Information Fusion 42 (2018) 158–173

Contents lists available at ScienceDirect

Information Fusion

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

Full Length Article

Deep learning for pixel-level image fusion: Recent advances and future prospects

Yu Liu[a], Xun Chen[a,b,*], Zengfu Wang[c], Z. Jane Wang[d], Rabab K. Ward[d], Xuesong Wang[e,*]

[a] Department of Biomedical Engineering, Hefei University of Technology, Hefei 230009, China
[b] Department of Electronic Science and Technology, University of Science and Technology of China, Hefei 230026, China
[c] Department of Automation, University of Science and Technology of China, Hefei 230026, China
[d] Department of Electrical and Computer Engineering, University of British Columbia, Vancouver, BC, Canada
[e] School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

ARTICLE INFO

Keywords:
Image fusion
Deep learning
Convolutional neural network
Convolutional sparse representation
Stacked autoencoder

ABSTRACT

By integrating the information contained in multiple images of the same scene into one composite image, pixel-level image fusion is recognized as having high significance in a variety of fields including medical imaging, digital photography, remote sensing, video surveillance, etc. In recent years, deep learning (DL) has achieved great success in a number of computer vision and image processing problems. The application of DL techniques in the field of pixel-level image fusion has also emerged as an active topic in the last three years. This survey paper presents a systematic review of the DL-based pixel-level image fusion literature. Specifically, we first summarize the main difficulties that exist in conventional image fusion research and discuss the advantages that DL can offer to address each of these problems. Then, the recent achievements in DL-based image fusion are reviewed in detail. More than a dozen recently proposed image fusion methods based on DL techniques including convolutional neural networks (CNNs), convolutional sparse representation (CSR) and stacked autoencoders (SAEs) are introduced. At last, by summarizing the existing DL-based image fusion methods into several generic frameworks and presenting a potential DL-based framework for developing objective evaluation metrics, we put forward some prospects for the future study on this topic. The key issues and challenges that exist in each framework are discussed.

1. Introduction

The aim of pixel-level image fusion[1] is to generate a composite image from multiple input images containing complementary information of the same scene [1]. The input images known as source images are captured from different imaging devices or a single type of sensor under different parameter settings. The composite image known as fused image should be more suitable for human or machine perception than any individual input. Due to this advantage, image fusion techniques exhibit great significance in a variety of applications that rely on two or more images of the same scene. For instance, in clinical diagnosis, physicians usually need medical images obtained by different modalities including computed tomography (CT), magnetic resonance (MR), single photon emission computed tomography (SPECT), etc. In this situation, integrating the important information of different source images into a composite image can often reduce the difficulty in achieving precise diagnosis. Another typical application of image fusion is digital photography, for example to extend the depth-of-field or dynamic range of a camera. Some other popular scenarios of image fusion include video surveillance, remote sensing, etc.

The study of image fusion has lasted for more than 30 years, during which hundreds of related scientific papers have been published [2]. In recent years, deep learning (DL) has gained many breakthroughs in various computer vision and image processing problems, such as classification [3], segmentation [4], super-resolution [5], etc. In the field of image fusion, the study based on deep learning has also become an active topic in the last three years. A variety of DL-based image fusion methods have been proposed for digital photography (e.g., multi-focus image fusion, multi-exposure image fusion) [6–9], multi-modality imaging (e.g., medical image fusion, infrared/visible image fusion) [10–12], and

* Corresponding authors at: Department of Biomedical Engineering, Hefei University of Technology, Hefei 230 0 09, China (X. Chen) and School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (X. Wang).
E-mail addresses: yuliu@hfut.edu.cn (Y. Liu), xunchen@ece.ubc.ca (X. Chen), zfwang@ustc.edu.cn (Z. Wang), zjanew@ece.ubc.ca (Z.J. Wang), rababw@ece.ubc.ca (R.K. Ward), wangxuesongcumt@163.com (X. Wang).

[1] It is known that image fusion can be grouped into three categories, namely, pixel-level fusion, feature-level fusion and decision-level fusion. In this paper, we focus on pixel-level image fusion. For simplicity, we omit the term pixel-level in most expressions later.

https://doi.org/10.1016/j.inffus.2017.10.007
Received 26 September 2017; Received in revised form 19 October 2017; Accepted 24 October 2017
Available online 11 November 2017
1566-2535/

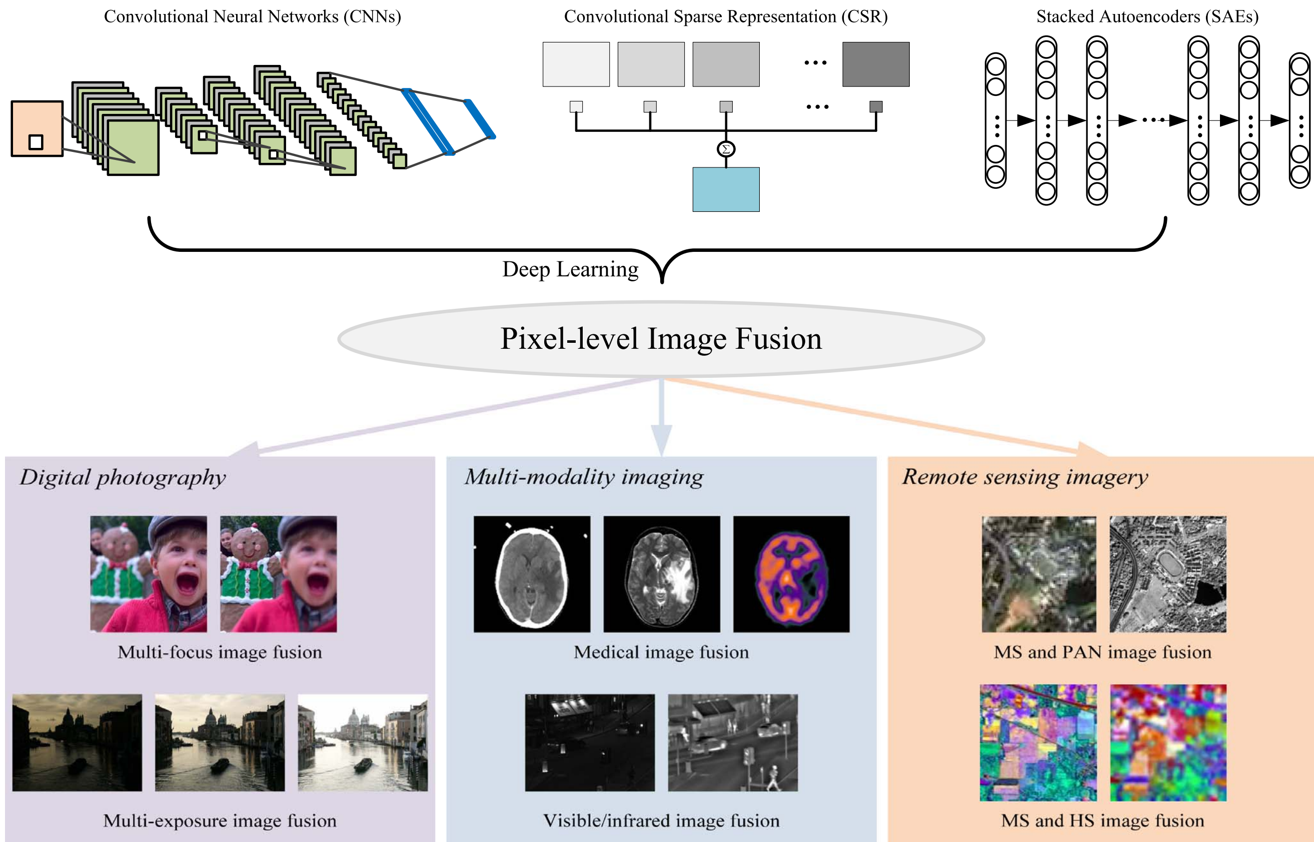

Fig. 1. An illustration of DL-based image fusion.

remote sensing imagery (e.g., multi-spectral (MS) and panchromatic (PAN) image fusion, MS and hyper-spectral (HS) image fusion) [13–19], showing advantages over conventional methods and leading to state-of-the-art results. Fig. 1 shows an illustration of DL-based image fusion. In this paper, we will review the recent advances related to DL-based image fusion and put forward some future prospects on this topic.

A number of representative survey works concerning image fusion have been proposed in the literature. In 1999, Zhang and Blum performed a detailed review on multi-scale decomposition (MSD)-based image fusion approaches [20]. Another influential survey on multi-resolution image fusion was presented by Piella in [21]. Liu et al. conducted a thorough study of the objective metrics used in fusion performance evaluation [22]. Zhang et al. gave a comprehensive review on the sparse representation (SR)-based image fusion methods [23]. Li et al. recently presented an all-round survey about image fusion covering methods, objective metrics and applications [2]. There are also arising some surveys that concentrate on the fusion issues in some specific application fields, such as medical imaging [24,25], remote sensing [26,27] and surveillance [28]. As DL-based image fusion has just been studied very recently, the related methods are not included in existing surveys related to image fusion. This paper presents a specific review of the recent achievements in DL-based image fusion, aiming to give a comprehensive introduction about the current progress in this field. Furthermore, we put forward several specific prospects for the future study of DL-based image fusion, hoping to provide some new thoughts for researchers in the field of image fusion.

In addition to the popularity of DL-based approaches for image processing, another important factor that motivates the study on DL-based image fusion is overcoming the difficulties faced by conventional image fusion research[2]. Therefore, we first summarize the related difficulties and describe the Advantages of DL for image fusion accordingly in Section 2. In Section 3, some commonly-used deep learning models in image fusion study are briefly introduced. Section 4 presents a detailed review of the recent advances in DL-based image fusion. In Section 5, we put forward some prospects for the future study of DL-based image fusion. Finally, concluding remarks are given in Section 6.

2. Motivations of DL for image fusion

2.1. Difficulties in conventional image fusion research

The study of image fusion generally contains two aspects: image fusion methods and objective evaluation metrics. In this subsection, we discuss the main difficulties that exist in these two aspects respectively. For better understanding of the discussions, some representative methods will be briefly introduced, while more comprehensive and more detailed reviews could be found in the survey papers mentioned above.

2.1.1. Image fusion methods

In view of its different application fields, image fusion can be categorized into different sub-problems, such as multi-focus image fusion, multi-exposure image fusion, medical image fusion, visible/infrared image fusion, remote sensing image fusion, etc. Among them, remote sensing image fusion, often known as pan-sharpening in most cases, has distinct differences with the other types of fusion problems. Pan-sharpening indicates the process of fusing a low-resolution MS image and a high-resolution PAN image to obtain a MS image with high spatial resolution. Thus, pan-sharpening can be viewed as a super-resolution problem for the MS image aided by a PAN image, and many methods originating from the field of image super-resolution have been applied to remote sensing image fusion. The situation of MS and HS image fusion is quiet similar, which can be regarded as a super-resolution

[2] The term *conventional image fusion research* in this paper indicates the research that is not based on deep learning approaches.

problem for the HS image with the help of the MS image. For other types of image fusion issues mentioned above, the roles of source images are generally equivalent in the fusion process. Although some general methods in the field of image fusion are applicable to remote sensing image fusion, the methods based on super-resolution have exhibited clear advantages and most methods recently proposed in this field belong to this category. Therefore, for the sake of clarity, we make a separate discussion by dividing image fusion methods into two categories, namely, *general image fusion*[3] and *remote sensing image fusion*.

1) General image fusion

According to [2], general image fusion methods can be categorized into four groups based on the image transform used: the multi-scale decomposition (MSD)-based methods, the sparse representation (SR)-based methods, the spatial domain based methods and the hybrid transform based methods. The MSD-based methods share a popular three-phase framework, namely, MSD decomposition, fusion and MSD reconstruction. Typical transforms used in MSD-based image fusion methods include image pyramids [29–32], wavelet-based transforms [33–35], multi-scale geometric transforms [36–39], spatial filtering based decompositions [40–44], etc. In the fusion phase, the activity level of source images is measured by the decomposed coefficients based on some certain pixel- or window-based approaches. Then, some pre-designed fusion rules such as choose-max and weighted-average are adopted to combine the coefficients of different source images. The basic assumption in the SR-based methods is that the activity level of source images can be measured in a sparse domain. Yang and Li first introduced a SR-based multi-focus image fusion method via orthogonal matching pursuit (OMP) for sparse coding and the max-L1 fusion strategy (choose the sparse vector which has the maximal L1-norm) for coefficient merging [45]. Since then, a variety of novel sparse representation models and related fusion strategies have been proposed for image fusion [46–53]. A comprehensive review about this topic is given in [23]. Unlike the MSD-based and SR-based methods, the spatial domain based methods accomplish the fusion task without explicitly performing a transform. A type of popular spatial domain methods is based on image blocking or segmentation. In these methods, the source images are first divided into a number of blocks or regions using some certain strategies, such as fixed block size setting in a manual way [54–56] or based on optimization approaches [57,58], quad-tree decomposition [59,60] and segmentation [61,62]. Then, the blocks or regions from different source images at the same spatial location are fused by the designed activity level measurements and fusion rules. In the past few years, many novel image fusion methods performed on pixel domain have also been proposed [63–71]. These methods tend to adopt relatively complicated fusion strategies to pursue high-quality fusion results. The hybrid transform based methods indicate the approaches which simultaneously apply more than one transforms in the fusion procedure, aiming to combine the advantages of different transforms. Representative methods belong to this category include curvelet and wavelet based method [72], wavelet and contourlet based method [73], multi-scale transform and sparse representation based method [74], etc.

