Multi-focus Image Fusion using Fully Convolutional Two-stream Network for Visual Sensors

Kaiping Xu, Zheng Qin, Guolong Wang, Huidi Zhang, Kai Huang and Shuxiong Ye
School of Software, Tsinghua University
Beijing, China
[e-mail: xu_kaiping@hotmail.com]
*Corresponding author: Zheng Qin

Received May 4, 2017; revised September 19, 2017; accepted November 24, 2017; published May 31, 2018

Abstract

We propose a deep learning method for multi-focus image fusion. Unlike most existing pixel-level fusion methods, either in spatial domain or in transform domain, our method directly learns an end-to-end fully convolutional two-stream network. The framework maps a pair of different focus images to a clean version, with a chain of convolutional layers, fusion layer and deconvolutional layers. Our deep fusion model has advantages of efficiency and robustness, yet demonstrates state-of-art fusion quality. We explore different parameter settings to achieve trade-offs between performance and speed. Moreover, the experiment results on our training dataset show that our network can achieve good performance with subjective visual perception and objective assessment metrics.

Keywords: multi-focus image fusion, fully convolutional neural network, deconvolution decoder, convolution fusion

http://doi.org/10.3837/tiis.2018.05.019 **ISSN : 1976-7277**

1. Introduction

Multi-focus image fusion is an important branch of multi-sensor image fusion. It is mainly employed in fusing substantial information from multiple images of a same scene to generate a clear composite image. Multi-focus image fusion can overcome the diversities and limitations of single sensor in spatial resolution, geometry, and spectrum, thus enhance the reliability of image processing tasks, such as feature extraction, edge detection, object recognition and image segmentation. Currently, multi-focus image fusion technology has a wide range of applications in transportation, medical imaging, military operations and machine vision [1]. In each application, the key task of image fusion is to find the accurate information from the source images. That is, the fused image containing all relevant objects in focus can be obtained by composing the clear regions or pixels. However, it is difficult to determine which regions or pixels are located in focus [2]. To solve this problem, many researchers have proposed information theory for performing fusion.

Generally, pixel, feature and decision levels are three levels of image fusion process [3-4]. Pixel-level fusion deals with pixels obtained from source images directly, which is the lowest level of image fusion and mainly concentrates on visual enhancement. It can preserve the original information in the scene more easily. Advantages of pixel level fusion are low complexity and high accuracy [5-6]. Feature-level fusion performs on features extracted from source images for analysis and processing, which can be support for decision-level fusion. In feature-level, features of images include size, edges, corners and textures [7-8]. Feature-level fusion does not require source images registration strictly. Moreover, only the image feature is processed, thus it is convenient for information compression and data transmission. Decision-level fusion is the highest level of image fusion, aiming to make the best decision with credibility criteria. Decision fusion can be defined as the process of fusing information from several individual data sources after each data source is preprocessed, extracted and classified [9]. In summary, pixel-level image fusion can preserve more detailed information than feature and decision level [10].

Pixel-level image fusion is categorized in two domains: spatial domain and frequency domain [11]. Spatial domain processes regions or pixels to combine relevant information directly with focused regions properties, such as focused pixels detection [46], point spread functions (PSFs) [38] and guided filtering (GF) [12]. In frequency domain, source images are transformed in frequency domain, then frequency coefficients are combined and conducted inverse transform to get clear images by fusion rules, such as non-subsampled contourlet transform (NSCT) [47], non-subsampled shearlet transform(NSST) [48] and discrete cosine transform (DCT) [49].

In Recent years, image fusion approaches are proposed using machine learning (ML) algorithms for the classification of focused image regions. Artificial neural network (ANN) and support vector machine (SVM) based fusion methods are explored with visibility, spatial frequency, and edge features [13-14]. Besides, another efficient variant of ANN, probabilistic neural network (PNN), is developed for image fusion [15]. C. M. Sheela Rani and V Vijayakumar et al.proposed an efficient block based feature level contourlet transform with neural network (BFCN) model for image fusion [16]. All the above mentioned approaches are focused on feature-level or decision-level fusion. Among the state-of-the-art methods, Convolutional Neural Network (CNN) has achieved record-breaking performance in computer vision and image processing tasks, ranging from detection, recognition, tracking to semantic segmentation, denoising and super-resolution. Jain and Seung proposed a novel CNN to denoise original images [17]. C Dong and CC Loy et al. explored a deep convolutional network for image super-resolution [18]. Their framework is the same as Fully Convolutional Neural Network (FCN) for image semantic segmentation [19]. Their networks accept an image as the input and produces an entire image as the output through hidden layers of convolution and deconvolution. The weights are learned by minimizing the difference between the whole-image inputs and corresponding whole-image groundtruths. Some novel image fusion methods based on CNN have been proposed. In [20], an algorithm is presented for both image fusion and super-resolution. The resolution is enhanced with CNN, and fusion rule is also images transformed in frequency domain. In [21], CNN is applied to output a score map and final fused image is obtained with pixel-wise weighted-average strategy. However, above two models are not deep end-to-end networks.

In this work, we investigate a fully convolutional two-stream network framework for pixel-level image fusion. The network consists of a chain of convolutional layers, fusion layer and deconvolutional layers. The inputs of our framework are two source images having different regions in focus, and output is a fused image. The convolution layers perform the feature extraction, which encode the primary contents of the pair of input images. The fusion layer combines the two stream networks with feature maps fusion. Then the deconvolutional layers as decoders process fused feature maps to recover the image content details.

Overall, the contributions of our study are mainly in three aspects:

1) The proposed framework is the first attempt to learn convolutional, fusion and deconvolutional mappings from different focus images to the clean version in an end-to-end network, instead of traditional image fusion schemes.

2) We establish a many-to-one mapping between input source images and output one. This mapping can provide guidance for design of the multi-stream networks structure.

3) We demonstrate that deep learning is useful for multi-focus image fusion, and can achieve good quality and speed.

The paper is organized as follows. We present the idea and architecture of performing image fusion network in section 2. Experimental results and analysis are provided in Section 3, followed by the conclusions in Section 4.

2. Fully convolutional two-stream network for image fusion

The proposed framework is composed of three parts : two-stream network of convolutional layers, fusion layer, and deconvolutional layers. Especially, two stream convolutional networks are combined through the convolutional fusion layer, as shown in **Fig. 1**.

2.1 Architecture

The framework of fully convolutional two-stream network is an encoder-decoder network essentially. Since our network has no full connection layer of standard CNN, the size of input images can be arbitrary[19]. The data in each layer has a three-dimensional array of size $h \times w \times d$, where h and w are spatial dimensions, and d is the channel dimension. The two inputs based on image registration have the same size, with pixel size $h \times w$, and d color channels.

The two-stream convolution network corresponds to feature extractor that transforms the input images to multidimensional feature representation, whereas the deconvolution network is a generator that restores an image from the feature maps fused from the fusion layer. The final output of the network is a clean image in the same size with input multi-focus images.

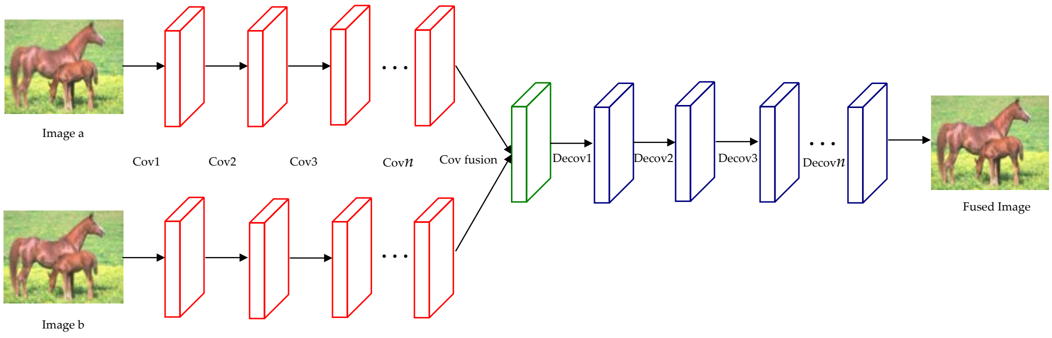


Fig. 1. The overview scheme of fully convolutional two-stream network. ʻCov’ is the short form for ‘convolution’, ‘Cov fusion ’ for ‘convolutional fusion’ and ‘Decov ’ for ‘deconvolution’ .

