNNSR_Pro performs the best, compared to the fusion methods mentioned here. Table 4 also indicates that CNNSR_Pro has high computational efficiency. The average computational time T of CNNSR_Pro is about half that of RSR_LR and SRCF, and is about one twentieth that of MRSR and DCNN.

Table 4. Performance of different methods on Fig. 8. Scores for the 10 pairs of input images in Fig.8 are averaged. The best and second scores obtained by different methods are marked by red and blue colors with bold in the table, respectively.

Method	Q_{MI}	Q_G	Q_Y	Q_{CB}	T (in Seconds)
DSIFT	1.2636	0.8077	0.9712	0.8133	**1.0650**
MF	1.2522	0.8074	0.9639	0.8009	1.6771
SR	1.1846	0.8052	0.9453	0.7863	16.5240
MRSR	1.2646	0.7964	0.9759	0.8096	29.0351
RSR_LR	1.2569	**0.8092**	0.9778	0.8152	3.5394
DCNN	1.2584	**0.8090**	0.9772	**0.8167**	40.7280
SRCF	**1.2923**	0.8076	**0.9800**	0.8149	3.0519
CNNSR_Pro	**1.2985**	0.8079	**0.9815**	**0.8223**	**1.6466**

5.4 Fusion of multi-focus color images

The proposed method can also be extended to the fusion of multi-focus color images. Similar to that in [2], the intensity component of input images is first obtained by simply averaging their Red (R), Green (G), and Blue (B) channels, respectively. Then a focus decision map is obtained by performing the proposed CNNSR_Pro method on the intensity component of input images. By using the decision map, the R, G, and B channels of the fused image are obtained, respectively, and the finally fused color image is constructed.

To demonstrate the validity of CNNSR_Pro on the fusion of multi-focus color images, a set of multi-focus color images are employed here, which are shown in Fig. 11[5]. In addition to the proposed CNNSR_Pro method, some fusion methods, including IMF [19], GFF [40], MWG [41], RSR_LR [1], DCNN [22] and SRCF [2], are performed on these images for comparisons.

Fig. 12[6] illustrates the fusion results of different methods on the input images in Fig. 11(a1) and Fig. 11(b1). Table 5 provides the averaging scores of different fusion methods on the 20 pairs of input images. The visual fusion results and the quantitative data in Table 5 indicate that the proposed CNNSR_Pro performs competitively with SRCF, DCNN and better than the other methods on the multi-focus color images in Fig. 11. Although CNNSR_Pro performs competitively with SRCF and DCNN, it has much higher computational efficiency than SRCF and DCNN. The average computational time T of CNNSR_Pro is about one seventh that of SRCF and DCNN for the test images in Fig. 11.

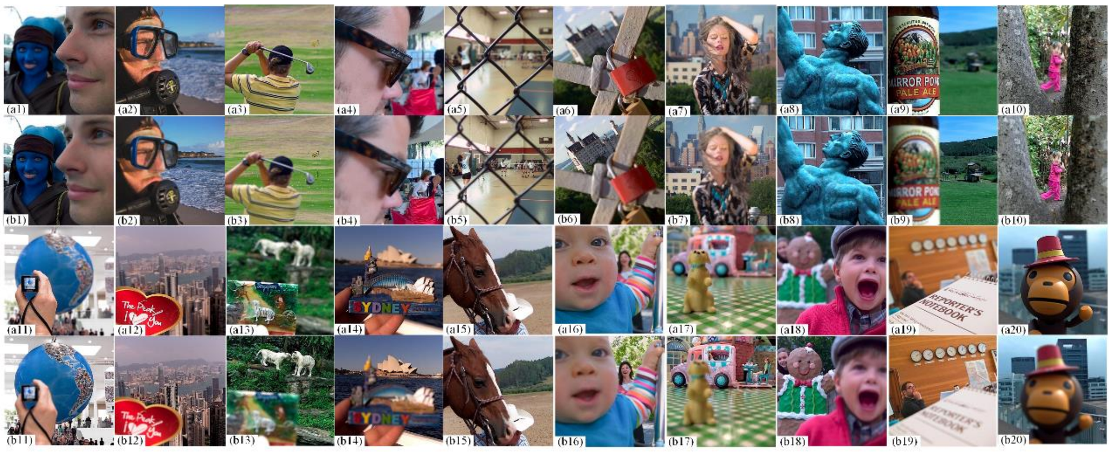

Fig. 11. 20 pairs of multi-focus color images. The first top row contains the first 10 input images with the focus on the front part, and the second row contains the corresponding input images with the focus on the back part. The third row contains the remaining 10 input images with the focus on the front part, and the bottom row contains the corresponding input images with the focus on the back part.