In [2], the authors reviewed each category of image fusion methods from two aspects: *image transform* and *fusion strategy* (please see Table 1 in [2]). The meaning of transform, which includes different multi-scale decompositions, various sparse representation models, non transforms (can be viewed as a special case) and combination of different transforms, is very straightforward. But the scope of fusion strategy is relatively large, which can be further split into two parts in most fusion methods, namely, *activity level measurement* and *fusion rule*. Actually, they are two well-known specific terms in the field of image fusion. The target of activity level measurement is to obtain quantitative information as the basis of assigning weights to different sources. Some typical examples include the absolute value of a decomposition coefficient or the sum of absolute values of all the coefficients within a local window in the MSD-based methods, the L1-norm of a sparse vector in the SR-based methods, the spatial frequency or some other similar measures of an image block in the image blocking based methods, etc. Based on the calculated activity level measurements, the fusion rule is used to determine the contribution of each source to the fused result. Choose-max and weighted-average are two popular fusion rules in image fusion. As choose-max is just an extreme case of weighted-average, fusion rule essentially plays the role of weight assignment. Specific approaches for weight calculation include simple proportion according to activity level measurements, machine learning based techniques [55,75,76], etc. Another optional component which also belongs to the scope of fusion strategy is known as *consistency verification* [33] in image fusion. It aims to refine the calculated weights on the basis of some priors like spatial consistency. In image fusion, consistency verification is often based on image filtering techniques, such as median filtering, majority filtering, small region filtering, edge-preserving filtering, etc.

Therefore, the study on image fusion methods consists of several critical components: image transform, activity level measurement, fusion rule, etc. To promote the development of image fusion, researchers make continuous efforts on all the above components by introducing more effective image transforms, more robust activity level measurements and more elaborate fusion rules. However, despite the significant progress achieved, some bottlenecks in this field have become increasingly clear in recent years. In most conventional image fusion methods, the above components are designed in a manual manner. In particular, it is not hard to find that activity level measurement and fusion rule are always the core issues which are carefully tackled in most newly proposed methods. To pursue better performance, these issues have the trend to become more and more complicated. Nevertheless, due to the limitations of many factors such as implementation difficulty and computational cost, it is quite difficult to manually bring up an ideal design which fully concerns the important issues in a fusion task.

In addition to the above difficulty, the lack of effective image transforms is another challenge in current image fusion research. As is well known, an effective image transform is the prerequisite of high-quality fusion methods. To a large extent, the progress on image fusion is mostly achieved along with the development of image representation theories, which makes the further improvements on fusion strategy possible. The transforms used in image fusion are mainly based on image representation theories such as pyramid decomposition, wavelet transform, multi-scale geometric transform and sparse representation, etc. However, there still exist many defects in these widely popularized approaches when used for image fusion [11,23,74]. Thus, it is of urgent significance to investigate some new image representation approaches which are more effective for image fusion.

2) Remote sensing image fusion

Pan-sharpening methods in the early stage mainly include the component substitution (CS)-based methods [77–79] like intensity hue saturation (IHS)-based ones and the principal component analysis (PCA)-based ones, and multi-resolution analysis (MRA)-based methods [80–82] like wavelet transform based ones and contourlet-based ones. However, these methods are likely to either suffer from severe spectral distortion or fail to preserve spatial details. In recent years, the model-based methods have become a dominant direction in this field. In this category of methods, pan-sharpening is viewed as a super-resolution problem, aiming to restore the high-resolution MS image from the low-resolution MS image assisted by the PAN image [83]. As a result, pan-sharpening can be modeled as inverse problems and solved by

[3] The term *general image fusion* in this paper indicates common image fusion issues apart from remote sensing image fusion. Although different issues have their own characteristics, the study on them has very high coherence in the field of image fusion. Many approaches can be used for multiple image fusion issues while just with differences on some certain aspects determined by specific problems. Thus, as many other works in the literature, the discussions on these image fusion issues are presented as a whole in this paper.

designing some regularization terms. A number of restoration-based pan-sharpening methods have been recently proposed based on Markov random field [84,85], variational approaches [86,87], compressive sensing [88,89], etc. More recently, following the example learning based natural image super-resolution methods such as the well-known SR-based one [90], some pan-sharpening methods based on sparse coding with coupled dictionaries have been proposed [91–93].

From a certain viewpoint, solving the problem of image super-resolution is essentially constructing a mapping from the low-resolution image to the high-resolution image [5]. Clearly, this idea is also valid for remote sensing image fusion problem which just has different input and output. Since the problem is highly ill-posed, this mapping relationship is likely to be non-linear and very complex, such that it is practically impossible to be mathematically expressed. In the restoration-based methods, this mapping is implicitly realized by introducing some assumptions or priors, such as degraded model, gradient constraint, sparse constraint, etc. However, due to the limitation of model complexity, it is very difficult to take all the crucial factors into consideration. In the example learning based methods, such as the methods based on sparse coding with couple dictionaries, the mapping mentioned above is mainly realized through learning from the training examples. The performance of these methods depends heavily on the effectiveness of the designed learning model which is also based on some related assumptions. Thus, although these learning-based methods are more likely to achieve higher performance, the above difficulty of the restoration-based methods still exists. Overall, the main difficulty in conventional remote sensing image fusion methods is that the employed models generally do not have sufficient representation ability to characterize the complex mapping relationship between the input (source) and targeting (fused) images.

2.1.2. Objective evaluation metrics

The target of objective evaluation in image fusion is to quantitatively assess the quality of a fused image. However, this is not an easy task as the reference image is unavailable in most image fusion problems, but an exception is remote sensing image fusion.

The objective metrics for general image fusion can be divided into two groups: the metrics that are based only on the fused image and the metrics that are based on both the fused image and source images. In the first group, some simple image quality measures like standard deviation, spatial frequency and entropy are usually employed. The metrics that belong to the second group are specific for image fusion issues and can be grouped into four categories [22]: the information theory based metrics [94–97], the image feature based metrics [98–100], the image structural similarity based metrics [101–103], and the human perception based metrics [104–106].

For remote sensing image fusion, the algorithm is usually tested on the artificially degraded data so that the original MS image can be used as the reference (Wald's protocol [107]). Popular evaluation metrics employed in remote sensing image fusion include root mean squared error (RMSE), correlation coefficient (CC), spectral angle mapper (SAM) [108], ERGAS [109], Q4 [110], etc. There also exist some non-reference metrics for pan-sharpening, such as the QNR index [111], which consists of a spectral distortion index and a spatial distortion index.

When compared with the research on fusion methods, the study of objective evaluation has received much less attention, as it is really not an easy task to develop a widely recognized fusion metric. In fact, even for the current existing metrics mentioned above, it is unpractical to say that a certain one is always better than others [22,112]. For different image fusion applications, the appropriate metrics may be different. Therefore, the objective evaluation of image fusion performance still remains a challenging issue.

2.1.3. Summary

Based on the above discussions, the difficulties that exist in conventional image fusion research can be summarized into the following four aspects:

1. the increasing challenge in designing image transforms and fusion strategies (activity level measurements, fusion rules, etc) to pursue state-of-the-art results,
2. the lack of effective image representation approaches for image fusion,
3. the limited representation ability of most existing models for characterizing the complex mapping relationship between the input and targeting images in remote sensing image fusion,
4. the lack of widely recognized fusion metrics for objective evaluation.

2.2. Advantages of DL for image fusion

Owing to the strong capability in feature extraction and data representation, deep learning (DL) has led to state-of-the-art results in many computer vision and image processing tasks. In this subsection, we mainly demonstrate that image fusion can also benefit from DL techniques. In particular, for each of the four difficulties mentioned above, the specific advantages of DL for image fusion are discussed.

For the first difficulty, the popular DL model convolutional neural networks (CNNs) can provide some novel ways for studying image fusion methods. This is because some image fusion issues can be considered as classification problems [55]. One classic example is multi-focus image fusion, which is usually based on the assumption that a local region is well-focused in only one source image. Therefore, the problem can be naturally interpreted as a classification issue by selecting the well-focused one among all the source images. For other image fusion issues, although this selection-based fusion strategy may be not appropriate, they can still be modeled as classification problems by defining each output class as a weight assignment (more discussions on this issue are given in Section 5.1). Based on the above consideration, we would like to make a more insightful comparison between image fusion and visual recognition. As is well known, conventional visual recognition methods generally contain three crucial steps, namely, feature extraction, feature selection and prediction. It is natural to match them with the three main steps of image fusion mentioned above. That is, the image transform, activity level measurement and fusion rule in image fusion can approximately correspond to feature extraction, feature selection and prediction in visual recognition, respectively. When considering the specific meaning of each step, the above correspondence is easy to understand. It is known that CNNs have achieved significant advances in visual recognition problems when compared with conventional methods since CNNs are capable of learning the most effective features from a large amount of training data. Therefore, CNNs also have great potential to be used for image fusion. The image transforms, activity level measurement and fusion rules (or part of them) can be jointly implemented in an implicit manner through learning a convolutional network. The main advantages of this type of approaches are similar to those of the CNN-based classification tasks, namely, avoiding the complexity of conventional handcrafted design and being more likely to obtain better performance. Several CNN-based image fusion methods on account of this idea have been recently proposed [6,8,12]. Most recently, CNNs have also been successfully employed for high dynamic range (HDR) imaging through modeling the complex fusion process of multi-exposure images in dynamic scenes [9]. Via the powerful modeling ability of deep convolutional networks, the method can overcome the difficulty faced by conventional HDR imaging (multi-exposure image fusion) in manually designing effective strategies to avoid motion artifacts.

For the second one, some image representation approaches derived from DL can be used as transforms for image fusion. For example, convolutional sparse coding (CSC) is the basic model of deconvolutional networks [113] for feature extraction. CSC can obtain a sparse representation for an entire image, rather than independently computing the representations for a set of overlapping image patches as the traditional standard sparse coding (SC) does, leading to several advantages over SC [114]. From the viewpoint of image representation, CSC is also known as convolutional sparse representation (CSR) [114]. In [11], a CSR-based fusion method has been presented for multi-focus image fusion and multi-modal image fusion, exhibiting advantages over conventional SR-based methods in detail preservation and robustness against mis-registration. Moreover, some CNN architectures for other image processing problems may be used as feature representation approaches for image fusion. A typical example which has been recently proposed is extending the convolutional network for super-resolution in [5] to image fusion [7], and the fusion process is performed on CNN feature maps. As a result, image fusion and super-resolution can be simultaneously accomplished under a unified CNN framework.

For the third one, DL-based approaches, and in particular CNN-based methods, have recently attracted great attention in the study of natural image/video super-resolution, leading to many state-of-the-art results [5,115–117]. The success is mainly due to the strong capability of DL models in characterizing complex relationship between different signals. In remote sensing image fusion, a series of pan-sharpening methods based on DL models such as stacked autoencoders (SAEs) [13,16] and deep convolutional networks [14,15,17,18] have been successfully proposed in the past three years. In addition to the traditional pan-sharpening problem which aims to fuse a MS image and a PAN image, the problem of MS and HS image fusion has also been addressed using a DL-based approach [19].

For the last one, most objective evaluation metrics in image fusion are based on measuring some certain types (e.g., original image, gradient map, etc) of local similarity between the fused image and the source images. In [118], a CNN-based method was proposed for measuring the local similarity between two image patches, and the experimental results demonstrated the superiority of the CNN-based method over traditional methods. Therefore, it is a promising direction to investigate the effectiveness of convolutional networks for designing objective fusion metrics. To the best of our knowledge, there is no related work that has yet emerged in the literature. In Section 5, we will present a potential CNN-based scheme for the design of image fusion metrics.

In summary, the key reasons that result in the popularity of DL-based methods for image fusion include the following aspects: 1) DL models can extract the most effective features automatically from data to overcome the difficulty of manual design. 2) DL models can characterize very complex relationship between input and targeting data. 3) DL community provides some potential image representation approaches which could be useful to the study of image fusion. 4) The availability of friendly DL libraries (e.g., Caffe, TensorFlow, Theano and MatConvNet) and large-scale image datasets (e.g., CIFAR, PASCAL VOC and ImageNet) ensures convenient investigations on this topic.