A main difference between our framework and FCN[19] is that our network includes convolutional, fusion and deconvolutional layers, but no pooling. The reason is that our aim is to fuse pixel-level images while preserving image details instead of learning image features for recognition or detection. Different from high-level applications, pooling typically eliminates the abundant image details and can degrades fusion performance.

2.2 Fusion scheme

In this section, we illustrate a scheme for fusing two stream feature maps. To put in the channel responses at the same pixel position correspondingly, there are two issues to be analyzed, spatial correspondence and channel correspondence. Spatial correspondence is easy to realize by stacking layers from one network on the other when two networks have same spatial resolution at the layers to be fused. By comparison, channel correspondence is relatively difficult, which deals with the correspondence of channel (or channels) in one network with channel (or channels) in the other network.

Concretely, we explore a way to fuse layers between two convolutional networks, and discuss the consequences of correspondence for each network.

$f : \mathrm{x}^a , \mathrm{x}^b$, y refers to the fusion function, which can fuse two feature maps $\mathrm{x}^a , \mathrm{x}^b \in \mathbb{R}^{H \times W \times D}$ and produce an output map $\mathrm{y} \in \mathbb{R}^{H^{'} \times W^{'} \times D^{'}}$, in which *H*, *W* and *D* denote height, width and channel numbers of respective feature maps. Applied at different points in the network, *f* aims to implement multiple layer fusion assuming $H = H^{'}$ and $W = W^{'}$.

Convolutional fusion. Firstly, we define a concatenation function. $\mathrm{y}^{\mathrm{cat}} = f^{\mathrm{cat}}(\mathrm{x}^a , \mathrm{x}^b)$ stacks two feature maps at the same spatial locations *i*, *j* across channel *d*:

$$y^{\mathrm{cat}}_{i,j,2d} = x^a_{i,j,d} \quad y^{\mathrm{cat}}_{i,j,2d-1} = x^b_{i,j,d} , \tag{1}$$

where $\mathrm{y} \in \mathbb{R}^{H \times W \times 2D}$ [22].

Then, Convolutional fusion function is as follows. $\mathrm{y}^{\mathrm{conv}} = f^{\mathrm{conv}}(\mathrm{x}^a , \mathrm{x}^b)$ stacks two feature maps at the same spatial locations *i*, *j* across feature channel *d* as shown in Eq.(1) and subsequently convolves the stacked data with a bank of fusion filters $\mathrm{f} \in \mathbb{R}^{1 \times 1 \times 2 \times D}$ and biases $\mathrm{b} \in \mathbb{R}^{D}$ [22].

$$\mathrm{y}^{\mathrm{conv}} = \mathrm{y}^{\mathrm{cat}} * \mathrm{f} + \mathrm{b} \tag{2}$$

Where D denotes the number of output channels, and the dimension of fusion filter is $1\times 1\times 2$. In this function, f is conducted to reduce the dimensionality from 2D to D and can be used to model weighted combinations of two feature maps x^a, x^b at the same spatial (pixel) location. The fusion feature map is obtained with weighted average of pixel values. When performed as a trainable filter, f learns correspondences of two feature maps that minimize the loss function in the network. The convolutional layers extracts feature patches with filers of 3×3 or 5×5, and each patch is represented as a high-dimensional vector. The convolutional fusion layer not only conducts to fuse two feature maps, but also nonlinearly maps each high-dimensional vector onto another high-dimensional vector. Each mapped vector can represent a high-resolution patch by learning. The non-linear mapping is on 3×3 or 5×5 patch of the feature map, therefore, we set fusion filter size only 1×1 in spatial domain. We apply Rectified Linear Unit (ReLU, max(0,x)) [23] on the filter responses [18].

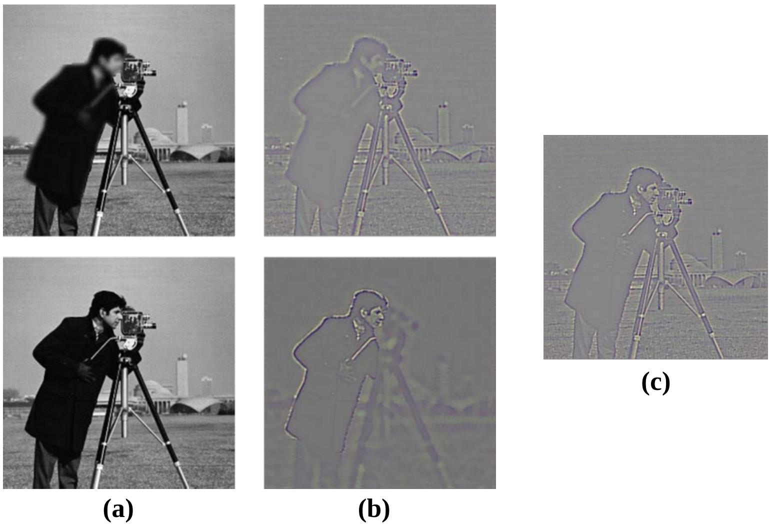


Fig. 2 Illustration of output feature maps convolutional and convolutional fusion layers.
(a) multi-focus source images ;
(b) feature maps of the second convolutional layers;
(c) feature map of convolutional fusion layer.

Fig. 2 shows correspondence feature maps of two source images and obtained convolutional fusion map.

2.3 Deconvlution decoder

The convolution performs as an encoder that maps multiple input activations to a single output activation within a filter window, whereas deconvolution corresponds to a decoder that associates a single input activation with multiple outputs, as show in Fig. 3. Such two operations are completely symmetrical, we need only reverses the forward and backward passes of standard convolutional neural network. Thus, transposed convolution can be performed for end-to-end learning by backpropagation from the pixel-wise loss.

Transposed convolution conducts a transformation going in the opposite direction of a normal convolution, from image that has the shape of the output of some convolution to image that has the shape of its input while maintaining a connectivity pattern that is compatible with said convolution. Our network uses such transformation as the decoding layer of a convolutional encoder to project feature maps to a higher-dimensional space [24].

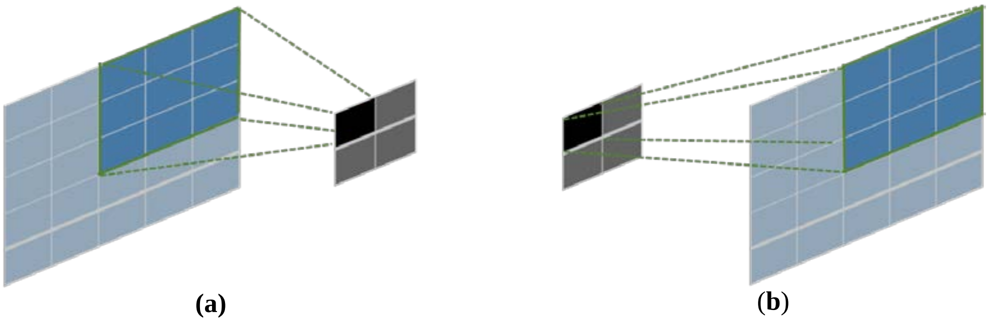


Fig. 3. Illustration of convolution and deconvolution operations. (a) Convolution; (b) Deconvolution.

In proposed model, the deconvolutional network is a key component for image fusion. Contrary to the deconvolution in performed on pixel-wise class probability for object segmentation, our model generates pixel-wise image detail restoration using deep deconvolution network, where a clean image is obtained by successive operations of deconvolution.

For our network, we use padding to make the input sources and corresponding output the same size, and the deconvolution filters are not fixed, but can be learned.

2.4 Network training

There are three types of layers in our network: convolution, fusion and deconvolution. We implement our framework by Caffe [25]. Each layer is followed by a Rectified Linear Unit (ReLU) activation function [23]. Let x be the input, the convolutional, fusion and deconvolutional layers are expressed as:

$$F(\mathrm{x}) = \max(0, w_k * \mathrm{x} + b_k) \tag{3}$$

where w_k and b_k refer to the filters and biases, and $*$ denotes convolution, fusion or deconvolution operation.

