IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 28, 2020, accepted July 12, 2020, date of publication July 20, 2020, date of current version August 4, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3010542

Non-Local Multi-Focus Image Fusion With Recurrent Neural Networks

ZHAO DUAN, TAIPING ZHANG, (Member, IEEE), JIN TAN, AND XIAOLIU LUO
School of Computer Science, Chongqing University, Chongqing 400044, China
Corresponding author: Taiping Zhang (tpzhang@cqu.edu.cn)

ABSTRACT Previous Convolutional Neural Networks (CNNs) based multi-focus image fusion methods rely primarily on local information of images. In this paper, we propose a novel deep network architecture for multi-focus image fusion that is based on a non-local image model. The motivation of this paper stems from local and non-local self-similarity widely shown in nature images. We build on this concept and introduce a recurrent neural network (RNN) that performs non-local processing. The RNN captures global and local information by retrieving long distant dependencies, hence augmenting the representation of each pixel with contextual representations. The augmented representation is beneficial to detect accurately focused and defocused pixels. In addition, we design a regression loss to address the influences of texture information. Experimental results demonstrate that the proposed method outperforms the state-of-the-art methods, both qualitatively and quantitatively.

INDEX TERMS Multi-focus image fusion, recurrent neural network, non-local self-similarity, texture information.

I. INTRODUCTION

Most image systems, such as digital cameras, have a limited depth-of-field and, as a result, only the objects in the depth-of-field are sharp, while others are blurred. Therefore, many researchers have designed various algorithms called multi-focus image fusion to integrate several images of the same scene into an all-in-focus image. The all-in-focus image can be used for further image-processing tasks and many applications, such as recognition, detection, and surveillance.

In the past decades, many algorithms have been proposed for multi-focus image fusion. In general, these algorithms can be categorized into four classes: transform domain-based methods, sparse representation-based methods, the CNN-based and spatial domain-based methods. Some representative transform domain-based methods include the Laplacian pyramid (LP) [1], the morphological pyramid (MP) [2], the discrete wavelet transform(DWT) [3], the dual-tree complex wavelet transform (DTCWT) [4], the curvelet transform (CVT) [5], the non-subsampled contourlet transform (NSCT) [6], [7] and the sparse representation (NSCT-SR) [8]. The above methods can effectively extract features of focused regions. However, most of them are complex and time-consuming [9].

Many sparse representation-based methods have been applied to image fusion [10]–[12]. The source image can be represented with sparse coefficients and an overcomplete dictionary. The sparse coefficients work well in representing the saliency information of the source image. A sequence of sparse representation-based algorithms performs successfully [13]–[17]. In [17], Zhu *et al.* first decomposed the source images into cartoon and texture components, aiming at describing the structure and detailed information. Then an energy-based and a sparse representation-based method were adopted to fuse cartoon and texture components, respectively.

Spatial domain-based methods have been applied widely to multi-focus image fusion [18]–[23]. The basic idea is to directly select clearer pixels or regions from the source image to construct the fused image. At present, the advanced pixel-based fusion methods include guided filtering-based (GFF) [24], the dense SIFT (DSIFT) [25], and self-similarity and depth information-based [26] methods. These fusion algorithms perform well in extracting and preserving image details. However, the focused detections could be inaccurate in the boundary between focused and defocused regions.

The associate editor coordinating the review of this manuscript and approving it for publication was Naveed Akhtar.

Recently, a variety of CNN-based image fusion methods [27]–[33] have been proposed. CNN-based fusion methods have gained many breakthroughs, which overcome the difficulty of manually designing complicated activity level measurement and fusion rules. Moreover, the activity level measurement and fusion rule can be jointly generated via learning a CNN model. A deep convolution neural network-based image fusion method firstly is presented by Liu [27]. Liu *et al.* view multi-focus image fusion as a classification task. Specifically, they designed a simple CNN to classify focused and defocused patches, generating an initial focus map from classification results. The all-in-focus image was generated after post-processing the initial focus map. Besides, to effectively learn the model, they created thousands of clean-blur image patch pairs for training. The patch size could influence the accuracy of the model. Moreover, the model also presented some inaccurate focused pixels when focused and defocused regions both existed in one patch. To solve this problem, Tang *et al.* proposed a p-CNN for multi-focus image fusion [30]. They created more precise classification labels and classified pixels into three classes: focused, defocused, unknown. Ma *et al.* proposed a boundary aware multi-focus image fusion approach based on deep neural networks [33], in which two different convolutional networks are employed to classify two kinds of patches, respectively. A kind of patch is far away from the boundary between focused and defocused regions and the other is near the boundary. Although the algorithm works well in fusing multi-focus images, it increases computational cost and storage. In [35], Lai *et al.* combined standard and dilated convolutional kernels to excavate more spatial information for segmenting focused and defocused pixels. Guo *et al.* utilized a conditional generative adversarial network to fulfill the image fusion [36]. They tried to train a discriminator to penalize the mismatched relationship between the focus map and source images, which guided their generator to produce a high confidence map. [34], [37] directly reconstruct the fused image instead of outputting the focus map. Specifically, in [34], Zhao *et al.* presented an end-to-end network that extracts low-level and high-level features to capture low-frequency content and high-frequency details, and then the multi-level features are combined to reconstruct the fused image. In [37], Zhang *et al.* introduced a general image fusion framework aiming at fusing multiple types of images without any finetuning procedures. They firstly used two convolutional layers to extract image features from multiple input images, and then the features of multiple input images were fused by a suitable fusion rule that is designed according to the type of input images. For instance, elementwise-maximum fusion rule is used to fuse multi-focus images, infrared and visual images, and medical images. Elementwise-mean fusion rule is applied to fuse the multi-exposure images. Afterwards, the fused image is produced by reconstructing the fused features. However, due to reconstructing the fused image without the focus map, the fused images obtained by the above two algorithms suffer from undesirable pixels that are irrelevant to the source images.

Although these recent advanced algorithms achieve appealing performance, most of them exploit local information to fulfill the fusion task, limiting fusion effect due to a lack of sufficient information. To improve fusion performance, exploiting global information of the image is an intuitive and important approach. Some researchers may have paid attention on global information. In our work, motivated by inherent local and non-local self-similarity, we propose a deep non-local network architecture for multi-focus image fusion that utilizes the self-similarity to generate more distinguishable representations. Specifically, we introduce the recurrent neural network (RNN) to retrieve local and global contextual dependencies, learning the relation between each pixel representation and augmenting the representation of each pixel according to their relations with the pixel representation. The augmented representation is very beneficial to segment pixels as focused or defocused. To the best of our knowledge, this is the first time that applying RNN to the multi-focus image fusion task. Our contributions in this work can be summarized as follows:

(1) We propose a novel deep network architecture for multi-focus image fusion. As opposed to most existing deep-learning algorithms for image fusion, which are based on local models, our model captures non-local contextual information from images.
(2) We unroll the RNN into the deep network architecture. The RNN can fully exploit local and global pixel dependencies.
(3) We design a regression loss that supervises the fused image by the ground-truth fusion image. In contrast to the majority of image fusion methods that are easily influenced by texture information, our fusion algorithm is capable of avoiding interference of texture information by utilizing the regression loss.
(4) Compared with most state-of-the-art fusion algorithms, the proposed fusion method can outperform them in terms of both visual and objective quality evaluation.

We describe the proposed method in Section II. A quantitative evaluation on public test datasets as well as visual illustrations are provided in Section III. The conclusions are given in Section IV.

II. THE PROPOSED METHOD

In this work, we pursue a non-local network architecture that generates a confidence map indicating the focused information for fusion tasks. The network has a great capacity for learning the local and non-local self-similarity features from training data. The core idea is that we explore the relation between features of each pixel to augment the representation of pixels according to the relation, providing a more accurate map. Next, we describe in detail the overall architecture of the proposed network.