3. Commonly-used DL models in image fusion

3.1. Convolutional neural networks (CNNs)

As a popular supervised DL model, CNN [119] is a trainable multilayer architecture aiming to learn a multistage feature representation of the input data, and each stage is composed of a number of feature maps. The coefficient in a feature map is known as a neuron. Feature maps at different stages are usually connected by performing several types of calculations including convolution, non-linear activation and spatial pooling. Unlike the traditional multilayer perception (MLP)-based neural networks, the neurons between two adjacent stages in a CNN are locally connected via a convolutional operation and a weight-sharing strategy, which dramatically decreases the amount of free parameters to be learned. Mathematically, the convolutional layer in a 2D-CNN can be expressed as

$$y^j = b^j + \sum_i k^{ij} * x^i, \tag{1}$$

where x^i is the ith input feature map and y^j is the jth output feature map. The network parameters k^{ij} and b^j are the 2D convolutional kernel and bias, respectively. The symbol * denotes 2D convolution. Recently, the 3D-CNN [120] has also attracted great attention in various vision-based applications that rely on image sequences or videos. In a 3D-CNN, each feature representation stage contains a set of 3D feature maps, and each 3D map indicates a certain number of 2D maps. The convolutional layer in a 3D-CNN can be also represented as Eq. (1), but with differences in terms of variable meanings. For a 3D convolutional layer, x^i and y^j are both 3D feature maps, k^{ij} is a 3D convolutional kernel, and the symbol * denotes 3D convolution. Actually, 3D-CNN is a more generalized model relative to 2D-CNN, as it will degrade into 2D-CNN when each 3D map only consists of a single 2D map. The non-linear activations widely used in CNNs include sigmoid function, rectified linear units (ReLUs), etc. For example, a convolutional layer and a ReLU layer can be jointly expressed as

$$y^j = \max\left(0,\, b^j + \sum_i k^{ij} * x^i\right). \tag{2}$$

In CNNs, a convolutional layer is always followed by a non-linear activation layer[4]. Another popular operation in CNNs is spatial pooling (sub-sampling), and in particular max-pooling, which can bring some desirable invariances including translation, rotation and scale into the model to a certain extent. It can also reduce the dimension of feature maps which is helpful to prevent overfitting, and the computational and memory costs of the model are also reduced accordingly.

Fig. 2(a) shows a typical CNN architecture for visual classification tasks. Clearly, it is an end-to-end system from the input image to the output prediction vector that contains the probability of each category. From the perspective of traditional visual classification approaches, the convolutional layers and max-pooling layers can be regards as the feature extraction/selection part of the system, while the fully-connected layers act as the role of a classifier. Because a fully-connected layer needs to pre-define the dimensions of input and output, the image fed to the network must have fixed size. For patch-based tasks like object detection, a possible way is to apply the sliding window technique to divide the whole image into a large number of overlapped patches and then fed each patch to the network, but it will introduce numerous repeated calculations as the patches are usually greatly overlapped. Fortunately, a fully-connected layer can be converted into an equivalent convolutional layer since they are both linear operators [4,121]. Concretely, the full connection is a special case of convolution when the spatial size of convolutional kernel equals to that of input data. By converting all the fully-connected layers into convolutional layers, the network can handle input image of arbitrary size as a whole to generate a dense prediction map. This technique has been widely used in object detection [121] and image segmentation [4].

Fully convolutional networks have also been employed as a regression model in many low-level problems such as image super-resolution. Dong et al. [5] first introduced a CNN model, termed SRCNN, into single-image super-resolution. The SRCNN model consists of three convolutional layers without max-pooling layers as the input and output have the same size. The mean squared error between network output and the ground truth high-resolution image is used to define the loss function. An

[4] For this reason, the non-linear activation layer will not be explicitly mentioned later, and it is regarded as being integrated into the convolutional layer.

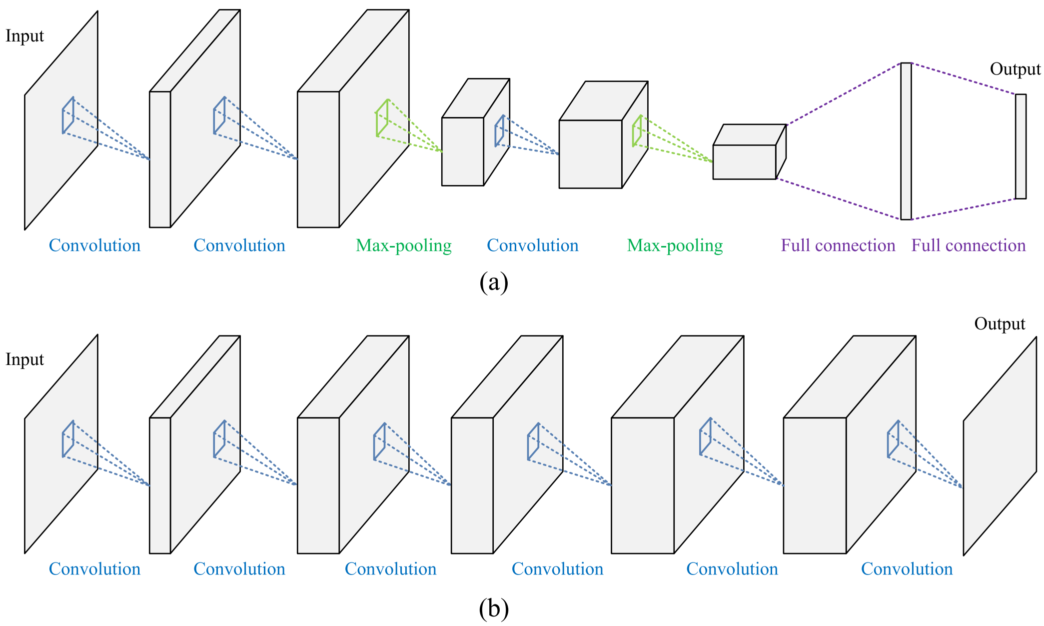

Fig. 2. Typical CNN architectures. (a) For visual classification. (b) For image super-resolution.

illustration of the CNN architecture for image super-resolution is given, as shown in Fig. 2(b). Recently, deep residual learning [122] has emerged as an active topic in image/video super-resolution. Kim et al. [115] proposed a 20-layer network for single-image super-resolution based on residual learning, namely, the residual image (the difference between the high-resolution image and the up-sampled low-resolution image) is actually applied to train the network. The residual image mainly contains the detail information, and most pixel values are approximate to zero. As a consequence, very deep network is applicable without the concern of occurring vanishing gradient problem, so that more powerful modelling capability can be obtained. Furthermore, the network with residual learning has a much higher convergence speed, so the training time is significantly shortened.

CNNs have also been applied to image processing issues that need more than one input. For example, Zagoruyko and Komodakis [118] proposed a patch similarity measure approach via convolutional networks. The two image patches are fed to a network and the output is the similarity measure. In [118], three basic architectures are presented, namely, siamese, pseudo-siamese and 2-channel. The siamese network has two branches with the same architecture to adopt the two input patches, respectively. Each branch has a series of convolutional layers and max-pooling layers for feature extraction. The outputs of two branches are finally concatenated as the input of following layers like fully-connected ones. In addition, the two branches share exactly the same weights in the siamese network. The pseudo-siamese network has similar architecture to the siamese network except that the weights of two branches are not constrained to be the same, so it provides more flexibility than the siamese one. The 2-channel network only has a trunk with no branches (just as conventional single-input networks) by concatenating the two patches before feeding them to the network. By jointly handling the two inputs at the starting point, the 2-channel network owns the highest flexibility among these three architectures [118].

Overall, CNNs have very powerful ability in both feature extraction and data representation. On one hand, it can learn the most effective features from the training data without the manual intervention. On the other hand, the complex relationship between different signals can be well modelled by a deep convolutional network. The design of CNN architecture is very flexible to make it applicable to a variety of classification and regression tasks. Furthermore, most existing DL libraries like Caffe, TensorFlow and MatConvNet develop the CNN related functions and interfaces very well, making the explorations based on CNNs very convenient.

3.2. *Convolutional sparse representation (CSR)*

The concept of convolutional sparse coding (CSC) originates from the deconvolutional networks proposed by Zeiler et al. [113]. The basic idea of CSC is to obtain a convolutional decomposition of an image under a sparsity constraint. The goal of deconvolutional networks is to learn a multistage feature representation of input image by building a hierarchy of such decompositions. The input image can be reconstructed from its decompositions in a layer-wise manner. Thus, deconvolutional network provides a promising image representation approach for both feature learning and reconstruction based problems. As the basic architecture of deep deconvolutional networks,[5] CSC has been successfully applied in a broad range of vision problems such as video background modeling [124], object detection [125], super-resolution [126], etc.

In CSC, given a set of dictionary filters d_m, an image s is decomposed into the sum of a set of convolutions as $s = \sum_m d_m * x_m$, where x_m is a set of unknown coefficient maps, as shown in Fig. 3. By regularizing x_m with sparsity prior, the CSC model is given by

$$\arg\min_{\{x_m\}} \frac{1}{2}\left\|\sum_m d_m * x_m - s\right\|_2^2 + \lambda \sum_m \|x_m\|_1, \tag{3}$$

where λ is the regularization parameter. The dictionary filters are typically learned from a number of training images using the following model

$$\arg\min_{\{d_m\}\{x_{k,m}\}} \tfrac{1}{2}\textstyle\sum_k \|\sum_m d_m * x_{k,m} - s_k\|_2^2 + \lambda \sum_k \sum_m \|x_{k,m}\|_1$$
$$\text{s.t} \quad \|d_m\|_2 = 1 \tag{4}$$

where s_k is a set of training images, and the constraint on d_m is mainly used to prevent the scaling ambiguity. In the past few years, many efficient algorithms based on the alternating direction method of multipliers (ADMM) framework have been proposed to solve the above optimization problems [114,127,128], which promotes the practical usage of CSC to a great extent.

In addition to being a fruit of DL community, CSC also has an origin in signal processing research, aiming to study shift-invariant sparse

[5] In addition, Papyan et al. recently demonstrated that there exists a tight connection between the CNNs and multi-layer CSC [123]. Therefore, although the CSC itself is not a "deep" architecture, it is viewed as a DL model within the scope of this paper.

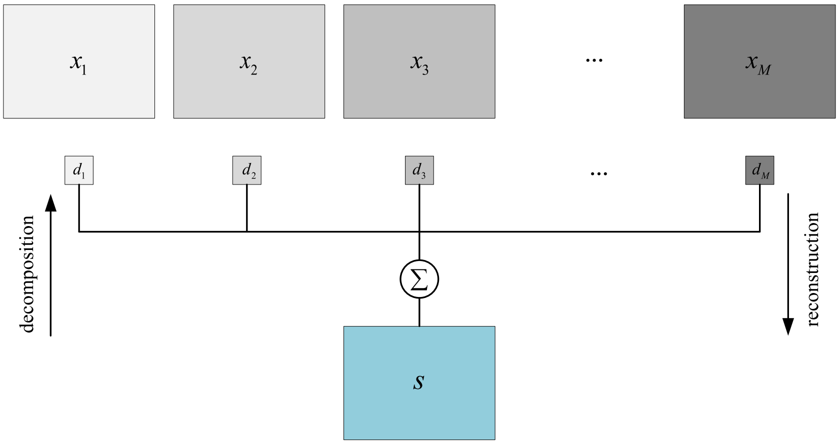

Fig. 3. The convolutional sparse coding model.

representation [114]. The convolutional form is the derivation rather than assumption in this stream. As an image representation approach, CSC is also termed convolutional sparse representation (CSR), owning several advantages over conventional sparse representation (SR) [114]. In conventional SR-based image processing approaches, sparse decomposition is independently performed on a set of overlapping patches extracted by the sliding window technique. As a consequence, the representation is multi-valued and not optimal with respect to the entire image. In contrast, the sparse representation of an entire image is computed in the CSR model, so that the obtained representation is single-valued and optimized over the entire image. Furthermore, CSR is a shift-invariant representation approach, which is a crucial property in many applications including image fusion. A popular CSR library developed by Wohlberg is available at [129].

3.3. Stacked autoencoders (SAEs)

SAEs are a popular category of deep neural networks in many visual classification and image restoration problems. The training procedure of an SAE generally consists of two main steps: unsupervised pre-training and supervised fine-tuning.