Learning the mappings from pairs of blurred images to clean ones needs to update the parameters $\Theta(w_k, b_k)$ represented by the convolutional, fusion and deconvolutional filters with standard back propagation. Specifically, $\{(\mathrm{x}_i^a, \mathrm{x}_i^b), \mathrm{y}_i\}$ refers to a collection of N training sample pairs, where $(\mathrm{x}_i^a, \mathrm{x}_i^b)$ denotes a pair of multi-focus images and y_i denotes the clean image as the groundtruth, and the collection is contained in the training dataset. We minimize the loss function Mean Squared Error (MSE), as followed.

$$L(\Theta) = \frac{1}{N}\sum_{i=1}^{N}\left\| F(\mathrm{x}_i^a, \mathrm{x}_i^b; \Theta) - \mathrm{y}_i \right\|_2^2 \tag{4}$$

The choice of the cost function is appropriate since Peak Signal-to-Noise Ratio (PSNR) is the main evaluation method of image restoration tasks and stands in monotonic relation with MSE. During the training stage, we update the weights and biases with standard back propagation [26-27].

Currently, the optimization of the loss function is dominated by the stochastic gradient descent (SGD) method [28]. Basically, at the t + 1th iteration they update the parameters Θ_{t+1} with the previous parameter update Λ_t and negative gradient $\nabla L(\Theta)$,

$$\Lambda_{\mathrm{t}+1} = a\Lambda_{\mathrm{t}} - b\nabla L(\Theta_t) \tag{5}$$

$$\Theta_{\mathrm{t}+1} = \Theta_t + \Lambda_{t+1} \tag{6}$$

where a, b are the momentum and learning rate, resp. One weakness of SGD is that the improvements gained from the optimization decrease rapidly with growing iteration steps. In such case, SGD may not be able to recover accurate details from blurred images pairs. This is the main reason why we adopt Adam estimation as our optimization method. The Adam method is stated as follows [29],

$$\Lambda_t = a_1\Lambda_{t-1} + (1-a_1)\nabla L(\Theta) \tag{7}$$

$$K_t = a_2K_{t-1} + (1-a_2)\nabla L(\Theta_t)^2 \tag{8}$$

where a_1, a_2 are exponential decay rates for the moment estimates and Θ_{t+1} is updated based on K_t, Λ_t,

$$\Theta_{t+1} = \Theta_t - b\frac{\sqrt{1-(a_2)^t}}{1-(a_1)^t}\frac{\Lambda_t}{\sqrt{K_t}+\varepsilon} \tag{9}$$

where b is the learning rate and ε is used to avoid explosion. We follow the recommended values in [29], where coefficient b is set to 0.001, a_1 set to 0.9, a_2 set to 0.999 and ε set to 10^{-8}. At the beginning of the iterations, the cost of $L(\Theta)$ converges considerably faster than SGD. Moreover, Eq. (9) shows that the magnitudes of parameter updates are independent of the rescaling of the gradient, therefore it provides a relatively fast convergence speed even after a large amount of iterations.

3. Experiment and Results

Firstly, we introduce our training dataset. Next, we examine and analyze the network on different parameters, including filter number, filter size, training patch size and network depth. At last, we compare our model with state-of-the-arts both quantitatively and qualitatively.

3.1 Training Dataset

Generally, deep learning benefits from data training as shown in the literature. In this work, we build a training dataset, choosing 1000 high-quality images from ILSVRC 2013 ImageNet detection partition [44] and Ava dataset [45]. For each image, the blurred version with different blurring level are set by Gaussian filtering [21-30]. Specifically, Gaussian filter set a standard deviation from 2 to 10 randomly. As shown in **Fig. 4**, images in first two rows with different blurred regions are processed from the original images with Gaussian filter. And, for each blurred image, patches of size are object region randomly sampled (human, horse, ship etc.). In this study, we totally obtain 1000 groups of images (1000 pair of blurred images and 1000 clean versions) from the open dataset, as show in **Fig. 4**.

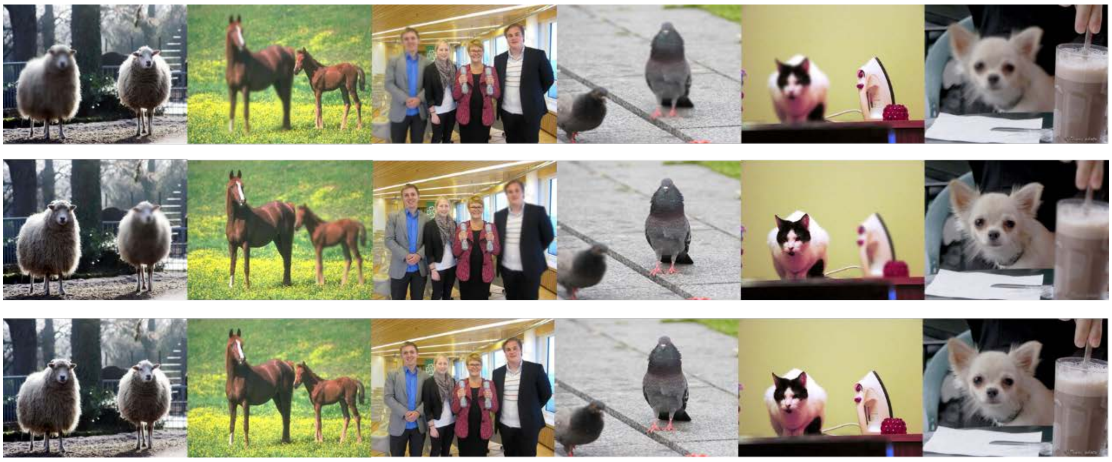


Fig. 4. Examples of training dataset. The top and the middle row is a pair of blurred images, The bottom row is the clean version.

3.2 Network Analysis

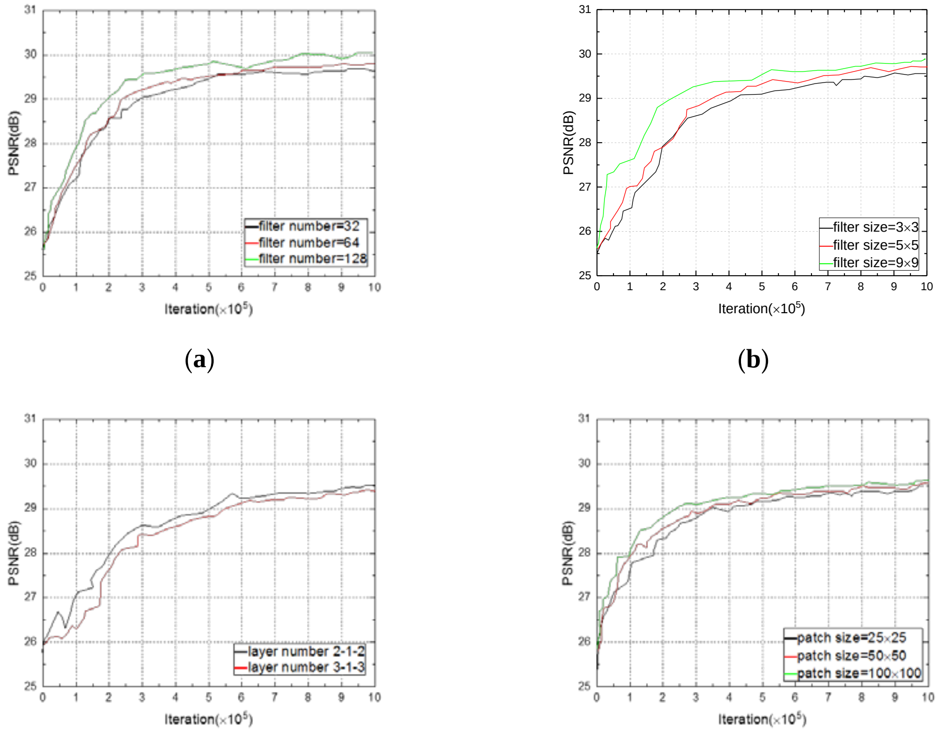
We modify some network parameters to explore the best trade-off between performance and speed, and find the relations between performance and parameters. In experiments, we convert training images into gray ones and evaluate the performance with Peak Signal-to-Noise Ratio (PSNR). We conduct the experiments on four aspects: (a) filter number, (b) filter size, (c) number of layers and (d) training patch size, to analyze the effects of different parameters.