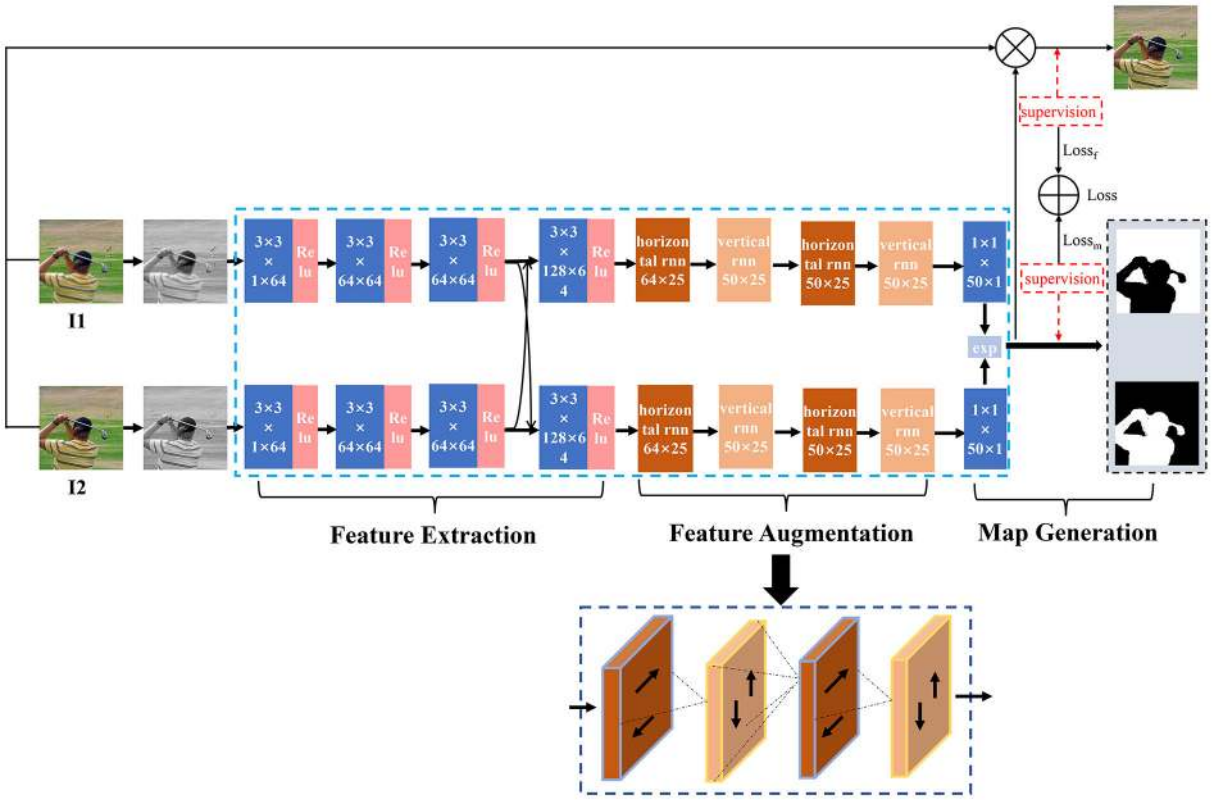


FIGURE 1. Schematic diagram of proposed non-local network architecture.

A. NETWORK ARCHITECTURE

As mentioned earlier, non-local models have been shown to generate superior results than their local counterparts [38], [39]. Their superiority in performance is mainly attributed to their ability of modeling complex image structures by acquiring richer information. This fact highly motivates us to design a non-local network for multi-focus image fusion that learns non-local self-similarity features to augment the representation of each pixel. The RNN is introduced to implement the non-local model and there are two reasons why we choose the RNN. Firstly, the RNN can learn the relation between non-local features of pixels by capturing long distance spatial dependencies. Secondly, the RNN is capable of augmenting the representation of each pixel with the long-term memory function. Based on the above analysis, we can notice that the relation of non-local features is obtained by the learning-based method instead of the nonparametric method, which may lead to better fusion performance. Moreover, the two processes of learning the relation of non-local features and augmenting the representation are integrated into a process by the RNN, which can be trained more effectively in an end-to-end manner.

We consider the situation that only two input images exist since the fusion process of multi-focus images can be achieved by fusing dual-focus images in succession. For a pair of multi-focus images of the same scene, our goal is to learn a CNN model aiming at extracting focused and defocused information of images and learning a map matrix ranging from 0 to 1. As shown in Fig.1, the proposed network is a siamese network, which comprises of three main sub-networks: the feature extraction sub-network, feature augmentation sub-network and map generation sub-network.

1) FEATURE EXTRACTION

As illustrated in Fig.1, we adopt four convolutional layers to extract local features. VGG16 [40] network was trained on ImageNet and obtained excellent performance for classification tasks. Therefore, we utilize the first two convolutional layers of the pre-trained VGG16 to extract the extensive low-level features. However, the extracted features by VGG16 could be not appropriate for the image fusion task, due to the reason that VGG16 stems from the classification task. Hence, we employ the third convolutional layer for exploring features of multi-focus images to suit the image fusion task. It is worth noting that the input images are fed separately to the three layers, which does not take into account the relativity between multi-focus images. Inspired by this fact, we concatenate the feature maps of the third layer of each branch with those of other branches, obtaining concatenated feature maps in a different order for each branch. Subsequently, the fourth convolutional layer is designed to extract the relative features by convoluting

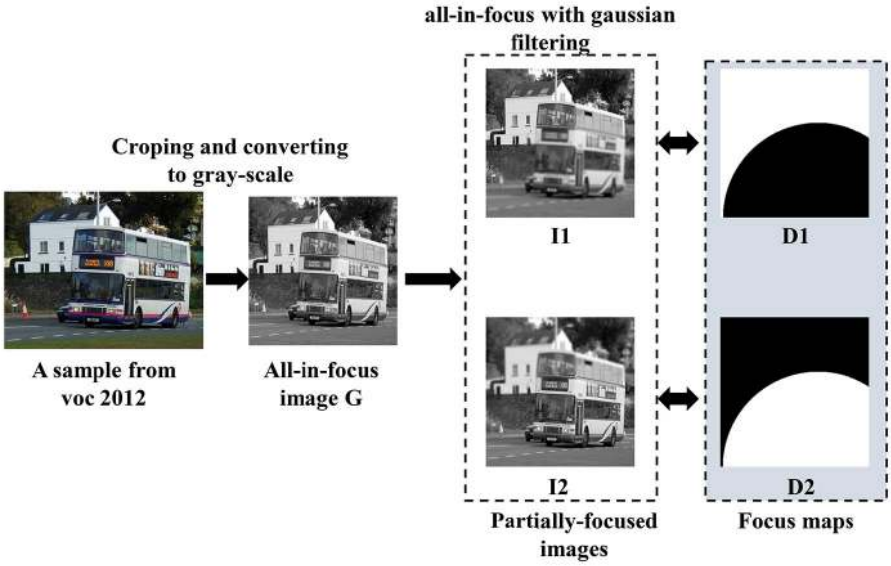


FIGURE 2. Schematic diagram of generating the training dataset.

concatenated feature maps. Note that the number of convolutional kernels could influence the discrimination capacity of features of images. A smaller number of kernels has lower computational complexity, however, it could not extract sufficient features. In contrast, a larger number of kernels could provide more extensive features, whereas it reduces computational efficiency. Therefore, we empirically set the number of kernels as 64 to achieve a trade-off between computational efficiency and the feature extraction.

2) FEATURE AUGMENTATION

Besides the local features of multi-focus images, we further need to augment the representation for each pixel of images using global contextual features. Based on the above analysis, The non-local module can be efficiently implemented using an RNN that is based on the Reseg [41] for semantic segmentation tasks. Next, we modify and extend it to perform the multi-focus image fusion task.

As shown in Fig.1, the non-local module contains four RNN layers consisting of two horizontal and vertical RNN layers, which scan along the horizontal and vertical directions of an image respectively, aiming at learning long range pixel dependencies. Considering that more layers will bring a higher burden of computational efficiency, in our work, four RNN layers are sufficient to obtain robust features. Specifically, let $X = \{x_{i,j}\}$ be the feature map from the feature extraction sub-network, where $x \in \mathbb{R}^{h\times w\times c}$ and h, w, c are the height, width, number of features, respectively. In our work, the horizontal RNN layer sweeps horizontally using two RNNs f_l and f_r, which move left-right and right-left respectively. Each RNN works along each row of X, receiving input from both a pixel $x_{i,j}$ and its hidden state, where i, j are the coordinates in the matrix of features. We choose an independent pixel instead of a patch as the input, mainly since the patch size could influence fused performance, especially when both focused and defocused information exist in one patch. At every time step, each RNN inputs the next pixel and its hidden state, outputs $o^*_{i,j}$ and updates its state $z^*_{i,j}$. This process of the horizontal RNN layer can be expressed as follows:

$$o^l_{i,j} = f_l(z^l_{i,j-1}, x_{i,j}), \quad for\ i = 1, \ldots, w \tag{1}$$
$$o^r_{i,j} = f_r(z^r_{i,j+1}, x_{i,j}), \quad for\ i = w, \ldots, 1 \tag{2}$$

where f_l and f_r return the activation of the recurrent hidden state, which contain N Gated Recurrent units [42] (25 units in our work), respectively. Based on the above process for all pixels of the input X, we can obtain a composite feature map $O_h = \{o^h_{i,j}\}$ by concatenating $o^l_{i,j}$ and $o^r_{i,j}$, where $o^h_{i,j} \in \mathbb{R}^{h\times w\times 2N}$.