---

[5] These images are downloaded from https://www.researchgate.net/publication/291522937_Lytro_Multi-focus_Image_Dataset.
[6] The visual results of different fusion methods on the rest of input images in Fig. 11 are provided in Supplementary files.

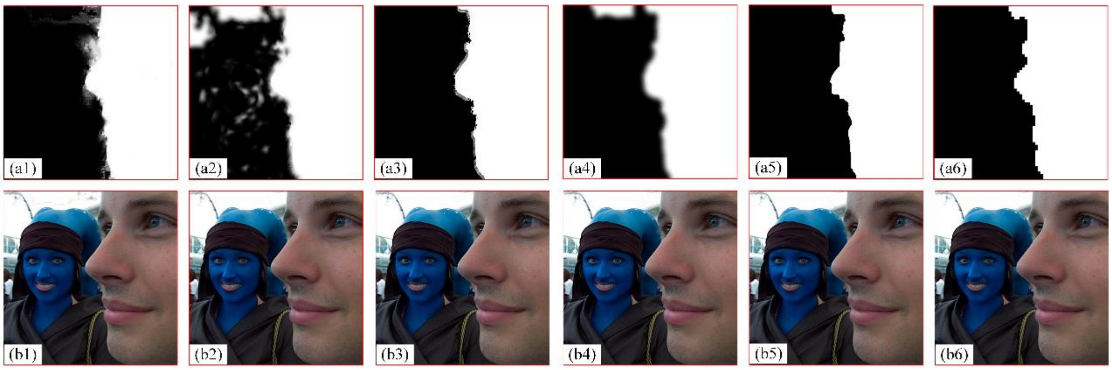

Fig. 12. Illustration of the fused results of different methods on Fig. 11(a1) and (b1). (a1)~(a6) Decision maps obtained by using IFM, GFF, MWG, DCNN, SRCF, and CNNSR_Pro, respectively. (b1)~(b6) Fused images obtained by using IFM, GFF, MWG, DCNN, SRCF, and CNNSR_Pro, respectively.

Table 5. Performance of different methods on Fig. 11. Scores for the 20 pairs of input images in Fig.11 are averaged. The best and second scores obtained by different methods are marked by red and blue colors with bold in the table, respectively.

Methods	Q_{MI}	Q_G	Q_Y	Q_{CB}	T (in Seconds)
IFM	1.1334	0.7845	0.9688	0.7861	**2.3747**
GFF	1.0980	0.7918	0.9821	0.7975	**0.5500**
MWG	1.1278	0.7819	0.9873	0.7974	7.5662
DCNN	1.1512	**0.7921**	0.9877	0.8084	167.0715
SRCF	**1.1929**	**0.7925**	**0.9890**	**0.8093**	132.0764
CNNSR_Pro	**1.1918**	0.7878	**0.9891**	**0.8103**	20.8199

6. Conclusions

We presented a non-negative sparse representation based multi-focus image fusion method, where the strong correlations among spatially adjacent image patches are fully considered. For that, we first construct a new NNSR model with a consistency constraint (CNNSR) on the representation coefficients for the fusion method. Then we present a patch-level consistency rectification strategy during the fusion process. The CNNSR model and patch-level consistency rectification make the proposed fusion method introduce very few spatial artifacts into the fused image. Moreover, owing to the patch-level consistency rectification, the input images may be divided into a set of non-overlapped patches, rather than a set of overlapped ones. This also makes the proposed fusion method have much computational efficiency in real applications. Additionally, we have constructed a compact non-negative dictionary for the CNNSR model. This further improves the fusion performance and the computational efficiency of the proposed fusion method to some extent. Finally, the proposed fusion method can be extended to the fusion of color images by some simple modifications. The proposed fusion method is experimentally shown to outperform some advanced SR-based fusion methods, such as MRSR, RSR_LR and SRCF. As well, it has the highest computational efficiency among these SR-based fusion methods.