We set filter number 32, 64 and 128 to be tested. Generally, the performance would boost if we extend the network width, i.e. adding more filers in a layer, at the cost of running time. It demonstrates that better performance could be achieved by extending the width, as show in **Fig. 5(a)**. However, if a fast image fusion speed is demanded, a narrow width network is popular and still can achieve good performance.

For experiments on filter size, we set filter size 3×3, 5×5 and 9×9 to examine the network sensitivity to different filter sizes. Experiments show the PSNR values as in **Fig. 5(b)**. This indicates that a properly larger filter size could capture richer structural information, which lead to better performance thereof. However, the running speed also decreases with the increase of filter size. The reason may be that for pixel-level image fusion, especially in deconvolutional layers for image restoration, larger filter need more information to recover larger region details. We can draw the conclusion that smaller filter size is beneficial for network convergence in such complex mappings, instead, larger filter make network more difficult to converge. From this perspective, we should balance between the speed and performance.

For number of layers, we gain that CNN could achieve good performance by increasing the depth of network moderately [31]. CNN from each layer have different nature of the features in the network [32]. Layer 2 shows the low-level features of corners and other edge/color conjunctions. Layer 3 shows complex invariances, extracting similar textures (e.g. mesh pattern). Layer 4 shows significant variation, capturing more class-specific (e.g. human face, horse legs). Layer 5 shows global features of objects with significant pose variation, e.g. clock, tree and dog. That is, CNN has larger receptive fields at higher-level layers. They can capture larger regions with lower spatial accuracy, which is beneficial for high-level tasks, e.g. image detection, classification and tracing. CNN at lower layers tends to localize features at smaller scale more precisely. As our task is feature maps fusion and restoration. In order to restore high quality image, we fix the framework with 2 convolutional layers, 1 fusion layer and 2 deconvolutional layers (2-1-2). We conduct two experiments, i.e. 2-1-2 and 3-1-3, as show in **Fig. 5(c)**. The initialization scheme and learning rate are all the same. We can observe that the network with more layers do not lead to better performance. The reason may be that image more details could be lost or corrupted by adding more layers. Meanwhile, the experiments indicate that deeper model make convergence more difficult. The same phenomenon is also described in [18-33], where increase of layers leads to speed sacrifice and performance degradation for image restoration.

For the training patch size, we set the filter number to be 64, filter size as 3×3. Then we test different training patch sizes of 25×25, 50×50, 100×100, as shown in **Fig. 5(d)**. Better performance is achieved with larger training patch size. Since the network performs pixel-level image fusion, large size patch contains more low feature information, and larger size of training patch contains more pixels that better capture the latent distributions to be learned.



(**a**) (**b**)

(**c**) (**d**)

Fig. 5. The performance on the validation set during training with different parameters: (a) The test of filter number; (b) The test of filter size; (c) The test of layer number; (d) The test of patch size.

3.3 Quantitative and Qualitative Evaluation

In this section, we verify the effectiveness of the proposed deep learning fusion method, and compare the results of our method with state-of-art methods quantitatively and qualitatively.

3.3.1 Experimental settings

In order to achieve good performance-speed trade-off, we train the network with patch size 50×50, convolutional and deconvolutional filter size 5×5, stride size 5 without overlapping, and set filter number 64, layer number 2-1-2, base learning rate of 10^{-5} . Training data are color and gray scale images with four channels. The network was trained on a single NVIDIA Tesla K40 GPU.

The proposed fusion scheme is compared with nine multi-focus image fusion methods, DWT-based method (DWT) [34], Laplacian Pyramid method (LP) [35] , Steerable Pyramid method (SP) [36], Ratio Pyramid method (RP) [37], Point Spread Functions method(PSFs) [38], Contrast Pyramid method (CP) [39], Guided filtering method(GF) [12], SRCNN method [20] and CNN method [21].

Objective evaluation is significant in pixel-level image fusion as the performance of a fusion sheme mainly assessed by the quantitative scores on multiple metrics. In this work, we select four metrics: Information Entropy (IE) [43], Mutual Information (MI) [40], Xydeas [41] and Piella [42]. Information entropy is generally applied to measure the amount of information. The more information entropy there is the better fusion result is obtained. Mutual information represents how much information obtained from the fusion of input images used to assess the performance of different image fusion algorithms. In addition, Xydeas and Piella metrics are applied to the assessment of the salient information transferred from the input images to the fused images. Piella metric takes the image correlation coefficient, mean luminance, contrast, and edge information into account in a comprehensive manner. The dynamic ranges of three Piella indexes Q, Q_w and Q_e are [0,1]. The larger the values are, the better the fusion performance is. We evaluate three traditional multi-focus image pairs “Clock”, “Pepsi” and “Camera”.

3.3.2 Fusion on multi-focus “Clock” images

In this section, a pair of 512 ×512 “Clock” multi-focus images with different focused regions are utilized to perform the proposed scheme and the comparative methods. In **Fig. 6, (a)** and **(b)** are source images. As the right clock is focused in ClockA, the left one is fuzzy. In comparison, the left clock is focused in ClockB and the right one is fuzzy. This pair of multi-focus images consists of both rich cartoon components and abundant texture components comparatively. In this section, we sketch some experiments to verify the performance of the proposed fusion approach.

Results using different fusion methods are shown in **Fig. 6**. The fused images are obtained by combining two multi-focus images by different methods. In **Fig. 5 (c)–(l)** denote the result acquired by DWT-based method (DWT), Laplacian Pyramid method (LP), Steerable Pyramid method (SP), Ratio Pyramid method (RP), Point Spread Functions method(PSFs), Contrast Pyramid method (CP) Guided filtering method(GF), SRCNN method, CNN method and the proposed method, respectively. The fused images in **Fig. 5 (c)–(f)** are obtained by integrating the cartoon and texture components. Although the fusion methods get high-quality fused images, the luminance distortion is obvious compared with the source images meanwhile, experimental results also demonstrate that some noise information has been introduced into the fused images obtained by DWT, LP and RP. In comparison, the fused images merged by PSFs, CP, GF, SRCNN,CNN and the proposed method contain much more detailed information. On the other hand, it is difficult to distinguish the diversity among those fused images acquired by PSFs, CP, CNN and the proposed method for human visual system. Thus, the objective indices are utilized to evaluate the fused images. The results of the assessment criteria are shown in **Table 1**.