The pre-training is conducted in a layer-wise unsupervised manner [133]. At each layer, an AE is trained to obtain a set of features by jointly learning an encoder and a decoder. Fig. 4(a) shows a typical AE architecture. Let $\{x^i\}_{i=1}^M$, $x^i \in R^N$ denote a set of input signals of the encoder. The encoding result of x^i is calculated by

$$y^i = f(Wx^i + b), \tag{5}$$

where $W \in R^{K\times N}$ is the encoding weight matrix, $b \in R^K$ is the encoding bias, and $f(\cdot)$ is the activation function of the encoder. The decoder is defined as

$$z^i = g(W'y^i + b'), \tag{6}$$

where $W' \in R^{N\times K}$ is the decoding weight matrix, $b' \in R^N$ is the decoding bias, and $g(\cdot)$ is the activation function of the decoder. The decoding matrix can be set as the transpose of the encoding matrix. Popular activation functions for encoder and decoder include sigmoid function $s(x) = (1 + \exp(-x))^{-1}$, hyperbolic tangent function $s(x) = \tanh(x)$, identity function $s(x) = x$, etc. To calculate the parameter set $\Theta = \{W, W', b, b'\}$, a commonly used loss function of AE is defined as

$$L_1(\Theta) = \frac{1}{M}\sum_{i=1}^{M} \|\hat{x}^i - z^i\|_2^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|W'\|_F^2), \tag{7}$$

where $\hat{x}^i$ is the desired output of the decoder. Actually, $\hat{x}^i$ is the original training example of the AE, while the input signal x^i is obtained from $\hat{x}^i$. The first term is the data term for minimizing the reconstruction error. The input signal x^i can be exactly the same as $\hat{x}^i$, which means that the reconstruction result is expected to be the input signal. This category of design often occurs in classification problems for feature extraction. In this situation, the dimension of the code y^i is generally not equal to that of the input signal to avoid learning an identity function (the code is just the input). Another way to prevent learning a trivial solution is applying the denoising AE (DAE) proposed by Vincent et al. [134], in which the input signal x^i is the degraded version (corrupted by noise) of $\hat{x}^i$. In this case, the AE model is trained to reconstruct the original high-quality signal from its degraded version. In [135], Xie et al. introduced the DAE model into image restoration issues including denoising and inpainting (the DAE firstly proposed in [134] is used for classification rather than denoising). The second term is known as a weight decay term for regularization. In fact, some other constraints can also be used to regularize the above problem, among which the sparsity is perhaps the most representative one. A popular sparsity term is based on the Kullback-Leibler (KL) divergence as

$$KL(\rho\|\hat{\rho}) = \rho\log\frac{\rho}{\hat{\rho}} + (1-\rho)\log\frac{1-\rho}{1-\hat{\rho}}, \tag{8}$$

where ρ is a sparsity parameter whose value is close to zero, and $\hat{\rho} = \frac{1}{M}\sum_{i=1}^{M}\bar{y}^i$ is the average activation of y^i (denoted as $\bar{y}^i$) averaged over all the inputs. The KL divergence gets the minimal value 0 only if $\hat{\rho} = \rho$, implying that most activations are close to zero. By applying this sparse constraint, the modified loss function for sparse AE is defined as

$$L_2(\Theta) = \frac{1}{M}\sum_{i=1}^{M}\|\hat{x}^i - z^i\|_2^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|W'\|_F^2) + \beta KL(\rho\|\hat{\rho}) \tag{9}$$

To build a deep architecture, the above training approach is iteratively performed to obtain SAEs. In visual classification problems, once an AE layer is trained, the decoder is removed, and the current encoding result of the original signal is employed in the next round of training [133,134]. In image restoration problems, both the encoder and decoder are preserved after a layer is trained. In [135], the current encoding results of the original high-quality signal and the original degraded signal are applied to train the next layer.

In the fine-tuning stage, an output layer is generally added on the top of trained SAEs for visual classification problems, and all the parameters in the network are fine-tuned to minimize the prediction error using the labeled training examples. A typical SAE architecture for classification problems is shown in Fig. 4(b). For image restoration issues, the fine-tuning is directly performed on the pre-trained SAEs without adding an output layer. Fig. 4(c) shows the related SAE architecture.

As a typical hierarchical data representation model, SAEs own the basic characteristics of DL models. In particular, due to the two phase (pre-training and fine-tuning) based training mechanism, SAEs have a high potential when the scale of labeled data for supervised learning is limited. As mentioned above, SAEs are applicable to both visual

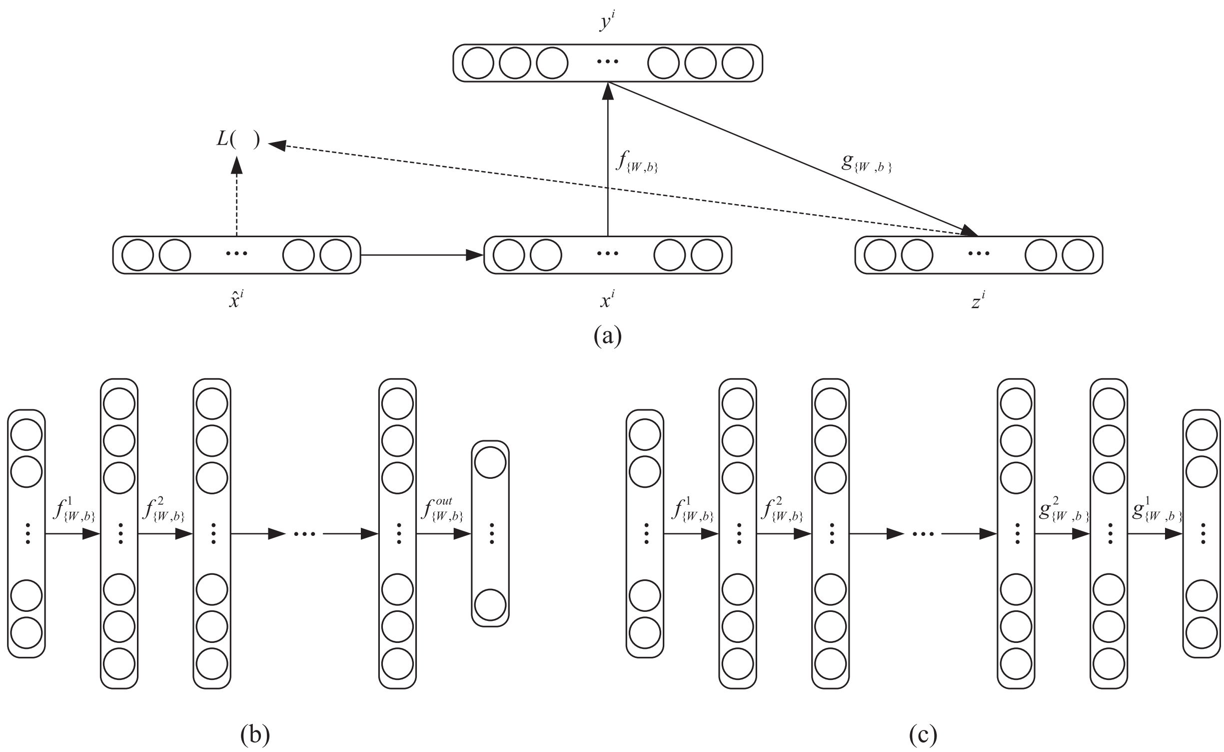

Fig. 4. Typical AE architectures. (a) The basic AE architecture. (b) The stacked AE architecture for visual classification problems. (c) The stacked AE architecture for image restoration problems.

classification and image restoration problems by adjusting the network architectures and adding constraints on the inner features. Moreover, SAEs can be easily implemented via DL libraries such as TensorFlow and Theano.

4. A review of DL-based image fusion study

In the last three years, a number of DL-based image fusion methods have been proposed. In accord with the contents in Section 2, the review in this section is also divided into general image fusion and remote sensing image fusion. Tables 1 and 2 list the outline of existing DL-based general and remote sensing image fusion methods, respectively. For each method, the adopted DL model, origin of training dataset, specific applications and major characteristics are listed.

4.1. General image fusion

Liu et al. [6] introduced a siamese convolutional network for multi-focus image fusion. In their method, multi-focus image fusion is regarded as a two-class classification problem and modeled by a CNN in spatial domain. Specifically, each branch of the siamese network is composed of three convolutional layers and one max-pooling layer, and the obtained feature maps of two branches are concatenated as the input of two fully-connected layers. In the training stage, two image patches of the same content but with different clarity are fed to the two branches of the siamese architecture, respectively. The output is a 2-dimension vector indicating the probability of each patch being the more focused patch. The training examples are created by performing multi-scale Gaussian filtering with different standard deviations on a large number of natural images from the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) dataset [130]. After training the network, the fully-connected layers are transformed into the equivalent convolutional layers to allow input of arbitrary size. In the fusion stage, two source images are fed to the network and the output is a dense map which contains the focus information of source images. In particular, each coefficient in the map indicates the focus property of a source patch pair at a corresponding location. By applying the popular aggregating and averaging strategy, a focus map of the same size as source images is obtained. Then, a binary segmentation is performed on the focus map to generate the decision map. Next, two commonly-used consistency verification techniques are applied to refine the decision map. Finally, the fused image is calculated using the pixel-wise weighted-average strategy. The main characteristic of this method is the two crucial issues in image fusion, namely, activity level measurement and fusion rule, can be jointly generated in an "optimal" manner through training a CNN, which provides a feasible way to overcome the difficulty of manual design in conventional fusion methods. The experimental results in [6] demonstrate that this method can achieve state-of-the-art multi-focus image fusion results. The source code of this method and trained CNN model are publicly available at [136].

Du and Gao [8] recently proposed an improved work by introducing a multi-scale input manner for the convolutional network. The network architecture in [8] generally follows that used in [6], namely, a siamese architecture in which each branch has three convolutional layers and one max-pooling layer. The main distinction is that the source image patches are extracted using a multi-scale manner and fed to the network individually. Concretely, the source images are divided into overlapping patches with three different sizes in their method, indicating three different scales. Each image patch is resized to the size of 16×16 to satisfy the network input. Thus, three focus maps are obtained by the network, and they are then averaged to generate a single merged focus map. Similar to [6], the entire pipeline of the multi-focus image fusion method in [8] also consists of four main steps: focus map generation via a CNN, initial segmentation, consistency verification and fusion. The authors experimentally show that the perfomance of the original CNN-based method [6] can be slightly improved using their multi-scale strategy, but the computational cost increases to a great extent since the patch-based manner is employed for network input.

Another work that follows [6] was presented in [12], which extends the usage of the CNN model proposed in [6] to multi-modal medical

Table 1
An outline of existing DL-based general image fusion methods.

Reference	DL model	Training dataset	Specific applications	Major characteristics
[6]	CNN	10000 natural images from the ILSVRC dataset [130]	multi-focus image fusion	A classification oriented CNN with siamese architecture is designed to jointly encode the activity level measurement and fusion rule in image fusion.
[8]	CNN	A number of natural images (not explicitly mentioned)	multi-focus image fusion	A multi-scale input manner is introduced based on the CNN-based image fusion framework [16].
[12]	CNN	10000 natural images from the ILSVRC dataset [130]	medical image fusion	The CNN model presented in [16] is extended to medical image fusion by incorporating it with MSD-based approaches.
[9]	CNN	A dataset specific for learning-based HDR released by the authors at [131]	multi-exposure image fusion	A fully convolutional network is learned to model the complex process of preventing the motion artifacts in dynamic scenes.
[11]	CSR	50 natural images	multi-focus image fusion, medical image fusion, visible/infrared image fusion	The CSR is used as the image representation approach for image fusion and a CSR-based fusion framework is introduced.
[10]	CNN	90 natural images from the ILSVRC dataset [130]	multi-focus image fusion, medical image fusion, visible/infrared image fusion	A simultaneous image fusion and super-resolution method under the UWT framework is presented and the SRCNN model [5] is used for the super-resolution of high-frequency bands.
[7]	CNN	A popular training dataset for image super-resolution (available at [132])	multi-focus image fusion	A more unified framework is designed for simultaneous image fusion and super-resolution and the fusion process is performed on the feature maps obtained by the SRCNN model.

Table 2
An outline of existing DL-based remote sensing image fusion methods.

Reference	DL model	Training dataset	Specific applications	Major characteristics
[13]	SAE	The source PAN image (IKONOS and QuickBird)	MS and PAN image fusion	A sparse denoising SAE model trained by the PAN image is applied to model the complex relationship between low- and high-resolution MS images.
[16]	SAE	The source PAN image (QuickBird and WorldView-3)	MS and PAN image fusion	The basic assumption in the SAE-based approach [6] is explained and the impacts of several key factors on the algorithm performance are studied.
[14]	CNN	A popular training dataset for image super-resolution (available at [132])	MS and PAN image fusion	The SRCNN model and GS transform are successively employed to enhance the spatial resolution of the MS image.
[15]	CNN	Same type as the testing MS/PAN images (IKONOS, GeoEye-1 and WorldView-2)	MS and PAN image fusion	An SRCNN architecture is used to model the pan-sharpening process as an end-to-end mapping and some domain-specific knowledge is introduced to modify the network input.
[17]	CNN	Same type as the testing MS/PAN images (LANDSAT 7 ETM +)	MS and PAN image fusion	A residual network with SRCNN architecture is applied to accomplish the pan-sharpening task.
[18]	CNN	Same type as the testing MS/PAN images (QuickBird and WorldView-2)	MS and PAN image fusion	A deep CNN which contains 11 convolutional layers is designed to take full advantage of residual learning for pan-sharpening.
[19]	CNN	Same type as the testing MS/HS images (ROSIS Pavia center dataset)	MS and HS image fusion	A 3D-CNN is introduced to model the process of MS and HS image fusion.

image fusion. In order to achieve perceptually better results, instead of performing the fusion process in spatial domain, the authors applied a multi-scale decomposition (MSD)-based framework since it is more consistent with human visual system and more preferred in multi-modal image fusion. Specifically, the source images are decomposed into Laplacian pyramids while the weight map obtained by the CNN is decomposed into a Gaussian pyramid. In addition, a local similarity based strategy is employed to adjust the fusion rule adaptively, which is also based on the characteristics of multi-modal source images. The experimental results in [12] show that this method outperforms several state-of-the-art medical image fusion methods.