Finally, it should be noted that the proposed fusion strategy is implemented in a patch-level way - the pixels in one patch will be determined to be all in-focus or all out-of-focus -. This is reasonable for most image patches. However, for those patches near the boundaries between the focused and defocused regions, the pixels in the same patch may belong to different classes, i.e., some pixels may be in-focus and some pixels may be out-of-focus. This is an inherent problem in the patch-based fusion methods. How to address such problem is of interest. We leave this for our future work.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 61773301, and by the Fundamental Research Funds for the Central Universities under Grant No. JBZ170401.

Appendix A

Appendix A details the description of the update scheme for solving Eq. (18) in the body.

(1) Update $\mathbf{X}$:

$$\begin{aligned} \mathbf{X}^{(t+1)} &= \arg\min_{\mathbf{X}} \alpha_1 \|\mathbf{X}\|_1 + \alpha_2 \mathrm{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^T) + \frac{\eta^{(t)}}{2}\left\| \mathbf{X} - \mathbf{H} + \frac{\mathbf{S}^{(t)}}{\eta^{(t)}} \right\|_F^2 \\ &= \arg\min_{\mathbf{X}} \alpha_1 \|\mathbf{X}\|_1 + Q(\mathbf{X}) \\ & s.t. \quad \mathbf{X} \geq \mathbf{0} \end{aligned} , \quad \text{(A1)}$$

where $Q(\mathbf{X}) = \alpha_2 \mathrm{tr}(\mathbf{X}\mathbf{L}\mathbf{X}^T) + \frac{\eta^{(t)}}{2}\left\| \mathbf{X} - \mathbf{H} + \frac{\mathbf{S}^{(t)}}{\eta^{(t)}} \right\|_F^2$. The problem can be solved in an iterated way by using the modified SpaRSA-based method [32], i.e,

$$\begin{aligned} \mathbf{X}^{(t+1)} &= \arg\min_{\mathbf{X}} \alpha_1 \|\mathbf{X}\|_1 + \frac{\gamma^{(t)}}{2}\left\| \mathbf{X} - \mathbf{X}^{(t)} \right\|_F^2 + \left\langle \nabla_{\mathbf{X}} Q(\mathbf{X}^{(t)}), \mathbf{X} - \mathbf{X}^{(t)} \right\rangle \\ &= \arg\min_{\mathbf{X}} \alpha_1 \|\mathbf{X}\|_1 + \frac{\gamma^{(t)}}{2}\left\| \mathbf{X} - \mathbf{X}^{(t)} + \frac{\nabla_{\mathbf{X}} Q(\mathbf{X}^{(t)})}{\gamma^{(t)}} \right\|_F^2 \\ & s.t. \quad \mathbf{X} \geq \mathbf{0} \end{aligned} , \quad \text{(A2)}$$

where $\gamma^{(t)} = 1.02\left(2\alpha_2 \|\mathbf{L}\|_F^2 + \eta^{(t)}\right)$ [42]. $\nabla_{\mathbf{X}} Q(\mathbf{X}^{(t)})$ is computed by:

$$\nabla_{\mathbf{X}} Q(\mathbf{X}^{(t)}) = 2\alpha_2 \mathbf{X}^{(t)}\mathbf{L} + \eta^{(t)}\left(\mathbf{X} - \mathbf{H}^{(t+1)} + \frac{\mathbf{S}^{(t)}}{\eta^{(t)}} \right). \quad \text{(A3)}$$

Eq. (A2) thus has the following solution [31]:

$$\mathbf{X}^{(t+1)} = \left[S_{\alpha_1/\gamma^{(t)}}\left(\mathbf{X}^{(t)} - \frac{\nabla_{\mathbf{X}} Q(\mathbf{X}^{(t)})}{\gamma^{(t)}} \right) \right]_+ . \quad \text{(A4)}$$