After the horizontal sweep, we present the vertical RNN layer that integrates the features of the input image to perform global contextual information. Similar to the horizontal sweep, the vertical RNN layer uses two new RNNs (top-down and bottom-up) to work along each column of the feature map O_h from horizontal sweep, generating the feature map $O_v = \{o^v_{i,j}\}$, where $o^v_{i,j} \in \mathbb{R}^{h\times w\times 2N}$. Through the joint process of the horizontal and vertical RNN layers, our non-local module transforms the local representation of the image to the global representation.

3) MAP GENERATION

Having local and global contextual representations of multi-focus images, we apply a non-linearity function to the output of the last convolutional layer of size 1×1 S, generating the confidence map M. The confidence map is computed as

$$M_r = exp(S_r)/(exp(S_1) + exp(S_2)), \quad r = 1, 2 \tag{3}$$

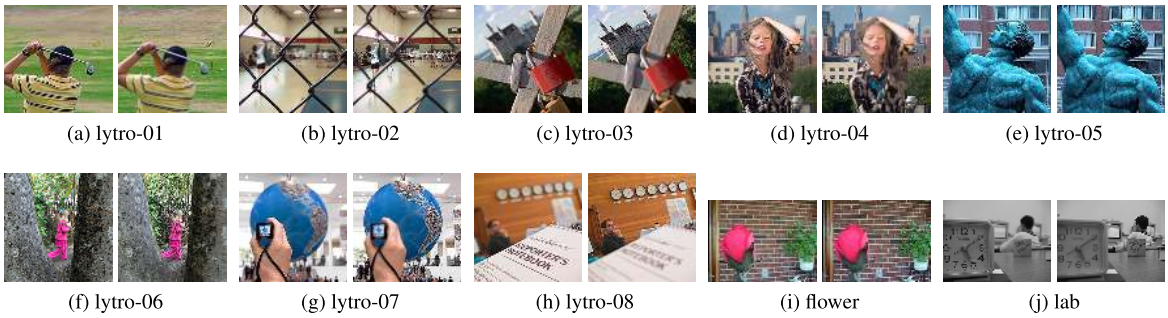


FIGURE 3. A portion of test images used in our experiments.

The generated map matrix ranges from 0 to 1. It estimates the focus degree of each pixel and 0, 1 imply defocused and focused information, respectively. As shown in Fig.1, the generated maps are supervised during training.

In the testing phase, we adopt a "choose-max" strategy to confidence maps for generating initial maps as below:

$$T_1(i,j) = \begin{cases} 1 & M_1(i,j) > M_2(i,j) \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where i and j are the coordinates in the matrix of the image.

$$T_2 = 1 - T_1 \tag{5}$$

Finally, the fused image is calculated with the following pixel weighted average rule.

$$F_t = I_1 T_1 + I_2 T_2 \tag{6}$$

B. LOSS FUNCTION

In order to conveniently describe the loss function, we assume there are two input images that are fed to our model. Given the training samples $\{\{I_1^n, D_1^n\}, \{I_2^n, D_2^n\}\}_{n=1}^N$, containing the training image $\{I_1^n, I_2^n\}$ and its ground-truth focus map $\{D_1^n, D_2^n\}$ from our training dataset, we compute the mean square error (MSE) between the generated focus map $\{M_1^n, M_2^n\}$ and the ground-truth $\{D_1^n, D_2^n\}$ as follows:

$$Loss_m = \frac{1}{Nhw} \sum_{n=1}^{N} ||M_1^n - D_1^n||^2 + ||M_2^n - D_2^n||^2 \tag{7}$$

where N is batch size. The above map loss can guide the proposed network to learn effectively focused and defocused features of images by the ground-truth. However, only utilizing the map loss to train the proposed network could limit the fusion effect, due to lack of the ground-truth fused image. For this problem, we add a pixel-wise regression loss that supervises the fused image F^n with the ground-truth fused image G^n, aiming at penalizing the map loss. The pixel-wise supervision is more proper than patch-wise supervision for avoiding the interference of texture information that is usually beneficial for the image classification task rather than the fusion task. Similarly, the MSE is adopted to compute the regression loss. The F^n can be directly generated by considering the pixel weighted average rule, that is

$$F^n = I_1^n M_1^n + I_2^n M_2^n \tag{8}$$

The regression loss is computed as

$$Loss_f = \frac{1}{Nhw} \sum_{n=1}^{N} ||F^n - G^n||^2 \tag{9}$$

Overall, our loss function consists of the map loss and regression loss, that is

$$Loss = Loss_m + Loss_f \tag{10}$$

As shown in Fig.1, we use $Loss$ as our final objective function to train our proposed model, which can generate better fusion results.

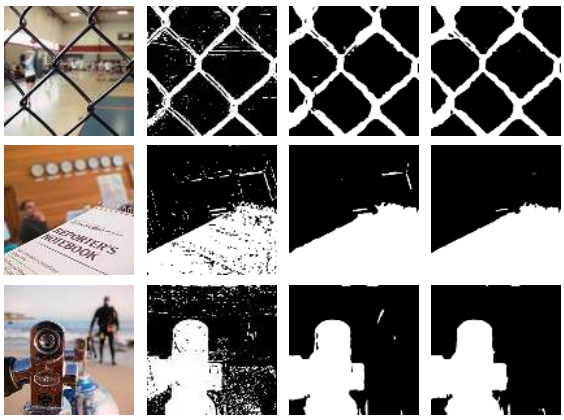
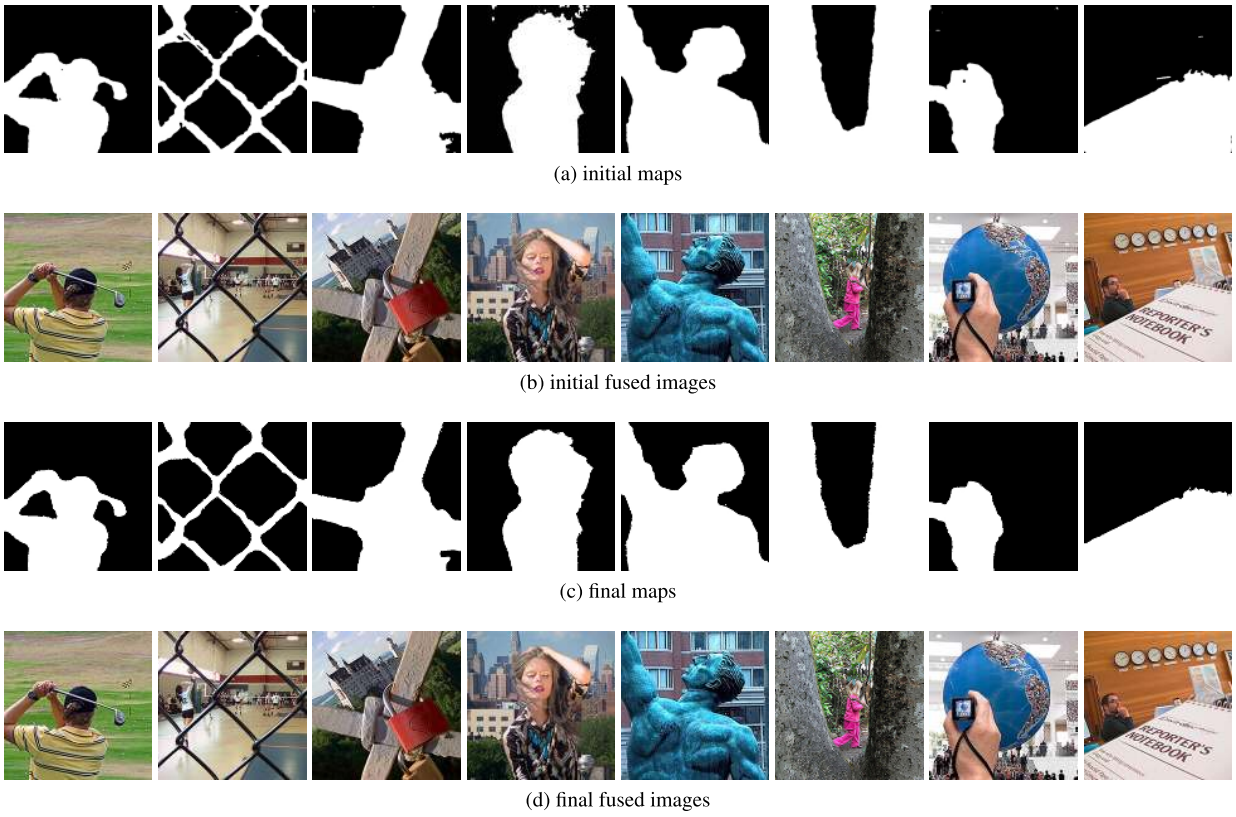