Kalantari and Ramamoorthi [9] recently proposed a CNN-based HDR imaging method to address the problem of motion artifacts in dynamic scenes. In their method, three low dynamic range (LDR) images are employed as the input. They are firstly aligned with optical flow based technique by setting the input with the medium exposure as the reference image. However, due to the existence of moving objects in a scene, it is practically impossible to obtain a strict alignment, leading to undesirable artifacts around motion boundaries and occluded regions in the final HDR image. Conventional approaches to tackle this problem are mainly based on detecting motion regions and excluding them with some manually designed strategies. The authors creatively proposed a learning-based scheme to model this complex process via CNNs. The three aligned LDR images are used as the input of a convolutional network to produce the HDR image directly. To this end, the authors designed a novel approach to create a dataset for CNN training, in which each training example contains three LDR images with motion and their ground truth HDR image. Specifically, they first capture three LDR images with different exposures in a static scene, and the ground truth HDR image is generated from these three images using a simple-yet-effective approach. Then, three bracketed exposure images of the same scene but with motion are captured. Finally, the medium exposure one in the dynamic set is replaced by the corresponding one in the static set. In this way, a dataset for CNN training is created, which is also of great significance to the future study in this field. The Euclidean distance between tone-mapped ground truth and estimated HDR images is used to define the loss function. In [9], the authors explored three different system architectures by adjusting the network output. The experimental results in [9] verify that this method consistently produces better results than several state-of-the-art approaches in dynamic scenes. The source code of this method and the created dataset are publicly available at [131].

In [11], Liu et al. introduced the CSR into the field of image fusion. The proposed method is applicable to several kinds of image fusion issues including multi-focus image fusion, visible/infrared image fusion and medical image fusion. In their method, each source image is decomposed into a detail layer and a base layer. For the detail layer, the CSC is first performed to obtain the sparse coefficient maps. Similar to the SR-based image fusion methods, the L1-norm of a sparse vector at each pixel location is assumed to contain the activity information of source images. To make the fusion process more robust to mis-registration, the local window summation based strategy is employed to calculate the activity level measurements. Then, the choose-max rule is applied to obtain the fused coefficients. Finally, the fusion result of detail layers is reconstructed by the fused sparse coefficient maps and dictionary filters. For the base layer, different fusion strategies are adopted to different image fusion issues. Particularly, the choose-max rule based on the activity level measurements is applied to fuse multi-focus images, while the averaging rule is used for multi-modal image fusion. The final fused image is reconstructed by the fused detail and based layers. In [11], the dictionary filters are learned from a set of high-quality natural images and the impact of the number of filters on the fusion performance is studied. The authors demonstrate that the CSR-based methods owns two main advantages over the SR-based methods, namely, more effective for detail preservation and less sensitivity to mis-registration. The source code of this method and learned dictionary filters are publicly available at [136].

Zhong et al. [10] proposed a simultaneous image fusion and super-resolution method based on CNNs. In their method, the source images are first upsampled and decomposed with undecimated wavelet transform (UWT). Then, the SRCNN model [5] is employed to enhance the spatial resolution of the high-frequency maps, and the traditional choose-max rule is applied to merge the enhanced coefficients based on their absolute values. The low-frequency coefficients are directly fused with the averaging rule. Finally, the fused image is reconstructed by the high-frequency and low-frequency coefficients. Unlike natural image super-resolution, the network in this study is trained by a large number of UWT high-frequency maps obtained from natural images in the ILSVRC dataset. The effectiveness of this method is experimentally verified by comparing with some well-known methods on multi-focus image fusion, visible/infrared image fusion, and medical image fusion.

It is easy to find that the CNN model only takes part in the super-resolution task in the above method [10], while does not directly contribute to the fusion process. A more unified framework for simultaneous image fusion and super-resolution is proposed by Yang et al. [7]. In their method, the SRCNN model [5] is also applied, but the first convolutional layer is duplicated to construct an architecture having two branches to accept two upsampled source images. The feature maps obtained by the first layer of two source images are combined using a weighted-average fusion rule. The weights are calculated based on spatial frequency with a novel two-scale strategy. The merged feature maps are fed to the remaining part of the CNN model to produce the final high-resolution fused image. The authors demonstrate that this method can obtain the fused images with better visual quality and acceptable computation efficiency when compared with some state-of-the-art methods.

4.2. Remote sensing image fusion

In [13], Huang et al. proposed a SAE-based pan-sharpening method which is the first work in this field. In their method, a sparse denoising AE is applied to model the complex relationship between low- and high-resolution MS image patches as a non-linear mapping. The authors assumed that this relationship is the same as that between low- and high-resolution PAN image patches, so the training examples are sampled from the PAN image and its upsampled low-resolution version. The AE model employed in the method generally follows the sparse denoising AE model presented in [135]. The input of the network is an upsampled low-resolution image patch generated from the ground truth, namely, its corresponding high-resolution patch. The loss function defined in Eq. (9) is used for training. The entire SAE after pre-training is finally fine-tuned to achieve better performance. In the testing phase, each band of the low-resolution MS image is divided into overlapping patches and each patch is fed to the trained network to produce the high-resolution patch. The targeting high-resolution MS image is reconstructed by averaging the overlapping patches. The experiments on IKONOS and QuickBird datasets demonstrate that this SAE-based method produces better results than some well-known pan-sharpening methods in terms of visual perception and objective measures.

Azarang and Ghassemian [16] presented a further study based on the above work [13]. In [16], the basic assumption that the relationship between low- and high-resolution MS images can be characterized by the PAN images is explained from the viewpoint of multi-resolution analysis (MRA)-based pan-sharpening framework. Furthermore, the authors conducted extensive experiments on QuickBird and World-View-3 datasets to investigate the impacts of some important factors including patch size, overlapping percentage of patches and the number of training samples on the algorithm performance.

Zhong et al. [14] proposed a pan-sharpening approach with the SRCNN model and Gran-Schmidt (GS) transform. In their method, the SRCNN model is employed to enhance the spatial resolution of the MS image. Concretely, the IHS transform is performed on the unsampled low-resolution MS image, and the intensity component is fed to the SRCNN model. The enhance MS image is obtained by inverse IHS transform. Then, the average band of the enhanced MS image is calculated as the reference of histogram matching to modify the PAN image. Finally, the GS transform is applied to generate the high-resolution MS image based on the enhanced MS image and the modified PAN image. The experimental results on QuickBird dataset show that this method can obtain better results than several conventional pan-sharpening methods.

Another work that applies CNNs to remote sensing image fusion was introduced by Masi et al. [15], in which the convolutional network is used to model the pan-sharpening process as an end-to-end mapping. The network employed also follows the SRCNN architecture, but the input is the stacking result of upsampled low-resolution MS image and the PAN image. The network output is the targeting high-resolution MS image. The training examples are artificially generated based on Wald's protocol [107], namely, the original MS images at hand are employed as the ground truth while the degraded source images are used as training input. In the fusion phase, the source images are entirely fed to the trained network to produce the result. In [15], the authors further analyzed the characteristics of the learned kernels and obtained feature maps. Based on some meaningful observations, some domain-specific knowledge in remote sensing imagery is introduced to further improve the algorithm performance. Specifically, some radiometric indices (e.g., water, vegetation, etc.) generated from the MS bands are added to the network input. The impacts of some key parameters such as the number and spatial size of convolutional filters are also fully studied in [15] for images captured by different remote sensing sensors. The authors experimentally demonstrate that this CNN-based method can obtain state-of-the-art pan-sharpening results on IKONOS, GeoEye-1 and WorldView-2 datasets. The source code of this method and trained CNN models are publicly available at [137].

Following the idea presented in [15], Rao et al. [17] et al. recently proposed a pan-sharpening method based on a residual convolutional network. The main difference is that the network output is the residual image between the ground truth high-resolution MS image and the upsampled low-resolution MS image. The network employed in [17] also follows the SRCNN model which only contains three convolutional layers. The experimental results on LANDSAT 7 ETM+ dataset demonstrate that some improvements can be achieved on several objective fusion metrics when compared with the method proposed in [15].

Another recent work that introduces residual learning into CNN-based pan-sharpening was proposed by Wei et al. [18], in which a much deeper network is employed to take full advantage of residual learning. Specifically, the network used in [18] contains 11 convolutional layers in total. Another distinction in [18] is that the residual output is actually not the final residual MS image, and there is an extra convolutional layer after the summation of the input and the residual output. The experimental results on QuickBird and WorldView-2 dataset show that this method exhibits clear advantages over the CNN-based method [15] without residual learning.

In addition to the conventional pan-sharpening issue (MS and PAN image fusion), some other remote sensing image fusion tasks have also benefited from the DL-based approaches. In [19], Palsson et al. recently proposed a MS and HS image fusion method based on a 3D-CNN. The target of MS and HS image fusion is to generate a high-resolution HS image. The HS images used in [19] come from the ROSIS Pavia Center dataset and the corresponding MS images are simulated from the HS images by averaging the HS bands according to the spectral response profiles of targeting MS bands. The basic idea of this method is similar to that of 2D-CNN-based pan-sharpening method [15]. To control the computational cost, the dimension of the HS image is firstly reduced using PCA. A 3D-CNN which contains three convolutional layers is trained by the MS image and the HS image after performing dimension reduction based on Wald's protocol [107]. In the fusion stage, dimension reduction is also required for the source HS image. After obtaining the network output, the final high-resolution HS image is reconstructed by performing inverse PCA transform. The experiments on the ROSIS Pavia Center dataset show that this 3D-CNN based method can obtain better performance than several existing model-based MS and HS image fusion methods.

5. Future prospects

Despite of the above achievements, the study of DL-based image fusion is currently at a beginning stage and there still exists great space for further development. In this section, we put forward some prospects for the future study on this topic mainly from two aspects. On one hand, we summarize the DL-based image fusion methods described above into several generic frameworks. On the other hand, a potential DL-based framework for designing objective fusion metrics is introduced. For each framework, the existing key issues and challenges are discussed.

5.1. General image fusion methods

It can be seen from Table 1 that the effectiveness of CNNs for general image fusion has been widely verified. In particular, both classification-type network and regression-type (fully convolutional) network have exhibited great potential in various image fusion issues. From our perspective, CNN-based image fusion research will still be an active topic in the future. A generic framework for CNN-based general image fusion is shown in Fig. 5.

The basic idea of classification-type CNN-based image fusion scheme is consistent with the work [6]. The convolutional network is used to generate a weight map directly from the source images, which avoids manually designing complicated activity level measurements and fusion rules. Specifically, each output neuron of the network represents a normalized weight assignment result (the sum of all the weights equals to 1), and its value indicates the corresponding possibility. Accordingly, the output vector of the network essentially denotes the probability distribution of weight assignment, and its mathematical expectation is the weight assignment to be calculated. In fact, the situation in [6] is just a special case that the output vector has two components, denoting two weight assignments "1-0" and "0-1", respectively. Therefore, in addition to multi-focus image fusion which can be viewed as a two-class classification problem, this approach may also make sense to some other image fusion issues such as multi-modal image fusion, since the more generalized multi-class output manner provides larger flexibility. Furthermore, besides the output weight map, the intermediate feature maps also contain the clarity information of source images, which is of great significance to various image fusion issues. In our opinion, to pursue high fusion performance, the classical image fusion techniques (e.g., multi-resolution analysis, consistency verification, etc) achieved in conventional research should not be ignored. On the contrary, according to the characteristics of a specific image fusion issue, they should be well adopted to design a complete fusion scheme together with the CNN-based approach. The input manner of the network is also not limited to the siamese architecture used in [6], while the effectiveness of other architectures like pseudo-siamese and 2-channel is deserved to be studied. In this kind of methods, one of the most challenging issues is how to create an effective training dataset. A possible way is introduced in [6] based on multi-scale Gaussian filtering, while further exploration is definitely needed to produce more elaborate strategies for more complex CNN models.