(2) Update $\mathbf{H}$:

$$\mathbf{H}^{(t+1)} = \arg\min_{\mathbf{H}} \frac{1}{2}\|\mathbf{Y} - \mathbf{D}\mathbf{H}\|_F^2 + \frac{\eta^{(t)}}{2}\left\| \mathbf{X}^{(t)} - \mathbf{H} + \frac{\mathbf{S}^{(t)}}{\eta^{(t)}} \right\|_F^2 . \quad \text{(A5)}$$

Its solution is computed by:

$$\mathbf{H}^{(t+1)} = \left(\mathbf{D}^T\mathbf{D} + \eta^{(t)}\mathbf{I}_M \right)^{-1}\left(\mathbf{D}^T\mathbf{Y} + \eta^{(t)}\mathbf{X}^{(t)} + \mathbf{S}^{(t)} \right) . \quad \text{(A6)}$$

Here, $\mathbf{I}_M$ is an identity matrix of size $M \times M$.

References

[1]. Q. Zhang, T. Shi, et al., "Roust sparse representation based multi-focus image fusion with dictionary construction and local spatial consistency," Pattern recognition 83 (2018) 299-313.

[2]. M. Nejati, S. Samavi, S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," Information

Fusion 25 (2015) 72-84.

[3]. J. Wright, A. Y. Yang, et al., “Robust face recognition via sparse representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2) (2009) 210-227.

[4]. B. Yang, S. Li, “Multifocus image fusion and restoration with sparse representation,” IEEE Transactions on Instrumentation and Measurement 59 (4) (2010) 884-892.

[5]. M. Kim, D. K. Han, H. Ko, “Joint patch clustering-based dictionary learning for multimodal image fusion,” Information Fusion 27(2016) 198-214.

[6]. F. Xiang, J. Zhang, et al., “Robust image fusion with block sparse representation and online dictionary learning,” IET Image Processing 12(3) (2108) 345-353.

[7]. Y. Yang, M. Ding, et al., “Multi-focus image fusion via clustering PCA based joint dictionary learning,” IEEE Access 5(2017) 16985-16997.

[8]. F. Fakhari, M. R. Mosavi, M. M. Lajvardi, “Image fusion based on multi-scale transform and sparse representation: an image energy approach,” IET Image Processing 11(11) (2017) 1041-1049.

[9]. B. Zhang, X. Lu, et al., “Multi-focus image fusion based on sparse decomposition and background detection,” Digital Signal Processing 58(2016) 50-63.

[10]. Q. Zhang, Y. Liu, et al., “Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: A review”, Information Fusion 40 (2018) 57-75.

[11]. J. Wang, J. Peng, et al., “Fusion method for infrared and visible images by using non-negative sparse representation,” Infrared Physics & Technology 67 (2014) 477-489.

[12]. Y. Liu, X. Chen, et al., “Image fusion with convolutional sparse representation,” IEEE Signal Processing Letters 23 (12) (2016) 1882-1886.

[13]. Q. Zhang, M. Levine, “Robust multi-focus image fusion using multi-task sparse representation and spatial context,” IEEE

Transactions on Image Processing 25 (5) (2016) 2045-2058.

[14]. S. Li, X. Kang, et al., “Pixel-level image fusion: a survey of the state of the art,” Information Fusin 33(2017) 100-112.

[15]. P. Hill, M. E. Al-Mualla, D. Bull, “Perceptual image fusion using wavelets,” IEEE Transactions on Image Processing 26 (3) (2017) 1076-1088.

[16]. H. Li, L. Liu, et al., “An improved fusion algorithm for infrared and visible images based on multi-scale transform,” Infrared Physics & Technology 74 (2016) 28-37.

[17]. H. Zhao, Z. Shang, et al., “Multi-focus image fusion based on the neighbor distance,” Pattern Recognition 46 (2013) 1002-1011.

[18]. Y. Chen, J. Guan, W. Cham, “Robust multi-focus image fusion using edge model and multi-matting,” IEEE Transactions on Image Processing 27 (3) (2018)1526-1541.

[19]. S. Li, X. Kang, B. Yang, “Image matting for fusion of multi-focus images in dynamic scenes,” Information Fusion, 14 (2) (2013) 147-162.