FIGURE 4. This is an ablation showing the visual effects of adding non-local module and the regression loss. The last three columns show the initial maps of the baseline model without the non-local module and the regression loss, the non-local model without the regression loss (B), and the final proposed model of this paper with both non-local module and the regression loss, respectively.

C. TRAINING

Our non-local model belongs to the image-to-image model, which requires substantial multi-focus images and



(a) initial maps

(b) initial fused images

(c) final maps

(d) final fused images

FIGURE 5. A comparison of initial and final segmentation maps (with and without post-processing) and corresponding fused images of the proposed method.

TABLE 1. This is an ablation showing the objective assessments of adding non-local module and the regression loss. The best results are shown in red color.

Source images	Indices	baseline	non-local model (B)	Proposed*
fence	MI	9.3989	9.2875	9.3213
	$Q^{AB/F}$	0.7182	0.7227	0.7229
	SF	18.2161	17.9598	17.9089
	$MSSIM$	0.9695	0.9703	0.9706
notebook	MI	8.5835	8.7319	8.7606
	$Q^{AB/F}$	0.7779	0.7822	0.7818
	SF	17.6372	17.4884	17.4453
	$MSSIM$	0.9796	0.9912	0.9925
diver	MI	9.0233	9.1518	9.1633
	$Q^{AB/F}$	0.7444	0.7534	0.7526
	SF	16.6834	16.6352	16.5618
	$MSSIM$	0.9762	0.9855	0.9867

all-in-focus images for training. To obtain the training dataset, an intuitive way is to create the partially-focused image pairs. We generated the training data using VOC 2012 JPEGImages [53] which consists of 17125 high-quality nature images. All the images were randomly cropped and their resulting size was 300×300.

As shown in Fig.2, we assume that the source image is the all-in-focus image G and a pair of multi-focus images is generated according to it. We randomly select one part of the all-in-focus to blur with Gaussian filtering, while the rest part does not do anything, generating a partially-focused image (such as $I1$ in Fig.2) and a focus map with the values of 0 and 1 in corresponding parts ($D1$ in Fig.2). To obtain the other one partially-focused image and its focus map of that pair, the similar process is applied to the all-in-focus image but on converse parts as shown in Fig.2 $I2$, $D2$. Therefore, we can generate thousands of multi-focus images and their focus maps by applying the above process to the source images. It is worth noting that the assumed all-in-focus images and generated multi-focus images are reasonable for training, since the focused region of a partially-focused image and its defocused region in the other one partially-focused image are simultaneously looked by our fusion network. During the process of creating the training dataset, our source images are randomly passed through different Gaussian filters with the kernel size range of 3 to 15. The standard deviation of Gaussian filters is computed from Gaussian kernels.

Our method is implemented based on PyTorch [51] framework and trained on a GeForce-GTX1080-8GB GPU. The Adam [43] is used to optimize the entire network and the learning rate 0.00001 is set constantly during the training

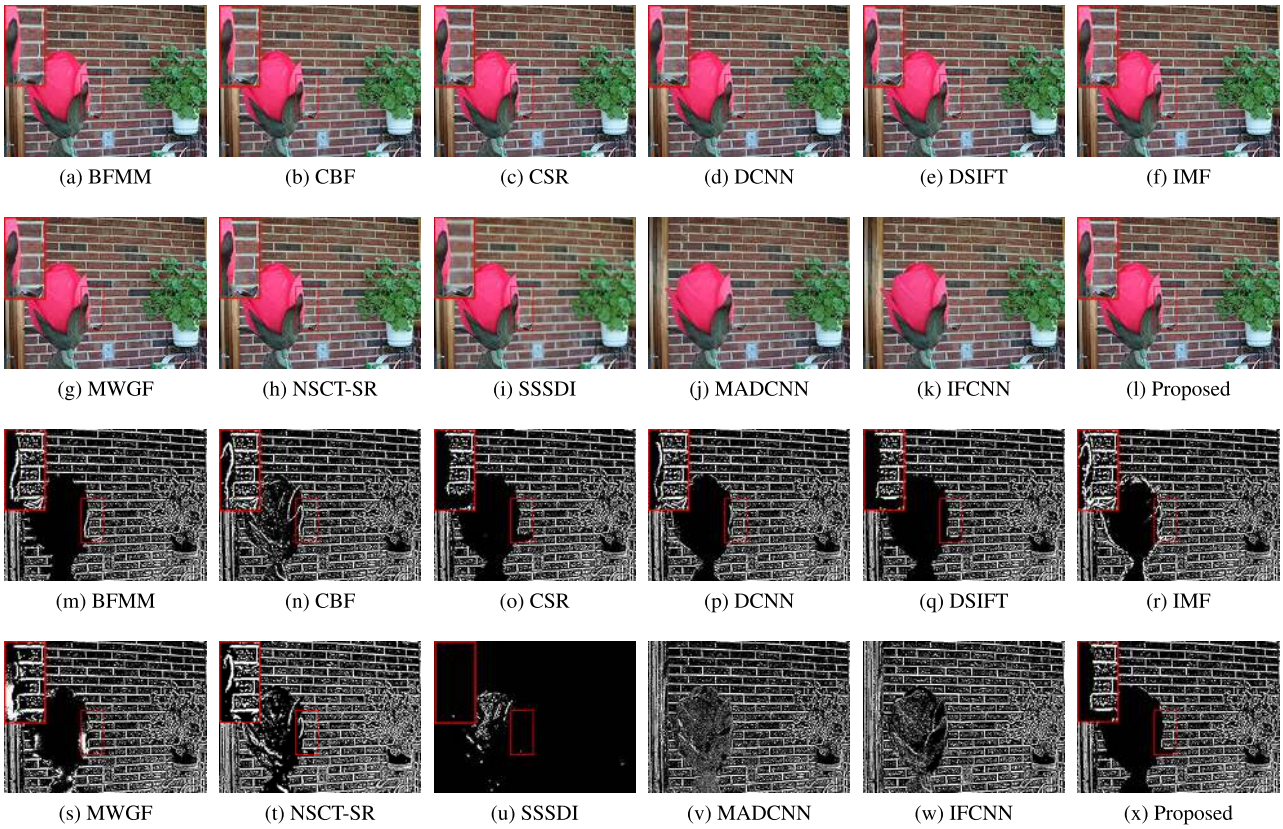


FIGURE 6. (a)-(l) Fusion results of "flower" obtained by different methods. (m)-(x) difference images obtained by different methods.

process. We use 17000 pairs of multi-focus images, iterating about 30 epochs to train our model. The batch size is 4.