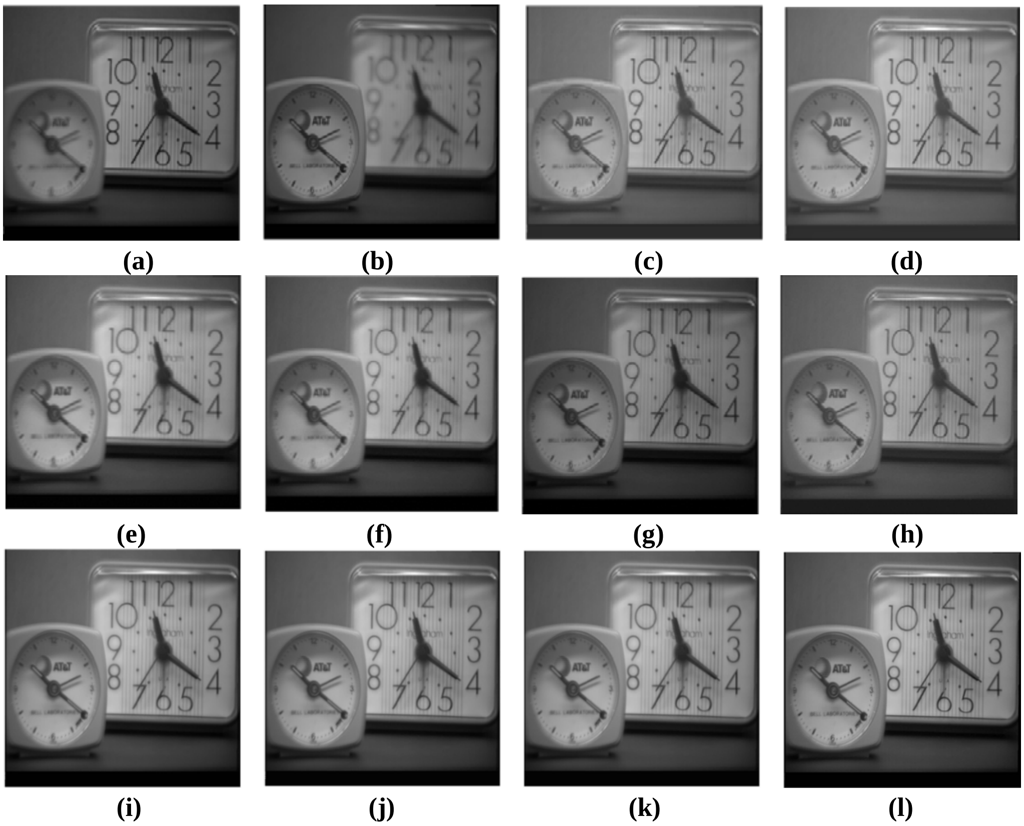


Fig. 6. The fused images obtained by different fusion methods for multi-focus “Clock” images. (a) and (b) Multi-focus source images：ClockA and ClockB ; (c) DWT-based method; (d) Laplacian Pyramid method; (e) Steerable Pyramid method; (f) Ratio Pyramid method; (g) Point Spread Functions method; (h) Contrast Pyramid method; (i) GF method; (j) SRCNN method; (k) CNN method; (l)Proposed method.

Table 1. The quantity assessment of fusion methods for multi-focus “Clock” images.

Metrics Methods	IE[43]	MI[40]	Xydeas[41]	Piella[42]		
				Q	Qw	Qe
DWT	5.5755	0.9073	0.6891	0.7307	0.8380	0.7318
LP	5.5174	1.0266	0.7560	0.7609	0.8539	0.7742
SP	5.7160	1.0568	0.4340	0.6874	0.7986	0.4822
RP	5.6579	1.0875	0.5315	0.7492	0.8163	0.6057
PSFs	5.5968	1.0714	0.7580	0.7970	0.8541	0.7734
CP	5.6047	1.0630	0.7522	0.7704	0.8562	0.7437
GF	5.7560	1.0789	0.7490	0.7793	0.8574	0.7792
SRCNN	5.7641	1.0984	0.7581	0.7887	0.8590	0.7940
CNN	**5.8764**	**1.2059**	0.7654	0.8045	0.8602	0.8177
Ours	5.8471	1.1574	**0.7691**	**0.8261**	**0.8754**	**0.8370**

As shown in **Table 1**, the proposed fusion approach outperforms other methods in terms of the evaluation criteria including Xydeas, and Piella. These quantitative assessments indicate that the fused image obtained by the proposed method contains more detail information and clarity. Although the value of quantitative metric IE and MI is smaller, the distinction is relatively tiny. In conclusion, the proposed fusion approach is superior to other methods through quantitative evaluation.

3.3.3 Fusion on multi-focus "Pepsi" images

In this work, some fusion experiments on a set of multi-focus "Pepsi" images with size of 512 ×512 pixels are sketched to illustrate the performance of the proposed fusion approach. PepsiA focuses on the right side of the device with white panel, some words and strides in the cover, the device clear, part Pepsi fuzzy. While PepsiB focuses on the left side of the part Pepsi with cylinder surface, and some words in the cover, part Pepsi clear, the device fuzzy. As mentioned above, this group of source images is abundant in cartoon components and texture components. To show the effectiveness of the different fusion methods, some experiments are implemented and demonstrate the merits of our method. The experimental results are shown in **Table 2**.

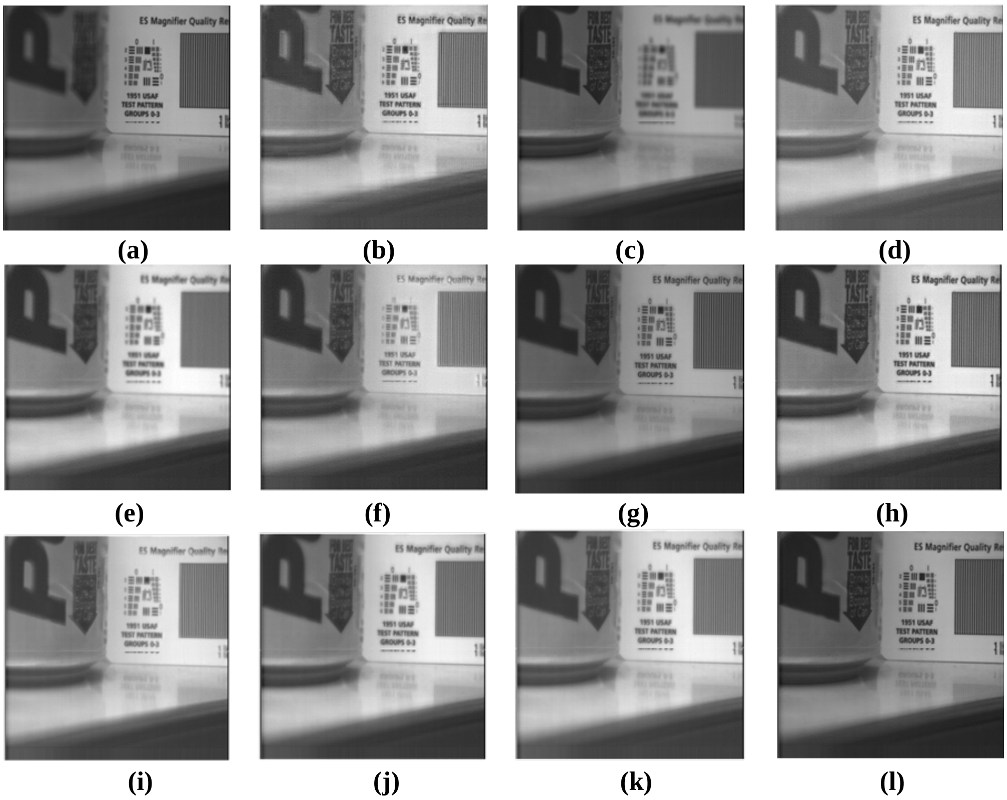


Fig. 7. The fused images obtained by different fusion methods for multi-focus "Pepsi" images. (a) and (b) Multi-focus source images: PepsiA and PepsiB; (c) DWT-based method; (d) Laplacian Pyramid method; (e) Steerable Pyramid method; (f) Ratio Pyramid method; (g) Point Spread Functions method; (h) Contrast Pyramid method; (i) GF method; (j) SRCNN method; (k) CNN method; (l)Proposed method.

The fused images obtained by the proposed fusion approach and the well-known methods are shown in **Fig. 6**. The source images can provide dissimilar information as shown in **Fig. 6 (a) and (b)**. From the fused results, visual observation illustrates that the fused images are satisfactory with contrast as shown in **Fig. 7 (d), (g), (h), (k)** and **(l)**, which are obtained by Laplacian Pyramid method, Steerable Pyramid method, Contrast Pyramid method CNN and the proposed method, respectively. Obviously, the fusion images by DWT-based method lose edge information on Pepsi, and the fused image by Ratio Pyramid method contains a fraction of noise on device.

As shown in **Table 2**, according to the evaluation metrics including IE and Piella, the proposed fusion method achieves better results when compared with other methods. Results of these assessment metrics demonstrate that the proposed fusion approach can capture much more information from source images. In particular, although the values of Xydeas are not the largest, the distinction is little. Therefore, the proposed fusion approach outperforms others according to visual intuition and quantitative evaluation.

Table 2. The quantity assessment of fusion methods for multi-focus “ Pepsi” images.