The regression-type convolutional networks can be used to learn an end-to-end mapping from the source images to the fused image. However, as a ground truth fused image generally not exists in an image fusion task, one main challenge in this category of methods is the generation of training examples. Fortunately, for some image fusion issues in which the source images are captured by the same imaging

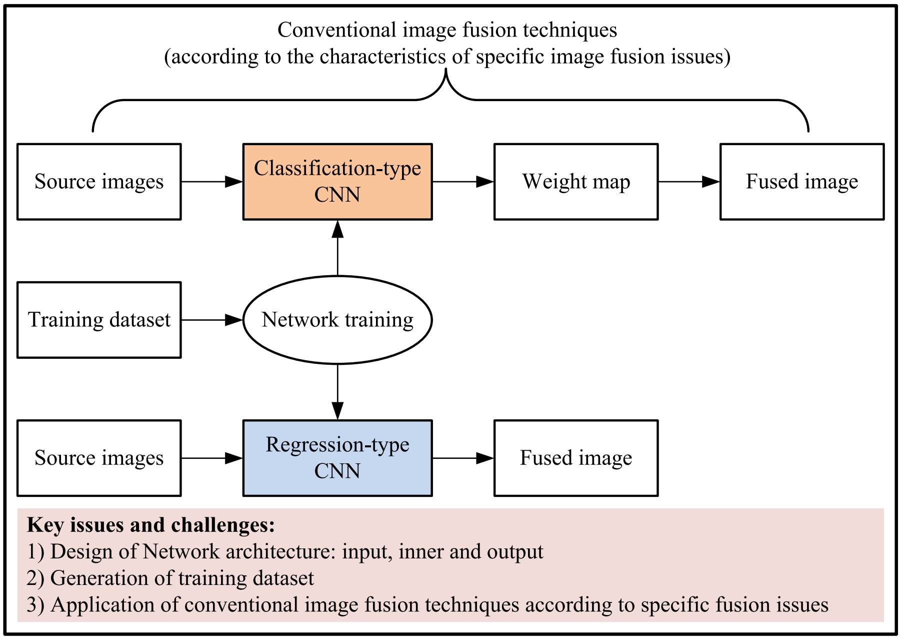

Fig. 5. A generic framework for CNN-based general image fusion.

modality, such as multi-exposure images and multi-focus images, it is possible to artificially create the ground truth for CNN training. A successful case has been introduced in multi-exposure fusion [9] by creating a dataset in which each example consists of three LDR images and the corresponding ground truth HDR image. This category of methods has the potential to overcome an open-ended challenge that exists in digital photography related fusion problems [2], namely, the effects of moving objects during different captures. By modeling the complex process of preventing motion/ghosting artifacts via a deep network, the CNN-based methods can avoid the difficulty of manually designing complicated ghost-removal strategies in conventional methods and are more likely to obtain better performance. In the future, much more efforts can be made in this direction.

Another future trend in this field is to derive effective image representation approaches from the DL community for image fusion. As an example, CSR has been successfully applied to image fusion [11], but it is just a preliminary work as many aspects require further investigation. Considering the great success achieved in the SR-based image fusion study, a promising future of the CSR-based image fusion is also expected. Fig. 6 shows a generic framework for CSR-based image fusion. First, a certain image transform (No transform performed is viewed as a special case) is performed on the source images. Then, the convolutional sparse coding is performed on some selected transformed bands via a set of off-line learned dictionary filters. Next, a fusion strategy including activity level measurement and fusion rule is designed to obtain the fused bands. More concretely, this step can accomplished with two different manners, just as the SR-based methods [23]. One approach is combining the source sparse coefficient maps firstly, and then the fused bands are reconstructed using the dictionary filters and merged coefficient maps. The other way is constructing the fused bands directly from the source bands according to the activity level measure calculated by the source sparse coefficients. Finally, the fused image is reconstructed by performing the corresponding inverse transform. The selections of image transform and fusion manner both depend on specific fusion problems. In the dictionary learning phase, the generation of training examples and the setting of related learning parameters are two key issues, which have significant influences on the fusion performance.

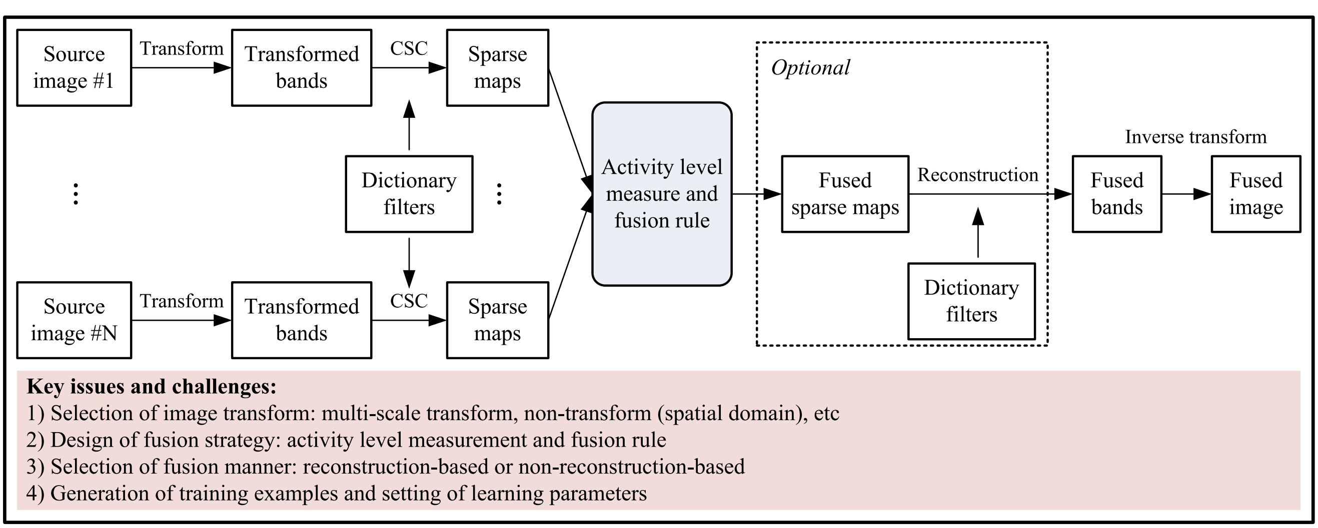

Fig. 6. A generic framework for CSR-based image fusion.

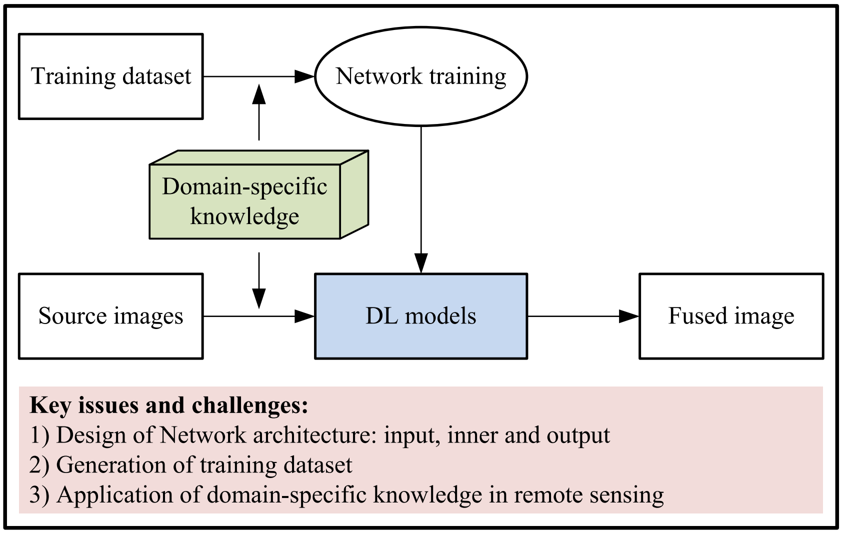

Fig. 7. A generic framework for DL-based remote sensing image fusion.

5.2. Remote sensing image fusion methods

Various SAE-based and CNN-based DL models have been applied to remote sensing image fusion. When compared with general image fusion, an advantage of DL for remote sensing image fusion is that the ground truth images for supervised learning are much easier to be created using Wald's protocol [107], which brings great convenience to related study.

Fig. 7 shows a generic framework for DL-based remote sensing image fusion. The design of network architecture will still be the focus in the future. According to the specific fusion tasks, such as MS/PAN fusion, HS/PAN fusion and MS/HS fusion, the issues including input manner, inner structure and output manner all need to be carefully considered. In particular, the underlying information contained across multiple neighbouring spectral bands may be an important clue for the fusion process. Therefore, some of the learning models for video or image sequence feature extraction may be effective in remote sensing image fusion, such as 3D-CNNs and recurrent convolutional networks. To improve the representation ability of applied networks, the residual learning may be preferred since it can lead to much deeper architectures.

Another factor that is worth paying attention to is the application of domain-specific knowledge in remote sensing imagery, as this is expected to improve the fusion performance of original models. A successful attempt that augments the network input with some class-specific radiometric indices has been carried in [15]. In the future, further exploration in this direction can be conducted to design more effective network architectures and fusion schemes.

5.3. Objective evaluation metrics

The objective assessment of image fusion is typically based on comparing some kinds of local similarities between the fused image and each source image. Although there is currently no published work about DL-based objective evaluation for image fusion issues, the CNN-based approach [118] for patch similarity measure motivates us to put forward a potential framework for fusion metric design based on CNNs, as shown in Fig. 8. The basic idea involves applying a classification-type CNN to measure a certain similarity of two image patches and each component of the network output denotes a pre-defined similarity score. The range of score can be defined as [0, 1] while the score 1 is obtained only when the two input patches are identical. Thus, the output vector of the network denotes the probability distribution of the measurement score, so its mathematical expectation is exactly equal to the score to be determined.

The generation of training dataset is one of the most crucial issues in this scheme. One possible solution is based on adding a random noise into a patch to create a training example which is composed of the original and modified patches. Different types and intensities of noises should be involved to ensure the variety of training set. For each example, a score is defined according to a pre-designed quantization strategy. Because the meaning of an evaluation metric is mostly given by the relative values rather than absolute values, this artificial assignment manner is likely to make sense. Once training is done, each fully-connected layer in the network is converted into its corresponding convolutional version so that the network can accept input images of arbitrary size.

In the testing phase, the target features of each source image and the fused image obtained by some pre-processing techniques (optional) are fed to the trained network to generate a score map. Each coefficient in the map contains the similarity score of a corresponding patch pair in the input images. The overall score of the two input images is the average value of all the coefficients in the score map. The final objective evaluation result is computed by averaging the scores obtained from all the source images.

It is worth mentioning that the above scheme only provides an idea for the future study in this field. The effectiveness of this scheme should be further verified with specific implementations and comprehensive experiments, but this goes outside the scope of this paper.

5.4. Summary

It can be observed from the above discussions that there exist several common challenges in DL-based image fusion research, such as the design of network architecture for a specific fusion task, the generation of dataset for network training (a practical difficulty in current DL-based image fusion is the lack of specific large-scale dataset), the application of domain knowledge or conventional fusion techniques according to characteristics of specific image fusion issues, etc. In practice, these issues are closely associated with each other and should be jointly considered. In addition to conducting further study on the frameworks mention above, it is also of great significance to develop more DL-based schemes for various image fusion tasks in the future.

6. Conclusions

The application of DL-based techniques to pixel-level image fusion has been progressing at a fast rate in recent years. This paper reviews the recent advances achieved in DL-based image fusion and puts forward some prospects for future study in this field. The primary

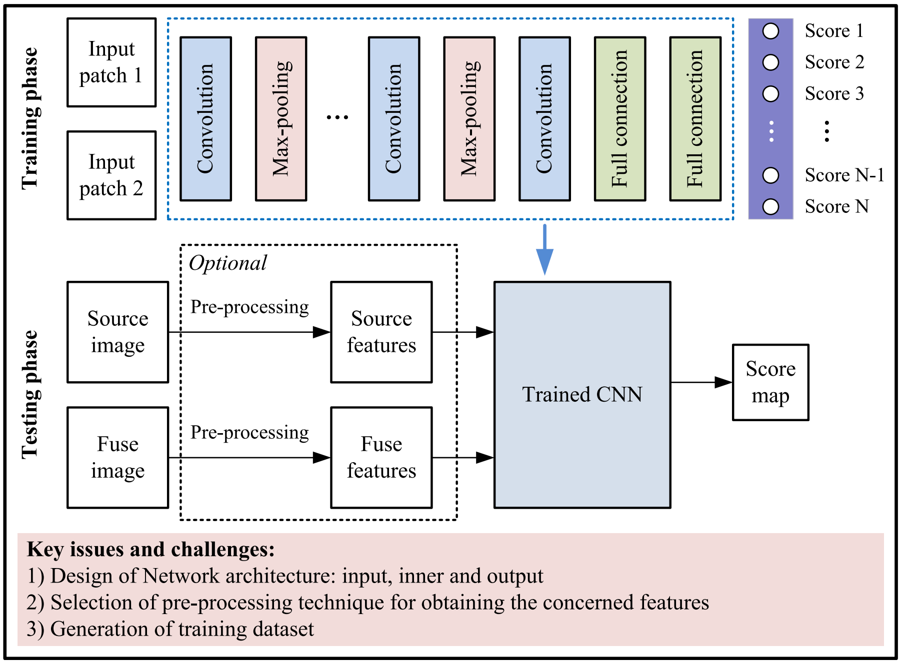

Fig. 8. A potential CNN-based framework for the design of image fusion metrics.

contributions of this work can be summarized as the following three points.