III. EXPERIMENTS

A. EXPERIMENTAL SETTING

In this section, we use 23 pairs of multi-focus images from the "Lytro" dataset [15] and other famous test multi-focus images such as "flower" and "lab" [52] for evaluation. A portion of the test images is shown in Fig.3. The proposed fusion method is compared with eleven state-of-the-art multi-focus image fusion methods, including the nonsubsampled contourlet transform and sparse representation (NSCT-SR) [8], the multi-scale weight gradient fusion (MWGF) [44], the image matting (IMF) [18], the dense SIFT (DSIFT) [25], the convolutional neural network (CNN) [27], self-similarity and depth information (SSSDI) [26], boundary finding (BFMM) [45], cross bilateral filter (CBF) [46], the convolutional sparse representation (CSR) [16], the multi-scale convolutional neural network (MADCNN) [35] and IFCNN [37]. In addition, in order to compare the proposed method with the eleven previous methods in fair condition, the default parameters of these methods are set to keep consistent with their original papers.

In order to evaluate the performance of different algorithms objectively, four metrics are adopted: mutual information (MI) [47], edge information preservation ($Q^{AB/F}$) [48], modified structural similarity metric($MSSIM$) [49] and spatial frequency (SF) [50]. MI can be used to reflect fused effects by measuring the similarity of global statistical characteristics on the gray level (via grayscale histograms) between input images and fused images. $Q^{AB/F}$ is a measure of fusion performance to evaluate the amount of edge information that is transferred from input images to the fused image. The greater the value of $Q^{AB/F}$ is, the better the fusion effect is. The $MSSIM$ is used to measure the degree of structure preservation by considering three aspects: structure, luminance, and contrast. SF indicates the overall activity level in an image. The larger the value of SF is, the better the fusion image is.

B. ABLATION STUDY

In this subsection, we conduct ablation experiments to analyze the influence of the non-local module and the regression loss. We design two different network architectures based on the proposed non-local network architecture for detailed comparisons: a baseline model without both non-local module and the regression loss (A), a non-local network architecture without the regression loss (B). In Fig.4, we can obviously observe that the maps of the non-local model (B) are better than those of the baseline model. It is indicated that the non-local module can improve visual perception results. However, the maps of the non-local model without using the regression

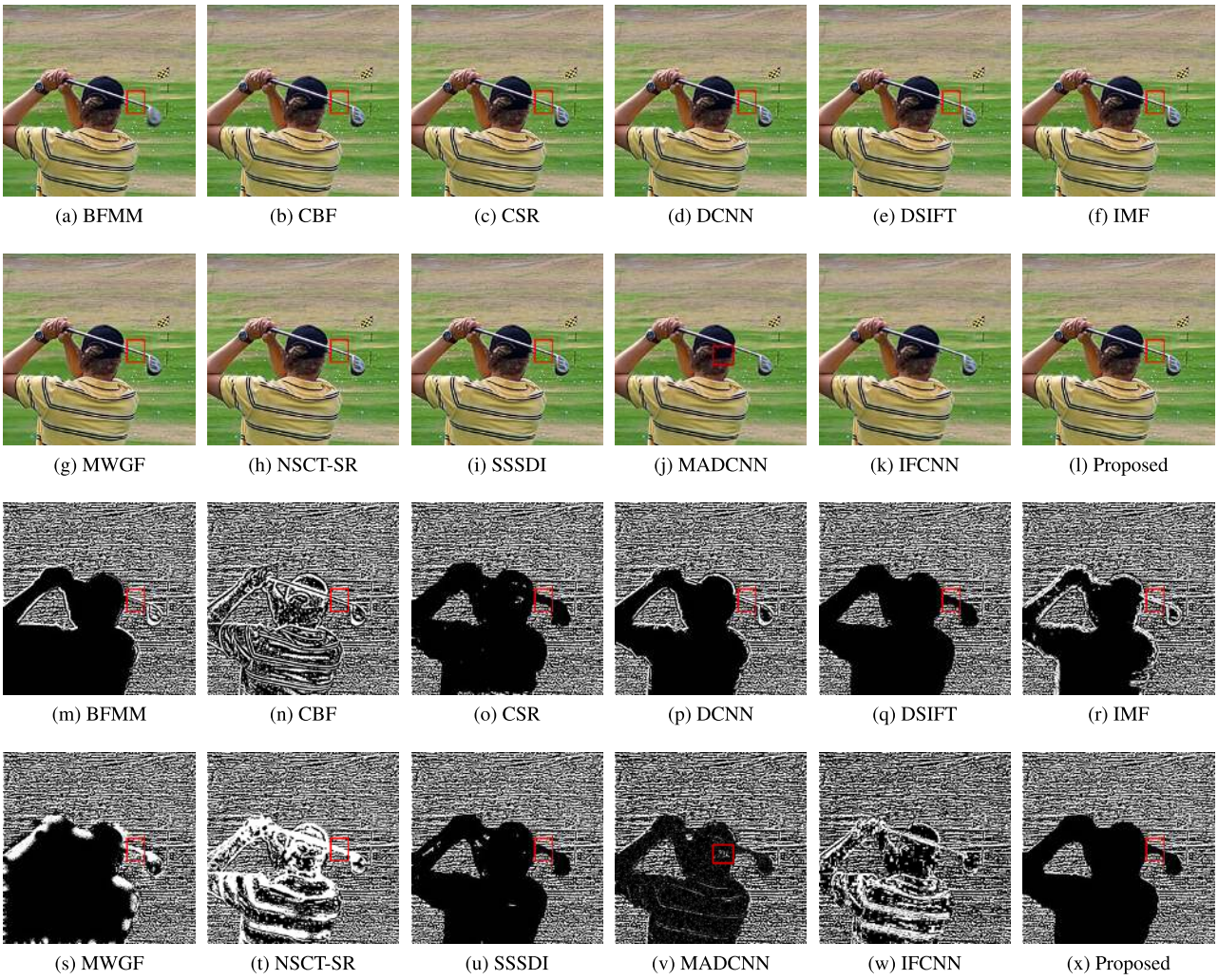


FIGURE 7. (a)-(l) Fusion results of "lytro-01" obtained by different methods. (m)-(x) difference images obtained by different methods.

loss (B) show more errors compared with the maps of the final proposed non-local model using the regression loss, which implies our model can obtain benefits by using the regression loss. Furthermore, comparing the objective results of the two network architectures and proposed model listed in Table.1, we can easily find that the objective results of the non-local model (B) are higher than those of the baseline model but are lower than those of the final proposed non-local model in most cases. From the analyses based on visual perception and objective metrics, adding non-local module and the regression loss achieved a significant improvement in fusion performance.

C. QUALITATIVE EVALUATION

First, we evaluate the effectiveness of the proposed multi-focus image fusion method by utilizing visual quality. As shown in Fig.5, there are two types of segmented maps, initial maps which are obtained by applying the "choose-max" strategy to immediate results of the proposed method and final maps which are obtained by applying some post-processing (such as small region removal and guided filtering) to the immediate results. Compared with final maps, the initial maps are already very accurate that most pixels are correctly classified, which shows the good capacity of augmenting features of our proposed model.

To further verify the effectiveness of the proposed method, we visualize difference images and fused images of four pairs test images obtained by different algorithms in Fig.6, Fig.7, Fig.8 and Fig.9. The difference images are produced by subtracting the first source image from the fusion images. The values of each difference image are normalized to the range of 0 to 1. When the focused region is completely detected, the difference image will not show any information of that. From Fig.6 (n), (t), (u), (v), (w), it's obvious that the CBF, NSCT-SR, MADCNN, IFCNN show undesirable artifacts and the SSSDI fails to detect the focused regions. In Fig.6 (r), (s), the pixels around the boundary regions between focused and defocused are white, illustrating the edge details are weakened. Besides, as shown in the red rectangular region of Fig.6 (m), (p), the edge of focused regions also contains