[20]. Y. Liu, S. Liu, Z. Wang, “Multi-focus image fusion with dense SIFT,” Information Fusion 23 (2015) 139-155.

[21]. C. Du, S. Gao, “Multi-focus image fusion with all convolutional neural network,” Optoelectronics Letters 14 (1) (2018) 0071-0075.

[22]. Y. Liu, X. Chen, et al., “Multi-focus image fusion with a deep convolutional neural network,” Information Fusion 36 (2017) 191-207.

[23]. Y. Yao, P. Guo, et al., “Image fusion by hierarchical joint sparse representation,” Cognitive Computation 6 (3) (2014) 281-292.

[24]. F. Yin, W. Gao, Z. Song, “Image fusion based on group sparse representation,” In: 8th International Conference on Digital Image Processing 2016, doi: 10.1117/12.2244879.

[25]. Y. Liu, Z. Wang, “Simultaneous image fusion and denoising with adaptive sparse representation,” IET Image Processing 9(5)

(2015) 347-357.

[26]. M. Aharon, M. Elad, A. Bruckstein, “K-SVD: an algorithm for designing overcomplete dictionaries for spare representation,” IEEE Transactions on Signal Processing 54 (11) (2006) 4311-4322.

[27]. W. Dong, F. Fu, et al., “Hyperspectral image super-resolution via non-negative structured sparse representation,” IEEE Transactions on Image Processing 25 (5) (2016) 2337-2352.

[28]. N. Guan, D. Tao , et al., “Manifold regularized discriminative nonnegative matrix factorization with fast gradient descent,” IEEE Transactions on Image Processing 20 (7) (2011) 2030-2048.

[29]. C. Bao, J. Cai, H. Ji, “Fast sparsity-based orthogonal dictionary learning for image restoration,” In: International Conference on Computer Vision (2013) 3384-3391.

[30]. S. Boyd, N. Parikh, et al. “Distributed optimization and statistical learning via the alternating direction method of multipliers,” Foundations and Trends in Machine Learning 3 (1) (2010) 1-122.

[31]. J. F. Cai, E. J. Candes, Z. Shen, “A singular value thresholding algorithm for matrix completion,” Siam Journal on Optimization 20 (4) (2010) 1956-1982.

[32]. S. J. Wright, R. D. Nowak, et al., “Sparse reconstruction by separable approximation,” IEEE Transactions on Signal Processing 57 (7) (2009) 3373-3376.

[33]. Z. Zhang, Rick S. Blum, “A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application,” Proceedings of the IEEE 87 (8) (1999) 1315-1326.

[34]. M. Hossny, S. Nahavandi, D. Creighton, “Comments on information measure for performance of image fusion,” Electronics Letters 44 (18) (2008) 1066-1067

[35]. V. Xydeas, V. Petrovic, “Objective image fusion performance measure,” Electronic Letters 36 (4) (2000) 308-309.

[36]. C. Yang, J. Zhang, X. Wang, et al., “A novel similarity based quality metric for image fusion,” Information Fusion 9 (2) (2008) 156-160.

[37]. Y. Chen, R. Blum, “A new automated quality assessment algorithm for image fusion,” Image and Vision Computing 27 (10) (2009) 1421-1432.

[38]. J. Mairal, F. Bach, J. Ponce, G. Sapiro, “Online learning for matrix factorization and sparse coding,” Journal of Machine Learning Research 11 (1) (2010) 19 – 60, 2010.

[39]. E. Shahrian, D. Rajan, B. Price, et al. “Improving Image Matting Using Comprehensive Sampling Sets,” In: Computer Vision and Pattern Recognition (2013) 636-643.

[40]. S. Li, X. Kang, J. Hu, “Image fusion with guided filtering,” IEEE Transactions on Image Processing 22 (7) (2013) 2864-2875.

[41]. Z. Zhou, S. Li, B. Wang, “Multi-scale weighted gradient-based fusion for multi-focus images,” Information Fusion 20 (2014) 60-72.

[42]. D. Tao, J. Cheng, et al., “Manifold ranking-based matrix factorization for saliency detection,” IEEE Transactions on Neural Networks and Learning Systems 27 (6) (2016) 1122-1134.