1. The difficulties that exist in conventional image fusion research are analyzed and the advantages of DL techniques for image fusion are discussed (Section 2). The relevant difficulties are summarized into four specific points in terms of fusion methods and objective evaluation metrics. For each point, the advantages of DL-based approaches are discussed.
2. A thorough overview about the current achievements in DL-based image fusion is conducted (Section 4). The survey covers more than a dozen recently proposed image fusion methods that are based on DL techniques including CNNs, CSR and SAEs. The basic ideas, main steps, specific applications and major characteristics of these methods are introduced.
3. Some prospects for the future study of DL-based image fusion are put forward (Section 5). Several generic DL-based frameworks for developing general image fusion methods, remote sensing image fusion methods and objective evaluation metrics are summarized and presented. The key issues and challenges that exist in each framework are discussed.

In conclusion, the recent progress achieved in DL-based image fusion exhibits a promising trend in the field of image fusion with a huge potential for future improvement. It is highly expected that more related researches would continue in the coming years to promote the development of image fusion.

Acknowledgements

The authors would like to thank the editors and anonymous reviewers for their insightful comments and constructive suggestions. This work is supported by the National Natural Science Foundation of China (Grants 61701160, 61501164 and 81571760) and the Fundamental Research Funds for the Central Universities (Grants JZ2017HGTA0176, JZ2016HGBZ1025 and JZ2016HGPA0731).

References

[1] A. Goshtasby, S. Nikolov, Image fusion: advances in the state of the art, Inf. Fusion 8 (2) (2007) 114–118.
[2] S. Li, X. Kang, L. Fang, J. Hu, H. Yin, Pixel-level image fusion: a survey of the state of the art, Inf. Fusion 33 (2017) 100–112.
[3] A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks, Proceedings of Advances in Neural Information Processing Systems (NIPS), (2012), pp. 1097–1105.
[4] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015), pp. 3431–3440.
[5] C. Dong, C. Loy, K. He, X. Tang, Image super-resolution using deep convolutional networks, IEEE Trans. Pattern Anal. Mach. Intell. 38 (2) (2016) 295–307.
[6] Y. Liu, X. Chen, H. Peng, Z. Wang, Multi-focus image fusion with a deep convolutional neural network, Inf. Fusion 36 (2017) 191–207.
[7] B. Yang, J. Zhong, Y. Li, Z. Chen, Multi-focus image fusion and super-resolution with convolutional neural network, Int. J. Wavelets Multiresolut. Inf. Process. 15 (4) (2017) 1750037:1–15.
[8] C. Du, S. Gao, Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network, IEEE Access 5 (2017) 15750–15761.
[9] N. Kalantari, R. Ramamoorthi, Deep high dynamic range imaging of dynamic scenes, ACM Trans. Graph. 36 (4) (2017) 144:1–12.
[10] J. Zhong, B. Yang, Y. Li, F. Zhong, Z. Chen, Image fusion and super-resolution with convolutional neural network, Proceedings of Chinese Conference on Pattern Recognition, (2016), pp. 78–88.
[11] Y. Liu, X. Chen, R. Ward, Z. Wang, Image fusion with convolutional sparse representation, IEEE Signal Process. Lett. 23 (12) (2016) 1882–1886.
[12] Y. Liu, X. Chen, J. Cheng, H. Peng, A medical image fusion method based on convolutional neural networks, Proceedings of 20th International Conference on Information Fusion, (2017), pp. 1–7.
[13] W. Huang, L. Xiao, Z. Wei, H. Liu, S. Tang, A new pan-sharpening method with deep neural networks, IEEE Geosci. Remote Sens. Lett. 12 (5) (2015) 1037–1041.
[14] J. Zhong, B. Yang, G. Huang, F. Zhong, Z. Chen, Remote sensing image fusion with convolutional neural network, Sensing Imaging 17 (2016) 10:1–16.
[15] G. Masi, D. Cozzolino, L. Verdoliva, G. Scarpa, Pansharpening by convolutional neural networks, Remote Sens. 8 (2016) 594:1–22.
[16] A. Azarang, H. Ghassemian, A new pansharpening method using multiresolution analysis framework and deep neural networks, Proceedings of 3rd International Conference on Pattern Recognition and Image Analysis, (2017), pp. 1–5.
[17] Y. Rao, L. He, J. Zhu, A residual convolutional neural network for pan-shaprening, Proceedings of International Workshop on Remote Sensing with Intelligent Processing, (2017), pp. 1–4.
[18] Y. Wei, Q. Yuan, H. Shen, L. Zhang, Boosting the accuracy of multispectral image pansharpening by learning a deep residual network, IEEE Geosci. Remote Sens. Lett. 14 (10) (2017) 1795–1799.
[19] F. Palsson, J. Sveinsson, M. Ulfarsson, Multispectral and hyperspectral image fusion using a 3-d-convolutional neural network, IEEE Geosci. Remote Sens. Lett. 14 (5) (2017) 639–643.
[20] Z. Zhang, R. Blum, A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application, Proc. IEEE 87 (8) (1999) 1315–1326.
[21] G. Piella, A general framework for multiresolution image fusion: from pixels to regions, Inf. Fusion 4 (4) (2003) 259–280.
[22] Z. Liu, E. Blasch, Z. Xue, J. Zhao, R. Laganiere, W. Wu, Objective assessment of multiresolution image fusion algorithms for context enhancement in night vision:

a comparative study, IEEE Trans. Pattern Anal. Mach. Intell. 34 (1) (2012) 94–109.
[23] Q. Zhang, Y. Liu, R. Blum, J. Han, D. Tao, Sparse representation based multi-sensor image fusion for multi-focus and multi-modality images: a review, Inf. Fusion 40 (2018) 57–75.
[24] A. James, B. Dasarathy, Medical image fusion: a survey of the state of the art, Inf. Fusion 19 (2014) 4–19.
[25] J. Du, W. Li, K. Lu, B. Xiao, An overview of multi-modal medical image fusion, Neurocomputing 215 (2016) 3–20.
[26] G. Vivone, L. Alparone, J. Chanussot, M. DallaMura, A. Garzelli, G. Licciardi, R. Restaino, L. Wald, A critical comparison among pansharpening algorithms, IEEE Trans. Geosci. Remote Sens. 53 (5) (2015) 2565–2586.
[27] H. Ghassemian, A review of remote sensing image fusion methods, Inf. Fusion 32 (2016) 75–89.
[28] X. Jin, Q. Jiang, S. Yao, D. Zhou, R. Nie, J. Hai, K. He, A survey of infrared and visual image fusion methods, Infrared Phys. Technol. 85 (2017) 478–501.
[29] A. Toet, Image fusion by a ratio of low pass pyramid, Pattern Recognit. Lett. 9 (4) (1989) 245–253.
[30] P. Burt, R. Kolcznski, Enhanced image capture through fusion, Proceedings of the 4th IEEE Interational Coference on Computer Vision, (1993), pp. 173–182.
[31] V. Petrovic, C. Xydeas, Gradient-based multiresolution image fusion, IEEE Trans. Image Process. 13 (2) (2004) 228–237.
[32] T. Mertens, J. Kautz, F.V. Reeth, Exposure fusion, Proc. Pacific Graphics, (2007), pp. 382–390.
[33] H. Li, B. Manjunath, S. Mitra, Multisensor image fusion using the wavelet transform, Graph. Models Image Process. 57 (3) (1995) 235–245.
[34] G. Pajares, J.M. de la Cruz, A wavelet-based image fusion tutorial, Pattern Recognit. 37 (9) (2004) 1855–1872.
[35] J. Lewis, R. OCallaghan, S. Nikolov, D. Bull, N. Canagarajah, Pixel- and region-based image fusion with complex wavelets, Inf. Fusion 8 (2) (2007) 119–130.
[36] L. Yang, B. Guo, W. Ni, Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform, Neurocomputing 72 (2008) 203–211.
[37] Q. Zhang, B. Guo, Multifocus image fusion using the nonsubsampled contourlet transform, Signal Process. 89 (7) (2009) 1334–1346.
[38] X. Qu, J. Yan, H. Xiao, Z. Zhu, Image fusion algorithm based on spatial frequency-motivated pulse coupled neural networks in nonsubsampled contourlet transform domain, Acta Automatic Sinica 34 (12) (2008) 1508–1514.
[39] G. Gao, L. Xu, D. Feng, Multi-focus image fusion based on non-subsampled shearlet transform, IET Image Proc. 7 (6) (2013) 633–639.
[40] Z. Farbman, R. Fattal, D. Lischinski, R. Szeliski, Edge-preserving decompositions for multi-scale tone and detail manipulation, ACM Trans. Graph. 27 (3) (2008) 67:1–10.
[41] J. Hu, S. Li, The multiscale directional bilateral filter and its application to multisensor image fusion, Inf. Fusion 13 (2012) 196–206.
[42] S. Li, X. Kang, J. Hu, Image fusion with guided filtering, IEEE Trans. Image Process. 22 (7) (2013) 2864–2875.
[43] H. Zhao, Z. Shang, Y. Tang, B. Fang, Multi-focus image fusion based on the neighbor distance, Pattern Recognit. 46 (3) (2013) 1002–1011.
[44] Z. Zhou, B. Wang, S. Li, M. Dong, Perceptual fusion of infrared and visible images through a hybrid multi-scale decomposition with gaussian and bilateral filters, Inf. Fusion 30 (2016) 15–26.
[45] B. Yang, S. Li, Multifocus image fusion and restoration with sparse representation, IEEE Trans. Instrum. Meas. 59 (4) (2010) 884–892.
[46] N. Yu, T. Qiu, F. Bi, A. Wang, Image features extraction and fusion based on joint sparse representation, IEEE J. Sel. Top Signal Process. 5 (5) (2011) 1074–1082.
[47] B. Yang, S. Li, Pixel-level image fusion with simultaneous orthogonal matching pursuit, Inf. Fusion 13 (1) (2012) 10–19.
[48] S. Li, H. Yin, L. Fang, Group-sparse representation with dictionary learning for medical image denoising and fusion, IEEE Trans. Biomed. Eng. 59 (2012) 3450–3459.
[49] J. Wang, H. Liu, N. He, Exposure fusion based on sparse representation using approximate k-svd, Neurocomputing 135 (2014) 145–154.
[50] Y. Liu, Z. Wang, Simultaneous image fusion and denosing with adaptive sparse representation, IET Image Proc. 9 (5) (2015) 347–357.
[51] M. Nejati, S. Samavi, S. Shirani, Multi-focus image fusion using dictionary-based sparse representation, Inf. Fusion 25 (1) (2015) 72–84.
[52] M. Kim, D.K. Han, H. Ko, Joint patch clustering-based dictionary learning for multimodal image fusion, Inf. Fusion 27 (2016) 198–214.
[53] Q. Zhang, M. Levine, Robust multi-focus image fusion using multi-task sparse representation and spatial context, IEEE Trans. Image Process. 25 (5) (2016) 2045–2058.
[54] S. Li, J. Kwok, Y. Wang, Combination of images with diverse focuses using the spatial frequency, Inf. Fusion 2 (3) (2001) 169–176.
[55] S. Li, J. Kwok, Y. Wang, Multifocus image fusion using artificial neural networks, Pattern Recognit Lett 23 (8) (2002) 985–997.
[56] A. Goshtasby, Fusion of multi-exposure images, Image Vision Comput. 23 (6) (2005) 611–618.
[57] J. Kong, K. Zheng, X. Feng, Multi-focus image fusion using spatial frequency and genetic algorithm, Int. J. Comput. Sci. Network Secur. 8 (2) (2008) 220–224.
[58] V. Aslantas, R. Kurban, Fusion of multi-focus images using differential evolution algorithm, Expert Syst. Appl. 37 (12) (2010) 8861–8870.
[59] I. De, B. Chanda, Multi-focus image fusion using a morphology-based focus measure in a quad-tree structure, Inf. Fusion 14 (2) (2013) 136–146.
[60] X. Bai, Y. Zhang, F. Zhou, B. Xue, Quadtree-based multi-focus image fusion using a weighted focus-measure, Inf. Fusion 22 (1) (2015) 105–118.
[61] M. Li, W. Cai, Z. Tan, A region-based multi-sensor image fusion scheme using pulse-coupled neural network, Pattern Recognit. Lett. 27 (16) (2006) 1948–1956.
[62] S. Li, B. Yang, Multifocus image fusion using region segmentation and spatial frequency, Image Vis. Comput. 26 (7) (2008) 971–979.
[63] R. Shen, I. Cheng, J. Shi, A. Basu, Generalized random walks for fusion of multi-exposure images, IEEE Trans. Image Process. 20 (12) (2011) 3634–3646.
[64] S. Li, X. Kang, Fast multi-exposure image fusion with median filter and recursive filter, IEEE Trans. Consum. Electron. 58 (2) (2012) 626–632.
[65] W. Zhang, W.-K. Cham, Gradient-directed multiexposure composition, IEEE Trans. Image Process. 21 (4) (2012) 2318–2323.
[66] S. Li, X. Kang, J. Hu, B. Y, Image matting for fusion of multi-focus images in dynamic scenes, Inf. Fusion 14 (2) (2013) 147–162.
[67] Y. Liu, S. Liu, Z. Wang, Multi-focus image fusion with dense sift, Inf. Fusion 23 (1) (2015) 139–155.
[68] D. Guo, J. Yan, X. Qu, High quality multi-focus image fusion using self-similarity and depth information, Opt. Commun. 338 (1) (2015) 138–144.
[69] Y. Liu, Z. Wang, Dense sift for ghost-free multi-exposure fusion, J. Vis. Commun. Image Represent. 31 (2015) 208–224.
[70] J. Ma, C. Chen, C. Li, J. Huang, Infrared and visible image fusion via gradient transfer and total variation minimization, Inf. Fusion 31 (2016) 100–109.
[71] Y. Zhang, X. Bai, T. Wang, Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure, Inf. Fusion 35 (2017) 81–101.
[72] S. Li, B. Yang, Multifocus image fusion by combining curvelet and wavelet transform, Pattern Recognit. Lett. 29 (9) (2008) 1295–1301.
[73] S. Li, B. Yang, Hybrid multiresolution method for multisensor multimodal image fusion, IEEE Sens. J. 10 (9) (2010) 1519–1526.
[74] Y. Liu, S. Liu, Z. Wang, A general framework for image fusion based on multi-scale transform and sparse representation, Inf. Fusion 24 (1) (2015) 147–164.
[75] S. Li, J. Kwok, I. Tsang, Y. Wang, Fusing images with different focuses using support vector machines, IEEE Trans. Neural Networks 15 (6) (2004) 1555–1561.
[76] Y. Yang, M. Yang, S. Huang, Y. Que, M. Ding, J. Sun, Multifocus image fusion based on extreme learning machine and human visual system, IEEE Access 5 (2017) 6989–7000.
[77] T. Tu, S. Su, H. Shyu, P. Huang, A new look at ihs-like image fusion methods, Inf. Fusion 2 (3) (2001) 177–186.
[78] S. Rahmani, M. Strait, D. Merkurjev, M. Moeller, T. Wittman, An adaptive ihs pan-sharpening method, IEEE Geosci. Remote Sens. Lett. 7 (4) (2010) 746–750.
[79] H.R. Shahdoosti, H. Ghassemian, Combining the spectral pca and spatial pca fusion methods by an optimal filter, Inf. Fusion 27 (2016) 150–160.
[80] J. Nunez, X. Otazu, O. Fors, A. Prades, V. Pala, R. Arbiol, Multiresolution-based image fusion with additive wavelet decomposition, IEEE Trans. Geosci. Remote Sens. 37 (4) (1999) 1204–1211.
[81] K. Amolins, Y. Zhang, P. Dare, Wavelet based image fusion techniques: an introduction, review and comparison, ISPRS J. Photogramm. Remote Sens. 62 (4) (2007) 249–263.
[82] V. Shah, N. Younan, R. King, An efficient pan-sharpening method via a combined adaptive pca approach and contourlets, IEEE Trans. Geosci. Remote Sens. 46 (2008) 1323–1335.
[83] Z. Li, H. Leung, Fusion of multispectral and panchromatic images using a restoration-based method, IEEE Trans. Geosci. Remote Sens. 47 (5) (2009) 1482–1491.
[84] M. Joshi, A. Jalobeanu, Map estimation for multiresolution fusion in remotely sensed images using an igmrf prior model, IEEE Trans. Geosci. Remote Sens. 48 (3) (2010) 1245–1255.
[85] M. Xu, H. Chen, P. Varshney, An image fusion approach based on markov random fields, IEEE Trans. Geosci. Remote Sens. 49 (12) (2011) 5116–5127.
[86] F. Fang, F. Li, C. Shen, G. Zhang, A variational approach for pan-sharpening, IEEE Trans. Image Process. 22 (7) (2013) 2822–2834.
[87] F. Palsson, J. Sveinsson, M. Ulfarsson, A new pansharpening algorithm based on total variation, IEEE Geosci. Remote Sens. Lett. 11 (1) (2014) 318–322.
[88] S. Li, B. Yang, A new pan-sharpening method using a compressed sensing technique, IEEE Trans. Geosci. Remote Sens. 49 (2) (2011) 738–746.
[89] M. Ghahremani, H. Ghassemian, A compressed-sensing-based pan-sharpening method for spectral distortion reduction, IEEE Trans. Geosci. Remote Sens. 54 (4) (2016) 2194–2206.
[90] J. Yang, T.H. J. Wright, Y. Ma, Image super-resolution via sparse representation, IEEE Trans. Image Process. 19 (11) (2010) 2861–2873.
[91] S. Li, H. Yin, L. Fang, Remote sensing image fusion via sparse representations over learned dictionaries, IEEE Trans. Geosci. Remote Sens. 51 (9) (2013) 4779–4789.
[92] X. Zhu, R. Bamler, A sparse image fusion algorithm with application to pan-sharpening, IEEE Trans. Geosci. Remote Sens. 51 (5) (2013) 2827–2836.
[93] C. Jiang, H. Zhang, H. Shen, L. Zhang, Two-step sparse coding for the pan-sharpening of remote sensing images, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 7 (5) (2014) 1792–1805.
[94] G. Qu, D. Zhang, P. Yan, Information measure for performance of image fusion, Electron. Lett. 38 (7) (2002) 313–315.
[95] M. Hossny, S. Nahavandi, D. Creighton, Comments on information measure for performance of image fusion, Electron. Lett. 44 (18) (2008) 1066–1067.
[96] Q. Wang, Y. Shen, A nonlinear correlation measure for multivariable data set, Physica D 200 (3) (2005) 287–295.
[97] M. Hossny, S. Nahavandi, D. Creighton, A. Bhatti, Image fusion performance metric based on mutual information and entropy driven quadtree decomposition, Electron. Lett. 46 (18) (2010) 1266–1268.
[98] C.S. Xydeas, V.S. Petrovic, Objective image fusion performance measure, Electron. Lett. 36 (4) (2000) 308–309.
[99] J. Zhao, R. Laganiere, Z. Liu, Performance assessment of combinative pixel-level image fusion based on an absolute feature measurement, Int. J. Innovat. Comput., Inf. Control 6 (A) (3) (2007) 1433–1447.