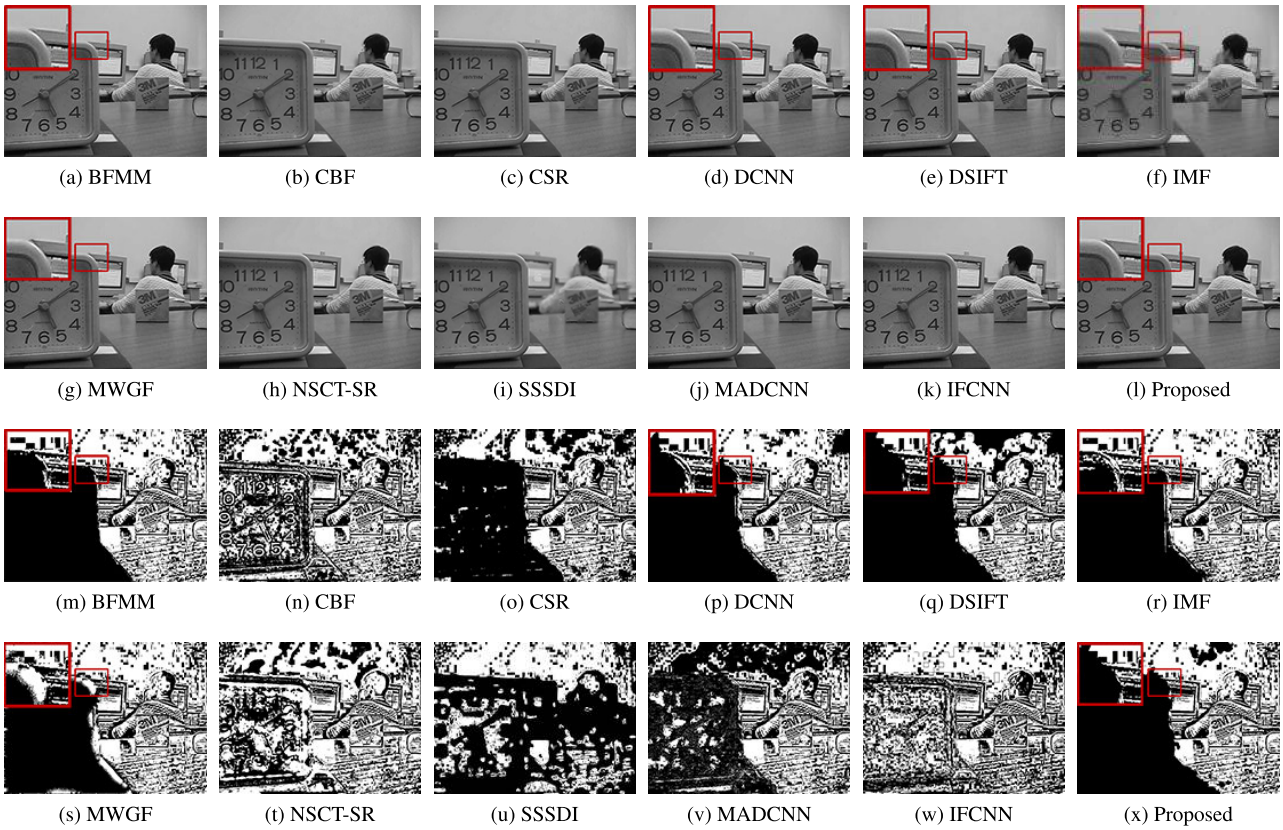


FIGURE 8. (a)-(l) Fusion results of "lab" obtained by different methods. (m)-(x) difference images obtained by different methods.

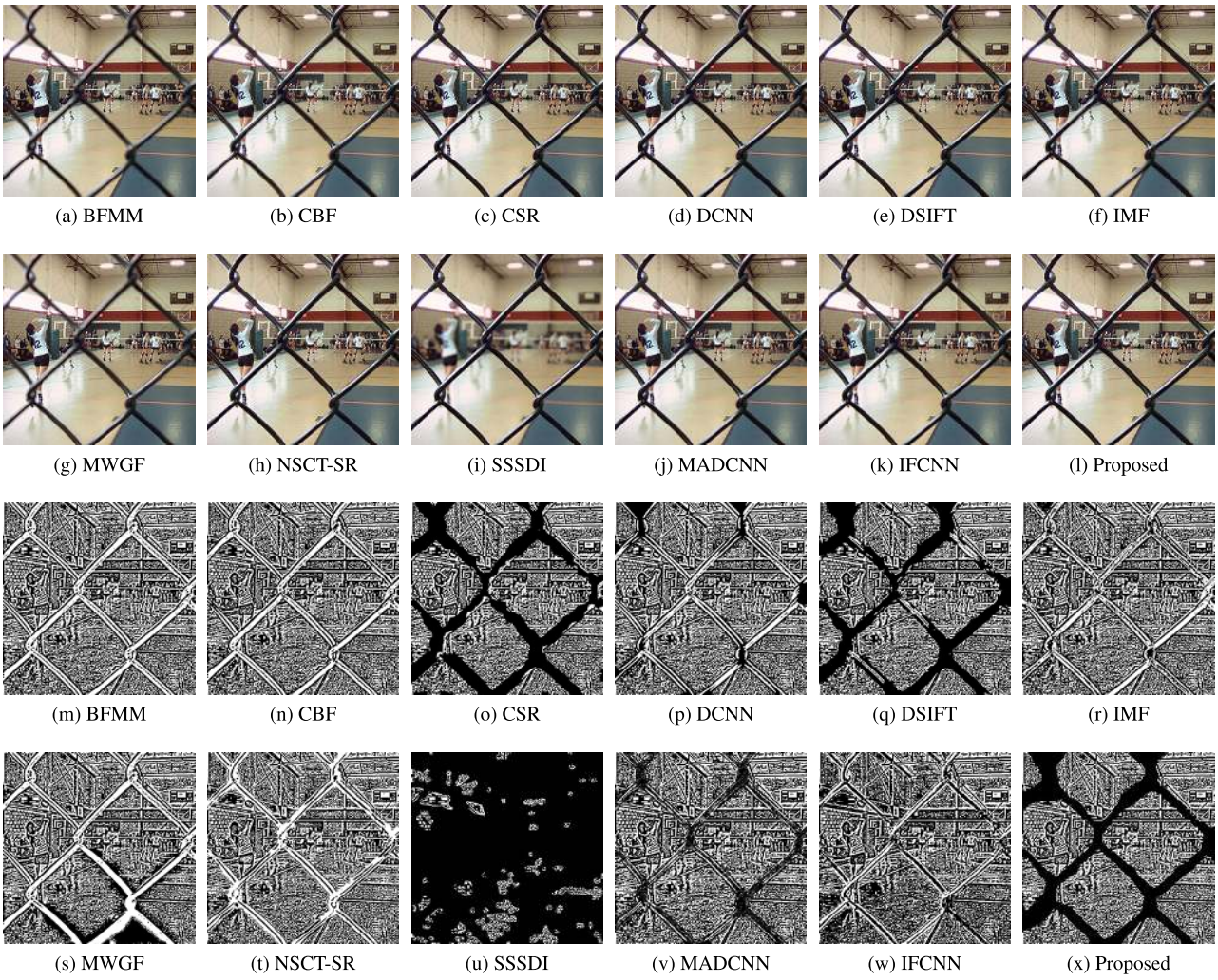
some artifacts. The CSR detects well the focused regions except for causing the coarse contour edge. Therefore, we can see that the DSIFT and proposed method perform well on the edge of focused and defocused regions and accurately detect focused regions.

In Fig.7, results similar to the observation above, the CBF (n), NSCT-SR (t), IFCNN (w) still show many obvious artifacts and the BFMM (m), IMF (r), MWGF (s), DCNN (p) make border regions of near focused regions blurry. In addition, the BFMM and MWGF incorrectly detect focused regions in the bottom left corner. Compared with the CSR and DSIFT, the SSSDI and proposed method have more accurate contour of near focused regions. However, we can see that Fig.7 (u) contains white holes inside the man, meaning the SSSDI produces some defocused pixels in focused regions. Fig.7 (v) also presents some undesirable noise-like pixels inside the man, which means the fusion result loss some information. Through the further comparative experiment, the proposed method can efficiently preserve the edges and reduce artifacts.

In Fig.8, it is obvious that the difference images from CBF, CSR, NSCT-SR, SSSDI, MADCNN and IFCNN show abundant residuals in the "clock" region, indicating many details of fusion results are lost. Observing the magnified red rectangular regions from BFMM, DCNN, IMF and MWGF, we can find that these methods incur some undesirable artifacts in the top right corner of the "clock". Besides, compared with DSIFT, the proposed method presents more accurate contour between focused and defocused regions. The above comparisons are enough to prove that the proposed method has a strong ability for fusing multi-focus gray images. To make a more challenging comparison, we perform the experiment on "lytro-02". In Fig.9, the focused regions are the fence. Obviously, all methods show artifacts in focused regions except the proposed method. These experiments above demonstrate that the proposed method reached a better visual perception result.