Metrics / Methods	IE	MI	Xydeas	Piella		
				Q	Qw	Qe
DWT	5.7644	0.9103	0.6703	0.7463	0.6887	0.7686
LP	5.7715	0.9814	0.7344	0.8554	0.8624	0.8340
SP	5.8198	1.0757	0.7655	0.8492	0.8564	0.8108
RP	5.7155	1.0501	0.5560	0.7410	0.7890	0.6288
PSFs	5.7811	1.0175	0.7888	0.8638	0.8630	0.8402
CP	5.8462	1.0751	**0.8262**	0.8640	0.8701	0.8411
GF	5.8667	1.0655	0.7810	0.8428	0.8561	0.8156
SRCNN	5.8813	1.1450	0.7899	0.8506	0.8423	0.8278
CNN	5.8952	1.1580	0.7938	0.8656	0.8599	0.8388
Ours	**5.9275**	**1.2834**	0.8066	**0.8795**	**0.8744**	**0.8472**

3.3.4 Fusion on multi-focus “Camera” images

In this section, the corresponding experiments on a pair of multi-focus “Camera” images with size 256 ×256, are implemented among the proposed scheme and other methods. The source images are shown in **Fig. 8(a)** and **(b)**. In CameraA, the trunk, head and right arm of the photographer are fuzzy, other contents of image are clear, while CameraB is the complete opposite to CameraA. In the experiments, we evaluate the performance of the proposed approach.

As shown in **Fig. 8**, from the perspective of human visual perception mechanism, Ratio Pyramid method has luminance distortion to some extent compared with the source images obviously, and there are some noise in **Fig. 8 (e)**, **(g)**, **(i)** and **(h)**, which is obtained by Steerable Pyramid method, Point Spread Functions method, Guided filtering method and Contrast Pyramid method. Fortunately, the fused images contain much more detail information including edges and contours in **Fig. 8(c)**, **(d)**, **(j)**, **(k)** and **(l)**, which are obtained by DWT-based method, Laplacian Pyramid method, SRCNN, CNN and the proposed method, respectively. There are some halo on cameraman’ shoulder as show in **Fig. 8** (j), (k) and (l).

The main reason of our method is that the halo is retained in feature map of convolutional fusion layer (see **Fig. 2**(c)). When operating deconvolution, the halo has been restored.

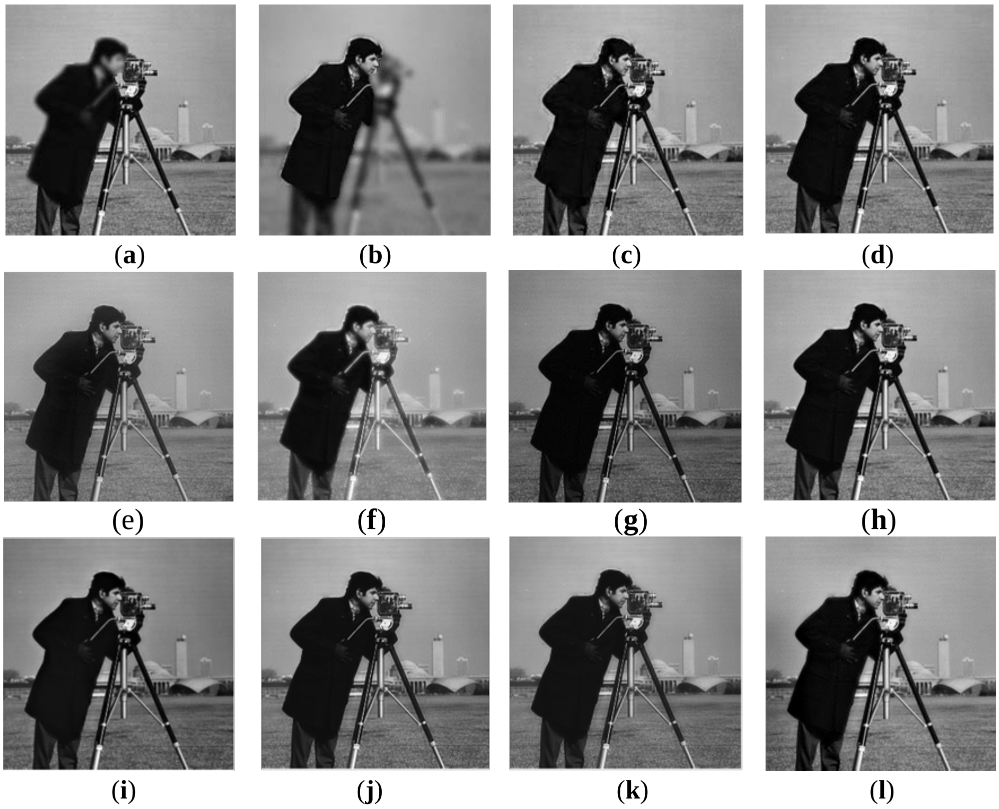


Fig. 8. The fused images obtained by different fusion methods for multi-focus “Camera” images. (a) and (b) Multi-focus source images: CameraA and CameraB; (c) DWT-based method; (d) Laplacian Pyramid method; (e) Steerable Pyramid method; (f) Ratio Pyramid method; (g) Point Spread Functions method; (h) Contrast Pyramid method; (i) GF method; (j) SRCNN method; (k) CNN method; (l)Proposed method.

Table 3. The quantity assessment of fusion methods for multi-focus “ Camera” images.

Metrics Methods	IE	MI	Xydeas	Piella		
				Q	Qw	Qe
DWT	5.3690	0.9149	0.6597	0.8004	0.8905	0.7807
LP	5.3881	0.8963	0.6471	0.8139	0.8964	0.7912
SP	5.4190	0.7502	0.6699	0.7972	0.8426	0.7268
RP	5.4427	0.7052	0.6716	0.7279	0.8178	0.7090
PSFs	5.3964	0.8974	0.6532	0.8154	0.8984	0.7935
CP	5.3590	0.9548	0.6577	0.8137	0.8854	0.7692
GF	5.3080	0.7693	0.6450	0.7675	0.8126	0.7190
SRCNN	5.3281	0.8021	0.6478	0.7720	0.8478	0.7588
CNN	5.3445	1.0890	0.6502	0.7821	0.8855	0.7750
Ours	**5.5714**	**1.2572**	**0.6790**	**0.8207**	**0.8971**	**0.7988**

As shown in **Table 3**, all quantitative metrics demonstrate that the proposed fusion approach can have much better performance from multi-focus source images. Since the test images “Camera” are similar with the training groups of images, whose object regions are blurred with Gaussian filters. Our network obtained much priori knowledge fitting for this sort of source images fusion. In brief, the results of subjective and objective evaluation illustrate that our fusion method performs better than other methods.

3.3.5 Fusion on multi-focus color images

In this section, fusion experiments on a pair of multi-focus color images blurred with different level by Gaussian filtering artificially are implemented to verify the property of the proposed the fusion approach. Observing multi-focus source color images with size 640 ×396 in **Fig.9(a) and (b)**, the calf is fuzzy in Color_imageA, other contents of image are clear, while Color_imageB is the complete opposite to Color_imageA. In this work, some experiments demonstrate the performance of fusion and high fused image quality.