[100] Y. Zheng, E. Essock, B. Hansen, A. Haun, A new metric based on extended spatial frequency and its application to dwt based fusion algorithms, Inf. Fusion 8 (2) (2007) 177–192.
[101] G. Piella, H. Heijmans, A new quality metric for image fusion, Proceedings of 10th IEEE International Conference on Image Processing (ICIP), (2003), pp. 173–176.
[102] N. Cvejic, A. Loza, D. Bull, N. Canagarajah, A similarity metric for assessment of image fusion algorithms, Int. J. Signal Process. 2 (3) (2005) 178–182.
[103] C. Yang, J. Zhang, X. Wang, X. Liu, A novel similarity based quality metric for image fusion, Inf. Fusion 9 (2) (2008) 156–160.
[104] H. Chen, P. Varshney, A human perception inspired quality metric for image fusion based on regional information, Inf. Fusion 8 (2007) 193–207.
[105] Y. Chen, R. Blum, A new automated quality assessment algorithm for image fusion, Image Vis. Comput. 27 (10) (2009) 1421–1432.
[106] Y. Han, Y. Cai, Y. Cao, X. Xu, A new image fusion performance metric based on visual information fidelity, Inf. Fusion 14 (2013) 127–135.
[107] L. Wald, T. Ranchin, M. Mangolini, Fusion of satellite images of different spatial resolutions: assessing the quality of resulting images, Photogramm. Eng. Remote Sens. 63 (6) (1997) 691–699.
[108] R. Yuhas, A. Goetz, J. Boardman, Discrimination among semi-arid landscape endmembers using the spectral angle mapper (sam) algorithm, Proceedings of 4th JPL Airborne Earth Science Workshop, (1992), pp. 147–149.
[109] L. Wald, Quality of high resolution synthesized images: is there a simple criterion? Proceedings of Third Conference Fusion of Earth Data: Merging Point Measurements, Raster Maps and Remotely Sensed Images, SEE/URISCA, (2000), pp. 99–103.
[110] L. Alparone, S. Baronti, A. Garzelli, F. Nencini, A global quality measurement of pan-sharpened multispectral imagery, IEEE Geosci. Remote Sens. Lett. 1 (4) (2004) 313–317.
[111] L. Alparone, B. Aiazzi, S. Baronti, A. Garzelli, F. Nencini, M. Selva, Multispectral and panchromatic data fusion assessment without reference, Photogramm. Eng. Remote Sens. 74 (2) (2008) 193–200.
[112] Z. Liu, E. Blasch, V. John, Statistical comparison of image fusion algorithms: recommendations, Inf, Fusion 36 (2017) 251–260.
[113] M. Zeiler, D. Krishnan, G. Taylor, R. Fergus, Deconvolutional networks, Proceedings of IEEE Conference Computer Vision and Pattern Recognition (CVPR), (2010), pp. 2528–2535.
[114] B. Wohlberg, Efficient algorithms for convolutional sparse representation, IEEE Trans. Image Process. 25 (1) (2016) 301–315.
[115] J. Kim, J. Lee, K. Lee, Accurate image super-resolution using very deep convolutional networks, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), pp. 1646–1654.
[116] Y. Huang, W. Wang, L. Wang, Bidirectional recurrent convolutional networks for multi-frame super-resolution, Proceedings of Advances in Neural Information Processing Systems (NIPS), (2015), pp. 235–243.
[117] D. Li, Z. Wang, Video super-resolution via motion com-pensation and deep residual learning, IEEE Trans. Comput. Imaging DOI: 10.1109/TCI.2017.2671360 (2017) 1–15.
[118] S. Zagoruyko, N. Komodakis, Learning to compare image patches via convolutional neural networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, (2015), pp. 4353–4361.
[119] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based leaning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[120] D. Tran, L. Bourdev, R. Fergus, L. Torresani, M. Paluri, Learning spatiotemporal features with 3d convolutional networks, Proceedings of International Conference on Computer Vision (ICCV), (2015), pp. 4489–4497.
[121] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, Y. L, Overfeat: integrated recognition, localizaton and detection using convolutional networks, arXiv:1312.6299v4 (2014) 1–16.
[122] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2016), pp. 770–778.
[123] V. Papyan, Y. Romano, M. Elad, Convolutional neural networks analyzed via convolutional sparse coding, arXiv:1607.08194v4 (2016) 1–50.
[124] B. Wohlberg, Endogenous convolutional sparse representations for translation invariant image subspace models, Proceedings of IEEE International Conference on Image Processing (ICIP), (2014), pp. 2859–2863.
[125] D. Carrera, G. Boracchi, A. Foi, B. Wohlberg, Detecting anomalous structures by convolutional sparse models, Proceedings of International Joint Conference on Neural Networks (IJCNN), (2015), pp. 1–8.
[126] S. Gu, W. Zuo, Q. Xie, D. Meng, X. Feng, L. Zhang, Convolutional sparse coding for image super-resolution, Proceedings of International Conference on Computer Vision (ICCV), (2015), pp. 1823–1831.
[127] H. Bristow, A. Eriksson, S. Lucey, Fast convolutional sparse coding, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2013), pp. 391–398.
[128] F. Heide, W. Heidrich, G. Wetzstein, Fast and flexible convolutional sparse coding, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), (2015), pp. 5135–5143.
[129] http://brendt.wohlberg.net/software/SPORCO/.
[130] http://www.image-net.org/challenges/LSVRC/.
[131] http://cseweb.ucsd.edu/%7eviscomp/projects/SIG17HDR/.
[132] http://mmlab.ie.cuhk.edu.hk/projects/SRCNN.html.
[133] Y. Bengio, P. Lamblin, D. Popovici, H. Larochelle, Greedy layer-wise training of deep networks, Proceedings of Advances in Neural Information Processing Systems (NIPS), (2007), pp. 153–160.
[134] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
[135] J. Xie, L. Xu, E. Chen, Image denoising and inpainting with deep neural networks, Proceedings of Advances in Neural Information Processing Systems (NIPS), (2012), pp. 341–349.
[136] http://www.escience.cn/people/liuyu1/Codes.html.
[137] http://www.grip.unina.it/research/85-image-enhancement/93-pnn.html.