D. QUANTITATIVE EVALUATION

To further verify the performance of the proposed method, the twelve algorithms perform on 22 pairs multi-focus images and MI, QAB/F, MSSIM and SF are used to quantitatively analyze the fusion results. We have listed the average scores for the various methods on the four metrics shown in Table.2. Each red, blue, cyan value in this table are the maximum, second maximum, third maximum in that corresponding row, respectively. In overall comparison, values in the table indicate that the proposed method shows better results than the other eleven methods. The proposed method, without any post-processing algorithms, is better than those of the others with or without post-processing algorithms. This is mainly because the proposed method could accurately find



(a) BFMM (b) CBF (c) CSR (d) DCNN (e) DSIFT (f) IMF

(g) MWGF (h) NSCT-SR (i) SSSDI (j) MADCNN (k) IFCNN (l) Proposed

(m) BFMM (n) CBF (o) CSR (p) DCNN (q) DSIFT (r) IMF

(s) MWGF (t) NSCT-SR (u) SSSDI (v) MADCNN (w) IFCNN (x) Proposed

FIGURE 9. (a)-(l) Fusion results of "lytro-02" obtained by different methods. (m)-(x) difference images obtained by different methods.

TABLE 2. Comparison of objective quality metrics of our proposed multi-focus image fusion method and the others. The column "Proposed*" and "Proposed" indicate the proposed method without and with any post-processing algorithms, respectively.

Indices	NSCT-SR	MWGF	DSIFT	DCNN	SSSDI	CSR	IMF	BFMM	CBF	MADCNN	IFCNN	Proposed*	Proposed
MI	7.0991	8.1564	8.8341	8.5454	8.9617	8.8034	8.4846	8.8030	7.5106	7.4856	6.4995	8.8428	8.8527
$Q^{AB/F}$	0.7436	0.7437	0.7578	0.7580	0.6950	0.7548	0.7497	0.7533	0.7509	0.7387	0.7112	0.7586	0.7580
SF	19.3177	19.0810	19.4790	19.2370	17.5254	19.4553	19.4565	19.1858	18.4217	19.4331	19.4535	19.5066	19.4904
$MSSIM$	0.9703	0.9779	0.9798	0.9806	0.9712	0.9776	0.9746	0.9802	0.9753	0.9750	0.9552	0.9793	0.9801

boundaries and detect focused regions. Besides, the proposed method with post-processing algorithms has more better consistency in focused regions, due to the processing of filtering, although a little border information of focused regions is lost as shown from $Q^{AB/F}$ metric.

Considering the above comparisons based on visual perception and evaluation metrics together, our proposed fusion method outperforms the other eleven state-of-the-art algorithms for multi-focus image fusion.

E. TIME COST COMPARISON

To evaluate the computational efficiency, we calculate and list the fusion time for producing one fusion image of size 520 × 520 from two input images in Table.3. Experiments

TABLE 3. Time cost comparison. The column "Proposed*" indicates the proposed method without any post-processing algorithms (Image size: 520 × 520, Time unit: second).

Algorithms	NSCT-SR	MWGF	DSIFT	DCNN	SSSDI	CSR	IMF	BFMM	CBF	MADCNN	IFCNN	Proposed*
Time Cost	98.46	5.16	4.48	90.44	43.28	74.43	3.01	1.18	12.76	5.19	2.58	11.97

are conducted on the platform with Intel Core i7-8700k CPU, 3.2GHz dominant frequency and 32G memory. Table.3 shows that our proposed algorithm is remarkably faster than NSCT-SR, DCNN, SSSDI, CSR and CBF, but is slightly slower than MWGF, DSIFT, IMF, BFMM, MADCNN and IFCNN. It means that the proposed algorithm can meet the requirement of many applications. Moreover, the running time of the proposed algorithm can be notably curtailed by running it on a GPU. Therefore, it is promising to employ our image fusion algorithm in practice.

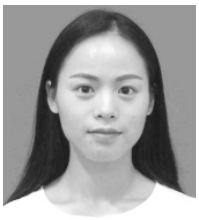
IV. CONCLUSION

In this work, we have proposed a non-local deep network architecture for multi-focus image fusion. It learns local and non-local self-similarity to achieve superior performance. We introduced RNN that retrieves long distance pixel dependencies to perform the non-local processing. To the best of our knowledge, the proposed method is first ever non-local deep learning method based on the RNN to perform multi-focus image fusion. In addition, we designed a regression loss to avoid the interference of texture information. Experimental results show that the proposed method can produce high-quality fusion results and outperform the state-of-the-art methods, both qualitatively and quantitatively. Currently, our non-local network performs well in fusing multi-focus images. For future research, an interesting direction is to explore some necessary modifications on the design of our current non-local model that allow it to be efficiently transplanted to multi-exposure image fusion and other image fusion applications.

REFERENCES

[1] P. Burt and E. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, no. 4, pp. 532–540, Apr. 1983.
[2] A. Toet, "A morphological pyramidal image decomposition," *Pattern Recognit. Lett.*, vol. 9, no. 4, pp. 255–261, May 1989.
[3] Y. Yang, "A novel DWT based multi-focus image fusion method," *Procedia Eng.*, vol. 24, pp. 177–181, Dec. 2011.
[4] J. J. Lewis, R. J. O'Callaghan, S. G. Nikolov, D. R. Bull, and N. Canagarajah, "Pixel- and region-based image fusion with complex wavelets," *Inf. Fusion*, vol. 8, no. 2, pp. 119–130, Apr. 2007.
[5] S. Li and B. Yang, "Multifocus image fusion by combining curvelet and wavelet transform," *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1295–1301, Jul. 2008.
[6] X.-B. Qu, G.-F. Xie, J.-W. Yan, Z.-Q. Zhu, and B.-G. Chen, "Image fusion algorithm based on neighbors and cousins information in nonsubsampled contourlet transform domain," in *Proc. Int. Conf. Wavelet Anal. Pattern Recognit.*, 2007, pp. 1797–1802.
[7] Z. Zhu, M. Zheng, G. Qi, D. Wang, and Y. Xiang, "A phase congruency and local Laplacian energy based multi-modality medical image fusion method in NSCT domain," *IEEE Access*, vol. 7, pp. 20811–20824, Feb. 2019.
[8] Y. Liu, S. Liu, and Z. Wang, "A general framework for image fusion based on multi-scale transform and sparse representation," *Inf. Fusion*, vol. 24, pp. 147–164, Jul. 2015.
[9] H. Li, X. Li, Z. Yu, and C. Mao, "Multifocus image fusion by combining with mixed-order structure tensors and multiscale neighborhood," *Inf. Sci.*, vols. 349–350, pp. 25–49, Jul. 2016.
[10] Y. Liu and Z. Wang, "Simultaneous image fusion and denoising with adaptive sparse representation," *IET Image Process.*, vol. 9, no. 5, pp. 347–357, May 2015.
[11] M. Kim, D. K. Han, and H. Ko, "Joint patch clustering-based dictionary learning for multimodal image fusion," *Inf. Fusion*, vol. 27, pp. 198–214, Jan. 2016.
[12] Q. Zhang and M. D. Levine, "Robust multi-focus image fusion using multi-task sparse representation and spatial context," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 2045–2058, May 2016.
[13] B. Yang and S. Li, "Visual attention guided image fusion with sparse representation," *Optik*, vol. 125, no. 17, pp. 4881–4888, Sep. 2014.
[14] H. Li, L. Li, and J. Zhang, "Multi-focus image fusion based on sparse feature matrix decomposition and morphological filtering," *Opt. Commun.*, vol. 342, pp. 1–11, May 2015.
[15] M. Nejati, S. Samavi, and S. Shirani, "Multi-focus image fusion using dictionary-based sparse representation," *Inf. Fusion*, vol. 25, pp. 72–84, Sep. 2015.
[16] Y. Liu, X. Chen, R. K. Ward, and Z. Jane Wang, "Image fusion with convolutional sparse representation," *IEEE Signal Process. Lett.*, vol. 23, no. 12, pp. 1882–1886, Dec. 2016.
[17] Z. Zhu, H. Yin, Y. Chai, Y. Li, and G. Qi, "A novel multi-modality image fusion method based on image decomposition and sparse representation," *Inf. Sci.*, vol. 432, pp. 516–529, Mar. 2018.
[18] S. Li, X. Kang, J. Hu, and B. Yang, "Image matting for fusion of multi-focus images in dynamic scenes," *Inf. Fusion*, vol. 14, no. 2, pp. 147–162, Apr. 2013.
[19] W. Zhang, W.-K. Cham, "Gradient-directed multiexposure composition," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2318–2323, Apr. 2012.
[20] S. Li and B. Yang, "Multifocus image fusion using region segmentation and spatial frequency," *Image Vis. Comput.*, vol. 26, no. 7, pp. 971–979, Jul. 2008.
[21] B. Gu, W. Li, J. Wong, M. Zhu, and M. Wang, "Gradient field multi-exposure images fusion for high dynamic range image visualization," *J. Vis. Commun. Image Represent.*, vol. 23, no. 4, pp. 604–610, May 2012.
[22] S. Li, J. T. Kwok, and Y. Wang, "Combination of images with diverse focuses using the spatial frequency," *Inf. Fusion*, vol. 2, no. 3, pp. 169–176, Sep. 2001.
[23] X. Bai, Y. Zhang, F. Zhou, and B. Xue, "Quadtree-based multi-focus image fusion using a weighted focus-measure," *Inf. Fusion*, vol. 22, pp. 105–118, Mar. 2015.
[24] S. Li, X. Kang, and J. Hu, "Image fusion with guided filtering," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2864–2875, Jul. 2013.
[25] Y. Liu, S. Liu, and Z. Wang, "Multi-focus image fusion with dense SIFT," *Inf. Fusion*, vol. 23, pp. 139–155, May 2015.
[26] D. Guo, J. Yan, and X. Qu, "High quality multi-focus image fusion using self-similarity and depth information," *Opt. Commun.*, vol. 338, pp. 138–144, Mar. 2015.
[27] Y. Liu, X. Chen, H. Peng, and Z. Wang, "Multi-focus image fusion with a deep convolutional neural network," *Inf. Fusion*, vol. 36, pp. 191–207, Jul. 2017.
[28] C.-B. Du and S.-S. Gao, "Multi-focus image fusion with the all convolutional neural network," *Optoelectron. Lett.*, vol. 14, no. 1, pp. 71–75, Jan. 2018.
[29] M. Amin-Naji, A. Aghagolzadeh, and M. Ezoji, "CNNs hard voting for multi-focus image fusion," *J. Ambient Intell. Hum. Comput.*, vol. 11, pp. 1749–1769, Apr. 2020.
[30] H. Tang, B. Xiao, W. Li, and G. Wang, "Pixel convolutional neural network for multi-focus image fusion," *Inf. Sci.*, vols. 433–434, pp. 125–141, Apr. 2018.
[31] Y. Yang, Z. Nie, S. Huang, P. Lin, and J. Wu, "Multilevel features convolutional neural network for multifocus image fusion," *IEEE Trans. Comput. Imag.*, vol. 5, no. 2, pp. 262–273, Jun. 2019.