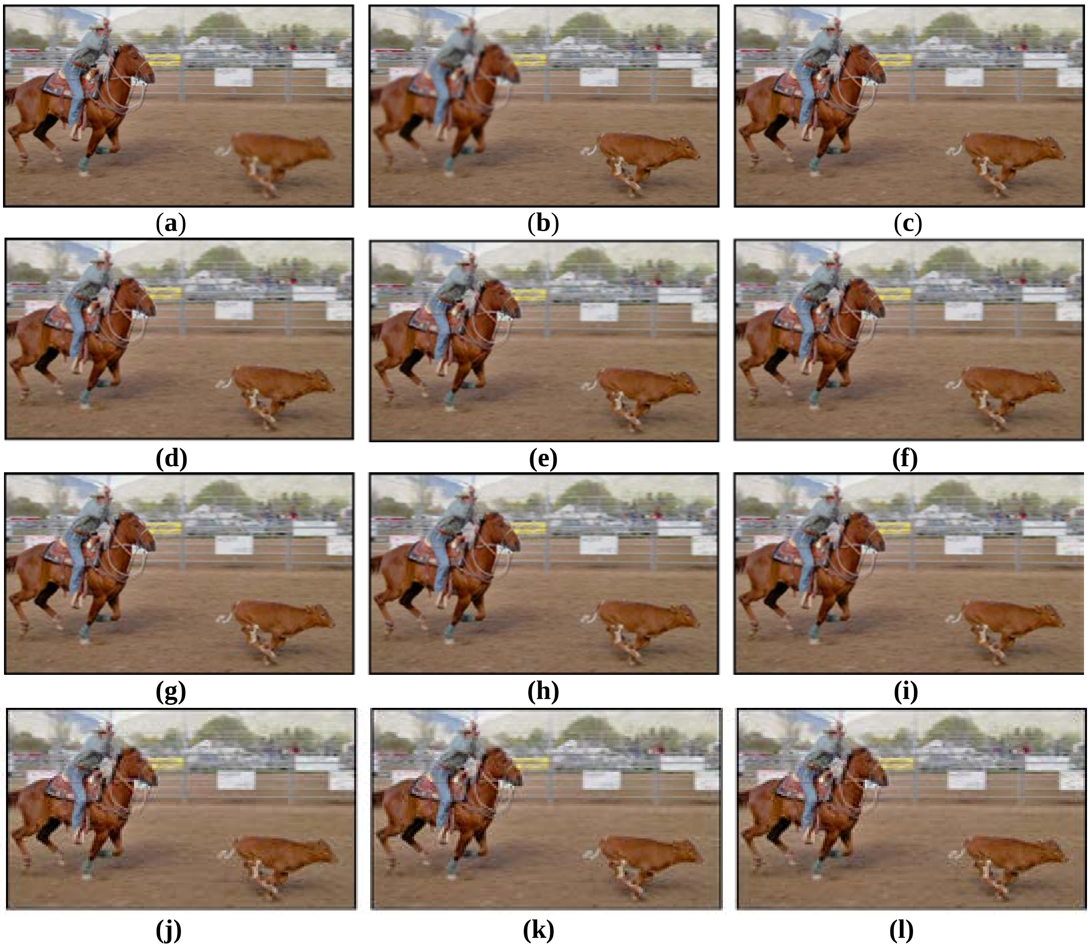


Fig. 9. The fused images obtained by different fusion methods for multi-focus color images. (a) and (b) Multi-focus source images: Color_imageA and Color_imageB; (c) DWT-based method; (d) Laplacian Pyramid method; (e) Steerable Pyramid method; (f) Ratio Pyramid method; (g) Point Spread Functions method; (h) Contrast Pyramid method; (i) GF method; (j) SRCNN method; (k) CNN method; (l)Proposed method.

As shown in **Fig. 9**, It can be seen that DWT-based method and Steerable Pyramid method don't perform very well in the fused image. The fusion quality of the GF method and SRCNN method are much better in terms of this issue, but the region of the horse's tail is still not well merged. The PSFs method, GF method and the proposed method all obtain fusion result in high quality. The difference between the three fused images is relatively small. But when carefully comparing the edge between horse and calf among all the fused images, we can see that the proposed method fuses more natural and smooth than the other methods.

Table 4 lists the objective performance of different fusion methods using the four metrics. All quantitative metrics show that the proposed fusion method has much better performance from multi-focus color images than all the other methods. The objective performance surpasses the above test multi-focus "Camera" images.

Table 4. The quantity assessment of fusion methods for multi-focus color images.

Metrics Methods	IE	MI	Xydeas	Piella		
				Q	Qw	Qe
DWT	5.1478	0.7315	0.6421	0.7654	0.8352	0.7607
LP	5.4784	0.7924	0.6550	0.8032	0.8567	0.7856
SP	5.2112	0.7358	0.6396	0.7641	0.8327	0.7288
RP	5.4371	0.7104	0.6584	0.7354	0.8398	0.7430
PSFs	5.4104	0.7742	0.6602	0.8108	0.8596	0.7935
CP	5.3585	0.7648	0.6523	0.8170	0.8804	0.7642
GF	5.3876	0.7720	0.6574	0.7570	0.8232	0.7274
SRCNN	5.3274	0.7956	0.6595	0.7727	0.8528	0.7620
CNN	5.3925	1.0074	0.6603	0.8127	0.8801	0.7950
Ours	**5.5921**	**1.2588**	**0.6819**	**0.8332**	**0.8994**	**0.8040**

5. Conclusion

In this paper, we have presented a novel deep learning model for multi-focus images fusion. We formulate fully convolutional network into a fully convolutional two-stream fusion network. The proposed method is the first attempt to learn end-to-end combining two stream convolutional networks and restoration a clean image, with convolutional filters, fusion filters and deconvolutional filters. Compared with several spatial domain, transform domain and machine learning based methods, experimental results and our analysis demonstrate that our model achieves better performance than state-of-the-art methods on image fusion with subjective visual perception and objective assessment metrics. The paper advances a many-to-one mapping idea between input source images and output one. Next, we will try to design the multi-stream networks structure. We assume that the further performance can be improved by exploring different fusion and training strategies.

References

[1] Abdipour, M. and M. Nooshyar, "Multi-focus image fusion using sharpness criteria for visual sensor networks in wavelet domain,“ *Computers & Electrical Engineering*, vol. 51, p. 74-88, 2016. Article (CrossRef Link)

[2] Petrović, V.S. and C.S. Xydeas, „Gradient-Based Multiresolution Image Fusion," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 13, no. 2, p. 228-237, 2004. Article (CrossRef Link)

[3] Piella, G., "A general framework for multiresolution image fusion: from pixels to regions," *Information Fusion*, vol. 4, no. 4, p. 259-280, 2003. Article (CrossRef Link)

[4] Goshtasby, A.A. and S. Nikolov, "Image fusion: Advances in the state of the art," *Information Fusion*. Vol. 8, no. 2, p. 114-118, 2007. Article (CrossRef Link)

[5] Zhang, Z. and R.S. Blum, "A categorization of multiscale-decomposition-based image fusion schemes with a performance study for a digital camera application," *Proceedings of the IEEE* , vol. 87, no. 8, p. 1315-1326, 1999. Article (CrossRef Link)

[6] Forster, B., et al., "Complex wavelets for extended depth-of-field: A new method for the fusion of multichannel microscopy images," *Microscopy Research & Technique*, vol. 65, no.v1-2, p. 33–42, 2004. Article (CrossRef Link)

[7] Mount, D.M., N.S. Netanyahu, and J.L. Moigne, "Efficient algorithms for robust feature matching, *Pattern Recognition*, vol. 32, no. 1, p. 17-38, 1999. Article (CrossRef Link)

[8] Stockman, G., S. Kopstein, and S. Benett, "Matching Images to Models for Registration and Object Detection via Clustering," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 4, no. 3, p. 229-41, 1982. Article (CrossRef Link)

[9] Benediktsson, J.A. and I. Kanellopoulos, "Classification of multisource and hyperspectral data based on decision fusion," *IEEE Transactions on Geoscience & Remote Sensing*, 37, no. 3, p. 1367-1377, 1999. Article (CrossRef Link)

[10] Wang, Z., et al., "A comparative analysis of image fusion methods," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 43, no. 6, p. 1391-1402, 2005. Article (CrossRef Link)

[11] Cvejic, N., et al., "Region-Based Multimodal Image Fusion using ICA Bases," in *Proc. of International Conference on Image Processing*, Atlanta, Georgia, USA. p. 1801-1804, 2006. Article (CrossRef Link)

[12] Li, S., X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 22, no. 7, p. 2864-2875, 2013. Article (CrossRef Link)

[13] Li, S., J.T. Kwok, and Y. Wang, "Multifocus image fusion using artificial neural networks," *Pattern Recognition Letters*, vol. 23, no. 8, p. 985-997, 2002. Article (CrossRef Link)

[14] Li, S., et al., "Fusing images with different focuses using support vector machines," *Neural Networks IEEE Transactions on*, vol. 15, no. 6, p. 1555-61, 2004. Article (CrossRef Link)

[15] Mamatha, S.G., S.A. Rahim, and C.P. Raj, "Feature-level multi-focus image fusion using neural network and image enhancement, Global," *Global Journal of Computer Science & Technology*, vol. 12, no. 10-F, 2012.