[32] C. Du and S. Gao, "Image segmentation-based multi-focus image fusion through multi-scale convolutional neural network," *IEEE Access*, vol. 5, pp. 15750–15761, 2017.
[33] H. Ma, J. Zhang, S. Liu, and Q. Liao, "Boundary aware multi-focus image fusion using deep neural network," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1150–1155.
[34] W. Zhao, D. Wang, and H. Lu, "Multi-focus image fusion with a natural enhancement via a joint multi-level deeply supervised convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 4, pp. 1102–1115, Apr. 2019.
[35] R. Lai, Y. Li, J. Guan, and A. Xiong, "Multi-scale visual attention deep convolutional neural network for multi-focus image fusion," *IEEE Access*, vol. 7, pp. 114385–114399, Aug. 2019.
[36] X. Guo, R. Nie, J. Cao, D. Zhou, L. Mei, and K. He, "FuseGAN: Learning to fuse multi-focus image via conditional generative adversarial network," *IEEE Trans. Multimedia*, vol. 21, no. 8, pp. 1982–1996, Aug. 2019.
[37] Y. Zhang, Y. Liu, P. Sun, H. Yan, X. Zhao, and L. Zhang, "IFCNN: A general image fusion framework based on convolutional neural network," *Inf. Fusion*, vol. 54, pp. 99–118, Feb. 2020.
[38] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
[39] L. Ran, Y. Zhang, and G. Hua, "CANNET: Context aware nonlocal convolutional networks for semantic image segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Quebec City, QC, Canada, Sep. 2015, pp. 4669–4673.
[40] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–14.
[41] F. Visin, A. Romero, K. Cho, M. Matteucci, M. Ciccone, K. Kastner, Y. Bengio, and A. Courville, "ReSeg: A recurrent neural network-based model for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Las Vegas, NV, USA, Jun. 2016, pp. 426–433.
[42] K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder–decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1–15.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–15.
[44] Z. Zhou, S. Li, and B. Wang, "Multi-scale weighted gradient-based fusion for multi-focus images," *Inf. Fusion*, vol. 20, pp. 60–72, Nov. 2014.
[45] Y. Zhang, X. Bai, and T. Wang, "Boundary finding based multi-focus image fusion through multi-scale morphological focus-measure," *Inf. Fusion*, vol. 35, pp. 81–101, May 2017.
[46] B. K. S. Kumar, "Image fusion based on pixel significance using cross bilateral filter," *Signal, Image Video Process.*, vol. 9, no. 5, pp. 1193–1204, 2015.
[47] G. Qu, D. Zhang, and P. Yan, "Information measure for performance of image fusion," *Electron. Lett.*, vol. 38, no. 7, pp. 313–315, Mar. 2002.
[48] V. Petrovic and C. Xydeas, "Objective image fusion performance characterisation," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, vol. 1, Oct. 2005, pp. 1866–1871.
[49] C. Yang, J.-Q. Zhang, X.-R. Wang, and X. Liu, "A novel similarity based quality metric for image fusion," *Inf. Fusion*, vol. 9, no. 2, pp. 156–160, Apr. 2008.
[50] A. M. Eskicioglu and P. S. Fisher, "Image quality measures and their performance," *IEEE Trans. Commun.*, vol. 43, no. 12, pp. 2959–2965, Dec. 1995.
[51] Pytorch. *Pytorch Framework*. Accessed: Jul. 4, 2019. [Online]. Available: https://pytorch.org
[52] *Multi-Focus Source Images*. Accessed: Jul. 23, 2019. [Online]. Available: https://github.com/yuliu316316/DSIFT-MFIF
[53] *VOC 2012 JPEGImages*. Accessed: Jul. 4, 2019. [Online]. Available: http://host.robots.ox.ac.uk/pascal/VOC/voc2012/VOCtrainval_11-May-2012.tar



ZHAO DUAN received the B.S. degree in electronic information science and technology from the Chengdu University of Technology, in 2017. She is currently pursuing the Ph.D. degree in computer science and technology with Chongqing University, Chongqing, China. Her research interests include image fusion and image processing.

TAIPING ZHANG (Member, IEEE) received the B.S. and M.S. degrees in computational mathematics, and the Ph.D. degree in computer science from Chongqing University, Chongqing, China, in 1999, 2001, and 2010, respectively. He is currently a Professor with the Department of Computer Science, Chongqing University, and a Visiting Research Fellow of the Faculty of Science and Technology, University of Macau. He has published extensively in the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS—PART B CYBERNETICS PUBLICATION INFORMATION, *Pattern Recognition*, and *Neurocomputing*. His research interests include pattern recognition, image processing, machine learning, and computational mathematics.

JIN TAN received the B.S. degree in telecommunication engineering and the M.S. degree in telecommunication and information systems from the University of Electronic Science and Technology of China, in 2005 and 2008, respectively. She is currently pursuing the Ph.D. degree in computer science with Chongqing University, Chongqing, China. She was a Visiting Scholar with The University of Texas at Dallas, Dallas, from 2015 to 2016. Her research interests include signal processing, pattern recognition, image processing, and machine learning.

XIAOLIU LUO received the B.S. degree in mathematics and applied mathematics from Ningxia University, in 2016. She is currently pursuing the Ph.D. degree in computer science with Chongqing University, Chongqing, China. Her research interests include pattern recognition, few-shot learning, graph learning, and machine learning.