[16] Rani, C.M.S., P.S.V.S. Rao, and V. Vijayakumar, "Improved Block Based Feature Level Image Fusion Technique Using Contourlet with Neural Network," *International Journal of Soft Computing & Engineering*, vol. 3, no. 4, 2012.

[17] Jain, V. and H.S. Seung," Natural Image Denoising with Convolutional Networks," in *Proc. of Conference on Neural Information Processing Systems*, Vancouver, British Columbia, Canada. p. 769-776, 2008.

[18] Dong, C., et al., "Image Super-Resolution Using Deep Convolutional Networks," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 38, no. 2, p. 295-307, 2016. Article (CrossRef Link)

[19] Long, J., E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA. p. 3431-3440, 2015. Article (CrossRef Link)

[20] Zhong, J., et al, "Image Fusion and Super-Resolution with Convolutional Neural Network," in *Proc. of Chinese Conference on Pattern Recognition*, Chengdu, China: Springer Singapore, 2016. Article (CrossRef Link)

[21] Liu, Y., et al., “Multi-focus image fusion with a deep convolutional neural network,“ *Information Fusion*, vol. 36, p. 191-207, 2017. Article (CrossRef Link)

[22] Feichtenhofer, C., A. Pinz, and A. Zisserman, “Convolutional Two-Stream Network Fusion for Video Action Recognition,“ in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA. p. 1933-1941, 2016.

[23] Nair, V. and G.E. Hinton, “Rectified Linear Units Improve Restricted Boltzmann Machines,“ in *Proc. of Icml*, p. 807-814, 2015.

[24] Noh, H., S. Hong, and B. Han, “Learning Deconvolution Network for Semantic Segmentation,“ in *Proc. of IEEE International Conference on Computer Vision* Santiago, Chile, p. 1520-1528, 2015. Article (CrossRef Link)

[25] Jia, et al., “Caffe: Convolutional Architecture for Fast Feature Embedding,“ in *Proc. of Eprint Arxiv*, p. 675-678, 2014. Article (CrossRef Link)

[26] Lecun, Y., et al., “Gradient-based learning applied to document recognition,“ in *Proc. of Proceedings of the IEEE*, vol. 86, no. 11, p. 2278-2324, 1998. Article (CrossRef Link)

[27] Rumelhart, D.E., G.E. Hinton, and R.J. Williams, “Learning representations by back-propagating errors,“ *Nature*, vol. 323, no. 6088, p. 533-536, 1986. Article (CrossRef Link)

[28] Bottou, L., “Large-Scale Machine Learning with Stochastic Gradient Descent,“ in *Proc. of International Conference on Computational Statistics*,Paris France, p. 177-186, 2010. Article (CrossRef Link)

[29] Kingma, D. and J. Ba, “Adam: A Method for Stochastic Optimization,“ in *Proc. of International Conference for Learning Representations*, San Diego, USA, 2015.

[30] Mao, X.J., C. Shen, and Y.B. Yang, “Image Restoration Using Convolutional Auto-encoders with Symmetric Skip Connections,“ in *Proc. of arXiv preprint arXiv*, 1606.08921, 2016.

[31] He, K. and J. Sun, Convolutional Neural Networks at Constrained Time Cost, *in IEEE Conference on Computer Vision and Pattern Recognition*, Boston, MA, USA. p. 5353-5360.

[32] Zeiler, M.D. and R. Fergus, “Visualizing and Understanding Convolutional Networks,“ in *Proc. of European Conference on Computer Vision*, Zurich, Switzerland, p. 818-833, 2014. Article (CrossRef Link)

[33] Glasner, D., S. Bagon, and M. Irani, “Super-resolution from a single image,“ in *Proc. of International Conference on Computer Vision*, Kyoto, Japan, p. 349-356, 2009. Article (CrossRef Link)

[34] Li, H., B.S. Manjunath, and S.K. Mitra, “Multi-sensor image fusion using the wavelet transform,“ in *Proc. of Image Processing, Proceedings. ICIP-94., IEEE International Conference*, 1994.

[35] Burt, P.J. and E.H. Adelson, “The Laplacian Pyramid as a Compact Image Code,“ *Readings in Computer Vision*, vol. 31, no. 4, p. 671-679, 1987. Article (CrossRef Link)

[36] Liu, Z., et al., “Image fusion by using steerable pyramid,“ *Pattern Recognition Letters*, vol. 22, no. 9, p. 929-939, 2001. Article (CrossRef Link)

[37] Toet, A., “Image fusion by a ratio of low-pass pyramid,“ *Pattern Recognition Letters*, vol. 9, no. 4, p. 245-253, 1989. Article (CrossRef Link)

[38] Aslantas V, Toprak A N, “A pixel based multi-focus image fusion method,“ *Optics Communications*, vol. 332, no. 4, pp. 350-358, 2014. Article (CrossRef Link)

[39] Toet, A., J.M. Valeton, and L.J. Van Ruyven, “Merging thermal and visual images by a contrast pyramid,“ *Optical Engineering*, vol. 28, no. 7, p. 789-792, 1989. Article (CrossRef Link)

[40] Hossny, M., S. Nahavandi, and D. Creighton, “Comments on 'Information measure for performance of image fusion,“ *Electronics Letters*, vol. 44, no. 18, p. 1066-1067, 2008. Article (CrossRef Link)

[41] Xydeas, C.S. and V. Petrovic, “Objective image fusion performance measure,“ *Military Technical Courier*, vol. 36, no. 4, p. 308-309, 2000. Article (CrossRef Link)

[42] Piella, G. and H. Heijmans, “A new quality metric for image fusion,“ in *Proc. of International Conference on Image Processing*, Barcelona, Catalonia, Spain, p. III-173-176, 2003. Article (CrossRef Link)

[43] Shannon C E., “A Mathematical Theory of Communication[J],“ *Bell Labs Technical Journal*, vol. 27, no. 4, pp. 379-423, 1948. Article (CrossRef Link)

[44] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, *ILSVRC-2012*, 2012. Article (CrossRef Link).

[45] Perronnin F, AVA, “A large-scale database for aesthetic visual analysis,“ in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, *IEEE Computer Society*, pp. 2408-2415, 2012.

[46] Li H, Chai Y, Li Z., “A new fusion scheme for multifocus images based on focused pixels detection,“ *Machine Vision & Applications*, vol. 24, no. 6, 1167-1181, 2013. Article (CrossRef Link)

[47] Li H, Chai Y, Li Z., “Multi-focus image fusion based on nonsubsampled contourlet transform and focused regions detection,“ *Optik - International Journal for Light and Electron Optics*, vol. 124, no. 1, 40-51, 2013. Article (CrossRef Link)

[48] Gao G, Xu L, Feng D.,“Multi-focus image fusion based on non-subsampled shearlet transform,“ *Iet Image Processing*, vol. 7, no. 6, pp. 633-639, 2013. Article (CrossRef Link)

[49] Haghighat M B A, Aghagolzadeh A, Seyedarabi H., “Multi-focus image fusion for visual sensor networks in DCT domain,“ *Computers & Electrical Engineering*, vol. 37, no. 5, pp. 789-797, 2011. Article (CrossRef Link)

Kaiping Xu is currently working toward the PhD degree in software engineering at Tsinghua University, Beijing, China. His research interests include computer vision, machine learning, and intelligent video content analysis.

Zheng Qin is a doctoral supervisor, professor and the director of the Software Engineering and Management Research Institute and Information Institute of Tsinghua University. Currently his research interests include Big data fusion and mobile computing, E-commerce and Internet of things, Image processing and Computer vison, Network security and software architecture.

Guolong Wang is currently working toward the PhD degree in software engineering at Tsinghua University, Beijing, China. His research interests include computer vision, machine learning, and intelligent video content analysis.

Huihui Zhang is currently working toward the Master degree in software engineering at Tsinghua University, Beijing, China. His research interests include computer vision, machine learning, and data analysis.

Kai Huang is currently working toward the Master degree in software engineering at Tsinghua University, Beijing, China. His research interests include computer vision, machine learning, and intelligent video content analysis.

Shuxiong Ye is currently working toward the Master degree in software engineering at Tsinghua University, Beijing, China. His research interests include computer vision, machine learning, and intelligent video content analysis.